Accuracy vs. Simplicity: A Complex Trade-Off*

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

January 2003

Abstract

Inductive learning aims at finding general rules that hold true in a database. Targeted learning seeks rules for the prediction of the value of a variable based on the values of others, as in the case of linear or non-parametric regression analysis. Non-targeted learning finds regularities without a specific prediction goal. We model the product of non-targeted learning as rules that state that a certain phenomenon never happens, or that certain conditions necessitate another. For all types of rules, there is a trade-off between the rule's accuracy and its simplicity. Thus rule selection can be viewed as a choice problem, among pairs of degree of accuracy and degree of complexity. However, one cannot in general tell what is the feasible set in the accuracy-complexity space. Formally, we show that finding out whether a point belongs to this set is computationally hard. In particular, in the context of linear regression, finding a small set of variables that obtain

*Earlier versions of this paper circulated under the title "From Cases to Rules: Induction and Regression." We thank Hal Cole, Joe Halpern, Bart Lipman, Yishay Mansour, and Nimrod Megiddo for conversations and references.

†Institut d'Anàlisi Econòmica, C.S.I.C. enriqueta.aragones@uab.es

‡Tel-Aviv University and Cowles Foundation, Yale University. Gilboa gratefully acknowledges support from the Israel Science Foundation. 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. schmeid@post.tau.ac.il

a certain value of R^2 is computationally hard. Computational complexity may explain why a person is not always aware of rules that, if asked, she would find valid. This, in turn, may explain why one can change other people's minds (opinions, beliefs) without providing new information.

"The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience." (Wittgenstein 1922, Proposition 6.363)

1 Motivation

Ann: "Russia is a dangerous country."

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. He did it by drawing her attention to "the democratic peace" phenomenon, sometimes attributed to Kant.[1] Observe, however, that Bob has not provided Ann with any new

[1] Kant (1795) wrote, "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." Ob-

factual information. Rather, Bob has pointed out a certain regularity in the cases that are known to both Ann and Bob. This regularity is, apparently, new to Ann. Yet, she had had all the factual information needed to observe it before meeting Bob. It simply did not occur to Ann to test the accuracy of the generalization suggested to her by Bob.

Much of human knowledge (and, indeed, all of mathematics) has to do with noticing facts and regularities that, in principle, could have been figured out based on existing knowledge, rather than with acquiring new information per se. Why do people fail to draw all the relevant conclusions from the information they possess? Sometimes the reason is that certain aspects of the observations they have simply do not occur to them. In the example above, it is quite possible that Ann never thought of the type of regime as an explanatory variable for the occurrence of wars. Sometimes a regularity involves a *combination* of several variables. For many real life prediction problems there are many potentially relevant variables. But the number of combinations of these variables is much larger than the number of variables (in fact, the latter increases in an exponential fashion as a function of the former), making it practically impossible to think of all possible regularities that might, once observed, prove true. A recent paper by La Porta, Lopez-de-Silanes, Shleifer, and Vishny (1998) on the quality of government states, "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." One level of difficulty is to come up with all the variables listed above as potential explanatory variables. A second level of difficulty is to consider sophisticated combinations as the one suggested in this paragraph.

In conclusion, there are many regularities that people do not notice, even though they might find them true once these regularities are pointed out to

serve that Kant wrote in favor of the republican constitution, rather than democracy per se. For recent documentations of this phenomenon, see Maoz and Russett (1992, 1993), Russett (1993), and Maoz (1998).

them. The main results of this paper attempt to explain this phenomenon more formally. Before we present these results, however, we attempt to convince the reader that the issue under discussion has more than anecdotal interest.

2 Belief Formation

Economic theory does not in general deal with the question of belief formation. Economic agents are assumed to have beliefs and to act on them, and these beliefs tend to take the form of a Bayesian prior probability. The literature offers axiomatizations of this paradigm, which may be used to elicit beliefs of agents who follow the axioms, but there is no description of a process by which one generates such beliefs.

There are several reasons for an interest inthe process of belief formation. Assume, first, that one takes a normative interpretation of expected utility theory. One may be convinced by Savage's (1954) axioms that one would like to be an expected utility maximizer. The next step would be to determinea utility function and a probability measure. At this step, one needsprobabilistic beliefs. Whereas *any* probability measure would do in order to be "rational" in the sense of satisfying Savage's axioms, one would normally like to choose a reasonable probability measure. Thus, one is faced with the question, what are reasonable beliefs to hold?

This question is also faced by organizations or teams, who attempt to aggregate beliefs of various individuals. Even if each of these individualshas a Bayesian prior, when these priors differ, the group mustdecide on a reasonable priorbased on their shared information.

There are also descriptive reasons to be interested in the process of belief formation. First, having a theory of this process might shed light on regularities in the prior probabilities that Bayesian agents entertain. One may be able to categorize economic problems by the type of information available in

them, and by the corresponding process by which the available information results in probabilistic beliefs. Second, a better understanding of this process can potentiallydelineate the scope of the Bayesian paradigm, and facilitate speculationabout the types of beliefs that people might have when they do not have a Bayesian prior.

The question of belief formation is beyond the scope of this paper, but the problem we deal with here can be viewed as a particular sub-problem of belief formation. Specifically, we ask how individual instances are generalized into rules.This inductive process is likely to be a part of any theory of belief formation. If the sun rises every day, we tend to believe that it would rise tomorrow. Naturally, only a small part of our beliefs can be attributed to such simple and clear regularities. Yet, slightly more complex and less accurate regularities also feed the belief formation process.

Moreover, there are many circumstances in which beliefs are captured byrules. For example, when one uses linear regression to generate predictions, one may be viewed as believingin the simple rule of the regression equation.[2] Organizations often have simple rules that may be viewed as reflecting the organization's beliefs. A credit card agency has to make daily decision on approval of potential card holders, where each such decision is an act that reflects beliefs regarding the applicants' financial credibility.But the agency is not free to select any probabilistic beliefs to guide its decisions. It may berestricted by organizational feasibility to rules that can be followed by many clerks, or by a software package. Furthermore, there are legal restrictions for which the agency has to be able to justify its decisions based on simple rules. (There may also belegal restrictions on the type of variables used in these rules.) Thus, to the extent that the agency's strategy reflects its beliefs, these beliefs are given by simple rules.

It follows that for some situations, the analysis of rules in this paper may be viewed as a simplistic theory of belief formation. More generally, we tend

[2]Indeed, Bray and Savin (1986) suggest to model agents' beliefs by linear regression.

to view the analysis of rules merely as a first step in any reasonable account of the process of belief formation. The motivating example above suggests that even this first step is, in reality, far from trivial.

3 Outline

As argued by Wittgenstein (1922), "The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience." (Proposition 6.363).[3] Thus, in performing induction, people have a general preference, other things being equal, for rules that are as simple as possible and as accurate as possible. Simple rules that accurately describe the data give reason to hope that one may be able to predict future observations. Generating such rules appears to be a rather modest condition that scientific theories are expected to satisfy. Simple and accurate rules can also be used to transfer information succinctly from one person to another, or to remember what seems to be the essence of a database.

One should normally expect to face a trade-off between simplicity and accuracy. Consider the example of linear regression.[4] Given a set of "predicting" variables, one attempts to provide a good fit to a "predicted" variable. A common measure of accuracy is the coefficient of determination, R^2. A reasonable measure of complexity is the number of predictors one uses. As is well known, sufficiently increasing the number of predictors will generically provide a perfect fit, namely, a perfectly accurate rule. But this rule will be complex. In fact, its complexity will be equal to the number of observations in the database, and it will be considered to have little predictive value. By contrast, a small number of predictors may not obtain a satisfactory level of

[3] Simplicity was mentioned by William of Occam seven centuries earlier. But Occam's razor is an argument with a normative flavor, whereas here we refer to a descriptive claim about the nature of human reasoning.

[4] While regression analysis is a basic tool of scientific research, it can also be viewed as a (somewhat idealized) model of non-professional human reasoning. See Bray and Savin (1986), who used regression analysis to model the learning of economic agents.

accuracy. Thus, the trade-off between simplicity and accuracy is inherent in the problem of induction.

Regression analysis is an example of *targeted learning*. It is geared to the prediction of a specific variable in a given set-up. Targeted learning must provide an answer to a prediction problem, typically because a decision is called for. But inductive learning can also be *untargeted*. Untargeted inductive learning is a process by which rules are observed for future use, without a concrete problem in mind. For example, the democratic peace example above may be modelled as a rule, "democracies do not engage in wars among themselves". Such a rule does not predict when wars would occur in general, nor does it state that dictatorships would necessarily declare wars on other countries. It does, however, provide an observation that may prove useful.

Observe that people engage in targeted as well as in untargeted learning. In fact, it may be hard to avoid untargeted learning when the data seem obviously to suggest an accurate and simple rule. One may speculate that the human brain has evolved to engage in untargeted learning for several reasons. First, since learning may be complex, it might be wise to perform some learning tasks "off-line", without a concrete goal in mind, and to retain their output for future use. Second, memory constraints may not allow one to retain all data. In such a case, a meaningful summary of the data is called for, and untargeted learning may result in simple and accurate rules that will be easy to remember, while retaining the most relevant parts of the information.[5]

In this paper we deal with targeted as well as with untargeted learning. We consider linear regression and non-parametric regression as models of targeted learning, and we offer two models of untargeted learning. Our focus in these models is on the trade-off between simplicity and accuracy. This

[5] Observe that this process may generate biases as well. In particular, one may tend to remember pieces of evidence that are consistent with the rules one has observed more than one would remember inconsistent evidence.

trade-off is especially conspicuous in a targeted learning problem, such as a prediction problem, where one needs to select a single method for prediction. But it is also implicitly present in untargeted learning, where one has to decide which rules are worthwhile rememberingfor future use. Our main results state that the trade-off between simplicity and accuracy is hard to perform. The reason is that, in these models, determining whether there exists a rule with pre-specified degrees of accuracy and of simplicity is a computationally hard problem. It follows that people cannot be expected to know all the possible rules that may apply to any database of more than trivial size. This, in turn, may explain why people may be surprised to learn rules that hold in a database they are already familiar with. Thus, the motivating examples above may be explained by the computational complexity of the accuracy-simplicity trade-off of inductive learning.

We employ a very simple model, in which observations (or past experiences) are 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. For the quality of government example above, one vector might represent information for a single country for a particular year, with the attributes/variables including the proportion of the population of different religious orientations, the linguistic and legal background of the country, the physical location of the country, etc.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.[6]

Targeted learning is modelled by *functional* rules: rules that point to a functional relationship between several ("predicting") variables and a given variable(the "predicted" variable). A well-known example of such a ruleis *linear regression*, where we take R^2 to be a measure of accuracy and the number of predictors to bea measure of complexity. However, we will also

[6]The degree to which an observation has a certain attribute will normally be in $[0, 1]$.

discuss *non-parametric regression*, where one may choose not only the set of predicting variables but also any function thereof in attempting to fit the predicted variable.

To model untargeted learning, we investigate two types of rules. The first are *exclusionary*: they state that certain phenomena *cannot* occur.[7] For instance, consider the rule "There are no instances in which a democratic country initiated war against another democratic country". Formally, an exclusionary rule is an assignment of values to a subset of columns, interpreted as claiming that this combination of values cannot occur. It is likely that there will often be counter-examples to conjectured rules. Indeed, the war between England and the US in 1812 was quoted as a counter-example to this rule. But such a counter-example need not render the rule useless. We define a notion of accuracy of rules that can roughly be thought of as the average degree to which the rule holds in the given knowledge base. As a measure of the rule's complexity we use the number of attributes it involves.

The second closely related type of rules are *conditional*: they state that *if* an observation has certain attributes, *then* it also has another. For instance, the rule "Democratic countries have universal schooling" can be thought of as saying that if attribute "democratic" has the value 1, then so will attribute "universal schooling". Formally, a conditional rule is a pair, where the first element is an assignment of values to certain columns, and the second is an assignment of a value to another column, interpreted as claiming that the first assignment necessitates the second. The complexity of a conditional rule will be defined as the number of attributes it involves.

Conditional rules can be stated as exclusionary rules. For instance, "Democratic countries have universal schooling" is equivalent to "There are no countries without universal schooling that are democratic". Conversely, exclusionary rules can also be stated as conditional ones. For instance, the

[7] This is in line with Popper's (1965) dictum, which suggested that a scientific theory be formulated by stating what *cannot* happen, highlighting the conditions under which the theory would be falsified.

democratic peace phenomenon can be restated as "If two countries are democratic, then they will not initiate wars against each other". Moreover, the simplicity of a conditional rule is identical to the simplicity of the corresponding exclusionary rule. However, we will argue below that the equivalence between exclusionary and conditional rules only holds at perfect accuracy. More generally, we find that natural definitions of the degree of accuracy of these two types of rules are rather different, and that one cannot easily reduce one type of ruleto the other.

A functional rule can be viewed as the conjunction of many conditional rules, each stating the value of the predicted variable for a particular combination of values of the predicting variables. Yet, functional rules deserve a separate discussion for several reasons. First, they are typically described in a much more parsimonious way than the conjunction of conditional rules. In fact, they are often represented by formulae that are defined for any combination of values of the predicting variables, and not only for the combinations that appear in the database.[8] Second, the way we assess their accuracy also tends to differ from the way we assess the accuracy of a collection of conditional rules. Finally, functional rules represent targeted learning, and they highlight the accuracy-simplicity trade-off when one is forced to make a choice among (prediction) rules.

Our aim is to demonstrate that finding "good" rules, of any of the types described above, 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 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

[8]The fact that a functional rule is formally represented by a function whose domain extends beyond the given database does not imply that the rule would necessarily hold for combinations of values that have not been encountered. Moreover, the rule may not hold even in future observations of combinations of values that have already been observed in the past. We return to this point below.

is polynomial, that can solve it. However, NP-Completeness means somewhat more than the fact that there is no 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 was found to date.

An appendix describes the notion of NP-Completeness more fully. For the time being, it suffices to mention that NP-Completeness, and the related concept of NP-Hardness, are the standard concepts of computational difficulty used in computer science, and that NP-Complete problems are generally considered to be intractable.

We show that finding simple rules is a computationally hard problem. Formally, once the concepts of a rule and its accuracy are defined, we prove that the question "Is there a rule employing no more than k attributes that has a given accuracy level?" is NP-Complete. This result holds for linear regression, for non-parametric regression, for exclusionary rules, and for conditional rules.

Our measures of the simplicity of rules is admittedly crude, but it offers a reasonable approximation to the intuitive notion of complexity, especially in the absence of additional structure that may distinguish among attributes.

We should emphasize that the rules we discuss here 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. Some of the examples and terminology we use may suggest that these rules are backed by theories, but we do not purport to model the entire process of developing and choosing among theories.

The rest of this paper is organized as follows. The next two sections are devoted to formal modelling of targeted and untargeted learning and to a statement of the complexity results. Section 6 concludes. It is followed by two appendices. Appendix A contains proofs of all results, and Appendix B contains an informal introduction to the theory of computational complexity and NP-Completeness.

4 Targeted Learning

4.1 Linear Regression

Assume that we are trying to predict a variable Y given the predictors $X = (X_1, ..., X_m)$. 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 $(x_{ij})_{i \leq n, j \in K}$.

Throughout this paper we assume that the data are given in their entirety, that is, that there are no missing values. Incomplete matrices or vectors introduce conceptual issues that are beyond the scope of this paper.

How does one select a set of predictors? That is, how does one select a functional rule? Let us first consider the feasible set of rules, projected onto the accuracy-complexity space. For a set of predictors 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 may be more convenient to define a connected set:

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

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

Insert Figure 1 about here

It seems reasonable that, other things being equal, people would prefer both simplicity (low k) and accuracy (high r). How does a person make this trade-off? One possibility is that a person may be ascribed a function $v : \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ that represents her preferences for simplicity and accuracy. For example, the widely used adjusted R^2 can be viewed as such a function.[9] 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 if she prefers both simplicity and accuracy, and trades them off as described by v. The optimization problem that such a person 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. More precisely, given X, Y, k, r, determining whether $(k,r) \in F'(X,Y)$ is computationally hard. Formally, we define

Problem LINEAR REGRESSION: Given a matrix X and a vector Y, a natural number $k \geq 1$, and a real number $r \in [0,1]$, is there a subset K of $\{X_1, ..., X_m\}$ such that $|K| \leq k$ and $R^2_K \geq r$?

Theorem 1 *LINEAR REGRESSION is an NP-Complete problem.*

This result shows that it is a hard problem to find the smallest set of variables that obtain a pre-specified level of R^2. Alternatively, it can be

[9] Observe that the adjusted R^2 is a function of the degree of accuracy, R^2, of the degree of complexity, k, as well as of the number of observations, n. As long as we compare rules given a fixed database, n is identical for all candidate prediction rules and need not appear explicitly as an argument of the function v. We return to this issue below.

viewed as pointing out that finding the highest R^2 for a pre-specified number of variables k is a hard problem.

Theorem 1 might explain 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.

While the focus of this paper is on everyday human reasoning, Theorem 1 can also be interpreted as a result about scientific research. It is often the case that a scientist is trying to regress a variable Y on some predictors $(X_j)_j$, and to find a small set of predictors that provide a good fit. Our result shows that many practicing scientists can be viewed as coping with a problem that is NP-Hard.

There is definitely more to scientific research than finding a small set of variables that provide a good fit. Yet, scientific research needs to take goodness of fit and complexity into account. While our model does not capture many other aspects of scientific research, it highlights the inherent difficulty of this one.

Our discussion here (and throughout the paper) presupposes a fixed database X. In reality, however, one may have to choose among prediction rules that were obtained given different databases. For instance, 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.[10] 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.[11]

We have shown that determining whether a given pair (k, r) is in $F'(X, Y)$ is a hard problem when the input is (X, Y, k, r). In the proof, however, we only use the value $r = 1$. Thus, we actually prove a stronger result, namely, that the following problem is also NP-Complete: "Given a database X, a variable Y, and a natural number k, is there a set of predictors that uses no more than k variables and that achieves $R^2 = 1$?" (Observe that this problem trivially reduces to LINEAR REGRESSION.) If we replace the value 1 above by a pre-determined degree of accuracy r, we obtain a family of problems parametrized by r. Our proof can be easily modified to show that these problems are also NP-Complete for any positive value of r.[12]

It follows that, for any positive value of r, it is also hard to determine

[10] 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.

[11] In sub-section 5.3 we discuss rules that vary in their degree of applicability even though they are derived from the same database.

[12] The results that follow are also stated for the case that r is given as input. All of them are proven for the case $r = 1$, and all of them can be strengthened to show that the respective problems are NP-Complete for ranges of values of r.

whether a given k is in the r-cut of $F'(X,Y)$ when the input is (X,Y,k). 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. Thus, 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 predictor or 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).

4.2 Non-Parametric Regression

Linear regression may not always be the most natural model of prediction. First, one may wish to consider non-linear functions. Second, one may not wish to commit to a particular functional form. Consider, for example the variable "the degree of corruption of the judiciary". Such a variable allows more than one obvious quantification. However, the choice of a quantification should be made in tandem with the choice of the functional form of the regression.

In this sub-section we focus on the informational content of the variables. Thus, we consider a regression problem in which one is free to choose *both* a

set of predicting variables *and a function* thereof that approximates a given variable. We refer to this problem as *non-parametric regression*.[13]

Consider a database consisting of observations, or cases $C = \{1, ..., n\}$, with attributes (predicting variables) $A = \{1, ..., m\}$. The data are $X = (x_{ij})_{i \leq n, j \leq m}$ and $Y = (y_i)_{i \leq n}$ where $x_{ij}, y_i \in \mathbb{R}$. We wish to predict the value of Y as a function of $(X_j)_{j \leq m}$.

Let a *predictor for* Y be a pair (E, g) where $E \subset A$ and $g : \mathbb{R}^E \to \mathbb{R}$. Given a predictor for Y, (E, g), it is natural to define its degree of inaccuracy in case $i \in C$ as the squared error:

$$SE((E, g), i) = (g(x_{ij})_{j \in E} - y_i)^2$$

and its degree of accuracy over the entire database as the mean squared error:

$$MSE(E, g) = \tfrac{1}{n} \textstyle\sum_{i \leq n} SE((E, g), i).$$

Thus, $MSE(E, g)$ corresponds to the mean of squared errors (MSE) in linear regression analysis.

Observe that if the matrix X and the vector Y are sampled from a continuous joint distribution, then, with probability 1 all values of each variable are distinct. In this case, every single variable X_j (defining $E = \{j\}$) will predict Y perfectly for an appropriately chosen function $g : \mathbb{R} \to \mathbb{R}$. The accuracy-simplicity trade-off is rather trivial in this situation. But this is not the case if, for instance, the variables can assume only finitely many values, and, in particular, if these are binary variables that indicate the existence of attributes in cases. In this situation, some values of each variable X_j are likely to recur, and it is no longer clear whether Y is a function of a particular

[13] We remind the reader that our use of the terms "predicting variables" and "predicted variable" should be taken with a grain of salt. This section addresses the formal problem of finding a set of variables that, *in a given database*, provides a good fit to another variable. Such a set of variables need not suggest a theory by which one can predict, let alone influence, the so-called "predicted variable".

X_j or of a combination of several X_j's. In interpreting the following result, the reader is asked to focus on models in which it is not implausible to find recurring values in the matrix X and the vector Y.[14]

We define the following problem:

Problem NON-PARAMETRIC REGRESSION: Given a matrix X and a vector Y, a natural number $k \geq 1$ and a real number $r \in \mathbb{R}$, is there a predictor for Y, (E, g), such that $|E| \leq k$ and $MSE(E, g) \leq r$?

Theorem 2 *NON-PARAMETRIC REGRESSION is an NP-Complete problem.*

Notice that this theorem is not implied by the previous one. That is, the fact that it is hard to tell whether a linear relationship (with parameters (k, r)) exists does not imply that it is hard to tell whether any functional relationship (with these parameters) exists. While a sub-class of functions is easier to exhaust, solution algorithms are not restricted to enumeration of all possible functions.

It follows that looking at a database and finding functional relationships in it is, in general, a hard problem. Correspondingly, one may be surprised to be told that such a functional relationship holds, even if, in principle, one had all the information needed to find it.

5 Untargeted Learning

5.1 Exclusionary Rules

Rules have traditionally been modelled by propositional logic, which may capture rather sophisticated formal relationships (see Carnap (1950)). By contrast, we offer here a very simplified model of rules, aiming to facilitate

[14] When X and Y are sampled from a continuous joint distribution, one may wish to restrict the function g (say, to be Lipschitzian with constant bounded away from 0) and obtain similar results.

the discussion of the process of induction, and to highlight its the similarity of rules to regression analysis.[15]

As above, let $C = \{1, ..., n\}$ be the set of *observations*, or *cases*, and $A = \{1, ..., m\}$ – the set of *variables*, or *attributes*. We assume that data are given by a matrix $X = (x_{ij})_{i \leq n, j \leq m}$ of real numbers in $[0, 1]$, such that, for all $i \leq n$, $j \leq m$, x_{ij} measures the degree to which case i has attribute j.[16] Observe that no predicted variable Y is given in this problem.

An *exclusionary rule* is a pair (E, φ) such that $E \subset A$ and $\varphi : E \to \{0, 1\}$.[17] The interpretation of the exclusionary rule (E, φ) is that there is no case whose values on the attributes in E coincide with φ. For instance, assume that the set of cases consists of countries at given years. Saying that "democratic countries provide universal education" would correspond to setting $E = \{democratic, universal\_education\}$ and $\varphi(democratic) = 1$, $\varphi(universal\_education) = 0$, stating that one cannot find a case of a democratic country that did not provide universal education.

To what extent does an exclusionary rule (E, φ) hold in a database X? Let us begin by asking to what extent the rule applies to a particular case i. We suggest that this be measured by[18]

[15] For other approaches to modelling learning processes, see Mitchell (1997).

[16] The restriction of x_{ij} to the unit interval is immaterial. It is designed to facilitate the interpretation of x_{ij} as the *degree* to which a case has an attribute. However, our results hold also if x_{ij} are unrestricted.

[17] One may also define $\varphi : E \to [0, 1]$, and allow rules to have intermediate values of the attributes. For some purposes it may even be useful to let φ assume interval values, that is, to exclude ranges of the attribute value. Our analysis can be extended to these more general cases.

[18] There are many alternatives to this measure. First, one may choose absolute value of the difference instead of its square. We chose the latter mostly for consistency with standard statistical measures. Second, sets of attributes can be used for evaluation of accuracy. Specifically, consider the measure

$$\theta((E, \varphi), i) = \max_{\varnothing \neq F \subset E} \frac{1}{|F|} \sum_{j \in F} (x_{ij} - \varphi(j))^2.$$

Our results hold for this measure as well.

$$\theta((E,\varphi),i) = \max_{j\in E}(x_{ij} - \varphi(j))^2.$$

Thus, if case i is a clear-cut counter-example to the rule (E,φ), i will be a case in which all attributes in E have the values specified for them by φ. That is, $x_{ij} = \varphi(j)$ for all $j \in E$, and then $\theta((E,\varphi),i) = 0$. By contrast, if at least one of the attributes in E is not shared by i at all, that is, $(x_{ij} - \varphi(j))^2 = 1$ for at least one $j \in E$, then $\theta((E,\varphi),i) = 1$, reflecting the fact that i fails to constitute a counter-example to (E,φ), and thus (E,φ) holds in case i.

Generally, the closer are the values $(x_{ij})_{j\in E}$ to $(\varphi(j))_{j\in E}$, the closer is i to being a counter-example to (E,φ). For instance, one might wonder whether the Falkland Islands war is a counter-example to the democratic peace rule. To this end, one must determine the extent to whichArgentina was a democracy at that time. The more is Argentina deemed democratic, the stronger is the contradiction suggested by this example to the general rule.

Given the degree to which a rule (E,φ) applies in each case i, it is natural to define *the degree of accuracy* of the rule (E,φ) given the entire database X as its average applicability over the individual cases:[19]

$$\theta(E,\varphi) = \tfrac{1}{n}\textstyle\sum_{i\le n}\theta((E,\varphi),i).$$

This definition appears to be reasonable when the database contains cases that were not selectively chosen. Observe that one may increase the value of $\theta(E,\varphi)$ by adding cases to C, in which rule (E,φ) has no bite and is therefore vacuously true. For instance, the veracity of the democratic peace phenomenon will be magnified if we add many cases in which no conflict occurred. We implicitly assume that only relevant cases are included in C. More generally, one may augment the model by a relevance function, and weight cases in $\theta(E,\varphi)$ by this function.[20]

[19] To simplify notation, we use the letter θ for different domains. Throughout this paper it is a measure of accuracy of a rule.

[20] The notion of conditional rules (in sub-section 5.3 below) offers another solution to this problem.

In untargeted learning, one need not select a single rule by which to perform prediction. Rather, one may maintain many rules that appear to be valid, and to use them when the need may arise. Yet, people tend to prefer simpler rules also in the context of untargeted learning. For instance, the rule "democratic countries provide universal education" is more complex than the rule "there is always universal education", corresponding to the set $E = \{universal\_education\}$ and $\varphi(universal\_education) = 0$. If, indeed, the column $universal\_education$ in the matrix X consisted of ones alone, it is more likely that one would come up with the generalization that there is always universal education than with the democracy–universal education relationship.

In our model, the *complexity* of a rule (E, φ) is naturally modelled by the number of attributes it refers to, namely, $|E|$. Thus, performing induction may be viewed as looking for a small set E that, coupled with an appropriate φ, will have a high degree of accuracy $\theta(E, \varphi)$.

Parallelingthe discussion in the context of linear regression, one may denote by $F(X)$ the set of pairs (k, r) for which there exists a rule with degree of accuracy r and degree of complexity k. Again, one may define

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

as the feasible set of (k, r) pairs. Out of this set, a subset of rules will be noticed and remembered. One may assume that, for a function $v : \mathbb{R}_+ \times [0, 1] \to \mathbb{R}$ that is decreasing in its first argument and increasing in the second, and for an appropriate constant $c \in \mathbb{R}$, the subset of rules that a person would notice and remember are those for which $v(k, r) \geq c$, as depicted in Figure 4.

Insert Figure 4 about here

A person who considers possible rules is confronted with the following problem: given a database X and a pair k, r, is it the case that $(k, r) \in F'(X)$? Or, differently stated:

Problem INDUCTION: Given a matrix X, a natural number $k \geq 1$ and a real number $r \in [0, 1]$, is there an exclusionary rule (E, φ) such that $|E| \leq k$ and $\theta(E, \varphi) \geq r$?

We can now state the following result.

Theorem 3 *INDUCTION is an NP-Complete problem.*

As mentioned above, the proof of Theorem 3 can be modified to show that the problem remains NP-Complete even if r is fixed, for a range of values of r.

5.2 Simple Exclusionary Rules

The induction problem stated above is computationally hard for two combinatorial reasons: first, there are many subsets E that may be relevant to the rule. Second, for each given E there are many assignments of values φ. Our intuitive discussion, however, focussed on the first issue: we claim that it is hard to find minimal regularities because there are many subsets of variables one has to consider. It is natural to wonder whether the complexity of problem INDUCTION is due to the multitude of assignments φ, and has little to do with our intuitive reasoning about the multitude of subsets.[21]

We therefore devote this sub-section to *simple exclusionary rules*, defined as exclusionary rules (E, φ) where $\varphi \equiv 1$. A simple exclusionary rule can thus be identified by a subset $E \subset A$. We denote $\theta((E, \varphi), i)$ by $\theta(E, i)$, and $\theta(E, \varphi)$ – by $\theta(E)$. Explicitly,

[21] This concern can only be aggravated by reading our proof: we actually use all attributes in the proof of complexity, relying solely on the difficulty of finding the assignment β.

$$\theta(E,i) = \max_{j\in E}(1-x_{ij})^2.$$

and

$$\theta(E) = \tfrac{1}{n}\sum_{i\leq n}\theta(E,i).$$

We now formulate the following problem:

Problem SIMPLE INDUCTION: Given a matrix X, a natural number $k \geq 1$ and a real number $r \in [0,1]$, is there a simple exclusionary rule $E \subset A$ such that $|E| \leq k$ and $\theta(E) \geq r$?

We can now state the following result.

Proposition 4 *SIMPLE INDUCTION is an NP-Complete problem.*

It follows that the difficulty of the problem INDUCTION is not an artifact of the function φ, but rather, has to do also with the selection of variables one considers.[22]

5.3 Conditional Rules

Rules are often formulated as conditional statements. For instance, it is more natural to state the rule "All ravens are black", or "*If* x is a raven, *then* x is black" then the rule "There are no non-black ravens". Obviously, the two formulations are equivalent, and they are both equivalent to "If x is not black, then x is not a raven". Indeed, this equivalence lies at the heart of Hempel's (1945) paradox of confirmation. Yet, this equivalence holds only when the

[22] SIMPLE INDUCTION deals with $\binom{m}{k}$ possible selections of subsets of k columns. INDUCTION, by contrast, deals with $\binom{m}{k}2^k$ selections of binary values for k columns. Yet, the fact that SIMPLE INDUCTION is NP-Complete does not mean that INDUCTION is as well. The reason is that one is not restricted to algorithms that exhaust all possible solutions. Using certain properties of a problem, one may be able to find a solution in a larger set of possible solutions more efficiently than in a smaller set.

rules are supposed to be perfectly accurate. By contrast, when we are trying to assess their degree of accuracy in general, this equivalence breaks down, as we explain shortly. We therefore introduce conditional rules as a separate formal entity.

Let a *conditional rule* be a pair $((E, \varphi), (j, b))$ such that $E \subsetneq A$, $\varphi : E \to \{0, 1\}$, $j \in A \backslash E$, and $b \in \{0, 1\}$. The interpretation of the conditional rule $((E, \varphi), (j, b))$ is that, whenever the attributes in E assume the corresponding values specified by φ, attribute j assumes the value b. Thus, the rule "All ravens are black" can be modelled by setting $E = \{raven\}$, $\varphi(raven) = 1$, $j = black$, $b = 1$. Similarly, the democratic peace phenomenon may be captured by $E = \{country\_1\_democratic, country\_2\_democratic\}$, $\varphi(country\_1\_democratic) = \varphi(country\_2\_democratic) = 1$, $j = war$, and $b = 0$.

Observe that, if $E = \varnothing$, the conditional rule states that attribute j always assumes the value b. It thus corresponds to the exclusionary rule that rules out the value $(1 - b)$ for attribute j.

To what extent does conditional rule $((E, \varphi), (j, b))$ hold in case i? We propose that the applicability of the rule in the case be measured by

$$\theta(((E, \varphi), (j, b)), i) = (x_{ij} - b)^2.$$

That is, in each particular case, the rule is judged solely by the degree to which the case agrees with the rule's consequent. However, not all cases are going to be equally relevant to the assessment of the rule given the entire database. Intuitively, the rule "All ravens are black" should be assessed based on ravens alone. Thus, a case is relevant to the assessment of a rule to the degree that it offers an example of the antecedent of the rule. Formally, let

$$w(((E, \varphi), (j, b)), i) = 1 - \max_{l \in E}(x_{il} - \varphi(l))^2.$$

(This expression is assumed to be 1 if $E = \varnothing$.)[23]

[23] Observe that, using the notation from sub-section 5.1, the weight w can be written as

Thus, if case i agrees with φ on all attributes specified in E, it is a very relevant test case for the rule. If, however, at least one of the attributes of case i is completely different from φ (that is, for one $l \in E$, $(x_{il} - \varphi(l))^2 = 1$), then case i is irrelevant to the rule, and should not be part of its evaluation.

With this definition, it is natural to define the degree to which conditional rule $((E,\varphi),(j,b))$ holds in the entire database by

$$\theta((E,\varphi),(j,b)) = \frac{\sum_{i \leq n} w(((E,\varphi),(j,b)),i)\theta(((E,\varphi),(j,b)),i)}{\sum_{i \leq n} w(((E,\varphi),(j,b)),i)}$$

in case the denominator does not vanish. If it does, we define $\theta((E,\varphi),(j,b)) = 1$.

Thus, a rule is evaluated by a weighted average of the extent to which it holds in each particular case, where the weights are defined by an endogenous relevance function. According to this definition, "All ravens are black" will be assessed based on the population of ravens, whereas "Everything that is not black is not a raven" will be judged by its correctness in the population of non-black objects. If one of these rules is true to degree 1, so is the other.[24] But if their degrees of correctness are in $[0,1)$, they need not be equal. In fact, they can be as different as this half-open interval allows.[25]

$w(((E,\varphi),(j,b)),i) = 1 - \theta((E,\varphi),i)$.

$\theta((E,\varphi),i)$ measures the degree to which case i satisfies the assignment φ over the set E. For an exclusionary rule, a high θ indicates that the case is a counter-example to the rule. For the antecedent of a conditional rule, a high θ indicates low relevance of the case to the rule.

[24] Observe that the populations are defined as fuzzy sets. For instance, each case i offers an example of a raven – say, attribute l – to a continuous degree $x_{il} \in [0,1]$.

For a conditional rule to be true to degree 1, every case i with $\sum_{i \leq n} w(((E,\varphi),(j,b)),i) > 0$ has to satisfy $x_{ij} = b$. In the example, "All ravens are black" holds to degree 1 if and only if every observation that is at least partly a raven is absolutely black. Equivalently, every example has to be absolutely black, or absolutely not a raven. This is equivalent to the fact that the exclusionary rule "There are no non-black ravens" holds to degree 1.

[25] For instance, in a database with $n-1$ black ravens and one white raven, "All ravens are black" will be true to degree $1 - \frac{1}{n}$ whereas "All non-black objects are not ravens" – to degree 0.

The definition above allows a conditional rule to be vacuously true (if $\sum_{i \leq n} w(((E, \varphi), (j, b)), i) = 0$) but it does not allow irrelevant examples to affect the truthfulness of a conditional rule in the presence of relevant examples. Thus, in a database that consists only of white shoes, "All ravens are black" will be deemed correct. But if there is but one case that is, to some degree, a raven, white shoes become irrelevant. One might view this formulation as a "resolution" of Hempel's paradox of confirmation, although this is not our purpose here.[26]

It is natural to define the *complexity* of a conditional rule $((E, \varphi), (j, b))$ by $|E| + 1$. (A conditional rule of complexity 1 would correspond to the case $E = \varnothing$.) The trade-off between accuracy and simplicity exists here as well. Indeed, finding whether, for pre-specified $k \geq 1$ and $r \in [0, 1]$, there exists a conditional rule $((E, \varphi), (j, b))$ of complexity k or less, that achieves $\theta((E, \varphi), (j, b)) \geq r$, is also an NP-Complete problem. We omit the formal statement of this result and its proof because both closely mimic the statement and the proof of Theorem 3.

As opposed to exclusionary rules, conditional rules do not apply to the entire database. The scope of a conditional rule depends on the cases that satisfy its antecedents, as measured by $\sum_{i \leq n} w(((E, \varphi), (j, b)), i)$. It follows that conditional rules differ from each other not only in their accuracy and in their complexity, but also in their generality, or scope of applicability. As discussed in sub-section 4.1, one would be expected to prefer, other things being equal, more general rules, which have a larger scope of applicability, and which are thus likely to be more useful in future decisions. An explicit model of the trade-off between accuracy, simplicity, and applicability is beyond the scope of this paper.

[26] There are many other resolutions of this paradox in the literature. One that we found particularly convincing is offered by Gilboa (1993). It argues that, to anyone whose intuition was shaped by Bayesian thinking, there is nothing paradoxical about Hempel's paradox, provided that one carefully distinguishes between weak and strict inequalities in comparing posterior to prior.

6 Conclusion

6.1 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 rule. Hence, this model can easily capture changes in 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 data base 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 com-

parison 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 insightful ways of succinctly summarizing many cases. The present paper may be viewed as an attempt to model the process by which rules are generated from cases.[27]

6.2 Discussion

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 attribute model above, 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

[27] Gilboa and Schmeidler (1999, 2002) attempt to model the process by which cases are used to form probabilistic beliefs.

respect to the predicate model.[28]

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.[29] This problem is, however, beyond the scope of the present paper.

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. Indeed, even polynomial problems can be difficult to solve when the database consists of many cases and many attributes. On the other hand, 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-Hard problems. Faced with the induction problems discussed here,

[28] 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.

[29] 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.

which are NP-Hard, people may use various heuristics to find rules, but they cannot be sure, in general, that the rules they find are the simplest ones.

6.3 Implications

Our results have several implications. First, we find that people may behave differently if their information consists of raw data as compared to rules that summarize these data. Importantly, the raw data may not be more informative than the rules that are derived from them, because this derivation is a non-trivial task. Second, the complexity results might explain why people may prefer to summarize information and make predictions using simple rules that do not employ more than a few variables. Whereas it is generally hard to find whether a certain degree of accuracy can be obtained with a given number of variables, one may find, in polynomial time complexity, the most accurate rule among those that use no more than, say, two variables.

Our model suggests two distinct reasons for which people who have access to the same information might entertain different beliefs and make different decisions. The first is that, due to the complexity problem, different people may happen to uncover different rules, while there is no reason to believe that any one of them can necessarily find the rules discovered by the others. The second reason has to do with personal tastes. Even if two people face the same set of rules, reflected in a set F as in Figure 2, they might have different preferences for the accuracy-simplicity trade-off (captured by the function v). Such preferences determine these individuals' beliefs, as reflected in the predictions and decisions that they are likely to make.

7 Appendix A: Proofs

Proof of Theorem 1:

It is easy to see that LINEAR REGRESSION 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, $(m+1)n$).[30] To show that LINEAR REGRESSION is NP-Complete, we use a reduction of the following problem, which is known to be NP-Complete (see Gary and Johnson (1979)):

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 $(y_i)_{i\leq n}$, a natural number k and a number $r \in [0,1]$ 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 $n = s+l+1$ observations of $m = 2l$ predicting variables. It will be convenient to denote the 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.[31] We now specify these vectors.

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

[30] 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.

[31] In terms of our formal model, the variables may well be defined by these vectors to begin with. However, in the context of statistical sampling, the variables are defined in a probabilistic model, and identifying them with the corresponding vectors of observations constitutes an abuse of notation.

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

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

For $j \leq l$, $x_{nj} = z_{nj} = 0$;

For $i \leq s + l$, $y_i = 1$ and $y_n = 0$.

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 $R^2_K \geq r \equiv 1$ 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

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

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

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

follows from our construction (assigning precisely one of $\{\beta_j, \gamma_j\}$ to 1 and the other – to 0). Obviously, $\alpha + \sum_{j \leq l} \beta_j x_{nj} + \sum_{j \leq l} \gamma_j z_{nj} = 0 = y_i = 0$. 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 $R^2_K = 1$.

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 $|K| \leq l$ for which $R^2_K = 1$. Equality for observation n implies that 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 $R^2_K = 1$. 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

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

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 $j \leq l$, the $s + j$ observation implies that $\beta_j + \gamma_j = 1$. This means that for every $j \leq l$, $\beta_j \neq 0$ 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 $j \leq l$ 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.

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:

We first show that NON-PARAMETRIC REGRESSION is in NP. To this end, it suffices to show that, for any given set of attributes $E \subset A$ with $|E| = k$, one may find in polynomial time (with respect to $n\times(m+1)$) whether there exists a function $g : \mathbb{R}^E \to \mathbb{R}$ such that $MSE(E, g) \leq r$. Let there be given such a set E. Restrict attention to the columns corresponding to E and to Y. Consider these columns, together with Y as a new matrix X' of size $n\times(|E|+1)$. Sort the rows in X' lexicographically by the columns corresponding to E. Observe that, if there are no two identical vectors $(x_{ij})_{j\in E}$ (for different i's in

C), there exists a function $g : \mathbb{R}^E \to \mathbb{R}$ such that $MSE(E, g) = 0$. Generally, $MSE(E, g)$ will be minimized when, for every vector $(x_{ij})_{j\in E} \in \mathbb{R}^E$ that appears in the matrix corresponding to $C \times E$, $g((x_{ij})_{j\in E})$ will be set to the average of (y_l) over all $l \in C$ with $(x_{lj})_{j\in E} = (x_{ij})_{j\in E}$. That is, for every selection of values of the predicting variables, the sum of squared errors is minimized if we choose the average value (of the predicted variable) for this selection, and this selection is done separately for every vector of values of the predicting variables. It only remains to check whether this optimal choice of g yields a value for $MSE(E, g)$ that exceeds r or not.

We now turn to show that NON-PARAMETRIC REGRESSION is in NP-Complete. We use a reduction of the following problem, which is known to be NP-Complete (see Gary and Johnson (1979)):

Problem COVER: Given a natural number p, a set of subsets of $S \equiv \{1, ..., p\}$, $\mathfrak{S} = \{S_1, ..., S_q\}$, and a natural number $t \leq q$, are there t subsets in $\mathfrak{S}$ whose union contains S?

(That is, are there indices $1 \leq j_1 \leq ... \leq j_t \leq q$ such that $\bigcup_{l\leq t} S_{j_l} = S$?)

Let there be given an instance of COVER: a natural number p, a set of subsets of $S \equiv \{1, ..., p\}$, $\mathfrak{S} = \{S_1, ..., S_q\}$, and a natural number t. Define $n = p + 1$, $m = q$, and $k = t$. Define $(x_{ij})_{i\leq n, j\leq m}$ and $(y_i)_{i\leq n}$ as follows. For $i \leq p$, and $j \leq m = q$, set $x_{ij} = 1$ if $i \in S_j$ and $x_{ij} = 0$ otherwise. For $i \leq n$, set $y_i = 1$. For all $j \leq m$, let $x_{nj} = 0$. Finally, set $y_n = 0$.

We claim that there is a predictor for Y, (E, g), with $|E| = k$ and $MSE(E, g) = 0$ iff there are $k = t$ subsets in $\mathcal{S}$ that cover S. Indeed, there exists a predictor for Y, (E, g), with $|E| = k$ and $MSE(E, g) = 0$ iff there are k columns out of the m columns in the matrix, such that no row in the matrix, restricted to these columns, consists of zeroes alone. (Observe that the first p observations of Y are 1. Thus the only problem in defining g to obtain a perfect match $MSE(E, g) = 0$ might occur if some of these vectors, restricted to the these k columns, is equal to the last vector, which consists of zeroes for the predicting variables and zero also for Y.) And this

holds iff there are k sets in $\mathcal{S} = \{S_1, ..., S_q\}$ that cover S.

Finally, observe that the construction is polynomial.□

Proof of Theorem 3:

That INDUCTION is in NP is simple to verify: given a suggested rule (E, φ), one may calculate $\theta(E, \varphi)$ in linear time in the size of the sub-matrix $C \times E$ (which is bounded by the size of the input, $|C \times A|$). That INDUCTION is NP-Complete may be proven by a reduction of the satisfiability problem:[32]

Problem SATISFIABILITY: Given a Boolean function f in Conjunctive Normal Form in the variables $y_1, ..., y_p$, is there an assignment of values ($\{0, 1\}$) to the variables for which $f = 1$?

Let there be given the function $f = \prod_{i \leq q} f_i$ where each factor f_i is the summation of variables y_j and their negations $\overline{y}_j$. (The variables are Boolean, summation means disjunction, multiplication means conjunction, and bar denotes logical negation.) Let $n = q$ and $m = p$. For each factor f_i, $i \leq q = n$, let there be a case i. For each variable y_j, $j \leq p = m$, let there be an attribute j. Define x_{ij} as follows:

If y_j appears in f_i, let $x_{ij} = 0$;

If $\overline{y}_j$ appears in f_i, let $x_{ij} = 1$;

Otherwise, let $x_{ij} = 0.5$.

We claim that there exists a valued rule (E, φ) with $|E| = k = n$ such that $\theta(E, \varphi) \geq r = 1$ iff f is satisfiable by some assignment of values to the variables $y_1, ..., y_p$. Observe that every rule (E, φ) with $|E| = n$ defines an assignment of values (0 or 1) to the variables $(y_j)_{j \leq p}$, and vice versa. We claim that every rule (E, φ) with $|E| = n$ obtains the value $\theta(E, \varphi) = 1$ iff the corresponding assignment satisfies f. To see this, let (E, φ) be a rule with $|E| = n$. Note that $\theta(E, \varphi) = 1$ iff for every case i we have

[32] SATISFIABILITY is the first problem that was proven to be NP-Complete. This was done directly, whereas proofs of NP-Completeness of other problems is typically done by reduction of SATISFIABILITY to these problems (often via other problems). See Gary and Johnson (1979) for definitions and more details.

$\theta((E,\varphi),i) = 1$, which holds iff for every case i there exists an attribute j such that $(x_{ij} - \varphi(j))^2 = 1$, that is, $x_{ij} = 1 - \varphi(j)$. By construction of the matrix X, $x_{ij} = 1 - \varphi(j)$ iff (i) y_j appears in f_i, and $\varphi(j) = y_j = 1$, or (ii) $\overline{y}_j$ appears in f_i and $\varphi(j) = y_j = 0$. (Observe that, if neither y_j nor $\overline{y}_j$ appear in f_i, $(x_{ij} - \varphi(j))^2 = 0.25$.) In other words, $x_{ij} = 1 - \varphi(j)$ iff the variable y_j (or its negation) satisfies the factor f_i. It follows that $\theta(E,\varphi) = 1$ iff the assignment defined by φ satisfies f. Observing that the construction above can be performed in polynomial time, the proof is complete.□

Proof of Proposition 4:

It is easy to see that SIMPLE INDUCTION is in NP. To show that SIMPLE INDUCTION is NP-Complete, we use a reduction of COVER again. Given and instance of COVER, n, $\mathfrak{S} = \{S_1, ..., S_q\}$, and t, define $n = p$, $m = q$, and $k = t$. Thus each member of S corresponds to a case $i \in C$, and each subset $S_j \in \mathfrak{S}$ – to an attribute $j \in A$. Let $x_{ij} = 1$ if $i \notin S_j$ and $x_{ij} = 0$ if $i \in S_j$. We argue that there is a rule $E \subset A$ such that $|E| \leq k$ and $\theta(E) \geq 1$ iff there are k subsets $\{S_{j_l}\}_{l \leq k}$ whose union covers S. Indeed, such a rule exists iff there is a set E of k attributes $\{j_l\}_{l \leq k}$ such that, for every i, $\theta(E,i) = 1$. This holds iff, for every i there exists an attribute $j_l \in E$ such that $x_{ij_l} = 0$. And this holds iff for each member of S there is at least one of the k sets $\{S_{j_l}\}_{l \leq k}$ to which it belongs.

Finally, observe that the construction of the matrix X is polynomial in the size of the data.□

8 Appendix B: Computational Complexity

A **problem** can be thought of as a set of legitimate inputs, and a correspondence from it into a set of legitimate outputs. For instance, consider the problem "Given a graph, and two nodes in it, s and t, find a minimal path from s to t". An input would be a graph and two nodes in it. These are assumed to be appropriately encoded into finite strings over a given alphabet. The corresponding encoding of a shortest path between the two nodes would be an appropriate output.

An **algorithm**, in the intuitive sense, is a method of solution that specifies what the solver should do at each stage. **Church's thesis** maintains that algorithms are those methods of solution that can be implemented by **Turing machines**. It is neither a theorem nor a conjecture, because the term "algorithm" has no formal definition. In fact, Church's thesis may be viewed as defining an "algorithm" to be a Turing machine. It has been proved that Turing machines are equivalent, in terms of the computational tasks they can perform, to various other computational models. In particular, a PASCAL (or BASIC) program run on a modern computer with an unbounded hard disk is also equivalent to a Turing machine and can therefore be viewed as a definition of an "algorithm".

It is convenient to restrict attention to **YES/NO problems**. Such problems are formally defined as subsets of the legitimate inputs, interpreted as the inputs for which the answer is YES. Many problems naturally define corresponding YES/NO problems. For instance, the previous problem may be represented as "Given a graph, two nodes in it s and t, and a number k, is there a path of length k between s and t in the graph?" It is usually the case that if one can solve all such YES/NO problems, one can solve the corresponding optimization problem. For example, an algorithm that can solve the YES/NO problem above for any given k can find the minimal k for which the answer is YES. Moreover, such an algorithm will typically also find a path that is no longer than the specified k.

Much of the literature on computational complexity focuses on **time complexity**: how many operations will an algorithm need to perform in order to obtain the solution and halt. It is customary to count operations of reading and writing numbers, as well as logical and algebraic operations on numbers as taking a single unit of time each. Taking into account the amount of time these operations actually take (for instance, the number of actual operations needed to add two numbers of, say, 10 digits) typically yields qualitatively similar results.

The literature focuses on **asymptotic** analysis: how does the number of operations grow with the size of the input. It is customary to conduct **worst-case** analysis, though attention is also given to average-case performance. Obviously, the latter requires some assumptions on the distribution of inputs, whereas worst-case analysis is free from distributional assumptions. Hence the complexity of an algorithm is generally defined as the order of magnitude of the number of operations it needs to perform, in the worst case, to obtain a solution, as a function of the input size. A problem is **polynomial** (or, "of polynomial complexity") if there exists an algorithm that always solves it correctly within a number of operations that is bounded by a polynomial of the input size. A problem is **exponential** if all the algorithms that solve it may require a number of operations that is exponential in the size of the input.

Polynomial problems are generally considered relatively "easy", even though they may still be hard to solve in practice, if the coefficients and/or the degree of the bounding polynomial are high. By contrast, exponential problems become intractable even for inputs of moderate sizes. To prove that a problem is polynomial, one typically points to a polynomial algorithm that solves it. Proving that a YES/NO problem is exponential, however, is a very hard task, because it is generally hard to show that there does *not* exist an algorithm that solves the problem in a number of steps that is, say, $O(n^{17})$ or even $O(2^{\sqrt{n}})$.

A **non-deterministic Turing machine** is a Turing machine that allows multiple computations for a given input. These computations can be thought of as paths in a tree, in which each node is a step in a computation, and the depth of the tree measures the time it takes the machine to reach a solution. A non-deterministic Turing machine can be loosely compared to a parallel processing modern computer with an unbounded number of processors. It is assumed that these processors can work simultaneously, and, should one of them find a solution, the machine halts. Consider, for instance, the Hamiltonian path problem: given a graph, is there a path that visits each node precisely once? A straightforward algorithm for this problem would be exponential: given n nodes, one needs to check all the $n!$ permutations to see if any of them defines a path in the graph. A non-deterministic Turing machine can solve this problem in linear time. Roughly, one can imagine that $n!$ processors work on this problem in parallel, each checking a different permutation. Each processor will therefore need no more than $O(n)$ operations. In a sense, the difficulty of the Hamiltonian path problem arises from the multitude of possible solutions, and not from the inherent complexity of each of them.

The class **NP** is the class of all YES/NO problems that can be solved in **P**olynomial time by a **N**on-deterministic Turing machine. Equivalently, it can be defined as the class of YES/NO problems for which the validity of a suggested solution can be verified in polynomial time (by a regular, deterministic algorithm). The class of problems that can be solved in polynomial time (by a deterministic Turing machine) is denoted **P** and it is obviously a subset of NP. Whether P=NP is considered to be the most important open problem in computer science. While the common belief is that the answer is negative, there is no proof of this fact.

A problem A is **NP-Hard** if the following statement is true ("the conditional solution property"): if there were a polynomial algorithm for A, there would be a polynomial algorithm for any problem B in NP. There may

be many ways in which such a conditional statement can be proved. For instance, one may show that using the polynomial algorithm for A a polynomial number of times would result in an algorithm for B. Alternatively, one may show a polynomial algorithm that translates an input for B to an input for A, in such a way that the B-answer on its input is YES iff so is the A-answer of its own input. In this case we say that B is **reduced** to A.

A problem is **NP-Complete** if it is in NP, and any other problem in NP can be reduced to it. It was shown that the **SATISFIABILITY** problem (whether a Boolean expression is not identically zero) is such a problem by a direct construction. The latter means that, for every NP problem B, there exists an algorithm that accepts an input for B, z, and generates in polynomial time a Boolean expression that can be satisfied iff the B-answer on z is YES. With the help of one problem that is known to be NP-Complete (**NPC**), one may show that other problems, to which the NPC problem can be reduced, are also NPC. Indeed, it has been shown that many combinatorial problems are NPC.

An NPC problem is NP-Hard, but the converse is not necessarily true. First, NP-Hard problems need not be in NP themselves, and they may not be YES/NO problems. Second, NPC problems are also defined by a particular way in which the conditional solution property is proved, namely, by reduction.

There are by now hundreds of problems that are known to be NPC. Had we known one polynomial algorithm for one of them, we would have a polynomial algorithm for each problem in NP. As mentioned above, it is believed that no such algorithm exists.

References

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

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

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.

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

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. (1993), "Hempel, Good, and Bayes", manuscript.

Gilboa, I., and D. Schmeidler (1999), "Inductive Inference: An Axiomatic Approach", *Econometrica*, forthcoming.

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

Gilboa, I., and D. Schmeidler (2002), "A Cognitive Foundation of Probability", *Mathematics of Operations Research*, **27**: 68-81.

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

Hempel, C. G. (1945). "Studies in the Logic of Confirmation I", *Mind* **54**: 1-26.

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.

Mitchell, T. (1997), *Machine Learning*. McGraw Hill.

Popper, K. R. (1965), *Conjectures and Refutations: The Growth of Scientific Knowledge (2nd edition)* New York: Harper and Row.

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

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

Figure 1

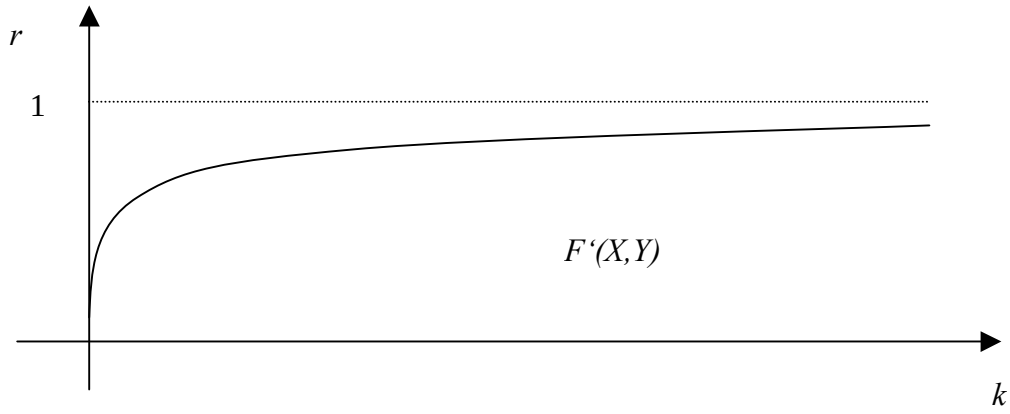

Figure 2

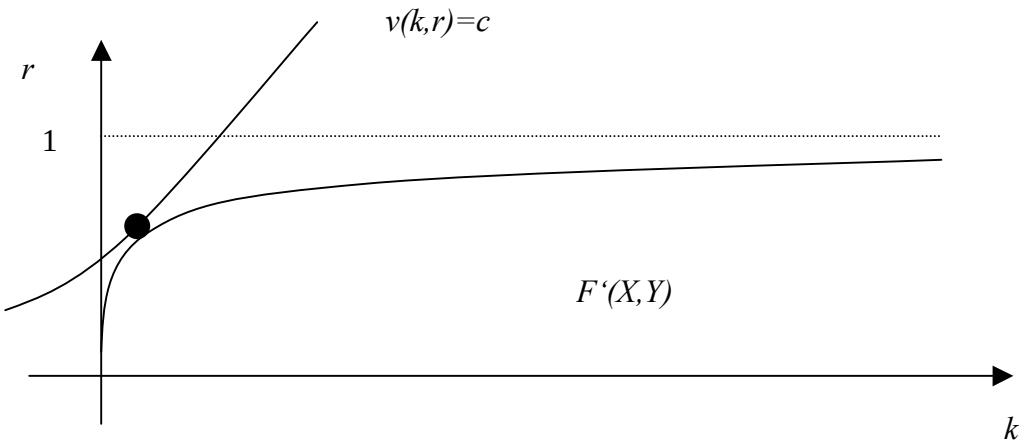

Figure 3

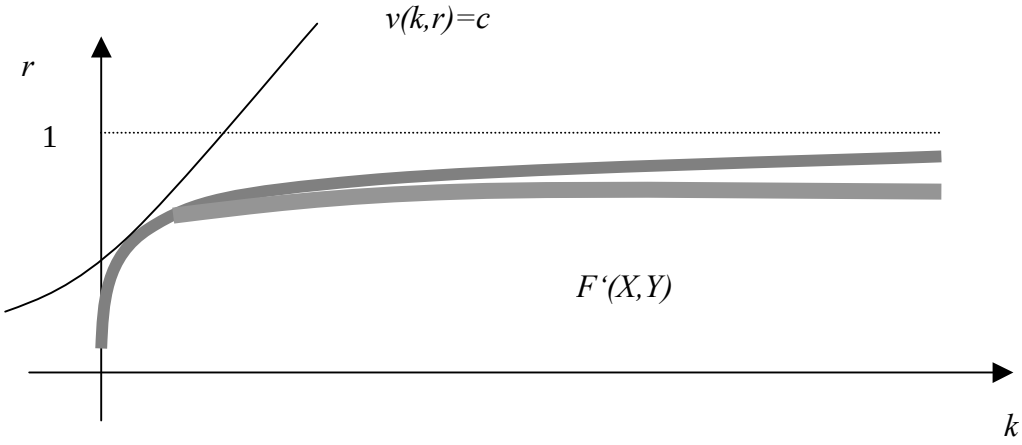

Figure 4