REAL-TIME DECENTRALIZED INFORMATION PROCESSING AND RETURNS TO SCALE

Timothy Van Zandt

Roy Radner

New York University, Leonard N. Stern School of Business
Department of Information, Operations, and Management Sciences
Henry Kaufman Management Center
44 West 4th Street
New York, NY 10012

Real-Time Decentralized Information Processing and Returns to Scale*

Timothy Van Zandt†
Princeton University

Roy Radner
Stern School of Business
New York University

May 6, 1996

Authors' addresses:

Timothy Van Zandt	Roy Radner
Department of Economics	Stern School of Business MEC 9-68
Fisher Hall	New York University
Princeton University	44 West Fourth Street
Princeton, NJ 08544-1021	New York, NY 10012
Voice: (609) 258-4050	Voice: (212) 998-0813
Fax: (609) 258-6419	Fax: (212) 995-4228
Email: tvz@princeton.edu	Email: rradner@stern.nyu.edu

*We would like to thank In-Koo Cho, Jacques Cremer, Mathias Dewatripont, George Mailath, Paul Milgrom and numerous seminar participants for helpful comments. Archishman Chakraborty, Sangjoon Kim and Bilge Yilmaz provided valuable research assistance. This paper appeared previously under the title, "Information Processing and Returns to Scale of a Statistical Decision Problem."

†This research supported in part by grants SES-9110973 and SBR-9223917 from the National Science Foundation, by a CORE Research Fellowship (1993–1994), and by grant 26 of the "Pôle d'Attraction Interuniversitaire" program of the Belgian Government.

Abstract

We study the properties of real-time decentralized information processing, as a model of human information processing in organizations, and use the model to understand how constraints on human information processing affect the returns to scale of firms. With real-time processing, decentralization does not unambiguously reduce delay, because processing a subordinate's report precludes processing current data. Because decision rules are endogenous, delay does not inexorably lead to eventually decreasing returns to scale; however, returns are more likely to be decreasing when computation constraints, rather than sampling costs, limit the information upon which decisions are conditioned. The results illustrate that the requirement of informational integration causes a breakdown of the replication arguments that are often used to establish non-decreasing returns.

Contents

1 Introduction

This paper explores the properties of real-time decentralized information processing, as a model of human information processing in organizations such as firms, and uses the model to study how constraints on human information processing affect the returns to scale of firms.

1.1 Real-time decentralized information processing

Information processing is the procedure of transforming data into decisions. Modeling the information processing of economic agents, whether these agents are individuals or organizations, is a way to understand how decision rules are different when there are information processing constraints from those predicted by models with unboundedly rational agents. Furthermore, in organizations—in which information processing is decentralized, i.e., performed jointly by the members of the organizations—information processing is itself an economic activity that is an important determinant of the structure of organizations and that uses significant resources. In fact, the members of a firm's administrative staff are hired precisely because there is no unboundedly rational entrepreneur who could control the firm in the same way.

There have been various approaches to modeling information processing in organizations. Models of costly communication among a fixed number of agents, such as the literature on static and iterative communication mechanisms[1] and team theory,[2] attempt only to explain the decision procedures of exogenously given agents. In fact, information transmission cannot explain why people with no prior private information relevant to a firm are hired to administer the firm; such hiring only adds to potential communication costs. The models in Williamson (1967) and Beckmann (1977) do have an endogenous number of managers, but profits are given by an exogenous function of managerial inputs. Although the managerial production functions are motivated by information processing, these papers do not explicitly model the source of managerial productivity.

More recent research has attempted to explain the source of managerial productivity, and hence not only what decision procedures managers use, but also why and how many administrators are hired to process information. Marschak and Reichelstein (1994, 1995) is a static model of communication mechanisms in which for each agent the marginal cost of sending and receiving messages is increasing (and the amount of information that she can send or receive may be bounded). This means, for example, that it may be useful to spread the task of aggregating four messages among a network of three managers, in which two managers read and aggregate two messages each and the third manager aggregates information received from these

[1] For example, Arrow and Hurwicz (1960), Hurwicz (1960), Malinvaud (1967), Mount and Reiter (1974) and Hurwicz (1977).

[2] For example, Marschak (1955), Radner (1961, 1962) and Marschak and Radne

two managers, because each manager only has to read two messages.[3] In Geanakoplos and Milgrom (1991), which is a static team theory model of hierarchical resource allocation, each manager can learn only a bounded amount of information about the environment, and hence it is valuable to hire more managers so that more information can be brought to bear on a problem.

One gap in these static models is that, in reality, individuals are not so much bounded in the size of information processing tasks they can handle, but rather in the size of the tasks they can perform in a given amount of time. Given enough time, a single person (or several generations working one at a time) can perform almost any computation task that a group of people can perform concurrently. To account properly for the time that information processing takes, a sequential model of computation is needed. Radner (1993) was the first to use such a model for organizations, in a paper that studies decentralized associative computation. There is a fixed computation problem whose data arrive at the same moment, and the time (the delay) between when the data arrive and the answer is produced is endogenous. Keren and Levhari (1979, 1983) can be interpreted as a reduced form of the Radner (1993) model, with restrictions on the regularity of hierarchies. In these three papers, the value of hiring *more* administrators is that they can perform an information processing task *faster*. In the models of periodic associative computation in Radner (1993), Bolton and Dewatripont (1994) and Van Zandt (1994), an organization may also hire more managers because they can handle a faster flow of tasks.

Presumably, the advantage of processing information more quickly and frequently is that decisions are based on more recent information, but these papers do not model this decision-theoretic value of timeliness. To do so requires a stochastic control problem in which the lag of information upon which decisions are based can be endogenous. In such a problem, new data become available each period and also new decisions are made each period. The adaptation of information processing to such a temporal decision problem is called *real-time* information processing. Radner and Van Zandt (1992) introduced such a model of real-time decentralized information processing, in which the decision problem is to predict the sum of a family of stochastic processes.

The current paper generalizes the model in Radner and Van Zandt (1992), and explores the differences between batch and real-time processing. We find that a model of real-time decentralized information processing does not merely provide a decision-theoretic derivation of the cost of delay for a batch processing model such as Keren and Levhari (1979, 1983) or Radner (1993). Instead:

1. The decision rule that is computed, and hence the computation task, is endogenous;

[3]These papers study mechanisms that minimize the sum of the individual communication costs (measured by increasing functions of the dimension of the message space) for a *fixed* set of agents, but endogenizing the number of agents would be a natural extension.

2. A decision can be based on data of heterogeneous lags, and so there is not even a single measure of delay;
3. The timing of decisions is fixed, and decentralization of information processing affects what recent information can be incorporated into decision rules.

Furthermore, with a model of real-time information processing, we can see that there is some truth to the statement that information processing in a firm is decentralized because the task of controlling a firm is too large for one person, and to the notion in Geanakoplos and Milgrom (1991) that decentralization allows decisions to be based on more information. A single person does not have forever to compute a decision rule, because the time decisions are made is part of the decision rule. Some decision rules that can be computed by several people cannot be computed by a single person, because the amount of current information a single person can learn each period is bounded.

We also find that there is a decision-theoretic cost to decentralization that cannot arise with batch processing. Because all data are available at the same time in batch processing and there is a single measure of delay, decentralization unambiguously reduces delay, up to a point. In contrast, with real-time computing, the trade-off between managerial resources and delay is ambiguous. Increased decentralization increases the amount of data of some lags that can be processed, but the subsequent aggregation of partial results displaces recent data. To illustrate this decision-theoretic cost, we give an example (Theorem 1) in which the value of the information in a subordinate's report is always lower than the value of a raw observation, and so it is optimal for each decision to be computed by a single manager, *even though managers are costless.* The intuition is that a manager may prefer to read today's newspaper instead of listening to a subordinate's summary of the news from the preceding five days, when the environment is changing very quickly.

This result suggests that real-time decentralized information processing may have interesting implications for organizational structure, but these are not the focus of this paper. Instead, we use the methodology to study returns to scale, and we will argue in Section 1.2 that it is well suited for this purpose. In subsequent work, Van Zandt (1995b) uses the paradigm in a model of hierarchical resource allocation. In that paper, organizational structure is at the forefront, and the multilevel hierarchies with decentralized decision making that appear there would not be possible (or at least not advantageous) in an analogous model with batch processing.

1.2 Modeling returns to scale

The classical study of technological returns to scale reached the conclusion that decreasing returns to scale should arise only when some input is held fixed. The reason is that a large firm can always imitate the production processes of several small firms, and thus should have average costs that are no higher than those of these small firms. Since the large firm may also be able to organize ways that separate small firms cannot, the large firm is likely to actually have lower long-run average costs. This conclusion is troubling because empirically we observe that most industries have several firms, rather than the natural monopolies that would arise with increasing returns to scale.

Economists have long hypothesized that there are organizational limits on returns to scale due to the problems of coordinating large numbers of activities within a single firm and of operating in diverse environments. Such limits are theoretically possible once we take into account decision procedures and information processing constraints because replication cannot be used to extend the scale of a firm at constant or decreasing unit cost. For a large firm to divide itself into subunits that replicate the activities of small firms means that the subunits cannot communicate, coordinate their activities, or allocate resources except as independent firms would do, such as through markets. There can be no headquarters that controls the subunits because such control would incur managerial costs and delays that the independent firms avoid. Such informationally disintegrated units could hardly constitute a single firm.

In Sections 4 and 5 of this paper, we use our model of real-time decentralized information processing to study whether and how information processing constraints can limit firm size. Real-time decentralized information processing is well suited to this task for several reasons:

1. Managerial resources are endogenous; hence, larger firms can hire more managers in order to cope with more complex decision problems.
2. Decision rules and information processing tasks are endogenous; hence, larger firms are not forced to bog themselves down with additional computation.
3. It best captures the effect of computational delay, which can potentially increase with the scale of the firm.

Mathematically, what we measure in this paper are the returns to scale of a statistical prediction problem that involves predicting the sum of an exchangeable collection of stochastic processes. This is not a complete model of a firm, but rather it is one decision problem with centralized decision making that a firm may face. For example, it is the control problem faced by a firm that sets its production level centrally in order to meet the total demand of a fixed collection of sales offices or customers,[4] or by a firm that predicts the average productivity of a firm's workers, machines or shops, based on past individual productivity indices.

We define the scale of the decision problem to be the number of stochastic processes. If each process is the demand at a sales office and if the expected demand at each sales office is the same, then the scale of the decision problem (as we measure

[4]E.g., Benetton's must respond quickly to changing market conditions at its many retail outlets in order to implement its just-in-time inventory management practices and thereby reduce inventory costs.

it) is proportional to the firm's output (the usual measure of the scale of a firm). If instead each process is a productivity index and if technological returns to scale are constant, then the scale of the decision problem is the number of workers, machines or shops, and it is again proportional to the firm's output. Therefore, in the examples given above, if technological returns to scale are constant, then the returns to scale of the firm are determined by the per-unit losses and computation costs for the prediction problem.

Like the other papers that have related information processing and returns to scale,[5] ours suffers from the weakness that the boundaries of a firm are defined by a given coordination problem or decision problem, and coordination among firms is not modeled. For example, the identification of firm size with the scale of our decision problem presumes that coordination *among* firms through market mechanisms is not possible, and that coordination *within* firms is as centralized as in our decision problem, in which a single decision is made each period. Clearly information integration is related to the boundaries of firms, but the dichotomy in our model and in the others is too extreme.

An ideal model would define the relationship between informational integration and the boundaries of the firm, and allow the level of integration within the firm to be endogenous. Furthermore, it would model both the information processing that coordinates activities within firms and the information processing involved in coordinating distinct firms through market mechanisms, and would subject these to the same constraints. It would then be possible to view optimal firm size as a question of what activities to coordinate within a firm and what activities to coordinate in markets, as framed by Williamson (1975, 1985).

Such an ideal model must be left for future research. In the meantime, at the very least, our model tells us about the returns to scale of centralized decision making. For example, if the decision problem involves setting a level of output, as in the two examples we have given, and if the firm has the option of dividing production among several plants, then our results can be interpreted as follows: If returns to scale are increasing, then firms should set production levels centrally (*ceterus paribus*) to take advantage of the increasing returns of centralized decision making. If returns to scale are decreasing, the firm should divide production among smaller plants to avoid the diseconomies of centralization, when for other reasons it should increase its scale.

[5]These include Williamson (1967), Keren and Levhari (1983), Geanakoplos and Milgrom (1991) and Radner (1993), along with a complementary literature on the loss of control due to incentive problems that arise because of decentralization that is motivated by (but not derived from) information processing constraints: McAfee and McMillan (1995), Melumad et al. (1995) and Mookherjee and Reichelstein (1995a, 1995b).

..3 Results on returns to scale

We find that returns to scale can be increasing or decreasing in our model with computation constraints, depending on our statistical assumptions. In order to distinguish between the effects of information processing constraints and the effects of the statistical assumptions, we also characterize the returns to scale of a benchmark model in which information may be costly but computation is unconstrained. We refer to this benchmark as the *sampling problem*, and to our main model as the *computation problem*. The sampling problem is a standard decision model without bounded rationality, but our characterization of returns to scale of the sampling problem is also new and of interest in its own right.

In spite of the various caveats that have already been given—we model just one of the decision problems a typical firm faces, we do not treat the boundaries of the firm in the ideal way described above, and the results are inconclusive about whether computation constraints limit firm size—the mechanisms by which our results follow from the model illustrate properties of information processing and returns to scale that we consider to be important and robust.

Returns to scale are more likely to be decreasing when computation constraints, rather than sampling costs, limit the information upon which decisions are conditioned, because of computational delay in aggregating information. This unites two themes that first appeared long ago in the economic literature on organizations. One is that delay and change are fundamental for understanding information processing constraints in organization. Kaldor (1934, p. 78) observed that coordination tasks arise only in changing, dynamic environments, and Robinson (1958, Chapter III) emphasized managerial delay as a limit to firm size. Hayek (1940, pp. 131–132)—in a criticism of the iterative planning procedures proposed by Lange (1936, 1937) and Dickinson (1939), which assume that the underlying economic data are constant—states:

> In the real world, where constant change is the rule, ... whether and how far anything approaching the desirable equilibrium is ever reached depends entirely on the speed with which the adjustments can be made. The practical problem is not whether a particular method would eventually lead to a hypothetical equilibrium, but which method will secure the more rapid and complete adjustment to the daily changing conditions ....

The second theme is that simply increasing the managerial staff along with the size of the firm does not eliminate organizational diseconomies of scale. As explained by Kaldor (1934, p. 68):

> You cannot increase the supply of co-ordinating ability available to an enterprise alongside an increase in the supply of other factors, as it is the essence of co-ordination that every single decisi[illegible]

on a comparison with all the other decisions made or likely to be made; it must therefore pass through a single brain.

In our model, as in Keren and Levhari (1983) and Radner (1993), it is not exactly that the brain through which a decision must pass is overloaded as the firm size increases, but rather that the aggregation of information, which is part of coordination as described by Kaldor, involves delay that increases with problem size even when there is decentralization of information processing.

This role of computational delay is the most important distinction between our model and the static team-theory model of Geanakoplos and Milgrom (1991). Their results on returns to scale depend on assumptions about what aggregate information is available exogenously, because their model does not allow the hierarchy to aggregate information. The assumption under which they conclude that returns to scale are decreasing—that no aggregate information is available—is extreme. However, the notion that aggregate information is less available or of poorer quality than disaggregate information is supported by our model; computational delay means that aggregate information cannot be as recent as disaggregate information.

Computational delay is also an important theme in Keren and Levhari (1983) and Radner (1993). In these papers, the impossibility of eliminating delay through decentralization is straightforward. The time it takes to sum n numbers, or perform some other associative operation on n items, is at least of order $\log n$, no matter how many managers are available to perform the task.[6] In our model, the phenomenon is more complex, as is seen in the proofs of Theorems 3 and 5. There is no single measure of delay, and a decision rule can always use some very recent information; however, delay in aggregating information imposes intertemporal constraints on the lags of data upon which a decision can be based, and, in particular, puts a bound on the information of any given lag that can be used. In contrast, in the sampling problem, it is possible to obtain and use unbounded amounts of data of a given lag, at a linear cost.

Another distinction between the batch processing models of Keren and Levhari (1983) and Radner (1993) and our real-time model is that the former are not based on a decision problem and the scale of the firm is defined to be the size of the computation problem. Delay increases unboundedly with problem size and hence with the scale of the firm. Combine this with a "cost" of delay derived from a temporal decision problem, such as ours, and the conclusion is very likely to be that delay leads inexorably to eventually decreasing returns to scale: Asymptotically decisions are based on such old information that unit costs are the same as if no information were processed. Even though this would also be true when the cost of delay is derived from our decision problem, we find (Theorem 4) that there are cases in our model when returns to scale are increasing in the computation problem

[6] In contrast, two vectors of length n can be added in a fixed amount of time that does not depend on n by hiring n managers who sum the n coordinates concurrently; however, this does not involve the aggregation of n items.

because *firms choose their decision rules* and do not have to give up using recent data just because the firm size increases. This is shown in Theorem 4 for assumptions under which the firm's task is to predict a component that is common to all the processes; the proof involves demonstrating that a firm can achieve unit costs that are lower than those of a smaller firm by imitating the decision procedure of the smaller firm.

Note that in that proof, a larger firm imitates a single smaller firm, rather than replicating the activities of several smaller firms. We stated in Section 1.2 that such replication does not work when we take into account information processing constraints. Our results illustrate this breakdown of replication arguments, and link it to the aggregation delay that results from informational integration. Under the statistical assumptions in Theorem 3, we show that there are constant returns to scale in the sampling problem because a firm should replicate the optimal sampling policy of a firm of size 1. Under the assumptions in Theorem 5, we show that there are eventually increasing returns to scale in the sampling problem because a firm of size KN can achieve per-unit costs lower than those of a firm of size N by dividing itself into K divisions of equal size that imitate the sampling policy of the firm of size N (returns are eventually increasing rather than simply constant because of a diversification effect). *Such replication strategies do not work in the computation problem because each division would compute its own prediction.* The aggregation of these predictions would introduce delay, and so the decision rule would use information that is older than the information used by the smaller firm. As a result, in the computation problem, there are eventually decreasing returns to scale under the statistical assumptions for Theorem 3, and there may be a firm size that minimizes unit costs under the assumptions of Theorem 5.

2 A real-time decision problem

2.1 The prediction problem

We study the real-time computation of the following prediction problem. There are N discrete-time stochastic processes: $\{X_{it}\}_{t=-\infty}^{\infty}$ for $i = 1, \ldots, N$. Their sum, $X_t = \sum_{i=1}^{N} X_{it}$, must be predicted each period. Let A_t be the prediction in period t, which is computed from past realizations of the processes. The performance is measured by a loss function $\psi(X_t - A_t)$, where $\psi(\cdot)$ is positive and $\psi(0) = 0$.

For example, this is the decision problem of a firm that has N sales offices with demands $\{X_{1t}, \ldots, X_{Nt}\}$ in period t and that controls the level of output centrally. There is a loss when output is not equal to the total demand. This is our leading example, and we sometimes use terminology from this example for concreteness. Another example is a firm that sets its output centrally and needs to estimate the average productivity of its N machines or shops, whose individual productivity indices are $\{X_{1t}, \ldots, X_{Nt}\}$ in period t. In both examples, the level of output, and hence the scale of the firm, is proportional to the number of stochastic processes whose sum is predicted.

Unlike most decision problems one sees in economic models of firms, such as setting output in response to a single demand parameter, this decision problem has a property that is common to a variety of decision problems a firm may face and that is fundamental both to the potential for decentralizing information process and to our results on returns to scale: It involves *aggregating* information about many activities, markets or parts of the firm, whose number varies with the scale of the firm.

2.2 Decentralized computation

The computation of decisions in organizations is decentralized, in that it is performed jointly by many managers or clerks, whose numbers and organization are determined endogenously. To model this, we need to specify how each potential manager processes information and how the managers communicate so that they can process information jointly.

Human information processing is complex, but we only need a model of computation whose sophistication matches that of the decision problem. We restrict the decision rules to be linear. Such decisions rules only require addition, multiplication and communication, and so these are the only capabilities with which we endow the managers. The managers in our model are thus not very clever, but this is a consequence of the decision problem, not of an inappropriate model of computation. Endowing the managers with the ability to do differential topology or prove theorems would only clutter the model, since these tasks are not useful for the decision problem. Endowing the managers with the ability to add and multiply arbitrarily quickly would make them as good as unboundedly rational managers for the decision problem.

Hence, it is not literally the inability of managers to do arithmetic quickly—which could hardly be an important constraint today even if it was before the invention of calculators and computers—that interests us; instead, our model is a proxy for more complex computations, such as when managers must aggregate capital budget requests based on soft information about capital needs and profitability or must predict demand or productivity based on soft information about markets or worker quality. Just as the simple decision problems in economic models help us understand real-world decision problems, so do the constraints on the ability to compute the simple decision rules that arise in such models help us understand the effects of the limited human capacity to make real-world decisions.

Most results in this paper do not make use of the minor details of the computation model, but we choose to work with a specific computation model so that we can illustrate the computation of decision rules and can use the model in some constructive proofs (Theorems 1 and 5). The model is an adaptation of the model of associative computation introduced by Radner (1993), which in turn is similar to various general models of parallel computation in the computer scier including the PRAM and the Log P model of Culler et al. (1993). (See Van Zandt (1995a) for a comparison.)

Each manager has two types of memory:

1. An *infinite addressed buffer* for receiving and storing incoming data (raw observations or partial results sent by other managers). This is analogous to the manager's *in-box.*[7]

2. A *single register* for storing results of computation. This is analogous to memory in the manager's *brain.*

In one cycle (the unit of time in the computation model), a manager can perform the following sequence:

1. Read the value x from one address in her buffer.

2. Calculate $\alpha_0 + \alpha_1 x + \alpha_2 y$, where y is the current value in the register and α_0, α_1 and α_2 are constants that may vary from one operation to another but do not depend on the realizations of any of the data and hence can be written into the manager's instructions.

3. Store the result in her register, replacing the existing value.

Such a sequence is called an *operation*, and the manager is then said to have *processed*, *aggregated* or *read* x (we use these terms interchangeably).

With this capability, a manager can compute a linear function $\beta_0 + \sum_{j=1}^{J} \beta_j x_j$ of a list $x_1, \ldots, x_J$ of J numbers in J cycles by storing $\beta_0 + \beta_1 x_1$ in the register during the first cycle, then adding $\beta_2 x_2$ to this in the second cycle, and so on. This model has the simplifying feature that the complexity of a linear function does not depend on whether any of the multiplicative constants are equal to 1 or whether the additive constant is equal to 0. Counting the addition and multiplication of constants as separate operations would not change the general results, although those proofs that are constructive (Theorems 1 and 5) would require modification.

To model decentralization, we need to specify how managers communicate and coordinate their actions. For the coordination of the managers, we assume that there is a clock that synchronizes the actions of the managers, and each manager has a list of instructions that specify the action taken at each moment of time. For communication, we assume that each manager can send the contents of his register to an address of one or more managers' buffers at the beginning of each cycle, and the value arrives instantly and without transmission costs. (The value sent is called a message or report.) There are still some implicit communication costs, which we illustrate in the next section. However, the results on returns to scale in this paper are due to the managers' computational delay, rather than to communication costs.

[7]We can think of the addresses as labels that allow the manager to pick out a message or a datum from his in-box.

Our assumptions about the input of the data $\{X_{1t}, \ldots, X_{Nt}\}$ and the output of the prediction A_t in each period t are as follows. To handle input, there is a device that distributes data instantly to the buffers of specified managers at the beginning of the period in which the data arrives. This is equivalent to assuming that the data are stored in an addressed buffer from which all managers can read values. To handle output, each manager can send the value of her register to an output device at the beginning of a cycle, and the message arrives instantly. The value stored in the output device remains the current prediction until it is replaced by a new message.

A decision procedure, which is given by the lists of instructions that the managers follow and by the distribution of data to the managers, must be such that each manager has only one instruction to follow at each moment, no two messages or data arrive at the same address of the same buffer at the same time, and no two messages are sent to the output device at the same time. It is then possible to trace through the operations and messages of the managers, keeping track of the values in each address and register and the messages sent to the output device, and thereby determine the decision rule that is implemented by a decision procedure. Some features of the model, such the managers' infinite addressed buffers to which the managers have random access, were selected so that we do not need to keep track of too many details of the computation in subsequent sections of this paper. For example, if a manager sends a message to another manager, it is always possible to come up with a non-conflicting address in which the message is stored until it is no longer needed by the recipient, and it is not necessary to keep track of which address is used for each message.

2.3 Constrained-optimal decision procedures

A decision procedure is a decision rule together with a computation procedure (algorithm) that implements (computes) the decision rule. The total cost of a decision procedure at date t is the expected decision-theoretic loss, $E[\psi(X_t - A_t)]$, plus the computation cost, wq_t, where w is the wage per manager and q_t is the number of managers employed at time t.[8] We measure the performance of a decision procedure by the long-run average of the total costs.

We have modeled the process by which decisions are computed, but not the process by which decision procedures are selected, i.e., by which organizations evolve or are designed. Some of our results just characterize the feasible set of decision rules. Others are about the feasible decision procedures that minimize the long-run costs. Such decision procedures are said to be constrained optimal, or simply optimal.

If the situation to which we apply the model is sufficiently stationary and the discount rate is sufficiently close to zero, then an optimal decision procedure can

[8]With linear decision rules, the computation costs are non-random; if operations depended on past observations, then the predictions would depend in a non-linear way on these observations.

be approximated by a stationary procedure with a finite description. The costs and delays of choosing a good procedure can be amortized over the life of the firm, whereas the costs and delays associated with the computation of a decision rule are incurred repeatedly. Furthermore, decision procedures can evolve slowly but still end up being approximately optimal. Therefore, under these (artificial) conditions, we may see decision procedures that are nearly constrained optimal. Otherwise, the constrained-optimal procedures simply provide a lower bound on the organizations' costs.

The examples in Section 3 hint at the enormous variety of computation procedures that are available. Optimal procedures need not be statistically optimal, need not modify the prediction each period, and need not be stationary even when the statistical model is stationary. This makes it difficult to fully derive the optimal procedures. In Radner and Van Zandt (1992), we described some classes of "good" decision procedures for a few special cases (of statistical assumptions). The classes were sufficiently restrictive so that it was possible to select optimal procedures from these classes. This was a useful exercise that further illustrated the mechanics of the decision procedures in this model. However, in this paper, we do not need an exact description of the constrained-optimal procedures, and instead we characterize returns to scale using basic properties of the computation constraints. One implication is that the results do not depend critically on the ability of organizations to derive constrained-optimal decision procedures themselves.

3 Some properties of real-time decision making

3.1 The costs and benefits of decentralization

In this section, we describe a few simple (but not necessarily good) decision procedures. One goal is to familiarize ourselves with the real-time computation of decisions. Another is to illustrate the costs and benefits of decentralization of information processing in such a dynamic decision problem.

In the examples, each prediction is the sum of a list of data. The predictors could instead be arbitrary linear functions of the data, by introducing multiplicative and additive constants, without otherwise modifying the decision procedures.

Recall the leading example of a firm that must predict the total demand of N sales offices. Suppose that the firm has four sales offices. Suppose also that one computation cycle is equal to one decision period.

Example 1 Consider the decision rule that sets output in each period to the total demand four periods earlier:

$$A_t = \sum_{i=1}^{4} X_{i,t-4} \,.$$

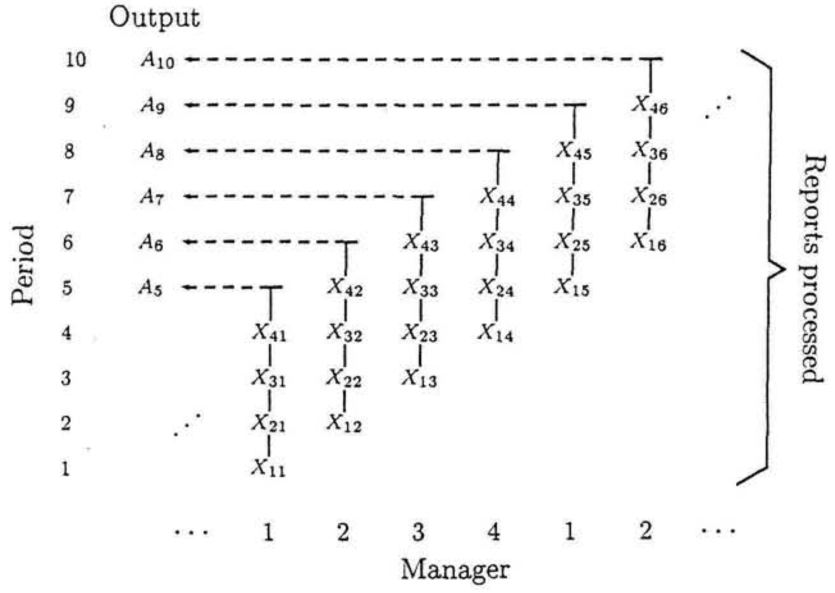

FIGURE 1. Computing the decision rule $A_t = \sum_{i=1}^{4} X_{i,t-4}$ with 4 managers (Example 1).

A decision procedure that computes this decision rule is depicted in Figure 1. The firm gives the sales reports of the four sales offices to one manager in each period. The manager who receives the reports in period 1 computes the sum in four periods and sends the result (A_5) to the production controller by period 5, as required. The manager is then free to process the reports that arrive in period 5. However, the reports that arrive in periods 2, 3 and 4 must be given to other managers. Thus, a total of four managers are needed to compute the decision rule.

Example 2 Although the task of computing the decision rule is shared among four managers in Example 1, computation is not truly decentralized because the managers do not communicate and each decision is computed by a single manager. We now consider a decision procedure with true decentralization.

The procedure is illustrated in Figure 2. In each period, the four latest sales reports are divided among two managers. Each manager receives two reports and computes their sum in two periods. One of the managers then sends her partial sum to the other manager, who adds it to his own partial sum in the next period. This manager can then send the total to the production controller, three periods after the reports were received. The decision rule that is computed is thus

$$A_t = \sum_{i=1}^{4} X_{i,t-3} .$$

This decision procedure uses five managers—one more than in in Example 1—because the reports that the managers send each other increase the workload. This is an example of the overhead cost of decentralization.

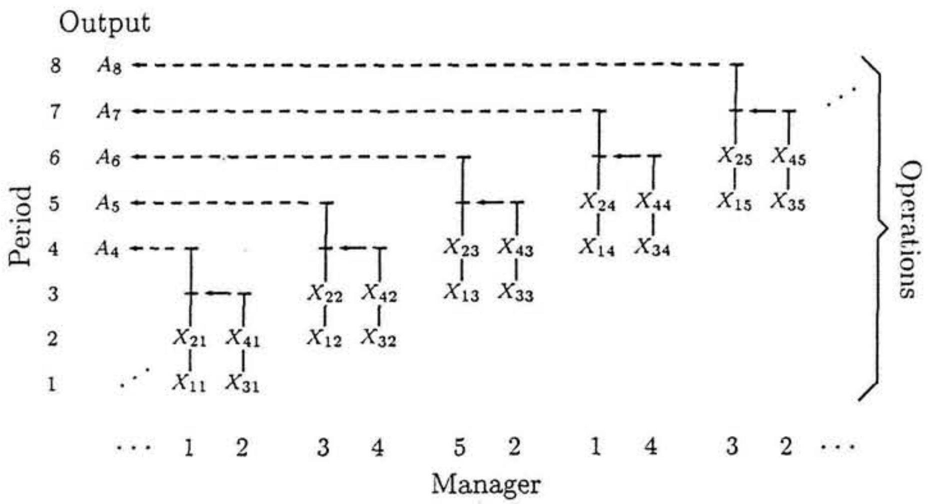

FIGURE 2. Computing the decision rule $A_t = \sum_{i=1}^{4} X_{i,t-3}$ with five managers (Example 2).

The potential decision-theoretic benefit of decentralized information processing is that delay is reduced. Each decision is computed from information that is three periods old in Example 2, compared to four periods old in Example 1. There are limits, however, to what can be achieved by decentralization. For example, the decision rule

$$A_t = \sum_{i=1}^{4} X_{i,t-2} ,$$

which sets output to the total demand two periods earlier, is computationally infeasible, because the four sales figures cannot be summed in two periods. (See Radner (1993) for bounds on speed-up from decentralization with batch processing.)

Example 3 Decentralization unambiguously reduces delay in "batch mode" models of computation, such as Keren and Levhari (1979, 1983, 1989), Radner (1993) and Van Zandt (1994), because all data are available at the same moment. The reduction in delay is also unambiguous when comparing Examples 1 and 2 because each decision is computed from data that have the same lag. However, this does not have to be the case in real-time computation.

In Example 1, manager 1 processes office 2's period-1 sales report (X_{21}) in period 2. The manager could instead process office 2's period-2 sales report (X_{22}). If we modify the procedure in Example 1 so that managers always use the most recent report when processing an office's sales report, then the following decision rule is computed:

$$A_t = X_{1,t-4} + X_{2,t-3} + X_{3,t-2} + X_{4,t-1} .$$

The decision procedure is illustrated in Figure 3. The potential advantage of this procedure over Example 1 is that some of the data is more recent.

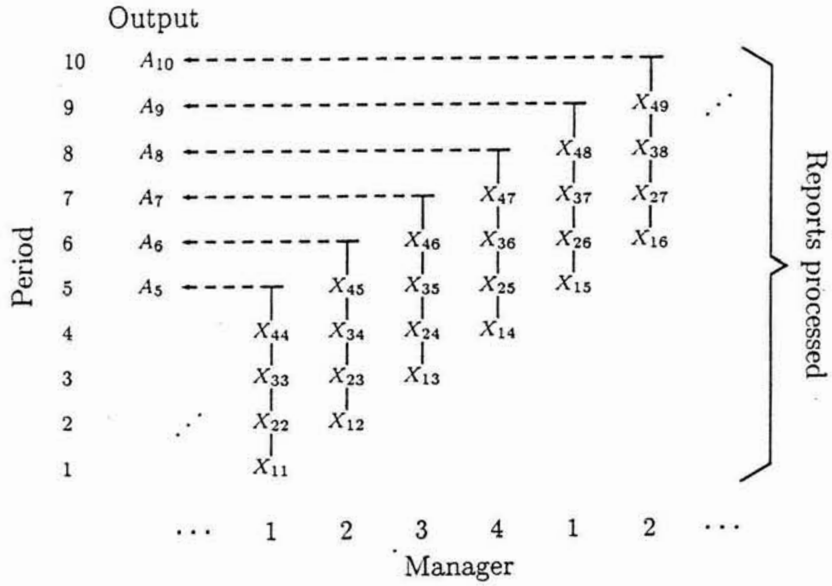

FIGURE 3. Computing the decision rule

$$A_t = X_{1,t-4} + X_{2,t-3} + X_{3,t-2} + X_{4,t-1}$$

with four managers (Example 3).

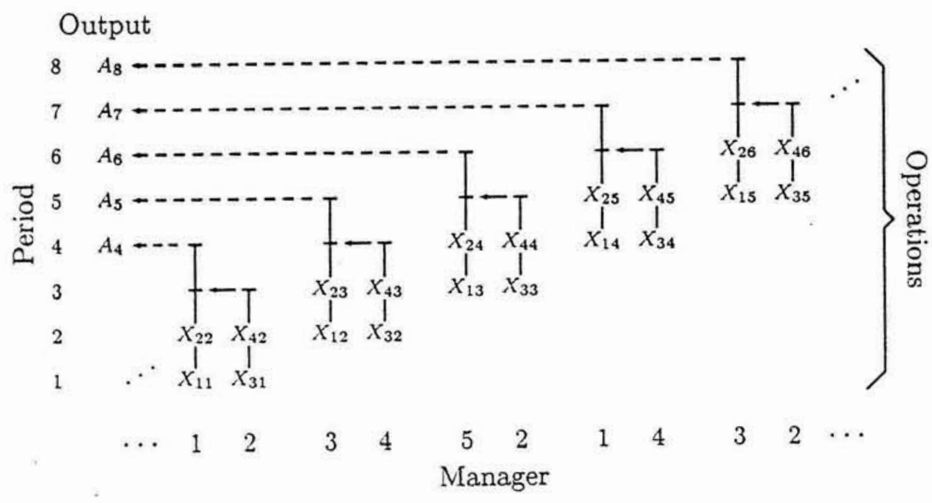

FIGURE 4. Computing the decision rule

$$A_t = X_{1,t-3} + X_{2,t-2} + X_{3,t-3} + X_{4,t-2}$$

with five managers (Example 4).

Example 4 If we also modify Example 2 so that managers use the most recent raw data, the decision rule

$$A_t = X_{1,t-3} + X_{2,t-2} + X_{3,t-3} + X_{4,t-2}$$

is computed. This is illustrated in Figure 4. Observe that the lags of the data in this procedure *do not* dominate the lags of the data in Example 3. In this procedure, the data have lags of 2 and 3. In Example 3, the data have lags that range from 1 to 4. In the last period before the decision is finished in this procedure, a manager is aggregating previously computed partial sums, whereas in Example 3 a manager is still reading in recent raw data. This illustrates a *decision-theoretic* cost of decentralization that arises in real-time parallel computation but not in models with batch processing: *When a manager listens to a subordinate, whose report may be based on a great deal of information, the manager foregoes processing raw data that are more recent than any of the data upon which the subordinate's report is based.*

Theorem 1 shows that this decision-theoretic cost of delegation can be dominant, so that there is no decentralization even when managers' wages are zero.

Theorem 1 *Assume:*

1. *The loss is quadratic:* $\psi(X_t - A_t) = (X_t - A_t)^2$.
2. *The processes are mutually independent and identically distributed, and each process is a first-order autoregressive process:*

   $$X_{it} = \gamma X_{i,t-1} + W_{it},$$

   where the random variables W_{it} *are i.i.d. across time and processes and have finite variance.*
3. *The length of a cycle is one period.*
4. *The managerial wage is 0.*

If $0 < |\gamma| \leq \sqrt{1/2}$, *then only one manager computes each action, no matter how large the problem size. If* $\sqrt{1/2} < |\gamma| < 1$, *then the number of managers that compute each action grows unboundedly as the problem size increases.*

(The proof is given in Section 5.)

Under the assumptions of the theorem, the constrained-optimal decision procedures are similar to the one illustrated in Figure 4. Each decision uses one observation from each process. Hence the workload—and the total number of managers employed—is roughly proportional to the firm size. For each decision, the processes are divided among the managers who compute the decision. The managers begin processing raw data at roughly the same time, sequentially aggregating the most recent observation from each assigned process. Then the managers aggregate the partial results hierarchically, as in the networks for associative computation studied by Radner (1993). (These properties are derived in the proof of the theorem.) When $|\gamma|$ is smaller, i.e., when the environment is changing more quickly, the decision-theoretic cost of decentralization is higher and the number of managers who compute each decision is lower. In the extreme case, when $|\gamma| \leq \sqrt{1/2}$, the value of a current raw observation is always greater than the value of all the information (none of which is current) that could be contained in a subordinate's report. Therefore, each decision is computed by a single manager.

3.2 Constrained optimality versus statistical optimality

Given a decision rule, let $\tilde{H}_t$ be the data upon which the prediction A_t is based. $\tilde{H}_t$ is the set of observations of the stochastic processes whose realizations affect A_t. In Example 4,

$$\tilde{H}_t = \{X_{1,t-3}, X_{2,t-2}, X_{3,t-3}, X_{4,t-2}\} .$$

For the moment, let us say that a decision rule is statistically optimal if it minimizes the expected loss conditional upon the information used by the decision rule:

$$A_t = \arg\min_{a_t} E\left[\psi(X_t - a_t) \mid \tilde{H}_t\right] .$$

Unlike in models without computational constraints, a constrained-optimal decision procedure is not necessarily statistically optimal, because two decision rules that use the same data can have different computation costs. It is even possible that one is feasible but the other is not.

Although this is an obvious general principle, it is not obviously relevant to our model, because we restrict attention to linear decision rules and use a computation model in which all linear functions of the same data have the same complexity. From here on, we say that a decision rule is *statistically optimal* if it minimizes the expected loss within the class of linear predictors. If only the prediction for a single period had to be computed, then any two linear decision rules using the same data would have the same computational complexity, and hence constrained-optimal decision procedures would be statistically optimal. However, because the decision problem is dynamic, a linear decision rule is not a single function, but rather a sequence of functions. Partial or final results in the computation of one prediction may be used to compute other predictions. For this reason, not all linear decision rules using the same data have the same computational complexity. We illustrate this in the following example.

Example 5 The decision rule in this example is a simple updating rule. There are four managers who do the same thing as the managers in Example 3, but instead of sending their answers to the production office, they send them to manager 0. Manager 0 is the only one who sends predictions to the production

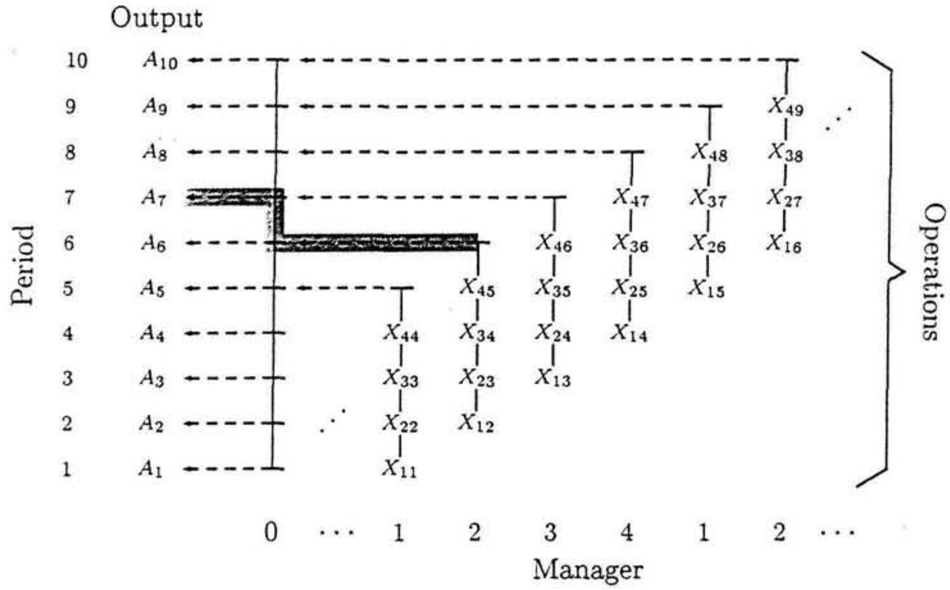

FIGURE 5. Computing the decision rule

$$A_t = \sum_{i=1}^{4} \alpha_i \left(\sum_{s=0}^{\infty} \beta^s X_{1,t-5-s+i} \right)$$

with five managers.

making prediction A_t, she receives $\sum_{i=1}^{4} \alpha_i X_{i,t-4+i}$ from one of the other managers and computes

$$A_{t+1} = \beta A_t + \sum_{i=1}^{4} \alpha_i X_{i,t-4+i}$$

in one cycle. This is the prediction for period $t+1$. The procedure is illustrated in Figure 5. Observe that each prediction is based on an infinite amount of data, even though only five managers are employed.

In this procedure, lags of the data in $\tilde{H}_t$ are stationary. Hence, if the stochastic processes are stationary, the decision rule can only be stochastically optimal if it is also stationary. This means that manager 0 must always use the same coefficient β and the other managers must always use the same coefficients α_1, α_2, α_3 and α_4. Then the following stationary decision rule is computed:

$$A_t \cdot = \sum_{i=1}^{4} \alpha_i \left(\sum_{s=0}^{\infty} \beta^s X_{i,t-5-s+i} \right) .$$

For some distributions of the stochastic processes, there may be values of α_1, α_2, α_3, α_4 and β such that this decision rule is actually statistically optimal. However, since there are only five parameters to vary, not all linear decision rules using this data can be computed this way. If we imagine that the purpose of all computation up to period 0 is to compute A_0, then by adjusting the coefficients used in that computation we can make A_0 statistically optimal, without hiring more managers. However, the changes to the coefficients will affect the predictions in period $t \neq 0$, and so it is typically not possible to make the predictions stochastically optimal in every period, without either using less data for each prediction or hiring more managers.[9] Therefore, this decision procedure can be constrained optimal even if it is not statistically optimal.

3.3 Sampling versus computation

As we shall see, the returns to scale are sensitive to the specific statistical distributions of the processes and the shape of the loss function. To distinguish the effect of the statistical assumptions from the effects of computation constraints, we also characterize the returns to scale for the standard approach to statistical decision problems—where computation is costless and instantaneous, but data is costly. We call this benchmark the *sampling problem*, and we call the main model of this paper, in which data is freely available but computation is constrained, the *computation problem*.

A special case of the sampling problem, which is included in all our results for this benchmark, is where all past observations are available at no cost. The computation problem thus consists of this special case combined with computation constraints. However, we also characterize the returns to scale of the sampling problem for more general assumptions about the sampling technology. We only assume that, in the sampling problem, the availability and cost of a datum is the same for all processes and depends only on the lag of the datum. For example, if it costs \$1 to observe yesterday's realization of one process, then it costs \$100 to observe yesterday's realization of 100 processes. For some lags, observing data may be free or impossible.

In the sampling problem, let $\tilde{H}_t$ be the data that has been sampled up through period t. $\tilde{H}_t \setminus \tilde{H}_{t-1}$ is the data that is sampled in period t. Let $s(d)$ be the cost of observing one of the processes with a lag of d. Then the sampling cost at date t is

$$\sum_{d=1}^{\infty} s(d) \left(\# \left\{ i = 1, \ldots, N \mid X_{i,t-d} \in \tilde{H}_t \setminus \tilde{H}_{t-1} \right\} \right) .$$

The total cost in each period is the decision-theoretic loss plus the sampling cost.

Because there is perfect recall and there are no computation constraints in the sampling problem, A_t is statistically optimal given $\tilde{H}_t$. This is one difference between the sampling problem and the computation problem. However, for returns to scale, the important difference is how much data of a given lag can be used in a decision. This is illustrated in Figure 6, for the case where one period equals one cycle.

[9]If $\{X_t\}$ has a rational spectrum, then $E[X_t \mid \tilde{H}_t]$ can be computed using a recursive updating rule, but the updating may involve averaging, in each period, a large but finite number k of past predictions. (See Yaglom (1962, Chapter 4) for details.) This will require at least $k - 1$ more

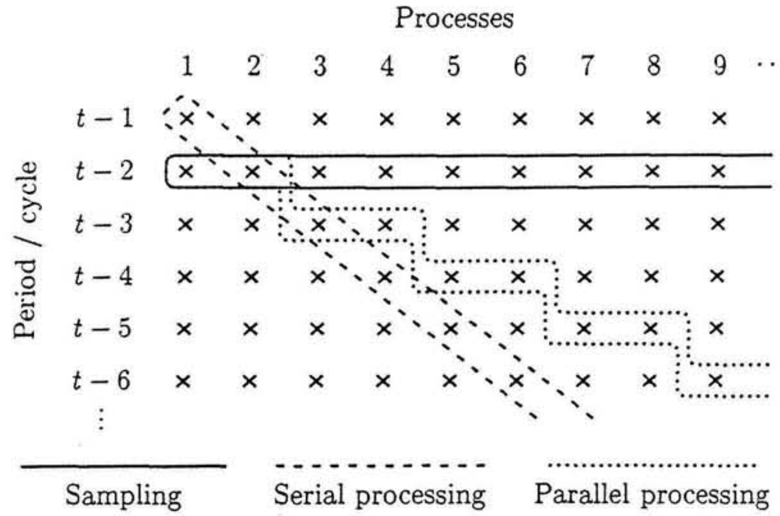

FIGURE 6. Bounds on aggregation speed. Each region shows an example of the data that could be incorporated into the prediction at time t. The region for parallel processing is for the case of two managers.

Suppose that, in the sampling problem, the prediction in period t of a firm of size 1 is based on $X_{1,t-2}$. For a firm of size N, it is possible to sample $X_{i,t-2}$ for all i, with the same per-unit sampling cost faced by the firm of size 1. The decision then uses the data surrounded by the solid line in Figure 6.

This is not possible in the computation problem. For example, suppose that each prediction is computed by a single manager. Then the decision at time t can be based on $X_{1,t-1}$, $X_{2,t-2}$, $X_{3,t-3}$, etc.; only one datum of each lag is used in the decision rule. The decision uses the data surrounded by the dashed line in Figure 6. If instead two managers compute each decision, with their partial results aggregated in the last cycle before the decision is made, then the decision in period t might be based on $X_{1,t-2}$, $X_{2,t-2}$, $X_{3,t-3}$, $X_{4,t-3}$, etc. The decision uses the data surrounded by the dotted line in Figure 6. This parallelization increases the amount of data of lag d used in the decision rule, for $d \geq 2$ (but decreases it for $d = 1$), but there are limits to the speed-up from parallelization.

It is known (e.g., Radner (1993)) that for this computation model, it takes at least $\lceil \log_2 n \rceil + 1$ cycles to aggregate n items. Inverting, this implies that a decision rule in the computation problem can use no more than 2^{d-1} observations with a lag of d or less (measured in cycles). It is this limit on the use of recent information that bounds firm size, compared to the sampling problem. Although the specific value of this limit does depend on the computational model, the existence of such a limit is a feature of all but the most unrealistic models of computation, and is certainly a property of actual human information processing. For this reason, we

managers than the procedure in Example 5, but the difference in expected loss may be arbitrarily small.

claim that the results on returns to scale that are due to the existence of such a limit are robust with respect to the model of computation.

4 Returns to scale

4.1 Statistical assumptions

For the characterization of returns to scale, we impose three assumptions on the infinite pool $\{X_{1t}\}$, $\{X_{2t}\}$, ... of potential stochastic processes.

Assumption 1 *The vector process $\{X_{1t}, X_{2t}, \dots\}$ is covariance-stationary.*

Motivation: So that long-run average costs are well-defined, so that the statistical analysis is simpler, and to be consistent with the focus on constrained-optimal decision procedures (see Section 2.3).

Assumption 2 *The processes $\{X_{1t}\}$, $\{X_{2t}\}$, ... are exchangeable (their joint distribution is symmetric).*

Motivation: So that the demand processes are statistically indistinguishable, and hence a larger firm is quantitatively larger than, but qualitatively similar to, a smaller firm.

Assumption 2 is satisfied, for example, if the stochastic processes can be written

$$X_{it} = Y_t + Z_{it} \ ,$$

where the processes $\{Y_t\}$, $\{Z_{1t}\}$, $\{Z_{2t}\}$, ... are independent and the processes $\{Z_{1t}\}$, $\{Z_{2t}\}$, ... are identically distributed. (The converse is true if the processes are Gaussian.) $\{Y_t\}$ is called the common component of the processes, and $\{Z_{it}\}$ is called the idiosyncratic component of process i. We will sometimes use this terminology even when we are not restricting attention to this case.

Assumption 3 *The vector process $\{X_{1t}, X_{2t}, \dots\}$ is purely indeterministic.*

Motivation: So that the value of information diminishes with its age.

A covariance-stationary process is called purely indeterministic if its Wold decomposition does not have a deterministic component (e.g., see Anderson (1971, pp. 420-424)), and hence can be written as an infinite-order moving average. Let H_t be the history of all processes ($i = 1, 2, \dots$) up to period t. A necessary and sufficient condition for Assumption 3 is that the covariance matrix of the error for the best linear prediction of $\{X_{1t}, X_{2t}, \dots\}$, conditional on H_{t-d}, converges to the unconditional covariance matrix of $\{X_{1t}, X_{2t}, \dots\}$ as $d \to \infty$. This property is used to prove Theorem 3. Another implication of this assumption is that if $\mathrm{Cov}(X_{it}, X_{jt}) > 0$ and

ϵ_{it} and ϵ_{jt} are errors for linear predictions of X_{it} and X_{jt}, respectively, using lagged data, then $\mathrm{Cov}(\epsilon_{it}, \epsilon_{jt}) > 0$. This property is used to prove Theorem 2.

In the theorems, we consider two possible loss functions. The first is the quadratic loss:

$$\psi(X_t - A_t) = (X_t - A_t)^2 .$$

Then we consider a case where the per-unit loss is a convex function of the per-unit error:

$$\tfrac{1}{N}\psi(X_t - A_t) = \Psi\left(\tfrac{1}{N}(X_t - A_t)\right) .$$

In the definition, Ψ is a convex function that does not depend on N, whereas ψ may depend on N. We refer to this as a "scalable loss" function. It includes the case where the quadratic loss is adjusted for firm size,

$$\psi(X_t - A_t) = \tfrac{1}{N}(X_t - A_t)^2$$

(in which case $\Psi(\epsilon) = \epsilon^2$), and it also includes the case where ψ is piecewise linear and does not depend on N:

$$\psi(X_t - A_t) = \begin{cases} \alpha_0 |X_t - A_t| & \text{if } X_t - A_t < 0 . \\ \alpha_1 |X_t - A_t| & \text{if } X_t - A_t \geq 0 . \end{cases}$$

4.2 Definitions of returns to scale

For both the computation and sampling problems, let $AC(N)$ be the per-unit cost for an organization of size N. That is, $AC(N)$ is the minimized long-run average of the decision-theoretic expected loss, plus computation or sampling cost, divided by N.

We use the following three notions of returns to scale:

Monotonic Returns to scale are *monotonically* increasing, decreasing or constant if $AC(N)$ is decreasing, increasing or constant.

Eventual Returns to scale are *eventually* increasing if $\limsup_{N' \to \infty} AC(N') < AC(N)$ for all N. They are *eventually* decreasing if $\liminf_{N' \to \infty} AC(N') > AC(N)$ for all N.

Asymptotic Returns to scale are *asymptotically* increasing, decreasing or constant, respectively, if $\lim_{N \to \infty} AC(N)$ equals 0, ∞ or K, for some $0 < K < \infty$, respectively.

An *optimal firm size* is one that minimizes $AC(N)$.

Monotonically and asymptotically increasing or decreasing returns to scale are both stronger properties than eventually increasing or decreasing returns to scale, but the existence of an *optimal firm size is determined by eventual returns to scale*. There is no optimal firm size if returns to scale are eventually increasing, and there is an optimal firm size if returns to scale are eventually decreasing.

	Statistical Assumptions	Returns to Scale: Sampling Problem	Returns to Scale: Computation Problem	
Quadratic Loss	mutually correlated	asymptotically decreasing	asymptotically decreasing	Thm 2
Quadratic Loss	mutually uncorrelated	constant (constant per-unit gain)	eventually decreasing (per-unit gain $\rightarrow 0$)	Thm 3
Scalable Loss	common process plus noise	monotonically increasing	monotonically increasing	Thm 4
Scalable Loss	general	eventually increasing	example with minimum per-unit cost	Thm 5

TABLE 1. Table of results.

4.3 Main theorems

In this section, we characterize the returns to scale of the computation and sampling problems under various statistical assumptions. The results are summarized in Table 1. All proofs are given in Section 5.

We first consider the case where the loss function is quadratic. If the per-unit error is constant, then the per-unit loss increases linearly with N. This leads to decreasing returns to scale, in both the computation and sampling problems, unless the processes are mutually uncorrelated:

Theorem 2 *If the loss is quadratic and* $\mathrm{Cov}(X_{it}, X_{jt}) > 0$ *for* $i \neq j$, *then returns to scale are asymptotically decreasing in both the sampling and the computation problems.*

However, when the processes are mutually uncorrelated, a law-of-large-numbers effect counterbalances the curvature of the loss function. This leads to constant returns to scale in the sampling problem, which is actually separable over the processes, as shown in the proof of Theorem 3. In the computation problem, however, returns to scale are still eventually decreasing—and the per-unit gain from information processing converges to zero—because of the constraints on the amount of recent information that can be aggregated. This is a negative informational externality between the sales offices; processing data about one sales office increases the lag with which data about other sales offices can be processed.

Theorem 3 *Assume that the loss is quadratic and that the processes* $\{X_{1t}\}$, $\{X_{2t}\}$, *... are mutually uncorrelated. Then returns to scale are constant in the sampling problem but eventually decreasing in the computation problem. In addition, in the* *computation problem, the per-unit loss converges (as $N \to \infty$) to the no-information per-unit loss.*

Radner and Van Zandt (1992) characterize the returns to scale and approximately optimal decision procedures in the computation problem for an example with a quadratic loss function and i.i.d. AR(1) processes, which fits the assumptions of Theorem 3.

With a quadratic loss function that does not change with firm size, larger firms have the same tolerance for errors of fixed magnitude as smaller firms. This does not hold if, for example, the loss when output exceeds demand comes from holding inventories and the inventory capacity is proportionate to firm size. The scalable loss function may then be more realistic, and so we consider it next.

When the processes are noisy versions of a common process, the task is to estimate the common process. This prediction is a "public good"; as the size of the firm grows, more data are available and the cost of the prediction can be spread among more processes. In particular, when the loss function is also scalable, a larger firm can achieve a strictly lower per-unit loss than a smaller firm simply by scaling the smaller firm's decision rule. This scaling does not increase the computational burden or the sampling cost, and thus the per-unit computation or sampling cost is strictly lower for the larger firm. Hence, in both the computation and sampling problems, returns to scale are monotonically increasing. This is a case where delay *does not* lead to decreasing returns to scale in the computation problem because the scale of the computation is endogenous.

Theorem 4 *If the loss is scalable and $X_{it} = Y_t + Z_{it}$, where $\{Y_t\}$ is a common process and the random variables Z_{it} are i.i.d. across i and t, then returns to scale are monotonically increasing in both the sampling and the computation problems.*

Radner and Van Zandt (1992) also characterize the returns to scale and approximately optimal decision procedures in the computation problem for an example that fits the assumptions of this theorem. The loss is piecewise-linear and the processes are noisy versions of a common AR(1) process.

The idea behind Theorem 4 is that a larger firm can achieve a lower per-unit loss than a small firm by imitating (but scaling) the decision rule of a *single* small firm. This is *not* an analogue of the principle that leads to non-decreasing technological returns to scale: A large firm can imitate the production processes of *several* small firms whose total size is the size of the large firm. In the sampling problem, the analogue of this—a large firm imitates the sampling policies of several small firms—does lead to eventually decreasing returns to scale when the loss function is scalable, under general statistical assumptions. This is the first part of Theorem 5. However, there is no such analogue for the computation problem. If a large firm imitates the decision rules of several small firms, it ends up with several predictions. If it attempts to aggregate these predictions, there is additional delay and so the decision rule uses information that is older than the information used by the small firms. The fact that this kind of replication is not possible does not imply that returns to scale are inexorably decreasing in the computation problem; this was shown in Theorem 4. However, we can find a robust example in which there is an optimal firm size in the computation problem, even though returns to scale are eventually increasing in the sampling problem. This is the second part of Theorem 5.

Theorem 5 *Assume that the loss function is scalable.*

In the sampling problem, per-unit costs are strictly lower for a firm of size KN than for a firm of size N, for any $K > 1$. Moreover, returns to scale are eventually increasing.

However, there is an optimal firm size in the computation problem (i) if one cycle equals one period, (ii) if the cost of managers is close to zero, (iii) if the processes have the decomposition $X_{it} = Y_t + Z_{it}$ described in Section 4.1, (iv) if $\{Y_t\}$ and $\{Z_{it}\}$ are AR(1) processes with autoregressive parameters close to 1 and with innovation terms whose variances are close to 2 and 1, respectively, and (v) if either $\Psi(\epsilon) = \epsilon^2$ or the processes are Gaussian.

5 Proofs

The following preliminary discussion is mainly relevant to the proof of Theorem 1.

Recall the definition of an operation in Section 2.2. We say that an operation affects a message if eliminating the operation could change the value of the message. To determine whether an operation affects a message, we do not have to determine the exact linear coefficients of the operations. Instead, we only have to distinguish between operations $\alpha_0 + \alpha_1 x + \alpha_2 y$ for which α_2 is zero, which we call a STORE operation because it replaces the value in the register and hence eliminates information that had previously been aggregated, and those for which $\alpha_2 \neq 0$, which we call an ADD operation because it adds a number to the value in the register and hence preserves any information that had previously been aggregated.

Now construct the following directed graph for a given procedure. The nodes are all the observations, messages and operations. There is an edge from an operation to a message if and only if the same manager performs the operation in cycle t and sends the message in cycle t', with $t' > t$, and the manager does not perform a STORE operation after cycle t and before cycle t' (that is, if and only if the operation is directly part of the calculation of the message by the manager). There is an edge from an observation or message to an operation if and only if the operation processes the observation or message. (Hence, the observation or message directly affects the operation.) Because (i) an operation can only be connected to a message that is sent at a later date, and (ii) a message or observation cannot be connected to an observation sent at an earlier date, the digraph is acyclic. The observations, messages and operations that affect a message are the predecessors of the message in the digraph.

PROOF OF THEOREM 1: When the managerial wage is zero, the organization can compute each prediction with a separate network of managers. (The only reason to use updating rules or otherwise use partial results from the computation of one prediction in the computation of another prediction is to reduce managerial costs.) Therefore, for a given period t, we can study the decision procedure for predicting X_t in isolation from predictions for other periods. In what follows, a manager means one who is involved in computing A_t, and an operation, message or observation is one that affects A_t.

When the computation of A_t is separated from that of other predictions, the linear coefficients for the operations that affect A_t can be chosen to minimize the expected loss in period t. Let $V(\tilde{H}_t)$ be the no-information minimum expected loss minus the minimized expected loss when $\tilde{H}_t$ is the information used to compute A_t. This is the value of the information $\tilde{H}_t$. The design problem is to find the procedure that maximizes $V(\tilde{H}_t)$.

Under the assumptions of this theorem, the restriction to linear estimates is not binding. Because the loss is quadratic, the prediction that minimizes the expected loss conditional on $\tilde{H}_t$ is $E[X_t|\tilde{H}_t]$; this expectation is linear because the processes are AR(1). If d_i is the lag of the most recent observation of process i in $\tilde{H}_t$, then

$$A_t \;=\; E[X_t|\tilde{H}_t] \;=\; \sum_{i=1}^{N} E[X_{it}|X_{i,t-d_i}] \;=\; \sum_{i=1}^{N} \gamma^{d_i} X_{i,t-d_i} \,.$$

(If there is no observation of process i in $\tilde{H}_t$, then we define $d_i = \infty$ and $\gamma^{d_i} = 0$.) It follows that for a constrained-optimal procedure, $\tilde{H}_t$ can contain at most one observation for each process.

The value of information in $\tilde{H}_t$ is

$$\begin{aligned} V(\tilde{H}_t) &= E[X_t^2] - E[(X_t - A_t)^2] \\ &= \sum_{i=1}^{N} E[X_{it}^2] - E[(X_{it} - \gamma^{d_i} X_{i,t-d_i})^2] \\ &= \operatorname{Var}(X_{it}) \sum_{i=1}^{N} \gamma^{2d_i} \,. \end{aligned}$$

Since $\operatorname{Var}(X_{it})$ is a constant in this formula, we can normalize $\operatorname{Var}(X_{it}) = 1$ to simplify notation. Let $\lambda = \gamma^2$, so that the value of information is $\sum_{i=1}^{N} \lambda^{d_i}$. Call λ^{d_i} the value of the observation of process i. If Ξ is the set of processes for which an observation affects a report, then the value of information in the report is defined to be $\sum_{i \in \Xi} \lambda^{d_i}$.

In Lemmas 1 through 7, we derive some properties of the computation of A_t for constrained-optimal procedures. The properties are used to prove the main results of this theorem. The lemmas also paint the following picture of the constrained-optimal procedures. Let $\mathcal{M}$ be the set of messages, including the final output, that affect A_t. Consider the directed graph whose nodes are $\mathcal{M}$ and where a message m is connected to a message m' if, in the digraph described at the beginning of this section, m is connected to an operation that is connected to m' (the manager who sends m' processes m in an operation that affects m'). This is a connected graph, since otherwise the messages not in the connected component containing the final output would not affect the final output. According to Lemma 7, each message is processed only once. Furthermore, the final output cannot be connected to any message. Therefore, there are $\#\mathcal{M} - 1$ edges, which implies that the connected graph is a tree. According to Lemma 3, each manager sends one and only one message, and so the tree also represents the communication between managers. That is, it is hierarchical as in the association computation networks in Radner (1993). Furthermore, as in Radner (1993), managers first process raw data and then process messages (Lemma 5). We can also show that, as in Radner (1993), all managers begin processsing raw data at approximately the same time.[10]

The proof of each lemma has the following format. It is shown that if a candidate procedure does not have the property stated in the lemma but has the other properties so far derived, then there is an alternate procedure for which the value $V(\tilde{H}_t)$ of information is higher.

Lemma 1 *In a constrained-optimal procedure, $\tilde{H}_t$ contains one observation for each process. Hence, $d_i < \infty$ for $i = 1, \ldots, N$.*

PROOF: Suppose that in a candidate procedure there is a process i for which no observation is in $\tilde{H}_t$ (hence, $d_i = \infty$). Pick a manager and let s be the first period the manager processes an observation for the calculation of A_t; there is such a period since A_t is calculated from finitely many observations. This manager can instead first store $X_{i,t-s-1}$ in period $t-s-1$, and add in period $t-s$ the information that was stored in the candidate procedure, without affecting subsequent operations. This change adds $X_{i,t-s-1}$ to $\tilde{H}_t$ but does not change any of the other observations in $\tilde{H}_t$. Hence, the value of information increases by λ^{s+1}. □

Lemma 2 *In a constrained-optimal procedure, if a manager processes $X_{i,t-s}$ in period $t - s'$ and the operation affects A_t, then $s = s'$.*

PROOF: That is, when a manager reads an observation for a process, she reads the most recent one. If not, then the manager can instead process $X_{i,t-s'}$ in the alternate procedure. If $X_{i,t-s}$ is processed in any other operation that affects A_t, then that operation should be eliminated. The information in the alternate procedure is then $\tilde{H}_t$, except that $X_{i,t-s}$ is replaced by $X_{i,t-s'}$. Hence, the value of information increases by $\lambda^{s'} - \lambda^s$. □

[10]If a Manager II starts processing two or more periods after a Manager I, then Manager II could first read an observation for the first process sampled by Manager I, and this would decrease the lag of the observation for that process.

Lemma 3 *In a constrained-optimal procedure, each manager sends one message that affects A_t.*

PROOF: Suppose that in a candidate procedure there is a manager I who sends one message in period $t - s_1$ and a subsequent message in period $t - s_2$, both of which affect A_t. Let manager II be the receiver of the first message and let $t - s_1'$ be the period in which she processes the message. Let $X_{i,t-s_1''}$ be one of the observations aggregated in the message. Then $1 \leq s_1' \leq s_1 < s_1''$. Consider an alternate decision procedure that differs in two ways. (1) Manager II processes $X_{i,t-s_1'}$ in period $t - s_1'$ instead of processing manager I's first message. (2) Manager I does not process $X_{i,t-s_1''}$ and does not send the first message. Instead, he changes any STORE executed between $t - s_1$ and $t - s_2 - 1$ to an ADD, so that the message he sends in period $t - s_2$ in the alternate procedure contains the same information as his first and second messages of the candidate procedure, except that it is missing $X_{i,t-s_1''}$. The only process whose observation is different for the two procedures is process i, for which the value of information has increased from $\lambda^{s'}$ to λ^{s}. Hence, the total value of information is higher in the alternate procedure. □

Lemma 4 *In a constrained-optimal procedure, the value of information in a report that is processed in period $t - s$ is greater than the value λ^s of a raw observation processed in that period.*

PROOF: Let Ξ be the set of processes for which a report is an aggregate of observations in a candidate procedure, and let $t - s$ be the period the report is processed by its recipient, manager I. Suppose that $\sum_{i \in \Xi} \lambda^{d_i} \leq \lambda^s$. Consider an alternate procedure in which manager I processes $X_{i,t-s}$ for some $i \in \Xi$ in period $t - s$ rather than processing the report. The value of the raw observation is λ^s, and so the total value of information increases if $\lambda^s > \sum_{i \in \Xi} \lambda^{d_i}$. Otherwise, the candidate and alternate procedures have the same value of information because $\lambda^s = \sum_{i \in \Xi} \lambda^{d_i}$. This implies that $\#\Xi > 1$ and the alternate procedure does not use an observation for every process. By Lemma 1, the alternate procedure is not constrained optimal and hence neither is the candidate procedure. □

Lemma 5 *In a constrained-optimal procedure, no manager processes a raw observation after processing a report.*

PROOF: Suppose that in a candidate procedure, manager I processes a report in period $t - s$ and a raw observation in period $t - s'$, with $1 \leq s' < s$. Let Ξ be the set of processes for which observations are aggregated in the report. From Lemma 4, the value $\sum_{i \in \Xi} \lambda^{d_i}$ of information in the report is greater than λ^s. Let j be the process for which manager I aggregates a raw item in period $t - s'$.

Consider the following alternate procedure which roughly involves exchanging the times in which the report and the raw observation are processed. All the managers involved in aggregating information for the report perform each of their operations $s - s'$ periods later than in the candidate procedure. By Lemma 3, these managers do not perform any operations after sending a message, and hence the only effect of this shift is that the report is sent $s - s'$ periods later. Manager I aggregates this report in period $t - s'$ rather than $t - s$, and aggregates $X_{j,t-s}$ in period $t - s$ rather than aggregating $X_{j,t-s'}$ in period $t - s'$.

The observations in the report in the alternate procedure are $s - s'$ periods newer, and so the value of the report increases from $\sum_{i \in \Xi} \lambda^{d_i}$ to $\sum_{i \in \Xi} \lambda^{d_i-(s-s')}$, which is an increase of $(\lambda^{-(s-s')} - 1) \sum_{i \in \Xi} \lambda^{d_i}$. The value of the observation of process j falls from $\lambda^{s'}$ to λ^{s}, which is a decrease of $(\lambda^{-(s-s')} - 1)\lambda^{s}$. Since $\sum_{i \in \Xi} \lambda^{d_i} > \lambda^{s}$, the total value of information is higher for the alternate procedure. □

Lemma 6 *Suppose that, in a constrained-optimal procedure, a manager processes K raw observations and either sends a message or processes his first report in period $t-s$. Then the manager processes the raw observations in periods $t-s-K$ to $t-s-1$, and the value of these observations is $\lambda^{s+1}(1 - \lambda^{K})/(1 - \lambda)$.*

PROOF: By Lemma 5, each manager first reads raw data, then perhaps reads reports, and then sends a single message. If a manager reads a raw observation in period $t - s'$, sends his report or processes a report in period $t - s$, and is idle in some period $t - s''$ between periods $t - s'$ and $t - s$, then the value of the manager's information is increased if the manager reads a raw observation in period $t - s''$ rather than in period $t - s'$. Thus, the raw observations are processed without interruption between periods $t - s - K$ and $t - s - 1$, and by Lemma 2, the value of these raw observations is

$$\sum_{k=1}^{K} \lambda^{s+k} = \lambda^{s+1} \frac{1 - \lambda^{K}}{1 - \lambda}.$$

□

Lemma 7 *In a constrained-optimal procedure, a manager processes a report only during the period it is received.*

PROOF: Suppose a manager II processes in period $t - s'$ a report that is received in period $t - s$, with $1 \leq s' < s$. It follows from Lemma 3 that all managers who aggregate information for the report can shift their operations forward by $s - s'$ periods, thereby increasing the value of the information in the report, and the report will still be ready by period $t - s'$. □

The remainder of this proof has two parts:

1. We show that if $\lambda \leq 1/2$, then the value of a subordinate's information is always lower then the value of a raw observation. That is, if a manager is reading in a partial sum set by another manager, the expected loss is reduced if instead the manager reads in a raw datum. Therefore, in the optimal procedures, a single manager computes each A_t. (Hence, $A_t = \sum_{i=1}^{N} \gamma^{i} X_{t-i}$.)

2. We show that if $\lambda > 1/2$, then there is an upper bound on the number of raw observations processed by each manager in the computation of each A_t. Since A_t should be a function of one observation from each process, the number of managers who compute each A_t is proportional to the problem size.

PART 1: Suppose that more than one manager computes A_t. Let period $t-s$ ($s \geq 1$) be the first period a report is sent, and let managers I and II be the receiver and sender, respectively. Since this is the first report sent, it is the aggregate only of raw observations processed by manager II. By Lemmas 6 and 7, the value of the information in this report is $\lambda^{s+1}(1-\lambda^K)/(1-\lambda)$, where K is the number of observations processed by manager II. By Lemma 4, the procedure is not constrained efficient if this is less than the value λ^s of a raw observation processed in period $t-s$, which is true if $\lambda/(1-\lambda) \leq 1$ or $\lambda \leq 1/2$.

PART 2: Suppose instead that $\lambda > 1/2$. Suppose manager I processes K raw observations. By Lemma 6, this takes place from periods $t-s-K$ to periods $t-s-1$, for some $s \geq 0$ and the value of this information is $\lambda^{s+1}(1-\lambda^K)/(1-\lambda)$. Call this procedure P.

Now consider a procedure like P, but in which there is an additional manager II who shares the workload of manager I. Managers I and II read observations from $\lceil K/2 \rceil$ and $\lfloor K/2 \rfloor$, respectively, of the processes that manager I was previously sampling. They do this from periods $t-s-1-\lceil K/2 \rceil$ and $t-s-1-\lfloor K/2 \rfloor$, respectively, to $t-s-2$, and then manager I adds in II's partial result in period $t-s-1$. The value of this information is

(1)

$$(K \bmod 2)\lambda^{s+1+\lceil K/2 \rceil} + 2\sum_{k=1}^{\lfloor K/2 \rfloor} \lambda^{s+1+k} \;=\; (K \bmod 2)\lambda^{s+1+\lceil K/2 \rceil} + 2\lambda^{s+2}\frac{1-\lambda^{\lfloor K/2 \rfloor}}{1-\lambda}$$

Call this procedure P^*. The difference between procedures P and P^* is like the difference between Examples 3 and 4.

Procedure P is constrained optimal only if its value of information is greater than or equal to the value (1) of information in P^*. That is:

$$\begin{aligned}
\lambda^{s+1}\frac{1-\lambda^K}{1-\lambda} &\geq (K \bmod 2)\lambda^{s+1+\lceil K/2 \rceil} + 2\lambda^{s+2}\frac{1-\lambda^{\lfloor K/2 \rfloor}}{1-\lambda} \\
(1-\lambda^K) &\geq (K \bmod 2)(1-\lambda)\lambda^{\lceil K/2 \rceil} + 2\lambda(1-\lambda^{\lfloor K/2 \rfloor}) \\
2\lambda - 1 &\leq 2\lambda^{\lfloor K/2 \rfloor + 1} - \lambda^K - (K \bmod 2)(1-\lambda)\lambda^{\lceil K/2 \rceil}\,.
\end{aligned}$$

The right-hand side converges to 0 as $K \to \infty$, and the left-hand side is positive since $\lambda > 1/2$. Therefore, this inequality can only hold for finitely many K.

This completes the proof of Theorem 1. □

Given random vectors x and I, let $\hat{E}[x|I]$ be the best linear predictor (that minimizes the mean-squared error) of x conditional on I.

PROOF OF THEOREM 2: A lower bound on the loss in both the computation and sampling problems is that of the best linear prediction $\hat{E}[X_t \mid H_{t-1}]$ conditional on H_{t-1}. (Recall that H_t is the history of all processes ($i = 1, 2, \ldots$) up to and including period t.) Let ϵ_{it} be the error for this prediction of X_{it}, i.e., $\epsilon_{1t} = \hat{E}[X_{it} \mid H_{t-1}] - X_{it}$. Then $\{\epsilon_{1t}, \epsilon_{2t}, \ldots\}$ are exchangeable. The aggregate error is $\sum_{i=1}^{N} \epsilon_{1t}$, and the per-unit expected loss is

$$\frac{1}{N} E\left[(\textstyle\sum_{i=1}^{N} \epsilon_{it})^2\right] = \operatorname{Var}(\epsilon_{it}) + (N-1)\operatorname{Cov}(\epsilon_{it}, \epsilon_{jt}) .$$

The theorem assumes $\operatorname{Cov}(X_{it}, X_{jt}) > 0$. By Assumption 3, $\operatorname{Cov}(\epsilon_{it}, \epsilon_{jt}) > 0$. Therefore, this lower bound on the per-unit expected loss increases linearly in N. □

PROOF OF THEOREM 3: Let $\tilde{H}_{it}$ be the information about process i in $\tilde{H}_t$, so that $\bigcup_{i=1}^{N} \tilde{H}_{it} = \tilde{H}_t$. Let S_{it} be the sampling cost at time t for the observations of process i, so that $\sum_{i=1}^{N} S_{it} = S_t$.

Because processes $\{X_{it}\}_{t=\infty}^{\infty}$ and $\{X_{jt}\}_{t=\infty}^{\infty}$ are mutually uncorrelated for $i \neq j$, the statistically-optimal decision rule is

$$A_t = \hat{E}[X_t|\tilde{H}_t] = \sum_{i=1}^{N} \hat{E}[X_{it}|\tilde{H}_t] = \sum_{i=1}^{N} \hat{E}[X_{it}|\tilde{H}_{it}] .$$

Because $\hat{E}[X_{it}|\tilde{H}_{it}]$ is not correlated with X_{jt} and $\hat{E}[X_{jt}|\tilde{H}_{jt}]$ for $j \neq i$, the expected loss for this decision rule is the sum of the individual mean-squared errors. The total cost is then

$$\sum_{i=1}^{N} E[(X_{it} - \hat{E}[X_{it}|\tilde{H}_{it}])^2] + S_{it} .$$

The sampling problem is thus separable over the processes. That is, the problem is to find the sampling policy for a single process that minimizes the long-run average of

$$E[(X_{it} - \hat{E}[X_{it}|\tilde{H}_{it}])^2] + S_{it} . \tag{2}$$

This policy should be used for all processes, whatever the size of the firm, and the per-unit cost for a firm is the minimized value of (2). Hence, returns to scale are constant.

For the computation problem, we show that, for any decision rule that satisfies the constraints on the amount of recent information that is processed, the per-unit gain from information processing converges to 0.

Let d_{it} be the lag of the most recent observation in $\tilde{H}_{it}$. Then the expected loss is at least

$$L_d \equiv E[(X_{it} - \hat{E}[X_{it}|X_{i,t-d}, X_{i,t-d-1}, \ldots])^2] .$$

Because of computational delay (see Sections 2 and 3), the number of processes i for which $d_{it} \leq d$ is bounded for all d, uniformly over t and N. Since each $\{X_{it}\}$ is purely indeterministic,

$$\lim_{d\to\infty} L_d = \operatorname{Var}(X_{it}) .$$

For $\delta > 0$, there is d such that L_d is within δ of $\mathrm{Var}(X_{it})$. Since $d_{it} < d$ for boundedly many i,

$$\liminf_{N\to\infty}(1/N)\sum_{i=1}^{N} L_{d_i} \geq \mathrm{Var}(X_{it}) - \delta\ .$$

This convergence is uniform over t. Since $\mathrm{Var}(X_{it})$ is also an upper bound on the per-unit loss, attained by not processing any information, the long-run per-unit expected loss converges to $\mathrm{Var}(X_{it})$ as $N \to \infty$. □

PROOF OF THEOREM 4: We will show that the per-unit expected loss for any fixed decision rule, scaled by firm size, is a decreasing function of firm size. Note that "scaled by firm size" means that the prediction in each period is multiplied by a constant that changes with the firm size, not that the data used or relative coefficients of the data change with firm size. Therefore, the computation or sampling cost of such a decision procedure is independent of firm size, and the per-unit computation or sampling cost is decreasing. Hence, a larger firm can always attain a lower per-unit cost than a smaller firm by imitating the decision procedure of the smaller firm.

For the purpose of this proof, a decision rule is simply a sequence $\{A_t\}$ of random variables such that A_t is independent of Z_{it} (because the $\{Z_{it}\}$ are serially independent). We normalize so that the scaled version for firm size N is $\{NA_t\}$. The per-unit error for a firm of size N is

$$\epsilon_t^N \equiv (Y_t - A_t) + (1/N)\sum_{i=1}^{N} Z_{it}\ .$$

Because A_t is independent of each Z_{it}, each ϵ_t^N is a noisy version of $Y_t - A_t$. Furthermore, because the riskiness of $\frac{1}{N}\sum_{i=1}^{N} Z_{it}$ is decreasing (in the Rothschild-Stiglitz sense) in N (since the $\{Z_{1t}, Z_{2t}, \dots\}$ are i.i.d.), the riskiness of ϵ_t^N is decreasing in N. Since Ψ is convex, $E[\Psi(\epsilon_t^N)]$ is decreasing in N. □

PROOF OF THEOREM 5: The sampling and computation problems are treated separately.

Sampling problem

We construct an upper bound $AC(K, N)$ on the per-unit costs of a firm of size KN such that $AC(K, N)$ is strictly decreasing in K and is always strictly less than the per-unit cost of a firm of size N.

Note that as $K \to \infty$, the difference between the per-unit costs of a firm of size KN and the per-unit costs of a firm of size $KN + j$, for $j = 1, \dots, N$, diminishes to zero. Therefore, it follows also that returns to scale are eventually increasing.

Let A_t^N be the optimal policy of a firm of size N, with sampling costs S_t^N and per-unit error ϵ_t^N. A firm of size KN can replicate the policy of a firm of size N as follows: It divides itself into K divisions of size N. Each division uses the sampling policy of the firm of size N and calculates an estimate of the sum of the processes in the division in the same way that firm N does. Let A_{kt}^N be the "decision rule" of division k, and let ϵ_{kt}^N be the per-unit error for this division. The firm then uses $\sum_{k=1}^K A_{kt}^N$ as its estimate[11] of X_t, and its per-unit error is $(1/K)\sum_{k=1}^K \epsilon_{kt}^N$. Let $AC(K,N)$ be the per-unit costs of a firm of size KN when it uses this sampling policy and decision rule.

As constructed, $\{\epsilon_t^1, \epsilon_t^2, \dots\}$ is an exchangeable sequence of random variables and the riskiness of $\frac{1}{K}\sum_{k=1}^K \epsilon_{kt}^N$ is decreasing in K, according to the following facts:

- If $\{x_1, x_2, \dots\}$ is an exchangeable sequence of random vectors and if f is a measurable function of $\{x_{i+1}, x_{i+2}, \dots, x_{i+N}\}$, then

$$\{f(x_1, \dots, x_N),\ f(x_{N+1}, \dots, x_{2N}),\ \dots\}$$

  is an exchangeable sequence of random variables.

- If $\{x_1, x_2, \dots\}$ is an exchangeable sequence of random variables, then the riskiness of $\frac{1}{K}\sum_{k=1}^K x_k$ is decreasing in K. (This is a simple extension of the same property for a sequence of i.i.d. random variables.)

The expected per-unit loss of a firm of size KN is $E\left[\Psi\left(\frac{1}{K}\sum_{k=1}^K \epsilon_{kt}^N\right)\right]$. The riskiness of $\frac{1}{K}\sum_{k=1}^K \epsilon_{kt}^N$ is decreasing in K and Ψ is convex; therefore, $E\left[\Psi\left(\frac{1}{K}\sum_{k=1}^K \epsilon_{kt}^N\right)\right]$ is decreasing in K. Hence, the expected per-unit loss is decreasing when firms replicate the decision rule of firm N. The per-unit sampling costs are constant. Hence, the overall per-unit cost $AC(K,N)$ is decreasing in K. This is the desired upper bound on per-unit costs of a firm of size KN. It is decreasing in K and it is strictly less than the per-unit costs of a firm of size N, since these are equal to $AC(1,N)$.

Computation problem

Under the additional assumptions for the computation problem, if additive constants are used optimally, the per-unit expected loss is an increasing function of the variance of the per-unit error:

- If $\Psi(\epsilon) = \epsilon^2$, then the additive constant is set so that the mean of the per-unit error is zero, and so the per-unit expected loss is equal to the variance of the per-unit error.

[11] Note that this estimate of X_t typically does not make optimal use of the information that has been sampled because the prediction of each process should be conditioned on all the information that has been sampled, rather than only the information sampled within the process's division.

- If instead the stochastic processes are Gaussian, then so are the per-unit errors. Let ϵ_1 and ϵ_2 be per-unit errors for decision rules that make optimal use of additive constants. Suppose $\mathrm{Var}(\epsilon_1) < \mathrm{Var}(\epsilon_2)$. Let η be a constant such that $\epsilon_1 + \eta$ and ϵ_2 have the same mean. Since variance is a measure of risk for Gaussian random variables and since Ψ is convex, $E[\Psi(\epsilon_1 + \eta)] < E[\Psi(\epsilon_2)]$. Since the decision rules use constants optimally, $E[\Psi(\epsilon_1)] \leq E[\Psi(\epsilon_1 + \eta)]$.

Hence, without loss of generality, we measure (ordinally) the per-unit loss by the mean of the squared per-unit error.

By assumption, $\{Y_t\}$ and $\{Z_{it}\}$ are AR(1) processes, which we write

$$\begin{aligned} Y_t &= \gamma Y_{t-1} + V_t \\ Z_{it} &= \beta Z_{i,t-1} + W_{it} \, . \end{aligned}$$

$\{V_t\}$ and $\{W_{it}\}$ are noise processes. Assume that $\gamma = \beta = 1$. The processes are then not stationary. However, we will claim later that our calculations vary continuously with γ and β, and we will only need γ and β close to 1. Setting $\gamma = \beta = 1$ now simplifies the calculations. Assume also that $\mathrm{Var}(V_t) = 2$ and $\mathrm{Var}(W_{it}) = 1$.

In this proof, firm N means a firm of size N.

Intuition: Firm 1 can make a good prediction of X_t just by observing $X_{1,t-1}$. Firm 1 cannot differentiate Y_t and Z_{it} with this data, but it does not need to. Firm N (for large N) cannot make such a good prediction of each X_{it}, because it cannot use recent data about most of the processes. However, for large N what is really important is predicting Y_t (a law-of-large-numbers effect diminishes the per-unit loss from errors in predicting the Z_{it}). The idea that predicting Y_t is all that matters for large N applies to the sampling problem as well, but in the sampling problem, if firm 1 can observe $X_{1,t-1}$, then firm N can observe $X_{i,t-1}$ for each i. Hence, for large N, it gets many noisy observations of Y_{t-1} and can get a precise estimate of Y_{t-1}. In the computation problem, however, the number of noisy observations of Y_{t-s} is bounded for each s, and so the prediction of Y_t may have a greater mean-squared error than firm 1's prediction of X_t.

MSE for firm 1: Firm 1 can compute $E[X_t | X_{1,t-1}] = X_{1,t-1}$ with one manager, and this is the statistically-optimal prediction since $X_t = X_{1t} = X_{1,t-1} + V_t + W_{1t}$. The mean-squared per-unit error is

$$E[(V_t + W_{1t})^2] \;=\; \mathrm{Var}(V_t) + \mathrm{Var}(W_{1t}) \;=\; 3 \, .$$

MSE for firm N: We construct a lower bound on the mean-squared per-unit error for large firms.

Consider the data from dates $t-1$ and $t-2$ that firm N can use to compute A_t:

Case 1 $X_{i,t-1}$ and $X_{i,t-2}$, for some i.
Case 2 $X_{i,t-1}$ and $X_{j,t-2}$, for some i and some $j \neq i$.
Case 3 $X_{i,t-2}$ and $X_{j,t-2}$, for some i and some $j \neq i$.

(See Figure 6.) In addition, the firm could use data from periods $t-3$ and earlier, but any such data is strictly less valuable than being able to condition the prediction on Y_{t-3} and $\{Z_{1,t-3}, \ldots, Z_{N,t-3}\}$. Thus, to construct a lower bound on the MSE, we assume that firm N can compute a statistically-optimal prediction of X_t, conditional on one of the pairs of data listed above together with Y_{t-3} and $\{Z_{1,t-3}, \ldots, Z_{N,t-3}\}$.

Consider Case 2. E.g., the firm uses $X_{1,t-1}$ and $X_{2,t-2}$. Given that the firm knows Y_{t-3} and $Z_{i,t-3}$, and hence $X_{i,t-3}$, the problem is to estimate

$$B \equiv (X_t - X_{t-3})/N$$

from

$$\begin{aligned} B_1 &\equiv X_{1,t-1} - X_{1,t-3} = V_{t-1} + W_{1,t-1} + V_{t-2} + W_{1,t-2} \\ B_2 &\equiv X_{2,t-2} - X_{2,t-3} = V_{t-2} + W_{2,t-2} \,. \end{aligned}$$

Let $\hat{B} = \hat{E}[B \mid B_1, B_2]$. Since $E[B] = E[B_1] = E[B_2] = 0$, $\hat{B} = \alpha_1 B_1 + \alpha_2 B_2$ for some constants α_1 and α_2 (i.e., there is no constant term). Given the estimate $\hat{B}$ of B, the firm sets $A_t = N\hat{B} + \sum_{i=1}^{N} X_{i,t-3}$. Then the mean-squared per-unit error $E[(B - \hat{B})^2]$ is equal to

$$\begin{aligned} (3)\quad & E\left[\left(V_t + V_{t-1} + V_{t-2} + (1/N)\textstyle\sum_{i=1}^{N}(W_{it} + W_{i,t-1} + W_{i,t-2}) - \alpha_1 B_1 - \alpha_2 B_2\right)^2\right] \\ &= E\left[(V_t + V_{t-1} + V_{t-2} - \alpha_1 B_1 - \alpha_2 B_2)^2\right] \\ &\qquad + (1/N^2) E\left[\left(\textstyle\sum_{i=1}^{N} W_{it} + W_{i,t-1} + W_{i,t-2}\right)^2\right] \\ &= E\left[\left(V_t + (1-\alpha_1)V_{t-1} + (1-\alpha_1-\alpha_2)V_{t-2}\right.\right. \\ &\qquad \left.\left. -\alpha_1(W_{1,t-1} + W_{1,t-2}) - \alpha_2 W_{2,t-2}\right)^2\right] \\ &\qquad\quad + (1/N)(\mathrm{Var}(W_{it}) + \mathrm{Var}(W_{i,t-1}) + \mathrm{Var}(W_i, t-2)) \\ &= 2 + 2(1-\alpha_1)^2 + 2(1-\alpha_1-\alpha_2)^2 + 2\alpha_1^2 + \alpha_2^2 + 3/N \,. \end{aligned}$$

Solving the first-order conditions for minimization yields $\alpha_1 = 4/7$ and $\alpha_2 = 2/7$. The minimized value of (3) is thus $3 + 1/7 + 3/N$, which is greater than 3. Similar calculations show that the minimized mean-squared per-unit error for Cases 1 and 3 are even larger ($3 + 1/3 + 3/N$ and $4 + 4/25 + 3/N$, resp.).

Hence, for large firms, the mean-squared per-unit error is greater than 3, which is the mean-squared per-unit error that we calculated for a firm of size 1.

Note that this prediction problem, using Y_{t-3} and $Z_{i,t-3}$, involves extrapolating the processes only finitely many periods. Therefore, this lower bound on the per-unit loss depends continuously on γ and β. The per-unit loss and the managerial costs for firm size 1 also depend continuously on these parameters and on the wage. By setting γ and β close enough to 1 and w close enough to 0, we can still find N' and $\delta' > \delta > 0$ such that the mean-squared per-unit error plus per-unit managerial costs of firms of size $N \geq N'$ is greater than δ' and the mean-squared error plus managerial costs of firms of size 1 is less than δ. Therefore, there is an optimal firm size. □

References

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* New York: John Wiley and Sons.

Arrow, K. and Hurwicz, L. (1960). Decentralization and computation in resource allocation. In R. W. Pfouts (Ed.), *Essays in Economics and Econometrics* (pp. 34–104). Chapel Hill: University of North Carolina Press.

Beckmann, M. J. (1977). Management production functions and the theory of the firm. *Journal of Economic Theory*, 14, 1–18.

Bolton, P. and Dewatripont, M. (1994). The firm as a communication network. *Quarterly Journal of Economics*, 109, 809–839.

Culler, D., Karp, R., Patterson, D., Sahay, A., Schauser, K., Santos, E., Subramonian, R., and von Eicken, T. (1993). Log P: Towards a realistic model of parallel computation. University of California, Berkeley.

Dickinson, H. D. (1939). *Economics of Socialism.* Oxford: Oxford University Press.

Geanakoplos, J. and Milgrom, P. (1991). A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and International Economies*, 5, 205–225.

Hayek, F. A. v. (1940). Socialist calculation: The competitive 'solution'. *Economica*, 7, 125–149.

Hurwicz, L. (1960). Optimality and informational efficiency in resource allocation mechanisms. In K. Arrow, S. Karlin, and P. Suppes (Eds.), *Mathematical Methods in the Social Sciences.* Stanford: Stanford University Press.

Hurwicz, L. (1977). On the dimensionality requirements of informationally decentralized pareto-satisfactory processes. In K. J. Arrow and L. Hurwicz (Eds.), *Studies in Resource Allocation Processes.* Cambridge: Cambridge University Press.

Kaldor, N. (1934). The equilibrium of the firm. *Economic Journal*, 44, 70–71.

Keren, M. and Levhari, D. (1979). The optimum span of control in a pure hierarchy. *Management Science*, 11, 1162–1172.

Keren, M. and Levhari, D. (1983). The internal organization of the firm and the shape of average costs. *The Bell Journal of Economics*, 14, 474–486.

Keren, M. and Levhari, D. (1989). Decentralization, aggregation, control loss and costs in a hierarchical model of the firm. *Journal of Economic Behavior and Organization*, 11, 213–236.

Lange, O. (1936). On the economic theory of socialism: Part one. *Review of Economic Studies*, 4, 53–71.

Lange, O. (1937). On the economic theory of socialism: Part two. *Review of Economic Studies*, 4, 123–142.

Malinvaud, E. (1967). Decentralized procedures for planning. In *Analysis in the Theory of Growth and Planning* (pp. 170–208).

Marschak, J. (1955). Elements for a theory of teams. *Management Science*, 1, 127–137.

Marschak, J. and Radner, R. (1972). *Economic Theory of Teams.* New Haven: Yale Univeristy Press.

Marschak, T. and Reichelstein, S. (1994). Network mechanisms, informational efficiency, and hierarchies. Haas School of Business, University of California, Berkeley.

Marschak, T. and Reichelstein, S. (1995). Communication requirements for individual agents in networks and hierarchies. In J. Ledyard (Ed.), *The Economics of Informational Decentralization: Complexity, Efficiency and Stability.* Boston: Kluwer Academic Publishers.

McAfee, R. P. and McMillan, J. (1995). Organizational diseconomies of scale. University of Texas (Austin) and University of California (San Diego).

Melumad, N., Mookherjee, D., and Reichelstein, S. (1995). Hierarchical decentralization of incentive contracts. *The Rand Journal of Economics*, 26, 654–672.

Mookherjee, D. and Reichelstein, S. (1995a). Budgeting and hierarchical control. Boston University and Haas School of Business (UC Berkeley).

Mookherjee, D. and Reichelstein, S. (1995b). Incentives and coordination in hierarchies. Boston University and Haas School of Business (UC Berkeley).

Mount, K. and Reiter, S. (1974). The informational size of the message space. *Journal of Economic Theory*, 8, 161–192.

Radner, R. (1961). *The Evaluation of Information in Organizations*, volume 1. Berkeley: University of California Press.

Radner, R. (1962). Team decision problems. *Annals of Mathematical Statistics*, 33, 857–881.

Radner, R. (1993). The organization of decentralized information processing. *Econometrica*, 62, 1109–1146.

Radner, R. and Van Zandt, T. (1992). Information processing in firms and returns to scale. *Annales d'Economie et de Statistique*, 25/26, 265–298.

Robinson, E. A. G. (1958). *The Structure of Competitive Industry*. Chicago: The University of Chicago Press.

Van Zandt, T. (1994). The scheduling and organization of periodic associative computation. Princeton University.

Van Zandt, T. (1995a). Organizations with an endogenous number of information processing agents. Princeton University.

Van Zandt, T. (1995b). Real-time hierarchical resource allocation. Princeton University.

Williamson, O. E. (1967). Hierarchical control and optimum firm size. *The Journal of Political Economy*, 75, 123–138.

Williamson, O. E. (1975). *Markets and Hierarchies, Analysis and Antitrust Implicatons*. New York: Free Press.

Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. New York: Free Press.

Yaglom, A. M. (1962). *An Introduction to the Theory of Stationary Random Functions*. Englewood Cliffs, NJ: Prentice-Hall.