Economic Theory **17**, 545–575 (2001)

Economic Theory

© Springer-Verlag 2001

# Real-time decentralized information processing and returns to scale[*]

**Timothy Van Zandt**[1,**] **and Roy Radner**[2]

[1] INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, FRANCE
(e-mail: tvz@econ.insead.fr)
[2] Stern School of Business MEC 9-68, New York University, 44 West Fourth Street, New York, NY 10012, USA (e-mail: rradner@stern.nyu.edu)

November 24, 1999; revised version: March 14, 2000

**Summary.** We use a model of real-time decentralized information processing to understand how constraints on human information processing affect the returns to scale of organizations. We identify three informational (dis)economies of scale: diversification of heterogeneous risks (positive), sharing of information and of costs (positive), and crowding out of recent information due to information processing delay (negative). Because decision rules are endogenous, delay does not inexorably lead to 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 how information processing constraints together with the requirement of informational integration cause a breakdown of the replication arguments that have been used to establish nondecreasing technological returns to scale.

**Keywords and Phrases:** Returns to scale, Real-time computation, Decentralized information processing, Organizations, Bounded rationality.

**JEL Classification Numbers:** D83, D23.

[*] We would like to thank Patrick Bolton, 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. Previous versions of this paper (which initially had the title "Information Processing and Returns to Scale of a Statistical Decision Problem") contained material that now appears in Van Zandt (199b).

[**] This author was supported in part by grants SES-9110973, SBR-9223917, and IRI-9711303 from the National Science Foundation, by CORE Research Fellowship (1993–1994), and by grant 26 of the "Pôle d'Attraction Interuniversitaire" program of the Belgian Government.

*Correspondence to*: T. Van Zandt

# 1 Introduction

## 1.1 Motivation

The purpose of this paper is to study formally whether and how human information processing constraints can limit the scale of centralized decision making in organizations with endogenous administrative staffs. This abstract question is relevant, for example, to the theory of firms and industrial organization, given that decision making appears to be more centralized when an industry is controlled by a single firm than when output is produced by independent firms. Hence, any advantages to decentralized decision making may limit the scale of firms.

We address this question by characterizing the average cost curve for a statistical decision problem that exhibits centralized decision making, in a model in which decisions are made in real time by an endogenous number of boundedly rational agents. A related model was introduced in Radner and Van Zandt (1992); here we develop a new axiomatic computation model, contrast it with a benchmark sampling problem, and provide more extensive and precise results. In the spirit of the theory of teams, we restrict attention to informational and computational decentralization, leaving aside issues of incentives and governance. The administrative agents in our model are boundedly rational because it takes them time to process and use information. This time represents both managerial wages that must be paid and also, more critically, decision-theoretic delay that constrains the use of recent information. The main theme of this paper is that such delay can lead to decentralization of decision making and bounded firm size – confirming, as stated by Hayek (1945, p. 524), that "we need decentralization because only thus can we ensure that the knowledge of the particular circumstances ... be promptly used".

## 1.2 Real-time decentralized information processing

We use a real-time computation model – that is, a model where computation constraints are embedded into a temporal decision problem in which data arrive and decisions are made at multiple epochs. Such a model, whose properties are explored in Van Zandt (1999b), captures in a sophisticated way the fact that human information processing constraints limit the use of recent information.[1]

The decision problem we study is the estimation in each period of the sum of $n$ discrete-time stochastic processes. This is one of the control problems faced by a firm or plant that sets its production level centrally in order to meet the uncertain total demand of $n$ sales offices or customers,[2] or by a firm or plant

[1] Marschak (1972) was the first economic model of real-time processing (that we are aware of). He studied how different price adjustment processes affect delay, but he did not study decentralization of information processing and the effects of problem size.

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

that needs to estimate the average productivity of $n$ workers (machines or shops) based on past individual productivity indices. This decision problem is also part of resource allocation problems – such as allocating capital to $n$ projects or assigning output orders to $n$ production shops – in which one of the steps is aggregating profit, cost or productivity indices in order to calculate a shadow price. The size or scale of the decision problem is $n$.

An essential ingredient of our model is the representation of constraints on individual administrators' ability to process information. Although the data in our model are numerical because we need to impose tractable statistical assumptions, the decision problem is a proxy for (and may actually include) the truly complex tasks performed by human administrators in organizations. Such tasks use soft information that must be substantiated in lengthy reports, and the process of reading reports, understanding the information, drawing conclusions, and communicating these conclusions to others is difficult and – most importantly – it takes *time*.

Whereas Radner and Van Zandt (1992) and Van Zandt (1999b) represent such constraints by an explicit model of computation, in this paper we represent them by axioms on an abstract set of decision procedures. This axiomatic approach has the advantage of highlighting the robust properties of the computation constraints that drive the main results, and it complements the concrete "hands-on" approach in the other two papers. The axioms allow for decentralized processing, meaning that multiple agents jointly calculate decision rules; therefore, the scale of centralized decision making (or of firms) is not artificially limited by a presumption that both large and small decision problems must be solved by a single person. Nevertheless, even with decentralized processing, there are bounds on the amount of recent information that can be incorporated into each decision. These bounds – which arise because information must be aggregated when decision making is centralized – are the fundamental constraint we impose on the set of feasible decision procedures.

### *1.3 Returns to scale of centralized decision making and of firms*

The exercise in this paper is to compare the overall costs of a decision problem of size $n$ with the overall costs when the processes are partitioned, thereby replacing the single decision problem with several smaller ones whose sizes sum to $n$. We consider that decision making is centralized within a single decision problem because a single decision is made each period; this is true even though information processing may be decentralized. We view the partitioning of a decision problem as decentralization of decision making, because multiple decisions are made using different information. Thus, one way to state the exercise in this paper is that we characterize when it is optimal to decentralize decision making and when there is a bound (uniform over $n$) on the size of each decision problem in optimal partitions.[3]

This exercise is relevant to the scale of firms. It has long been understood that limits to the scale of production in a firm cannot be explained by production technology alone, because a large firm could *replicate* the production processes of several small firms and thereby achieve nondecreasing returns to scale. Instead, these limits must be due to differences in the organization properties of one large firm compared to several small firms. One of these differences is that the scale of coordination and centralized decision making is greater within a single large firm – with its headquarters and tight bureaucratic procedures for coordinating the parts and making such common decisions as total output.[4] Even if such centralization is dysfunctional at large scales, a large firm cannot overcome this by *replicating the organizational features* of several smaller firms because such replication would literally turn the large firm into several smaller firms.[5] Thus, diseconomies to centralized decision making may also limit firm size.

We emphasize that by a firm we mean an enterprise – as this term is used in Chandler (1966) – rather than merely a legal entity. For example, in the construction of a large building, many independent contractors work together. During the project, they continue to maintain their independent identities, but they also give up some of their autonomy because of the tight coordination that is required by the project. This paper considers whether there are organizational limits to the scale of such enterprises. As another example,[6] consider two farmers each owning a piece of land and a small tractor. Suppose that they decide to buy a big tractor and cultivate all the land together (not merely share the tractor). Then they have formed an enterprise that did not exist prior to the merger, even if each farmer continues to own his or her own land or to maintain a separate business identity for certain purposes. The farmers would nearly always form a legal partnership after such a merging of operations, but a lack of legal status would not eliminate the economic status of the enterprise. The joint operations involve collective decision making about the cultivation of the land, and the aggregation of information about soil qualities of the two pieces of land and the markets served by the two farmers. Leaving the technological returns aside, such collective decision making may have certain benefits, such as the sharing of

[3] A limitation of this exercise is that the class of decision problems permits only a stark view of decentralization: Problems can only be split into components among which there is no coordination. Geanakoplos and Milgrom (1991) study “internally” decentralized decision making in a resource allocation problem, but in a static team theory model. In subsequent research by Van Zandt (1998b, 1999c, d), decentralized decision making is studied in a temporal version of the resource allocation problem with real-time decentralized computation.

[4] This is documented, for example, in Kaldor (1934), Coase (1937), Robinson (1958), and Chandler (1966).

[5] For example, the subunits could not communicate, coordinate their activities, or allocate resources except as independent firms would do. Even if there continues to exist a common entity that owns the subunits, these subunits would be independent firms, just as the common ownership of the many publicly traded corporations by overlapping sets of stockholders and investments firms does not erase the boundaries between these corporations.

[6] This example is borrowed from comments of an anonymous referee.

information, and certain disadvantages, such as delays in aggregating information. These are precisely the issues we study in this paper.

To capture in a simple and concrete way that decision making is more centralized within a single large firm than among multiple small firms, we identify each decision problem in our model with a single firm. This is consistent with the examples of the decision problem given in Section 1.2, where the decision variable for each problem is a level of output. There is always some centralized control over a firm's total output, but very little coordination of output levels of different firms in the same industry. In these examples, our measure *n* of scale is proportional to the level of output, which is the usual measure of scale.

Our approach sheds new light on how bounded rationality limits firm size, but it has limitations which could be addressed in future research. First, our identification of a firm with a single centralized decision problem introduces two biases. On the one hand, because there is also decentralized decision making within firms, which is not allowed for by our model, we may underestimate the scale of firms. On the other hand, because there is some coordination among firms – through anonymous market interactions and also through contractual relationships – which is also not captured by our model, we may overestimate the scale of firms.[7]

Second, it would be useful to integrate our complexity-based modeling of organizational decision making with the incentives-based property-rights theory of firms (see Hart (1995) for an overview). The latter literature explicitly models coordination both within and among firms, and emphasizes that the legal ownership of the assets that make up firms is explained by the control rights and residual claims conferred by such ownership. On the other hand, it does not model the computational burden that such control entails, nor does it explain the structure of organizational decision making within large corporations – a structure that, to many observers (e.g. Chandler, 1966, 1990), defines the boundaries and internal structure of enterprises. Although the two paradigms capture quite different aspects of firms, they are not inconsistent. That ownership confers control (decision) rights means that decision making is more centralized when all assets and activities are within a single firm, compared to when they are dispersed across firms.

[7] While recognizing these limitations, we note that most other models of organizational returns to scale are also based on an ad hoc identification of a firm as some informationally integrated unit. For example, Williamson (1967) defines a firm to be a hierarchy with an exogenous managerial production function. Keren and Levhari (1983) define a firm to be a hierarchy with coordination delay that could be derived from a model of associative computation. Radner (1993, Section 7) defines a firm to be a network for aggregating cohorts of data. Geanakoplos and Milgrom (1991) and Van Zandt (1999d) define a firm to be a group of shops to which resource allocations are coordinated by a hierarchy.

### *1.4 Summary of results*

We identify three determinants of returns to scale:

**Diversification effect** The variance of the total demand increases more slowly with firm size – and hence returns to scale tend to be higher – when markets are subjected to independent rather than common shocks.

**Arrow effect** When information about one stochastic process is useful for estimating other processes, information processing costs can be amortized over a larger number of processes in large organizations, leading to higher returns to scale. Arrow (1974, Chapter 2) highlighted such information sharing as a source of positive returns to centralization and mergers.

**Aggregation delay effect** Computation delay imposes constraints on the amount of recent information that can be incorporated into decisions, and it creates a *negative* externality among the stochastic processes in the computation problem: Recent information about one process crowds out recent information about other processes, leading to lower returns to scale.

In order to distinguish between the effects of information processing constraints and the effects of statistical assumptions, we also characterize the returns to scale of a benchmark model in which information may be costly but computation is unconstrained. This benchmark is similar in spirit to Wilson (1975), who studied the Arrow effect in a statistical decision model of firm. We refer to this benchmark as the *sampling problem*, and to our main model – in which data are freely available but computation is constrained – as the *computation problem*. The key difference between the sampling and computation problems is that the aggregation delay effect is not present in the former.

Under the assumptions of one of our theorems (Theorem 3), only the diversification and Arrow effects matter, and returns to scale are increasing in both the computation and sampling problems. This result illustrates that, even in the computation problem, delay does not increase inexorably with the scale of the decision problem. Instead, because decision rules are endogenous and can use data of heterogeneous lags, organizations can use recent information even for large problem sizes. The proof of this theorem involves showing that a large firm can achieve lower costs than a small firm by *imitating* (not replicating) the small firm's computation or sampling procedure.

In contrast, in Theorems 2 and 4, the negative informational externality due to delay is important. As a result, returns to scale are more likely to be decreasing when computation constraints, rather than sampling costs, limit the information upon which decisions are conditioned. The proofs of these two theorems illustrate how replication arguments, which have been used to show nondecreasing technological returns to scale, also work in the sampling problem but break down in the computation problem. Furthermore, the proofs link this breakdown to aggregation delay and the informational integration implied by centralized decision making.

Specifically, under the assumptions of Theorem 2, we show that there are constant returns to scale in the sampling problem because a firm should replicate the optimal sampling procedure of a firm of size 1. Under the assumptions of Theorem 4, we show that there are eventually increasing returns to scale in the sampling problem because a firm of size $mn$ can achieve average costs lower than those of a firm of size $n$ by dividing itself into $m$ divisions of equal size that imitate the sampling procedure of the firm of size $n$. *Such replication strategies do not work in the computation problem because each division would compute only its own forecast*. The aggregation of these forecasts would introduce delay, and so the decision rule would use information that is older than the information used by the smaller firm. Consequently, in the computation problem, there are eventually decreasing returns to scale under the assumptions of Theorem 2 and there may be a firm size that minimizes average costs under the assumptions of Theorem 4.

Empirical research in this area beyond case studies is limited. Brynjolfsson et al. (1994) measure the impact of information technology (IT) on firm size and find that it is linked to smaller firm size. Heuristically, if we claim that firm size is limited in part by managerial delay, then improvements in IT should instead lead to larger firm size (although we do not perform such a comparative statics exercise). However, in a general equilibrium model, improvements in IT also mean that each firm's competitive environment is changing more quickly, and this aggravates the effect of managerial delay. Brynjolfsson and Hitt (1998) find positive correlation between demand for IT and decentralization of decision making within firms. This is a link between hardware and the structure of *human* decision making that our model is not rich enough to capture, but heuristically this might contradict our conclusion that information processing constraints limit centralized decision making. Alternatively, it may mean tha t firms that operate in rapidly changing environments respond by both decentralizing decision making and improving IT. Further research is needed to resolve these theoretical and empirical issues.

## 2 The decision problem

We study the real-time computation of a family of forecasting problems that are parameterized by their size or scale $n$, a strictly positive integer. Our goal is to compare decision problems of different sizes. In the definitions that follow, the exogenous components that vary with $n$ are indexed by $n$, whereas the endogenous components are not.

Let $\mathbb{Z}$ denote the set of integers and $\mathbb{N}$ the set of strictly positive integers. We fix once and for all a countably infinite set of potential discrete-time stochastic processes, indexed by $i \in \mathbb{N}$, from which the processes that enter into each decision problem are drawn. Process $i$ is denoted by $\{X_{it}\}_{t=-\infty}^{\infty}$ or simply $\{X_{it}\}$. The decision problem of size $n$ involves forecasting the sum $X_t^n \equiv \sum_{i=1}^{n} X_{it}$ of the first $n$ processes at the beginning of each period $t \in \mathbb{N}$, based on their past realizations.[8]

A *forecast* $A_t$ of $X_t^n$ is a random variable measurable with respect to the history $\{X_{1,t-d}, \ldots, X_{n,t-d}\}_{d=1}^{\infty}$. A *policy* is a sequence $\{A_t\}_{t=1}^{\infty}$ of forecasts (also denoted $\{A_t\}$). The *long-run loss* of a policy $\{A_t\}$ is defined as follows. There is a *loss function* $\psi^n$ satisfying $\psi^n(0) = 0$ and $\psi^n(\epsilon) > 0$ if $\epsilon \neq 0$. (Additional assumptions on $\psi^n$ are stated in Section 4.2.) The period-$t$ loss is $\psi^n(X_t^n - A_t)$. Let $L_t \equiv E[\psi^n(X_t^n - A_t)]$ be the expected loss for $t \in \mathbb{N}$. Then the long-run loss of $\{A_t\}$ is denoted $\Gamma(\{L_t\})$. (The domain and other properties of the function $\Gamma$ are specified in Section 4.3.)

Several interpretations of this decision problem were given in Section 1.2. One was that the decision problem is of a firm that has $n$ sales offices or markets with demands $\{X_{it}\}_{i=1}^{n}$ in period $t$ and that controls the level of output centrally. There is a loss when output is not equal to the total demand. We shall use terminology from this example for concreteness – in particular, keeping in mind the application to returns to scale of firms, we shall identify the decision problem with a single firm. See Section 1.3 for an explanation and caveats with regard to this last point.

Unlike most decision problems one sees in economic models of firms (e.g., setting output in response to a single demand parameter), our forecasting problem has a property that is common to a variety of decision problems a firm may face and that is fundamental to our results on returns to scale: It involves *aggregating* information about many of the firm's activities (or markets or parts) whose number varies with the scale of the firm.

## 3 Computation and sampling problems

### 3.1 Decision procedures and performance

We axiomatically define decision procedures for two models, reflecting two types of constraints. In the main model, which we call the *computation problem*, information is costless but computation is constrained. As a benchmark that helps us distinguish between the effects of computation constraints and the effects of statistical assumptions, we also study a conventional model, called the *sampling problem*, in which information may be costly but computation is unconstrained.

In either model, $\Pi^n$ denotes the set of decision procedures for the decision problem of size $n$, and $\Pi \equiv \bigcup_{n=1}^{\infty} \Pi^n$ is the set of all potential decision procedures. Each decision procedure $\pi \in \Pi$ has an administrative cost $C(\pi)$ and generates a policy $\{A_t^{\pi}\}$. The *total cost* of a decision procedure $\pi \in \Pi^n$ when used in a decision problem of size $n$ is the sum $\mathrm{TC}^n(\pi) \equiv C(\pi) + \Gamma(\{L_t\})$ of its administrative cost and long-run loss (where $L_t \equiv E[\psi^n(X_t^n - A_t^{\pi})]$ for $t \in \mathbb{N}$).

Sections 3.2 and 3.3 impose restrictions, which are different for the two models, on the mapping from decision procedures to costs and policies. The

[8] Even though the forecasting begins in period 1, we assume a double infinity of time periods for the processes in order to simplify the statement of certain statistical assumptions.

following notation is used to state these restrictions. For $n \in \mathbb{N}$ and $t \in \mathbb{N}$, let $\Phi_t^n$ be the set of indices of all realizations of the stochastic processes in a firm of size $n$ up to but not including period $t$, that is,

$$\Phi_t^n \equiv \{\langle i, s\rangle \mid i \in \{1, 2, \ldots, n\},\ s \in \{\ldots, t-2, t-1\}\}\ .$$

For $\pi \in \Pi^n$, $\Phi_t^\pi \subset \Phi_t^n$ denotes the indices of the observations of the stochastic processes that are used by decision procedure $\pi$ for the period-$t$ forecast; the random-vector representation of this information is $H_t^\pi \equiv \{X_{is} | \langle i, s\rangle \in \Phi_t^\pi\}$ and $A_t^\pi$ is a measurable function of $H_t^\pi$.

### 3.2 The computation problem

In the computation problem, $\pi \in \Pi$ represents a *computation procedure* – a specification of the bureaucratic procedures managers follow in order to calculate forecasts from available data. $C(\pi)$ is the long-run cost, including managerial wages, of this information processing. $\Phi_t^\pi$ denotes the indices of the data used to calculate $A_t^\pi$.

We model the computation constraints axiomatically rather than constructively. The assumptions allow for decentralized computation, that is, computation performed jointly by many managers or clerks whose numbers and activities are determined endogenously. This property, which is analogous to parallel or distributed processing by networks of machines, cannot be suppressed when studying returns to scale, because managerial resources must be allowed to vary with the scale of the firm. The assumptions stated ar e satisfied by the computation model in Radner and Van Zandt (1992) and in Van Zandt (1999b), which study this same decision problem, and by most other distributed processing models, including those that have been used in economics, such as Mount and Reiter (1990) and Reiter (1996).[9]

The fundamental constraint we want to capture is that information processing – which includes the reading and preparation of reports and aggregation of non-numerical information – takes time. To motivate this, the numerical data in our decision problem should be viewed as a proxy for the complex data used by human administrators in actual organizations, or the reader should imagine that the data is not available in a simple numerical format and instead is difficult to understand and substantiate and must be communicated through lengthy reports. We emphasize that our use of a numerical decision problem as a proxy for more realistic human decision problems is standard in economics and derives from the need to impose statistical assumptions, rather than from our need to impose computation constraints. Van Zandt (1999b) explains that the information processing constraints we impose are qualitatively similar to the ones we would impose for more realistic problems.

[9] Kenneth Mount and Stanley Reiter have advocated decentralized information processing as a model of human organizations since 1982. See Van Zandt (1998a, 1999a) for surveys of the use of such models in the economic theory of organizations.

This time constraint has two effects. First, it adds an administrative cost (reflected in $C(\pi)$) to the calculation of any policy owing to the time managers spend processing information. Second, it restricts the set of feasible policies; in particular, it limits the amount of recent data that can be incorporated into decisions. This second effect is the more important one for this paper, and is captured by the following "iron law of delay".

**Assumption 1** *For each lag $d \in \mathbb{N}$, there is a uniform bound on the amount of data whose lag is $d$ or less on which any forecast can depend. Formally, there is a function $B : \mathbb{N} \to \mathbb{N}$ such that $\#\{\langle i, s\rangle \in \Phi_t^\pi | s \geq t - d\} \leq B(d)$ for $d \in \mathbb{N}$, $\pi \in \Pi$, and $t \in \mathbb{N}$.*

This bound comes from the delay in aggregating information. For example, suppose that policies are computed by having agents perform elementary operations that can have at most $k$ inputs (which can be any previous results, raw data, or constants) and that produce an arbitrary number of outputs. Suppose each operation takes at least $\delta$ units of time. Either $A_t$ is a constant or the value of a raw datum, or it is the output of an elementary operation that was begun by time $t - \delta$. Thus, $A_t$ can depend on at most 1 datum that is first available after $t - \delta$. If $A_t$ is the result of an elementary operation begun by time $t - \delta$, then each of the $\leq k$ inputs is either a constant or a raw datum, or is itself the output of an operation begun by time $t - 2\delta$. Hence, $A_t$ can depend on at most $k$ data first available after $t - 2\delta$. Repeating this argument inductively, $A_t$ can depend on at most $k^2$ data first available after $t - 3\delta$, and on at most $k^{\nu-1}$ data first available after $t - \nu\delta$ for $\nu \in \mathbb{N}$. This implies that, for $d \in \mathbb{N}$, $A_t$ can depend on at most $k^{\lfloor d/\delta \rfloor}$ observations from period $t - d$ or later. The bound would also hold if the delay comes from reading and interpreting raw data and messages; see Van Zandt (1999b) for a discussion.

Note how the bound holds even though we allow for decentralization of information processing. This is because information must be *aggregated* to make a forecast. This is related to the centralization of decision making, as is illustrated by the following example. Suppose there are $n$ firms, each of which must choose the best of two potential projects. Suppose there are potential administrators each of whom can perform a pairwise ranking of any two projects in one period. Then, however large $n$ is, all of the projects can be selected in one period. This is because the $n$ decision problems are independent and each can be performed in the same period by a different administrator. Now suppose the $n$ firms merge, with the intention of selecting the best $n$ of the $2n$ projects. This selection problem cannot be decomposed into operations that can all be performed concurrently. Instead, it takes more than $\log_2 n$ periods to select the $n$ projects by pairwise rankings.

The next assumption states that policies can be scaled without changing the processing costs.

**Assumption 2** *For $n \in \mathbb{N}$, $\pi \in \Pi^n$, and $\alpha > 0$, there is $\pi' \in \Pi^n$ such that $C(\pi') = C(\pi)$ and $A_t^{\pi'} = \alpha A_t^\pi$ for $t \in \mathbb{N}$.*

Note that the scaling factor $\alpha$ is constant over time and independent of the realizations of the stochastic processes. The scaling corresponds to a change in the units used to measure demand and/or to control production.

The third assumption about computation procedures is that there are procedures that process no information and have no administrative cost.

**Assumption 3** *For $n \in \mathbb{N}$ and $a \in \mathbb{R}$, there is $\pi \in \Pi^n$ such that $C(\pi) = 0$ and $A_t^\pi = a$ for $t \in \mathbb{N}$.*

Finally, we assume that a large organization can mimic the decision procedure of a smaller organization.

**Assumption 4** *For $n \in \mathbb{N}$, $\Pi^n \subset \Pi^{n+1}$.*

### 3.3 The sampling problem

In the sampling problem, $\pi \in \Pi$ represents a *sampling procedure* – a specification of the information to be gathered each period and to be stored from one period to the next. $C(\pi)$ is the long-run cost of obtaining and storing that information. $\Phi_t^\pi$ denotes the indices of the data that have been sampled up through period $t-1$ and are available when making the period-$t$ forecast.

We neither assume nor preclude perfect recall ($\Phi_t^\pi \subset \Phi_{t+1}^\pi$). The rationality (no processing constraints) assumption is that when a sampling procedure $\pi$ is used for a decision problem of size $n$, the period-$t$ forecast $A_t^\pi$ minimizes $E[\psi^n(X_t^n - A_t)]$ subject to the constraint that $A_t$ be $H_t^\pi$-measurable. This is called statistical optimality, which may be formalized as follows.

**Assumption 5** *For $n \in \mathbb{N}$, $\pi \in \Pi^n$, and $t \in \mathbb{N}$, $A_t^\pi \in \arg\min_{a \in \mathbb{R}} E[\psi^n(X_t^n - a)|H_t^\pi]$ a.e.*

Our only assumption on sampling (and data storage) costs is that they are additive and symmetric across stochastic processes. 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. To state this formally, we identify for each $\pi \in \Pi$ and $i \in \mathbb{N}$ the dates of the information about process $i$ provided by $\pi$ by letting $\phi_{it}^\pi \equiv \{s \in \mathbb{Z} \mid \langle i, s\rangle \in \Phi_t^\pi\}$ for $t \in \mathbb{N}$ and $\phi_i^\pi \equiv \{\phi_{it}^\pi\}_{t=1}^\infty$. We refer to $\phi_i^\pi$ as a *single-process information structure*.[10] (If a process is not sampled at all, then its information structure is $\phi_{\text{null}} \equiv \{\emptyset\}_{t=1}^\infty$.) Our assumption is then that (a) there is a set $\tilde{\phi}$ of single-process information structures with associated costs, (b) a sampling procedure specifies a single-process information structure in $\tilde{\phi}$ for each process, and (c) sampling costs are summed over the processes.

**Assumption 6** *There is a set $\tilde{\phi}$ of single-process information structures such that, for $n \in \mathbb{N}$, $\{\phi_1, \dots, \phi_n\} \subset \tilde{\phi}$ if and only if there is $\pi \in \Pi^n$ such that $\phi_i^\pi = \phi_i$ for $i \in \{1, \dots, n\}$. Furthermore, there is $S : \tilde{\phi} \to \mathbb{R}_+$ such that, for $n \in \mathbb{N}$ and $\pi \in \Pi^n$, $C(\pi) = \sum_{i=1}^n S(\phi_i^\pi)$. Also, $\phi_{\text{null}} \in \tilde{\phi}$ and $S(\phi_{\text{null}}) = 0$.*

[10] The structure $\phi_{it}^\pi$ is an element of $2^{\{\dots, t-2, t-1\}}$, and so formally we define a single-process information structure to be any element of $\prod_{t=1}^\infty 2^{\{\dots, t-2, t-1\}}$.

### *3.4 Comparison of the computation and sampling problems*

The policies in the computation problem do not minimize the expected loss conditional on all available information, since they do not even depend on all available information. A weaker notion of statistical optimality of a computation procedure $\pi \in \Pi^n$ is that $A_t^\pi$ minimizes $E[\psi^n(X_t^n - a)|H_t^\pi]$ almost surely. As discussed in Van Zandt (1999b), a constrained-optimal computation procedure (one that minimizes total costs on $\Pi$) need not be statistically optimal in the computation problem because it may be more costly (or impossible) to compute the statistically-optimal decision rule that uses the same information as $\pi$. This is one potential difference between the sampling problem and the computation problem.

However, this difference is not relevant to our results. In fact, we never preclude statistical optimality in the computation problem. Instead, the important difference is how much data of a given lag can be used in a forecast. Suppose that, in the sampling problem, the forecast in period $t$ of a firm of size 1 is based on $X_{1,t-2}$. Then, for a firm of size $n$, it is possible to sample $X_{i,t-2}$ for all $i \in \{1, \ldots, n\}$ with the same average sampling cost faced by the firm of size 1, so that the forecast uses the data surrounded by the solid line in Figure 1.

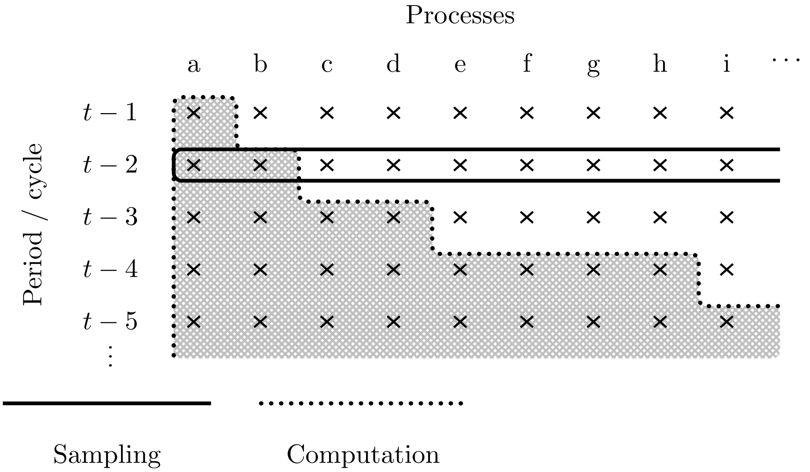

**Figure 1.** Bounds on aggregation speed. The solid outline shows data that could be incorporated into the period-$t$ prediction in the sampling problem. The dotted outline shows a limit on the data of various lags that could be incorporated into the period-$t$ prediction in the computation problem

This is not possible in the computation problem because of aggregation delay (Assumption 1). For example, Figure 1 shows the bound on the data of any given lag for the case where $B(d) = 2^{d-1}$. Thus, in the computation problem, aggregation delay creates a negative informational externality among the processes – data of a given lag about one process crowds out data of that lag about other processes.

## 4 Returns to scale: Assumptions and definitions

### 4.1 Statistical assumptions

For $t \in \mathbb{Z}$, the vector $\{X_{1t}, X_{2t}, \dots\}$ is denoted by $\mathbf{X}_t$; then $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ or simply $\{\mathbf{X}_t\}$ denotes the vector process. For $t \in \mathbb{Z}$, $H_t$ denotes the history of $\mathbf{X}_t$ up through period $t$.

We assume that the processes have finite variance and are stationary and exchangeable.

**Assumption 7** *For all $t \in \mathbb{N}$ and $i \in \mathbb{N}$, $0 < \mathrm{Var}(X_{it}) < \infty$.*

**Assumption 8** *The vector process $\{\mathbf{X}_t\}$ is stationary.*

**Assumption 9** *The family $\{\{X_{1t}\}, \{X_{2t}\}, \dots\}$ of processes is exchangeable (the joint distribution of the processes is symmetric).*

Exchangeability ensures that the processes are statistically indistinguishable. In particular, if we partition a set of stochastic processes into groups whose sums must be forecast independently, then the statistical properties of each group depend only on the number of processes and not on their identities. A canonical decision problem is thus the estimation of the sum of the first $n$ processes, as in Section 2. This assumption precludes, for example, a locational model in which the demand indices of nearby markets are more correlated than those of distant markets.

*Remark 1* A sufficient condition for the processes $\{\{X_{1t}\}, \{X_{2t}\}, \dots\}$ to be exchangeable is that they can be written $X_{it} = Y_t + Z_{it}$, where the processes $\{\{Y_t\}, \{Z_{1t}\}, \{Z_{2t}\}, \dots\}$ are independent and the processes $\{\{Z_{1t}\}, \{Z_{2t}\}, \dots\}$ are identically distributed. (This condition is also necessary if the processes are Guassian.) We then call $\{Y_t\}$ the *common* component of the processes and $\{Z_{it}\}$ the *idiosyncratic* component of process $i$.

### 4.2 Loss functions

We consider two classes of loss functions.

**Quadratic loss** The first is the quadratic loss, whose form is the same for all $n$:

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

**Scalable loss** The other case is where the average loss is a convex function of the average error:

$$\tfrac{1}{n}\psi^n(X_t^n - A_t) = \Psi\left(\tfrac{1}{n}(X_t^n - A_t)\right) \ ,$$

where $\Psi$ is a convex function not depending on $n$ such that $\Psi(0) = 0$, $\Psi(\epsilon) > 0$ if $\epsilon \neq 0$, and $E[\Psi(X_{it} - E[X_{it}])] < \infty$. We refer to this as a *scalable* loss function.

A leading example of the scalable loss is when $\psi$ is piecewise linear and does not depend on $n$:

$$\psi^n(X_t^n - A_t) = \begin{cases} \gamma_0 \mid X_t^n - A_t \mid & \text{if } X_t^n - A_t < 0 \ , \\ \gamma_1 \mid X_t^n - A_t \mid & \text{if } X_t^n - A_t \geq 0 \ . \end{cases}$$

For example, this is the loss when a firm has to make up for excess demand (resp., supply) by buying (resp., selling) output at a price that exceeds (resp., is less than) the firm's unit production cost. The scalable loss also includes the case where a quadratic loss is adjusted for firm size, $\psi^n(X_t^n - A_t) = \frac{1}{n}(X_t^n - A_t)^2$, in which case $\Psi(\epsilon) = \epsilon^2$.

### *4.3 Long-run loss*

As explained in Section 3.1, the function $\Gamma$ aggregates period-by-period expected losses into a measure of long-run loss. We denote the domain of $\Gamma$ by $\mathscr{L}$, which must contain the sequence of expected losses for any policy that is generated by a decision procedure (such a policy is said to be *allowable*). Our next assumption restricts the domain $\mathscr{L}$ and assumes that $\Gamma$ is linear and strictly monotone.

**Assumption 10** *If $A_t$ is an allowable policy for a decision problem of size n and $L_t = E[\psi^n(X_t^n - A_t)]$ for $t \in \mathbb{N}$, then $\{L_t\} \in \mathscr{L}$. Furthermore:*

1. *$\mathscr{L}$ is the positive cone of a linear subspace of $\mathbb{R}^{\mathbb{N}}$ containing the constant sequences;*
2. *$\Gamma$ is a linear functional;*[11]
3. *if $\{L_t\}$ and $\{L'_t\}$ belong to $\mathscr{L}$ and $L_t < L'_t$ for $t \in \mathbb{N}$, then $\Gamma(\{L_t\}) < \Gamma(\{L'_t\})$.*

As a normalization, we also assume that if $\{L_t\}$ is constant then $\Gamma(\{L_t\})$ is equal to the constant value of $\{L_t\}$.

The purpose of the linearity assumption is to make comparisons across problems of different size meaningful (e.g., if the expected loss in each period scales linearly with problem size, then so does the long-run loss). This assumption holds if $\mathscr{L}$ is the set of bounded sequences in $\mathbb{R}^{\mathbb{N}}$ and $\Gamma(\cdot)$ is the discounted present value with respect to a summable sequence of discount factors. It is also consistent with the case where $\Gamma(\{L_t\})$ is the long-run average value of $\{L_t\}$, in which case there is an implicit restriction on the set of decision procedures. Specifically, the set $\mathscr{L}$ must then contain only sequences whose long-run averages are well-defined and $\mathscr{L}$ cannot contain two sequences $\{L_t\}$ and $\{L'_t\}$ such that $L_t < L'_t$ for $t \in \mathbb{N}$ and such that both have the same long-run average (true if $\mathscr{L}$ contains only constant or cyclic sequences). See Van Zandt and Radner (1999) for further discussion and a sketch of how to weaken the monotonicity condition.

[11] That is, $\Gamma$ can be extended to a linear functional on the subspace spanned by $\mathscr{L}$.

### 4.4 Definitions of returns to scale

For both the computation and sampling problems, we assume that there is a cost-minimizing decision procedure for all $n$.

**Assumption 11** *For $n \in \mathbb{N}$, there is a $\pi \in \Pi^n$ such that* $\mathrm{TC}^n(\pi) \leq \mathrm{TC}^n(\pi')$ *for $\pi' \in \Pi^n$.*

Such a decision procedure is said to be *constrained-optimal* or simply optimal. Let $\mathrm{TC}(n)$ be the minimum cost and let $\mathrm{AC}(n) \equiv \mathrm{TC}(n)/n$ be the average cost for a firm of size $n$.

Although we have defined a firm of size $n$ to be the decision problem with the *first $n$* stochastic processes, it would be trivial to generalize the model to decision problems with *any $n$* of the stochastic processes. Given the symmetric distribution of the stochastic processes and the symmetry of the sampling costs with respect to the processes, and adding additional symmetry assumptions to the computation model, the minimum costs of a firm would depend only on its size and not on the identity of its processes. Then, for any partition $\{n_1, \ldots, n_k\}$ of $n$ (a list of strictly positive integers that sum to $n$), the total cost in a market of size $n$ that is served by $k$ firms with sizes $n_1, \ldots, n_k$ is equal to $\mathrm{TC}(n_1) + \cdots + \mathrm{TC}(n_k)$. A partition of $n$ is *optimal* if it has the lowest total costs. Our goal is to characterize the size of firms in optimal partitions. (The interpretation of this exercise in terms of returns to scale of centralized decision making and of firms was given in Section 1.3.)

An integer $\bar{n}$ is a bound on firm size if the size of any firm in any optimal partition (for any $n$) is no greater than $\bar{n}$.

**Definition 1** *For $\bar{n} \in \mathbb{N}$, $\bar{n}$ is a* bound on firm size *if, for all $n \in \mathbb{N}$ and optimal partitions $\{n_1, \ldots, n_k\}$ of $n$,* $\max\{n_1, \ldots, n_k\} \leq \bar{n}$. *Firm size is bounded and returns to scale are eventually decreasing if there is a bound on firm size.*

We say that firm size is unbounded not simply if there is no bound on firm size, but also if, heuristically, all large markets contain large firms.

**Definition 2** *Firm size is unbounded and returns to scale are eventually increasing if, for all $\bar{n} \in \mathbb{N}$, there is an $n' \in \mathbb{N}$ such that, for all $n \geq n'$ and all optimal partitions $\{n_1, \ldots, n_k\}$ of $n$,* $\max\{n_1, \ldots, n_k\} > \bar{n}$.

We define monotonic returns to scale in the usual way.

**Definition 3** *Returns to scale are* monotonically (weakly/strictly) increasing, decreasing, *or* constant *if* $\mathrm{AC}(n)$ *is (weakly/strictly) decreasing, increasing, or constant.*

If returns to scale are monotonically strictly increasing, then the only optimal partition of $n$ is $\{n\}$ and firm size is unbounded.

## 5 Returns to scale: Results

For each of the two classes of loss functions (quadratic and scalable), we characterize the returns to scale of the computation and sampling problems for two

**Table 1.** Table of results

| | Statistical assumptions | Returns to scale | | |
|---|---|---|---|---|
| | | Sampling problem | Computation problem | |
| Quadratic loss | mutually dependent | bounded firm size<br>($\lim_{n\to\infty}$ AC (n) = $\infty$) | bounded firm size<br>($\lim_{n\to\infty}$ AC (n) = $\infty$) | Thm 1 |
| | mutually independent | constant<br>(constant per-unit gain) | bounded firm size<br>(per-unit gain ⟶ 0) | Thm 2 |
| Scalable loss | common process plus noise | monotonically increasing | monotonically increasing | Thm 3 |
| | general | unbounded firm size<br>(replication works) | example with bounded firm size | Thm 4 |

sets of additional statistical assumptions. This yields four theorems, which are summarized in Table 1. The reader may wish to refer back to Section 1.4, which contains a more extensive summary and interpretation of the results. *All proofs are given in the Appendix.*

### *5.1 Quadratic loss and mutually correlated processes*

We first dispense of a case that has no interesting contrast between the computation and sampling problems – the quadratic loss with an assumption that rules out statistical independence of the stochastic processes. The quadratic loss function is not favorable to increasing returns because if the average error is constant then the average loss increases linearly with $n$. Theorem 1 shows that this leads to decreasing returns to scale in both the computation and sampling problems if (heuristically) there is a common component that cannot be perfectly forecasted from past data.

**Theorem 1** *Assume the loss is quadratic and that* $E[\mathrm{Cov}(X_{it}, X_{jt}|H_{t-1})] > 0$ *for* $i, j \in \mathbb{N}$ *such that* $i \neq j$.[12] *In both the sampling and the computation problems,* $\lim_{n\to\infty} \mathrm{AC}(n) = \infty$ *and firm size is bounded.*

[12] Recall that $E[\mathrm{Cov}(X_{it}, X_{jt}|H_{t-1})] = E[(X_{it} - E[X_{it}|H_{t-1}])(X_{jt} - E[X_{jt}|H_{t-1}])]$. If the decomposition in Remark 1 holds and if $\{\mathbf{X}_t\}$ or simply $\{Y_t\}$ is regular (see Remark 2), then $E[\mathrm{Cov}(X_{it}, X_{jt}|H_{t-1})] > 0$ if and only if the processes are mutually dependent. We conjecture but have not verified that this holds without the decomposition.

### *5.2 Quadratic loss and mutually uncorrelated processes*

When the loss function is quadratic but the processes are mutually *independent*, a diversification effect counterbalances the curvature of the loss function. This leads to constant returns to scale in the sampling problem. As shown in the proof of Theorem 2, the selection of a sampling procedure is separable over the processes and any firm should replicate an optimal procedure of a firm of size 1.

In the computation problem, such replication is impossible because the firm would compute $n$ forecasts, which must then be aggregated, thereby incurring additional delay. In fact, the aggregation delay implies that the data about "most" processes is "old" in large firms. In Theorem 2, we assume that information becomes useless as it gets older. (Specifically, we assume $\{\mathbf{X}_t\}$ is regular; see Remark 2 immediately after Theorem 2.) Hence, as firm size grow s, the average cost converges to the *no-information average cost*. This is defined to be the average cost of the decision procedure that (a) has no administrative cost, (b) makes the same forecast each period, and (c) has an expected loss each period of $\min_{a\in\mathbb{R}} E[\psi^n(X_t^n - a)]$. Such a procedure corresponds to no computation or no sampling. Its existence is implied by Assumption 3 for the computation problem and Assumption 6 for the sampling problem; hence its average cost is an upper bound on AC($n$).

**Theorem 2** *Assume that the loss is quadratic and that the processes* $\{\{X_{1t}\}, \{X_{2t}\}, \ldots\}$ *are mutually independent. Then returns to scale are constant in the sampling problem. However, if also* $\{\mathbf{X}_t\}$ *is regular, then in the computation problem the average cost converges (as* $n \to \infty$*) to the no-information average cost; furthermore, if for some firm size* $n \in \mathbb{N}$ *there is a computation procedure for which the average cost is lower than the no-information average cost, then firm size is bounded.*

*Remark 2* Regularity is defined as follows: Given an underlying probability space on which the process $\{\mathbf{X}_t\}$ is defined, let $\{\mathscr{F}_t\}$ be the filtration generated by $\{\mathbf{X}_t\}$. Then $\{\mathbf{X}_t\}$ is *regular* if and only if the tail $\sigma$-field $\bigcap_{t=0}^{-\infty} \mathscr{F}_t$ is trivial, meaning that it contains only events with probability 1 and 0. It follows (from Martingale convergence theorems; see Loève (1978, p. 75)) that $E[X_{it}|H_{t-d}] \to E[X_{it}]$ as $d \to \infty$, where the convergence is a.e. and in the $L^p$-norm for $1 \leq p < \infty$.

Radner and Van Zandt (1992) characterize returns to scale for a specific computation model under assumptions (quadratic loss and i.i.d. AR(1) processes) that are consistent with those of Theorem 2.

### *5.3 Scalable loss and noisy common processes*

With a quadratic loss function that does not change with firm size, larger firms have the same tolerance as smaller firms for errors of fixed magnitude. However, this may not hold if, for example, the loss when output exceeds demand comes from holding inventories and the inventory capacity is proportionate to firm size.

In this case the scalable loss function may be more realistic. The scalable loss also includes the piecewise linear loss.

Our first theorem regarding the scalable loss, Theorem 3, assumes that the processes are noisy versions of a common process. The task is to estimate the common process, and this forecast is a "public good"; as the size of the firm grows, more data are available and the cost of the forecast can be spread among more processes (Arrow effect). In particular, when the loss function is also scalable, a larger firm can achieve a strictly lower average 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 *average* 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 one case in which the increasing returns to scale due to information sharing, studied by Arrow (1974) and Wilson (1975), arise even with computation constraints. In the extreme case in which there are no computation or sampling costs to be shared, a diversification effect (assumption (ii) or (iii) in the theorem) can make returns strictly rather than simply weakly increasing.

**Theorem 3** *Assume that the loss is scalable, that the decomposition $X_{it} = Y_t + Z_{it}$ in Remark 1 holds, and that the random variables $Z_{it}$ are i.i.d. across $i \in \mathbb{N}$ and $t \in \mathbb{Z}$. Then returns to scale are monotonically weakly increasing in both the sampling and the computation problems. They are monotonically strictly increasing if also either: (i) for all $n \in \mathbb{N}$, there is an optimal decision procedure $\pi^n \in \Pi^n$ such that $C(\pi^n) > 0$; (ii) $\Psi$ is strictly convex and $\mathrm{Var}(Z_{it}) > 0$; or (iii) the Lebesgue measure is absolutely continuous with respect to the distribution of $Z_{it}$.*

Interestingly, delay *does not* lead to decreasing returns in the computation problem because the amount of data used in the computation is endogenous and, in particular, does not have to increase with the size of the firm. In Section 6.2, we contrast this with the eventually decreasing returns that may obtain in a benchmark batch processing model.

Radner and Van Zandt (1992) characterize the returns to scale for a specific computation model under assumptions (piecewise linear loss and processes that are noisy versions of a common AR(1) process) that are consistent with those of Theorem 3.

### *5.4 Scalable loss and general processes*

The idea behind Theorem 3 is that a larger firm can achieve a lower average loss than a small firm by imitating the decision procedure of a *single* small firm. This is *not* an analog of the principle that leads to nondecreasing 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. However, in the sampling problem with scalable loss, the analog of this principle – a large firm imitates the sampling procedures of several small firms – does lead to eventually decreasing returns to scale under general statistical assumptions. This is the first part of Theorem 4.

**Theorem 4A** *Assume that the loss function is scalable and Assumption 12 (stated in the Appendix) holds. In the sampling problem, firm size is unbounded and* $\mathrm{AC}(kn) < \mathrm{AC}(n)$ *for* $n, k \in \mathbb{N}$ *such that* $k > 1$.

There is no such analog for the computation problem. If a large firm imitates the policies of several small firms, it ends up with several forecasts each period. If it attempts to aggregate these forecasts, there is additional delay and so the policy uses information that is older than the information used by the small firms. This does not imply that returns to scale are never increasing in the computation problem, as was shown in Theorem 3. However, the second part of Theorem 4 presents a robust example in which firm size is bounded in the computation problem. This result shows how aggregation delay in a centralized decision problem may subvert the Arrow effect.

**Theorem 4B** *Assume that the loss function is scalable and Assumption 13 (stated in the Appendix) holds. In the computation problem,* $\mathrm{AC}(1) < \mathrm{AC}(n)$ *for* $n \geq 2$ *so* $1$ *is a bound on firm size.*

Assumptions 12 and 13 in Theorems 4A and 4B, respectively, are stated in the Appendix because they are rather technical. Assumption 12 is a weak statistical assumption that plays the following role. We obtain the inequality $\mathrm{AC}(kn) \leq \mathrm{AC}(n)$ in the sampling problem by showing that if the firm of size $kn$ replicates the sampling procedure of a firm of size $n$, then the average sampling cost of the large firm and the small firm are the same, and the average expected loss of the large firm is as low as that of the small firm. To obtain the *strict* inequality $\mathrm{AC}(kn) < \mathrm{AC}(n)$, we appeal to the diversification effect, but this requires, for example, that the processes not be perfectly correlated. Assumption 12 rules out this and similar trivial cases.

For the computation problem, Assumption 13 specifies a detailed but robust example. It assumes, for example, that the processes can be decomposed as $X_{it} = Y_t + Z_{it}$, and that each of the components is a first-order autoregressive processes. When the statistical conditions in Assumption 13 are satisfied, so is Assumption 12; hence, the contrast between the sampling and computation problems is real. Assumption 13 also states restrictions on the computation technology, which are satisfied, for example, if the computation model is the one in Van Zandt (1999b) (an adaptation of the one in Radner, 1993) and if one cycle equals one period in that model.

## 6 Related literature

### 6.1 Historical

Collectively, our results (summarized in Section 1.4) show that returns to the scale of firms and decentralized decision making 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.

The first is that delay and change are fundamental for understanding information processing constraints in organizations. 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. In a criticism of the iterative planning procedures of Lange (1936, 19) and Dickinson (1939) – which assume that the underlying economic data are constant – Hayek (1940, pp. 131–132) stated:

> In the real world, where constant change is the rule, ... 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 decision should be made 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 literally 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.

### *6.2 Other information processing models of returns to scale*

Keren and Levhari (1983) and Radmer (1993, Sect. 7) also study when aggregation delay limits the scale of firms and of centralized decision making. These papers are based on decentralized computation models that are consistent with the constraints in this paper, but they study *batch* rather than *real-time* processing. In batch processing, an exogenously given function must be computed and delay is measured by the time between the beginning and completion of computation. The authors study returns to scale by positing an exogenously given cost function – a function that depends on the scale of the problem, the computational costs, and the delay. Reiter (1996) is also a batch processing model that examines limits to firm size and centralization, but under the postulate that there are bounds on the size of the informational inputs of any organizational unit.

Real-time control is a different, and in some ways richer, methodology for studying the effects of delay on decision making. First, because it is based on a temporal decision problem, we can implicitly derive a “cost of delay” from the degradation of the quality of decisions that are based on old information. Furthermore, because decision rules are endogenous, we do not artificially limit centralization by forcing organizations with large-scale decision problems to bog themselves down with computation and only use old data. Compare this with a benchmark model obtained by embedding a batch processing model into our decision problem. Following Keren and Levhari (1983) and Radner (1993, Sect. 7), in which all data are collected for a decision at the same point in time and the amount of data is equal to the scale of the firm, we would consider only computation procedures in which the firm calculates the period-$t$ decision from $\{X_{i,t-d}\}_{i=1}^{n}$ for some delay $d$. The computation constraints require that $n \leq B(d)$ and hence $d \to \infty$ as $n \to \infty$. Consider the assumptions of Theorem 3, with a negligible idiosyncratic component. The problem is then to forecast $Y_t$ from $\{X_{i,t-d}\}_{i=1}^{n}$. Assuming that $\{Y_t\}$ is regular, the average expected loss in the benchmark model is approximately equal in the limit (as $n \to \infty$ and $d \to \infty$) to the average expected loss when there is no information processing. One can thus construct specific examples (see Van Zandt and Radner, 1999) in which firm size is bounded in the benchmark model, whereas Theorem 3 shows that returns to scale are monotonically increasing in our model.

The model by Geanakoplos and Milgrom (1991) is a team-theory model of resource allocation in which an endogenous administrative apparatus hierarchically disaggregates resource allocations. Their model has the advantage of allowing for internal decentralization of decision making, with coordination among the decision-making nodes. Theirs is a static approach that does not explicitly model the hierarchical aggregation of information; rather, there are constraints on information acquisition for individual agents that represent information processing constraints. Hence, their results on returns to scale depend on assumptions about what aggregate information is available exogenously. 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. Van Zandt (1999c, d) studies a temporal version of their decision problem, but with real-time information processing.

The work of Orbay (1996) and Meagher (1996) is also related, but with interesting differences. They consider a problem of forecasting a fixed stochastic process without variations in the scale of the decision problem or operations of the firm. However, the amount of data sampled about the process for calculating each decision is endogenous. Because of computational delay, the trade-off is between basing each forecast on a large amount of old information or on a small amount of recent information. The size of the administrative apparatus is roughly proportional to the amount of data incorporated into each decision, so this exercise considers the optimal size of the administrative apparatus for a firm whose scale of production is fixed. They find, for example, that the administrative apparatus tends to be smaller the more quickly the environment is changing.

### *6.3 Other models of decentralized decision making*

Decentralized decision making has been studied formally in the communication mechanism (planning and message space) literatures and in statistical team theory [see Van Zandt (1999a) for references, which for team theory include Marschak and Radner (1972)]. With the exception of Geanakoplos and Milgrom (1991), these formal literatures have focused attention on communication costs as a motive for decentralized decision making. The idea is simple. If a set of exogenously given agents are endowed with private information and it is costly to pool this information, then decisions may be delegated to various agents.[13] Van Zandt (1999c) also mentions a few recent papers that attempt to explain decentralized decision making with incentives, by relaxing the assumptions of the revelation principle. Again, decisions are decentralized to agents who are endowed with private information.

In contrast, the main constraint in the current paper [and its predecessor, Radner and Van Zandt (1992)] is human delay in processing information. We show that this can limit the scale of centralized decision making, even in the absence of information transmission costs and incentive problems, and without relying on the existence of agents who *a priori* have private information. Our explanation of decentralized decision making is meant to complement the others.[14]

## Appendix: Proofs

We refer to a typical firm of size $n$ as “firm $n$”. We use symbols such as $\ell_t$ to denote the *average* loss in period-$t$ ($\ell_t = L_t/n$). Because $\Gamma$ is linear, if a firm's average loss in each period $t$ is $\ell_t$ then its average long-run loss is $\Gamma(\{\ell_t\})$. For any $\pi \in \Pi^n$, let $\mathrm{AC}^n(\pi) \equiv \mathrm{TC}^n(\pi)/n$.

*Proof of Theorem 1.* A lower bound on $\mathrm{AC}(n)$ is the average expected loss of firm $n$ when the forecast in each period $t \in \mathbb{N}$ minimizes the expected loss given information $H_{t-1}$. This forecast is $E[X_t^n|H_{t-1}] = \sum_{i=1}^n E[X_{it}|H_{t-1}]$ and the expected loss each period is $E[(\sum_{i=1}^n \epsilon_{it})^2]$, where $\epsilon_{it} = X_{it} - E[X_{it}|H_{t-1}]$. The stationarity and exchangeability of the underlying stochastic processes imply that the processes $\{\{\epsilon_{1t}\}, \{\epsilon_{2t}\}, \ldots\}$ are stationary and exchangeable. Hence, $\mathrm{Var}(\epsilon_{it})$ and $\mathrm{Cov}(\epsilon_{it}, \epsilon_{jt})$ are the same for all $t \in \mathbb{N}$ and all $i, j \in \mathbb{N}$ with $i \neq j$; denote these values by $k_1$ and $k_2$, respectively. Thus,

[13] Nevertheless, Marschak (1996) is a message-space model in which centralization (mergers) can decrease certain communication costs.

[14] However, Radner (1992) argues that pure transmission costs are much less important today than human processing costs.

$$\frac{1}{n}E\left[\left(\sum_{i=1}^{n}\epsilon_{it}\right)^{2}\right]=\frac{1}{n}\sum_{i=1}^{n}\left(\operatorname{Var}(\epsilon_{it})+\sum_{j\neq i}\operatorname{Cov}(\epsilon_{it},\epsilon_{jt})\right)=k_1+(n-1)k_2\ .$$

The theorem assumes $k_2 > 0$; hence, this lower bound on AC$(n)$ increases linearly in $n$. □

*Proof of Theorem 2.* For $i \in \mathbb{N}$ and $t \in \mathbb{N}$, let $H_{it}$ be the history of process $i$ up through period $t$ and, for $\pi \in \Pi$, let $H_{it}^{\pi} \equiv \{X_{is} | \langle i,s\rangle \in \Phi_t^{\pi}\}$ be the information about $i$ in $H_t^{\pi}$.

Consider first the sampling problem. Fix a firm size $n \in \mathbb{N}$ and a decision procedure $\pi \in \Pi^n$. Because the loss is quadratic, $A_t^{\pi} = E[X_t^n | H_t^{\pi}] = \sum_{i=1}^{n} E[X_{it} | H_t^{\pi}]$. Because processes $\{X_{it}\}$ and $\{X_{jt}\}$ are mutually independent for $i \neq j$, it follows that $E[X_{it}|H_t^{\pi}] = E[X_{it}|H_{it}^{\pi}]$. Furthermore,

$$E[(X_t^n - A_t^{\pi})^2] = E\left[\left(\sum_{i=1}^{n} X_{it} - E\left[X_{it}|H_{it}^{\pi}\right]\right)^2\right] = \sum_{i=1}^{n} E\left[\left(X_{it} - E\left[X_{it}|H_{it}^{\pi}\right]\right)^2\right]\ . \tag{1}$$

For $i \in \{1,\dots,n\}$, firm 1 could use a sampling procedure $\pi' \in \Pi^1$ such that $\phi_1^{\pi'} = \phi_i^{\pi}$ (Assumption 6), and would then have the sequence

$$\left\{E\left[(X_{it} - E\left[X_{it}|H_{it}^{\pi}\right])^2\right]\right\}_{t=1}^{\infty}$$

of expected losses. Hence, this sequence belongs to $\mathscr{S}$. Given also the linearity of $\Gamma$ (Assumption 10) and the additive separability of sampling costs (Assumption 6), the total cost is

$$\mathrm{TC}^n(\pi) = \sum_{i=1}^{n}\left(\Gamma\left(\left\{E\left[\left(X_{it} - E[X_{it} \mid H_{it}^{\pi}]\right)^2\right]\right\}_{t=1}^{\infty}\right) + S(\phi_i^{\pi})\right)\ .$$

The sampling problem is thus additively separable over the stochastic processes. That is, the problem is to find a single-process information structure $\phi^* \in \tilde{\phi}$ that minimizes

$$\Gamma\left(\left\{E\left[\left(X_{it} - E\left[X_{it}|\{X_{is}|s \in \phi_t\}\right]\right)^2\right]\right\}_{t=1}^{\infty}\right) + S(\phi)\ , \tag{2}$$

and then to use a sampling procedure $\pi \in \Pi^n$ such that $\phi_i^{\pi} = \phi^*$ for $i = 1,\dots,n$. The average cost for any firm is the minimized value of equation (2) and returns to scale are constant.

For the computation problem, we show that the average gain from information processing converges to 0 as $n \to \infty$. For $d \in \mathbb{N}$, let $\lambda_d = E[\left(X_{it} - E[X_{it}|H_{i,t-d}]\right)^2]$, which does not depend on $i$ or $t$ because the stochastic processes are exchangeable and stationary. Let $n \in \mathbb{N}$ and $\pi \in \Pi^n$. The right-hand side of equation (1) is a lower bound on the period-$t$ expected loss for the policy $\{A_t^{\pi}\}$. (This lower bound may not actually be attained, because the decision procedure is not necessarily statistically optimal.) Furthermore, a lower bound on $E\left[(X_{it} - E\left[X_{it}|H_t^{\pi}\right])^2\right]$ is given by $\lambda_{d_{it}^{\pi}}$, where $d_{it}^{\pi} = t - \max\{s|\langle i,s\rangle \in \Phi_t^{\pi}\}$

is the minimum lag of the data in $H_t^\pi$ (or $d_{it}^\pi = \infty$ and $E\left[X_{it} | H_{i,t-d_{it}}\right] = E\left[X_{it}\right]$ if $H_{it}^\pi$ is null). Hence, the period-$t$ expected loss for $\{A_t^\pi\}$ is at least $\sum_{i=1}^n \lambda_{d_{it}^\pi}$.

Let $B : \mathbb{N} \to \mathbb{N}$ be the bound in Assumption 1, and let $\{d_i\}_{i=1}^\infty$ be the sequence such that $d_i = 1$ for the first $B(1)$ terms, $d_i = 2$ for the next $B(2)$ terms, and so on. This sequence is such that, for $n \in \mathbb{N}$, $\pi \in \Pi^n$, $t \in \mathbb{N}$, and $d \in \mathbb{N}$,

$$\#\left\{i \in \{1, \ldots, n\} \,|d_i \le d\right\} \ge \#\left\{i \in \{1, \ldots, n\} \,|d_{it} \le d\right\} \ .$$

Hence, because $\lambda_d$ is decreasing in $d$, $\sum_{i=1}^n \lambda_{d_{it}^\pi} \ge \sum_{i=1}^n \lambda_{d_i}$. Therefore, $\mathrm{AC}(n) \ge \frac{1}{n}\sum_{i=1}^n \lambda_{d_i}$. Since each stochastic process $\{X_{it}\}$ is regular, $\liminf_{d\to\infty} \lambda_d = \mathrm{Var}(X_{it})$ (Remark 2). Since also $\lim_{i\to\infty} d_i = \infty$, we have $\lim_{i\to\infty} \lambda_{d_i} = \mathrm{Var}(X_{it})$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \lambda_{d_i} = \mathrm{Var}(X_{it})$. Consequently, $\liminf_{n\to\infty} \mathrm{AC}(n) \ge \mathrm{Var}(X_{it})$. Because $\mathrm{Var}(X_{it})$ is the no-information average cost and is an upper bound on $\mathrm{AC}(n)$, $\lim_{n\to\infty} \mathrm{AC}(n) = \mathrm{Var}(X_{it})$.

Suppose also that there is $n \in \mathbb{N}$ such that there is a computation procedure whose average costs are lower than the no-information average cost. Then $\mathrm{AC}(n) < \mathrm{Var}(X_{it})$ and there exists an $\bar{n} \in \mathbb{N}$ such that, for $m \ge \bar{n}$,

$$\lfloor m/n \rfloor n\mathrm{AC}(n) + (m \bmod n)\mathrm{Var}(X_{it}) < \mathrm{AC}(m) \ .$$

The left-hand side of this inequality are the total costs when $m$ processes are partitioned into $\lfloor m/n \rfloor$ firms of size $n$ and $m$ mod $n$ firms of size 1. Hence, $\bar{n}$ is a bound on firm size. □

The following lemma is used in the proofs of Theorems 3 and 4. It combines minor extensions of this well-known fact: If $\{x_1, x_2, \ldots\}$ are i.i.d. random variables and if $f : \mathbb{R} \to \mathbb{R}$ is convex, then $E\left[f(\frac{1}{n}\sum_{i=1}^n x_i)\right] \ge E\left[f(\frac{1}{n+1}\sum_{i=1}^{n+1} x_i)\right]$ for $n \in \mathbb{N}$ (because $\frac{1}{n}\sum_{i=1}^n x_i$ is a mean-preserving spread of $\frac{1}{n+1}\sum_{i=1}^{n+1} x_i$). We want to replace "i.i.d." by "exchangeable" and add conditions so that the weak inequality is strict. Here and further below $\mathscr{B}$ denotes the Borel $\sigma$-field of $\mathbb{R}$.

**Lemma 1** *Let $\{\epsilon_1, \epsilon_2, \ldots\}$ be an exchangeable sequence of random variables and let $\Psi : \mathbb{R} \to \mathbb{R}$ be convex. For $n \in \mathbb{N}$, let $\epsilon^n \equiv \frac{1}{n}\sum_{i=1}^n \epsilon_i$. Then $E\left[\Psi(\epsilon^n)\right] \ge E\left[\Psi(\epsilon^{n+1})\right]$ for $n \in \mathbb{N}$. Assume also that one of the following two pairs of assumptions hold.*

1. *(a) $\Psi$ is strictly convex. (b) For $i, j \in \mathbb{N}$ such that $i \ne j$,* $\mathrm{Prob}\left[\epsilon_i \ne \epsilon_j\right] > 0$.
2. *(a) $\Psi$ is not affine. (b) Let $n \in \mathbb{N}$ and let $P : \mathbb{R} \times \mathscr{B} \to [0, 1]$ be a regular conditional probability of $\epsilon^n$ given $\epsilon^{n+1}$. Then there is a $B \in \mathscr{B}$ such that* $\mathrm{Prob}\left[\epsilon^{n+1} \in B\right] > 0$ *and such that, for $\epsilon \in B$ and for each open $U \subset \mathbb{R}$,* $P(\epsilon, U) > 0$.

*Then $E[\Psi(\epsilon^n)] > E[\Psi(\epsilon^{n+1})]$ for $n \in \mathbb{N}$.*

*Proof.* See Van Zandt and Radner (1999). □

*Remark 3* Suppose the random variables $\{\epsilon_1, \epsilon_2, \ldots\}$ can be written $\epsilon_i = y + z_i$ for $i \in \mathbb{N}$, where $\{y, z_1, z_2, \ldots\}$ are independent and $\{z_1, z_2, \ldots\}$ are identically distributed. Then $\{\epsilon_1, \epsilon_2, \ldots\}$ are exchangeable. Furthermore, assumption

1(b) in Lemma 1 holds if $\mathrm{Var}(z_i) > 0$ and assumption 2(b) holds if the Lebesgue measure is absolutely continuous with respect to the distribution of $z_i$ (i.e., $\mathrm{Prob}\,[z_i \in U] > 0$ for any open $U \subset \mathbb{R}$).

*Proof of Theorem 3.* We show that, for $n \in \mathbb{N}$ and $\pi^n \in \Pi^n$, there is an "imitation" $\pi^{n+1} \in \Pi^{n+1}$ of $\pi^n$ such that $\mathrm{AC}^n(\pi^n) \geq \mathrm{AC}^{n+1}(\pi^{n+1})$ (with strict inequality under assumption (i), (ii), or (iii)). By letting $\pi^n$ be the optimal decision procedure for firm $n$, we obtain $\mathrm{AC}(n) = \mathrm{AC}^n(\pi^n) \geq \mathrm{AC}^{n+1}(\pi^{n+1}) \geq \mathrm{AC}(n+1)$, with the first inequality being strict under assumption (i), (ii), or (iii).

We begin with a preliminary result that is common to the sampling and computation problem. Namely, we show that the average expected loss in period $t \in \mathbb{N}$ when firms use the same period-$t$ forecast, but scaled by firm size, is a decreasing function of the firm size. Note that "scaled by firm size" means that the forecast is multiplied by a constant that changes with the firm size, not that the data (or relative coefficients of the data) change with firm size. For the purpose of this result, a forecast is simply a random variable $A_t$ such that $A_t$ and $\{Z_{it}\}_{i\in\mathbb{N}}$ are independent (because $A_t$ must be a function of $H_{t-1}$ and because each process $\{Z_{it}\}$ is serially independent). We normalize so that the scaled version for firm $n$ is $nA_t$. Let $\epsilon^n \equiv (X_t^n - nA_t)/n$ be the average error and let $\ell^n \equiv E[\Psi(\epsilon^n)]$ be the average expected loss in period $t$ when firm $n$ uses the decision rule $nA_t$. We want to show that $\left\{\ell^1, \ell^2, \ldots\right\}$ is weakly or strictly decreasing.

For $i \in \mathbb{N}$, let $\epsilon_i \equiv X_{it} - A_t$ so that $\epsilon^n = \frac{1}{n}\sum_{i=1}^n \epsilon_i$. We can then write $\epsilon_i = (Y_t - A_t) + Z_{it}$. Because $Y_t - A_t$ and $\{Z_{it}\}_{i\in\mathbb{N}}$ are independent and $\{Z_{it}\}_{i\in\mathbb{N}}$ are identically distributed, the sequence $\{\epsilon_1, \epsilon_2, \ldots\}$ is exchangeable. According to Lemma 1, since $\Psi$ is convex, $\ell^n \geq \ell^{n+1}$ for $n \in \mathbb{N}$. Furthermore, it follows from Remark 3 that assumption (ii) or (iii) of Theorem 3 implies assumption 1 or 2 (respectively) of Lemma 1, and hence that $\ell^n > \ell^{n+1}$ for $n \in \mathbb{N}$.

To rest of the proof is different for the computation and sampling problems.

**Computation Problem** Assumptions 2 and 4 imply that there is $\pi^{n+1} \in \Pi^{n+1}$ such that $A_t^{\pi^{n+1}} = ((n+1)/n)A_t^{\pi^n}$ for $t \in \mathbb{N}$ and such that $C(\pi^{n+1}) = C(\pi^n) \equiv C^*$. For $t \in \mathbb{N}$, let $\ell_t^n \equiv E[\psi^n(X_t^n - A_t^{\pi^n})]/n$ and $\ell_t^{n+1} \equiv E[\psi^{n+1}(X_t^{n+1} - A_t^{\pi^{n+1}})]/(n+1)$, so that

$$\begin{aligned}\mathrm{AC}^n(\pi^n) &= \Gamma\left(\{\ell_t^n\}\right) + C^*/n \\ \mathrm{AC}^{n+1}(\pi^{n+1}) &= \Gamma\left(\{\ell_t^{n+1}\}\right) + C^*/(n+1)\ .\end{aligned}$$

We showed above that $\ell_t^n \geq \ell_t^{n+1}$ for $t \in \mathbb{N}$, and hence that $\Gamma\left(\{\ell_t^n\}\right) \geq \Gamma\left(\{\ell_t^{n+1}\}\right)$ and $\mathrm{AC}^n(\pi^n) \geq \mathrm{AC}^{n+1}(\pi^{n+1})$. We also showed that if either assumption (ii) or (iii) held then $\ell_t^n > \ell_t^{n+1}$ for $t \in \mathbb{N}$, and hence, by Assumption 10 (part 3), $\Gamma\left(\{\ell_t^n\}\right) > \Gamma\left(\{\ell_t^{n+1}\}\right)$. If instead assumption (i) holds, then $C^* > 0$ and hence $C^*/n > C^*/(n+1)$. In either case, $\mathrm{AC}^n(\pi^n) > \mathrm{AC}^{n+1}(\pi^{n+1})$.

**Sampling Problem** We let $\pi^{n+1} \in \Pi^{n+1}$ be a sampling procedure such that $\phi_i^{\pi^{n+1}} = \phi_i^{\pi}$ for $i \in \{1, \ldots, n\}$ and $\phi_{n+1}^{\pi^{n+1}} = \phi_{\mathrm{null}}$. According to Assumption

6, such a procedure $\pi^{n+1}$ exists and $C(\pi^{n+1}) = C(\pi^n) \equiv C^*$. For $t \in \mathbb{N}$, let $A'_t \equiv ((n+1)/n)A^\pi_t$. We showed above that $E[\psi^n(X^n_t - A^{\pi^n}_t)]/n$ is (weakly or strictly) greater than $E[\psi^{n+1}(X^{n+1}_t - A'_t)]/(n+1)$; the latter is an upper bound on $E[\psi^{n+1}(X^{n+1}_t - A^{\pi^{n+1}}_t)]/(n+1)$, since $A'_t$ is a function of $H^{\pi^{n+1}}_t$. The rest of the proof is like the one for the computation problem. □

The following assumption ensures that, in Theorem 4A, the diversification effect is present.

**Assumption 12** *In the sampling problem, one of the following two conditions holds.*

1. *(a) $\Psi$ is strictly convex. (b) For $i, j \in \mathbb{N}$ such that $i \neq j$ and for $t \in \mathbb{N}$, there are no functions $f_i$ and $f_j$ of $H_{t-1}$ such that $X_{it} - f_i(H_{t-1}) = X_{jt} - f_j(H_{t-1})$ a.e.*
2. *For $i, t \in \mathbb{N}$, if $P$ is a regular conditional probability of $X_{it}$ given $H_t \setminus \{X_{it}\}$, then with strictly positive probability $H_t \setminus \{X_{it}\}$ is such that the conditional probability $P(H_t \setminus \{X_{it}\}, \cdot) : \mathscr{B} \to \mathbb{R}$ does not have a support that is bounded above or below.*

*Proof of Theorem 4A.*
**Overview of main step:** The main idea of this proof is that firm $kn$ can achieve lower average costs than firm $n$ by *replicating* the sampling procedure and policy of firm $n$. Specifically, let $n \in \mathbb{N}$ and $\pi \in \Pi^n$. For $t \in \mathbb{N}$, let $\ell_t$ be the average period-$t$ expected loss of firm $n$ given $\pi$. For $k > 1$, we define a sampling procedure $\pi^k$ for firm $kn$, which replicates $\pi$, such that $C(\pi^k) = kC(\pi)$. For $t \in \mathbb{N}$, let $\ell^k_t$ be the average period-$t$ expected loss for firm $kn$ given $\pi^k$. We define an upper bound $\overline{\ell}^k_t$ on $\ell^k_t$ such that $\ell_t > \overline{\ell}^k_t$.

**Why this proves the theorem:** It follows that $\ell_t > \ell^k_t$ for $t \in \mathbb{N}$ and hence $\Gamma(\{\ell_t\}) > \Gamma(\{\ell^k_t\})$. That is, firm $n$'s average long-run loss given $\pi$ is greater than firm $kn$'s given $\pi^k$. Both firms' average sampling costs are $(1/n)C(\pi)$. Hence, $\mathrm{AC}^n(\pi) > \mathrm{AC}^{kn}(\pi^k)$. By letting $\pi$ be an optimal sampling procedure for firm $n$, so that $\mathrm{AC}^n(\pi) = \mathrm{AC}(n)$, we have shown that $\mathrm{AC}(n) > \mathrm{AC}^{kn}(\pi^k) \geq \mathrm{AC}(kn)$.

We can then conclude that firm size is unbounded. Let $\overline{n} \in \mathbb{N}$, let $n' \equiv 1 + \sum_{n=1}^{\overline{n}} n$, and let $n \geq n'$. Any partition of $n$ either has a firm whose size is greater than $\overline{n}$ or has two firms of the same size. In the latter case, these two firms can be combined to reduce average costs and so the partition is not optimal. Hence, the maximum firm size of any optimal partition of $n$ is greater than $\overline{n}$.

**Construction of replication strategies:** Let $k \in \mathbb{N}$. Heuristically, firm $kn$ divides the stochastic processes into $k$ divisions, labeled $j \in \{1, \ldots, k\}$, such that division $j$ contains the processes $\{j(n-1)+1, \ldots, jn\}$ and mimics the sampling procedure $\pi$. This means that, for $j \in \{1, \ldots, k\}$ and $i \in \{1, \ldots, n\}$, $\phi^{\pi^k}_{j(n-1)+i} = \phi^\pi_i$. According to Assumption 6, there is such a sampling procedure $\pi^k$ in $\Pi^{kn}$ and $C(\pi^k) = kC(\pi)$.

Let $j \in \mathbb{N}$. For any $k \geq j$ (i.e., $k$ such that firm $kn$ has a division $j$), the period-$t$ information for division $j$, as a random object, is

$$\hat{H}_t^j \equiv \{X_{j(n-1)+i,s} | i \in \{1,\ldots,n\} \text{ and } s \in \phi_i^\pi\} \ .$$

Then $\hat{H}_t^1 = H_t^\pi$ and, for $j \geq 2$, $\hat{H}_t^j$ is like $H_t^\pi$ except that the indices for the stochastic processes are increased by $n(j-1)$.

**Construction of the upper bound:** For $t \in \mathbb{N}$, let $f_t$ be the function such that $A_t^\pi = f_t(H_t^\pi)$ and, for $k \in \mathbb{N}$, let $\overline{A}_t^k \equiv \sum_{j=1}^k f_t(\hat{H}_t^j)$. The interpretation of the decision rule $\overline{A}_t^k$ is that firm $kn$ replicates the decision rule of firm $n$, so that each division calculates a forecast of the sum of its own processes in the same way as firm $n$ does, and then these $k$ independent forecasts are summed. This decision rule is not necessarily statistically optimal because it does not pool the information, but it provides an upper bound on the expected loss. That is, since $\overline{A}_t^k$ is a function of $H_t^{\pi^k}$, $\overline{\ell}_t^k \equiv E[\Psi((X_t^{kn} - \overline{A}_t^k)/kn)]$ is an upper bound on $\ell_t^k$.

**The sequence of upper bounds is strictly decreasing:** For $j \in \mathbb{N}$, let

$$\epsilon_t^j \equiv \frac{1}{n}\left(\left(\sum_{i=(j-1)n+1}^{jn} X_{it}\right) - f_t(\hat{H}_t^j)\right) \ .$$

For $k \in \mathbb{N}$, the average error for the decision rule $\overline{A}_t^k$ when used by firm $kn$ is $(1/k)\sum_{j=1}^k \epsilon_t^j$, and so $\overline{\ell}_t^k = E[\Psi\Big((1/k)\sum_{j=1}^k \epsilon_t^j\Big)]$.

The sequence $\{\epsilon_t^1, \epsilon_t^2, \ldots\}$ of random variables is exchangeable according to the following fact: If $\{x_1, x_2, \ldots\}$ is an exchangeable sequence of random objects with sample space $\langle \mathscr{X}, \mathscr{B} \rangle$ and if the function $f : \mathscr{X}^n \to \mathbb{R}$ is measurable, then the sequence

$$\{f(x_1,\ldots,x_n),\ f(x_{n+1},\ldots,x_{2n}),\ \ldots\}$$

is exchangeable. Therefore, $\overline{\ell}_t^k$ is weakly decreasing according to Lemma 1. One can show that part 1 (resp., part 2) of Assumption 12 implies assumption 1 (resp., 2) in Lemma 1, and hence $\overline{\ell}_t^k$ is a strictly decreasing. Since $\overline{\ell}_t^1 = \ell_t$, we have $\ell_t^k < \ell_t$ for $k > 1$. □

**Assumption 13** *In the computation problem:*

i. *Assumption 1 holds for some $B : \mathbb{N} \to \mathbb{N}$ such that $B(1) = 1$ and $B(2) = 2$;*
ii. *for $\pi \in \Pi$, $\{A_t^\pi\}$ is a linear policy;*
iii. *there is $w$ close to zero,*[15] *such that, for $\alpha \in \mathbb{R}$, there exists $\pi \in \Pi^1$ such that $A_t^\pi = \alpha + X_{1,t-1}$ for $t \in \mathbb{N}$ and $C(\pi) \leq w$.*
iv. *the processes have the decomposition $X_{it} = Y_t + Z_{it}$ described in Remark 1;*
v. *$\{Y_t\}$ 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*
vi. *either $\Psi(\epsilon) = \epsilon^2$ or the stochastic processes $\{\{X_{1t}\}, \{X_{2t}\}, \ldots\}$ are Gaussian.*

[15] This informal terminology *parameter $\rho$ is close to $\eta$* has the usual meaning in Theorem 4B: Given the remaining assumptions, we can find neighborhoods of the values such that, when the parameters are in their respective neighborhoods, there is an optimal firm size.

*Proof of Theorem 4B.*

**Notation:** By assumptions (iv) and (v), we can write $X_{it} = \mu + Y_t + Z_{it}$, where $\{Y_t\}$ and $\{Z_{it}\}$ are mean-zero AR(1) processes. We write these as $Y_t = \gamma Y_{t-1} + V_t$ and $Z_{it} = \beta Z_{i,t-1} + W_{it}$, where $\{V_t\}$ and $\{W_{it}\}$ are noise processes.

**Parameter values:** Assume that $\gamma = \beta = 1$, that $\mathrm{Var}(V_t) = 2$ and $\mathrm{Var}(W_{it}) = 1$, and that $w = 0$. We will claim that our calculations vary continuously with these parameters, so that the results hold when the values are close to the ones given. In particular, the fact that the processes are not stationary when $\gamma = \beta = 1$ does not invalidate the calculations and results. Assume first that $\Psi(\epsilon) = \epsilon^2$. We will later explain how to adapt the calculations to the Gaussian case.

**Intuition:** Firm 1 can make a good forecast of $X_t^1$ simply by observing $X_{1,t-1}$. Firm 1 cannot differentiate $Y_t$ and $Z_{it}$ with this data, but it does not need to. For large $n$, firm $n$ cannot make as good a forecast of each $X_{it}$ because it cannot use recent data about most of the processes. However, what it mainly needs is to forecast $Y_t$ (a law-of-large-numbers effect diminishes the average loss from errors in forecasting the idiosyncratic terms). For each $s \in \mathbb{N}$ and $i \in \mathbb{N}$, $X_{i,t-s}$ is a noisy observation of $Y_{t-s}$. Unlike in the sampling problem, the number of these noisy observations used in a forecast is bounded for each $s$, and so the forecast of $Y_t$ may have a greater expected loss than firm 1's forecast of $X_t^1$.

**Loss for firm 1:** By assumption (iii), firm 1 can compute $A_t = X_{1,t-1}$. Since $X_t^1 = X_{1t} = X_{1,t-1} + V_t + W_{1t}$, the expected error is

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

**Loss for firm n:** Let $n \geq 2$ and $\pi \in \Pi^n$. Assumption (i) implies that the data from dates $t-1$ and $t-2$ that may be included in $H_t^\pi$ is at most one of the following:

**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 $j \neq i$.

In addition, $H_t^\pi$ may include data from periods $t-3$ and earlier.

Consider Case 2. For example, $H_t^\pi$ includes $X_{1,t-1}$, $X_{2,t-2}$, and data from periods $t-3$ and earlier. To construct a lower bound on the expected loss, we can assume that $H_t^\pi$ includes $X_{1,t-3}$ and $X_{2,t-3}$. Since $A_t^\pi$ is a linear decision rule, there are constants $\alpha_1, \alpha_2 \in \mathbb{R}$ such that $A_t^\pi = n\alpha_1 B_1 + n\alpha_2 B_2 + B_3$, where

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

and $B_3$ is a measurable function of $H_{t-3}$. Let

$$B \equiv \frac{X_t^n - X_{t-3}^n}{n} = V_t + V_{t-1} + V_{t-2} + \frac{1}{n}\sum_{i=1}^{n}(W_{it} + W_{i,t-1} + W_{i,t-2})\ .$$

Then the average error is $\epsilon \equiv B + X^n_{t-3}/n - \alpha_1 B_1 - \alpha_2 B_2 - B_3/n$. Because $B_3$ and $X^n_{t-3}$ are independent from $B$, $B_1$, and $B_2$, we have

$$E[\epsilon^2] \geq \mathrm{Var}(B - \alpha_1 B_1 - \alpha_2 B_2) = E[(B - \alpha_1 B_1 - \alpha_2 B_2)^2] \ .$$

This, in turn, is equal to

$$\begin{aligned}(3) \qquad & E\Bigg[\bigg(V_t + V_{t-1} + V_{t-2} + \frac{1}{n}\sum_{i=1}^{n}(W_{it} + W_{i,t-1} + W_{i,t-2}) \\ & \quad -\alpha_1(V_{t-1} + W_{1,t-1} + V_{t-2} + W_{1,t-2}) - \alpha_2(V_{t-2} + W_{2,t-2})\bigg)\Bigg] \\ & = E\Bigg[\bigg(V_t + (1-\alpha_1)V_{t-1} + (1-\alpha_1-\alpha_2)V_{t-2} \\ & \quad +(1/n - \alpha_1)(W_{1,t-1} + W_{1,t-2}) + (1/n - \alpha_2)W_{2,t-2} \\ & \quad +\frac{1}{n}\sum_{i=1}^{n} W_{it} + \frac{1}{n}\sum_{i=2}^{n} W_{i,t-1} + \frac{1}{n}\sum_{i=3}^{n} W_{i,t-2}\bigg)\Bigg] \\ & = 2 + 2(1-\alpha_1)^2 + 2(1-\alpha_1-\alpha_2)^2 + 2\alpha_1^2 + \alpha_2^2 + (3 - 4\alpha_1 - 2\alpha_2)/n \ .\end{aligned}$$

The minimum value $g(n)$ of equation (3) is thus a lower bound on the average expected loss. Solving the first-order conditions for minimization yields $\alpha_1 = 4/7 + 2/7n$ and $\alpha_2 = \frac{2}{7} + \frac{1}{7}n$. We can then show that $g(2) = 3\frac{1}{28}$, that $g(n) > g(2)$ for $n \geq 3$, and that $\lim_{n\to\infty} g(n) = 3\frac{1}{7}$. With similar calculations we can derive even higher lower bounds ($3\frac{1}{3}$ and $4\frac{2}{5}$, resp.) on the expected loss for cases 1 and 3; see Van Zandt and Radner (1999) for details.

**Perturbing the parameters:** Hence, the average expected loss for firm $n > 1$ is at least $\delta' \equiv 3\frac{1}{28}$, whereas firm 1 can attain an average expected loss of $\delta \equiv 3$. Note that these bounds depend continuously on $\gamma$ and $\beta$ and also on $\mathrm{Var}(V_t)$ and $\mathrm{Var}(W_{it})$. Hence, by setting $\gamma$ and $\beta$ close enough to 1, $\mathrm{Var}(V_t)$ and $\mathrm{Var}(W_{it})$ close enough to 2 and 1 (respectively), and the administrative cost close enough to 0, we can still find $\delta' > \delta > 0$ such that $\mathrm{AC}(1) \leq \delta$ and $\mathrm{AC}(n) \geq \delta'$ for $n > 1$.

**The Gaussian case:** Rather than assuming $\Psi(\epsilon) = \epsilon^2$, suppose that the stochastic processes are Gaussian. We impose the initial assumptions on the parameter values stated previously and show that there are $\delta' > \delta > 0$ and a computation procedure for firm 1 whose expected loss is no greater than $\delta$, whereas the average expected loss of any computation procedure for any firm $n > 1$ is at least $\delta'$. The perturbations to the parameter values are handled in the same way as before.

Choose $\alpha \in \mathbb{R}$ in order to minimize $E[\Psi\big(X_{1t} - (\alpha + X_{1,t-1})\big)]$. According to assumption (iii), there is a $\pi \in \Pi^1$ such that $A^\pi_t = \alpha + X_{1,t-1}$. Let $\epsilon^1 \equiv X_{1t} - (\alpha + X_{1,t-1})$ be the error for firm 1 when it uses the procedure $\pi$. We have already calculated that $\mathrm{Var}(\epsilon^1) = 3$.

Let $\epsilon^*$ be a Gaussian random variable with mean $E[\epsilon^1]$ and variance $3\frac{1}{28}$. Choose $\alpha^* \in \mathbb{R}$ in order to minimize $E[\Psi(\alpha^* + \epsilon^*)]$. Because (a) $\alpha^* + \epsilon^*$ and $\alpha^* + \epsilon^1$

are Gaussian and have the same mean, (b) $\mathrm{Var}(\alpha^* + \epsilon^*) > \mathrm{Var}(\alpha^* + \epsilon^1)$, and (c) $\Psi$ is strictly convex and not affine it follows that, $E\left[\Psi(\alpha^* + \epsilon^*)\right] > E\left[\Psi(\alpha^* + \epsilon^1)\right]$. Since $\alpha$ was chosen to minimize the expected loss, $E[\Psi(\alpha^* + \epsilon^1)] \geq E[\Psi(\epsilon^1)]$. Therefore, $\delta' \equiv E\left[\Psi(\alpha^* + \epsilon^*)\right] > E\left[\Psi(\epsilon^1)\right] \equiv \delta$.

Now let $\epsilon^n$ be the error for some computation procedure for firm $n$. As we have shown, $\mathrm{Var}(\epsilon^n) \geq 3\frac{1}{28}$. Using the same argument as in the previous paragraph, we can show that $E[\Psi(\epsilon^n)] \geq E[\Psi(\alpha^* + \epsilon^*)]$. Thus $\mathrm{AC}(n) \geq \delta'$ for $n > 2$, whereas $\mathrm{AC}(1) \leq \delta$. □

## References

Arrow, K. J.: The limits of organization. New York: Norton 1974

Brynjolfsson, E., Hitt, L. M.: Information technology and organizational design: Evidence from micro data. MIT Sloan School of Management and Wharton School (1998)

Brynjolfsson, E., Malone, T. W., Gurbaxani, V., Kambil, A.: Does information technology lead to smaller firms? Management Science **40**, 1628–1644 (1994)

Chandler, A. D.: Strategy and structure. New York: Doubleday 1966

Chandler, A. D.: Scale and scope: The dynamics of industrial capitalism. Cambridge, MA: Harvard University Press 1990

Coase, R.: The nature of the firm. Economica **4**, 386–405 (1937)

Dickinson, H. D.: Economics of socialism. Oxford: Oxford University Press 1939

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

Hart, O.: Firms, contracts, and financial structure. Oxford: Oxford University Press 1995

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

Hayek, F. A. v.: The use of knowledge in society. American Economic Review **35**, 519–530 (1945)

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

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

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

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

Loève, M.: Probability theory II. New York: Springer 1978

Marschak, J., Radner, R.: Economic theory of teams. New Haven, CT: Yale University Press 1972

Marschak, T.: Computation in organizations: The comparison of price mechanisms and other adjustment processes. In: McGuire, C. B., Radner, R. (eds.) Decision and organization, chapter 10, pp. 237–281. Amsterdam: North-Holland 1972 (2nd edn. published in 1986 by University of Minnesota Press)

Marschak, T.: On economies of scope in communication. Economic Design **2**, 1–31 (1996)

Meagher, K. J.: How to chase the market: An organizational and computational problem in decision making. Australian National University (1996)

Mount, K., Reiter, S.: A model of computing with human agents. Discussion Paper No. 890, Center for Mathematical Studies in Economics and Management Science, Northwestern University (1990)

Orbay, H.: Hierarchy size and environmental uncertainty. Koç University (1996)

Radner, R.: Hierarchy: The economics of managing. Journal of Economic Literature **30**, 1382–1415 (1992)

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

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

Reiter, S.: Coordination and the structure of firms. Northwestern University 1996

Robinson, E. A. G.: The structure of competitive industry. Chicago: University of Chicago Press 1958

Van Zandt, T.: Organizations with an endogenous number of information processing agents. In: Majumdar, M. (ed.) Organizations with incomplete information. Cambridge: Cambridge University Press 1998a

Van Zandt, T.: Real-time hierarchical resource allocation. Discussion Paper No. 1231, Center for Mathematical Studiies in Economics and Management Science, Northwestern University (1998b)

Van Zandt, T.: Decentralized information processing in the theory of organizations. In: Sertel, M. (ed.) Contemporary economic issues, Vol.4. Economic design and behavior, chapter 7. London: Macmillan Press 1999a

Van Zandt, T.: Real-time decentralized information processing as a model of organizations with boundedly rational agents. Review of Economic Studies **66**, 633–658 (1999b)

Van Zandt, T.: Real-time hierarchical resource allocation with quadratic costs. INSEAD (1999c)

Van Zandt, T.: Structure and returns to scale of real-time hierarchical resource allocation. INSEAD (1999d)

Van Zandt, T., Radner, R.: Real-time decentralized information processing and returns to scale: Supplementary notes. INSEAD and Stern School of Business, New York University (1999)

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

Wilson, R.: Informational economies of scale. Bell Journal of Economics **6**, 184–195 (1975)