MATHEMATICAL MODELS OF THE MANUFACTURING LEARNING CURVE

By

Paul Speaker

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Applied Mathematics

2010

ABSTRACT

MATHEMATICAL MODELS OF THE MANUFACTURING LEARNING CURVE

By

Paul Speaker

The industrial learning curve is the widely observed process whereby, through manufacturing experience, improvement in quality metrics improve over time. Such quality metrics include cost, mean product lifetime, and error rate.

This dissertation motivates, builds, describes, and investigates a *learning curve* model based on ideas of information theory and statistical mechanics. We first examine the relationship between the learning curve and learning in other contexts. Then we create a model of the learning curve based upon the ideas of statistical mechanics. We then build two continuous models for the learning curve. Several additional features of learning curves, such as multi-dimensional learning and forgetting, are realized by the model.

The first continuous model is based on a new mathematical concept: *weak convergence of sequences of sets*. The idea of weak convergence of sequences of sets is developed for both deterministic and random sets. Several functional analytic-like results are then proven for such weak limits. The weak limits are then used to construct a model of learning. In doing so, the general approach of creating a continuous model from a discrete model via such a weak limit is described.

The second continuous model to be described is a partial differential equation model. By making the learning process Poisson in time, it is shown that the learning model is a discretization of a transport equation. Aspects of this transport equation are then explored and interpreted.

Further, some simulations of the model are made. It is shown that the model simulations generate the generally observed form of the learning curve. The dissertation concludes with an exploration of further directions to be pursued in future investigation of the learning curve phenomenon.

Dedication

To Liisa, Isadora, and Evangeline.

Acknowledgements

For a large, interdisciplinary project such as this one, there are many people to thank. This investigation started with a project on the manufacturing learning curve for Michigan State University students sponsored by General Motors. I would first like to thank David Field and Randy Urbance of General Motors for their numerous helpful suggestions on the topic. I would also like to thank the other members of my team, Michael Olson and Rahul Datta for their help in getting me thinking about the project. Peter Magyar of Michigan State was the project supervisor, and his guidance both during and after the project time period.

I would also like to thank my dissertation committee members: Roger Calantone, Gabor Francsics, Baisheng Yan, and Keith Promislow. Roger Calantone pointed me towards the research on organizational learning by James March, which was very helpful. Gabor Franciscs and Baisheng Yan had numerous helpful suggestions with the chapter on the partial differential equation model for the learning curve. Finally, Keith Promislow was very helpful helping me gain an understanding of some of the physical analogies for random sets, in particular, with the capacitor example.

I would also like to thank the outside help. I would certainly like to thank Igor Molchanov. His seminal book on random sets, as well as his suggestions for future directions, were very helpful in my formulating the learning problem in the context of random sets. I would also like to thank the people at the Econophysics Colloquium in Ancona who made helpful suggestions for connecting the learning curve problem to statistical mechanics.

Finally, I would like to thank most deeply Charles R. MacCluer for the multitude of suggestions, his patience with this work, and his overall focus on what this dissertation should look like. The suggestions he has provided are too numerous to list here, but they permeate the entire dissertation. Without his tireless efforts, this dissertation would possess the breath or depth which it has.

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter 1

Literature Review

1.1 Introduction

It is generally observed that manufacturing measurements of quality (*quality metrics*) exhibit an improvement over time. The improvement given by this *learning curve* is generally described as a function of a single parameter—the total quantity produced (called the *accumulated production*) [Dutton 1984]. Various models for this learning curve have generally been expressed as either a power law of the accumulated production q:

$$C(q) = C_0 q^{-\alpha}, \tag{1.1}$$

or as an exponential law of the form

$$C(q) = C_0 e^{-\lambda q}, \tag{1.2}$$

where $C(q)$ is the metric associated with the learning curve and q is the accumulated quantity produced [Kantor Zangwill 1998]. This metric $C(q)$ is a quantitative measure that a manager would wish to minimize, such as manufacturing cost or production errors. Past models have been little more than a choice between which function is a better fit to the data, rather than a causal explanation for why this might be observed. The goal of this work is to provide such an explanation

Several deviations from these analytic forms (1.1) and (1.2) have also been observed. For example, some initial rates of learning occur less quickly than either model would suggest. Initial downward concavity has been observed in a number of studies [Garg Milliman 1961][Muth 1986]. A second, more significant deviation from the standard forms (1.1) and (1.2) is that, after a period of time, improvement ceases, so that the learning curve asymptotically approaches a non-zero level, rather than approaching infinitesimal costs, as predicted by (1.1) and (1.2) [Conway Schultz 1959]. This *plateau phenomenon* is a fundamental characteristic of any learning curve over the long term, as costs are never expected to approach zero, even asymptotically, a fact ignored by many practitioners of the learning curve.

Attempts to explain the learning curve fall into two broad categories: *deterministic* and *stochastic*. Moreover, there have been observed deviations from both deterministic models and stochastic models.

1.2 Deterministic Models

Several deterministic models have been proposed in the past, such as in [Levy 1965]. The leading deterministic models were proposed by Kantor and Zangwill in a pair of papers [Zangwill Kantor 1998] [Kantor Zangwill 2000]. These papers propose that the improvement in a metric C follows the intuitively plausible differential equation

$$\frac{dC}{dq} = -cE(q)C(q), \tag{1.3}$$

where $E(q)$ is a quantitative measure of management effectiveness and where q is the accumulated production. Therefore, the rate of improvement is proportional both to the management effectiveness and to the amount by which the metric may still be decreased. To generate both power law and exponential models for the learning curve, the authors make the further assumption that $E(q)$ should itself have a power law form as a function of $C(q)$, that is,

$$E(q) \propto (C(q))^{\alpha}. \tag{1.4}$$

The choice between a power law and an exponential law then depends on α, the power of $C(q)$, since $\alpha = 0$ yields an exponential while a positive α yields a power law for C.

1.3 Stochastic Models

Several stochastic models have been proposed for the learning curve. For example in [Levy 1965] a model was proposed wherein the learning curve was solely dependent on the remaining amount by which improvement can be made. This idea yielded the differential equation

$$\frac{d(1/x)}{dn} = \mu(P - 1/x), \tag{1.5}$$

where x is the quality metric, P is the eventual plateau level, and μ is a parameter of the equation. This differential equation has the solution

$$x(n) = \left[P - \left(P - \frac{1}{x_0}\right) e^{-\mu n}\right]^{-1}, \tag{1.6}$$

where x_0 is the initial level for the quality metric. However, as stated in [Muth 1986] this model is not capable of reproducing the power laws typically seen in actual industrial settings.

An influential stochastic model for the learning curve is given in [Muth 1986]. In this paper, Muth takes a "search theory" approach whereby "[c]ost reductions are realized through independent random sampling from a space of . . . alternatives." In this model, the distribution of such alternatives is assumed to approach a power law in some lower bound. The probability of picking an alternative depends only on its distribution, that is, all choices are equally likely, and the evolution is driven merely by the distribution of choices.

Muth then proceeds to explain deviations from a power law relationship. He explains the plateau effect as being caused by an increase in the search cost over time, so that at some point searching for cost savings ceases being cost effective. He explains the initial deviation from a power law as being caused by searching prior to production, in line with what is argued in [Garg and Milliman 1961]. It is this type of learning curve model that will serve as a springboard for both the discrete and continuous models discussed below in this work.

Chapter 2

Insights from Other Research into Learning

Both the deterministic and stochastic models have their appeal. The deterministic model resembles the sort of scientific law that leads to predictive modeling. Muth's stochastic model on the other hand uses the intuitive idea of sampling of all possible cost reductions. However, both theories require the assumption of a power law distribution for some part of the theory, and hence neither are based on first principles.

A more foundational approach would incorporate the underlying causes that lead to the observed learning curve. As the very term implies, the primal cause of this evolution is learning. Workers on the plant floor learn to do their tasks better over time. Managers learn how better to allocate labor and capital. Over time, this learning allows the manufacturing to be done in a more efficient way.

2.0.1 The Analogy to Artificial Intelligence

It is reasonable then to turn to learning theory for sources of a complete model of the learning curve. The field of Artificial Intelligence (AI) has a robust model of learning. While many acts of innovation have occurred via AI [Russell Norvig 2002], the innovation process itself has not been viewed through the AI lens. Applying AI, which takes place inside a computer, to a system of humans and technology might seem too simplistic to analogize to the field of manufacturing, which employs many different resources–both human and mechanical—in many different ways.

On the other hand, a distributed computer network also has many distinct components, networked over long distances, working on different aspects of a problem. Thus, an AI network has such distinct parts learning diverse tasks to improve performance.

In AI a crucial distinction is made between *supervised* and *unsupervised* learning [Engel and Van den Broeck 2001]. Supervised learning occurs when a machine attempts to generalize examples to fit a general rule. Examples of supervised learning are *regression* [Hastie et al. 2001] and *classification* [Sperduti and Starita 1997]. In regression analysis, example data is given to a system, and the system then makes a best fit of the data to an assumed form such as linear or logistic models. The system can use the resulting model to make predictions of future data (extrapolations). Similarly, classification (which can be understood as discrete regression) uses a set of examples to make the best fit for future data by determining to which classification they belong.

Unsupervised learning does not assume a single preconceived model of the output for the data. Instead, the learning system has to decide on a model to be used, as well as a way to fit the given data to that model. In this way, unsupervised learning is much less structured or dependent on fixed probabilities than structured learning.

An example of unstructured learning occurs in the study of clustering of points, which is very important in image recognition. For clustering, there may be many different categories of clusters, such as dense (many points clustered together), interacting (for example, two clusters which appear to be attracted to each other), or uniform (constant density throughout a cluster). The AI system has to decide which of the models has best fit, or even create a new category to fit existing data, as well as make predictions based upon this choice on the location of additional features. The unstructured learning AI system can also modify these categories based upon experience (how dense does a cluster have to be to be considered is dense, for example?) and reassign probabilities for which a possible clustering pattern might be met.

This dichotomy between supervised and unsupervised learning is directly analogous to the distinction between *exploitation* and *exploration* [March 1991]. Exploitation is associated with mechanistic structures, routinization, bureaucracy, and control [He and Wong 2001]. Similarly, supervised learning uses a fixed, mechanistic structure to build and refine outputs. On the other hand, both exploration and unstructured learning are more open-ended processes, in that the output goal can change and be modified over time, based upon, for example, external inputs (instructions from a manager).

2.0.2 The Application of Statistical Mechanics to Artificial Intelligence

Two aspects of artificial intelligence theory of learning merit close attention: *energy minimization* and the *connection with information theory.* Learning in the AI context has been described as a stochastic minimization of energy in an information-theoretic setting [Seung 1992][Watkin 1993]. Similarly, in the manufacturing setting, one attempts for example to attain a minimization of cost or errors in the manufacturing process.

Information theory is undoubtedly relevant to understanding the learning curve. The evolution of the learning curve is driven by learning, and any learning in the formal sense is the acquisition of information. The landmark work done by Claude Shannon demonstrated that the principles of statistical mechanics are applicable to the act of learning [Shannon 1948]. Later studies demonstrate that the rate of learning is in fact very predictable using the principles of statistical mechanics [Seung 1992][Watkin 1993]. Statistical mechanics is now the standard model for explaining learning in the AI context [Engel Van den Broeck 2001].

March's model of learning [March 1991] and its refinement over the years have all the outward appearance of modeling in statistical mechanics. March's theory has four key aspects [Miller 2006]:

i. The environment of an organization is modeled as an m-dimensional belief vector with each coordinate randomly assigned a value of 1, 0, or -1, with m equal to the number of individuals in the network.

ii. Each individual's belief about reality is represented by 1 (positive belief), 0 (no belief), or -1 (disbelief).

iii. Learning within the firm causes change in the values of these belief values.

iv. The organizational code adjusts over time to reflect the dominant belief among better-performing members of the organization.

Therefore, according to March's theory, learning occurs via interaction between members of a firm and the organizational code, and these interactions are probabilistic in nature. One refinement of this theory is described in [Miller 2006], where learning is also allowed to take place between individuals in a firm, with a probabilistic bias toward learning by and between "proximate neighbors."

This type of modeling is exactly how physicists use statistical mechanics because it attempts to treat a large number of discrete objects in a mechanical, yet probabilistic, way. In statistical mechanics the first order approximation for interactions between individual particles is with a *mean field*. A mean field is an averaged effect of all the interactions a particle and the system. For example, in studying the gravitational field of the earth, one does not look at the gravitational field of each atom, but one instead looks at the mean field averaged over the entire earth.

In the instance of March's theory, the organizational code plays the role of the mean field. After this level of approximation has been examined, one can look at higher orders of approximation, which generally focus on interactions between individuals. In these higher orders of approximation, it is the interaction between nearest neighbors that is typically important, while other interactions are negligible.

This hierarchical approach in statistical mechanics can be seen, for example, in the modeling of magnetic domains [Kodama Berkowitz 1999], protein molecules [Schilk 2000], and galaxies [Binney Tremaine 1988], but these steps are ubiquitous in all modeling problems involving statistical mechanics. Therefore, since modeling learning within a firm has already aped the process of statistical mechanics, it is worth exploring how much further statistical mechanics can explain learning.

2.1 Statistical Mechanics in Other Areas

As noted in [Demsetz], the firm occupies a anomalous position in standard economic theory. Specifically, the firm is considered to be a "black box" which acts as a single economic actor (an *agent*) and competes with other actors in the marketplace. Management theory on the other hand looks at the forces of different actors (agents) within a firm and seeks to find ways that their utility may be maximized. Such a disconnect is puzzling since it is apparent that the utility maximization point of a firm need not coincide with the utility maximization point of actors within the firm.

While the problems studied in management theory have not entered the realm of general economic modeling, the last twenty years have witnessed an explosion of attempts to apply economic theory to management problems [Donaldson I 1990]. As shown in [Donaldson I 1990][Barney 1990], this spread of economic ideas has been met with some skepticism. This skepticism has been attributed to concerns about the methodologies of economics [Barney 1990][Donaldson II 1990], as well as to protection by an academic group of its own turf [Barney 1990]. However, this diffusion of economic ideas into management theory has not engendered a theory of the effects that such organizational economics might have on the overall macroeconomic picture.

2.1.1 Differences between Management Theory and Organizational Economics.

In 1990 Lex Donaldson published a pair of papers outlining the connection between management theory and economics [Donaldson I 1990][Donaldson II 1990]. Donaldson argued that while traditional economic theory had something to add to the theory of managerial organization, organizational economics was overly concerned with the actions and motivations of the individual within the firm. In contrast, Donaldson explained that management theory incorporated action on the collective level as well as individual motivations.

Central to creation of the emergent, collective level is the issue of *trust*. An simple example of how trust creates a collective scale can be seen with the standard form of the *prisoners' dilemma*. Without trust, the two prisoners act as individuals, and there is no collective level of action. On the other hand, if the prisoners trust each other, they work as a single unit, and motivations have to be interpreted as a type of aggregate behavior. Donaldson states that, in the standard theory of organizational economics, trust within an organization only exists through enforcement mechanisms—"It is 'trusting' someone only after you have them firmly under control" [Donaldson II 1990]. In contrast, management theory allows for trust of managers even in the absence of these type of controls.

In general, organizational economists discount the idea that an organization is an entity unto itself—"organizations, as fiction, have been invented by humans to make sense of the world." [Donaldson I 1990] In contrast, as Donaldson says, "... systems analysis is at the heart of management theory because managers are responsible for the conduct and performance of organizational systems." [Donaldson II 1990] These type of collective models are the central aspect of any organizational theory that does treats the organization itself as more than a convenient fiction.

2.1.2 Econophysics and Learning.

Physicists readily recognize and understand the issues brought up by Donaldson and March. This problem is one of *scale formation.* Economics, in particular neoclassical economics, examines issues related to actions occurring on the single scale of the individual. In contrast, organizational theorists have shown that many actions are best viewed as occurring at larger scales, or even multi-scales [March 1991].

Since the ideas of statistical mechanics have been successfully used to model learning in some contexts, it is natural to investigate whether it may be used to model learning in other contexts. In both artificial intelligence and an organizational structure, learning typically occurs as through a coupled interaction of subsystems to external stimuli. Further, this coupling between subsystems is tighter than what one might see between individuals in a general economic context. Finally, since the principles of management have become more uniform and standardized over the years [Dillard 2004], more predictability should be evident.

2.2 Further Directions for Statistical Mechanics in Management Theory

Just as management theory might learn something from econophysics, econophysics might gain some breadth from venturing into management theory. As we have argued above, the regularity of relations within a firm should make it a more natural field for a study that depends on this regularity. Perhaps more important in the long run, however, is the centrality of learning in management theory. As some have noted, one severe deficiency in most of econophysics is that nearly all of the models given are exchange-only models, with no discussion of increases in wealth [Gallegati 2006]. This limitation is in contrast to the idea—a central to economics since at least Adam Smith—that exchanges occur to increase the wealth of each participant. Learning, on the other hand, necessarily involves an increase in a valuable commodity. Since learning has already been extensively explored within the framework of statistical mechanics, the learning paradigm might be a natural avenue through which to incorporate growth.

Management theory is a natural fit to the ideas of econophysics. In fact, as described above, some management theorists have already anticipated these ideas. It is hoped that this work will spur the development of the application of econophysics to management theory.

Part II

The Mathematical Model

Chapter 3

The Discrete Case

3.1 The Parameters of the Model

3.1.1 Derivation of the Probability

Suppose there are n lessons that a manager or production worker can learn that will improve the quality metric. Let us fix ideas by supposing this quality metric to be unit cost, so that learning the i-th lesson will—from that moment forward—yield a cost saving of c_i dollars per unit. The ideas developed in this way will carry over naturally to improvements in other quality metrics.

While in some cases learning one lesson might affect the probability of learning some other lessons, in general, most lessons are independent of all other specific lessons. Learning a lesson at one stage of the manufacturing process will not in general affect decisions made regarding another stage of the manufacturing process. While in practice many lessons will depend on others, the probability of one given decision relying on a second given decision is low. Therefore, for purposes of this model *there is no correlation between lesson choices.*

Also, any improvement, once made, cannot be made a second time. But how does this removal of a decision choice affect the probabilities of choosing decisions not already made? One possible answer would be to have the probabilities of the other choices increase, so that the partition function is only taken over the remaining choices. This idea (the *elimination hypothesis*) has an intuitive basis, since some managerial choices will become more obvious once other choices are taken.

However, the elimination hypothesis does not survive closer scrutiny. The elimination hypothesis guarantees that, at each increment of manufacturing, learning will take place. In contrast, learning the last lessons will take more steps than the first lessons. It becomes harder and harder to learn the lessons later. The methods of manufacturing change less and less over time. Therefore, we reject the elimination hypothesis.

So the question becomes how to reconcile the two above rules. If improvement cannot be made a second time, how can the probabilities of learning the lesson stay the same the manufacturing lifecycle? The resolution is by way of scoring the learning. We will keep the lesson space, and their probabilities, fixed over time. However, if a lesson which has been previously learned is chosen a second time, no additional improvement in the quality metric will be achieved. Over time, as more and more lessons are learned, the probability of choosing a lesson learned previously will increase. This rule implies that improvements to the process will decrease over time, and that fewer and fewer new lessons will be learned.

Some cost-saving ideas are more obvious than others; let p_i be the likelihood that the i-th lesson will be learned. Again, we will assume that *there is no correlation between lesson choices.* Therefore, a lesson learned at one step will be statistically independent from a lesson learned at the previous or subsequent step.

Finally, we assume that *the rate that lessons are drawn from the pool of possible* *lessons is proportional to production rate.* Learning proceeds through production experience. So for example, as the k-th hundredfold unit is completed, the k-th draw is taken from the lesson pool. This assumption is borne out in the learning curve literature, as learning patterns are observed over production cycles, instead of the length of time making products.

Based on these rules, we may make the following conclusion regarding the structure of learning.

Theorem 3.1 *The expected cost savings $C(k)$ after k draws from the pool of n lessons is given by the rule*

$$C(k) \;=\; \sum_{i=1}^{n} (1-(1-p_i)^k)c_i. \tag{3.1}$$

Proof. We proceed via induction. After one draw the expected savings are clearly

$$C(1) \;=\; \sum_{i=1}^{n} p_i c_i. \tag{3.2}$$

The total expected savings after k draws is the expected total savings from the previous $k-1$ draws plus the expected savings for the k-th draw (which assumes individual lessons have not been drawn in the previous $k-1$ draws). In symbols,

$$C(k) \;=\; C(k-1) \;+\; \sum_{i=1}^{n} p_i(1-p_i)^{k-1}c_i \tag{3.3}$$

$$= \sum_{i=1}^{n} (1-(1-p_i)^k)c_i. \tag{3.4}$$

Corollary 3.1 *The cost U per unit can be expected to decrease with production by* *the rule*

$$U(k) \ = \ U_0 \ - \ \sum_{i=1}^{n} (1-(1-p_i)^k)c_i, \tag{3.5}$$

where U_0 is the initial cost per unit when production commences.

Example 3.1. Suppose that all savings $c_i = 1$ and that each lesson is equally probable: $p_i = 1/n$. Then the evolution of the expected unit cost U is given from (10) by

$$U(k) \ = \ U_0 - n\big(1 - (\frac{n-1}{n})^k\big). \tag{3.6}$$

Thus as one would expect, for large k, well along in the production, cost per unit is falling and approaching its plateau $U_0 - n$. Because $U''/U' = -\log k$ for lesson draws $k = 1, 2, \ldots$, initial relative upward concavity is small in contrast to the abrupt empirical models (3.1) and (3.2).

3.1.2 How lesson probabilities are determined

Let us now uncover why some production lessons are more likely to be learned than others. The form of the learning curve across industries has been remarkably uniform, suggesting that such a determination must exist, and that it is likely to be very simple. Learning is driven by the push by managers to improve some quality metric, such as production cost. In artificial intelligence, a fairly strong correlation has been noted between the likelihood of learning a lesson and the benefit accrued from that lesson (Russell and Norvig, 2002). These observations indicate that *the decision probability can be treated solely as a function of the corresponding cost savings*, that is

$$p_i = p(c_i), \tag{3.7}$$

where $p = p(c)$ is some smooth function of the nonnegative real variable c.

Theorem 3.2 *The probability p_i that the i-th lesson will be drawn is given by the rule*

$$p_i \;=\; \frac{e^{\beta c_i}}{\sum_{j=1}^{n} e^{\beta c_j}}, \tag{3.8}$$

where the constant β is determined by the effectiveness of the manager.

Proof. The following argument will be familiar from statistical mechanics: Suppose the n lessons $x_1, x_2, \ldots, x_n$ deliver the respective cost savings $c_1, c_2, \ldots, c_n$, and suppose each individual lesson x_i will be chosen with probability p_i. The probabilities p_i should not depend on the currencies used to value the cost savings c_i, so let us fix once and for all the value of the average cost savings after one draw

$$C(1) \;=\; \sum_{i=1}^{n} p_i \, c_i. \tag{3.9}$$

Let us draw one lesson after another, returning each to the lesson pool after each draw. The relative frequency of drawing the i-th lesson will of course asymptotically approach p_i as the trials proceed. By the Boltzmann H-theorem (Feynman, 1972), the largest number of distinct trials—the freest and most realistic experiment—occurs simultaneously with the maximum value of the entropy

$$H \;=\; -\sum_{i=1}^{n} p_i \log p_i \tag{3.10}$$

subject to the obvious constraint

$$p_1 + p_2 + \cdots + p_n \;=\; 1 \tag{3.11}$$

and subject as well to

$$p_1c_1 + p_2c_2 + \cdots + p_nc_n \; = \; C(1). \tag{3.12}$$

Applying the method of Lagrange multipliers to $Q = H + \lambda \sum p_i + \beta \sum p_i c_i$, we see that H subject to its two constraints will maximize when

$$\frac{\partial H}{\partial p_i} \; = \; -\log p_i - 1 \; = \; -\lambda - \beta c_i, \tag{3.13}$$

which is $p_i = \alpha \exp(\beta c_i)$ with $\alpha = \exp(1 + \lambda)$. Because of (3.13), the quantity α is the reciprocal of the denominator of right side of (3.9).

Observation. The expected cost savings (3.12) after one draw is similar to a familiar thermodynamic quantity:

$$C(1) \; = \; \frac{1}{\beta} \sum_{i=1}^{n} p_i \log p_i \; - \; \frac{1}{\beta} \log \alpha, \tag{3.14}$$

where $1/\alpha$ is the *partition function* (Feynman, 1972).

Let us combine the previous two theorems.

Theorem 3.3 *As production of a good proceeds, lessons are drawn (with replacement) independently from the pool of n possible lessons. Suppose the i-th lesson, once learned, yields an ongoing cost saving of c_i per unit manufactured. Once this saving is realized, it is unavailable for future improvement. Then the expected cost per unit must decrease by the rule*

$$U(k) \; = \; U_0 - \sum_{i=1}^{n} [1 - (1 - \alpha e^{\beta c_i})^k] c_i, \tag{3.15}$$

where U_0 is the initial cost per unit, where k is proportional to accumulated production *q, where*

$$\frac{1}{\alpha} = \sum_{j=1}^{n} e^{\beta c_j}, \tag{3.16}$$

and where β is the effectiveness of the manager.

Remark 3.1: To understand how the model (3.15) applies to learning, one can look at the various limits of the *effectiveness factor* β , which can be considered a measure of managerial focus. In the limit of a small effectiveness factor (which corresponds to a high temperature in thermodynamics) the probability factors will be very close together. Therefore, in this limit, which corresponds to low managerial efficiency, a manager would be equally likely to choose an option that yields little improvement or one that yields great improvement. Therefore, in the limit of low managerial efficiency, the rate of improvements is just a matter of luck, rather than any skill on the part of the managers.

On the other hand, in the limit of a high effectiveness factor (which corresponds to a low temperature setting in thermodynamics), the manager is very likely to make the decisions which yield the best improvements first. Therefore, a high managerial efficiency yields a much more rapid rate of improvement than does a lower managerial efficiency. Rather than being a matter of luck, the rate of improvement comes directly from the skill of the managers and all others who are making the decisions that cause any improvements in a plant.

3.1.3 Entropy

The concept of *entropy* is central to statistical mechanics and information theory. While entropy is often qualitatively considered to be a measure of the disorder of a system, it is more accurately expressed to be the tendency of a system towards dispersion of the possible states. The equation for the entropy H of a system is given by the following definition.

Definition 3.1 *The* entropy H *of a system of states with corresponding probabilities* p_i *is given by*

$$H = -k \sum_{i=1}^{n} p_i \ln p_i, \tag{3.17}$$

where value k *is called the* Stefan-Boltzmann *constant.*

The Stefan-Boltzmann constant is ubiquitous in statistical mechanics. In a thermodynamic system (3.9), the parameter $\beta = 1/kT$, where T is the *temperature* of the system.

Using the form (3.9) for the probabilities given for our system and $\beta = 1/kT$, the entropy for our learning system becomes

$$H = -\frac{1}{T} \sum_{i=1}^{n} \Delta C_i \, p(\Delta C_i) + \Omega, \tag{3.18}$$

where Ω is constant. Equation 3.11 shows that a change in entropy for learning one lesson is directly proportional to expected cost savings for this lesson. According to Boltzmann's H-Theorem, changes that increase the entropy of the system will be statistically favored. Therefore, this model predicts that learning will tend to increase the expected value of cost savings.

3.1.4 Interpretation of Results

Equation (3.15) for the expected quality metric $C(k)$ yields some important consequences. The form is not a simple power law or exponential. It is in fact a sum of exponentials, which in fact is a form suggested in [Zangwill and Kantor 1998]. But can it accurately match observed learning curves?

As noted in section 1.1, one characteristic of learning curves missing from most standard models is the plateau effect. In contrast, the plateau effect is present in our model (3.15). This can be seen by looking at the asymptotic limit of the cost evolution formula as k becomes very large.

Theorem 3.4 *As k becomes very large, the expected total cost $C_0 - \overline{C}(k)$ has the asymptotic behavior*

$$\lim_{k\to\infty} C_0 - \overline{C}(k) = C_0 + \sum_{i=1}^{n} \Delta C_i. \qquad (3.19)$$

Proof. Clear.

Remark 3.2: After many decisions, all options of cost savings have been exhausted, and the cost of production levels out to a finite, non-zero value. This is consistent both with what is seen in long term learning curves and what is intuitively expected. The cost of production is never expected to go to zero, since the production of a product will always have, if anything, the commodity costs of raw materials in making the product. *Thus our model (3.15), in contrast all previous models, incorporates a plateau effect.*

Remark 3.3: The second characteristic of many observed learning curves is the initial lack of an upward concavity. While both the exponential and power law functional forms maintain an upward concavity throughout the curve, many observed learning curves either have no initial concavity (linear) or have a downward concavity. This phenomenon is sometimes described as an *initial rate of learning* which was slower than what would be expected from the overall shape of the curve [Muth 1986].

One can examine the second derivative (with respect to k) of equation (3.15) to determine whether the graph of equation (3.15) has the proper characteristics of concavity. The power-law forms of the learning curves of previous authors have asymptotically large second derivatives for small values of the amount of production. In contrast, the function of (3.15) has a finite derivative that becomes very small as the number of choices n is increased. This fact can be seen in the case where all improvements have equal weighting, which, as will be proven later in Theorem 5.3, is maximal for this type of expression.

Lemma 3.1 *When probabilities are equal, in the limit of a large number of lessons learned, the second derivative of the learning curve*

$$C_0 - \overline{C}(k)$$

is $O(n^{-1})$.

Proof. The second derivative is given by

$$\frac{d^2\overline{C}}{dk^2} = \sum_{i=1}^{n} \Delta C_i (1-p_i)^k \ln^2(1-p_i). \tag{3.20}$$

The equation for the second derivative of $\overline{C}(k)$ for equal probabilities becomes

$$\frac{d^2\overline{C}}{dk^2} = n\Delta C \left(1 - \left(1 - \frac{1}{n}\right)^k\right) \ln^2\left(1 - \frac{1}{n}\right). \tag{3.21}$$

For large n, this becomes the asymptotic relation

$$\frac{d^2\overline{C}}{dk^2} \approx n\Delta C \frac{1}{n^2} = O(n^{-1}). \tag{3.22}$$

Remark 3.4: Thus, the value of the second derivative vanishes as n goes to infinity. Furthermore, this phenomenon of a vanishingly small second derivative can be intuitively explained by the fact that when there are a large number of choices, removing a few does not significantly change the expected value of the cost savings for each step. This leads to a approximately linear form at the outset of production. Thus, *this second defect of previous models is corrected by our model.*

The results of this chapter are summarized in the following theorem.

Theorem 3.5 *For a learning system with n lessons $\{l_1, \ldots, l_n\}$ to be learned with corresponding cost savings $\{\Delta C_1, \ldots, \Delta C_n\}$, the evolution of the learning curve for a quality metric C as function of the accumulated production k with managerial effectiveness parameter β is given by*

$$C - \overline{C}(k) = \sum_{i=1}^{n} (1 - (1 - p_i)^k) \Delta C_i, \tag{3.23}$$

where

$$p_i = \frac{e^{\beta \Delta C_i}}{\sum_{j=1}^{n} e^{\beta \Delta C_j}}. \tag{3.24}$$

3.2 Summary

We have derived a model for the learning curve with a discrete number of decisions. As important as this step is, it is likely to be intractable in practice given the large number of decisions which might come into play. Therefore, our next step will be to develop a continuous model which will be more tractable. In fact, two different such models are developed in the next two chapters. But before departing the discrete model, we will examine several possible augmentations.

3.3 Multi-Dimensional Learning

3.3.1 Introduction and Motivation

Any analysis of learning curves is complicated by the fact that managers typically have more than one quality metric that they seek to improve. For example, improving manufacturing cost alone might not be the only goal; decreasing a defect rate would also be a desirable goal.

To date there has been really no analytical explanation of how this *trade-off effect* occurs. Several empirical studies in fact have been done. For example in [Gino 2006] a study was made of learning of a new technology by cardiac surgical units. This study found a tradeoff effect between efficiency and innovation.

This trade-off effect is especially familiar for anyone who researches mathematics education. In performing arithmetic, students display a natural trade-off between speed and accuracy. While both speed and accuracy are desirable goals, it appears throughout that improving one goal often hampers development the other, if not making the other goal worse.

In manufacturing, March's exploration/exploitation distinction is very important. Some of the research talks about a trade-off between exploration and exploitation. However, this trade-off is analytically different from the other trade-offs discussed above, since in the trade-off between exploration and exploitation there is only a single quality metric considered.

Since there has been no analytical explanation of how the trade-off effect works, the trade-off effect has been purely an empirical observation. Furthermore, confusion is caused within the literature by referring to two somewhat distinct phenomena as a *trade-off*. A trade-off between exploration and exploitation is different from a trade-off between striving to improve two different quality metrics, since the exploration/exploitation trade-off still only involves the improvement of a single quality metric. Henceforth the trade-off effect will be restricted to the effect of a multi-dimensional quality metric.

3.3.2 Modeling Multi-Dimensional Learning

The learning curve model based on (3.13) is able to incorporate the trade-off effect in a very natural manner.

Theorem 3.6 *For a learning curve where two quality metrics are being simultaneously improved then the probability of a lesson to be learned is given by an exponential of a linear combination of the improvements of the metrics for that lesson, that is*

$$p_i \propto e^{-(\beta_1 \Delta Q_i + \beta_2 \Delta R_i)}. \tag{3.25}$$

Proof. By Theorem 3.2, the probability of learning the i-th lesson with improvements in the two quality metrics is equal to Q_i and R_i has to satisfy both

$$p_i \propto e^{\beta_1 \Delta Q_i} \tag{3.26}$$

and

$$p_i \propto e^{\beta_2 \Delta R_i} \tag{3.27}$$

for some constant values β_1 and β_2.

Remark 3.5: In the one-dimensional model (3.13) the quality metric took the place of energy in statistical mechanics. The present model (3.25) results in the most logical extension to a multidimensional setting—that a linear combination of the quality metrics take the place of energy. In both the one-dimensional and multi-dimensional models a minimization of the energy analogue, but additional features are present in the multi-dimensional case. In the one-dimensional case, the minimum of the energy is independent of β. However, in the multi-dimensional case the minimum will depend on the relative values of the positive constants denoted by β_i.

3.4 Learning with Forgetting

It is well known that lessons which have been learned can also be forgotten. This forgetting can occur in two formats: *while production is occurring*, and *during breaks in production* [Bailey 1989].

It is in fact very simple to model both incarnations of forgetting in the context of the learning curve. We first examine the case of forgetting during suspensions of production. Rather than the "forgetting" curve being driven the accumulated production, the forgetting curve will be a function of time t. Since we are now using a continuous rather than discrete variable, it is natural to expect that the forgetting of previously learned lessons will be a Poisson process with rate λ_f, a forgetting rate. For any lesson learned, the probability p_i that a lesson will be forgotten then will obey the rule:

$$p_i \propto e^{-\lambda_f \Delta C_i}. \tag{3.28}$$

In this expression it is important to note that the change in the quality metric ΔC_i for a forgotten lesson will be equal in magnitude but opposite in sign for a learned lesson. This gives rise to a reversal of probabilities, with respect to the probabilities in learning. In learning, the most valuable lessons were the ones most likely to be learned; in forgetting, however, it is the most valuable lessons are the ones *least* likely to be forgotten, and vice versa. This idea for forgetting has been explored for example in [Bailey 1989].

This type of forgetting process can be extended to the situation of forgetting intermingled with learning. Two simultaneous events occur; learning occurs as a function of accumulated production, while forgetting is a Poisson process as a function of time. The faster the production rate is, the more that learning will relatively outstrip production. And for learning, the limit of the production rate going to zero will yield the forgetting process described previously.

Chapter 4

Continuous Model I: Weak Convergence of Sequences of Sets

One approach to making the discrete model of Chapter 3 into a continuous model leads to some interesting new mathematics. We will first develop this mathematics, and then show how it relates to the model.

4.1 Introduction and Motivation

The basic idea of a *dynamic random set* can be illustrated with the following metaphor.

Thought Experiment: Consider raindrops hitting a sheet of paper. The distribution of raindrops in time over the sheet is random and is therefore given by some probability distribution. Over time, the water from the raindrops will completely cover the sheet of paper. In this model, the raindrops all have a minimum positive size, and only a finite number of raindrops are needed to cover the sheet of paper. However, the diversity of shapes of the possible raindrops makes it difficult to determine the amount of the sheet of paper that is covered by raindrops. If a limit is taken so that the number of raindrops increases, while the size of the raindrops goes to zero (with the product of the number of raindrops and the size of each raindrop stays constant), we might arrive at a continuous model of the problem.

However, the approach has fundamental problems. This analysis is done as a limit of the sequence of sets. However, in the limit, this sequence of sets converges (in the Hausdorff metric) only to a countable number of singletons. At each stage in the limit, we have a positive measure, and the limit of the measures is also positive. However, the measure of the limit is zero. In the limit, therefore, the measure disappears. Therefore, an alternative approach is needed.

If these problems can be surmounted, this type of model could have numerous applications. For example, this problem can be viewed as a type of random growth model, explored in many publications, such as [Witten 1981]. Many manifestations of solid state physics deal with the growth of a specific characteristic in a material over time, such as with the *Ising model* [Griffiths 1969]. In mathematical biology, the modeling of tumor growth is also done according to random growth models. The standard approach to such problem involves the use of partial differential equations. However, an alternate approach, based upon the theory of random sets and an appropriate limit, might yield some new insights.

The theory of random sets was rigorously laid out by [Matheron 1975]. Since then, several works have presented examples of set-valued random processes, such as in [Harris 1976] and in [Papageorgiou 1989]. Physical applications of random sets have been explored, for example in [Koukiou 1990].

However, it was only until very recently that time-dependent random sets have been explored [Capasso and Villa 2006]. Perhaps because of this fact, no physical applications of dynamic random sets have been investigated. Time-dependent sets are often seen in very simple applications. One example is modeling the region of a capacitor plate that is covered with charge—a time-dependent set. However, there has been little-to-no exploration of time-dependent random sets. A set-based approach, rather than the conventional function-based approach, might yield some fruitful results.

Another important area in which such a theory might be applied is information theory. Information theory deals with the quantification of data and measures the rate at which such data may be communicated from a transmitter to a receptor. When such information is transmitted through a noisy channel, each piece of information has the potential to be lost. The signal that is received by a receptor can then be characterized by a random set. Over time, if the signal is repeated, the receptor will receive more and more of the actual message, so that over time the complete message might be ascertained by the receptor. Again, the amount received at any time can then be characterized as a random set.

A continuous-time model of such a random set could be very useful as a limiting process. This paper will solve this problem by developing a distributional view of this limit. This process of defining a distributional set is exactly analogous to the limit which yields a delta function in a distributional sense. Based upon this view, which is applicable to sequences of both deterministic and random sets, a rigorous mathematical model of a dynamic random set will be defined and explored.

4.2 Weak Convergence of Sequences of Sets.

For purposes of this construction the random sets under consideration will be closed subsets of a fixed compact subset of the real numbers $\mathbb{R}$. However, the idea and results described below can be easily generalized to any compact subset of $\mathbb{R}^n$.

In order to build a different concept of convergence, we will make an analogy to generalized functions, for which the notions of a weak convergence were first explored. A delta function, for example, can be seen as a linear functional which is the weak limit of a sequence of classical functions f_n. For our weak limit, we must also start with a concept of an inner product. Weak convergence in the ordinary context only requires that the real-valued sequence of inner product converges, that is,

$$\langle f_n, g \rangle \to \langle f, g \rangle \tag{4.1}$$

for all suitable *test* functions g.

We therefore start by defining an "inner product" between two measurable sets.

Definition 4.1 *Given two measurable sets A and B, the* pseudo-inner product $\langle A, B \rangle$ *is given by the rule*

$$\langle A, B \rangle = \mu(A \cap B), \tag{4.2}$$

where μ is Lebesgue measure.

Remark 4.1: This pseudo-inner product can also be expressed as

$$\langle A, B \rangle = \int_B \chi_A \, dx, \tag{4.3}$$

where χ_A is the characteristic function of A.

Definition 4.2 *Given a measurable set A, the* norm of A *is*

$$||A|| = \sqrt{\mu(A)}. \tag{4.4}$$

Lemma 4.1 *For the pseudo-inner product of (4.2) and norm (4.4), the following properties are true for any measurable sets* A*,* B*,* C*:*

$$i)\quad \|A\| \geq 0, \tag{4.5}$$

$$ii)\quad \|A\| = 0 \Longleftrightarrow A = \emptyset \;\; a.e., \tag{4.6}$$

$$iii)\quad \langle A, B\rangle = \langle B, A\rangle, \tag{4.7}$$

$$iv)\quad \|A \cup B\| \leq \|A\| + \|B\|, \tag{4.8}$$

$$v)\quad \langle A, B\rangle \leq \|A\|\|B\|. \tag{4.9}$$

Proof. Clear.

Remark 4.2: In particular, the triangle inequality property (iv) directly follows from the subadditivity property of measures. Also, the Schwarz inequality (v) is a straightforward application of Hölder's inequality to the characteristic function of the sets A and B. Two measurable sets are *orthogonal* if they are disjoint up to a set of measure zero.

Remark 4.3: This pseudo-inner product is *sublinear*:

$$\langle A \cup B, C\rangle \leq \langle A, C\rangle + \langle B, C\rangle, \tag{4.10}$$

with equality only if the three sets have an intersection of measure zero. Furthermore, a concept of scalar multiplication is missing. The lack of these properties prevent the set of measurable sets from being considered a Hilbert space. Nevertheless, this pseudo-inner product provides an example of a *sublinear functional* on the set of measurable sets.

Definition 4.3 *A* sublinear functional *is a function $f^* : \Sigma \to [0, \infty)$, where Σ is the set of sets of finite measure, such that for any two sets E_1 and E_2 in Σ,*

$$f^*(E_1 \cup E_2) \leq f^*(E_1) + f^*(E_2). \tag{4.11}$$

For example, for any measurable set A, we can define the associated sublinear functional f_A such that

$$f_A(B) = \langle A, B \rangle = \int_B \chi_A \, dx \tag{4.12}$$

for any set B of bounded measure.

In this example, (4.12) suggests a way of defining a functional without regard to a measurable set. For any integrable function f, we can define a sublinear functional $\hat{f}$ such that

$$\hat{f}(B) = \int_B f \, dx. \tag{4.13}$$

This functional is sublinear for the same reason as the pseudo-inner product (4.12) is sublinear. Now we consider a concept of a weak convergence with respect to this pseudo-inner product.

Definition 4.4 *A sequence $\{A_n\}$ of measurable subsets is said to converge weakly to a sublinear functional f^* if*

$$\langle A_n, B \rangle \to f^*(B), \tag{4.14}$$

for every set B of finite measure.

The set of sets of finite measure then takes the place of the set of test functions in this analog of conventional distributions. It is also evident that if a sequence of measurable sets $\{A_n\}$ converges in the normal sense, i.e. the limit infimum equals the limit supremum, then the sequence also weakly converges, by the Lebesgue dominated convergence theorem.

This weak limit makes sense in the context of sequences of both the deterministic and the random sets to be defined in Section 4.3.

Example 4.2: Consider the following sequence of subsets of [0,1]. Let $A_1 = [0, 1/2]$, A_2=$[0, 1/4] \cup [1/2, 3/4]$, $A_3 = [0, 1/8] \cup [1/4, 3/8] \cup [1/2, 5/8] \cup [3/4, 7/8]$, and so forth. The measure of each set in the sequence is 1/2. Also, a limit of the sequence does not exist, since the limit infimum of the sequence is a subset of the rationals, while the limit supremum of the sequence is the interval $[0, 1]$. The sequence fails to converge in the Hausdorff sense as well.

However, it is evident that

$$\langle A_n, B \rangle \longrightarrow 1/2\mu(B \cap [0, 1]) \tag{4.15}$$

for every measurable subset B. Therefore, the sequence $\{A_n\}$ weakly converges to the sublinear functional f^* as in (4.12), where

$$f^*(B) = 1/2 \int_B \chi_{[0,1]} \, dx = \mu(B \cap [0, 1])/2 \tag{4.16}$$

for every measurable set B, even though it does not converge in any usual sense.

This construction may be generalized as follows.

Lemma 4.2 *Given any step function of the form $u_a = a\chi_E$ with a in $[0, 1]$, E a* *bounded measurable set, there exists a weakly convergent sequence of measurable sets $\{A_n\}$ with limit u^* such that for any set B of finite measure,*

$$u^*(B) = \int_B u_a \, dx. \tag{4.17}$$

Proof. We make the proof by explicit construction. First, given a set E of nonzero but finite measure we define a continuous nondecreasing function

$$m(x) = \frac{\mu([-\infty, x] \cap E)}{\mu(E)}. \tag{4.18}$$

For any $0 \leq c \leq 1$ let $m^{-1}(c)$ denote the least value x for which $m(x) = c$. Define the sequence

$$\begin{array}{rcl} A_1 & = & E \bigcap [-\infty, m^{-1}(a)], \\ A_2 & = & E \bigcap \big([-\infty, m^{-1}(a/2)] \cup [m^{-1}(1/2), m^{-1}(1/2 + a/2)]\big), \\ \vdots & & \\ A_n & = & E \bigcap \big(\bigcup_{i=0}^{n-1} [m^{-1}(i/n), m^{-1}(i/n + a/n)]\big), \\ & & \vdots \\ & & . \end{array}$$

Every term in the sequence has measure $a\mu(E)$. Furthermore, it is clear that for any measurable set B,

$$\langle A_n, B \rangle \longrightarrow f^*(B) = \int_B a\chi_E \, dx, \tag{4.19}$$

completing the proof.

Examining the density function

$$\rho_B(x) = \lim_{h \to 0} \frac{\mu(B \cap [x-h, x+h])}{2h} \tag{4.20}$$

sheds further light on weak convergence of a sequence of sets. It is a result of analysis that one can use the density function to determine the measure of the set by the rule

$$\mu(B) = \int_B \rho_B(x)\, dx, \tag{4.21}$$

which is directly analogous to (4.12) above. For these reasons, the weak limit can be viewed as generating a generalized density function.

Remark 4.4: The relation (4.21) of course holds for characteristic functions as well, that is,

$$\mu(B) = \int_B \chi_B(x)\, dx. \tag{4.22}$$

However, it is more natural to employ the density function to realize weak limits because characteristic functions can only take the values 0 or 1, whereas a density function can take any intermediate value, albeit only on a set of measure zero. Therefore, the density function is closer in spirit to the idea of the weak limit, as illustrated in Example 4.2 above.

4.2.1 Some Functional Analytic Results on Sequences of Sets

We now extend several familiar results from functional analysis to the context of sequences of sets. First we associate a function with the weak limit of a weakly convergent sequence of sets.

Theorem 4.1 *For every sequence of measurable sets $\{A_n\}$ that weakly converges to a sublinear functional f^*, there exists a measurable function $\rho^* : \mathbb{R} \to [0,1]$ (a* generalized density function*) such that*

$$f^*(B) = \int_B \rho^*(x)\,dx \tag{4.23}$$

for every set B of finite measure.

Proof. Suppose that a sequence of sets $\{A_n\}$ weakly converges to the sublinear functional f^*. Since the terms in the corresponding sequence of characteristic functions $\{\chi_{A_n}\}$ are bounded in the L^∞ sense, they possess a weak* convergent subsequence, which we will denote $\{\chi_{A'_n}\}$. Denote its weak* limit by ρ^*. Then for every set B of finite measure,

$$\int_B \rho^* dx = \int \rho^* \chi_B dx \tag{4.24}$$

$$= \lim_{n\to\infty} \int \chi_{A'_n}\chi_B dx = \lim_{n\to\infty} \langle A'_n, B\rangle \tag{4.25}$$

$$= \lim_{n\to\infty} \langle A_n, B\rangle = f^*(B). \tag{4.26}$$

Corollary 4.1 *Let $\{A_n\}$ be a sequence of measurable sets. Then $\{A_n\}$ contains a weakly convergent subsequence.*

Proof. As seen in (4.24)–(4.26), the weak* convergence of the $\{\chi_{A'_n}\}$ to ρ^* implies the weak convergence of $\{A'_n\}$ to a sublinear functional with density ρ^*

Next we prove that the set of all possible measurable density functions, whose range is contained in $[0,1]$, occupies a large portion of the space of functions, the range of which is contained in $[0,1]$.

Theorem 4.2 *Given a compact set C, let M be the set of generalized density functions ρ^* guaranteed by Theorem 4.1 as the representation of the weak limit f^* of* *a sequence of sets $\{A_n\}$, with all $A_n \subset C$. Then M is dense with respect to the L^1 norm in the set of absolutely integrable functions with support C and with range contained in* $[0, 1]$.

Proof. Without loss of generality, we may assume $C = [0, 1]$. Recall that from Lemma 4.2, that any constant real-valued function u_a with support E and range in $[0, 1]$ has a corresponding sequence of measurable sets which weakly converge to a sublinear functional with density u_a. Combining such a construction using n disjoint sets E_i of finite measure, one can construct a simple function $s(x)$ given by

$$s(x) = \sum_{i=1}^{n} a_i \chi_{E_i}, \tag{4.27}$$

where $0 \leq a_i \leq 1$, and where $s(x)$ is the representation of the sublinear functional of the weak limit of a sequence of sets. But any integrable function g supported on C with range in $[0, 1]$ is clearly well approximated by simple functions of the type (4.29).

4.3 A Distributional Weak Limit of Random Sets

Let us next carry over the idea of weak convergence of sets to the regime of random sets. A random set is a random variable X whose values lie in the family of measurable subsets of $\mathbb{R}$. More formally, we state the following definition.

Definition 4.5 *Let $\mathcal{F}$ be a family of measurable sets of $\mathbb{R}^n$. Let $\Sigma(\mathcal{F})$ be a sigma algebra on $\mathcal{F}$ and π be a probability measure on $\Sigma(\mathcal{F})$, giving that $(\mathcal{F}, \Sigma(\mathcal{F}), \pi)$ is a probability space. A* random measurable set *(or simply* random set*) X of $\mathbb{R}^n$ is the identity function $X : \mathcal{F} \to \mathcal{F}$.*

Example 4.5: Consider the collection $N = \{\{1\}, \{2\}, \ldots, \{m\}\}$ of singletons of the first m natural numbers, each corresponding to a decision d_i that can be made in learning. Assign to each singleton (decision) $\{i\}$ its probability p_i of being learned. Let $\mathcal{F}$ equal the collection of measurable sets of $\mathbb{R}$ and let $\Sigma(\mathcal{F})$ equal the smallest sigma algebra containing every element of $\mathcal{F}$. For any set S of $\Sigma(\mathcal{F})$, define the probability measure π by the rule

$$\pi(S) = \sum_{\{i\} \in N \cap S} p_i.$$

Our random set X is then simply the identity map on $\mathcal{F}$.

Example 4.6: Let p be a measurable probability density on $\mathbb{R}$. Let $\mathcal{F}$ equal the collection of measurable sets of $\mathbb{R}$ and let $\Sigma(\mathcal{F})$ equal the smallest sigma algebra containing every element of $\mathcal{F}$. We define a function M such that for any collection of measurable sets S, $M(S)$ equals the minimal set of real numbers which has all of the singleton members of S as a subset. We define $\pi(S) = \int_{M(S)} p(x)\, dx$. Then the identity map $X : \mathcal{F} \longrightarrow \mathcal{F}$ is a random set of $\mathbb{R}$.

Example 4.7: Let p be a measurable probability density on $\mathbb{R}$ and let b be a fixed positive real number. Let $\mathcal{F}$ equal the collection of measurable sets of $\mathbb{R}$. For any collection S of sets in $\mathcal{F}$, define a function M such that $M(S)$ equals the set of left endpoints of those members of S which are closed intervals of length b. Let $\Sigma(\mathcal{F})$ equal the collection of all collections S of sets such that $M(S)$ is Lebesgue measurable. We define $\pi(S) = \int_{M(S)} p(x)\, dx$. Then the identity map $X : \mathcal{F} \longrightarrow \mathcal{F}$ is a random set of $\mathbb{R}$.

Example 4.8: Let p be a measurable probability density on $\mathbb{R}^2$. Let $\mathcal{F}$ equal the collection of all measurable set of $\mathbb{R}$. For any collection S of sets in $\mathcal{F}$, define a function M such that $M(S)$ equals the set of ordered pairs (x, y) where x is the left endpoint of any member of S that is a closed interval with length y. Let $\Sigma(\mathcal{F})$ equal the collection of all collections S of sets such that $M(S)$ is Lebesgue measurable. Define $\pi(S) = \int\int_{M(S)} p(x, y)\, dy\, dx$. Then the identity map $X : \mathcal{F} \longrightarrow \mathcal{F}$ is a random set of $\mathbb{R}$.

Example 4.9: Let p be a measurable probability density on $\mathbb{R}^{n+1}$. Let $\mathcal{F}$ equal the collection of all measurable sets of $\mathbb{R}^n$ and let $\Sigma(\mathcal{F})$ equal the smallest sigma algebra containing every element of $\mathcal{F}$. We define a function M such that for any S of $\Sigma(\mathcal{F})$, $M(S)$ equals the set of points in $\mathbb{R}^{n+1}$ where the first n coordinates locate the center of an n-ball in S and where the last coordinate is the radius of that n-ball. Define $\pi(S) = \int \ldots \int_{M(S)} p(x_1, \ldots, x_{n+1})\, dx_{n+1} \ldots dx_1$. Then the identity map $X : \mathcal{F} \longrightarrow \mathcal{F}$ is a random set of $\mathbb{R}^n$.

Example 4.10: Let p be a measurable probability density on $\mathbb{R}$. Let $\mathcal{F}$ equal the collection of all measurable sets of $\mathbb{R}$ and let $\Sigma(\mathcal{F})$ equal the smallest sigma algebra containing every element of $\mathcal{F}$. For any S of $\Sigma(\mathcal{F})$ we define S_1 to equal the set of all members of S which are the union of three closed intervals, each of length 1/3. For each such member of S_1 we associate a point $s_1 = (x_1, x_2, x_3)$ in $\mathbb{R}^3$ where the coordinates are the left endpoints of these three intervals, with $x_1 \leq x_2 \leq x_3$. We define a function M such that $M(S)$ equals the set of all such points s_1 associated with S in $\mathbb{R}^3$. Define $\pi(S) = \int\int\int_{M(S)} p(x_1)p(x_2)p(x_3)\, dx_3 dx_2 dx_1$. Then the identity map $X : \mathcal{F} \longrightarrow \mathcal{F}$ is a random set of $\mathbb{R}^n$.

Henceforth we will adopt a standard abuse of notation: X will denote both the random set itself as well as the map. The context will indicate which is meant. Now we introduce the fundamental idea of a *capacity functional.*

Definition 4.6 *Given a random set X, the* capacity functional $T_X(x)$ *is the probability that the point x is contained in the random set X. That is, under the notation of Definition 4.5,*

$$T_X(x) = \pi(X \in \mathcal{F}; x \in X) = \int_{\mathcal{F}} \chi_X(x)\, d\pi(X). \tag{4.28}$$

The capacity functional is a basic tool for characterizing random sets. For example, it can be used to calculate the expected value of the measure of any random set. To do this, it is important to realize that the measure of a random set is itself a real-valued random variable [Molchanov 2005]. We now quote a fundamental result due to Robbins published in 1944.

Theorem 4.3 (Robbins's Theorem) *Let X be a random set in $\mathbb{R}^n$ and μ the Lebesgue measure on $\mathbb{R}^n$. Then $\mu(X)$ is a random variable with*

$$E(\mu(X)) = \int_{\mathbb{R}^n} T_X(x)\, d\mu(x). \tag{4.29}$$

Proof sketch. By Fubini's Theorem and (4.30),

$$\begin{aligned} E(\mu(X)) &= \int_{\mathcal{F}} \mu(X)\, d\pi(X) = \int_{\mathcal{F}} \int_{\mathbb{R}^n} \chi_X(x)\, d\mu(x) d\pi(X) && (4.30) \\ &= \int_{\mathbb{R}^n} \int_{\mathcal{F}} \chi_X(x)\, d\pi(X) d\mu(x) = \int_{\mathbb{R}^n} T_X(x)\, d\mu(x). && (4.31) \end{aligned}$$

Corollary 4.2 *For a random closed set in $\mathbb{R}^n$ and μ the Lebesgue measure, the* *second moment of $\mu(X)$ is given by*

$$E(\mu^2(X)) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} T_X(y) T_X(x) \, d\mu(y) d\mu(x). \tag{4.32}$$

Higher order moments are given analogously.

Given a measurable set A and random set X we can define an "intersection" random set, $X \cap A$.

Definition 4.7 *Let $\mathcal{F}_A$ be the collection of measurable subsets of A. For each σ in $\Sigma(\mathcal{F})$, let*

$$\sigma_A = \{f_A \in \mathcal{F}_A; \exists f \in \sigma, f_A = \{y \cap A; y \in f\}\}.$$

Let $\Sigma_A(\mathcal{F}_A)$ equal the set of all such σ_A. Then $\Sigma_A(\mathcal{F}_A)$ is a sigma algebra given for the intersection random set. The probability measure π_A of S_A is now given by the rule $\pi_A(\sigma_A) = \pi(\sigma)$. The intersection random set will be the identity map $X : \mathcal{F}_A \to \mathcal{F}_A$.

Given this definition, the following is evident.

Corollary 4.3 *For a random set X of $\mathbb{R}^n$ and measurable set A in $\mathbb{R}^n$, the expected value of the measure of the intersection of X with A is given by*

$$E(\mu(X \cap A)) = \int_A T_X(x) \, d\mu(x). \tag{4.33}$$

The expected value of the measure of a random set is therefore the same in form as the integral formula for the inner product in (4.21), with the generalized density function being replaced by a probability. In other words, the deterministic case given in (4.21) is the same as (4.32), with the points contained in the deterministic set having probability one. Therefore, the integral formulation (4.32) for a weak limit of a sequence of random sets closely matches what has been previously presented. Thus Robbin's Theorem motivates the exploration of the weak limit of a sequence of random sets. First, we extend the definition from Section 4.2 of a sublinear product to an inner product between a random set and a set of finite measure.

Definition 4.8 *The* sublinear product *between a random set X and a set B of finite measure is defined to be*

$$\langle X, B\rangle = E(\mu(X \cap B)) = \int_B T_X(x)\, d\mu(x). \tag{4.34}$$

Next, we define weak convergence of a sequence of random sets.

Definition 4.9 *A sequence $S = \{X_n\}$ of random sets* weakly converges *to a sublinear functional X^* if for every set B of finite measure we have $\langle X_n, B\rangle \to X^*(B)$.*

Corollary 4.4 *For every weakly convergent sequence $S = \{X_n\}$ of random sets converging to the sublinear functional X^* there exists a measurable function T_S : $\mathbb{R} \to [0,1]$ such that*

$$X^*(B) = \int_B T_S(x)\, d\mu(x). \tag{4.35}$$

Proof. The weak* convergence argument from the proof of Theorem 4.1 gives that for the sequence of corresponding capacity functionals T_{X_n}, there exists a bounded measurable function $T_S(x)$ of modulus at most 1 such that

$$X^*(B) = \lim_{n\to\infty} \langle X_n, B\rangle = \lim_{n\to\infty} \int_B T_{X_n}(x)\, d\mu(x) = \int_B T_S(x)\, d\mu(x). \tag{4.36}$$

Remark 4.5: Thus, the weak limit of (4.37) gives rise to the concept of *generalized random sets* X that are determined by their corresponding capacity functional T_S.

As will be seen in an example below, it is possible to construct a generalized random set which is not itself a random set.

Finally, a *dynamic random set* will be defined as follows.

Definition 4.10 *A* dynamic random set *is a triple* $(S_0, D, \mathbb{S})$*, where* $S_0 = \{X_n\}$ *is a fixed weakly convergent sequence of random sets,* $\mathbb{S}$ *is a collection of weakly convergent sequences of random sets that contains the sequence* S_0*, and where* $D(t)$ *is a one-parameter family of operators on* $\mathbb{S}$*, where* t *is a non-negative real number, such that for any element of* S *in* $\mathbb{S}$ *with the limiting capacity functional* T_S*,*

i. $D(0)S = S$*,*

ii. $S = D(t)S_0$ *for some* $t \in [0, \infty)$*,*

iii. $D(t+s)S = D(t)D(s)S$*, and*

iv. $D(t)$ *is continuous at* $t = 0$*, that is*

$$\lim_{t \to 0} \int_B |T_{D(t)S}(x) - T_S(x)| \, d\mu(x) = 0 \tag{4.37}$$

for every set B *of finite measure.*

The parameter t will be interpreted later as time. This definition is directly parallel to that of a dynamical system or a semigroup. As in dynamical systems, for every dynamic random set and value t, there exists a weakly convergent sequence $S(t)$, hence its limit $X^*(t)$.

We now define what will be shown to be an example of a Dynamic Random Set, and which will be applied to learning.

Definition 4.11 (The Principal Dynamic Random Set) *Let f be a differentiable probability density on $\mathbb{R}$ with compact support. Let $\mathcal{F}$ equal the collection of all measurable sets of $\mathbb{R}$ and let $\Sigma(\mathcal{F})$ equal the smallest sigma algebra containing every element of $\mathcal{F}$. Fix n. Then for any $S^{(n)}$ of $\Sigma(\mathcal{F})$ we define $S_1^{(n)}(t)$ to equal the set of all members of $S^{(n)}$ which are the union of n closed intervals, each of length $\lambda t/n$, with λ a fixed positive number and t a nonnegative real parameter. For each such member of $S_1^{(n)}(t)$ we associate a point $s_t^{(n)} = (x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$ where the coordinates are the left endpoints of these n intervals which are ordered so that $x_1 \leq x_2 \leq \cdots \leq x_n$. We define a function M_t such that $M_t(S^{(n)})$ equals the set of all points $s_t^{(n)}$ in $\mathbb{R}^n$. Define*

$$\pi^{(n)}(S^{(n)}) = \int \ldots \int_{M_t(S^{(n)})} \Big(\prod_{i=1}^{n} f(x_i)\Big)\, dx_n \ldots dx_1. \tag{4.38}$$

Then the identity map $X_n(t) : \mathcal{F} \longrightarrow \mathcal{F}$ is a random set of reals with respect to the probability measure $\pi_t^{(n)}$. Let S_0 be the sequence of random sets $X_n(t)$, with $t = 0$, and where $D(t')(S_0)$ is the corresponding sequence with $t = t'$. In general, $D(t')(\{X_n(t)\}) = \{X_n(t + t')\}$. Finally, let $\mathbb{S}$ denote all $D(t)S_0$.

Theorem 4.4 *Let $(S_0, D(t), \mathbb{S})$ be defined as Definition 4.10. Then for any $t \geq 0$, the sequence $D(t)S_0 = S(t)$ will weakly converge to the limit X_t^*, where this limit has the property that for any set of finite measure B,*

$$X_t^*(B) = \int_B (1 - e^{-\lambda t f(x)}) dx. \tag{4.39}$$

Furthermore,

$$\lim_{t \to 0} \int_B |T_{D(t)S}(x) - T_S(x)|\, d\mu(x) = 0 \tag{4.40}$$

and therefore the triple $(S_0, D(t), \mathbb{S})$ forms a dynamic random set.

Proof. We need to show the existence of the sublinear functional X_t^* such that for any measurable set B

$$X_t^*(B) = \lim_{n\to\infty} \int_B T_{X_{n(t)}}(x) d\mu(x). \tag{4.41}$$

In order to show this existence, we examine the limiting behavior of the hitting probability for the given sequence of random sets. Fix x and t. Recall that $X_n(t)$ takes only values which are the union of n closed intervals which are each of length $\lambda t/n$. For each $X_n(t)$ in the sequence, the probability of being in an interval of the type comprising $X_n(t)$ is equal to

$$\int_{x-\lambda t/n}^{x} f(x')d\mu(x'). \tag{4.42}$$

The probability of *not* being in the union of n such intervals is therefore given by

$$\Big(1 - \int_{x-\lambda t/n}^{x} f(x')d\mu(x')\Big)^n, \tag{4.43}$$

and so

$$T_{X_{n(t)}}(x) = 1 - \Big(1 - \int_{x-\lambda t/n}^{x} f(x')d\mu(x')\Big)^n. \tag{4.44}$$

For large n, the differentiability and the compact support of f ensures that

$$\int_{x-\lambda t/n}^{x} f(x')d\mu(x') = \frac{\lambda t f(x)}{n} + O(1/n^2). \tag{4.45}$$

An expression of the form

$$\Big(1 - \frac{a}{n} + O(1/n^2)\Big)^n \tag{4.46}$$

converges pointwise to e^{a}. Therefore, from (4.44)-(4.46) we have

$$\lim_{n\to\infty} T_{X_{n(t)}}(x) = 1 - e^{-\lambda t f(x)}. \tag{4.47}$$

By our definition of the measure of the limit of a weakly convergent sequence of random sets and the use of the Lebesgue Dominated Convergence Theorem, we obtain a sublinear functional which represents the limiting measure of the sequence such that for any measurable set B,

$$X_t^*(B) = \int_B (1 - e^{-\lambda t f(x)})dx. \tag{4.48}$$

Hence, at $t = 0$, we have a capacity function equal to 0 for any singleton. In contrast, for $\lambda > 0$, the capacity functional for any given point approaches 1 for large t.

To show that the Principal Dynamic Random Set is indeed a Dynamic Random Set, we may observe that Properties (i)-(iii) for a Random Dynamic Set are apparent. Property (iv) follows from (4.48). To see this we examine the limit

$$\lim_{t'\to 0} \int_B |T_{D(t')S}(x) - T_S(x)|\, dx. \tag{4.49}$$

In the present case, we obtain that this limit is equal to

$$\lim_{t'\to 0} \int_B |(1 - e^{\lambda(t+t')f(x)}) - (1 - e^{\lambda t f(x)})|\, dx = 0. \tag{4.50}$$

Therefore, the Principal Dynamic Random Set is indeed a Dynamic Random Set.

Remark 4.6: We can now observe that the sequence X_n does not strongly converge to any random set $X(t)$ for a fixed nonzero value of λt. To see this, suppose some such random set exists, $X(t)$. For any interval E which contains a set of positive measure on which f is nonzero the hitting probability $T_{X(t)}(E)$ is 1. However, consider a sequence of intervals E_n which converge to a singleton E_∞ which is in the support of f. The limit of the hitting probabilities will be 1, while the hitting probability of the limiting set will be $1-e^{-\lambda t f(x)}$. This contradicts the result of Choquet's Theorem (Matheron 1975) which requires the upper semicontinuity property

$$\lim_{n\to\infty} T(E_n) \leq T(\lim_{n\to\infty} E_n) \tag{4.51}$$

for every random set. It is unknown whether the sequence weakly converges to a random set.

4.4 Properties of the Principal Dynamic Random Set

Given a set B which is the minimal compact support of $f(x)$, define $C_f(t)$ as

$$C_f(t) = X_t^*(B) = \int_B \left(1 - e^{-\lambda t f(x)}\right) dx. \tag{4.52}$$

Theorem 4.5 *$C_f(t)$ has the following properties:*

i. $\frac{dC_f(t)}{dt}|_{t=0} = \lambda\mu(B)$*;*

ii. $\frac{dC_f(t)}{dt} > 0$ *for all t; and*

iii. $\frac{d^2C_f(t)}{dt^2}$ *is negative for all t.*

Proof. Clear.

Remark 4.7: Of course λ need not be a constant, but could instead depend of position or even time. If however $\lambda = \lambda(x)$, the dependence would be indistinguishable from a position dependence of the probability distribution. Therefore, one could redefine the probability distribution to incorporate this dependence, and leave λ as a constant. On the other hand, if λ depends on time, the derivative properties from Theorem 4.5 might no longer hold. Therefore, we will assume λ to be a constant.

Another very important property is that $C_f(t)$ is maximized when $f(x)$ is uniform. This result is intuitively plausible since a uniform probability minimizes the chances of overlap over time.

Theorem 4.6 (Maximizing Distribution) *If f is a probability distribution density with compact support B, a subset of $\mathbb{R}$, with λ not dependent on time, then the maximum value for $C_f(t)$, t is positive and fixed, is obtained when is f is constant with respect to x.*

Proof. To maximize C_f is to minimize

$$\int_B \exp(-\lambda t f)\, dx.$$

This result is a special case of the following. Define a functional J such that

$$J(f) = \int_B G(f(x))dx,$$

where G is non-increasing and f satisfies

$$\int_B f(x)dx = 1.$$

Then

$$J(f) \geq \int_B G(\sup f(x))dx,$$

and the lower bound is realized when

$$f(x) \equiv 1/\mu(B)$$

almost everywhere.

Remark 4.8: Given a principal random dynamic set $X_t(t)$, the complement $X_t(t)$ has a nice pseudo-semigroup property. Recall that for standard semigroup R we have the properties

$$\begin{aligned} &i) \quad R(0) = I \\ &ii) \quad R(t+s) = R(t)R(s). \end{aligned}$$

Let us define an operator $R(t)$ acting on measurable subsets E by the rule $R(t)E$ yields the subset $E \setminus X(t)$. It is clear from the properties of a Dynamic Random Set that property (i) is met. However, property (ii) must be understood in a different way than is standard for a semigroup. The second property can be rewritten as

$$R(t+s)E = (R(t)E) \cap (R(s)E). \tag{4.53}$$

However, this expression is at least correct with the expectation value of the measures. It is trivial to show that

$$\mu(R(t+s)E) = \mu\big((R(t)E) \cap (R(s)E)\big) \tag{4.54}$$

and therefore much of the content of the semigroup concept is seen in the properties of Dynamic Random Sets.

4.4.1 The Weak Limit as a Random Functional

A weak limit X^* of a sequence of random sets $\{X_n\}$ (see Definition 4.8) can alternatively be thought of as a random functional that acts upon measurable sets to produce a real-valued random variable. To obtain this result, we will first define a real-valued random variable arising from a single random set. Second, we will then derive a random functional from this real-valued random variable. Third, we will use a sequence of such random functionals to create a limit random functional.

Definition 4.12 *Let X be a random set of the triple $(\mathcal{F}, \Sigma(\mathcal{F}), \pi)$ with $\mathcal{F}$ equal to the set of all measurable sets of $\mathbb{R}$, $\Sigma(\mathcal{F})$ a sigma algebra on $\mathcal{F}$, and π a probability measure on $\Sigma(\mathcal{F})$. Further suppose that, for any measurable set $\mathcal{M}$ of non-negative real numbers, the set D of all sets which have measure m for some $m \in \mathcal{M}$ is in $\Sigma(\mathcal{F})$. Consider the probability space $(\Omega, \Sigma(\Omega), \pi_X)$ where*

i. Ω is the set of real numbers,

ii. $\Sigma(\Omega)$ is the sigma algebra of all Lebesgue measurable subsets of $\mathbb{R}$, and

iii. for any measurable set $S \in \mathcal{F}$, the probability measure π_X of S equals $\pi(D)$, where D is the set of all sets with measure m as m ranges over all of S.

Then the identity map $X^\dagger : \Omega \to \Omega$ is a real-valued random variable of the triple $(\Omega, \Sigma(\Omega), \pi_X)$. We will call this random variable $X^\dagger$ the measure random variable associated with *X.*

We next define a functional with a domain consisting of all measurable subsets of $\mathbb{R}$ and range equal to a subset of the collection of all real-valued random variables on $\mathbb{R}$.

Definition 4.13 *Given a random set X suppose that for any measurable subset B of $\mathbb{R}$, there exists a measure random variable $(X \cap B)^\dagger$ for the intersection random set $X \cap B$ (see Definition 4.7). Then we define the* measure random functional f_X *to be a function with a domain consisting of all measurable subsets of $\mathbb{R}$ and range equal to a subset of the set of real-valued random variables, such that for any measurable set B,*

$$f_X(B) = (X \cap B)^\dagger. \tag{4.55}$$

Thus this random functional f_X operates on a measurable set B such that $f_X(B)$ is a real-valued random variable.

Example 4.11: Choose the random set X to be that given in Example 4.7, where $p(x)$ is a uniform distribution on $[0, 1]$, and $b = 1/2$. Then we have a random variable $f_X([0, 1/2])$ with cumulative distribution function

$$C(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} + x & 0 \leq x \leq \frac{1}{2} \\ 1 & x > \frac{1}{2} \end{cases}. \tag{4.56}$$

In general, the relationship between the random set X, the intersection random set $X \cap B$, and the linear functional is captured by the commutative diagram of Figure 4.1.

$$\begin{tikzcd}
X \arrow[r, "\text{intersection}"] \arrow[d, "\text{functional}"'] & X \cap B \arrow[d, "\text{functional}"] \\
f_X \arrow[r, "\text{action of functional}"] & f_X(B) = (X \cap B)^\dagger
\end{tikzcd}$$

Fig. 4.1. The commutative diagram for intersection random sets and the measure functional.

Now we define the limit random functional $X_\infty^\dagger$.

Definition 4.14 *Assume we are given a sequence of random sets $\{X_n\}$ with a corresponding sequence of measure random functionals $\{f_{X_n}\}$ such that the sequence of real-valued random variables $\{f_{X_n}(B)\}$ converges in probability for every measurable set B. Then we define the limit random functional $X_\infty^\dagger$ so that for every measurable set B,*

$$X_\infty^\dagger(B) = \lim_{n\to\infty} f_{X_n}(B). \tag{4.57}$$

Since expected values pass in the limit for convergence in probability, the expected value of $X_\infty^\dagger(B)$ then will be given by the sublinear product $X^*(B)$ given above in (4.35):

$$E(X_\infty^\dagger(B)) = X^*(B) = \int_B \lim_{n\to\infty} T_{X_n}(x) d\mu(x), \tag{4.58}$$

again directly analogous to the expected value given in Robbin's Theorem. Higher order moments will also be given with analogous formulas. If $P(x \in X \text{ and } y \in y)$ is the probability that x and y are both in random set X, then we obtain for the second moment

$$E((X_\infty^\dagger(B))^2) = \int_B \int_B \lim_{n\to\infty} \lim_{m\to\infty} P(x \in X \text{ and } y \in X) d\mu(x) d\mu(y). \tag{4.59}$$

Finally, the variance of $X_\infty(B)$ will be given by

$$\sigma^2(X^\dagger_\infty(B)) = (E(X^\dagger_\infty(B)))^2 - E((X^\dagger_\infty(B))^2) \tag{4.60}$$

In the context of dynamic random sets, we can similarly define a one-parameter set of random functionals $X^\dagger_\infty(t)$. When we examine the random functional associated with the Principal Dynamic Random Set given in Definition 4.10, one discerns a very important property.

Theorem 4.7 *Let $X^\dagger_\infty(t)$ be the random functional associated with a Principal Dynamic Random Set of Definition 4.11 with f and a measurable set B that is the minimal compact support of f, the variance of $X^\dagger_\infty(t)(B)$ is zero.*

Proof. Using the formulas (4.58)–(4.59) for the first and second moment of $X^\dagger_\infty(t)(B)$ we obtain

$$\begin{aligned}
E((X^\dagger_\infty(t)(B))^2) &= \int_B \int_B \lim_{m,n\to\infty} T_{X_n}(x) T_{X_m}(y) dxdy && (4.61)\\
&= \int_B \int_B (1 - e^{-\lambda t f(x)})(1 - e^{-\lambda t f(y)})\, dydx && (4.62)\\
&= \int_B (1 - e^{-\lambda t f(x)})\, dx \int_B (1 - e^{-\lambda t f(y)})\, dy && (4.63)\\
&= (E(X^\dagger_\infty(t)(B)))^2. && (4.64)
\end{aligned}$$

Therefore, the random variable $X^\dagger_\infty(t)(B)$ will have zero variance.

Since the random variable $X^\dagger_\infty(t)(B)$ is a limit of measures within of random sets, the result can be interpreted to mean that the measure of the principal random set has zero variance is therefore deterministic. As shown by Theorem 4.4, its expected value will be given by $C_f(t)$. Therefore, the measure of $X^\dagger_\infty$ is deterministic, even though the random sets in the sequence $X_n(t)$ do not have a deterministic measure.

Remark 4.9: Theorem 4.7 is a special case of the more general case that if the probabilities of the random set containing two distinct points x and y are uncorrelated, then the variance of its measure will be zero. For the Principal Dynamic Random Set, given λ and t, find an n_0 such that for all $n > n_0$ implies that $\lambda t/n < |x - y|$. Therefore, after the first n_0 terms in the sequence, x and y will never be in the same interval, but instead in uncorrelated intervals. This lack of correlation between the probability of containing two distinct points is an important motivation for the use of the principal dynamic random set to model learning.

4.5 Applications of dynamic random sets

Many of the possible applications to this problem can be seen in the field of statistical mechanics. However, the simplest application is from a beginning physics course.

4.5.1 Charging a Capacitor

Consider a constant source of voltage V, across which is connected a resistor of R Ohms in series with a parallel-plate capacitor of C farads. When this circuit is first connected, electrons flow from the battery onto one plate of the capacitor, giving that plate a negative charge. Electrostatic repulsion pushes electrons from the other plate into the other segment the circuit, giving the second plate a positive charge which is equal in magnitude to the charge on the first plate. Over time, the current i, which equals the product of the charge of an electron and the rate that the electrons are collecting on the first plate, decreases, and the magnitude of the charge on the plate asymptotically approaches a maximum value $Q_{max} = CV$.

The standard explanation for the eventual slowing down of charge flow is that there is effectively a smaller applied voltage across the resistor R in the circuit, since the charge accumulated on the plates creates a potential difference between the plates. Macroscopically, this is certainly true. However, the use of random sets affords us a microscopic view of what is actually happening at the capacitor plates.

Rather than a single current i we will consider two currents: an incoming current i_inc, and a current i_refl reflected back from the capacitor towards the electron source. The existence of two separate currents is merely heuristic and is not meant to represent two separate, distinct flows of electrons. The incoming current will depend only on the applied voltage V and the resistance R:

$$i_{inc} = \frac{V}{R}. \tag{4.65}$$

The macroscopic current i will be the difference between these two, since they are traveling in opposite directions:

$$i = i_{inc} - i_{refl}. \tag{4.66}$$

The incoming current i_{inc} will be constant. The reflected current i_{refl} on the other hand will not be constant. Reflected current will be created when charge attempts to go to a part of the plate where charge already resides. Since the capacitor will initially be assumed to be uncharged, i_{refl} will initially be zero. As the plate fills with electrons, the probability that electrons will attempt to travel to parts of the plate already occupied will increase, and i_{refl} will correspondingly increase.

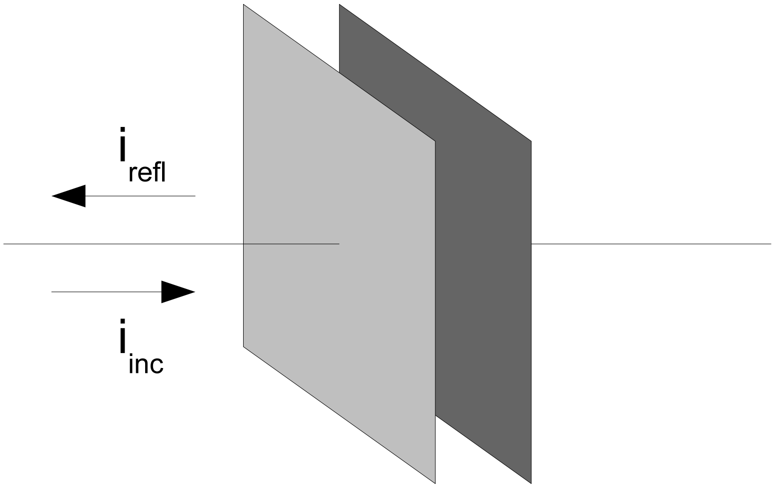

Figure 4.1: The currents involved in charging a capacitor.

To make this description more precise, we will use the Principal Dynamic Random Set given previously in Section 4.4. Electrons will be equally likely to occupy any area on the capacitor. The amount of space an electron will take up on the capacitor plate will equal the area of the plate divided by the number of electrons which can fit on the plate, or

$$A_{e_-} = A_{plate} \frac{q_{e_-}}{CV} \tag{4.67}$$

Since this number is vanishingly small in comparison to the size of the plates, this phenomenon can be modeled well by the principal dynamic random set, which looks at the limit of a large number of intervals covering a space. We will assume that $f(x)$, the probability for the random dynamic set to cover a point x, to be constant:

$$f(x) = \frac{1}{A_{plate}} \tag{4.68}$$

In the present case, the Principal Dynamic Random Set will be 2-dimensional rather than 1-dimensional, but the n-th term of the sequence of random sets will merely be formed by the union on n squares rather than the union of n intervals.

The parameter λ represents the coverage rate, the rate at which electrons cover the plate:

$$\lambda = \frac{i_{inc}}{Q_{max}} = \frac{i_{inc}}{CV} = \frac{1}{RC} \tag{4.69}$$

Given $f(x)$ and λ, the area covered by electrons can be found by (4.67)

$$\begin{aligned} A(t) &= \int_{A_{plate}} \left(1 - e^{-\lambda t f(x)}\right) dA & (4.70) \\ &= A_{plate}(1 - e^{-t/RC}), & (4.71) \end{aligned}$$

which yields the charge $Q(t)$ as a function of time to be

$$\begin{aligned} Q(t) &= \frac{A(t)}{q_{e-} A_{e-}} & (4.72) \\ &= CV(1 - e^{-t/RC}), & (4.73) \end{aligned}$$

which exactly matches the rate observed in the charging of a capacitor. The total current can be found by

$$i(t) = \frac{dQ(t)}{dt} = \frac{V}{R} e^{-t/RC}, \tag{4.74}$$

again exactly matching experiments.

Finally, the current reflected from the capacitor will be given by the expression

$$i_{refl}(t) = i_{inc} - i(t) = \frac{V}{R}(1 - e^{-t/RC}). \tag{4.75}$$

Therefore, in the the random set model, the charging of a capacitor plate slows not because less current is incoming at the plates, but because arriving charge becomes more and more likely to attempt to go to spaces already occupied by charge, causing a reflected current.

4.5.2 Growth Models

Several scientific problems involve the growth of a body to fill up a space. The best example is that of a tumor [Holash et al.]. A tumor starts as a very small body with one or very few cells. Over time, it can grow to macroscopic sizes, even hundreds of cubic centimeters. The typical cluster model involves discrete steps of adding parts on the cluster, with blood vessel proximity increasing the probability of additional tumor cells at a site.

This type of problem seems to be particularly well suited to dynamic random sets. A dynamic random set can explicitly deal with the problem of tumor growth in continuous time, rather than the discrete time as in the model in [Holash et al. 1999]. The fact that different places have different growth probabilities is naturally handled in the case of random dynamic sets. This potential application for random sets is a topic I certainly hope to explore in the future.

4.6 Creating a Continuous Model of the Learning Curve via Distributional Random Sets

4.6.1 A Deterministic Model of Learning a Continuum of Lessons

We will now use random sets to build a continuous model of learning. To this end we must overlay a cost structure over a space. Suppose that each lesson to be learned has been labeled by exactly one real number x so that the cost savings realized by learning the lessons labeled by points lying in a measurable set B is given by the Stieltjes integral

$$\Delta C(B) = \int_B dC(x), \tag{4.76}$$

where the *cumulative cost savings function* $C(x)$ is an everywhere defined, nondecreasing, right-continuous function that is continuously differentiable between isolated jumps.

Example 4.12: Suppose there are exactly n lessons to be learned, labeled by $x_i = 1, x_2 = 2, \ldots, x_n = n$. Each lesson $x_i = i$ yields a cost improvement of c_i. Then the total cost improvement obtained by learning the lessons in a set B is

$$\Delta C(B) = \int_B dC(x) = \sum_{i \in B} c_i. \tag{4.77}$$

Example 4.13: Suppose cumulative cost function C is everywhere continuously differentiable on the open set G. Then the cost savings contributed by an infinitesimal neighborhood of $x \in G$ is $dC(x) = C'(x)\,dx = c(x)\,dx$, giving that the cost savings for lessons in $B \subset G$ is

$$\Delta C(B) = \int_B dC(x) = \int_B c(x)\ dx = \int_G \chi_B(x)c(x)\ dx. \tag{4.78}$$

Remark 4.10: Points of differentiability of the cumulative cost function C should be thought of as *incremental improvements,* while jump discontinuities are *paradigm shifts.* For the incremental movements, learning is done incrementally, and improvement happens continuously over time. As discussed in Section 2.1, this type of learning is what March called "exploitation" [March 1991] and is considered "supervised learning" in the artificial intelligence context [Russell and Norvig 2002].

On the other hand, the paradigm shift corresponds to March's idea of the "exploration" aspect of learning, which is considered "unsupervised learning" in the artificial intelligence context. Learning under the category of exploration includes, for example, new technology, new worker insight to production, new cost structure, material replacement, etc. Rather than continuous improvement, this type of learning takes the form of sudden leaps forward. While continuous improvement may be modeled with a continuum of learning choices, explorative learning takes the form of a finite number of breakthroughs.

As was argued for the discrete case in Chapter 3, the probability $\mathrm{prob}(X)$ of choosing the lessons forming the measurable set X is a smooth function $p = p(y)$ of the cost savings $y = \Delta C(X)$ resulting from these lessons:

$$\mathrm{prob}(X) = p\bigg(\int_X\ dC(x)\bigg). \tag{4.79}$$

Further, from Chapter 3's use of the Boltzmann's H-theorem [Feynman 1972], it follows that $p(y) = \alpha \exp(\beta y)$, where the normalization α will be determined below.

In particular, if F is the *cumulative probability that all lessons indexed by x or less will be chosen,* then

$$F(x) \;=\; \alpha \exp\big(\beta \int_{-\infty}^{x} dC(y)\big) \;=\; \alpha \exp(\beta C(x)\big), \tag{4.80}$$

where

$$\frac{1}{\alpha} \;=\; \exp\big(\beta \int_{-\infty}^{\infty} dC(x)\big) \;=\; \exp\big(\beta C(\infty)\big). \tag{4.81}$$

Note that the corresponding probability density function must be

$$f(x) \;=\; F'(x) \;=\; \alpha\beta \exp(\beta C(x))c(x) \;=\; \beta F(x)c(x), \tag{4.82}$$

where

$$c(x) \;=\; C'(x) \tag{4.83}$$

at points x of differentiability of C. Jumps of C of height c_0 will yield jumps in F of height $\alpha e^{\beta c_0}$.

But now suppose X is a *random set of lessons*; that is, X is a random set. Each random set X possesses a *hitting probability* $T_X(x)$, which is the probability that x lies in X. This hitting probability is obtained by integrating the characteristic function $\chi_X(x)$ over the probability space of the set-valued random variable X.

For any integrable real-valued function g of a real variable x using Robbin's theorem:

$$E[\int_X g(x)dx] \;=\; \int_{-\infty}^{\infty} g(x)T_X(x)\,dx \;=\; E[g(x)|\,x \in X]. \tag{4.84}$$

Remark 4.11: At first glance, the expected cost savings from learning all the lessons from this random sample of lessons X would from (4.76) be

$$\Delta\overline{C}(X) \;=\; E[\int_X dC(x)] \;=\; \int_{-\infty}^{\infty} T_X(x)\, dC(x). \tag{4.85}$$

The second equality of (4.84) is indeed a correct statement of probability. However, lesson choice must be independent for this to be the learning observed in manufacturing, as shown in Chapter 3. Moreover, this independence is necessary to obtain the probability structure (4.84). Unfortunately, a random set X may possess correlated points, that is, where $x_1 \neq x_2$ yet

$$\text{prob}(x_1 \in X \text{ and } x_2 \in X) \;\neq\; T_X(x_1)T_X(x_2).$$

But when correlation is absent, (4.85) holds for random lesson choice.

Theorem 4.8 *If lesson choice from X is uncorrelated, then the expected cost savings from learning the random set of lessons of X is indeed given by*

$$\Delta\overline{C}(X) \;=\; E[\int_X dC(x)] \;=\; \int_{-\infty}^{\infty} T_X(x)\, dC(x). \tag{4.86}$$

Our task then is to construct an evolving random lesson sampling technique that grows over time to eventually cover all lessons, a process suggested by Robbins's 1944 modeling of airfield carpet bombing. This dynamic model of learning must allow *uncorrelated sampling of lessons* as well as *additivity of the resulting cost savings* when compound lessons are refined into independent sublessons. Both properties are necessary if we hope to retain the probability structure (4.80)–(4.81).

4.6.2 Expected Savings from a Random Set of Lessons.

We now exploit the machinery of the Principal Dynamic Random Set to build a continuous model of learning. There are two reasons why the Principal Dynamic Random Set gives the right framework for a continuous model of learning:

First, using the Principal Dynamic Random Set allows for infinite uncorrelated sampling of lessons. Recall that one of the central premises of Chapter 3 was that the lesson learning had to be uncorrelated (Assumption 3.4). The formation of the Principal Dynamic Random Set involves a sequence with a limit which qualitatively resembles an infinite number of infinitesimally small intervals. However, at each term of the sequence, λt is the measure of what is sampled. Because of the structure of the construction, there is no correlation of the probability of whether two distinct points are within the sampled space. There is no other process that samples an infinite number of points (indeed, a set of positive measure) but does so in a completely uncorrelated way.

Second, sampling using the Principal Dynamic Random Set naturally yields *infinite additivity*. A useful property of the construction of the probabilities is that the probabilities were additive. Take two lessons L_1 and L_2. Additivity means that if L_1 and L_2 are conjoined, the probability of picking the $L_1 + L_2$ lesson scales uniformly with to the product of the probabilities of learning L_1 and L_2 individually. With the Principal Dynamic Random Set we can subdivide lessons infinitely many times (this relates to my construction of the discrete set of lessons with each lesson comprising an interval). In other words, the Principal Dynamic Random Set carries additivity to its most extreme limit. Also, as will be shown later, a side benefit of the infinite additivity is that it allows for partial learning of lessons.

In short, we will sample via the Principal Dynamic Random Set because it allows

i. *Infinite Uncorrelated Learning*, and

ii. *Infinite Additivity.*

4.6.3 The Smooth Case and Principal Dynamic Random Sets

We present here only the finite incremental learning case, where cumulative cost saving $C = C(x)$ is everywhere continuously differentiable and where $C'(x) = c(x)$ has compact support. Suppose that learning is taking place at the fixed Poisson rate λ so that the lessons that comprise the measurable set L require $t = \mu(L)/\lambda$ seconds to learn, where μ is Lebesgue measure.

The construction. For each natural number k let $X_k(t)$ denote the random set of lessons that is the union of k intervals of type $[a, a + \lambda t/k)$, where the k left end points a are the result of k independent draws with probability given by the density function f of (4.86). Intuitively, and as the next result shows,

$$E[\mu(X_k(t))] \ \leq \lambda t, \tag{4.87}$$

that is, the lessons of $X_k(t)$ require on average at most t seconds to learn. Thus $X_k(t)$ is a random sample of k compound lessons that can be learned in time t.

Lemma 4.3 *Let $T_{X_k(t)}(x)$ denote the hitting probability of $X_k(t)$, namely the prob-*

ability that x will lie in $X_k(t)$. Then

$$T_{X_k(t)}(x) \; = \; 1 \; - \; \big[1 - \int_{x-\lambda t/k}^{x} f(y)\, dy\big]^k. \tag{4.88}$$

Proof. The point x will lie in the random interval $[a, a + \lambda t/k)$ exactly when $x \geq a > x - \lambda t$ with probability

$$\text{prob}\big(a \in (x - \lambda t/k, x]\big) \; = \; \int_{x-\lambda t/k}^{x} f(y)\, dy. \tag{4.89}$$

Thus x lies in none of the k intervals $[a, a + \lambda t)$ forming $X_k(t)$ with probability

$$\big[1 - \int_{x-\lambda t/k}^{x} f(y)\, dy\big]^k. \tag{4.90}$$

Lemma 4.4 *Lesson choice is asymptotically independent: Let*

$$T_{X_k(t)}(x_1, x_2) \; = \; prob\big(x_1 \in X_k(t) \text{ and } x_2 \in X_k(t)\big). \tag{4.91}$$

If $x_1 \neq x_2$, then for all sufficiently large k,

$$T_{X_k(t)}(x_1, x_2) \; = \; T_{X_k(t)}(x_1) \cdot T_{X_k(t)}(x_2). \tag{4.92}$$

Proof. When $\lambda t/k < |x_2 - x_1|$ it is impossible for both x_1 and x_2 to belong to the same one of the k random intervals $[a, a + \lambda t/k)$ that comprise $X_k(t)$.

Thus as k increases, the sample $X_k(t)$ consisting of k ever-shortening compound lessons begins to reassemble an independent sample of individual lessons, all of which can be learned in time t.

Lemma 4.5 *The hitting probabilities converge pointwise. In fact, for each x and*

$t \geq 0$, *we have*

$$\lim_{k \longrightarrow \infty} T_{X_k(t)}(x) \;=\; 1 - e^{-\lambda t f(x)}. \tag{4.93}$$

Proof. Since the probability density f has compact support,

$$\int_{x-\lambda t/k}^{x} f(x)\, dx \;=\; \frac{\lambda t f(x)}{k} + O(1/k^2). \tag{4.94}$$

Therefore,

$$\lim_{k \longrightarrow \infty} \big[1 - \int_{x-\lambda t/k}^{x} f(y)\, dy\big]^k \;=\; \lim_{k \longrightarrow \infty} \big[1 - \frac{\lambda t f(x)}{k} + O(1/k^2)\big]^k \tag{4.95}$$

$$=\; \exp(-\lambda t f(x)). \tag{4.96}$$

It is worth noting that this construction matches what is present above in Definition 4.11. Therefore, we may say the following.

Theorem 4.9 *The construction (4.87)-(4.96) yields a Principal Dynamic Random Set whose limiting hitting probability has the form (4.96). The resulting expected costs savings (4.98) asymptotically reduce to the discrete case (3.13), as shown below in Section 4.6.5.*

4.6.4 Prediction of unit cost

As seen in Lemma B, our random sample $X_k(t)$ of k compound lessons, learnable in t seconds, approaches for large k an independent sampling of individual lessons. By Lemma C, the corresponding hitting probability $T_{X_k(t)}$ approaches the limiting simple expression (4.93). Therefore, it is more than plausible that in the limit we have obtained a prediction of the decrease in cost-per-unit during manufacturing.

Theorem 4.10 *Assume cumulative cost saving $C = C(x)$ is continuously differen-*

tiable and that its derivative has compact support. If learning is proceeding at the Poisson rate λ, then the expected production cost U per unit is given by the rule

$$U(t) \;=\; U_0 - \int_{-\infty}^{\infty} (1 - e^{\lambda t f(x)})\, dC(x), \tag{4.97}$$

where U_0 is the initial cost per unit at the onset of production and where the probability density f is given by

$$f(x) \;=\; \alpha \exp(\beta C(x)) C'(x). \tag{4.98}$$

Proof. The unit cost is decreased from its initial cost by the expected cost saving per unit from learning during time t, i.e.,

$$U(t) = U_0 - \Delta \overline{C}(t). \tag{4.99}$$

But as argued above,

$$\Delta \overline{C}(t) \;=\; \lim_{k \longrightarrow \infty} E[\int_{X_k(t)} dC(x)]. \tag{4.100}$$

Hence (4.93) will then yield (4.96).

4.6.5 Devolution to the discrete case

As further evidence that Theorem B gives the correct model for unit cost evolution resulting from incremental learning, let us demonstrate that it cuts back to the discrete case of Theorem A. Let $I = [a, b]$ be the smallest closed interval containing the compact support of f. We may, as always, translate and rescale our lesson labeling system at will, so that in this case $I = [0, 1]$. Let us partition I in the usual way into n congruent subintervals $[x_{i-1}, x_i]$, with $0 = x_0 < x_1 < \cdots < x_n = 1$ and $\Delta x_i = x_i - x_{i-1} = 1/n$. Then as in (4) the cost saving accrued by the compound lesson $[x_{i-1}, x_i)$ is

$$c_i \;=\; \int_{x_{i-1}}^{x_i} dC(x), \tag{4.101}$$

which will be chosen via (7) with probability

$$p_i \;=\; \int_{x_{i-1}}^{x_i} f(x)\,dx \;=\; \frac{f(x_i^*)}{n}. \tag{4.102}$$

We take snapshots in time t_k under the time scaling where one compound lesson $[x_{i-1}, x_i)$ is learned on average per second. Then because λ is the average rate of learning per lesson length, $t_k\lambda = k/n$. Therefore the Riemann-Stieltjes sum approximation of the integral of (4.96) takes on the form

$$\begin{aligned}
\int_{-\infty}^{\infty} 1 - e^{-\lambda t_k f(x)}\,dC(x) &= \int_0^1 \left[1 - e^{-kf(x)/n}\right] dC(x) && (4.103)\\
&= \sum_{i=1}^{n} \left[1 - e^{-kf(x_i^*)/n}\right] c_i \;+\; O(1/n^2) && (4.104)\\
&= \sum_{i=1}^{n} \left[1 - \big(1 - \frac{f(x_i^*)}{n}\big)^k\right] c_i \;+\; O(k/n^2) && (4.105)\\
&= \sum_{i=1}^{n} [1 - \big(1 - p_i\big)^k]\, c_i \;+\; O(k/n^2), && (4.106)
\end{aligned}$$

which is consistent with the discrete result in Chapter 3.

This random set model for learning has features which are not possible in other models. As will be shown in Chapter 5, the only viable learning model which uses a partial differential equation is one that aggregates states, so it measures only the fraction of lessons learned, not which lessons might be learned. The random set model allows for an infinite uncorrelated sampling of an infinite number of lessons.

As mentioned in the previous paragraph, the random set model also naturally allows for the partial learning of a finite number of lessons.

As discussed in Remark 4.6, the sequence of random sets associated with the construction of the dynamic random set does not converge strongly to any random set. It is unknown whether it converges weakly to a random set, however. It seems doubtful that this weak convergence occurs either. Therefore, we make the following conjecture:

Conjecture 4.1: The sequence of random sets associated with the construction of the dynamic random set does not converge weakly to any random set.

Chapter 5

Continuous Model II: A Partial Differential Equation for Learning

5.1 . The Model as a Limit of a Markov Process

5.1.1 . Construction of a Master Equation

Let us build intuition by first examining a discrete toy model of learning with a small number n of lessons to be learned. For the moment we will focus on counting the probable number $m(k)$ of distinct lessons learned after k steps (draws) with replacement. The central idea of discrete learning models is that at each step either one lesson is learned or nothing is learned. Of any of the particular m lessons (say the i-th lesson), the probability p_i of learning this i-th lesson was found in Chapter 3 to be

$$p_i \propto e^{\beta \Delta C_i}, \tag{5.1}$$

where β is a parameter of the learning process and ΔC_i is the improvement in the quality metric when the i-th lesson is learned [1].

The relationship (5.1) means that the more that the choice improves the quality, the more likely it is to be chosen. This fact is supported by much in the artificial intelligence literature, as several studies in artificial intelligence have demonstrated that the most valuable rules are learned first, and any less important rules are learned later [Russell and Norvig 2002]. If at any stage the roll of the dice yields a lesson that has already been learned, no learning takes place during that step. Thus it is harder to find the lessons that have not already been learned if many lessons have been learned. The probabilities used in the model do not change over time, which is characteristic of a Markov chain.

In order to create a Markov chain model, we must first define states. In Chapter 3, it was assumed that a state is described by particular coordinates of an n-tuple v in which the entries are either 0 or 1. Each entry represents a discrete lesson which can be learned to improve the metric. The initial state $v_0 = (0, 0, \dots, 0)$ is the 0 vector, while the desired state $v_{2^n-1} = (1, 1, \dots, 1)$ is the state where every possible lesson has been learned. For n lessons, there are of course 2^n possible states.

Definition 5.1 *Assume there are n lessons to be learned. Let $\mathbf{q}^{(0)}$ be the 2^n-tuple $(1, 0, \dots, 0)^T$. Then define $\mathbf{q}^{(k)} = (q^{(k)}0, \dots, q^{(k)}_{2^n-1})$ by the recursion*

$$\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} + A\mathbf{q}^{(k)} = (I + A)^k \mathbf{q}^{(0)} \tag{5.2}$$

where A is a 2^n by 2^n matrix governing the change of probabilities between iterations and $q_i^{(k)}$ is the probability for the system to be in state v_i after k iterations of the

[1] Recall that ΔC is not necessarily the change in the magnitude of the quality metric. For example, if the quality metric is cost, than a decrease ΔC in the quality metric is an improvement. On the other hand, if the quality metric is mean life of the product, that an increase ΔC in the quality metric is an improvement.

equation.

Recall that at each step in the discrete process, one of the n possible lessons is chosen. If the chosen lesson has not been previously learned then a transition occurs. Otherwise, there is no change to the state. Recursion (5.2) is the *master equation* governing the evolution of the probabilities of the states of learning.

The master equation formulation of a Markov chain is preferable over a transition matrix formulation since we wish to pass to an infinite number of possible lessons to be learned; a master equation formulation passes to the limit as a differential equation, as we will soon see. In any case, given a master equation

$$\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} + A\mathbf{q}^{(k)}$$

with resulting state probabilities $\mathbf{q}^{(k)} = \{q_1^{(k)}, \ldots, q_{2n}^{(k)}\}^T$ at step k, and matrix A from (5.2), the traditional *transition matrix* B for the system of learning states is given by

$$B = (A + I)^T,$$

a matrix with nonnegative entries where each row sums to 1.

Example 5.1: Let us examine what the model would look like for $n = 4$ lessons, with the i-th lesson having probability p_i to be learned at each step. In this case, there are 16 possible states, each with distinct probabilities q_j to account for each of the 4 lessons which could be turned off or on:

1. $q_0^{(k)}$ is the probability that the system after k steps is in the state where no lessons have been learned.

2. $q_1^{(k)}$ is the probability that the system after k steps is in the state where only

lesson 1 has been learned.

3. $q_2^{(k)}$ is the probability that the system after k steps is in the state where only lesson 2 has been learned.

4. $q_3^{(k)}$ is the probability that the system after k steps is in the state where only lesson 3 has been learned.

5. $q_4^{(k)}$ is the probability that the system after k steps is in the state where only lesson 4 has been learned.

6. $q_5^{(k)}$ is the probability that the system after k steps is in the state where only lessons 1 and 2 have been learned.

7. $q_6^{(k)}$ is the probability that the system after k steps is in the state where only lessons 1 and 3 have been learned.

8. $q_7^{(k)}$ is the probability that the system after k steps is in the state where only lessons 1 and 4 have been learned.

9. $q_8^{(k)}$ is the probability that the system after k steps is in the state where only lessons 2 and 3 have been learned.

10. $q_9^{(k)}$ is the probability that the system after k steps is in the state where only lessons 2 and 4 have been learned.

11. $q_{10}^{(k)}$ is the probability that the system after k steps is in the state where only lessons 3 and 4 have been learned.

12. $q_{11}^{(k)}$ is the probability that the system after k steps is in the state where only lessons 1, 2, and 3 have been learned.

13. $q_{12}^{(k)}$ is the probability that the system after k steps is in the state where only lessons 1, 2, and 4 have been learned.

14. $q_{13}^{(k)}$ is the probability that the system after k steps is in the state where only lessons 1, 3, and 4 have been learned.

15. $q_{14}^{(k)}$ is the probability that the system after k steps is in the state where only lessons 2, 3, and 4 have been learned.

16. $q_{15}^{(k)}$ is the probability that the system after k steps is in the state where all lessons have been learned.

The initial probability for the system to be in the 'ground state' $q_0^{(0)}$ will be 1; all other $q_i^{(0)}$ will thus be zero. In this example, the evolution of probability of states will be given by the iteration

$$\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} + A\mathbf{q}^{(k)}, \tag{5.3}$$

where the matrix A is given by the block matrix

$$\begin{pmatrix} -1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ A_1 & A_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_3 & A_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_5 & A_6 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & A_7 & 0 \end{pmatrix}, \tag{5.4}$$

where

$$A_1 = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}, \tag{5.5}$$

$$A_2 = \begin{pmatrix} -(1-p_1) & 0 & 0 & 0 \\ 0 & -(1-p_2) & 0 & 0 \\ 0 & 0 & (-1-p_3) & 0 \\ 0 & 0 & 0 & -(1-p_4) \end{pmatrix}, \tag{5.6}$$

$$A_3 = \begin{pmatrix} p_2 & p_1 & 0 & 0 \\ p_3 & 0 & p_1 & 0 \\ p_4 & 0 & 0 & p_1 \\ 0 & p_3 & p_2 & 0 \\ 0 & p_4 & 0 & p_2 \\ 0 & 0 & p_4 & p_3 \end{pmatrix}, \tag{5.7}$$

$$A_4 = \begin{pmatrix} -(p_3+p_4) & 0 & 0 & 0 & 0 & 0 \\ 0 & -(p_2+p_4) & 0 & 0 & 0 & 0 \\ 0 & 0 & -(p_2+p_3) & 0 & 0 & 0 \\ 0 & 0 & 0 & -(p_1+p_4) & 0 & 0 \\ 0 & 0 & 0 & 0 & -(p_1+p_3) & 0 \\ 0 & 0 & 0 & 0 & 0 & -(p_1+p_2) \end{pmatrix}, \tag{5.8}$$

$$A_5 = \begin{pmatrix} p_3 & p_2 & 0 & p_1 & 0 & 0 \\ p_4 & 0 & p_2 & 0 & p_1 & 0 \\ p_4 & p_3 & 0 & 0 & p_1 & 0 \\ 0 & 0 & p_4 & p_3 & p_2 & 0 \end{pmatrix}, \tag{5.9}$$

$$A_6 = \begin{pmatrix} -p_4 & 0 & 0 & 0 \\ 0 & -p_3 & 0 & 0 \\ 0 & 0 & -p_2 & 0 \\ 0 & 0 & 0 & -p_1 \end{pmatrix}, \tag{5.10}$$

and

$$A_7 = (p_4, p_3, p_2, p_1) \tag{5.11}$$

The block entries (5.4)–(5.11) follow from examining the one step evolution of any one ‘pure state’ (i.e. elements of the natural basis).

The system of equations (5.3)–(5.11) trivially meets the *mass conservation re-*

quirement that the columns of A sum to zero,

$$\sum_{i=1}^{2^n} A_{ij} = 0. \tag{5.12}$$

A system such as (5.2)-(5.10) is given by a lower triangular matrix A and is thus explicitly solvable. However, it is apparent that problems arise when n is increased without limit. The number of equations is $O(2^n)$, and the somewhat irregular structure of A makes it extremely difficult to obtain even qualitative results for the system.

In the general case, with n lessons to be learned, the states and their probabilities will then be ordered in the following way: the 'ground state' in which no lessons have been learned will come first. The next n states are where just one lesson has been learned. Then the next $n(n+1)/2$ states are all the possible states where 2 lessons have be learned, and so on.

5.1.2. Reduction of Order

This problem can be better analyzed once we reduce the order of the system. A natural approach is to amalgamate the states. Examining the master equation, one can see that there is a block relationship among the state probabilities. Specifically, *transitions between states only occur between those states which have one more or one fewer lesson to be learned.*

Therefore, one can redefine the states in the following way: again let n be the number of lessons.

There are $n+1$ aggregated states, where the i-th aggregate state
has had exactly i lessons learned, and where i can run from 0 to n.

Definition 5.2 *Given n lessons to be learned, let $\mathbf{q}^{(k)} = (q_0^{(k)}, q_1^{(k)}, \ldots, q_{2^k-1}^{(k)})^T$ be* *the 2^n-tuple for the state probabilities according to Definition 5.1. Then $\mathbf{u}^{(k)} = (u_0^{(k)}, \ldots, u_n^{(k)})^T$ is to be the $(n+1)$-tuple with $u_i^{(k)}$ equal to the probability that exactly i lessons have been learned at the k-th iteration.*

Lemma 5.1 *In the formulation of Definition 5.2, $u_0^{(k)} = q_0^{(k)}$. Moreover, for $i > 0$, the i-th component of $\mathbf{u}^{(k)}$ will be given by*

$$u_i^{(k)} = \sum_{j=w}^{w+\binom{n}{i}-1} q_j^{(k)}, \tag{5.13}$$

where

$$w = \sum_{k=0}^{i-1} \binom{n}{k}. \tag{5.14}$$

Proof: Clear, since the the number of $q_i^{(k)}$ aggregated into the state will be $\binom{n}{i}$.

To determine the evolution equation for these aggregated states and their associated probabilities, we need an expression which yields the probability of transition from the aggregated state i to the aggregated state $i+1$. For the rest of this chapter we will only consider aggregated states. We will first examine the simple case, where the probabilities p_i are all equal, and make the model continuous in time. Then we will examine the case of unequal probabilities.

5.2 The Continuous Case

5.2.1. The Continuous Case with Equal Weighting

Making the Model Continuous in Time

Let us examine the special case of the aggregated state model where all lessons have an equal probability of being learned. If lessons have already been learned, there is a likelihood that a transition will not occur. If i out of n lessons have already been learned, then the probability of a transition due to a new lesson is

$$\frac{n-i}{n}. \tag{5.15}$$

This transition probability yields the transition equation

$$\mathbf{u}^{(\mathbf{k+1})} = \mathbf{u}^{(\mathbf{k})} + U\mathbf{u}^{(\mathbf{k})} = (I+U)\mathbf{u}^{(\mathbf{k})}, \tag{5.16}$$

where

$$U = \begin{pmatrix} -1 & 0 & 0 & \dots & & \\ 1 & -\frac{n-1}{n} & 0 & 0 & \dots & \\ 0 & \frac{n-1}{n} & -\frac{n-2}{n} & 0 & 0 & \dots \\ \vdots & 0 & \ddots & \ddots & 0 & \dots \\ 0 & 0 & 0 & \frac{n-i+1}{n} & -\frac{n-1}{n} & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots \end{pmatrix}. \tag{5.17}$$

Let us now derive a model which is continuous in time. We do this by having learning occur as a Poisson process with rate λ in place of the step variable k. A Poisson process is a model of a random occurrence process, characterized by a rate constant λ, where the probability of exactly k occurrences during a time period Δt

is given by

$$\frac{(\lambda\Delta t)^j}{j!}e^{-\lambda\Delta t}. \tag{5.18}$$

Lemma 5.2 *Given the transition process*

$$\begin{aligned}\mathbf{u}^{(\mathbf{k+1})} &= \mathbf{u}^{(\mathbf{k})} + M\mathbf{u}^{(\mathbf{k})} && (5.19)\\ &= (I+M)\mathbf{u}^{(\mathbf{k})}, && (5.20)\end{aligned}$$

where the transitions are instead made via a Poisson process with rate λ, the expected state after a time Δt, starting from the initial state $\mathbf{u}$, will be given by

$$E(\mathbf{u}(\Delta t)) = \mathbf{u} + \lambda\Delta t M\mathbf{u} + O(\Delta t^2). \tag{5.21}$$

Proof. The expected value will be given by

$$\begin{aligned}E(\mathbf{u}(\Delta t)) &= \sum_{j=0}^{\infty}\frac{(\lambda\Delta t)^j}{j!}e^{-\lambda\Delta t}(I+M)^j\mathbf{u} && (5.22)\\ &= e^{-\lambda\Delta t}(I+\lambda\Delta t(I+M))u + \mathbf{v}, && (5.23)\end{aligned}$$

where

$$\mathbf{v} = \sum_{j=2}^{\infty}\frac{(\lambda\Delta t)^j}{j!}e^{-\lambda\Delta t}(I+M)^j\mathbf{u}. \tag{5.24}$$

The 1-norm of v is easily estimated.

$$\begin{aligned}||\mathbf{v}||_1 &\le e^{-\lambda\Delta t}(\lambda\Delta t)^2||I+M||_1^2\sum_{j=0}^{\infty}\frac{(\lambda\Delta t)^j}{(j+2)!}||I+M||_1^j||\mathbf{u}||_1 && (5.25)\\ &\le e^{-\lambda\Delta t}(\lambda\Delta t)^2||I+M||_1^2\sum_{j=0}^{\infty}\frac{(\lambda\Delta t)^j}{j!}||I+M||_1^j||\mathbf{u}||_1 && (5.26)\\ &= \frac{1}{2}e^{-\lambda\Delta t}(\lambda\Delta t)^2||\mathbf{u}||_1||I+M||_1^2e^{\lambda\Delta t||I+M||_1}. && (5.27)\end{aligned}$$

The first term in the expression (5.23) can be reduced as follows

$$\begin{aligned} & e^{-\lambda\Delta t}(I+\lambda\Delta t(I+M))\mathbf{u} && (5.28) \\ = & \left(\sum_{j=0}^{\infty}\frac{(-\lambda\Delta t)^j}{j!}\right)(I+\lambda\Delta t(I+M))\mathbf{u} && (5.29) \\ & = \mathbf{u}+\lambda M\mathbf{u}+\mathbf{w}, && (5.30) \end{aligned}$$

where $||\mathbf{w}||_1$ is clearly $O(\Delta t^2)$.

We have thus proved that in the Poisson limit the iteration matrix changes from $I+M$ to the infintesimal generator λM. Therefore, the discrete-time system (5.19) becomes the continuous-time system $\dot{u} = \lambda M u$.

Theorem 5.1 *When lessons are learned at a Poisson stochastic rate, the aggregated state probabilities evolve by the rule*

$$\dot{u} = \lambda U u, \tag{5.31}$$

where U is given by

$$U = \begin{pmatrix} -1 & 0 & 0 & \dots & & \\ 1 & -\tfrac{n-1}{n} & 0 & 0 & \dots & \\ 0 & \tfrac{n-1}{n} & -\tfrac{n-2}{n} & 0 & 0 & \dots \\ \vdots & 0 & \ddots & \ddots & 0 & \dots \\ 0 & 0 & 0 & \tfrac{n-i+1}{n} & -\tfrac{n-1}{n} & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots \end{pmatrix}. \tag{5.32}$$

Proof: Clear.

Making the Model Continuous in Space

We next pass to a limit of a continuous spatial variable. *We define the spatial variable x to be the fraction of all lessons which have been learned.* Thus the domain of this problem will be the unit interval $[0, 1]$, where $x = 0$ represents the state where no lessons have been learned and $x = 1$ represents the state where all lessons have been learned. With such continuous learning, the number of lessons to be learned become infinite. Before, at any time t, the mean number of steps taken will be λt. To reach more than an infinitesimal fraction x, we take a limit where λ goes to infinity, so that k, i, and n also go to infinity, and define a new parameter λ' according to the rules

$$\begin{aligned} \lambda &\rightarrow \infty, && (5.33) \\ n &\rightarrow \infty, && (5.34) \\ i &\rightarrow \infty, && (5.35) \\ \lambda' &= \frac{\lambda}{n}. && (5.36) \end{aligned}$$

The variable u will then become a function of the spatial variable x as well as of t. The probability of having learned a fraction of lessons between x and $x + dx$ then is equal to udx.

Lemma 5.3 *Define a differential operator P as follows*

$$Pu = -\lambda' \frac{\partial}{\partial x} \left((1 - x)u \right), \tag{5.37}$$

where $(1-x)u$ belongs to $\mathbb{H}^1([0, 1])$. Then (5.32) is the backwards Euler discretization of the operator P.

Proof: We discretize the operator P given in (5.37) according to a backwards Euler method the spatial derivative in P over n intervals, so that

$$\frac{\partial}{\partial x}\left((1-x)u\right)|_i \approx \frac{(1-x_i)u_i-(1-x_{i-1})u_{i-1}}{1/n}, \tag{5.38}$$

so that on the interval $[0,1]$ where $x_i = i/n$ we obtain

$$Pu|_i \approx n\lambda' \left(\frac{n-i+1}{n}u_{i-1} - \frac{n-i}{n}u_i\right), \tag{5.39}$$

or

$$Pu|_i \approx \lambda \left(\frac{n-i+1}{n}u_{i-1} - \frac{n-i}{n}u_i\right), \tag{5.40}$$

from (5.29). Thus the approximation for Pu_i yields the right hand side of (5.32).

Theorem 5.2 *Taking the limits (5.33)–(5.36) yields the continuous aggregated states model of the fraction x of lessons learned,*

$$\dot{u} = -\lambda' \frac{\partial}{\partial x}\left((1-x)u\right). \tag{5.41}$$

We also note that the initial condition of the system will be given by

$$u(x,0) = \delta(x). \tag{5.42}$$

This initial condition is dictated from the modeling requirement that the initial state of the system has no lessons which have been learned and that the probability of finding the system in some state $x \in [a,b]$ is given by

$$\int_a^b u(x,t)\,dx. \tag{5.43}$$

Remark 5.1. Model (5.41) has the form of a transport equation with a singular point at $x = 1$. As with any standard transport equation, the term $\lambda'(1-x)$ represents the *velocity* of the state. This velocity is positive and monotonically decreasing on $[0,1]$, reflecting the fact that learning happens continuously for as long as there are lessons to be learned, and that the learning rate decreases as learning progresses.

Properties of the Transport Model of Learning

Lemma 5.4 *Given the impulsive condition $u(x,0) = \delta(x)$, the solution of (5.41) becomes*

$$u(x,t) = \delta(x - (1 - e^{-\lambda' t})). \tag{5.44}$$

Proof. We will now solve for the characteristics:

$$\begin{aligned} \frac{dt}{ds} &= 1, && (5.45)\\ \frac{dx}{ds} &= \lambda'(1-x), && (5.46)\\ \frac{du}{ds} &= \lambda' u. && (5.47) \end{aligned}$$

This system when solved yields the following

$$\begin{aligned} t &= s, && (5.48)\\ x &= 1 - (1 - x_0)e^{\lambda' t}, && (5.49)\\ u &= u(x_0)e^{\lambda' t}. && (5.50) \end{aligned}$$

When these equations are solved to eliminate x_0 and u_0, we obtain

$$u(x,t) = e^{\lambda' t}\delta(1-(1-x)e^{\lambda' t}), \tag{5.51}$$

which equals

$$u(x,t) = \delta(x-(1-e^{-\lambda' t})), \tag{5.52}$$

via the identity

$$a\delta(ax+b) = \delta(x+b/a). \tag{5.53}$$

When the system is within the initial state, (no improvements have been made at the initial time),

$$u(x,0) = \delta(x). \tag{5.54}$$

Remark 5.2: The discrete case for at least two lessons to be learned involves some variance in the probabilities at any positive time (i.e., for all iterations, all u_i are strictly less than 1). However, the continuous model has no variance in the time evolution of learning, as $u(x,t)$ retains the form of a delta function for all t. The disappearance of the variance matches the result obtained in the application of weak convergence of random sets to the problem in Chapter 4, where the variance vanished, as guaranteed by the examination of the associated random functionals in Section 4.4.1.

The results of this section are summarized in the following theorem.

Theorem 5.3 *Suppose all lessons have an equal probability of being chosen. Then the evolution of the fraction x of the lessons learned by time t is governed by the partial differential transport equation*

$$\dot{u} = -\lambda' \frac{\partial}{\partial x}\left((1-x)u\right), \tag{5.55}$$

where $u = u(x,t)$ is for each $t > 0$ the probability density

$$prob(a \leq x \leq b) = \int_a^b u(x,t)dx. \tag{5.56}$$

The initial condition of the system will be given by

$$u(x,0) = \delta(x), \tag{5.57}$$

which is dictated from the modeling requirement that the initial state of the system has no lessons which have been learned.

5.2.2. The Continuous Case with Unequal Weighting

Suppose now that the lessons are no longer equally weighted, but instead each lesson has its own distinct probability p of being learned. We will now examine more closely the set of differential equations which govern this unequally-weighted case. We again replace the discrete steps with a Poisson process of rate λ. The transitions are governed by:

$$\begin{aligned}
\dot{u_0} &= -\lambda u_0, && (5.58)\\
\dot{u_1} &= \lambda(u_0 - f(1)u_1), && (5.59)\\
&\vdots && (5.60)\\
\dot{u}_m &= \lambda(f(m-1)u_{m-1} - f(m)u_m), && (5.61)\\
&\vdots && (5.62)\\
\dot{u}_n &= \lambda f(n-1)u_{n-1}, && (5.63)
\end{aligned}$$

or in vector form

$$\dot{\mathbf{u}} = V\mathbf{u}, \tag{5.64}$$

where

$$V = \lambda \begin{pmatrix} -1 & 0 & 0 & \dots \\ 1 & -f(1) & 0 & \dots \\ 0 & f(1) & f(2) & \dots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \tag{5.65}$$

and where $f(i)$ equals the probability that a lesson not previously learned will be learned at the next iteration, given that i lessons have already been learned.

We next examine a limit with a continuous spatial variable. At any time t, the mean number of steps taken will be λt. We take a limit where λ goes to infinity, so that i and n also go to infinity, and λ' is defined as follows:

$$\lambda \rightarrow \infty, \tag{5.66}$$

$$n \rightarrow \infty, \tag{5.67}$$

$$i \rightarrow \infty, \tag{5.68}$$

$$\frac{\lambda}{n} \rightarrow \lambda'. \tag{5.69}$$

The function f will now have a domain of $[0, 1]$, with $x \in [0, 1]$ representing the fraction of the total number of lessons which have already be learned.

We can then again see that the matrix V is a backwards Euler discretization of the operator Q, where

$$Qu = -\lambda' \frac{\partial}{\partial x} \left(f(x)u \right). \tag{5.70}$$

Thus our learning curve model becomes:

$$\dot{u} = Qu, \tag{5.71}$$

with Q given by (5.69).

Thus the limiting case of the discrete system (5.65)-(5.68) when lessons are not equally weighted leads to the modified transport equation of the form

$$\frac{\partial u}{\partial t} = -\lambda' \frac{\partial}{\partial x}(f(x)u), \tag{5.72}$$

where the product $\lambda' f(x)$ represents the characteristic velocity of transport for the system, as will be proven in Theorem 5.4.

Theorem 5.4 *In a continuous learning model where there are unequal probabilities for each lesson to be learned, the learning will follow the equation*

$$\frac{\partial u}{\partial t} = -\lambda' \frac{\partial}{\partial x}(f(x)u), \tag{5.73}$$

with initial condition $u(x, 0) = \delta(x)$ and learning velocity $\lambda' f(x)$.

5.2.3. Properties of the Learning Velocity

One important property of the resulting partial differential equation (5.80) is that the learning velocity is maximized in the case when the probabilities of learning each lesson are equal. In the finite learning case, the probabilities of all the lessons to be learned together constitute an n-dimensional parameter in the function $m(k)$. In more detail, the probability of learning the i-th lesson after k attempts (with replacement) is

$$1-(1-p_i)^k. \tag{5.74}$$

Summing this expression over all lessons gives the expected number of distinct lessons learned after k chances:

$$m(k)=\sum_{i=0}^{n} 1-(1-p_i)^k, \tag{5.75}$$

as given by (3.15).

This expression is strictly monotonic and therefore invertible, even though it cannot be expressed in closed form. In general, the inverse of an expression for an expected value of a random variable is not invertible; however, if the variance is zero, the random variable is almost surely a single value, and thus deterministic. Therefore, one can find the inverse relation simply by inversion. The inverse $k(m)$ therefore asymptotically yields the expected number of draws it takes to have m distinct objects.

Lemma 5.5 . *The expression* $m(k)=\sum_{i=0}^{n} 1-(1-p_i)^k$ *is maximized at each* $k \geq 1$ *when the probabilities* p_i *are equal.*

Proof. This Lemma is a special case of the following.

Lemma 5.6 *Suppose* $H = H(\mathbf{x})$ *is a symmetric convex real-valued function on the convex set* $K \subset \mathbb{R}^n$ *such that* H *has a minimum on* K *at a point* $\mathbf{x}$. *Suppose further that* K *is closed under any permutation of the coordinates of the points in* K. *Then this minimum is taken at a point of* K *with all coordinates equal.*

Proof. Suppose H takes on a global minimum value m at some point $x = (x_1, \ldots, x_n)$ of K. Since H is symmetric, $H(\sigma x) = m$ where $\sigma x = (x_{\sigma 1}, \ldots, x_{\sigma n})$ is the vector obtained by permuting the coordinates of x by the permutation σ. Let

$$\overline{x} = \frac{1}{n!} \sum_{\sigma} \sigma x, \tag{5.76}$$

where the summation is over all $n!$ possible permutations. Clearly $\sigma\overline{x} = \overline{x}$. But because H is convex, $H(\overline{x}) = m$.

The proof of Lemma 5.5 then is the special case where H is given by the expression in Lemma 5.5 with domain $p_1 + \cdots + p_n = 1$, $p_i \geq 0$. As a corollary, since the expression

$$\sum_{i=1}^{n} q^k = \sum_{i=1}^{n} (1 - p_i)^k \tag{5.77}$$

is minimized when the probabilities are equal, the expression

$$m(k) = \sum_{i=1}^{n} 1 - (1 - p_i)^k \tag{5.78}$$

is maximized for equal probabilities.

Remark 5.2: For the present investigation into learning, this Lemma 5.2 guarantees that the inverse function $m(k)$ will be maximized when the probabilities are all equal. It should be pointed out that while having equal probabilities does maximize the number of lessons learned, it does not necessarily maximize the benefits of learning. As discussed previously, an artificial intelligence model of learning implies a direct correlation between the benefit of learning a lesson and the probability of learning that lesson, such as rule 5.1 given previously, namely

$$P(\Delta C) \propto e^{\beta \Delta C}. \tag{5.79}$$

As can be seen by this rule, the limit as β goes to zero leads to the case of equal probability, and consequently leads to a maximal rate to learning.

Remark 5.3: In the preceding form (5.72) of the transport equation for an arbitrary distribution of choices, $f(x)$ is determined by the distribution of decisions. As with any transport equation the factor $\lambda' f(x)$ represents the velocity for the evolution of the state. As described in Section 5.2, $f(x) = 1 - x$ for a uniform distribution. For a non-uniform distribution, $f(x) < 1 - x$, since Lemma 5.5 requires that the maximum learning rate occurs in the case of equal probabilities. Specifically, since the inequality is true in the discrete case for every n, the inequality passes through the limit to the continuous case. Therefore, the equal-probability case places an upper limit on the learning rate for every case.

Lemma 5.7 *Characteristic curves for the partial differential equation (5.72) move with velocity v given by*

$$v = \lambda' f(x). \tag{5.80}$$

Proof. Since (5.56) is linear and of first order, the characteristics are given by

$$\begin{aligned} \frac{dt}{ds} &= 1, && (5.81) \\ \frac{dx}{ds} &= \lambda f(x), && (5.82) \\ \frac{du}{ds} &= 0. && (5.83) \end{aligned}$$

This system of differential equations yields that solutions of (5.72) are of the form

$$u(x,t) = u(x - vt), \tag{5.84}$$

where

$$v = \lambda' f(x). \tag{5.85}$$

Since $f(0) = 1$, $f(1) = 0$, and $f(x)$ is monotonic on $[0, 1]$, the model reflects the intuition that the rate at which lessons will be learning is monotonically decreasing, and will only asymptotically approach $x = 1$.

Remark 5.4: Since (5.41) is a standard transport equation, it also obeys a conservation law. Specifically, the characteristic equations imply that the equation preserves mass, that is,

$$\frac{d}{dt}||u||_1 = 0. \tag{5.86}$$

5.3 . Cost Evolution during Learning

The ultimate goal of any learning curve model is to predict the evolution of an associated given quality metric, such as cost. This determination can be made by examining the expected value of cost savings at each state x. In the continuous case, the expected cost savings would then be given by

$$\overline{C}(u) = \int_0^1 u(x,t)\, dP(x), \tag{5.87}$$

where $dP(x) = p(x)dx$ is the differential cost savings between x and $x + dx$. By employing our model (5.55), we see that for $t > 0$

$$\frac{d}{dt}\overline{C}(u) = \frac{d}{dt}\int_0^1 u(x,t)\, dP(x) = \int_0^1 u_t(x,t)\, dP(x) \tag{5.88}$$

$$= \int_0^1 -\lambda\frac{\partial}{\partial x}(f(x)u)\, dP(x) = \lambda\int_0^1 f(x)u(x)\, p'(x)dx \tag{5.89}$$

$$= \lambda\int_0^1 p'(x)f(x)u(x)\, dx, = \lambda E(p'(x)f(x)) \tag{5.90}$$

since $u(x)$ gives the probability of fraction x at each x. Note that the boundary terms in the integration by parts disappear since $u(0) = u(1) = 0$ for $t > 0$. The expected value is over the probabilities u, and is therefore time-dependent.

Remark 5.5: While having the probabilities equal to each other may maximize the total number of options taken, it will not necessarily result in a maximal improvement in the quality metric. Early in the run, when the probability of choosing a new decision is very high, it is wise to make the probabilities so that the options with the best cost savings are most likely to be chosen. Later on, when the best options have been chosen, the improvement in the quality metric is optimized by having the probabilities more equal. The point at which a transition is made between the two regimes will depend on the underlying distribution of available options for improvements to the quality metric. Finally, it is also worth noting the similarity of (5.85) to the model of cost savings (4.86) from Chapter 4.

5.4 . Forgetting

We now examine the phenomenon of *forgetting* in learning. Forgetting can take place in two forms: during *lapses of production* or *intermingled with learning during production.* The first type of forgetting is modeled well as negative learning, with evolution determined by

$$\frac{\partial u}{\partial t} = -\lambda_f f(x) \frac{\partial u}{\partial x}, \tag{5.91}$$

where λ_f is a *forgetting rate* that is (typically) smaller than the learning rate. In this situation, $f(0) = 0$ (if everything is forgotten, then no more forgetting will occur) and $f(1) = -1$ (if every lesson has been learned, something will surely be forgotten).

The more interesting type of forgetting is forgetting intermingled with learning during production. In this case, there are simultaneous possibilities of going backwards or forwards. Although a comprehensive theory of how these ideas might interlink does not exist, we can examine some qualitative aspects of such a model. We define a_i to represent a *forgetting rate* if i lessons have already been learned. The state space equations have the form

$$\dot{\mathbf{u}} = \lambda \begin{pmatrix} -1 & a_1 & 0 & \dots & \\ 1 & -a_1 - f(1) & a_2 & 0 & \dots \\ 0 & f(1) & -a_2 - f(2) & a_3 & \dots \\ \vdots & 0 & \ddots & \ddots & \ddots \end{pmatrix} \tag{5.92}$$

With three distinct terms for almost every equation, it is apparent that a first order partial differential equation will not form a continuous model. Instead, the continuous model should have the form

$$\dot{u} = Pu, \tag{5.93}$$

where

$$Pu = a(x)u_x + b(x)u_{xx}. \tag{5.94}$$

From even this simple qualitative treatment, several facts are evident. First, we have a parabolic model, where more complicated behaviors are to be expected. Second, the introduction of a second derivative introduces a dispersive effect on the evolution of the state. That the model is parabolic also implies that some variance in the expected value of the state will exist–in contrast to the model which does not incorporate forgetting (which becomes deterministic in the continuous limit).

In fact, some other characteristics of this spreading can be determined through qualitative analysis of the coefficients a and b. For example, it is evident that $b(0) = 0$ and $b(1) = 0$, since otherwise there would be a nonzero probability of reaching either a negative value of x or a value which is greater than 1. Also, $a(1) \leq 0$ (again, no transport past $x = 1$). However, unlike the previous case, a does not have to be 0 at the right endpoint, although a will be monotonic on $[0, 1]$. What is likely to occur is that on $(0, 1)$, there will exist some state x in which it is equally likely to learn a lesson as it is to forget a lesson.

5.5 . Conclusion

A partial differential equation model of learning has been presented. Instead of having to deal with an impossibly large set of parameters to analyze, this continuous limit involves only a few functional parameters, namely λ' and $f(x)$ that govern the process. The analysis of these parameters will yield accurate predictions in the analysis of real-life learning curves.

Part III

Simulations and Applications

Chapter 6

Simulations of Models

6.1 Introduction and Methodology

In this chapter several different simulations of the model will be presented. The first two sections will concentrate on the standard formulation of the model, and the *multi-dimensional quality metrics* will be discussed in the subsequent section. The simulations will be exclusively that of the discrete model, since we have shown in Sections 4.6 and 5.2 that the continuous models has a first order discretization equal to the discrete model

These simulations were conducted using a program written by the author in the C language. For each simulation several inputs must be specified. One such input is the managerial effectiveness parameter β. The second input is a *floor level*, the best obtainable value for the quality metric. The last (and most complex) input is the set of the quality improvements accrued by each lesson to the quality metric. Typical sizes for the set of possible lesson improvement values range between 50 to 500. When running these simulations, it is generally noticed that overall results were not sensitive to the size of the set of possible improvement values, although, as expected, more variance was observed for those smaller choice spaces.

To obtain the evolution of the quality metric, a Monte Carlo sampling of the choice space was performed, according to the specified probabilities. Although for the simple learning curve model (3.23)–(3.24),

$$C - \overline{C}(k) = \sum_{i=1}^{n}(1 - (1 - p_i)^k)\Delta C_i, \tag{6.1}$$

where

$$p_i = \frac{e^{\beta \Delta C_i}}{\sum_{j=1}^{n} e^{\beta \Delta C_j}}, \tag{6.2}$$

an analytical solution can be expressed functionally, there are three reasons for proceeding with Monte Carlo experiments. First, such a stochastic format allows the exploration of variances in learning rates (The analytical expressions (6.1)–(6.2) are merely expressions of the expected learning rates). Second, it can be observed that the expression (6.1) involves the sum of a large number of components, each of which involve multiplication of pairs of small numbers, typically much less than one, all yielding a likelihood of significant rounding errors being introduced. In contrast, the Monte Carlo simulation avoids these large sums and therefore avoids the problems with rounding errors. Finally, we will extend the Monte Carlo simulations to certain situations, such as where a varying managerial effectiveness parameter is present and multi-dimensional learning, where no analytical expression yet exists. Monte Carlo simulations to prepare for the more complex future analyses.

The process for the Monte Carlo simulation is as follows: once a possible lesson is learned, a subsequent choice of the same lesson would yield no further improvement. The experiment was run several times, with the results averaged. Since the

Table 6.1: Pseudocode for the Monte Carlo Experiments.

variance was smaller for larger sample spaces (and the computational needs of the program were not extensive), most of the experiments were performed with larger sets of improvement values.

The data thus produced were then exported to *Excel* for graphing and data analysis.

6.2 Results

First we present a pseudocode for the simulation algorithm used in the one-dimensional context:

Table 6.1: Pseudocode for the Monte Carlo Experiments.

```
% Initialize lesson set (the possible values of the improvement in the quality metric, $\Delta C_i$)
initialize $\beta$
initialize initial value of the quality metric $\{\Delta C_1, \Delta C_2, ...\Delta C_n\}$
assign a pre-probability weight to lessons by the rule $p_i = e^{-\beta \Delta C_i}$
determine all cumulative sums $s_i = p_1 + p_2 + \cdots + p_i$ of pre-probability weights
    $p_j$, $1 \leq i \leq n$
repeatedly simulate a manufacturing learning experiment (typically 10 times)
    {
    pick a random number $x$, uniformly between 0 and $s_n$
    % find lesson between 0 and $s_n$ associated with the random number $x$
```

```
        loop $i = 1$ to $n$
        {
        If $s_{i-1} < x < s_i$
            then i-th lesson is chosen
        }
    if lesson just found previously learned
        {
        then no change in the quality metric is found
        else, change value of quality metric based on value of lesson learned.
        }
    }
average the learning curves over all experiments
end
```

Experiment 6.1: Below in Figure 6.1 is a graph of the results of a typical simulation of decreasing cost per unit. In this graph, the horizontal axis represents accumulated production; in this case, the number of lessons attempted to be learned. The initial cost was set at \$200, $\beta = 3$, and the ΔC_i were increasing in uniform steps: $\{0, 0.1, 0.2, ..., 6.9, 7.0\}$, that is, $\Delta C_i = .1(i-1), 1 \leq i \leq 71$.

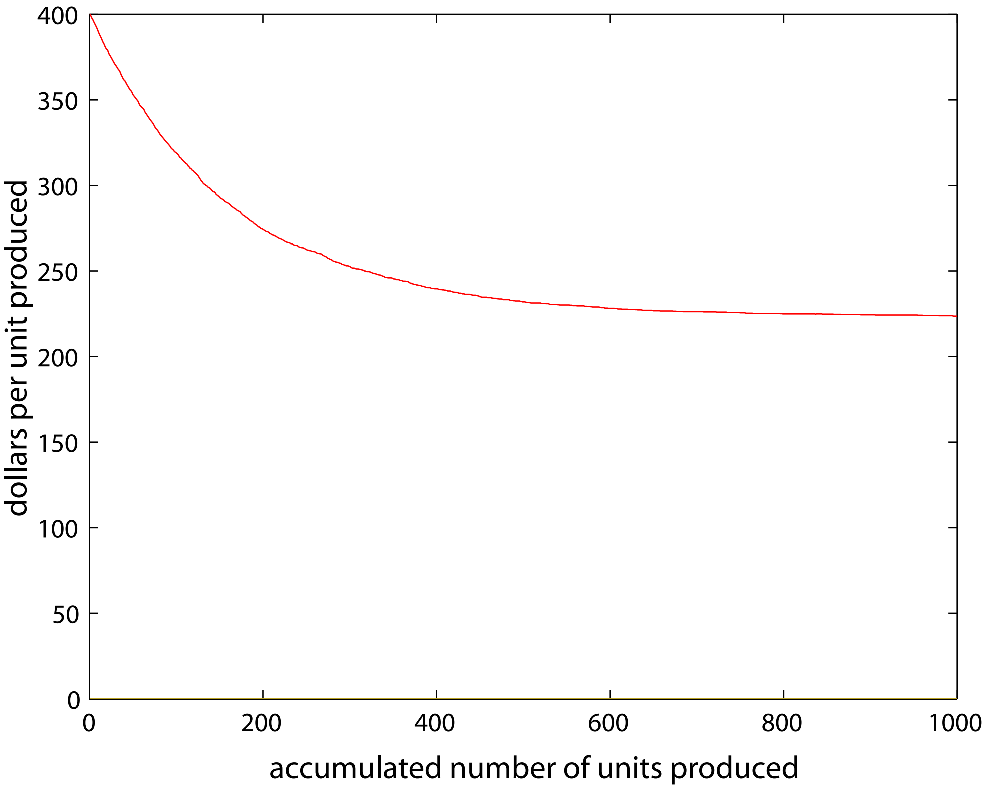

Figure 6.1: Sample learning curve plotting cost per unit produced versus accumulated number of units produced, showing decreasing cost per unit during production.

Figure 6.1 has many of the typical results seen in all such simulations. An analysis of an initial segment indicates that this part is almost perfectly linear. Later on, once the most profitable cost decisions have been taken, the learning rate slows down.

6.2.1 Regression Comparisons

A regression analysis of the data displayed in Figure 6.1 and other, similar curves are fit very well with a power law. For example, the data displayed in Figure 6.1 fits a power law with $R^2 = 0.988$. In fact, as discussed in Chapter 2, the best fit seems to be a two-piece curve, where initial stages of learning are modeled with an exponential law, and later stages are modeled with a power law. This provides again an explanation why some empirical curves are exponential and some are power laws — *the exponential is indicative of a learning process that has a substantial amount of savings decisions still possible, while a power law is indicative of a more mature learning process.*

Experiment 6.2: A suite of simulations was performed to see how varying the efficiency parameter β changed the power law fit. In these experiments the cost of production started at 400 with a plateau of 100. There were 251 lessons to be learned, with individual improvements to the quality metric ranging from 0 to 2.2 in increasing, equally-spaced intervals. Table 6.2 below shows the results for different managerial effectiveness parameters.

Table 6.2: Regression results for learning curves for different managerial effectiveness parameters β.

β	Power Law Exponent	R^2
0	-0.2597	0.9836
0.005	-0.2687	0.9907
0.01	-0.2897	0.9884
0.015	-0.2961	0.9836
0.02	-0.3442	0.9848
0.025	-0.3362	0.9885
0.03	-0.3758	0.9828
0.04	-0.4083	0.9869
0.05	-0.4785	0.976

As can be seen in Table 6.2, the larger β is, the more negative the power law exponent becomes. This demonstrates the general rule discussed above, *increasing β increases the learning rate.*

6.2.2 Comparison of Lesson Structures

Now we compare the results of a unform structure of cost savings to that of a non-uniform structure.

Experiment 6.3: Uniform Structure of Cost Savings.

For this simulation, the following assumptions were in force:

i. The initial cost of production is 350 dollars per unit.

ii. The set of decisions have an associated set of cost savings which are uniformly spaced over an interval. Specifically, the cost savings ranged from 0 to 3 dollars, at increments of 10 cents between the potential cost savings.

iii. The time to implement decisions is approximately constant for all decisions.

Multiple Monte Carlo simulations of the model (6.1)–(6.2) under these assumptions were averaged for Figure 6.2 below. As can be seen from Figure 6.2, the overall shape does indeed match what is commonly seen in a learning curve. Furthermore, the common deviations from the analytic forms of the learning curve are all readily observed. In contrast to the exponential or power law curves, the learning curve in Figure 6.2 exhibits the observed lack of significant upward concavity at the beginning of production. Also, this simulated learning curve can be seen to plateau at a non-zero value, which indicates that neither the power law nor the exponential form are appropriate in the long run.

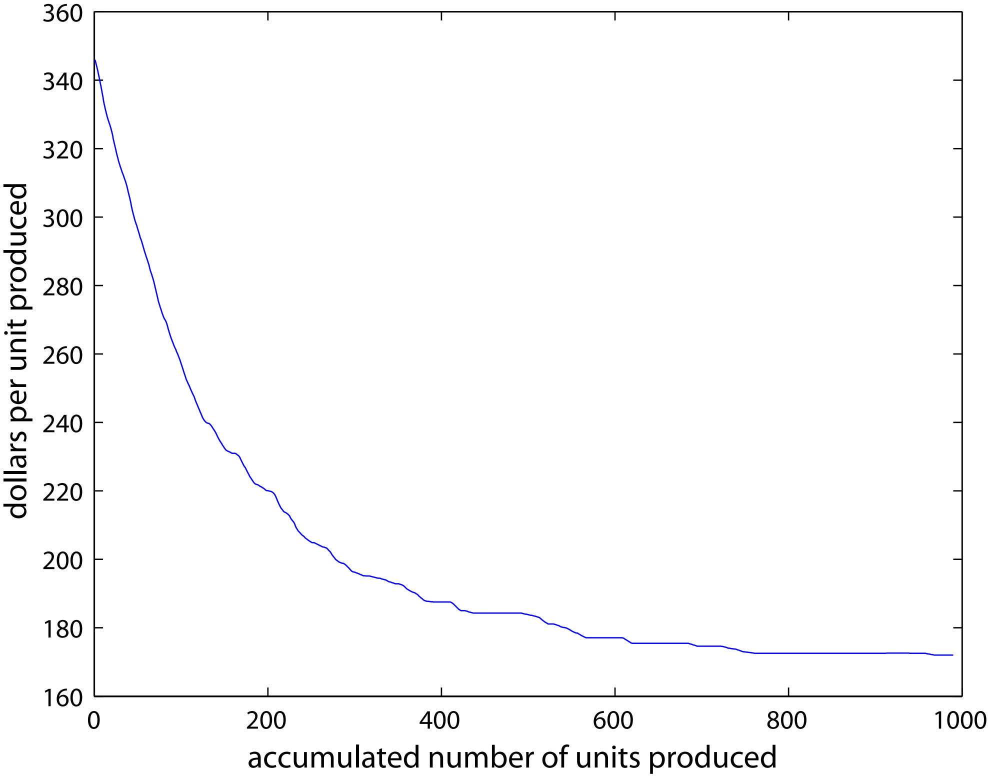

Figure 6.2: Example of simulated learning curve for a linearly increasing cost structure with $\beta = 0.8$.

The results shown in Figure 6.2 were found to fit a power law ($R^2 = 0.960$) much better than an exponential ($R^2 = 0.741$). This R^2 value is well within the bounds of deviation from a power law that actual learning curves exhibit. On the other hand, if the initial cost is adjusted so that the plateau level is much smaller than the initial cost, it is found that an exponential fits the curve much better than a power law.

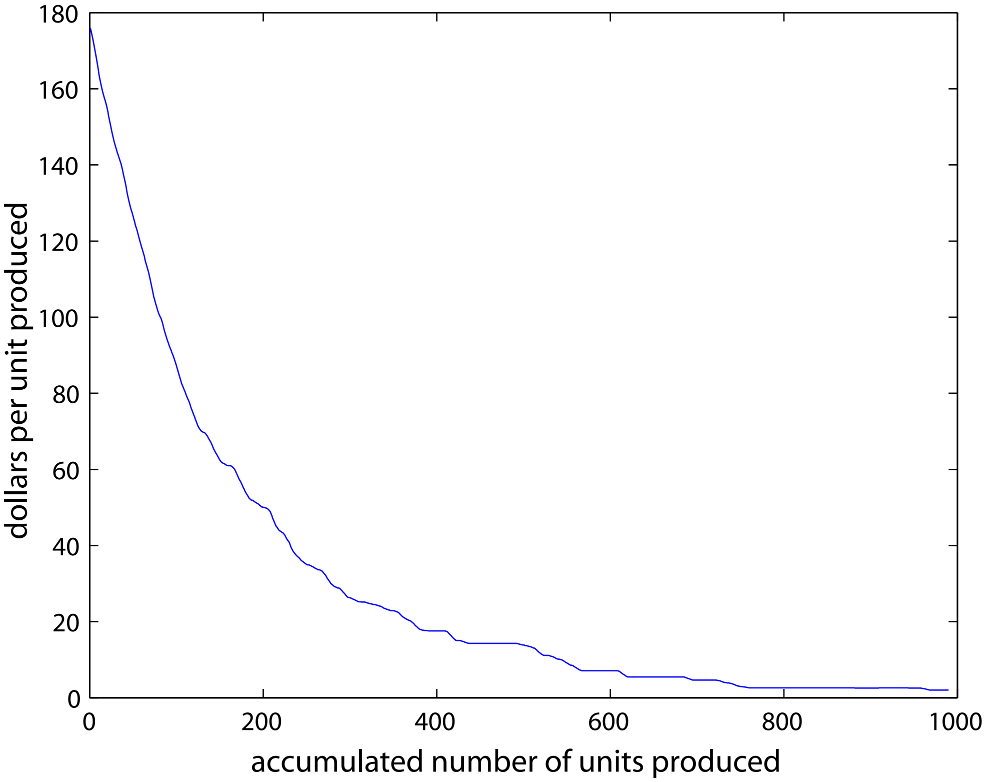

Figure 6.3: Simulated learning curve for a linearly increasing cost savings space with a low plateau, with $\beta = 0.8$.

The R^2 value for an exponential fit to the low-plateau learning data graphed in figure 6.3 is 0.979, while for a power law fit R^2 is 0.932. This result indicates that whether a graph can be better fit by a power law or exponential does not depend on the raw values of the costs savings, but instead depends on how much further cost savings can push down the cost on a percentage basis.

Therefore, for a relatively new industry, where the space of cost savings has only begun to be explored, an exponential learning curve is likely to be observed, while in a more mature industry, a power law decrease is more likely to be observed.

Experiment 6.4: (Non-Uniform Structure of Cost Savings) Let us now simulate a situation with a non-uniform structure of cost savings options, that is, where the cost savings options are not equally spaced, but are instead more concentrated, typically on the lower end of the saving values. In this simulation, it is again assumed that the initial cost of production is 40 dollars. However, in this case, the decisions has associated cost savings which have a cost savings structure such that the spacing between cost savings is inversely proportional to the value of the cost savings. This distribution should give a more realistic model, since in an actual manufacturing setting there will be many decisions that yield a small decrease in cost, while there are only a few possible decisions that will effect a large decrease in cost. Again, the time to implement decisions is roughly constant for all decisions. And again, for the sake of simplicity, the probability of no successful decision is equal to the probability of all decisions previously made.

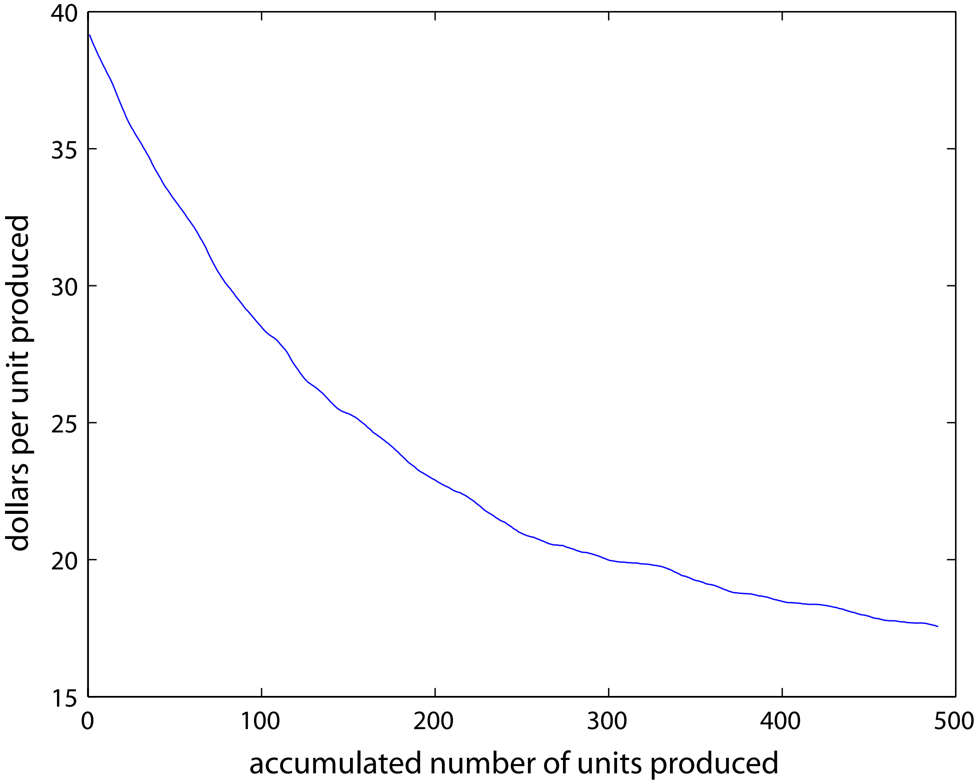

Figure 6.4: Example of simulated learning curve for a non-uniformly spaced cost savings space, with $\beta = 0.8$.

As can be seen in Figure 6.4, the evolution of the learning curve fits a power law exceedingly well. The R^2 value for this fit is 0.9957, while an exponential curve fits the data badly ($R^2 = 0.7531$). However, if the initial cost is adjusted in this case, so that the plateau level is much lower than the initial level, it is again the exponential curve that is a much better fit to the data, as seen in Figure 6.5,

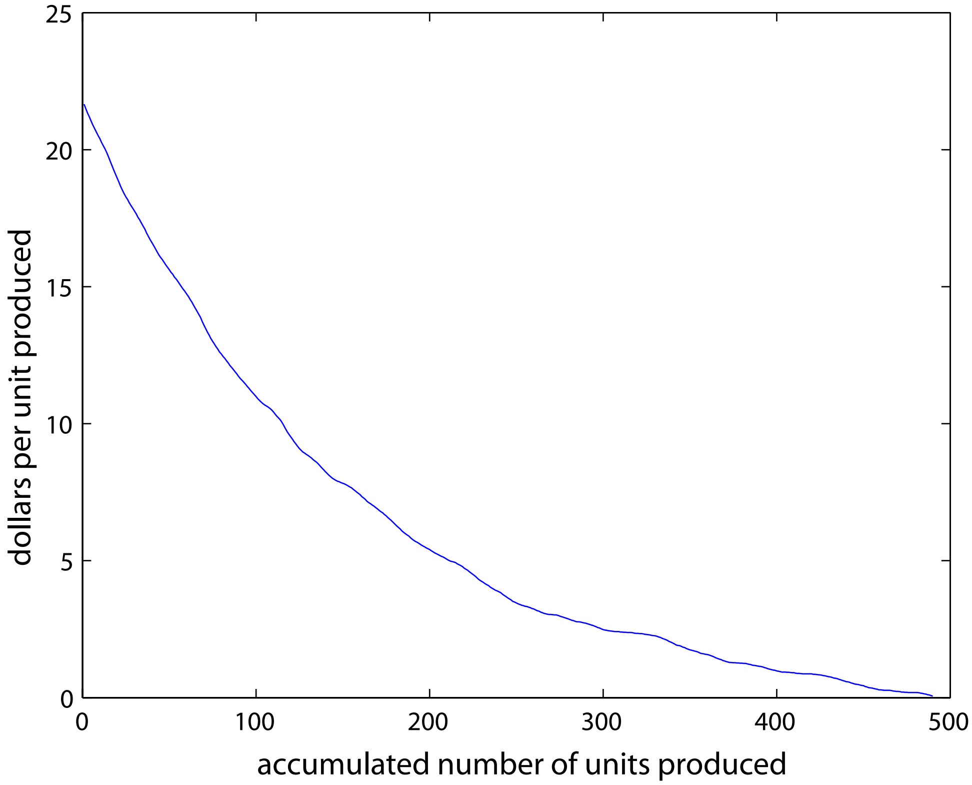

Figure 6.5: Simulated learning curve for a non-uniformly distributed cost savings space with a plateau very small relative to initial cost, with $\beta = 0.8$.

For the small-plateau data exhibited in Figure 6.5 the exponential fit had a R^2 value of 0.9908, while the power law had one of only 0.8956. Therefore, when the cost savings can drive down the cost to a small fraction of the original cost, the exponential curve is a much better fit to the data than the power-law fit.

Therefore, we can see that explicit connections can be made between the structure of the learning curve model (3.26)–(3.27) and observed learning curves. First, the form of the learning curve conforms to the level of learning plateau, relative to the starting point. Also, an explicit relationship can be seen between the managerial effectiveness parameter β and the parameters in the exponential/power law fits. That the model presented herein fit observed learning curves so well serves as a partial validation of the model.

The results of the theory developed in Chapter 3 have proven to explain many of the exact details of real-life learning curves. In contrast to standard learning curve analysis, which treats the exponential and power laws as mere tools of regression, these simulations demonstrate that the form of learning curves is attributable to specific causes in the learning space. The choice of a learning curve form between an exponential and a power law is attributable to the level of the plateau, relative to the initial level of cost. The coefficients in the power law or exponential relate to the structure of the cost savings space; again, rather than being merely outputs of the regression, these parameters are tied to specific causes. The most important results are summarized below.

For a relatively new industry, where the space of cost savings has only begun to be explored, an exponential learning curve is likely to be observed, while in a more mature industry, a power law de-crease is more likely to be observed.

Next we examine the predictions of our theory on multi-dimensional learning.

6.3 Simulations of Multi-Dimensional Learning

This section will examine some variations on the standard learning curve model presented heretofore. First, some simulations of learning with a two-dimensional quality metric will be examined. Next, some simulations with forgetting will be examined.

Recall that for multidimensional learning the probability of learning the i-th lesson with different improvements to the quality metrics Q and R is related to those improvements by

$$p_i \propto e^{\beta_1 \Delta Q_i + \beta_2 \Delta R_i}. \tag{6.3}$$

Experiment 6.5: To keep ideas grounded, we will assume the first quality metric to be cost and the second quality metric to be defect rate. In the first set of simulations, no correlation is assumed between the improvements in the two metrics. Specifically, there are 441 lessons arranged in a uniform square grid, from $(0, 0)$ to $(20, 20)$, where the first coordinate measures the improvement (in dollars) in the manufacturing cost, and the second coordinate measures the improvement (in percent) in the defect rate. Therefore, for example, lesson (1,2) has 1 dollar improvement in the cost and 2 percent defect improvement. In this experiment, then, the lesson set has no correlation between metrics. Also, for this experiment, $\beta_1 = 0.4$ and $\beta_2 = 0.6$.

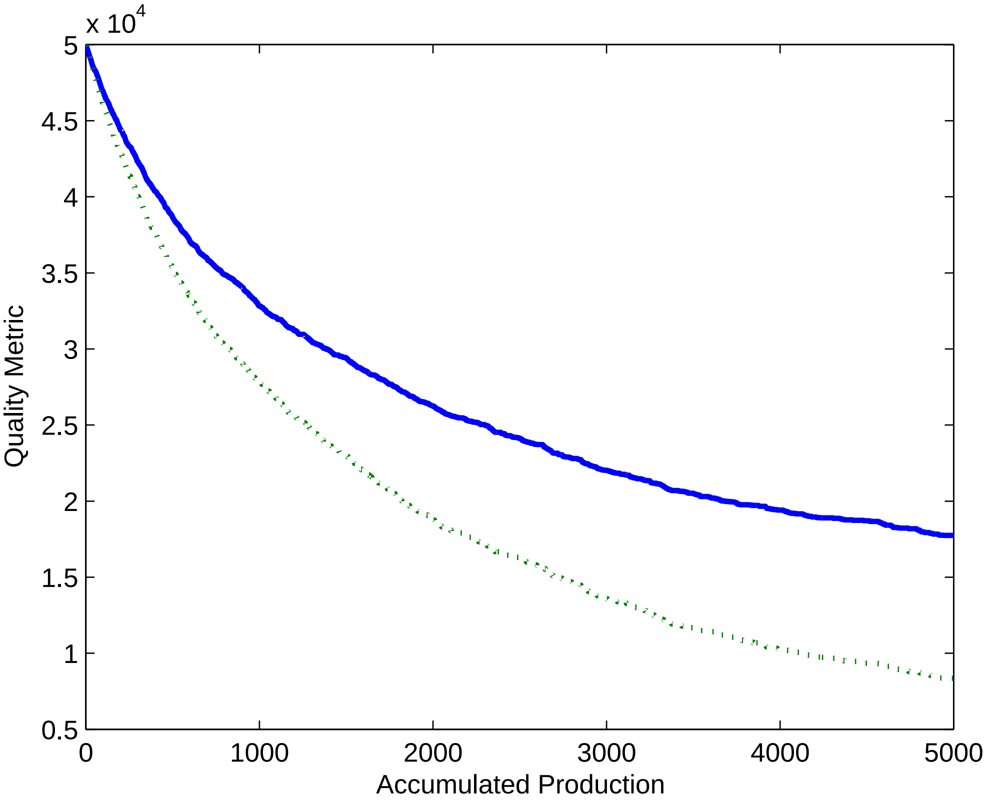

Figure 6.6: Learning curve for two metrics where lesson set has no correlation between metrics (the solid line represents improvement in cost, measured in 10^4 dollars, with $\beta_1 = 0.4$, and the dotted line represents the defect rate (in percent)), with $\beta_2 = 0.6$.

What is apparent from the results in Figure 6.6 is that there is not any trade-off effect shown—both metrics decrease in a similar form. Improvement in one metric does not suffer due to improvement in another metric. However, since cost has the higher β-factor, it experiences the more rapid improvement. As expected, since there is no correlation between payoffs, both cost saving and defect reductions will follow the form of the learning curves obtained previously in Experiments 6.1-6.2.

However, this lack of correlation is unrealistic. More likely is the situation wherein choices which improve cost are less likely to improve defect rate. If one hurries the production line to improve cost, the defect rate is likely to suffer. If one is painstaking enough to remove all defects, cost is likely to suffer. This reasoning points to the fact that there will likely be a negative correlation among the choices between quality improvement and cost improvement.

Experiment 6.6: Next we examine the more realistic situation where the lesson space has an inverse correlation between the two quality metrics is introduced; that is, the larger a lesson improves one quality metric, the less it improves the second quality metric. Below in Figure 6.7 is one such case. In this case, the lessons have cost and defect improvement rate components which lie on the line $\Delta Q_i + \Delta R_i = 10$, so that a lesson which brings a larger cost improvement will have a smaller improvement in the defect rate, and vice versa. Also in this case, the β-factor is much larger for cost ($\beta_1 = 0.8$) than for defect rate ($\beta_2 = 0.2$). In a business setting, this difference in β represents heavy pressure to improve cost, but very little pressure to improve quality.

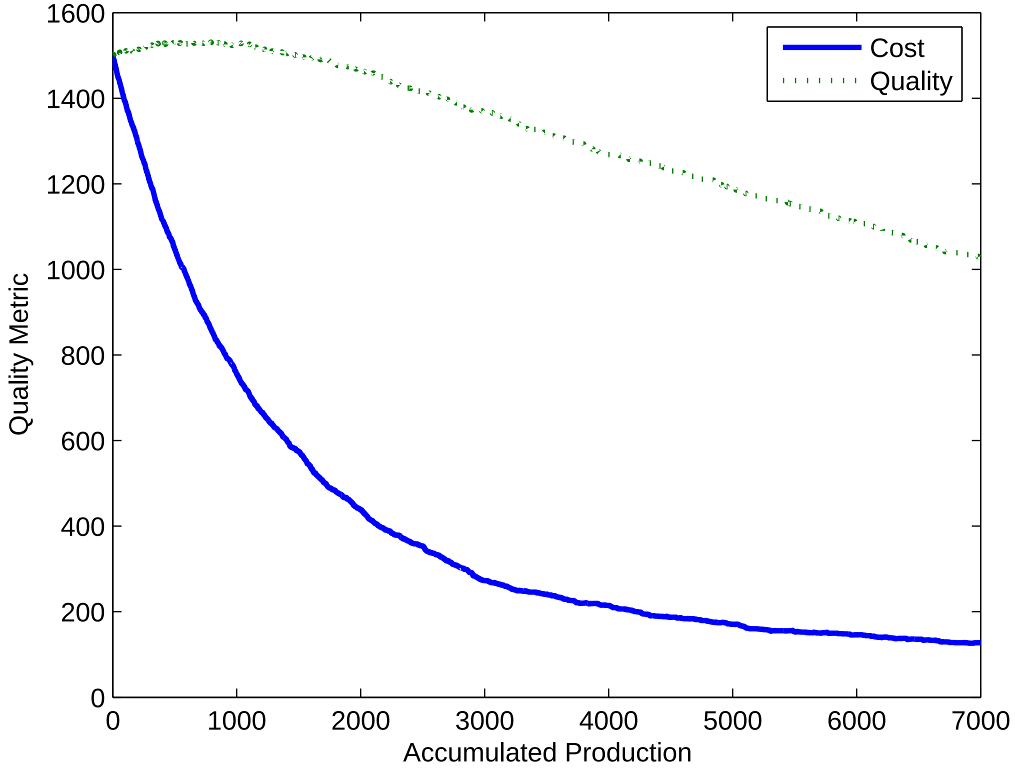

Figure 6.7: Learning curve for two competing metrics with negative correlation (solid line is cost, dotted line is defect rate), with $\beta_1 = 0.8$ and $\beta_2 = 0.2$.

Several characteristics can be observed in Figure 6.7. Initially costs plummet dramatically, while quality is very level. It is only after costs have begun to plateau that any improvement is seen in quality. Even at this stage, the improvement rate for quality is very slow, much slower than is exhibited with cost. The relative weightings of the β factors has forced the cost improvements to occur first, and this result is borne out in the graph.

Experiment 6.7: Contrast this previous simulation with a simulation shown in Figure 6.8 with the same lesson space but where the two quality metrics are instead treated equally, i.e., they have the same β values ($\beta_1 = \beta_2 = 0.08$):

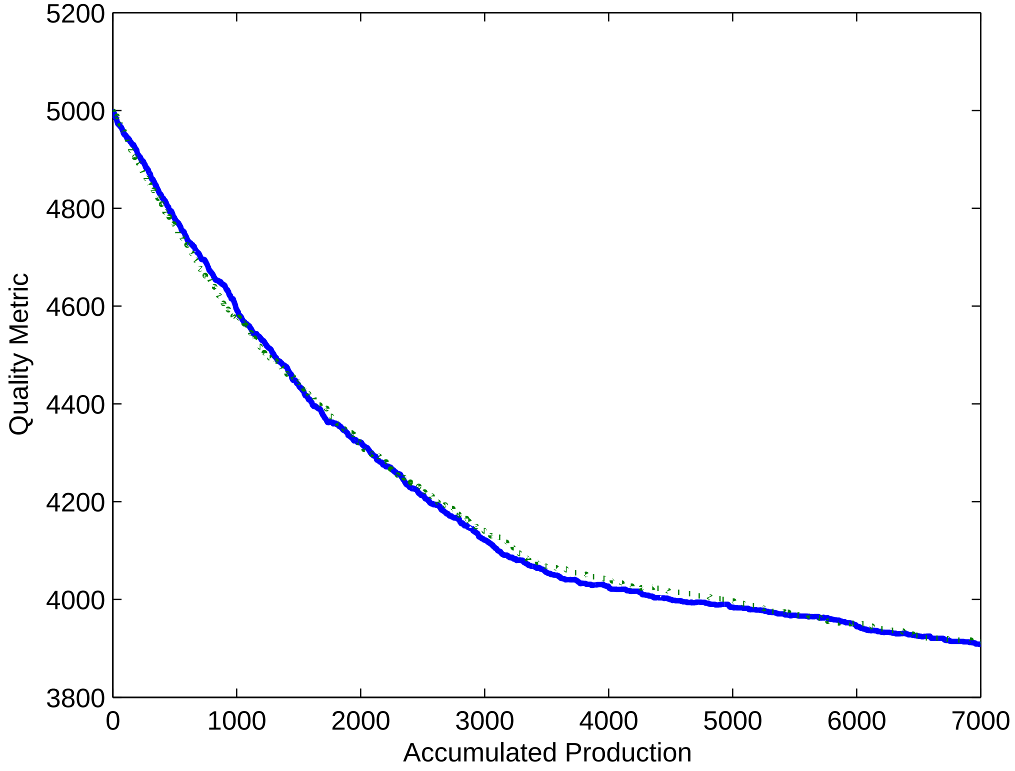

Figure 6.8: Learning Curve for two equally preferred metrics.

As should be expected, the two quality metrics improved at almost exactly the same rate.

Experiment 6.8: Next we examine the case where a change in emphasis between two quality metrics occurs in the middle of the production process—see Figure 6.9. In this case there is initially a strong preference for improvement in the cost metric ($\beta_1 = 1.5$, $\beta_2 = 0.1$). Also present here are lessons which greatly improve one metric, while actually making the other metric worse. In this case, as should be expected, the cost metric initially improves greatly, while little improvement occurs for the quality metric. Midway through production, the preferences are reversed ($\beta_1' = 0.1$, $\beta_2' = 1.5$). Cost then becomes worse, while quality suddenly improves greatly.

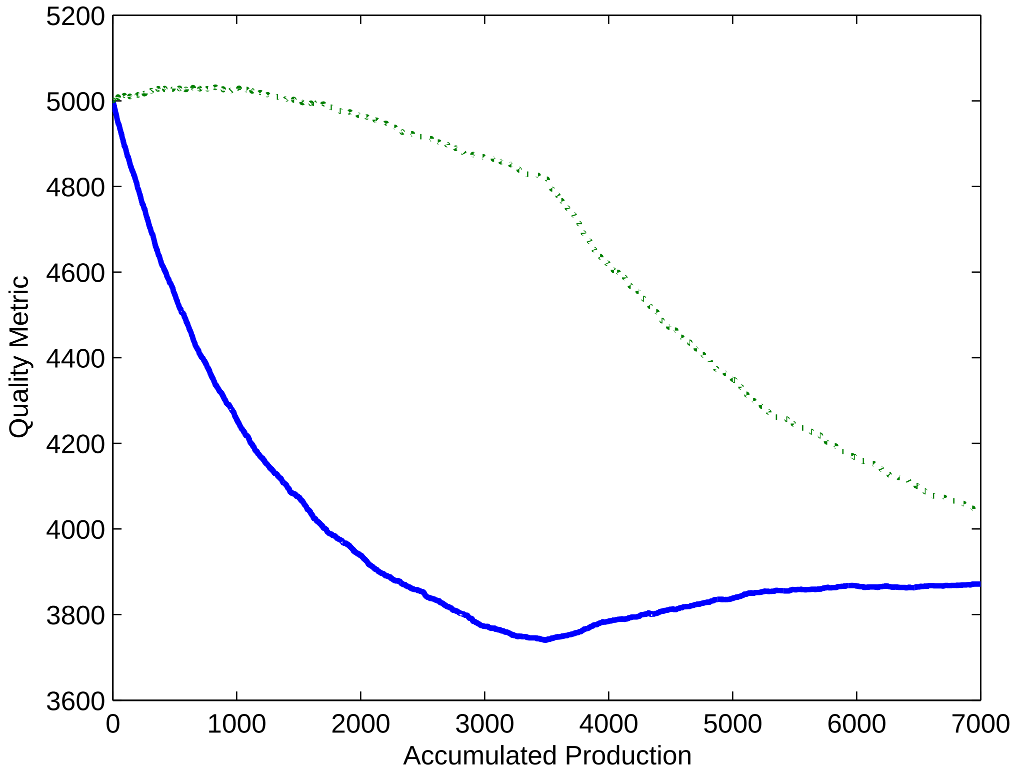

Figure 6.9: Learning Curve for two competing metrics with switch between preference for quality metrics.

Experiment 6.9: Now we can look at the when the two quality metric are initially weighted the same ($\beta_1 = \beta_2 = 0.08$), but in the middle of the production process a decision is made to weight one much more heavily than the other ($\beta_1 = .15$, $\beta_2 = 0.01$). Below in Figure 6.10 is a graph of the results:

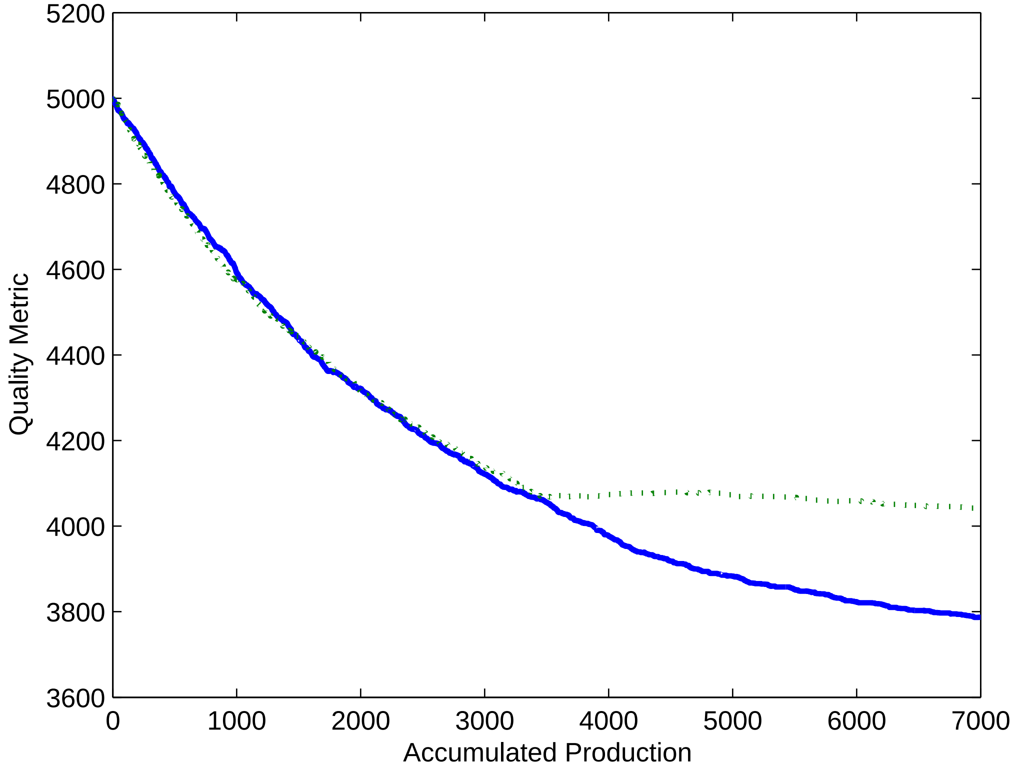

Figure 6.10: Learning Curve for Two Competing Metrics with Switch from no Preference to Strong Preference for Quality Metrics.

We can see from Figure 6.10 that the switch causes one of the quality metric to improve slightly more rapidly, while the second one is allowed to stagnate.

6.4 Conclusion

What the above graphs have shown is that the gross form of the predicted learning curves matches the learning curves seen in manufacturing. Like stated previously, our learning curve theory explains the observed data in existing learning curves. As the previous section describes, this theory is also very useful in exploring an area which the literature barely touches—multidimensional learning. The usefulness of the theory in explaining any possible types of learning goes a long way to validate the theory, as well as merit further explorations into where the theory can go.

Furthermore, the theory has the capability to explain many more details of manufacturing learning curves than any previous theory. The effect of managers is captured in a single parameter β. The theory can explain whether and to what extent the learning in two different metrics interacts. Finally, it gives a testable equation for the evolution of learning. The testing of the theory will be explored in the first section of the next chapter.

Chapter 7

Further Directions

This thesis has laid out a complete analytical theory for the learning curve, as well as some practical mathematical formulations for understanding the structure of the problem. In this final chapter a few further directions for the ideas in this paper will be proposed.

7.1 Empirical Analysis of Learning Curves

Empirical validation of the learning curve results described herein is an important goal. Several factors are important in considering the effectiveness of such an effort, such as data, assumptions of choice space, and choice of statistical model.

First, the data problem is dire. Without reliable data, no proposed model can be verified. Many of the published studies with learning curve time series have a very small number of data points, which makes fitting data to a model problematical. Fortunately, recent data-tracking techniques which allow a real-time tracking of internal cost and price data make this scarcity of data problem less paramount. For example, managerial dashboards allow managers to track data for production cost and production numbers on a daily basis [Irgens 2008]. The data problem is likely to be less of an issue as more companies compete with such analytical tools.

Next, we must examine the choice of model. Should we use a discrete or continuous model? If discrete, how many lessons do we provide for? If continuous, what labeling subset of the reals do we use, and what will be the cumulative cost savings $C(x)$? At first glance, these problems might seem insurmountable. However, recall that one of the properties for learning with random sets was *infinite additivity.* This property allows the choice probability of an aggregate lesson to be directly related to the probabilities of each of the sub-lessons. If we were to use a discrete case, then, to some extent it should not matter how many possible lessons we choose. It will only affect the result to the extent that the variance of the stochastic process would be larger or smaller, depending on the size of the lesson space.

Finally, we briefly examine the possible statistical methods available for verifying our model. For the discrete case, a regression is sufficient. Specifically, we will have the cost-per-unit function is given by

$$\overline{C}(k) = C_0 + \sum_{i=1}^{n} \big(1 - (1 - e^{k\beta \Delta C_i})\big) \Delta C_i, \tag{7.1}$$

where k is the accumulated production and n is the number of cost lessons assumed. The values which are needed to be found through regression are β and the ΔC_i. Therefore, for n lessons, $n + 2$ values must be found through regression.

The continuous case has some very interesting aspects. Recall that the form of the continuous learning curve is given by

$$\begin{aligned}
\overline{\Delta C}(t) &= \int_G (1 - e^{-\lambda t f(c(x))})\, dC(x) && (7.2)\\
&= \int_G (1 - e^{-\lambda t f(c(x))})\, c(x)dx && (7.3)\\
&= \int_G c(x)dx - \int_G e^{-\lambda t f(c(x))}\, c(x)dx. && (7.4)
\end{aligned}$$

If we choose f so that $f(c(x)) = x$ and define $t' = \lambda t$ we obtain

$$\overline{\Delta C}(t) = \int_G c(x)dx - \int_0^\infty e^{-xt'}\, c(x)\chi_{G(x)}dx. \tag{7.5}$$

The second integral can be immediately recognized as a Laplace Transform. Thus, if the cost is related to a Laplace Transform of the cost space, the cost space may be obtained via an inverse Laplace Transform. It is known that numerically inverting a Laplace Transform is very difficult [Cohen 2007] Nonetheless, some interesting aspects of this problem may still be explored.

7.2 Further Directions for Exploring Dynamic Random Sets

This section sketches some applications of the dynamic random sets constructed in Chapter 4. The example given earlier of raindrops falling on a sheet of paper is prototypical. Although the theory described in chapter 4 is in a one-dimensional setting, it may be easily generalized to higher dimensions. In this case, *f(**x**)* becomes the probability distribution of a raindrop hitting a point $\boldsymbol{x}$ on the sheet of paper. The paper is initially dry, but over time, the measure of “wetness” of the sheet asymptotically approaches the measure of entire sheet, as it becomes more and more likely that each point has been covered over time.

This random set process can also be viewed as a random growth model. For example, [Holash et al. 1999] used random sets created via a Poisson model to model tumor growth. As a Poisson process, however, the growth model was discontinuous. Our model has no such discontinuities.

The deterministic concept of a weak convergence of sets can also be used to model the process of *coarse graining* which is pervasive in statistical mechanics—see for example [Arnold Avez 1968]. Coarse graining is the method by which one examines the average properties within small subsets called grains, rather than the properties at each point. In the example of Arnold and Avez, if we stir vigorously a solution of 20 percent rum and 80 percent $^{\mathrm{TM}}$Coke, every part of the solution will consist of 20 percent rum and 80 percent $^{\mathrm{TM}}$Coke. Even though at some level the components might not be completely mixed, such a level is too small to be relevant. This process of coarse graining is exactly what is being given in the weak convergence of deterministic sets and is very similar given in the example of the sequence of deterministic sets given in section II above.

There are many possible generalizations of the principal dynamic random set. The strongest assumption we have imposed is the independence of the n intervals which form each term of the sequence, the weak limit of which is the example process. Many growth models would likely have some dependence between the intervals. Also, as is evident from the description above, many possible applications are available for a modeler.

Chapter 8

Bibliography

Argote L. and Epple D., Learning Curves in Manufacturing, Science 247 (1990) 920-924.

Arnold V. I. and Avez A. (1968) Ergodic Problems in Statistical Mechanics, W. A. Benjamin, Inc., New York.

Bailey, C., Forgetting and the Learning Curve: A Laboratory Study, Management Science 35 (1989) 340-352.

Barney J., The debate between Traditional Management Theory and Organizational Economics: Substantive Differences or Intergroup Conflict? Academy of Management Review 15 (1990) 383-393.

Binney J. and Tremaine S. (1988) Galactic Dynamics. Princeton.

Capasso V. and Villa E., Continuous and Absolutely Continuous Random Sets, Stochastic Analysis and Applications 24 (2006) 381-397.

Cohen A. (2007) Numerical Methods for Laplace Transform Inversion. Springer, New York.

Conway R. and Schultz A., The Manufacturing Progress Function, Journal of Industrial Engineering 10(1) (1959) 39-53.

Demsetz H., The Firm in Economic Theory: a Quiet Revolution, American Economic Review 87 (1997) 426-429.

Dillard J., Riggsby, J. T. et al., The Making and Remaking of Organization Context: Duality and the Institutionalization Process, Accounting, Auditing and Accountability Journal 17 (2004) 506-542.

Donaldson L., The Ethereal Hand: Organizational Economics and Management Theory, Academy of Management Review 15 (1990) 369-381.

Donaldson L., A Rational Basis for Criticisms of Organizational Economics: A Reply to Barney, Academy of Management Review 15 (1990) 394-401.

Engel A. and Van den Broeck C. (2001) Statistical Mechanics of Learning, Cambridge.

Feynman, R. P. (1972) Statistical Mechanics. Addison Wesley, New York.

M. Gallegati, S. Keen, T. Lux and P. Ormerod, Worrying Trends in Econophysics, Physica A 370 (2006) 1-6.

Garg, A. A. and Milliman, P., The Aircraft Progress Curve Modified for Design Changes, J. Industrial Engineering 12(1) (1961) 23-27.

Griffiths R. B., Nonanalytic Behavior Above the Critical Point in a Random Ising Ferromagnet, Phys. Rev. Let., 23 (1969) 17-19.

Hastie T. and Tibshirani R. and Friedman J. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Haussler et al., Rigorous Learning Curve Bounds from Statistical Mechanics. Machine Learning, 25(2) (1986) 195-236.

He Zi-Linp and Wong Poh-Kam, Exploration vs. Exploitation: An Empirical Test of the Ambidexterity Hypothesis, Organization Science 15(4) (2001) 481-494.

Holash J. and Wiegand S. J., Yancopoulos G. D., New model of tumor angiogenesis: dynamic balance between vessel regression and growth mediated by angiopoietins and VEGF, Nature 18(38) (1999) 5356-5362.

Irgens et al., Optimization for Operational Decision Support: The Rig Management Case. SPE Annual Technical Conference and Exhibition (2008) 116616-MS.

Jaber M. Y. and Bonney M., Production Breaks and the Learning Curve: The

Forgetting Phenomenon, Applied Mathematical Modelling 20(2) (1996) 162-169.

Kodama R. H. and Berkowitz A. E., Atomic-Scale Magnetic Modeling of Oxide Nanoparticles. Physical Review B 59(9) (1999) 6321-6336.

Levy, F. K., Adaption in the Production Process, Management Science 11(6) (1965) 136-154.

March, J. G. (1964) A Primer on Decision Making, Macmillan, New York.

March, J. G., Exploration and Exploitation in Organizational Learning. Organization Science 2(1) (1991) 71-87.

Miller K. D., Zhao M., and Calantone R., Adding Interpersonal Learning and Tacit Knowledge to March's Exploration-Exploitation Model. Academy of Management Journal 49(4) (2006) 709-722.

Muth, J., Search Theory and the Manufacturing Progress Function. Management Science 32(8) (1986) 948-962.

Robbins H. E., On the Measure of a Random Set, Annals of Mathematical Statistics 15 (1944) 70-74.

Russell, Stuart; Norvig, Peter (2002) Artificial Intelligence: A Modern Approach, Prentice-Hall, New York.

Schilk T. (2000) Molecular Modeling and Simulation. Springer, New York.

Seung, H. S., Statistical Mechanics of Learning from Examples, Physical Review A 45(8) (1992) 6056-6091.

Shannon C. E., A Mathematical Theory of Communication, Bell Systems Technical Journal, 27 (1948) 379-423; 623-656.

Speaker P., Using Random Sets to Model Learning in Manufacturing, submitted to European Journal of Operations Research

Speaker P. and MacCluer C. R., A Distributional Construction of Random Sets, submitted SIAM Journal of Applied Mathematics.

Sperduti A. and Starita A., Supervised Neural Networks for the Classification of Structures, IEEE Transactions on Neural Networks 8(3) (1997) 714-735.

Tushman, M. L. and Anderson, P., Technological Discontinuities and Organizational Environments, Administrative Science Quarterly 31(3) (1986) 439-465.

Watkin T., Rau A., and Biehl M., The Statistical Mechanics of Learning a Rule, Reviews of Modern Physics 65(2) (1993) 499-556.

Witten, T. A., Diffusion-Limited Aggregation, a Kinetic Critical Phenomenon, Physics Review Letters 47, (1981) 1400-1403.

Yelle L. E., The Learning Curve: A Comprehensive Survey, Decision Sciences 10 (1979) pp. 302-328.

Zangwill, W. I. and Kantor, P. B., Toward a Theory of Continuous Improvement and the Learning Curve. Management Science 44(7) (1998) 910-920.

Zangwill, W. I. and Kantor, P. B., The Learning Curve: A New Perspective. International Transactions in Operations Research 7(6) (2000) 595-607.