

Effects of Feedback Delay on Learning

Hazhir Rahmandad

Post-doctoral Associate, Sloan
School of Management, M.I.T
E53-364A, 30 Wadsworth Ave.,
Cambridge, MA 02142
617-253-3865; hazhir@mit.edu

Nelson Repenning

Associate Professor, Sloan
School of Management, M.I.T
E53-335, 30 Wadsworth Ave.,
Cambridge, MA 02142
617-258-6889; nelson@mit.edu

John Sterman

Professor, Sloan School of
Management, M.I.T
E53-351, 30 Wadsworth Ave.,
Cambridge, MA 02142
617-253-1951; jsterman@mit.edu

Abstract

Learning figures prominently in many theories of organizations. Understanding barriers to learning is therefore central to understanding firms' performance. This essay investigates the role of time delays between taking an action and observing the results in impeding learning. These delays, ubiquitous in real-world settings, can introduce important tradeoffs between the long-term and the short-term performance. In this essay, four learning algorithms, with different levels of complexity and rationality, are built and their performances in a simple resource allocation task are analyzed. The study focuses on understanding the effect of time delays on learning. Simulation analysis shows that regardless of the level of rationality of the organization, misperceived delays can impede learning significantly.

Introduction

Learning figures prominently in many perspectives on organizations (Cyert and March 1963). Organizational routines and capabilities fundamental to firm performance are learnt through an adaptive process of local search and exploration (Nelson and Winter 1982). Researchers have highlighted learning as an important source of competitive advantage (DeGeus 1988; Senge 1990), central to the firm's performance in high-risk industries (Carroll, Rudolph et al. 2002), and underlying many efficiency gains by firms (Argote and Epple 1990). Therefore understanding the functional and dysfunctional effects of learning processes are vital to understanding firms' performance.

Management scholars have studied both positive and negative aspects of learning. On the one hand, this literature portrays learning as an adaptive process (Cyert and March 1963), which brings useful knowledge (Huber 1991), increases efficiency (Argote and Epple 1990), and is

central to the success of organizations in the fast-changing world (DeGeus 1988; Senge 1990). At the individual level, it is argued that learning helps individuals find the best strategies, and therefore, this argument is used to justify the individual rationality assumption in economics (Friedman 1953).

On the other hand, if learning mechanisms were always functional and could find the optimum strategies fast enough¹, all organizations in similar market niches would acquire similar strategies and capabilities. Therefore, to understand heterogeneity in a population of firms, we need to understand the failure and path-dependence of different learning processes, along with the population-level dynamics of selection and reproduction. Organizational learning scholars have studied several factors that hinder learning or result in different outcomes for learning, including payoff noise, complexity, defensive routines, and delays.

Noise and ambiguity in the payoff that an organization receives increase the chance of superstitious learning (Argyris and Schön 1978; Levitt and March 1988), where the organization draws wrong conclusions from its experience. Moreover, stochasticity in results creates a bias against more uncertain strategies for which an unlucky first experience prohibits future experimentation (Denrell and March 2001).

Another challenge to functional learning lies in the interaction between the competency evolution and the search process. One can conceptualize learning as a process of (virtual or actual) experimentation by an agent (or by other agents in the case of vicarious learning), observing the resulting payoff, and adjusting (mental models and) actions to enhance performance. In this framework, each alternative's payoff not only depends on the inherent promise of that alternative, but also depends on the agent's experience with that alternative. As a result, initially-explored alternatives pay off better, regardless of their inherent promise, and therefore they close the door of experimentation to other, potentially better, alternatives (Levinthal and March 1981; Herriott, Levinthal et al. 1985). Such competency traps (Levitt and March 1988) also inhibit the organization from exploring better strategies once the current routines of the firm become obsolete in face of changing environment, thereby turning core capabilities into core rigidities (Leonard-Barton 1992).

¹ Here the speed of learning should be compared to the speed of environmental change and processes of selection and reproduction.

Complexity of the payoff landscape on which organizations adapt raises another challenge to organizational learning. The complementarities among different elements of a firm's strategy (Milgrom and Roberts 1990) result in a complex payoff landscape in which improvements in performance often come from changing multiple aspects of the strategy together. In other words, after some local adaptation to find an internally consistent strategy, incremental changes in one aspect of the strategy are often not conducive to performance gains, so that the firm rests on a local peak of a rugged payoff landscape (Busemeyer, Swenson et al. 1986; Levinthal 1997). Even though discontinuous change can take the organization out of such local peaks (Tushman and Romanelli 1985; Lant and Mezias 1990; March 1991; Lant and Mazias 1992), this spatial complexity of payoff landscape can indeed contribute to the observed firm heterogeneity (Levinthal 2000; Rivkin 2000). Further studies in this tradition have shed light on the challenges of imitation and replication in the presence of spatial complexity (Rivkin 2001), the adaptation processes empirically used by organizations under these conditions (Siggelkow 2002), and the role of cognitive search in adaptation on rugged landscapes (Gavetti and Levinthal 2000).

Individuals play a central role in organizations; therefore cognitive biases and challenges for individuals' learning have important ramifications for organizational learning. Studies have shown how dynamic complexity and delays can significantly hamper decision-making (Sterman 1989; Paich and Sterman 1993; Diehl and Sterman 1995) and result in self-confirming attribution errors in organizations (Repenning and Sterman 2002). Moreover, the activation of individual defensive routines closes the communication channels in the organization, creates bias in intra-organizational information flows, seals off flawed theories-in-practice from external feedback, and promotes group-think (Argyris and Schön 1978; Janis 1982; Argyris 1999).

Temporal challenges to learning such as delays between action and payoff have received little attention in organizational learning research. In contrast, as early as in the 1920's the importance of temporal contiguity of stimulus and response was recognized by psychologists studying conditioning (Warren 1921). However, despite evidence of adverse effects of delays on individual learning (Sengupta, Abdel-Hamid et al. 1999; Gibson 2000) and some case-specific evidence at the organizational level (Repenning and Sterman 2002), with a few exceptions (Denrell, Fang et al. 2004), the effects of delays on organizational learning have not been studied. In fact, few formal models of organizational learning capture such delays explicitly.

Denrell and colleagues (2004) build on the literature of learning models in artificial intelligence to specify a Q-learning (Watkins 1989) model of a learning task in a multi-dimensional space with a single state with non-zero payoff. Consequently the learner needs to learn to value different states based on their proximity to the goal state. Their analysis informs the situations in which several actions with no payoff need to be taken before payoff is observed. This setting is conceptually similar to having a delay between an action and the payoff where the intermediate states that the system can occupy are distinguishable by the decision-maker and lead to no payoff. The study therefore highlights the importance of information about such intermediate states and elaborates on the central problem of credit assignment, how to build associations between actions and payoff in the presence of delays. This study, however, does not discuss the existence of payoff values in the intermediate states, does not consider the tradeoffs between short-term and long-term actions central to real-world delay effects, and uses an information-intensive learning algorithm that requires thousands of action-payoff cycles to discover a short path to a single optimum state.

We expand these results by examining learning in a resource allocation setting that resembles the real-world challenges of an organization. We study how much performance and learning weaken as a result of delays, how sensitive the results are to the rationality of the decision-maker, and what the mechanisms are through which delays impede learning. A formal model enables us to pin down the effects of time delays on learning by removing other confounding factors that potentially influence learning. Specifically, we will remove the effects of feedback noise, multiple peaks and competency traps, discount rate, interpersonal dynamics, and confounding contextual cues on learning. Running controlled experiments on the simulation model, we can also gain insight into the mechanisms through which delays impede learning. Moreover, using a formal model, we are able to run a large number of controlled experiments and investigate them closely at a relatively cheap cost, which improves the internal validity of the study compared to using real experiments. Nevertheless, using simulations raises the question of the external validity of the model. Specifically, it is not clear how closely our algorithms of learning correspond to human learning patterns or organizational learning practices, and therefore how well the results can be generalized to real social systems.

To meet this challenge we draw on the current literature on learning in organizational behavior, psychology, game theory, and attribution theory to model learning, and investigate four

different learning procedures with different levels of rationality, information processing requirements, and cognitive search ability. By including these different algorithms, we can distinguish the behavioral patterns that are common across different algorithms and therefore derive more valid conclusions about the processes that extend to real social systems. Moreover, using these different algorithms, we can differentiate the effects of decision-making rationality from learning problems arising from delayed feedback.

Our results confirm the hypothesis that delays complicate the attribution of causality between action and payoff and therefore can hinder learning (Einhorn and Hogarth 1985; Levinthal and March 1993). Furthermore, our analysis shows that performance can still be significantly sub-optimal in very easy learning tasks, specifically, even when the payoff landscape is unchanging, smooth, and has a unique optimum. Moreover, the organization can learn to believe that the sub-optimal performance is really the best it can do. Finally, the results are robust to different learning algorithms and the difficulty of learning in the presence of delays appears to be rooted in the misperception of the delays between actions and outcomes, rather than the specific learning procedures we examined.

These results suggest interesting research opportunities, including extending these algorithms; exploring interaction of delays with feedback noise, perception lags, and ruggedness of landscape; investigating the possibility of learning about the structure of time delays; and explaining empirical cases of learning failure. Moreover, the results have practical implications in the design of accounting and incentive systems as well as evaluation of firm performance.

In the next section, we describe the model context and structure and discuss the different learning procedures in detail. The “Results and Analysis” section presents a base case demonstrating that all four learning procedures can discover the optimum allocation in the absence of action-payoff delays. Next we analyze the performance of the four learning procedures in the presence of action-payoff delays, followed by tests to examine the robustness of these results under different parameter settings. We close with a discussion of the implications, limitations, and possible extensions.

Modeling learning with delayed feedback

Resource allocation problems provide an appropriate context for studying the effect of time delays on learning. First, many important situations in multiple levels of analysis involve allocating resources among different types of activities with different delays between allocation and results. Examples include an individual allocating her time between work and education, a factory allocating resources between production, maintenance, and process improvement, and a government allocating budget between subsidies, building roads, and building schools.

Moreover, these situations often involve tradeoffs between short-term and long-term results and therefore raise important questions about the short-term vs. long-term success of different social entities. For example, the factory can boost output in the short run by cutting maintenance, but in the long run, output falls as breakdowns increase. Other examples include learning and process improvement (Repenning and Sterman 2002), investment in proactive maintenance (Allen 1993), and environmental degradation by human activity (Meadows and Club of Rome 1972). These tradeoffs suggest that individuals, organizations, and societies can often fail to learn from experience to improve their performance in these allocation and decision-making tasks.

Previous studies motivated formulation of our model. Repenning and Sterman (2002) found that managers learned the wrong lessons from their interactions with the workforce as a result of different time delays in how various types of worker activity influence the system's performance. Specifically, managers seeking to meet production targets had two basic options: (1) increase process productivity and yield through better maintenance and investment in improvement activity; or (2) pressure the workforce to “work harder” through overtime, speeding production, taking fewer breaks, and, most importantly, by cutting back on the time devoted to maintenance and process improvement. Though the study found, consistent with the extensive quality improvement literature, that “working smarter” provides a greater payoff than working harder, many organizations find themselves stuck in a trap of working harder, resulting in reduced maintenance and improvement activity, lower productivity, greater pressure to hit targets, and thus even less time for improvement and maintenance (Repenning and Sterman 2002).

In parallel to this setting, our model represents an organization engaged in a continuous-time resource allocation task. The organization must allocate a fixed resource among different activities. The payoff generated by these decisions can depend on the lagged allocation of resources, and there may be different delays between the allocation of resources to each activity

and its impact on the payoff. The organization receives outcome feedback about the payoff and past resource allocations, and attempts to learn from these actions how to adjust resource allocations to improve performance. As a concrete example, consider a manufacturing firm allocating the organizational resources (e.g., employee time) among three activities: production, maintenance, and process improvement. These activities influence production, but with different delays. Time spent on production yields results almost immediately. There is a longer delay between a change in maintenance activity and machine uptime (and hence production). Finally, it takes even longer for process improvement activity to affect output.

The organization gains experience by seeing the results of past decisions and seeks to increase production based on this experience. It has some understanding of the complicated mechanisms involved in controlling production captured in the mental models of individual members and in organizational routines, and it may be aware of the existence of different delays between each activity and observed production. Consequently, when evaluating the effectiveness of its past decisions, the organization takes these delays into consideration (e.g., it does not expect last week's process improvement effort to enhance production today). However, the organizational mental model of the production process may be imperfect, and there may be discrepancies between the length of the delays it perceives and the real delays.

1- Allocation and payoff

The organization continuously allocates a fraction of total resources to activity j of m possible activities at time t , $F_j(t)$ ² where:

$$\sum_j F_j(t) = 1 \quad \text{For } j:1,\dots,m \quad (1)$$

In our simulations we assume $m = 3$ activities, so the organization has two degrees of freedom. Three activities keep the analysis simple while still permitting different combinations of delay and impact for each. Total resources, $R(t)$, are assumed to be constant so $R(t) = R$, and resources allocated to activity j at time t , $A_j(t)$ are:

$$A_j(t) = F_j(t) * R \quad (2)$$

² The model is formulated in continuous time but simulated by Euler integration with a time step of 0.125 period. Sensitivity analysis shows little sensitivity of the results to time steps < 0.2 .

These allocations influence the payoff with a delay. The payoff at time t is determined by the Effective Allocation, $\hat{A}_j(t)$, which can lag behind A_j with a payoff generation delay of T_j . We assume a fixed, non-distributed delay for simplicity³:

$$\hat{A}_j(t) = A_j(t - T_j) \quad (3)$$

The payoff generation delays are fixed but can be different for each activity.

In our manufacturing firm example, R is the total number of employees, $F_j(t)$ is the fraction of the fraction of people working on activity j (j : producing, maintenance, process improvement) and $A_j(t)$ is the number of people working on activity j . In our example, the delays in the impact of these activities on production (the payoff) are presumably ordered approximately as $0 \approx T_{production} < T_{maintenance} < T_{improvement}$.

For simplicity we assume the payoff, $P(t)$, to be a constant-returns-to-scale, Cobb-Douglas function of the effective allocations, which yields a smooth landscape with a single peak, allowing us to eliminate learning problems that arise in more complicated landscapes :

$$P(t) = \prod_j \hat{A}_j(t)^{\alpha_j} \quad , \quad \sum_j \alpha_j = 1 \quad (4)$$

2- Perception delays

The organization perceives its own actions and payoff and uses this information to learn about the efficiency of different allocation profiles and to develop better allocations. We assume the organization accounts for the delays between past allocations and payoffs, but recognize that its estimate of the length of these delays may not be correct. In real systems perceived payoff and payoff, $P(t)$, are different, but to give the learning algorithms the most favorable circumstances, we assume that measurement and perception to be fast and unbiased. The organization accounts for the delays between allocations and payoff based on its beliefs about the length of the payoff generation delays, \bar{T}_j . It attributes the current observed payoff, $P(t)$, to allocations made \bar{T}_j periods ago, so the action associated to the current payoff, $\bar{A}_j(t)$, is:

$$\bar{A}_j(t) = A_j(t - \bar{T}_j) \quad (5)$$

³ Different types of delay, including first- and third-order Erlang delays, were examined; the results were qualitatively the same.

The values of $\bar{A}_j(t)$ and $P(t)$ are used as inputs to the various learning and decision-making procedures representing the organization's behavior.

3- Learning and decision-making procedures

Having perceived its own actions and payoff streams, the organization learns from its experience, that is, updates its beliefs about the efficiency of different allocations and selects what it believes is a better set of allocations. We developed four different learning models to explore the sensitivity of the results to different assumptions about how organizations learn. The inputs to all of these algorithms are the perceived payoff and action associated with that payoff, $P(t)$ and $\bar{A}_j(t)$; the outputs of the algorithms are the allocation decisions $F_j(t)$.

The learning algorithms differ in their level of rationality, information processing requirements, and assumed prior knowledge about the shape of the payoff landscape. Here, rationality indicates organization's ability to make the best use of the information available by trying explicitly to optimize its allocation decisions. Information processing capability indicates organization's capacity for keeping track of and using information about past allocations and payoffs. Prior knowledge about the shape of the payoff landscape determines the ability to use off-line cognitive search (Gavetti and Levinthal 2000) to find better policies.

We use four learning algorithms⁴ denoted as *Reinforcement*, *Myopic Search*, *Correlation* and *Regression*. In all the learning algorithms, the decision-maker has a mental representation of how important each activity is. We call these variables "Activity Value," $V_j(t)$. The activity values are used to determine the allocation of resources to each activity (equations 12-14 explain how allocation of resources is determined based on Activity Values.) The main difference across the different learning algorithms is how these activity values are updated. Below we discuss the four learning algorithms in more details. The complete formulation of all the learning algorithms can be found in the technical appendix (1-1).

1- *Reinforcement learning*: In this method, the value (or attractiveness) of each activity is determined by (a function of) the cumulative payoff achieved so far by using that alternative. Attractiveness then influences the probability of choosing each alternative in the future.

⁴ We tried other algorithms, including Q-learning model used by Denrell et al. (2004), and decided to focus on the four reported here because they are more efficient. The discussion and some results for the Q-learning model are reported in the technical appendix (1-1). In summary, to converge to the optimal policy, that model requires far more data than the reported algorithms.

Reinforcement learning has a rich tradition in psychology, game theory and machine learning (Erev and Roth 1998; Sutton and Barto 1998). It has been used in a variety of applications, from training animals to explaining the results of learning in games and designing machines to play backgammon (Sutton and Barto 1998)⁵.

In our model, each perceived payoff, $P(t)$, is associated with the allocations believed to be responsible for that payoff, $\bar{A}_j(t)$. We increase the value of each activity, $V_j(t)$, based on its contribution to the perceived payoff. The increase in value depends on the perceived payoff itself, so a small payoff indicates a small increase in the values of different activities while a large payoff increases the value much more. Therefore large payoffs shift the relative weight of different activity values towards the allocations responsible for those better payoffs.

$$\frac{d}{dt} V_j(t) = P(t)^{\text{ReinforcementPower}} * \bar{A}_j(t) - V_j(t) / \text{Reinforcement Forgetting Time} \quad (6)$$

Our implementation of reinforcement learning includes a forgetting process, representing the organization's discounting of older information. Discounting old information helps the algorithm adjust the activity values better. Equation 6 shows the main formulation of the *Reinforcement* algorithm. Here both "Reinforcement Power" and "Reinforcement Forgetting Time" are parameters specific to this algorithm. "Reinforcement Power" indicates how strongly we feedback the payoff as reinforcement, to adjust the "Activity Values" and therefore determines the speed of converging to better policies. "Reinforcement Forgetting Time" is the time constant for depreciating the old payoff reinforcements.

Details of the model formulations are included in the technical appendix (1-1), table 1. The *Reinforcement* method is a low information, low rationality procedure: it continues to do what has worked well in the past, adjusting only slowly to new information, and does not attempt to extrapolate from these beliefs about activity value to the shape of the payoff landscape or to use gradient information to move towards higher-payoff allocations.

⁵ The literature on reinforcement learning has different goals and algorithms in different fields. In the context of machine learning, reinforcement learning is developed as an optimization algorithm. Game theory literature on reinforcement learning focuses on providing simple algorithms that give good descriptive accounts of individual learning in simple games. Our use of reinforcement learning here is more aligned with the game theoretic reinforcement learning models. For a survey of machine learning literature on reinforcement learning see Sutton and Barto's book (1998) and Kaelbling et al. (1996). Also see technical appendix (1-1) for the results on a Q-learning model motivated by machine learning literature.

2- *Myopic search*: In this method the organization explores neighboring regions of the decision space (at random). If the last allocation results in a better payoff than the aspiration (based on past performance), the action values are adjusted towards the explored allocation, if the exploration results in a worse payoff, the action values remain unchanged. This procedure is a close variation of Levinthal and March's (1981) formulation for adaptation of search propensity for refinement and innovation. It is also similar to the underlying process for many of the stochastic optimization techniques where, unaware of the shape of the payoff landscape, the algorithm explores the landscape and usually moves to better policies upon discovering them.

Optimization models assume decisions switch instantly to better policies, if found, for the next step. In reality, even if individual decisions can change fast, underlying value system adjusts slowly, due to the time required to aggregate new information, process the information, and make changes to the routines; that is, to the factors that lead to organizational inertia. To be behaviorally more realistic, we assume activity value, $V_j(t)$, adjusts gradually towards the value-set suggested by the last allocation, $V_j^*(t) = \bar{A}_j / \sum \bar{A}_k$ (Equation 7). $V_j^*(t)$ is the last allocation if the payoff improved in the last step and remains to be the current allocation value if the payoff did not change significantly or decreased.

$$\frac{d}{dt}V_j(t) = (V_j^*(t) - V_j(t))/\lambda \quad (7)$$

where λ is the Value Adjustment Time Constant. The formulation details for *Myopic search* method can be found in the technical appendix (1-1), table 2. The myopic method is a low rationality method with medium information processing. It compares the previous payoff with the result of a local search and does not attempt to compare multiple experiments; it also does not use information about the payoffs in the neighborhood to make any inferences about the shape of the payoff landscape.

3- *Correlation*: This method uses principles from attribution theories to model learning. In our setting, learning can be viewed as how organizational decision-makers attribute different allocations to different payoffs and how these attributions are adjusted as new information about payoffs and allocations are continuously perceived. Several researchers have proposed different models for explaining how people make attributions. Lipe (1991) reviews these models and concludes that all major attribution theories are based on the use of counterfactual information. However it is difficult to obtain counterfactual information (information about contingencies that

were not realized) so she proposes the use of covariation data as a good proxy. The correlation algorithm is based on the hypothesis that people use the covariation of different activities with the payoff to make inferences about action-payoff causality.

In our model, the correlations between the perceived payoff and the actions associated with those payoffs let the organization decide whether performance would improve or deteriorate if the activity increases. A positive (negative) correlation between recent values of $\bar{A}_j(t)$ and $P(t)$ suggests that more (less) of activity j will improve the payoff. Based on these inferences the organization adjusts the activity values, $V_j(t)$, so that positively correlated activities increase above their current level and negatively correlated activities decrease below the current level.

$$\frac{d}{dt}V_j(t) = V_j(t) \cdot f(\text{Action_PayoffCorrelation}_j(t)) / \lambda \quad f(0)=1, f'(x)>0 \quad (8)$$

The formulation details for the correlation algorithm are found in technical appendix (1-1), Table 3. At the optimal allocation, the gradient of the payoff with allocation will be zero (the top of the payoff hill is flat) and so will the correlation between activities and payoff. Therefore the change in activity values will become zero and the organization settles on the optimum policy.

The correlation method is a moderate information, moderate rationality approach: more data are needed than required by the myopic or reinforcement methods to estimate the correlations among activities and payoff. Moreover, these correlations are used to make inferences about the local gradient so the organization can move uphill from the current allocations to allocations believed to yield higher payoffs, even if these allocations have not yet been tried.

4- *Regression*: This method is a relatively sophisticated learning algorithm with significant information processing requirements. We assume that the organization knows the correct shape of the payoff landscape and uses a correctly specified regression model to estimate the parameters of the payoff function.

By observing the payoffs and the activities corresponding to those payoffs, the organization receives the information needed to estimate the parameters of the payoff function. To do so, after every few periods, it runs a regression using all the data from the beginning of the

learning task⁶. From these estimates the optimal allocations are readily calculated. Raghu and colleagues (Raghu, Sen et al. 2003) use a similar learning algorithm to model how simulated fund managers learn about effectiveness of different funding strategies. For the assumed constant returns to scale, Cobb-Douglas function, the regression is:

$$\log(P(t)) = \log(\alpha_0) + \sum_j \alpha_j \cdot \log(\bar{A}_j(t)) + e(t) \quad (9)$$

The estimates of α_j , α_j^* , are evaluated every “Evaluation Period,” E . Based on these estimates, the optimal activity values, $V_j^{**}(t)$ are given by:

$$V_j^{**}(t) = \text{Max}(\alpha_j^* / \sum_j \alpha_j^*, 0)^7 \quad (10)$$

The organization then adjusts the action values towards the optimal values (see Equation 7). See Table 4 in technical appendix (1-1) for equation details.

The regression algorithm helps test how delays affect learning over a wide range of rationality assumptions. The three other algorithms represent an organization ignorant about the shape of the payoff function, while in some cases organization have at least partial understanding of the structure and functional forms of the causal relationships relating the payoff to the activities. Although in feedback-rich settings, mental models are far from perfect and calculation of optimal decision based on the understanding of the mechanisms is often cognitively infeasible (Sterman 1989; Sterman 1989), the regression algorithm offers a case of high rationality to test the robustness of our results. Table 1 summarizes the characteristics of the different algorithms on the three dimensions of rationality, information processing capacity and prior knowledge of payoff landscape.

Table 1- Sophistication and rationality of different learning algorithms

Dimension Algorithm	<i>Rationality</i>	<i>Information Processing Capacity</i>	<i>Prior Knowledge of Payoff Landscape</i>
Reinforcement	Low	Low	Low
Myopic Search	Low	Medium	Low
Correlation	Medium	Medium	Low
Regression	High	High	High

⁶Discounting older information does not make any positive difference here since the payoff function does not change during the simulation.

⁷ In the case of negative α_j^* , resulting $V_j^{**}(t)$'s are adjusted to a close point on the feasible action space. This adjustment keeps the desired action values ($V_j^{**}(t)$'s) feasible (positive).

The tradeoff between exploration and exploitation is a crucial issue in learning (March 1991; Sutton and Barto 1998). On one hand the organization should explore the decision space by trying some new allocation policies if she wants to learn about the shape of the payoff landscape. On the other hand pure exploration leads to random allocation decisions with no improvement. So exploration is needed to learn about payoff landscape and exploitation is required if any improvement is to be perceived.

We use random changes in resource allocation to capture the exploration/exploitation issue in our learning models. The “Activity Values” represent the accumulation of experience and learning by the organization. In pure exploitation, the set of activity values, $V_j(t)$, determines the allocation decision. Specifically, the fraction of resources to be allocated to activity j would be:

$$F_j'(t) = V_j(t) / \sum_j V_j(t) \quad (11)$$

The tendency of the organization to follow this policy shows its tendency to exploit the experience it has gained so far. Deviations from this policy represent experiments to explore the payoff landscape. We multiply the activity values by a random disturbance to generate Operational Activity Values, $\hat{V}_j(t)$ ’s, which are the basis for the allocation decisions.

$$\hat{V}_j(t) = V_j(t) + \hat{N}_j(t) \quad (12)$$

$$\hat{N}_j(t) = N_j(t) * \sum_j V_j(t) \quad (13)$$

$$F_j(t) = \hat{V}_j(t) / \sum_j \hat{V}_j(t) \quad (14)$$

$N_j(t)$ is a first order auto-correlated noise term specific to each activity. It generates a stream of random numbers with autocorrelation. In real settings, the organization experiments with a policy for some time before moving to another. Such persistence is both physically required (organizations cannot instantly change resource allocations), and provides organizations with a large enough sample of data about each allocation to decide if a policy is beneficial or not. “Activity noise correlation time”, δ , captures degree of autocorrelation in the stream $N_j(t)$.

High values of δ represent organizations who only slowly change the regions of allocation space they are exploring; a low value represents organizations who explore quickly from one allocation

to another in the neighborhood of their current policy⁸. We use Adjusted Pink Noise, $\hat{N}_j(t)$ for evaluating the Operational Action Values, $\hat{V}_j(t)$'s, because different algorithms have different magnitudes of Action Value and comparable exploration terms need to be scaled.

The standard deviation of the noise term, which determines how far from the current allocations is explored, depends on where the organization finds itself on the payoff landscape. If its recent explorations have shown no improvement in the payoff, it concludes that it is near the peak of the payoff landscape and therefore extensive exploration is not required (alternatively, it concludes that the return to exploration is low and reduces experimentation accordingly). If, however, recent exploration has resulted in finding significantly better regions, it concludes that it is in a low-payoff region and there is still room for improvement so it should keep on exploring, so $\text{Var}(N_j(t))$ remains large:

$$\text{Var}(N_j(t)) = g(\text{Recent Payoff Improvement}(t)),$$

$$g(0) = \text{Minimum Exploration Variance}, g'(x) > 0 \quad (15)$$

In this formulation if the current payoff is higher than recent payoff, *Recent Payoff Improvement* will be increased, if not, it will decay towards 0. Details of exploration equations can be found at Table 5 of technical appendix (1-1).

Results and Analysis

In this section we investigate the behavior of the learning model under different conditions. We first explore the basic behavior of the model and its learning capabilities when there are no delays. The no-delay case helps compare the capabilities of the different learning algorithms and provides a base against which we can compare the behavior of the model under other conditions. Next we introduce delays between activities and payoff and analyze the ability of the different learning algorithms to find the optimal payoff, for a wide range of parameters.

In running the model, the total resource, R, is 100 units and the payoff function exponents are set to 1/2, 1/3 and 1/6 for activities one, two and three, respectively. The optimal allocation

⁸ We set the mean of $N_j(t) = 0$, so the decision-maker has no bias in searching different regions of the landscape. We also truncate the values of PN so that $N_j(t) \geq -1$ to ensure that $\hat{V}_j(t) \geq 0$ (Equation 12).

fractions are therefore $1/2$, $1/3$ and $1/6$, and optimal payoff is 36.37. Starting from a random allocation, the organization tries to improve its performance in the course of 300 periods.⁹

We report two aspects of learning performance: (1) How close to optimal the simulated organization gets, and (2) how fast it converges to that level of performance, if it converges at all. The percentage of optimal payoff achieved by the organization at the end of simulation is reported as Achieved Payoff Percentage. Monte-Carlo simulations with different random noise seeds for the exploration term (equation 13) and for the randomly chosen initial resource allocation give statistically reliable results for each scenario analyzed. To facilitate comparison of the four learning algorithms, we use same noise seeds across the four models therefore any differences in a given run are due only to the differences among the four learning procedures and not the random numbers driving them.

1- Base case

The base case represents an easy learning task where there are no delays between actions and the payoff. The organization is also aware of this fact and therefore she has a perfect understanding of the correct delay structure. In this setting, we expect all the learning algorithms to find the optimum solution. Figure 1 shows the trajectory of the payoff for each learning algorithm, averaged over 100 simulation runs. The vertical axis shows the percentage of the optimal payoff achieved¹⁰ by each learning algorithm.

⁹ The simulation horizon is long enough to give the organization opportunity to learn, while keeping the simulation time reasonable. In the manufacturing firm example we have chosen different time constants so that a period represents one month. Therefore the 300 period horizon would represent about 27 years, ample time to examine how much the organization learns.

¹⁰ Note that by optimum we mean the best payoff one can achieve over the long-term by pursuing any policy which need not to be the highest possible payoff. For example, with a one period delay for activity 1 and no delay for the two others, the organization can allocate all 100 units to activity 1 during the current period and allocate all the resource between two other activities during the next period. Under these conditions, it can achieve higher than optimum payoff for the next period, at the expense of getting no payoff this period (because activities 2 and 3 receive no resources) and the period after the next (because activity 1 receives no resources in that period). A constant returns to scale payoff function prevents such policies from yielding higher payoffs than the constant allocations in longterm.

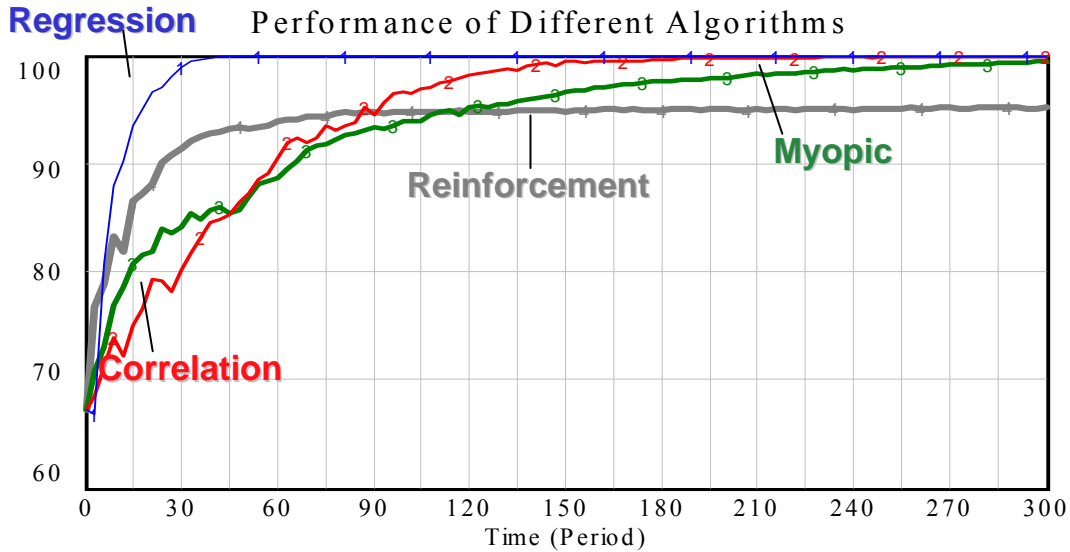


Figure 1- Percentage of payoff relative to optimal in the base case, averaged over 100 runs.

When there are no delays, all four learning algorithms converge to the optimal resource allocation. For comparison, the average payoff achieved by a random resource allocation strategy is 68% of optimal (since the 100 simulations for each learning algorithm start from randomly chosen allocations, all four algorithms begin at an average payoff of 68%.)

An important aspect of learning is how fast the organization converges to the allocation policy it perceives to be optimal. In some settings and with some streams of random numbers it is also possible that the organization does not converge within the 300 period simulation horizon. Table 2 reports the payoffs, the fraction of simulations that have converged as well as the average convergence time¹¹ (for those that converged) in the base case.

Regression, correlation, and myopic algorithms find the optimum almost perfectly and reinforcement goes a long way towards finding the optimum payoff. Majority of simulations converge prior to the 300 period horizon and the average convergence times range from a low of 38 periods for the regression algorithm to a high of 87 periods for the myopic model.

¹¹ We consider a particular learning procedure to have converged when the variance of the payoff falls below 1% of its historical average. If later the variance increases again, to 10% of its average at the time of convergence, we reset the convergence time and keep looking for the next instance of convergence.

Table 2- Achieved Payoff, Convergence Time and Percentage Converged for the Base case. Statistics from 400 simulations.

Learning Algorithm	Regression		Correlation		Myopic		Reinforcement	
Variable	μ	σ	μ	σ	μ	σ	μ	σ
Achieved Payoff percentage at period 300	99.96	0.04	99.91	0.07	99.43	1.57	95.18	5.02
Convergence Time	37.97	3.84	75.39	32.68	87.26	60.34	40.28	29.25
Percentage Converged ¹²	99.75		99.75		98.75		90	

2-The Impact of delays

Most models of organizational learning share an assumption that actions instantly payoff and the results are perceived by the organization with no delay. Under this conventional assumption, all our learning algorithms reliably find the optimum allocation and converge to it. In this section we investigate the results of relaxing this assumption.

A basic variation is to introduce a delay in the impact of one of the activities, leaving the delay for the other two other activities at zero. We simulate the model for 9 different values of the “Payoff Generation Delay” for activity one, T_1 , ranging from 0 to 16 periods, while we keep the delays for other activities at 0. In our factory management example the long delay in the impact of activity 1 is analogous to process improvement activities that take a long time to bear fruit. Because activity one is the most influential in determining the payoff, these settings are expected to highlight the effect of delays more clearly. Delays as high as 16 period are realistic in magnitude, when weighed against the internal dynamics of the model. For example, in the base case, it takes between 38 (for regression) to 87 (for myopic) periods for different learning models to converge when there is no delays, suggesting that dynamics of exploration and adjustment delays unfold in longer time horizons than different delays tested (See technical appendix (1-1), Table 6, for an alternative measure of speed of internal dynamics).

We analyze two setting for the perceived payoff generation delay. First, we examine the results when the delays are correctly perceived: we change the perceived payoff generation delay

¹² Percentage of simulations converging to some policy at the end of 300 period simulations. These numbers are reported based on 400 simulations.

for activity one, \bar{T}_1 at the same values as T_1 , so that we have no misperception of delays: $T_j = \bar{T}_j$. This scenario informs the effect of delays when they are correctly perceived and accounted for. Then, we analyze the effect of misperception of delays by keeping the perceived payoff generation delay for all activities, including activity one, at zero. This setting corresponds to an organization that believes all activities affect the payoff immediately.

Figure 2 reports the behavior of the four learning algorithms under the first setting, where delays are correctly perceived. Under each delay time the Average Payoff Percentage achieved by organization at the end of 300 periods and the Percentage Converged¹³ are reported. In the “Average Payoff Percentage” graph, the line denoted “Pure Random” represents the case where the organization selects its initial allocation policy randomly and sticks to that, without trying to learn from its experience. Results are averaged over 400 simulations for each delay setting.

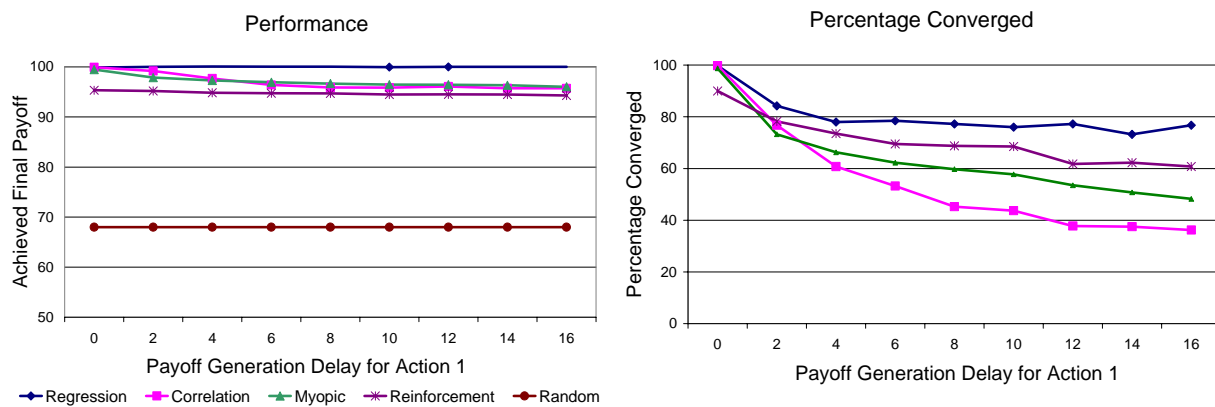


Figure 2- Payoff Percentage and Percentage Converged with different, correctly perceived, time delays for first activity. Activities 2 and 3 have no delay in influencing the payoff.

The performances of all the four learning algorithms are fairly good. In fact regression model always finds the optimum allocation policy under all different delay settings if it has a correct perception of delays involved and the average performance of all the algorithms remain over 94% performance line. The convergence percentages declines, as continuously exploring the optimal region, a significant number of simulations fail to converge to the optimal policy, even though their performance levels are fairly high.

¹³ Convergence times are (negatively) correlated with the fraction of runs that converge and therefore are not graphed here.

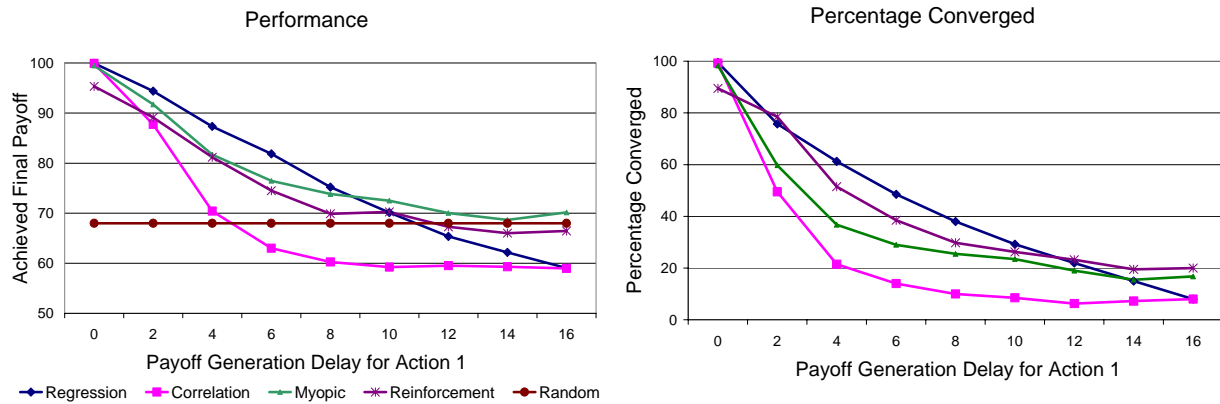


Figure 3- Average Payoff Percentage and Percentage Converged with different time delays for first activity, when perceived delays are all zero. Results are averaged over 400 simulations.

Figure 3 shows the performance of four learning algorithms under the settings with misperceived delays. This test suggests that the performances of all of the learning algorithms are adversely influenced with the introduction of misperceived delays. It also shows a clear drop in the convergence rates as a consequence of such delays.

In summary, the following patterns can be discerned from these graphs. First, learning is significantly hampered if the perceived delays do not match the actual delays. The pattern is consistent across the different learning procedures. As the misperception of delays grow all four learning procedures yield average performance worse than or equal to the random allocation policy (68% of optimal under this payoff function). However, correctly perceived delays have a minor impact on performance.

Second, under biased delay perception the convergence percentages decrease significantly, suggesting that algorithms usually keep on exploring as a result of inconsistent information they receive. More convergence happen when the delays are perceived correctly. Meanwhile, still a notable fraction of simulation runs do not converge under correctly perceived delays, which suggest that existence of delays by themselves complicate the task of organization for converging on the optimal policy. More interestingly, under misperceived delays, some simulations converge to sub-optimal policies for different algorithms. This means that the organization can under some circumstances end up with an inferior policy concluding that this is the best payoff it can get, and stops exploring other regions of the payoff landscape, even though there is significant unrealized potential for improvement.

Third, different learning algorithms show the same qualitative patterns. They also show some differences in their precise performance, where under misperceived delays, the more rational learning algorithms, regression and correlation, perform better in short delays and under-perform in long delays. In fact regression and correlation models under-perform the random allocation ($p < 0.0001$), while regression always finds the optimal allocation when the delay structure is correctly perceived.

In short, independent of our assumptions about the learning capabilities of the organization, a common failure mode persists: when there is a mismatch between the true delay and the delay perceived by the organization, learning is slower and usually does not result in improving the performance. In fact the simulated organization frequently concludes that an inferior policy is the best it can achieve, and sometimes, trying to learn from their experience, the organizations reach equilibrium allocations that yield payoffs significantly lower than the performance of a completely random strategy. To explore the robustness of these results, we investigate the effect of all model parameters on the performance of each learning algorithm.

3-Robustness of results

Each learning model involves parameters whose values are highly uncertain. To explore the sensitivity of the results to these parameters we conducted sensitivity analysis over all the parameters of each learning procedure. In this analysis we aim at investigating the robustness of the conclusions we have derived in the last section about the effect of delays on learning and answer the following questions:

- Is there any asymmetry in the effect of misperceived delays on performance? In reality organizations are more prone to underestimating, rather than overestimating, the delays (e.g. managers usually don't overestimate how long it takes the process improvement to payoff), so finding out about such asymmetries is important and consequential.
- Is there any nonlinear effect of delays on learning? In specific, there should be a limit to negative effects of delays on performance. Do such saturation effects exist and if so at what delay levels do they become important?

We conducted a Monte Carlo analysis, selecting each of the important model parameters randomly from a uniform distribution over a wide range (One to four times smaller/bigger than the base case. See the complete listing and ranges in the technical appendix (1-1), table 7). We carried out 3000 simulations, using random initial allocations in each.

Simple graphs and tables are not informative about these multidimensional data. We investigate the results of this sensitivity analysis using regressions with three dependent variables: Achieved Payoff Percentage, Distance Traveled, and Probability of Convergence. Distance traveled measures how far the algorithm has traveled on the action space, and is a measure of the cost of search and exploration.

For each of these variables and for each of the learning algorithms, we run a regression over the independent parameters in that algorithm (See technical appendix (1-1) for complete listing). The regressors also include the Perception Error, which is the difference between the Payoff Generation Delay and its perceived value; Absolute Perception Error; and Squared Perception Error. Having both Absolute Perception Error and Perception Error allows us to examine possible asymmetric differences for misperceiving errors. The second order term will allow inspection of nonlinearities and saturation of delay effects. To save space and focus on the main results, we only report the regression estimates for the main independent variables (Delay, Perception Error, Absolute Perception Error, and Squared Perception Error), even though all the model parameters are included on the right hand side of the reported regressions. OLS is used for the Achieved Payoff Percentage and Distance Travelled; logistic regression is used for the Convergence Probability. Tables 3–5 show the regression results, summary statistics are available in the technical appendix (1-1), Table 8. All the models are significant at $p < 0.001$. The following results follow from these tables:

- Increasing the absolute difference between the real delay and the perceived delay always decreases the achieved payoff significantly (Table 4, row 4). The coefficient for this effect is large and highly significant, indicating the persistence of the effect across different settings of parameters and different learning algorithms. Absolute perception error also usually increases the cost of learning (Distance Traveled), even though the effect is not as pronounced as that of achieved payoff (Compare Tables 3 and 4, Rows 4).
- Delay alone appears to have much less influence on payoff. The regression model is insensitive to correctly perceived delays and the rest of the models show a small but significant decline of performance (Table 3, Row 3).
- Three of the models show asymmetric behavior in presence of misperceived delays. For Regression and Correlation, underestimating the delays is more detrimental than overestimating them (Table 3, Row 5) while Myopic is better off underestimating the

delays rather than overestimating them. In all models the absolute error has a much larger effect than perception error itself (Table 3, Compare Rows 4 and 5).

- There is a significant nonlinear effect of misperceived delays on performance (Table 3, Row 6). This effect is in the expected direction, so higher perception errors have smaller incremental effect on learning and the effect saturates (second order effect neutralizes the first order effect) when the delay perception error is around 14 (except for regression which has a much smaller saturation effect).
- Convergence is influenced both by delay and by error in perception. So higher delays, even if perceived correctly, make it harder for the algorithms to converge (Table 5).

These results are in general aligned with the analysis of delays offered in the last section and highlight the robustness of the results. The analysis shows that the introduction of delay between resource allocations and their impact causes sub-optimal performance, slower learning, and higher costs for learning.

Table 3- Regression for Achieved Payoff Percentage

Variable \ Algorithm		Regression			Correlation			Myopic			Reinforcement		
1	Adj. R ² \ Model DF	0.269	12		0.172	13		0.074	11		0.075	12	
		Mean	S.D.	P*	Mean	S.D.	P*	Mean	S.D.	P*	Mean	S.D.	P*
2	Intercept (Est, S.D., Significance)	91.819	1.971	<.0001	91.540	3.155	<.0001	88.259	2.425	<.0001	87.105	2.322	<.0001
3	Payoff Generation Delay [a1]	0.027	0.068	0.694	-0.285	0.104	0.006	-0.199	0.087	0.022	-0.112	0.081	0.163
4	Absolute Perception Error	-1.857	0.204	<.0001	-4.718	0.309	<.0001	-2.592	0.260	<.0001	-2.402	0.240	<.0001
5	Perception Error	-0.704	0.049	<.0001	-0.418	0.074	<.0001	0.221	0.062	0.0004	0.034	0.057	0.559
6	Perception Error ²	0.025	0.012	0.039	0.208	0.019	<.0001	0.125	0.016	<.0001	0.099	0.015	<.0001

Table 4- Regression for Distance Traveled

Variable \ Algorithm		Regression			Correlation			Myopic			Reinforcement		
1	Adj. R ² \ Model DF	.695	12		0.662	13		0.686	11		0.682	12	
		Mean	S.D.	P*	Mean	S.D.	P*	Mean	S.D.	P*	Mean	S.D.	P*
2	Intercept (Est, S.D., Sig.)	18.482	0.822	<.0001	20.301	0.871	<.0001	20.932	0.852	<.0001	19.806	0.896	<.0001
3	Payoff Generation Delay [a1]	0.099	0.029	0.001	0.055	0.029	0.055	0.150	0.031	<.0001	0.115	0.031	0.000
4	Absolute Perception Error	0.469	0.085	<.0001	0.136	0.085	0.111	0.230	0.091	0.012	0.254	0.093	0.006
5	Perception Error	0.016	0.020	0.422	0.019	0.020	0.359	-0.079	0.022	0.0003	-0.026	0.022	0.237
6	Perception Error ²	-0.014	0.005	0.006	-0.006	0.005	0.219	-0.013	0.006	0.022	-0.011	0.006	0.054

Table 5- Logistic Regression for Convergence Probability

Variable \ Algorithm		Regression				Correlation				Myopic				Reinforcement			
	Chi-Sqr \ number converged	594	707			401	367			492	535			528	600		
2	Est/S.D./OddsR/ Sig.	Est.	S.D.	Sig.	Odd R	Est.	S.D.	Sig.	Odd R	Est.	S.D.	Sig.	Odd R	Est.	S.D.	Sig.	Odd R
3	Intercept	-0.690	0.429	0.108		-1.32	0.501	0.008		-1.06	0.438	0.016		-1.221	0.450	0.007	
4	Payoff Generation Delay [a1]	-0.049	0.015	0.001	0.952	-0.122	0.017	<.0001	0.885	-0.074	0.016	<.0001	0.929	-0.070	.0157	<.0001	0.932
5	Absolute Perception Error	-0.071	0.047	0.130	0.931	-0.150	0.051	0.003	0.861	-0.122	0.049	0.012	0.885	-0.143	0.0487	.0033	0.867
6	Perception Error	-0.045	0.011	<.0001	0.956	-0.009	0.012	0.448	0.991	0.040	0.011	0.000	1.041	.014	.0113	0.207	1.014
7	Perception Error ²	-0.003	0.003	0.348	0.997	0.006	0.003	0.051	1.006	0.006	0.003	0.055	1.006	0.00417	0.00301	0.166	.004

Discussion

The results show that learning is slow and ineffective when the organization underestimates the delay between an activity and its impact on performance. It may be objected that this is hardly surprising because the underlying model of the task is misspecified (by underestimation of the delays). A truly rational organization would not only seek to learn about better allocations, but would also seek to test its assumptions about the true temporal relationship of actions and their impact; if its initial beliefs about the delay structure were wrong, experience should reveal the problem and lead to a correctly specified model. We do not attempt to model such second-order learning here and leave the resolution of this question to further research. Nevertheless, the results suggest such sophisticated learning is likely to be difficult.

First, estimating delay length and distribution is difficult and requires substantial data¹⁴; in some domains, such as the delay in the capital investment process, it took decades for consensus about the length and distribution of the delay to emerge (see Sterman 2000, Ch. 11 and references therein). Results of an experimental study are illuminating: in a simple task (one degree of freedom) with delays of one or two periods, Gibson (2000) shows that individuals had a hard time learning over 120 trials. A simulation model fitted to human behavior and designed to learn about delays, learns the delay structure only after 10 times more training data.

Second, the experimental research suggests people have great difficulty recognizing and accounting for delays, even when information about their length and content is available and salient (Sterman 1989 (a); Sterman 1989 (b); Sterman 1994); in our setting there are no direct cues to indicate the length, or even the presence, of delays. Third, the outcome feedback the decision maker receives may be attributed to changing payoff landscape and noise in the environment, rather than to a problem with the initial

¹⁴ The Q-learning model discussed in the technical appendix (1-1) allows us to show mathematically why the complexity of the state-space (which correlates well with time and data needed to learn about delay structure) grows exponentially with the length of delays. Details can be found in technical appendix (1-1); however, the basic argument is that if the action space has the dimensionality $O(D)$, then the possibility for one period of delay requires us to know the last action as well as the current one to realize the payoff, increasing the dimensionality to $O(D.D)$ and for delays of length K to $O(D^{K+1})$.

assumptions about the delay structure. Decline in convergence and learning speed should be the main feedback for the organization to decide that its assumptions about the delay structure are erroneous. However, in the real world, where the payoff is not solely determined by the organization's actions, several other explanatory factors, from competitors' actions to sheer good luck and supernatural beliefs, come first in the explanation of the inconsistent payoff feedback.

Our results are robust to rationality assumption of learning algorithm. Effects of time delays persist for a large range of model complexity and rationality that is attributable to human organizations; therefore, one can have higher confidence in applicability of these results to real-world phenomena. This is not to claim that no learning algorithm can be designed to account for delays¹⁵; rather, for the range of learning algorithms and experiment opportunities realistically applicable to organizations, delays between action and results become a significant barrier to learning. This conclusion is reinforced by the amount of experimentation needed to learn about a mismatch between the real-world delay structure and what the aggregate mental model of the organization suggests.

A few different mechanisms contribute to failure of learning in our setting. First, in the absence of a good mental model of the delay structure, the organization fails to make reasonable causal attributions between different policies and results (the credit assignment problem (Samuel 1957; Minsky 1961)). Moreover, in the context of resource allocation, an increase in resources allocated to any activity naturally cuts down the resources allocated to the other activities; therefore, observing payoff of long-term activities requires going through a period of low payoff. For example, investing resources in process improvement will cut down the short-term payoff by reducing the time spent on producing. When delays are not correctly perceived, the transitory payoff shortfall can be attributed to lower effectiveness of the long-term activity, thereby

¹⁵ There is a rich literature in machine learning, artificial intelligence, and optimal control theory that searches for optimum learning algorithms in dynamic tasks. (See some examples at: Kaelbling 1993; Barto, Bradtke et al. 1995; Kaelbling, Littman et al. 1996; Santamaria, Sutton et al. 1997; Tsitsiklis and VanRoy 1997; Bertsekas 2000). The focus of this literature is on optimal policies, rather than behavioral decision/learning rules and therefore these models are often higher in information-need and level of rationality than the range of empirically plausible decision-rules for human subjects or organizations.

providing a mechanism for learning wrong lessons from experience. Finally, learning may be difficult even when the delays are correctly perceived by the organization, especially when the organization does not have a perfectly specified model of the payoff landscape. In the presence of delays, even if they are correctly perceived, the information provided through payoff feedback may be outdated and be relevant only for a different region of payoff landscape. Consider the case where the delay is three periods and correctly perceived. In this case the organization properly attributes the payoff to the decisions of 3 periods ago and correctly finds out in which direction it could have changed the allocation decision 3 periods ago to improve the payoff. However, during the intervening 3 periods it has been exploring different policies and therefore the indicated direction of change in policy may no longer indicate the best direction to move. Nevertheless, this type of error is not large in our experimental setting as a result of relatively flat payoff surface near the peak.

Our modeling of learning includes simplifying assumptions that need to be clear in light of their possible influence on learning performance and speed. First, in line with most other learning models in the literature, these algorithms use stochastic exploration of different possible actions, to learn from the feedbacks that they receive. This encourages a relatively large number of exploratory moves compared to what most real-world situations allow and provides the simulated organization with more data to learn from than most real settings. Moreover, costs and resources are not modeled here and therefore there is no bound on exploration and no pressure on the simulated organization to converge to any policy. These effects may increase the possibility of convergence of the real-world organization and increase the speed of learning; however, it is conceivable that their effect on quality of learning is not positive, possibly leading to more learning of wrong lessons than the current analysis suggests. Such reinforcement of harmful strategies is both theoretically interesting and practically important.

For a concrete example, consider a firm facing a production shortfall. Pressuring the employees to work harder generates a short-run improvement in output, as they reallocate time from improvement and maintenance to production. The resulting decline in productivity and equipment uptime, however, comes only after a delay. It appears to be difficult for many managers to recognize and account for such delays, so they conclude

that pressuring their people to work harder was the right thing to do, even if it is in fact harmful. Over time such attributions become part of the organization's culture and have a lasting impact on firm strategy. Repenning and Sterman (2002) show how, over time, managers develop strongly held beliefs that pressuring workers for more output is the best policy, that the workers are intrinsically lazy and require continuous supervision, and that process improvement is ineffective or impossible in their organization—even when these beliefs are false. Such convergence to sub-optimal strategies as a result of delays in payoff constitutes an avenue through which temporal complexity of payoff landscape can create heterogeneity in firm performance (Levinthal 2000).

Future extensions of this research can take several directions. First, there is room for developing more realistic learning models that capture the real-world challenges that organizations face. Possible extensions can be designing more realistic payoff landscapes (e.g., modeling the plant explicitly) and going beyond simple stimuli-response models (e.g., including culture formation). Second, it is intriguing to look into whether there is any interaction effect between delays and rugged landscape. Third, it is possible to investigate the effect of delays on learning in the presence of noise in payoff and measurement/perception delays. Fourth, distribution of payoff over time may have important ramifications for routines learnt by organizations. For example, it is conceivable that payoff that is distributed smoothly through time does not pass the organizational attention threshold and is thus heavily discounted. Therefore, activities that lead to distributed payoff can be under-represented in learnt routines. Fifth, these results can help us better explain some empirical cases of learning failure. Sixth, one can look at learning with delays in a population of organizations. Finally, it is important to look at the feasibility of the second-order learning, i.e., learning about the delay structure.

Our research also has important practical implications. The results highlight the importance of employing time horizons longer than typical strategies for evaluating performance. This is important both in designing incentive structures inside the organization as well as in the market-level evaluation of the firm's performance. Empirical evidence highlights this conclusion. For example, Hendricks and Singhal (2001) show the inefficiency of financial markets, in that stock prices do not effectively reflect the positive effects of total quality management programs on firm performance at

the time the information about the quality initiative becomes known to the public. Moreover, more effective data-gathering and accounting procedures can be designed that explicitly represent the delays between different activities/costs and the expected/observed payoff, and therefore improve the chances of organizational learning in the presence of delays.

In sum, the results of our analysis suggest that learning can be significantly hampered and slowed when the delays between action and payoff are not correctly taken into account. In fact, under relatively short delays we repeatedly observe that all learning algorithms fail to achieve the performance of a random allocation policy. The robustness of these results under a large parameter space and their independence from level of rationality and information processing of models adds to their external validity and generalizability. This research highlights the importance of explicitly including the time dimension of action-payoff relationship in learning research and offers a new dimension to explain several cases of learning failure in the real world and the heterogeneity of firms' strategies.

References:

- Allen, K. 1993. Maintenance diffusion at du pont: A system dynamics perspective. *Sloan School of Management*. Cambridge, M.I.T.
- Argote, L. and D. Eppe. 1990. Learning-curves in manufacturing. *Science* **247**(4945): 920-924.
- Argyris, C. 1999. *On organizational learning*. Malden, Mass., Blackwell Business.
- Argyris, C. and D. A. Schön. 1978. *Organizational learning : A theory of action perspective*. Reading, Mass., Addison-Wesley Pub. Co.
- Barto, A. G., S. J. Bradtke and S. P. Singh. 1995. Learning to act using real-time dynamic-programming. *Artificial Intelligence* **72**(1-2): 81-138.
- Bertsekas, D. P. 2000. *Dynamic programming and optimal control*. Belmont, Mass., Athena Scientific.
- Busemeyer, J. R., K. N. Swenson and A. Lazarte. 1986. An adaptive approach to resource-allocation. *Organizational Behavior and Human Decision Processes* **38**(3): 318-341.
- Carroll, J. S., J. W. Rudolph and S. Hatakenaka. 2002. Learning from experience in high-hazard organizations. *Research in Organizational Behavior, Vol 24* **24**: 87-137.
- Cyert, R. M. and J. G. March. 1963. *A behavioral theory of the firm*. Englewood Cliffs, N.J., Prentice-Hall.
- DeGeus, A. P. 1988. Planning as learning. *Harvard Business Review* **66**(2): 70-74.
- Denrell, J., C. Fang and D. A. Levinthal. 2004. From t-mazes to labyrinths: Learning from model-based feedback. *Management Science* **50**(10): 1366-1378.
- Denrell, J. and J. G. March. 2001. Adaptation as information restriction: The hot stove effect. *Organization Science* **12**(5): 523-538.
- Diehl, E. and J. Sterman. 1995. Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes* **62**(2): 198-215.
- Einhorn, H. J. and R. M. Hogarth. 1985. Ambiguity and uncertainty in probabilistic inference. *Psychological Review* **92**(4): 433-461.
- Erev, I. and A. E. Roth. 1998. Xxxpredicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* **88**(4): 848-881.
- Friedman, M. 1953. *The methodology of positive economics. Essays in positive economics*. Chicago, IL, Chicago University Press.
- Gavetti, G. and D. Levinthal. 2000. Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly* **45**(1): 113-137.
- Gibson, F. P. 2000. Feedback delays: How can decision makers learn not to buy a new car every time the garage is empty? *Organizational Behavior and Human Decision Processes* **83**(1): 141-166.
- Hendricks, K. B. and V. R. Singhal. 2001. The long-run stock price performance of firms with effective tqm programs. *Managemet Science* **47**(3): 359-368.
- Herriott, S. R., D. Levinthal and J. G. March. 1985. Learning from experience in organizations. *American Economic Review* **75**(2): 298-302.
- Huber, G. 1991. Organizational learning: The contributing processes and literature. *Organization Science* **2**(1): 88-115.

- Janis, I. L. 1982. *Groupthink : Psychological studies of policy decisions and fiascoes*. Boston, Houghton Mifflin.
- Kaelbling, L. P. 1993. *Learning in embedded systems*. Cambridge, Mass., MIT Press.
- Kaelbling, L. P., M. L. Littman and A. W. Moore. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* **4**: 237-285.
- Lant, T. K. and S. J. Mazias. 1992. An organizational learning model of convergence and reorientation. *Organization Science* **3**: 47-71.
- Lant, T. K. and S. J. Mezas. 1990. Managing discontinuous change - a simulation study of organizational learning and entrepreneurship. *Strategic Management Journal* **11**: 147-179.
- Leonard-Barton, D. 1992. Core capabilities and core rigidities - a paradox in managing new product development. *Strategic Management Journal* **13**: 111-125.
- Levinthal, D. 2000. Organizational capabilities in complex worlds. *The nature and dynamics of organizational capabilities*. S. Winter. New York., Oxford University Press.
- Levinthal, D. and J. G. March. 1981. A model of adaptive organizational search. *Journal of Economic Behavior and Organization* **2**: 307-333.
- Levinthal, D. A. 1997. Adaptation on rugged landscapes. *Management Science* **43**(7): 934-950.
- Levinthal, D. A. and J. G. March. 1993. The myopia of learning. *Strategic Management Journal* **14**: 95-112.
- Levitt, B. and J. G. March. 1988. Organizational learning. *Annual Review of Sociology* **14**: 319-340.
- Lipe, M. G. 1991. Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin* **109**(3): 456-471.
- March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science* **2**(1): 71-87.
- Meadows, D. H. and Club of Rome. 1972. *The limits to growth; a report for the club of rome's project on the predicament of mankind*. New York., Universe Books.
- Milgrom, P. and J. Roberts. 1990. The economics of modern manufacturing - technology, strategy, and organization. *American Economic Review* **80**(3): 511-528.
- Minsky, M. 1961. Steps toward artificial intelligence. *Proceedings Institute of Radio Engineers* **49**: 8-30.
- Nelson, R. R. and S. G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, Mass., Belknap Press of Harvard University Press.
- Paich, M. and J. Sterman. 1993. Boom, bust, and failures to learn in experimental markets. *Management Science* **39**(12): 1439-1458.
- Raghu, T. S., P. K. Sen and H. R. Rao. 2003. Relative performance of incentive mechanisms: Computational modeling and simulation of delegated investment decisions. *Management Science* **49**(2): 160-178.
- Repenning, N. P. and J. D. Sterman. 2002. Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly* **47**: 265-295.
- Rivkin, J. W. 2000. Imitation of complex strategies. *Management Science* **46**(6): 824-844.

- Rivkin, J. W. 2001. Reproducing knowledge: Replication without imitation at moderate complexity. *Organization Science* **12**(3): 274-293.
- Samuel, A. 1957. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**: 210-229.
- Santamaria, J. C., R. S. Sutton and A. Ram. 1997. Experiments with reinforcement learning in problems with continuous state and action spaces. *Adaptive Behavior* **6**(2): 163-217.
- Senge, P. M. 1990. *The fifth discipline: The art and practice of the learning organization*. New York, Currency Doubleday.
- Sengupta, K., T. K. Abdel-Hamid and M. Bosley. 1999. Coping with staffing delays in software project management: An experimental investigation. *Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans* **29**(1): 77-91.
- Siggelkow, N. 2002. Evolution toward fit. *Administrative Science Quarterly* **47**(1): 125-159.
- Sterman, J. D. 1989. Misperception of feedback in dynamic decision making. *Organizational behavior and human decision processes* **43**: 301-335.
- Sterman, J. D. 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**(3): 321-339.
- Sterman, J. D. 1994. Learning in and about complex systems. *System Dynamics Review* **10**(2-3): 91-330.
- Sutton, R. S. and A. G. Barto. 1998. *Reinforcement learning: An introduction*. Cambridge, The MIT Press.
- Tsitsiklis, J. N. and B. VanRoy. 1997. An analysis of temporal-difference learning with function approximation. *Ieee Transactions on Automatic Control* **42**(5): 674-690.
- Tushman, M. L. and E. Romanelli. 1985. Organizational evolution: A metamorphosis model of convergence and revolution. *Research in Organizational Behavior* **7**: 171-122.
- Warren, H. C. 1921. *A history of the association psychology*. New York Chicago etc., C. Scribner's sons.
- Watkins, C. 1989. Learning from delayed rewards. *Ph.D. Thesis*. Cambridge, UK, Kings' College.