Cost optimal robust-schedules

Policy search in continuous action space using stochastic feedback

Nico Roos

Department Knowledge Engineering, Maastricht University
P.O.Box 616, 6200 MD Maastricht, The Netherlands
email: roos@maastrichtuniversity.nl

Abstract

This paper investigates the application of *stochastic approximation theory* to the problem of learning optimal temporal buffers between activities in order to make a schedule more robust. We investigate this problem for the domain of Airport Ground Handing services. Because of incidents, the duration of these services may sometimes take longer than expected. This may have consequences for preceding services or the planned takeoff of an aircraft. To avoid rescheduling of services, temporal buffers can be inserted between activities.

The optimal buffer size depends on the occurrence of uncertain incidents. Stochastic approximation theory enables us to learn an optimal buffer based on observed incident costs. Convergence is however slow. The paper investigates the reason for the slow convergence and proposes an improvement that gives a speedup of a factor 30.

1 Introduction

Airport Ground Handling (AGH) is well-known for incidents disrupting the execution of the ground handling services an aircraft needs during its turnaround process (from landing till take-off). Because of incidents, some ground handling services have to be rescheduled, causing additional costs. These costs can be reduced by adding temporal buffers between services. The temporal buffers increase the time of the turnaround process and thereby also introducing additional costs. The total cost of inserting a temporal buffer after a ground handling service is the sum of the delay cost caused by the buffer and the rescheduling cost if the buffer is too small.

The challenge is to choose a buffer size that minimizes its expected total cost. Of course, the optimal buffer size depends on environmental factors such as the weather and the crowdedness of the airport. Hence, we have an optimization problem with a continuous domain of possible actions (choosing a buffer size) and a continuous state space (determined by the weather, the crowdedness, etc.).

Abstracting from the application domain of AGH, we can say that we have to learn an optimal policy mapping from a continuous domain of states to a continuous domain of actions based on observations about incidents. Here an incident is described by the additional time needed to finish a ground handling service. The cost of the current buffer size and the rescheduling cost if the buffer size is too small to handle an incident, is the reward of the current buffer size. Reinforcement learning seems therefore the obvious choice to learn an optimal policy.

To simplify the problem we discretize the state space of the learning problem. We can introduce states for different weather conditions and different levels of crowdedness at an airport. We do not wish to discretize the action space. A discrete action space requires learning a quality measure for every action. We are however only interested in learning the optimal action. For this reason, we will not consider reinforcement learning techniques such as Q-learning [5]. Instead, we will focus on a technique underlying all learning techniques based on temporal difference learning, namely, *stochastic approximation theory*.

Stochastic approximation theory was introduced by Robins and Monro [4]. It has subsequently been extended and the convergence conditions have been relaxed by several authors, e.g. [2, 1]. We will investigate

the practical applicability of stochastic convergence theory to our problem of learning optimal buffer sizes for airport ground handling services.

In the next section, a formalization of our learning problem is given together with a specification of the environment in which the learning task has to be preformed. Section 3 analyzes the possibility of learning an optimal buffer size using different approximation methods. In Section 4, the experiments that were carried out are described as well as the results of the experiments. Section 5 concludes the paper.

2 Problem description

The processing time of ground handling services may take longer because of incidents. Incidents may force rescheduling of the remaining ground handling services, resulting in additional rescheduling cost. We may reduce the additional rescheduling cost by reducing the need for rescheduling through the insertion of a temporal buffer after a service. The insertion of a buffer after a service results in a larger processing time of the whole job / project. The marginal cost of inserting a buffer is added to the cost of the service after which the buffer is inserted. Also the rescheduling cost in case the buffer is still to small, is added to the cost of this service.

Notation:

- a : the buffer size (we use 'a' because of the *action* of adding a buffer);
- s : the current state;
- $p(x \mid s)$: the probability that ground handling service requires an additional x time units processing time given the current state of the world s;
- $dc(s, a)$: the cost of adding a buffer a (delay costs), which might depend the on state s;
- $rs(s, a, x)$: the (expected) rescheduling cost in state s given a buffer a and an incident duration of x;
- $etc(s, a)$: the expected total cost of adding a buffer of size a after a service in state s.

The expected *total cost* is given by:

$$etc(s, a) = dc(s, a) + \int_a^\infty p(x \mid s) \cdot rs(s, a, x)\, dx$$

Since $rs(s, a, x) = 0$ if $x \leq a$, we can rewrite the equation to:

$$etc(s, a) = dc(s, a) + \int_0^\infty p(x \mid s) \cdot rs(s, a, x)\, dx$$

The optimal buffer a given in state s is a buffer size for which:

$$\frac{\partial}{\partial a} etc(s, a) = 0$$

where:

$$\begin{aligned}
\frac{\partial}{\partial a} etc(s, a) &= \frac{\partial}{\partial a} dc(s, a) + \frac{\partial}{\partial a} \int_0^\infty p(x \mid s) \cdot rs(s, a, x)\, dx \\
&= \frac{\partial}{\partial a} dc(s, a) + \int_0^\infty \frac{\partial}{\partial a} \left(p(x \mid s) \cdot rs(s, a, x) \right)\, dx \\
&= \frac{\partial}{\partial a} dc(s, a) + \int_0^\infty p(x \mid s) \cdot \frac{\partial}{\partial a} rs(s, a, x)\, dx
\end{aligned}$$

The optimal policy $\pi^* : S \to A$ selects the optimal buffer size:

$$\frac{\partial}{\partial a} etc(s, \pi^*(s)) = 0$$

Note that in the domain of buffer sizes that we are interested in, we can assume that the function $etc(s, a)$ is convex. Therefore, there will be no local optima beside the global optimum.

Cost functions and the probability of incidents In order to study the ability of learning an optimal buffer size independent of the rescheduling algorithm [3] and to be able to change the rescheduling cost, all experiments have been carried out with predefined functions. The following functions were used in the experiments.

- The probability of an incident extending the duration of a service with a time period x is determined by an exponential probability distribution.

$$p(x \mid s) = \frac{1}{\mu(s)} e^{\frac{-x}{\mu(s)}}$$

  Here, $\mu(s)$ is the average duration of an incident, which may depend on the state s.

- The cost of a buffer of size a is described by a linear function.

$$dc(s, a) = u(s) \cdot a$$

  For a fixed state s, $u(s)$ is a constant.

- The rescheduling cost of an incident x given a buffer of size a is also described by a linear function

$$rs(s, a, x) = \begin{cases} v(s) \cdot (x - a) + w(s) & \text{if } x > a \\ 0 & \text{otherwise} \end{cases}$$

  For a fixed state s, $v(s)$ and $w(s)$ are constants. Note that the rescheduling cost is discontinuous if $w(s) \neq 0$. This is often the case at an airport where the aircraft for which services have to be rescheduled have to wait.

3 Learning

Given a state s, we must learn an optimal policy $\pi(s)$, that is, the policy such that $\pi^*(s) = a^*$ and $etc(s, a^*)$ is minimal. As was pointed out above, the expected total cost function $etc(s, a)$ is convex. Therefore learning the optimal policy for a given state seems to be a simple hill-climbing problem. Unfortunately, because of the underlying stochastic process, the occurrence of incidences, standard hill-climbing is not an option. Stochastic approximation theory offers an alternative for standard hill-climbing.

Stochastic approximation theory introduced by Robins and Monro [4] addresses learning the root of expected value of a stochastic function. Here, the stochastic function is described the random variable $Y(a, x) = \frac{\partial}{\partial a}(rs(s, a, x) + dc(s, a))$ assuming that the state s is fixed. The expected value of this random variable is a function of the buffer size a:

$$\mu'(a) = E_x(Y(a, x)) = \frac{\partial}{\partial a} etc(s, a)$$

The value a^* such that $\mu'(a^*) = 0$ is approximated by the sequence $a_1, a_2, a_3, \ldots$ where a_1 is an arbitrary value,

$$a_{n+1} = a_n - \lambda_n \cdot y_n,$$

and y_n is an instance of the random variable $Y_n(a_n, x)$. Robins and Monro prove that this sequence converges to a^* if $\lambda_n = \frac{1}{n}$ and some additional requirements. Blum [1] relaxed the requirements proving convergence *with probability 1* if:

1. $\sum_{i=1}^{\infty} \lambda_i = \infty$,
2. $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$,
3. $\mu'(a) < 0$ for $a < a^*$ and $\mu'(a) > 0$ for $a > a^*$,
4. $\mu'(a) \leq c + d \cdot |a|$ for some $c, d \geq 0$,
5. $E_x((x - \mu(a))^2 \leq \sigma^2 < \infty$,
6. $\inf_{e_1 \leq |a - a^*| \leq e_2} |\mu(a)| > 0$ for every $0 < e_1 < e_2 < \infty$.

The problem in using this approach lays in determining $Y(a,x) = \frac{\partial}{\partial a}\left(rs(s,a,x) + dc(s,a)\right)$. We can calculate $\frac{\partial}{\partial a}dc(s,a)$ but $\frac{\partial}{\partial a}rs(s,a,x)$ must be derived from samples $rs(s,a,x) + dc(s,a)$. Especially because the rescheduling cost function $rs(s,a,x)$ need not be continuous at $x = a$, approximating $Y(a,x)$ is hard.

Kiefer and Wolfowitz [2] adapted the Robbins-Monro procedure to approximate the maximum of the expected value of a stochastic function. In this case, the stochastic function is described the random variable $Z(a,x) = -(rs(s,a,x) + dc(s,a))$ assuming that the state s is fixed. The expected value of this random variable is:

$$\mu(a) = E_x(Z(a,x)) = etc(s,a)$$

The value a^* such that $\mu(a^*)$ is *maximal* is approximated by the sequence $a_1, a_2, a_3, \ldots$ where the a_1 is an arbitrary value,

$$a_{n+1} = a_n + \frac{\lambda_n}{\eta_n} \cdot (z_{2n} - z_{2n-1}),$$

and z_{2n-1} and z_{2n} are instance of random variables $Z(a_n - \eta_n, x)$ and $Z(a_n + \eta_n, x)$, respectively. Blum [1] relaxes the convergence condition of Kiefer and Wolfowitz and proves convergence *with probability 1* if

1. $\lim_{n\to\infty} \eta_n = 0$,
2. $\sum_{n=1}^{\infty} \lambda_n = \infty$,
3. $\sum_{n=1}^{\infty} \left(\frac{\lambda_n}{\eta_n}\right)^2 < \infty$,
4. $\mu(a)$ is strictly decreasing for $a < a^*$ and strictly decreasing for $a > a^*$,
5. there exists two positive numbers ρ and r such that $|a' - a''| < \rho$ implies $|\mu(a') - \mu(a'')| < r$,
6. for every $d > 0$ there exists a $e > 0$ such that $|a - a^*| > d$ implies

$$\inf_{d/2>\epsilon>0} \frac{\mu(a+\epsilon) - \mu(a-\epsilon)}{\epsilon} > e.$$

Two important parameters of the Kiefer-Wolfowitz procedure are λ_n and η_n. Here, we choose for λ_n and η_n the functions: $\lambda_n = n^{-x}$ and $\eta_n = n^{-y}$. A series $\sum_{n=1}^{\infty} n^{-z}$ converges to a finite value for $z > 1$, and goes to infinity otherwise. Therefore, the first three requirements of the Kiefer-Wolfowitz procedure imply that $0.5 < x \leq 1$. If $x = 1$, then $0 < y < 0.5$, and if x approximates 0.5, then y must approximate 0. Experiments showed the best convergence results for the Kiefer-Wolfowitz procedure if we choose x close to 0.5 and y close to 0.

Despite optimizing the choices for λ_n and η_n, experiments with the Kiefer-Wolfowitz procedure showed that convergence is still slow. The problem is caused by a high variance in the total cost of an incident x: $Z(a,x) = -(rs(s,a,x) + dc(s,a))$. This causes a high variance in the approximation of the derivative $(z_{2n} - z_{2n-1})/\eta_n$, which is insufficiently damped by λ_n. To cope with this problem, we investigated the possibility of reducing the variance of $Z(a,x)$.

$Z(a,x) = -(rs(s,a,x) + dc(s,a))$ is the total cost of the current buffer size and rescheduling cost of the incident x. If instead of the current incident x, we use the average total cost over the last M incidents, we may significantly reduce the variance. Hence, we propose a new random variable

$$Z'_n(a) = \frac{1}{M}\sum_{i=1}^{M} Z(a, x_{M\cdot n+i})$$

The experiments will show that this enables much faster learning of the optimal buffer size.

4 Experiments

To evaluate the ability to learn an optimal buffer size, a series of experiments has been carried out. Below we will report on some of these experiments. The reported results are all based on the following settings:

- $p(x \mid s) = \frac{1}{3}e^{\frac{-x}{3}}$
- $dc(s, a) = a$
- $rs(s, a, x) = \begin{cases} 3 \cdot (x - a) + 5 & \text{if } x > a \\ 0 & \text{otherwise} \end{cases}$

Given these settings we can calculate the expected total cost $etc(s, a)$ for different buffer sizes. Figure 1 shows this cost. As we can see in the figure, the optimal buffer size to be learned in this example is 4.6.

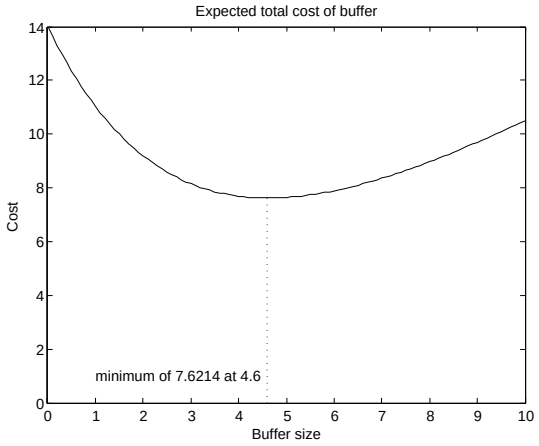

Figure 1: The expected total cost as a function of the buffer size.

We first tested the Kiefer-Wolfowitz procedure in its standard setting. That is, using the random variable $Z_n(a_n, x_n)$ which is equal to $Z'_n(a_n)$ for $M = 1$. Given these settings we evaluated different choices for the parameters λ_n and η_n. Figure 2 shows the results for $\lambda_n = n^{-0.51}$ and $\eta_n = n^{-0.005}$, for $\lambda_n = n^{-1}$ and $\eta_n = n^{-0.01}$, and for $\lambda_n = n^{-1}$ and $\eta_n = n^{-0.49}$. The figure shows the average buffer size over 100 experiments and standard deviation. Since we have the best result for $\lambda_n = n^{-0.51}$ and $\eta_n = n^{-0.005}$, we will use this setting in the following experiments.

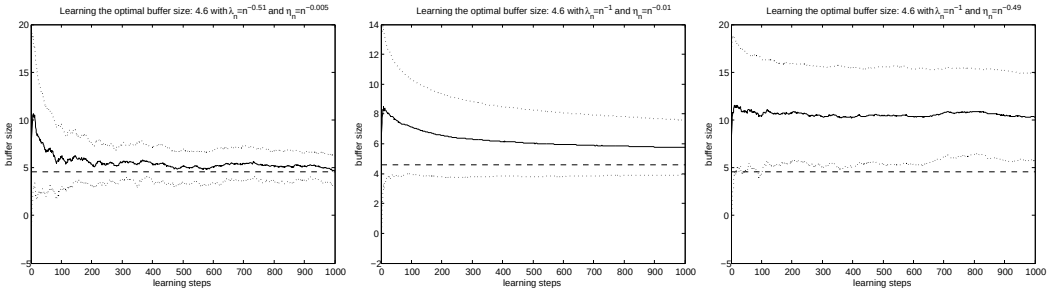

Figure 2: Learning the optimal buffer size using different choices for λ_n and η_n.

After determining the optimal choice for λ_n and η_n, we continued testing the Kiefer-Wolfowitz procedure in its standard setting. That is, using the random variable $Z_n(a_n, x_n)$ which is equal to $Z'_n(a_n)$ for $M = 1$. Figure 3 shows the results. The result shows a high variation in the total cost of an incident, and a significant variation in the buffer size around the optimal buffer size.

The next experiment shows the effect of averaging the total cost over M incidents given a buffer size a. Figure 4 shows the results for $M = 10$ (left) and $M = 100$ (right). Note that we adapted the number of learning steps such that in each experiment 20000 observations are used. As we can observe in the figures,

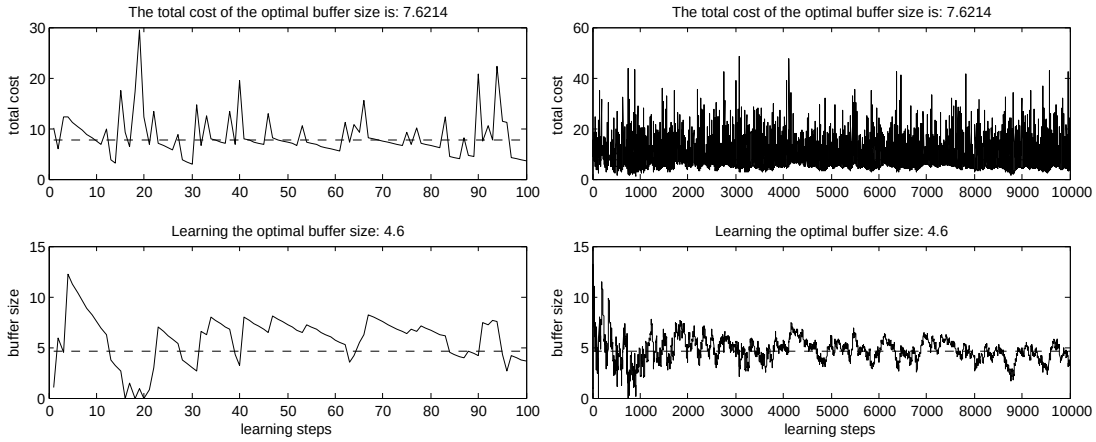

Figure 3: The total cost and the buffer size as function of the learning steps.

larger values for M result in a lower variance of $Z'(a)$ and a better approximation of a^*. Only 10 learning steps are needed to learn a good approximation of the buffer size where each learning step requires 200 observations of incidents (or the absence thereof). So, in total the information of 2000 observations are needed for a good approximation of the optimal buffer size.

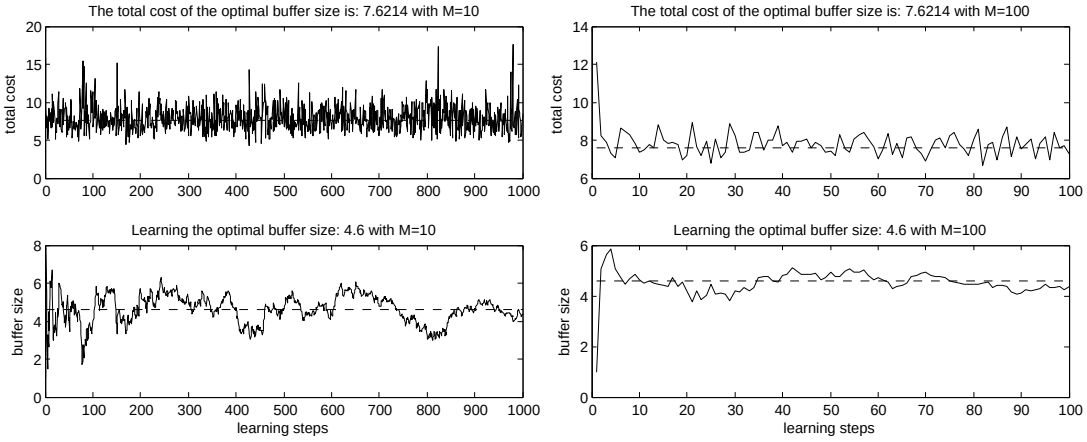

Figure 4: The estimated total cost and the buffer size as function of the learning steps.

Figure 4 shows a single learning sequence. The convergence shown in the figure could be a coincident. Therefore, in the next experiments, the average over 100 learning sequences is determined for $M = 1$, $M = 10$, $M = 100$ and $M = 1000$. Figure 5 shows the average buffer size that is learned as well as the standard deviation for $M = 1$, $M = 10$, $M = 100$ and $M = 1000$ (shown from left to right and from top to bottom, respectively). In every experiment 20000 observations are used.

Comparing the results of the experiments shown in Figure 5, we see that

- higher values for M reduce the standard deviation of the learning sequences;
- higher values for M decrease the learning speed;
- between $M = 10$ and $M = 100$ we have an optimal balance between the reduction of the standard deviation and the decrease of the learning speed.

The experiments also show that after 2000 observations, the difference between $M = 1$ and $M = 100$ is not in the average value of the 100 learning sequences but in the reduction of the standard deviation with a factor of about 3.5. 2000 observations correspond with 1000 and with 10 learning steps, respectively. Figure 5 also shows that using more observations does not significantly change this ratio.

In the last experiment we investigated how many learning steps are needed to using $M = 1$ to get the same standard deviation as we get using 10 learning steps and $M = 100$. The result in Figure 6 shows that

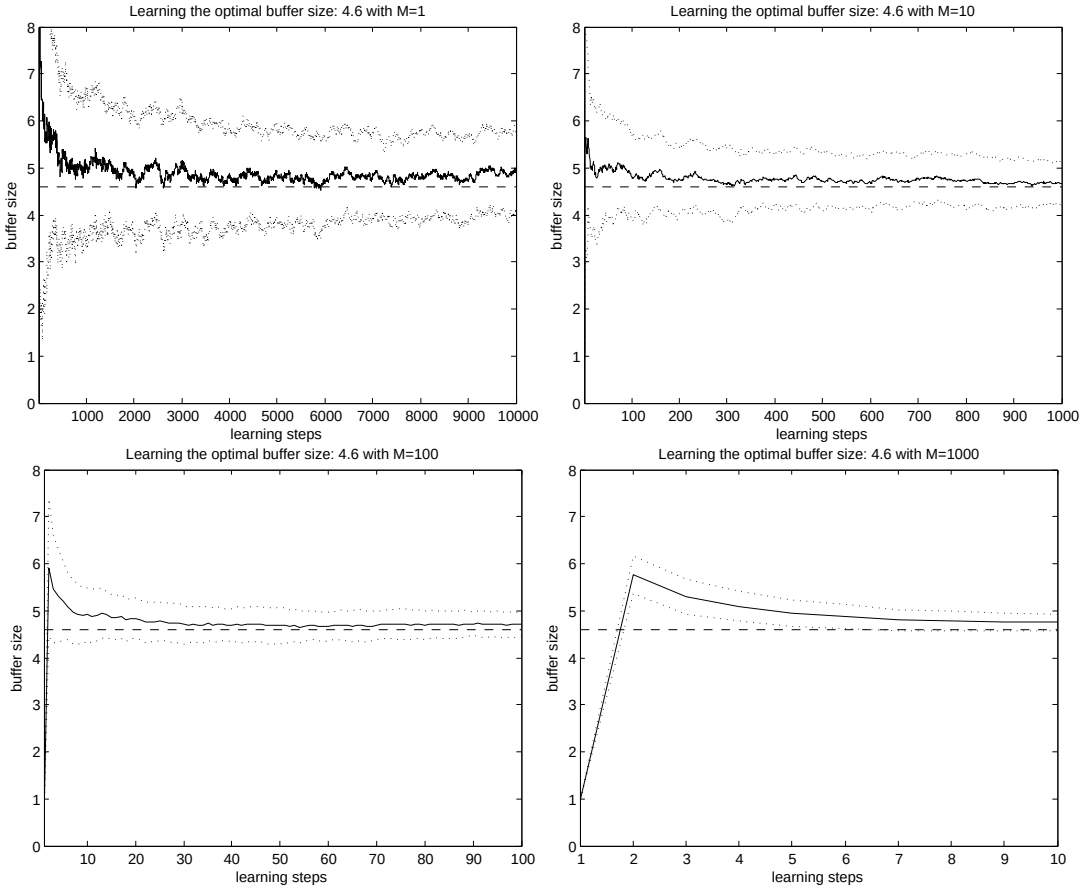

Figure 5: The average learned buffer size and the standard deviation over 100 learning sequences for $M = 1$, $M = 10$, $M = 100$ and $M = 1000$.

for $M = 1$ we need about $30,000$ learning steps to get more or less the same standard deviation. Hence, averaging over 100 observations for each learning step clearly improves the result with a factor 30.

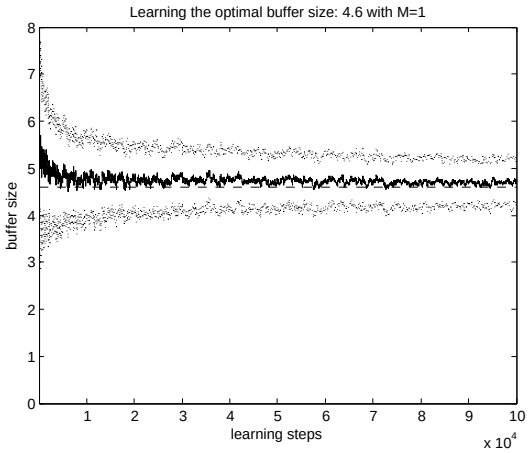

Figure 6: The average learned buffer size and the standard deviation over 100 learning sequences for $M = 1$.

5 Conclusion

This paper addressed an application of the stochastic convergence theory to the optimization problem of learning an optimal temporal buffer between airport ground handling services. The optimal buffer size must be learned using stochastic information of the costs of incidents. The result shows that because of the large variance in the stochastic information, convergence is very slow and the standard deviation over multiple experiments is high. By averaging the stochastic information before a learning step is made, the convergence can be accelerated and the standard deviation can be reduced. Future research will address dynamically determining the amount of stochastic information over which we average before making a learning step.

References

[1] Julius R. Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954.

[2] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[3] X. Mao, N. Roos, and A. Salden. Stable multi-project scheduling of airport ground handling services with heterogeneous agents. In *AAMAS 2009*, pages 537–544, 2009.

[4] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[5] C.J.C.H Watkins. *Learning from delayed rewards*. PhD thesis, Cambridge University, United Kingdom, 1989.