Solving Non-Stationary Bandit Problems by Random Sampling from Sibling Kalman Filters

Ole-Christoffer Granmo and Stian Berg

Department of ICT, University of Agder, Grimstad, Norway

Abstract. The multi-armed bandit problem is a classical optimization problem where an agent sequentially pulls one of multiple arms attached to a gambling machine, with each pull resulting in a random reward. The reward distributions are unknown, and thus, one must balance between exploiting existing knowledge about the arms, and obtaining new information. Dynamically changing (non-stationary) bandit problems are particularly challenging because each change of the reward distributions may progressively degrade the performance of any fixed strategy.

Although computationally intractable in many cases, Bayesian methods provide a standard for optimal decision making. This paper proposes a novel solution scheme for bandit problems with non-stationary normally distributed rewards. The scheme is inherently Bayesian in nature, yet avoids computational intractability by relying simply on updating the hyper parameters of sibling Kalman Filters, and on random sampling from these posteriors. Furthermore, it is able to track the better actions, thus supporting non-stationary bandit problems.

Extensive experiments demonstrate that our scheme outperforms recently proposed bandit playing algorithms, not only in non-stationary environments, but in stationary environments also. Furthermore, our scheme is robust to inexact parameter settings. We thus believe that our methodology opens avenues for obtaining improved novel solutions.

Keywords: Bandit Problems, Kalman Filter, Bayesian Learning.

1 Introduction

The conflict between exploration and exploitation is a well-known problem in reinforcement learning, and other areas of artificial intelligence. The multi-armed bandit problem captures the essence of this conflict, and has thus occupied researchers for over fifty years [1]. In [2], a new family of *Bayesian* techniques for solving the classical two-armed Bernoulli bandit problem was introduced, akin to the *Thompson Sampling* [3] principle, and empirical results that demonstrated its advantages over established top performers were reported.

The above mentioned solution schemes were mainly designed for stationary bandit problems, leading to fixed arm selection strategies. However, in many real-life applications, such as web polling [4], the associated bandit problems are changing with time, making them non-stationary. Thus, if an algorithm with a

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part III, LNAI 6098, pp. 199–208, 2010.

fixed strategy is applied to bandit problem that is changing, each change may progressively degrade the performance of the algorithm [5].

In this present paper, we address the *Non-Stationary* Multi-Armed Normal Bandit (MANB) problem. In brief, we propose a Bayesian solution for non-stationary normally distributed rewards, that has sufficient flexibility to track the better actions as changes occur over time.

1.1 The Non-Stationary Multi-Armed Normal Bandit Problem

The MANB problem is a classical optimization problem that explores the trade off between exploitation and exploration in, e.g., reinforcement learning. The problem consists of an agent that sequentially pulls one of multiple arms attached to a gambling machine, with each pull resulting in a reward. The reward obtained from each $Arm\ i$ has been affected by Gaussian noise of variance σ_{ob}^2 (observation noise), with the true unperturbed reward r_i being unknown.

This leaves the agent with the following dilemma: Should the arm that so far seems to be associated with the largest reward r_i be pulled once more, or should an inferior arm be pulled in order to learn more about *its* reward? Sticking prematurely with the arm that is presently considered to be the best one, may lead to not discovering which arm is truly optimal. On the other hand, lingering with the inferior arm unnecessarily, postpones the harvest that can be obtained from the optimal arm.

The non-stationary MANB problem renders the above problem even more intriguing because it allows the true unperturbed rewards r_i to change with time. In this paper, we address problems where each reward r_i is modified by independent Gaussian perturbations of constant variance σ_{tr}^2 (transition noise) at each time step.

In effect, a solution scheme for non-stationary MANB problems must thus both determine which action is the best one, as well as tracking any reward distribution changes that might occur.

1.2 Contributions and Paper Organization

The contributions of this paper can be summarized as follows. In Sect. 2, we briefly review a selection of the main MANB solution approaches. Then, in Sect. 3, we present our Kalman Filter based solution to MANB (KF-MANB). The KF-MANB scheme is inherently Bayesian in nature, even though it only relies on simple updating of hyper parameters and random sampling. Thus, the MANB solution scheme takes advantage of the Bayesian perspective in a computationally efficient manner. In addition, the scheme is able to track the best arms when the problem is non-stationary, relying on a set of sibling Kalman filters. In Sect. 4, we provide extensive experimental results that demonstrate that the KF-MANB scheme outperforms established top performers, both for stationary (!) and non-stationary bandit problems. Accordingly, from the above perspective, it is our belief that the KF-MANB scheme represents the current state-of-the-art and a new avenue of research. Finally, in Sect. 5 we list open KF-MANB related research problems and provide concluding remarks.

2 Related Work

From a broader point of view, one can distinguish two distinct fields that address bandit like problems, namely, the field of Learning Automata and the field of Bandit Playing Algorithms. A myriad of approaches have been proposed within these two fields, and we refer the reader to [5] and [6] for an overview and comparison of schemes. We here provide a brief review of selected top performers in order to cast light on the distinguishing properties of KF-MANB.

Learning Automata (LA) have been used to model biological systems [5] and have attracted considerable interest in the last decade because they can learn the optimal action when operating in (or interacting with) unknown stochastic environments. Furthermore, they combine rapid and accurate convergence with low computational complexity. More notable approaches include the family of linear updating schemes, with the Linear Reward-Inaction (L_{R-I}) automaton being designed for stationary environments [5]. In short, the L_{R-I} maintains an arm probability selection vector $\bar{p} = [p_1, p_2]$, with $p_2 = 1 - p_1$. The question of which arm is to be pulled is decided randomly by sampling from $\bar{p}$, which initially is uniform. Upon receiving a reward, a linear updating rule increases the probability of selecting the rewarded arm in the future, allowing the scheme to achieve ϵ-optimality [5]. A *Pursuit scheme* (P-scheme) makes this updating scheme more goal-directed by maintaining Maximum Likelihood (ML) estimates of the unperturbed rewards associated with each arm. Instead of using the rewards that are received to update $\bar{p}$ directly, they are rather used to update the ML estimates. The ML estimates, in turn, are used to decide which arm selection probability p_i is to be increased.

The **ϵ-greedy rule** is another well-known strategy for the bandit problem [7]. In short, the arm with the presently highest average reward is pulled with probability $1 - \epsilon$, while a randomly chosen arm is pulled with probability ϵ. In other words, the balancing of exploration and exploitation is controlled by the ϵ-parameter. Note that the ϵ-greedy strategy persistently explores the available arms with constant effort, which clearly is sub-optimal for the MANB problem. It is even sub-optimal for non-stationary MANB problems because the ϵ-greedy strategy maintains strongly converging ML estimates of the unperturbed rewards. As a remedy for the former (but not the latter) problem, ϵ can be slowly decreased, leading to the ϵ_n-greedy strategy described in [8]. The purpose is to gradually shift focus from exploration to exploitation.

A promising line of thought is the **interval estimation** methods, where a confidence interval for the unperturbed reward of each arm is estimated, and an "optimistic reward estimate" is identified for each arm. The arm with the most optimistic reward probability estimate is then greedily selected [6, 9]. The IntEst (Interval Estimation) scheme [9] was one of the first schemes to use optimistic reward estimates to achieve exploration. Many variants of the IntEst scheme have been proposed — one for normally distributed rewards is described in [6]. In [8], several confidence interval based algorithms are analysed. These algorithms provide logarithmically increasing *Regret*, with Ucb1-Normal targeting normally distributed rewards.

The application of a **Bayesian philosophy** for searching in such probabilistic spaces has a long and well-documented path through the mathematical literature, probably pioneered by Thompson [3] even as early as 1933, in the form of the Thompson sampling principle. According to this principle, each arm should be chosen a fraction of the time that corresponds to the probability that the action is optimal. This principle was recently rediscovered by other authors, e.g., in [1], such a Bayesian philosophy appears in the so-called *probability matching algorithms*. By using conjugate priors, these authors have resorted to a Bayesian analysis to obtain a closed form expression for the probability that each arm is optimal given the prior observed rewards/penalties. More recently, however, Wang *et al* [10] combined so-called sparse sampling with Bayesian exploration, where the entire process was specified within a finite time horizon. By achieving this, they were able to search the space based on a sparse look-ahead tree which was produced based on the Thompson Sampling principle. In [11], the author derived optimal decision thresholds for the multi-armed bandit problem, for both the infinite horizon discounted reward case and for the finite horizon undiscounted case. Based on these thresholds, he then went on to propose practical algorithms, which can perhaps be perceived to be enhancements of the *Thompson Sampling* principle. Finally, the authors of [12] take advantage of a Bayesian strategy in a related domain, i.e., in Q-learning. They show that for normally distributed rewards, in which the parameters have a prior normal-gamma distribution, the posteriors also have a normal-gamma distribution, rendering the computation efficient. They then integrate this into a framework for Bayesian Q-learning by maintaining and propagating probability distributions over the Q-values, and suggest that a non-approximate solution can be obtained by means of random sampling for the normal distribution case. In a similar vein, a scheme for Gaussian Temporal Difference learning is proposed in [13], with on-line learning of the posterior moments of a value Gaussian process. It would be interesting to investigate the applicability of these results for non-stationary MANB problems.

A more recent technique, the "Price of Knowledge and Estimated Reward" (POKER) algorithm proposed in [6], attempts to combine the following three principles: (1) Reducing uncertainty about the arm rewards should grant a bonus to stimulate exploration; (2) Information obtained from pulling arms should be used to estimate the properties of arms that have not yet been pulled; and (3) Knowledge about the number of rounds that remains (the horizon) should be used to plan the exploitation and exploration of arms. We refer the reader to [6] for the specific algorithm that incorporates these three principles.

3 Kalman Filter Based Solution to Non-Stationary Normal Bandit Problems (KF-MANB)

Bayesian reasoning is a probabilistic approach to inference which is of importance in machine learning because it allows quantitative weighting of evidence supporting alternative hypotheses, with the purpose of allowing optimal decision making. It also provides a framework for analyzing learning algorithms [14].

We here present a scheme for solving the non-stationary MANB problem that inherently builds upon the Bayesian reasoning framework, taking advantage of the tracking capability of Kalman filters [15]. We thus coin the scheme *Kalman Filter Based Solution to MANB* (KF-MANB). Essentially, KF-MANB uses the Kalman filter for two purposes. First of all, the Kalman filter is used to provide a *Bayesian estimate* of the rewards associated with each of the available bandit arms. Secondly, a novel feature of KF-MANB is that it uses the Kalman filters as the basis for a *randomized* arm selection mechanism. Being based on the Kalman filter, the normal distribution is central to KF-MANB:

$$f(x_i; \mu_i, \sigma_i) = \alpha \ e^{-\frac{1}{2}\left(\frac{(x_i - \mu_i)^2}{\sigma_i^2}\right)}$$

with μ_i being the mean of the distribution, σ_i^2 being its variance, and α is a normalization constant. The following algorithm contains the essence of the KF-MANB approach. Note that in order to emphasize the simplicity of the KF-MANB algorithm, the Kalman filters are incorporated into the parameter updating of the algorithm itself, and do not appear as distinct entities.

Algorithm: KF-MANB
Input: Number of bandit arms q; Observation noise σ_{ob}^2; Transition noise σ_{tr}^2.
Initialization: $\mu_1[1] = \mu_2[1] = \cdots = \mu_q[1] = A$; $\sigma_1[1] = \sigma_2[1] = \cdots = \sigma_q[1] = B$; *# Typically, A can be set to* 0*, with B being sufficiently large.*
Method:
For $N = 1, 2, \ldots$ **Do**

1. For each *Arm* $j \in \{1, \ldots, q\}$, draw a value x_j randomly from the associated *normal* distribution $f(x_j; \mu_j[N], \sigma_j[N])$ with the parameters $(\mu_j[N], \sigma_j[N])$.
2. Pull the *Arm* i whose drawn value x_i is the largest one:

$$i = \underset{j \in \{1, \ldots, q\}}{\text{argmax}} \ x_j.$$

3. Receive a reward $\tilde{r}_i$ from pulling *Arm* i, and update parameters as follows:
   - *Arm* i:

$$\mu_i[N+1] = \frac{\left(\sigma_i^2[N] + \sigma_{\text{tr}}^2\right) \cdot \tilde{r}_i + \sigma_{\text{ob}}^2 \cdot \mu_i[N]}{\sigma_i^2[N] + \sigma_{\text{tr}}^2 + \sigma_{\text{ob}}^2}$$

$$\sigma_i^2[N+1] = \frac{\left(\sigma_i^2[N] + \sigma_{\text{tr}}^2\right) \sigma_{\text{ob}}^2}{\sigma_i^2[N] + \sigma_{\text{tr}}^2 + \sigma_{\text{ob}}^2}$$

   - *Arm* $j \neq i$:

$$\mu_j[N+1] = \mu_j[N]$$

$$\sigma_j^2[N+1] = \sigma_j^2[N] + \sigma_{\text{tr}}^2$$

End Algorithm: KF-MANB

As seen from the above KF-MANB algorithm, N is a discrete time index and the parameters $\phi^N = \langle(\mu_1[N], \sigma_1[N]), (\mu_2[N], \sigma_2[N]), \ldots, (\mu_q[N], \sigma_q[N])\rangle$ form an infinite $2 \times q$-dimensional continuous state space, with each pair $(\mu_i[N], \sigma_i[N])$ giving the prior of the unknown reward r_i associated with *Arm i*. Within Φ the KF-MANB navigates by transforming each prior normal distribution into a posterior normal distribution, based on the reward obtained at that time step, the observation noise, and the transition noise.

Since the state space of KF-MANB is both continuous and infinite, KF-MANB is quite different from both the *Variable Structure-* and the *Fixed Structure* LA families [5], traditionally referred to as *Learning Automata*. In all brevity, the novel aspects of the KF-MANB are listed below:

1. In traditional LA, the action chosen (i.e, arm pulled) is based on the so-called action probability vector. The KF-MANB does not maintain such a vector, but chooses the arm based on the *distribution* of the components of the *Estimate* vector.
2. The second difference is that we have not chosen the arm based on the *a posteriori* distribution of the estimate. Rather, it has been chosen based on the magnitude of a *random sample* drawn from the *a posteriori* distribution, and thus it is more appropriate to state that the arm is chosen based on the *order of statistics* of instances of these variables[1].

In the interest of notational simplicity, let *Arm* 1 be the Arm under investigation. Then, for any parameter configuration $\phi^N \in \Phi$ we can state, using a generic notation[2], that the probability of selecting *Arm* 1 is equal to the probability $P(X_1^N > X_2^N \wedge X_1^N > X_3^N \wedge \cdots \wedge X_1^N > X_q^N | \phi^N)$ — the probability that a randomly drawn value $x_1 \in X_1^N$ is greater than all of the other randomly drawn values $x_j \in X_j^N, j \neq i$, at time step N, when the associated stochastic variables $X_1^N, X_2^N, \ldots, X_q^N$ are *normally* distributed, with parameters $(\mu_1[N], \sigma_1[N]), (\mu_2[N], \sigma_2[N]), \ldots, (\mu_q[N], \sigma_q[N])$ respectively. In the following, we will let $p_1^{\phi^N}$ denote this latter probability.

Note that the probability $p_1^{\phi^N}$ can also be interpreted as the probability that *Arm* 1 is the optimal one, given the observations ϕ^N. The formal result that we will derive in the unabridged paper shows that the KF-MANB will gradually shift its arm selection focus towards the arm which most likely is the optimal one, as the observations are received.

4 Empirical Results

In this section we evaluate the KF-MANB by comparing it with the best performing algorithms from [8, 6], as well as the Pursuit scheme, which can be seen

[1] To the best of our knowledge, the concept of having automata choose actions based on the *order of statistics* of instances of estimate distributions, has been unreported in the literature.

[2] By this we mean that P is not a fixed function. Rather, it denotes the probability function for a random variable, given as an argument to P.

as an established top performer in the LA field. Based on our comparison with these "reference" algorithms, it should be quite straightforward to also relate the KF-MANB performance results to the performance of other similar algorithms.

Although we have conducted numerous experiments using various reward distributions, we here report, for the sake of brevity, results for 10-armed stationary and non-stationary MANB problems, measured in terms of *Regret*. *Regret* is simply *the difference between the sum of rewards expected after N successive arm pulls, and what would have been obtained by only pulling the optimal arm.* For these experiment configurations, an ensemble of 1000 independent replications with different random number streams was performed to minimize the variance of the reported results. In each replication, 2000 arm pulls were conducted in order to examine both the short term and the limiting performance of the evaluated algorithms.

We first investigate the effect observation noise has on performance in stationary environments. The difficulty of stationary bandit problems is determined by the "signal-to-noise ratio" of the problem, that is, the ratio between arm *reward difference* and the *standard deviation* of the observation noise. In brief, if the arms are ranked according to their index i, we let the difference in reward between *Arm* i and *Arm* $i+1$ be 50.0 ($r_i - r_{i+1} = 50.0$). With the scale thus being set, we vary observation noise to achieve a wide range of signal-to-noise ratios: $\sigma_{\text{ob}} \in [\frac{1}{4} \cdot 50.0, \frac{1}{3} \cdot 50.0, \frac{1}{2} \cdot 50.0, \frac{2}{3} \cdot 50.0, \frac{4}{5} \cdot 50.0, 50.0, 1\frac{1}{4} \cdot 50.0, 1\frac{1}{2} \cdot 50.0, 2 \cdot 50.0, 3 \cdot 50.0, 4 \cdot 50.0]$. Thus, for $\sigma_{\text{ob}} = \frac{1}{4} \cdot 50.0$ the observation noise is small compared to the difference between rewards, and distinguishing which arm is best is correspondingly easier. Conversely, an observation noise of $\sigma_{\text{ob}} = 4 \cdot 50.0$ makes the true difference between arms in the ranking fall within 1/4 standard deviation of the noise. Accordingly, discriminating between the arms in this case is correspondingly more challenging.

Table 1 reports normalized *Regret* after 2000 arm pulls for each algorithm, with suitable parameter settings. As seen in the table, KF-MANB is

Table 1. Normalized regret for transition noise $\sigma_{tr} = 0$ (stationary environments)

Algorithm / σ_{ob}	**12.5**	**16.7**	**25.0**	**33.3**	**50.0**	**75.0**	**100.0**	**150.0**	**200.0**
BLA Kalman	*5.1*	*5.2*	*5.4*	*5.8*	*7.1*	*10.6*	*16.1*	*28.5*	*45.0*
Ucb1 Normal	135.0	135.0	135.2	136.3	140.7	151.8	168.9	213.9	270.5
IntEst 0.1	608.5	595.1	537.2	515.0	450.3	367.3	298.5	229.4	188.9
IntEst 0.2	581.7	579.7	559.3	549.9	526.1	442.8	422.3	357.6	320.2
IntEst 0.05	558.6	585.6	522.0	484.3	420.1	325.4	239.7	158.1	150.0
Pursuit 0.050	45.6	41.2	45.4	47.0	50.8	52.7	71.0	85.7	114.6
Pursuit 0.010	52.3	52.2	52.3	53.0	53.3	55.5	58.6	69.0	78.1
Pursuit 0.005	101.7	101.9	102.2	102.4	103.3	105.0	106.4	112.4	122.9
Poker	5.5	5.7	5.5	7.2	10.6	19.1	29.5	42.6	58.7
ϵ_n-Greedy $c = 0.3$	29.0	28.6	28.1	30.5	33.5	39.8	50.2	74.1	97.8
ϵ_n-Greedy $c = 0.6$	37.9	37.7	37.9	38.3	40.4	46.1	54.1	68.0	84.4
ϵ_n-Greedy $c = 0.9$	65.3	65.5	65.5	65.2	65.4	68.1	70.3	86.5	99.0

superior under all tested degrees of observation noise, with the POKER algorithm being a close second best. Note that POKER is given the total number of arm pulls to be used in each run. This information is used to prioritize exploitation when further exploration will not pay off due to the limited number of arm pulls remaining. The Pursuit scheme and the e_n-Greedy-schemes provide significantly higher *Regret*. Furthermore, IntEst clearly provides the highest *Regret* among the compared algorithms. Corresponding results concerning the probability of selecting the best arm after each arm pull can be found in the unabridged paper. These results show that KF-MANB converges to selecting the best arm more accurately and more quickly than the other schemes.

The fact that KF-MANB outperforms the other schemes in stationary environments is rather surprising since most of the competing algorithms also assume normally distributed rewards (which is the sole assumption of KF-MANB in this experiment). This performance advantage can perhaps be explained by KF-MANBs capability of reasoning with the known observation noise. However, as will be seen below, the performance advantage of KF-MANB is not overly sensitive to incorrectly specified observation noise.

The advantage of KF-MANB becomes even more apparent when dealing with bandit problems that are non-stationary. In Table 2, normalized *Regret* is reported for $\sigma_{\text{ob}} = 50.0$, with varying transition noise: $\sigma_{\text{tr}} \in [0.0, \frac{1}{4} \cdot 50.0, \frac{1}{3} \cdot 50.0, \frac{1}{2} \cdot 50.0, \frac{2}{3} \cdot 50.0, \frac{4}{5} \cdot 50.0, 50.0, 1\frac{1}{4} \cdot 50.0, 1\frac{1}{2} \cdot 50.0, 2 \cdot 50.0, 3 \cdot 50.0, 4 \cdot 50.0]$. As seen in the table, KF-MANB is superior for the non-stationary MANB problems. Also, notice how the degree of transition noise affects its performance to a very small degree, demonstrating the power of the KF-MANB scheme. Conversely, the other schemes, being unable to track reward changes, are obtaining a progressively worse *Regret* as the number of arm pulls increases.

Table 2. Normalized regret for observation noise $\sigma_{ob} = 50.0$, with varying transition noise σ_{tr} (non-stationary environments)

Algorithm / σ_{tr}	**0.0**	**12.5**	**16.7**	**25.0**	**33.3**	**50.0**	**75.0**	**100.0**	**150.0**	**200.0**
BLA Kalman	*7.2*	*43.0*	*44.9*	*45.9*	*47.6*	*48.4*	*49.3*	*49.8*	*49.3*	*49.3*
Ucb1 Normal	140.6	260.5	269.4	271.1	286.0	281.7	287.5	286.2	278.6	280.2
IntEst 0.1	456.4	488.1	470.2	460.6	476.3	475.3	470.6	481.9	471.5	468.6
IntEst 0.2	506.0	536.7	523.2	519.5	519.7	521.0	498.4	514.7	499.6	509.0
IntEst 0.05	429.1	446.9	431.3	439.4	441.2	463.9	447.7	454.8	449.4	460.0
Pursuit 0.050	51.8	389.0	407.5	437.9	485.9	487.7	497.7	515.4	516.0	521.3
Pursuit 0.010	53.0	365.7	400.0	407.1	451.2	456.0	448.4	452.9	456.5	440.9
Pursuit 0.005	103.1	393.3	418.8	422.0	462.1	453.5	447.5	463.0	448.7	455.4
Poker	12.0	297.2	332.8	361.4	391.2	416.1	409.7	422.9	416.8	419.7
ϵ_n-Greedy $c = 0.3$	33.8	304.0	336.3	341.7	379.9	379.1	371.3	381.6	374.9	386.4
ϵ_n-Greedy $c = 0.6$	39.3	304.7	320.8	335.0	368.7	374.3	374.0	378.9	364.6	370.2
ϵ_n-Greedy $c = 0.9$	65.7	320.0	338.3	338.3	370.3	367.3	370.7	371.4	366.0	371.3

Table 3. Sensitivity to incorrectly specified observation noise with $\sigma_{tr} = 0.0$

Belief $\hat{\sigma}_{ob}$:	$\frac{1}{4}\sigma_{ob}$	$\frac{1}{3}\sigma_{ob}$	$\frac{1}{2}\sigma_{ob}$	$\frac{2}{3}\sigma_{ob}$	$\frac{4}{5}\sigma_{ob}$	σ_{ob}	$1\frac{1}{4}\sigma_{ob}$	$1\frac{1}{2}\sigma_{ob}$	$2\sigma_{ob}$	$3\sigma_{ob}$	$4\sigma_{ob}$
Regret:	13.2	*10.3*	*5.7*	*5.9*	*6.4*	*7.1*	*8.3*	*10.2*	14.9	27.1	41.3

Table 4. Sensitivity to incorrectly specified transition noise with $\sigma_{ob} = 50.0$

Belief $\hat{\sigma}_{tr}$:	$\frac{1}{4}\sigma_{tr}$	$\frac{1}{3}\sigma_{tr}$	$\frac{1}{2}\sigma_{tr}$	$\frac{2}{3}\sigma_{tr}$	$\frac{4}{5}\sigma_{tr}$	σ_{tr}	$1\frac{1}{4}\sigma_{tr}$	$1\frac{1}{2}\sigma_{tr}$	$2\sigma_{tr}$	$3\sigma_{tr}$	$4\sigma_{tr}$
Regret:	*168.8*	*113.5*	*66.6*	*46.5*	*46.1*	*48.4*	*59.6*	*74.6*	*106.4*	*159.0*	*225.6*

One of the advantages of KF-MANB is its ability to take advantage of knowledge about the observation noise and transition noise affecting a MANB problem. In cases where the exact observation noise and transition noise is unavailable, KF-MANB may have to operate with incorrectly specified noise. Table 3 summarizes the effect incorrect specification of observation noise has on normalized *Regret* for KF-MANB. As seen, the KF-MANB algorithm is still the top performer for the tested stationary MANB problems, provided the incorrectly specified observation noise has a standard deviation $\hat{\sigma}_{ob}$ that lies within 33% and 150% of the true standard deviation σ_{ob} (!). Also observe that the performance even improves when $\hat{\sigma}_{ob}$ is set slightly below σ_{ob}. A reasonable explanation for this phenomenon is that the KF-MANB scheme is too conservative when planning its arm pulls, probably because it does not consider the so-called *gain in information*, only the present probability of selecting the correct arm. Therefore, it is likely that a slightly too low $\hat{\sigma}_{ob}$ compensates for this by making the KF-MANB more “aggressive”.

We observe similar results when transition noise is incorrectly specified, as seen in Table 4. Again, the KF-MANB is surprisingly unaffected by incorrectly specified noise. Indeed, it provides reasonable performance when operating with a standard deviation $\hat{\sigma}_{tr}$ that lies within 50% and 150% of σ_{tr}.

From the above results, we conclude that KF-MANB is the superior choice for MANB problems in general, providing significantly better performance in the majority of the experiment configurations.

5 Conclusion and Further Work

In this paper we presented a*Kalman Filter based solution scheme to Multi-Armed Normal Bandit* (KF-MANB) problems. In contrast to previous LA and *Regret* minimizing approaches, KF-MANB is inherently Bayesian in nature. Still, it relies simply on updating the hyper parameters of sibling Kalman Filters, and on random sampling from these. Thus, KF-MANB takes advantage of Bayesian estimation in a computationally efficient manner. Furthermore, extensive experiments demonstrate that our scheme outperforms recently proposed algorithms, dealing particularly well with non-stationary problems. Accordingly, in the above perspective, it is our belief that the KF-MANB represents a new promising avenue of research, with a number of interesting applications. In further work, we intend to study systems of KF-MANB from a game theory point of view, where multiple KF-MANBs interact forming the basis for multi-agent systems.

References

1. Wyatt, J.: Exploration and Inference in Learning from Reinforcement. PhD thesis, University of Edinburgh (1997)
2. Granmo, O.C.: Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. To Appear in the International Journal of Intelligent Computing and Cybernetics (2010)
3. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25, 285–294 (1933)
4. Granmo, O.C., Oommen, B.J., Myrer, S.A., Olsen, M.G.: Learning Automata-based Solutions to the Nonlinear Fractional Knapsack Problem with Applications to Optimal Resource Allocation. IEEE Transactions on Systems, Man, and Cybernetics, Part B 37(1), 166–175 (2007)
5. Narendra, K.S., Thathachar, M.A.L.: Learning Automata: An Introduction. Prentice-Hall, Englewood Cliffs (1989)
6. Vermorel, J., Mohri, M.: Multi-armed bandit algorithms and empirical evaluation. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 437–448. Springer, Heidelberg (2005)
7. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
8. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning 47, 235–256 (2002)
9. Kaelbling, L.P.: Learning in Embedded Systems. PhD thesis, Stanford University (1993)
10. Wang, T., Lizotte, D., Bowling, M., Scuurmans, D.: Bayesian sparse sampling for on-line reward optimization. In: Proceedings of the 22nd International conference on Machine learning, pp. 956–963 (2005)
11. Dimitrakakis, C.: Nearly optimal exploration-exploitation decision thresholds. In: Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4131, pp. 850–859. Springer, Heidelberg (2006)
12. Dearden, R., Friedman, N., Russell, S.: Bayesian q-learning. In: AAAI/IAAI, pp. 761–768. AAAI Press, Menlo Park (1998)
13. Engel, Y., Mannor, S., Meir, R.: Reinforcement learning with gaussian processes. In: Proceedings of the 22nd International conference on Machine learning, pp. 956–963 (2005)
14. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
15. Russel, S., Norvig, P.: Artificial Intelligence - A Modern Approach, 2nd edn. Prentice-Hall, Englewood Cliffs (2003)