bioRxiv preprint doi: https://doi.org/10.1101/117598; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Bridging Computational Neuroscience and Machine Learning on Non-Stationary Multi-Armed Bandits

**George Velentzas**
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
geovelentzas@gmail.com

**Costas Tzafestas**
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
ktzaf@cs.ntua.gr

**Mehdi Khamassi**
Institute of Intelligent Systems and Robotics
Sorbonne Universités, UPMC Univ Paris 06, CNRS
F-75005 Paris, France
mehdi.khamassi@upmc.fr

## Abstract

Fast adaptation to changes in the environment requires both natural and artificial agents to be able to dynamically tune an exploration-exploitation trade-off during learning. This trade-off usually determines a fixed proportion of exploitative choices (i.e. choice of the action that subjectively appears as best at a given moment) relative to exploratory choices (i.e. testing other actions that now appear worst but may turn out promising later). The problem of finding an efficient exploration-exploitation trade-off has been well studied both in the Machine Learning and Computational Neuroscience fields. Rather than using a fixed proportion, non-stationary multi-armed bandit methods in the former have proven that principles such as exploring actions that have not been tested for a long time can lead to performance closer to optimal – bounded regret. In parallel, researches in the latter have investigated solutions such as progressively increasing exploitation in response to improvements of performance, transiently increasing exploration in response to drops in average performance, or attributing exploration bonuses specifically to actions associated with high uncertainty in order to gain information when performing these actions. In this work, we first try to bridge some of these different methods from the two research fields by rewriting their decision process with a common formalism. We then show numerical simulations of a hybrid algorithm combining bio-inspired meta-learning, kalman filter and exploration bonuses compared to several state-of-the-art alternatives on a set of non-stationary stochastic multi-armed bandit tasks. While we find that different methods are appropriate in different scenarios, the hybrid algorithm displays a good combination of advantages from different methods and outperforms these methods in the studied scenarios.

**Keywords:** Bandits; Decision-making; Meta-learning; Active exploration; Kalman filter; Reinforcement learning; Multi-armed bandit

### Acknowledgements

This research work has been partially supported by the EU-funded Project BabyRobot (H2020-ICT-24-2015, grant agreement no. 687831), by the Agence Nationale de la Recherche (ANR-12-CORD-0030 Roboergosum Project and ANR-11-IDEX-0004-02 Sorbonne-Universités SU-15-R-PERSU-14 Robot Parallearning Project), and by Labex SMART (ANR-11-LABX-65 Online Budgeted Learning Project).

bioRxiv preprint doi: https://doi.org/10.1101/117598; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 1 Introduction

Selecting an action from a number of distinctive alternatives in unknown environments, where each action may return a reward in a stochastic manner, can be a challenging problem when the environment is non-stationary (i.e the rewards are changing in time). The adaptive learning strategy of the decision maker, especially when facing abrupt changes, requires a sophisticated method for the well known exploration-exploitation trade-off. Modeling the problem for implementing intelligent decision-makers is a subject mainly studied from the machine learning field, while the computational neuroscience field is concerned about finding models that fit human's or animal's behavior efficiently. In the former the task is to find a decision-making algorithm whose actions achieve optimal cumulative regret - the regret being defined as the difference between optimal performance and the accumulated reward. Solutions to the stationary problem have been studied thoroughly for more than a decade, but the non-stationary approach was a more recent subject of research [1,2], while its recent recapturing of attention is particularly due to its characteristic as a keystone in reinforcement learning. In parallel, computational neuroscience researches have also progressed in combining methods from reinforcement learning with exploration bonuses [3] and sometimes change-point detectors [4] to tackle the problem of fitting a human's or an animal's action decisions on bandit tasks with altering rewards in time.

With the dual goal of conceiving adaptive machines with cognitive components (e.g for human-robot interaction environments) and of generating new experimental predictions for neuroscience, we explored the impact of integrating approaches from both fields at the lower level of analysis for a simple artificial decision-maker. We tested the performance of several alternative algorithms on a set of different non-stationary bandit setups where we variated the most crucial components of a stochastic and changing environment. Here we present empirical evidence that our hybrid approach outperforms some state-of-the-art alternatives, rising questions about how the integration of neuroscience-based models and machine learning-models could have a positive impact on performance.

# 2 Problem Formulation and Algorithms

Stochastic multi armed-bandits can be considered as having a set of arms $\mathcal{K} = \{1, 2, ..., K\}$ of a casino machine, where a decision-maker chooses an action $a \in \mathcal{K}$ at every timestep $t \in \mathcal{T} = \{1, 2, ..., T\}$ of its lifespan $T$ (also called the time horizon). He then receives a reward $r_t(a)$ with probability $p_t(a)$ and zero otherwise. While interested in arm's expected value of reward, without loss of generality, we can assume Bernoulli arms where $\forall(a, t) \in \mathcal{K} \times \mathcal{T}$ the rewards are binary, $r_t(a) \in \{0, 1\}$. By choosing Bernoulli arms, the expected value $\mathbb{E}[r_t(a)]$ for every arm reward is equal to the probability $p_t(a)$. When the environment is stationary, $p_{t+1}(a) = p_t(a)$ for all time steps. For non-stationary environments the above rule does not stand. Specifically in drifting environments $|p_{t+1}(a) - p_t(a)| < \epsilon$, where $\epsilon$ is a small constant, while in abruptly changing environments $\exists t : |p_{t+1}(a) - p_t(a)| > \delta$, where $\delta$ is a sufficiently large value in terms of probabilities. Both $\epsilon$ and $\delta$ can be used as a measure of the environment dynamics. Denoting $a^*$ the optimal arm to choose, the regret at every time step is then defined as $\mathbb{E}[r_t(a^*) - r_t(a)]$.

A first important note is that in most bandit methods, from either machine learning or computational neuroscience, the action $a_t$ to be taken at the time episode $t$ can be chosen after comparing between all actions a combination of an empirical measure of each action's value $\mu_t(a)$ and some measure of uncertainty for this value $\sigma_t(a)$, according to the following general equation:

$$a_t = \mathcal{F}(\beta_t(\mu_t(a) + \phi\sigma_t(a))) \tag{1}$$

In the classical family of bandit methods initially developed in machine learning - called *UCB* for Upper Confidence Bound [5] -, function $\mathcal{F}$ is an argmax, parameters $\beta$ and $\phi$ are fixed to 1, and $\mu_t(a)$, $\sigma_t(a)$ are empirically estimating the average reward yielded by action $a$ and the uncertainty associated to this estimate. Then the method differ in the way these estimates are updated, using for instance a *Sliding Window* (SW-UCB) or *Discounted* estimates (D-UCB) [1]. In a more recent work, the *Kalman Filter Multi-Armed Normal Bandit* (KF-MANB) proposed in [2] is also another empirically optimal algorithm using an observational noise parameter $\sigma_{ob}$ and a transitional noise parameter $\sigma_{tr}$ to estimate $\mu_t(a)$ and $\sigma_t(a)$ at each timestep, which are then used as the parameters of a normal distribution $\mathcal{N}(\mu_t(a), \sigma_t(a))$ for each arm. Then a sample $s_a$ from each distribution is taken, and the action choices are made according the rule $a_t = \arg\max_a (s_a)$.

In computational neuroscience models for bandit tasks [3, 6], function $\mathcal{F}$ is a Boltzmann softmax followed by sampling from the resulting action probability distribution, parameters $\beta$ and $\phi$ are either fixed or estimated from each human subject's behavior during the task, $\mu_t(a)$ is the *Q-value* of action $a$ learned through reinforcement learning - most often with a discounted factor $\gamma = 0$ because the task is not sequential - and $\sigma_t(a)$ is an exploration bonus. In some models, $\mu_t(a)$ can be transiently adapted faster through the use of change point detectors which identify abrupt task shifts that produce a higher noise than the expected observational noise [4]. Interestingly, this has also been recently applied to bandit methods in Machine Learning (*e.g.* the *Adapt-Eve* algorithm [7]). Finally, some models propose to dynamically tune through *meta-learning* the algorithm's parameters as a function of variations of performance [8, 9]: for instance, increasing the inverse temperature $\beta_t$ in the softmax function to promote exploitation when the average performance improves, and decreasing it to promote exploration when performance drops indicate abrupt changes in the task.

bioRxiv preprint doi: https://doi.org/10.1101/117598; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In previous work, we have applied this meta-learning principle in an algorithm here referred to as *MLB* to dynamically tune $\beta_t$ in a simple multi-armed bandit scenario involving the interaction between a simulated human and a robot [10]. Formally, function $\mathcal{F}$ is a Boltzmann softmax with parameter $\phi = 0$ (*i.e.* no uncertainty term is used as exploration bonus here). The immediate reward $r_t$ received from action $a = a_t$ is then used to update the estimated action's expected value $\mu_t(a_t)$ of the action taken, using the update rule of a low-pass filter, as $Q_{t+1}(a) = (1 - \alpha_Q)Q_t(a) + \alpha_Q r_t$. Furthermore, following [8], a similar update rule applies to estimate mid-term and long-term reward running averages according to:

$$\overline{r}_t = \alpha_m r_t + (1 - \alpha_m)\overline{r}_{t-1} \text{ and } \overline{\overline{r}}_t = \alpha_\ell \overline{r}_t + (1 - \alpha_\ell)\overline{\overline{r}}_{t-1} \tag{2}$$

where $\alpha_m$ and $\alpha_\ell$ are two different learning rates. $\beta_t$ is then updated as

$$\beta_{t+1} = max\{\beta_t + \eta(\overline{r}_t - \overline{\overline{r}}_t)), 0\} \tag{3}$$

With the above update rules, $\beta_t$ increases when the mid-term reward is greater than the long-term. One can view this as an evidence that the recent actions are more optimal, therefore exploitation may be increased. In the opposite case when $\overline{r}_t < \overline{\overline{r}}_t$, the recent actions denote that the present policy has lower performance in getting rewards compared to the past, therefore exploration should be encouraged by reducing $\beta_t$.

## 3 Hybrid Meta-learning with Kalman Filters - MLB-KF

After showing a common way to formalize bandit methods in machine learning and computational neuroscience, the objective of the present paper is to illustrate how a hybrid method combining ideas from the two fields that we have recently proposed in [11] can yield a promising performance in a variety of non-stationary multi-armed bandit tasks. The hybrid model integrates the sibling kalman filters of KF-MANB and the core of MLB. To do so, action values learned through reinforcement learning in *MLB* are here simply replaced within the softmax function with a linear combination (with $\phi > 0$) of the mean $\mu_t(a)$ and the standard deviation $\sigma_t(a)$ of each arm's distribution according to the sibling kalman filters of KF-MANB. The algorithm also uses two constants $\sigma^2_{ob}$ and $\sigma^2_{tr}$, which relate to the stochastic part and the non-stationary part of the environment respectively. The values for means $\mu_1(a)$, variances $\sigma^2_1(a)$, and $\beta_t$ are initialized. Following [2], but writing equations down on a different way, the reward $r_t(a)$, for $a = a_t$, is then used to update the parameters of the respective arm with

$$\mu_{t+1}(a) = v_t(a)r_t(a) + (1 - v_t(a))\mu_t(a) \text{ and } \sigma^2_{t+1}(a) = v_t(a)\sigma^2_{ob} \tag{4}$$

while for arms $a \neq a_t$ the means $\mu_{t+1}(a)$ maintain their previous value, and a transitional variance is added to $\sigma^2_{t+1}(a)$ in order to enhance exploration of non-chosen arms as shown below

$$\mu_{t+1}(a) = \mu_t(a) \text{ and } \sigma^2_{t+1}(a) = \sigma^2_t(a) + \sigma^2_{tr} \tag{5}$$

In the above equations $v_t(a)$ are dynamically tuned learning rates for each arm, which may mirror some adaptation procedures of learning rate studied in a computational neuroscience bandit tasks (*e.g.* [12]). Here, for KF-MANB, it follows the update rule:

$$v_t(a) = (\sigma_t(a)^2 + \sigma^2_{tr})/(\sigma_t(a)^2 + \sigma^2_{tr} + \sigma^2_{ob}) \tag{6}$$

The mid- and long-term rewards are also calculated like in [8] and used to update the inverse temperature parameter $\beta_t$ according to the update rule of MLB [10]. When the uncertainty of an arm's action value increases, the respective arm should be explored more. However when the environment suggests that no evidence of change is observed, KF-MANB suffers from exploration since a transitional variance is always added to the non-pulled arms. This can also be seen in Fig.1 (right) of simulations section. MLB is also expected to display large regret at cases where a non-optimal arm becomes optimal after long periods of stationarity, since the "rising" arm may never be chosen. The above implementation is a trade-off between these two cases, as we empirically present evidence that it incorporates both the benefits of MLB's fast adaptation and small regret on many test cases, as also the robustness and good performance of KF-MANB.

## 4 Numerical Simulations

Testing an algorithm on a scenario that was also used in order to tune its parameters, should restrict its evaluation of performance to concern only this specific environment. Here we avoided this bias by testing the algorithms on many different scenarios, altering the most crucial components of a stochastic and non-stationary multi armed bandit, while keeping the parameters that were tuned on a simple mid-case. For the mid-case we considered 5 arms, with initial arm probabilities of 0.8, 0.4, 0.3, 0.2, 0.1 for each arm respectively, with a circular shift of those at every 2000 time steps, in a total of 10000. After optimization, the parameters chosen for D-UCB were $\xi = 0.22$ and $\gamma = 0.9999$, for SW-UCB $\xi = 0.3$ and $\tau = 998$, for Adapt-EvE $\delta = 0.13$, $\lambda = 40$, $t_m = 50$ and used UCBT as decision-makers and as meta-bandit. For KF-MANB $\sigma_{ob} = 0.2$, $\sigma_{tr} = 0.01$, initializing all means and variances to 0.5. For MLB $\alpha_Q = 0.14$, $\alpha_m = 1/15$, $\alpha_\ell = 1/350$, $\eta = 0.44$. For MLB-KF we kept the parameters found for MLB and KF-MANB and only tuned $\phi = 1.5$.

bioRxiv preprint doi: https://doi.org/10.1101/117598; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Algorithm 1** Hybrid MLB-KF

Choose parameters $\alpha_m$, $\alpha_\ell$, $\eta$, $\phi$, $\sigma_{ob}$, $\sigma_{tr}$
Initialize $\beta_1$, and $\mu_1(a)$, $\sigma_1(a)$ $\forall a \in \mathcal{K}$
**for** $t = 1, 2, ..., T$ **do**
  select action $a_t \in \mathcal{K}$ from distribution of (1) with softmax
  observe the immediate reward $r_t$ and update $\overline{r}_t$, $\overline{\overline{r}}_t$ with (2)
  for $a = a_t$ update $\mu_{t+1}(a)$, $\sigma^2_{t+1}(a)$ with (4),(6)
  for all arms $a \neq a_t$ update $\mu_{t+1}(a)$, $\sigma^2_{t+1}(a)$ with (5)
  update $\beta_{t+1}$ with (3)
**end for**

For problem set I, we considered that a global change point $cp$, may occur at every time step $t$ with a constant probability $h$ (i.e $p_t(cp) = h$). When a change point occurs, all arm probabilities are then re-sampled from a uniform distribution in [0, 1]. For each subproblem (using only the set of probabilities generated for each arm with the above type of random walk), we run each algorithm for 20 sub-sessions, then regenerated another problem set with the same probability $h$, and repeated the procedure for 100 hyper-sessions. We then calculated the average total cumulative regret with the assumption that the results are representative due to central limit theorem. We increased $h$ and repeated all the above, for $h \in [2/10000, 8/10000]$ with a step of $1/10000$. Therefore, an average of 2 to 8 global change points occur (we have an horizon of 10000) for every step. As seen in Fig 1 (left), MLB exhibited the best performance for small $h$, KF-MANB demonstrated both robustness and low regret in all cases and MLB-KF inherited the best of these performances by following the regret dynamics of the most optimal algorithm between the two. In problem set II we explored cases where the difference $\Delta_g$ between the expected reward of the optimal action and the second best action varied. The best optimal arm always had a probability of $0.8$, the second best $0.8-\Delta_g$, the third $0.8-2\Delta_g$ and so forth. For our experiments we begun with $\#cp = 1$ (where $\#cp$ denotes the number of change points). This change point occurred at $t = T/2$, when the probability of the best arm dropped from 0.8 to 0.8-$\Delta_g$, the probability of the second best dropped from 0.8-$\Delta_g$ to 0.8-$2\Delta_g$ (and so forth for other intermediate arms), while the probability of the worst arm increased from 0.8-$4\Delta_g$ to 0.8. We altered $\Delta_g$ from 0.02 to 0.2 with a step size of 0.02 for each independent test case, while we simulated all algorithms on each case for 200 sessions, observing the average final cumulative regret. We then increased $\#cp$ by one, fixed the change point time steps to be evenly distributed over the timesteps, and repeated the procedure. In Fig.1 (middle) a visualization of the performance for each algorithm is presented, for different gaps $\Delta_g$ (rows), and $\#cp$ (columns). The results once again provided evidence that MLB-KF combines the behavior of both MLB and KF-MANB in a satisfactory manner, denoting the best performance on average. Finally, cases where the best optimal action becomes the less optimal, cases where the less optimal action becomes the optimal as also sequential alteration of those were tested in problem set III. MLB adapted greatly in Best-to-Worst scenarios and D-UCB in Worst-to-Best. However the overall average performance of MLB-KF was the empirically optimal as described in [11] (not shown here).

In order to highlight how parameters $v_t(a)$ change in time, a simple two-armed bandit case using KF-MANB, where $p_t(a_1)$ was changing from 0.75 to 0.25 every 50 timesteps and $p_t(a_2) = 1 - p_t(a_1)$, is also shown in Fig. 1 (right) for an average of 5 sessions. For the interval between timesteps 50-100, meta-parameter $v_t(a_1)$ increases since the most chosen arm was $a_2$. Even though the estimated reward rate of arm $a_1$ in this interval differs from the true rate, a more precise estimation is not needed in order to make the most optimal decision. However, evidence about the arm's reward rate in this interval is not provided, therefore $v_t(a_1)$ increases, and then decreases as the arm is once again frequently chosen.

## 5 Conclusion and Discussion

In this work we presented empirical evidence that a hybrid algorithm which makes use of a classical computational neuroscience approach to properly tune exploration with meta-learning, combined with sibling Kalman filters proposed in Machine Learning to estimate each arm's action value and add a measure of uncertainty as an exploration bonus, can lead to an adaptive learning strategy which efficiently manages the exploration-exploitation trade-off dilemma. However we still need to do the comparison more systematically for many different tasks (different numbers of arms, different stochasticity levels, different amplitudes between best and second-best arm, abrupt task shifts versus progressive drifts). Nevertheless, such a bridge between the fields of computational neuroscience and machine learning seems promising.

Besides, each development of a new algorithm displaying efficient adaptation performance is an opportunity to sketch some new experimental predictions for neuroscience. Here, an interesting aspect of the hybrid algorithm is that it combines dynamic tuning of a global exploration rate $\beta_t$ through meta-learning and adaptive tuning of action specific learning rates $v_t(a)$ based on uncertainty variations in a Kalman filter. The resulting two timeseries may be less colinear than if they had been generated by the same method, which can be useful to generate regressors for model-based analyses of neurocience data. In addition, the resulting learning rates $v_t(a)$ may follow a different dynamics - briefly illustrated in Fig. 1 (right) - than other computational methods of learning rate studied in neuroscience bandit tasks, such as adapta-

bioRxiv preprint doi: https://doi.org/10.1101/117598; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tions as a function of reward volatility [12]. Again, this may constitute an interesting experimental prediction to explore and compare in the future with human behavioral and neural data.

Finally, it is interesting to note recent progresses in machine learning using meta-learning combined with deep reinforcement learning [13]. Their deep learning approach is promising to not only dynamically tune some parameters but learn at a more global level the dynamics/statistics of the task. In addition, the method produces generalization to different but statistically related bandit tasks. All these results also suggest that a common ground between meta-learning and reinforcement learning can help find general methods to not only solve single-step bandit tasks, but also generalize to multi-step reinforcement leaning problems.

## References


[1] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," *Algorithmic Learning Theory*, pp. 174–188, 2011.

[2] O.-C. Granmo and S. Berg, *Solving Non-Stationary Bandit Problems by Random Sampling from Sibling Kalman Filters*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 199–208.

[3] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, 2006.

[4] R. C. Wilson, M. R. Nassar, and J. I. Gold, "Bayesian online learning of the hazard rate in change-point problems," *Neural computation*, vol. 22, no. 9, pp. 2452–2476, 2010.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[6] M. J. Frank, B. B. Doll, J. Oas-Terpstra, and F. Moreno, "Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation," *Nature neuroscience*, vol. 12, no. 8, pp. 1062–1068, 2009.

[7] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, "Multi-armed bandit, dynamic environments and meta-bandits," in *NIPS 2006 Workshop on Online trading between exploration and exploitation*, Whistler, Canada, 2006.

[8] N. Schweighofer and K. Doya, "Meta-learning in reinforcement learning." *Neural Networks*, vol. 16, no. 1, pp. 5–9, 2003.

[9] M. Khamassi, P. Enel, P. Dominey, and E. Procyk, "Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters," *Progress in Brain Research*, vol. 202, pp. 441–464, 2013.

[10] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas, "Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task," in *IEEE Robotic Computing 2017*, Taipei, Taiwan, 2017.

[11] G. Velentzas, C. Tzafestas, and M. Khamassi, "Bio-inspired meta-learning for active exploration during non-stationary multi-armed bandit tasks," in *IEEE Intelligent Systems Conference 2017*, London, UK, 2017.

[12] T. E. Behrens, M. W. Woolrich, M. E. Walton, and M. F. Rushworth, "Learning the value of information in an uncertain world," *Nature neuroscience*, vol. 10, no. 9, pp. 1214–1221, 2007.

[13] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick, "Learning to reinforcement learn," *arXiv preprint arXiv:1611.05763*, 2016.


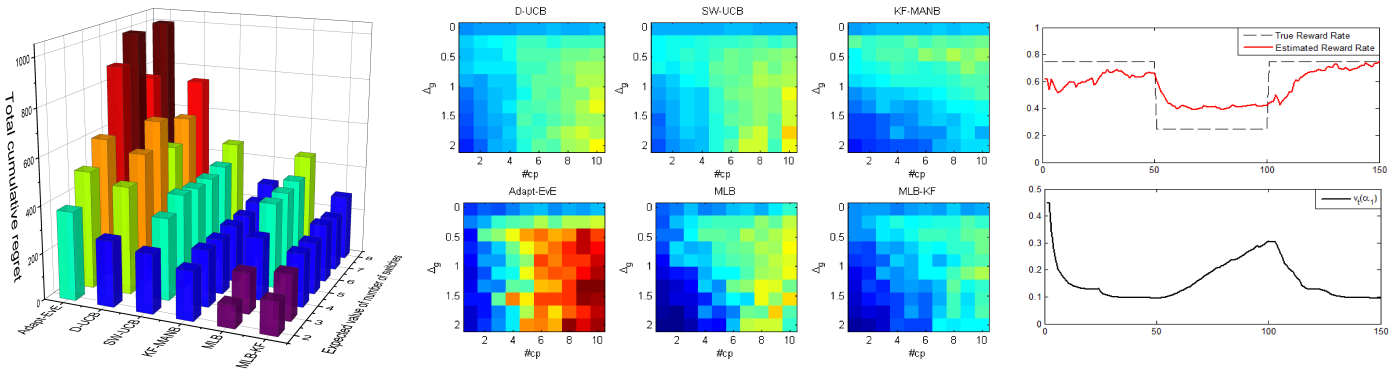


Figure 1: left: total regret for different values of expected number of change points $h$, middle: total cumulative regret for different gaps $\Delta_g$ and number of switches $\#cp$, right: estimated action value $\mu(a_1)$ vs true value (top) and parameter $v_t(a_1)$ of KF-MANB for a two-armed bandit case (bottom)