On Balancing Exploration vs. Exploitation in a Cognitive Engine for Multi-Antenna Systems

Haris I. Volos and R. Michael Buehrer
Mobile and Portable Radio Research Group (MPRG), Wireless@Virginia Tech
Bradley Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Email: {hvolos, buehrer}@vt.edu

***Abstract*—In this paper, we define the problem of balancing exploration vs. exploitation in a cognitive engine controlled multi-antenna communication system in terms of the classical multi-armed bandit framework. We then employ the ϵ-greedy strategy and Gittins' indices methods for addressing the problem in a system with no prior information. Results show that the Gittins' indices assuming a normal reward process had the best overall performance compared to the Gittins' indices with a Bernoulli reward process and the ϵ-greedy strategy. The latter was found to be more consistent albeit inefficient for most of the cases except in the case of both a low number of trials and a low SNR in which it was found to have better performance than the other methods. Nevertheless, the Gittins' indices method should be generally preferred as it is more consistent than the ϵ-greedy strategy across different scenarios.[1].**

I. Introduction

A radio has traditionally used a fixed set of communication methods selected by its operator. Today, however, a radio is expected not only to use a large number of the plethora of current communication methods, but also to be able to select the method that best meets its goal under the current operating environment.

Typically, the radio designer will analyze each communication method in terms of the desired goal under assumed channel models. Then, the designer will combine the analysis in order to arrive at a set of adaptation rules that best meets the goal for the radio. Performing the required analysis is usually a lengthy task requiring a significant amount of effort, especially when a large number of different methods are involved. Furthermore, if the channel models do not hold or the communication methods perform in a way not expected, the design becomes moot. The same applies if the desired goal changes.

The question then arises: what if a radio could be designed in such a way that it could determine on its own which communication method to use to meet its goals? The scope of our work [1], [2] is to propose methods that can make such a design possible. This work is based on the pioneering work of Mitola who first described the "cognitive radio". While Mitola's ideal Cognitive Radio (CR) is not only capable of optimizing its own wireless capabilities, but also of self-determining its goals by observing its human operator's behavior [3], we are only interested in optimizing the wireless capabilities of the radio. Research such as [4], [5] borrowed ideas from the fields of machine learning and Artificial Intelligence (AI) in an effort to make that possible. Although current research made steps in the right direction, the main channel metrics considered were the Signal-to-Noise Ratio (SNR) and basic channel statistics which are not adequate for multiple-antenna systems. The latter systems are also affected by other factors such as spatial correlation. Also, [5] addressed a range of issues concerning a CR, where the Physical (PHY) layer issues received partial attention. Our overall work focuses on multi-antenna PHY layer aspects. Furthermore, we look closely at the two main tasks needed to enable the radio to optimize its own performance: learning its own capabilities and optimizing to meet its goals. We want the radio to use the most applicable and efficient methods for learning and optimizing. Therefore, our aim is to design a Cognitive Engine (CE), the software package that makes the desired behavior possible. *In this paper we focus on the CE's performance while it is learning.*

In [2] we have proposed a CE design that employs learning and optimization techniques in order to learn the capabilities of the radio and optimize for different goals. Our CE design differs from existing designs [5]–[8] in that the tasks of learning and optimizing are separate. Our CE is not only learning what is the best method, given the goal and the channel conditions, but it also learns the abilities of the radio independent of the goal. Should the goal change, the known abilities of the radio can be used to speed up the optimization process and minimize the need for new learning. Our method can be seen as analogous to the *Actor-Critic* [9] methods used in reinforcement learning, where the actor (optimization unit) suggests possible solutions, and the critic (learning unit) responds with the desirability of the proposed solution based on its experience. In our previous work [2], we have assumed that the learning unit had enough examples to learn what is needed to provide the optimization unit with the necessary information. Furthermore, we established that there may be a negative impact when not enough samples are available. The extent of the negative effect is dependent upon the learning and optimization techniques used. However, performance during learning (collecting examples) was not evaluated. Even if the CE is assumed to go through prolonged learning sessions, it

[1] This work was supported by the National Science Foundation under Grant No. 0520418.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

is practically impossible to expose it to all possible channel conditions *a priori*. Consequently, it is reasonable to expect that the CE sooner or later will face unknown conditions. For example, if the radio is operating in a critical mission may not have the luxury of time to learn what is best before operating; it has to establish a connection and learn at the same time. Optimally balancing exploration vs. exploitation ensures that the negative effects of learning will be kept to a minimum. Therefore, we need to evaluate the performance of the radio controlled by the CE during learning.

In this paper we seek to evaluate the CE's performance during learning and to introduce synergies between learning and optimization (i.e., match learning with the optimization goal(s)). To put it in more classical terms, we address the problem of balancing the exploration vs. exploitation. *Exploration* refers to trying options with unknown but potential beneficial outcome. On the other hand, *exploitation* refers to using what is already known to have the highest performance metric. Methods focusing on exploitation instead of exploration are usually known as *greedy* or *myopic* methods. Exploration involves unknown risk, while exploitation tends to be safer in terms of expectations. Depending on the situation, exploitation might ultimately limit the long-term performance. In a few words the exploration vs. exploitation problem is: do we choose an option that guarantees short-term performance (exploitation) or do we choose an option that could either hurt short-term performance but improve the long term performance of the system? How do we balance those two conflicting objectives? This is a universally occurring problem and we borrow the results from other fields such as re-reinforcement learning [9] and dynamic programming [10], [11] and apply them to the context of selecting the best communication technique in a multi-antenna communication system. [12] applied similar concepts on finding unused channels under the context of Dynamic Spectrum Sharing (DSS).

The goal of this paper is to investigate and apply exploration vs. exploitation balancing techniques namely the simple, yet effective, ϵ-greedy strategy and the more complex, but optimal Gittins' [13] dynamic allocation index method.

Section II presents some background information and formulates the problem. Section III explains our test setup. Section IV presents and discusses the test results, and Section V provides some concluding remarks.

II. Background and Problem Formulation

The problem of exploration vs. exploitation is classically studied by using the mathematical framework of the Multi-Armed Bandit (MAB) problem. We introduce the MAB problem and two possible approaches, namely the ϵ-greedy strategy and the Gittins' indices. Finally, we define how the distribution parameters are estimated in a recursive way.

A. Our Communication Problem

We have a communication system with K options. Each option uses a combination of modulation, coding, and MIMO techniques. The system is controlled by a CE that can learn and optimize performance subject to the collection of data samples under the different channel conditions. In this paper we limit the scope of our results in the case of achieving maximum spectral efficiency (capacity) and assume that the channel statistics do not change during the operating interval. We plan on expanding to more goals and varying channel statistics in our next publication.

B. The Multi-Armed Bandit Problem

The MAB problem gets its name from the slot machines (bandits) found in casinos. A typical slot machine has a single arm that when pulled returns a reward with a certain probability. In the multi-armed bandit problem it is assumed that the player is faced with either multiple machines or a single machine with multiple arms and his goal is to get the maximum reward by using the machines. Generally it is assumed that the player has little or no information about the bandits and he has to decide between exploring for the most rewarding machines and using the machine that was found to yield the higher reward. Essentially this is an information acquisition problem and the player is always faced with the same options.

Adapting the description found in [10], let $\mathcal{Y}$ be the set of K slot machines (comm. options), and let W_y be a random variable that gives the amount of the reward returned (capacity) if we use the machine y. Also let μ_y be the unknown true mean of W_y and σ_y^2 be the variance. Finally, let $(\bar{\mu}_y^n, \hat{\sigma}_y^2)$ be the estimate of the mean and the variance of W_y after n iterations and s a belief state about the random variable W_y. The estimates $(\bar{\mu}_y^n, \hat{\sigma}_y^2)$ can be an example of a belief state.

Let $x_y^n = 1$ if the y^{th} machine is played at iteration n and W_y^n the reward returned on that round. Also let

$$N_y^n = \sum_{y=1}^{n} x_y^n \tag{1}$$

be the total number of times the y^{th} machine was used.

In the MAB problem we are looking for a policy that maximizes the expected return $V(s)$:

$$V(s) = E_s \sum_{n=1}^{N} \gamma^n W^n \tag{2}$$

where N is the maximum number of plays (often assumed to be ∞), E_y is the expectation operator over the belief state s, γ is a discount factor $0 < \gamma < 1$, and W^n the return at time n. The discount factor is used to ensure a finite return when $N \rightarrow \infty$. Another interpretation is to treat $1 - \gamma$ as the probability that the process is going to stop [14]. Therefore, the discount factor is a way to express our expectation on the duration of the optimization horizon. A low value discount factor discounts future returns with a higher rate. As result, when balancing exploration vs. exploitation the latter has a higher weight. On the other hand, a high valued discount factor ($\gamma \rightarrow 1$) will make future rewards more important and exploration will have a higher weight than the previous case.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

Finding the policy that maximizes (2) is a K-dimensional problem. Two key methods of addressing this problem are the ϵ-greedy strategy and the use of the Gittin's indices.

C. The ϵ-greedy Strategy

The ϵ-greedy strategy [9] is a simple strategy that is exploiting by using the best method ($y_{gready} = \arg\max_y \bar{\mu}_y^n$) 1-$\epsilon$ ($\epsilon \in [0, 1]$) of the time (greedy). However, with probability ϵ it explores by using a random y uniformly selected. As $n \to \infty$ by the law of large numbers $\bar{\mu}_y^n$ is going to converge to the true mean. The ϵ-greedy methods guarantee that all the options are explored as the horizon tends to infinity. The ϵ parameter controls how fast exploration is performed. A higher ϵ will cause a faster exploration and arrive more quickly at an optimal or near-optimal option. However, the high exploration rate may cause reduced overall returns because of the higher exploration cost.

The $\epsilon-$greedy strategy has two main variations: the $\epsilon-$first strategy and the $\epsilon-$decreasing strategy. The former explores for ϵN trials and exploits for the remaining $(1-\epsilon)N$ trials, where N is the number of total trials. The latter variation decreases the exploration rate by decreasing ϵ as the number of trials is increasing.

In this paper we consider only the classic version of the $\epsilon-$greedy strategy. However, we have some prior information about the communication system that should be used: we know the potential return of each option (capacity) and we also know an upper bound of the capacity that can be achieved under the current channel. Therefore, we restrict the exploration to machines that potentially can outperform the current y_{gready}. We also restrict exploration on methods with a capacity $\leq C_{\max}$, the maximum Shannon estimated capacity, $C_{\max}$, for the current channel conditions as given by [15]:

$$C_{\max} = \sum_{i=1}^{\min\{N_t, M_r\}} \log_2 \left(1 + \frac{\text{SNR}}{N_t}\lambda_i\right) \tag{3}$$

where N_t and M_r is the number of the transmit and receive antennas respectively. λ_i is ith eigenvalue of HH^T, where H is the $M_r \times N_t$ channel matrix.

D. The Gittins Index

Gittins in [13] showed we can solve the K-dimensional problem of (2) by using a dynamic allocation index method that breaks the problem in a series of K one-dimensional problems. The Gittins' index ν_y at a belief state s is given by:

$$\nu_y(s) = \sup_{n \leq N} \frac{E_s \sum_{n=1}^{N} \gamma^n W_y^n}{E_\pi \sum_{n=1}^{N} \gamma^n} \tag{4}$$

which can be interpreted as the maximum expected reward per unit of expected discount time. The result, albeit simple and straightforward, has been proven in a non-trivial way to be optimal. The interested reader can find more information in the paper [13] and book [11] authored by Gittins.

The optimal policy is simply to use the option y with the highest ν_y. The Gittins' index is dependent upon the underlying distribution of W_y. In this work we consider the Gittins' index for the Normal Reward Process (NRP) and the Bernoulli Reward Process (BRP). It may be noted that in the application examined in this work the underlying process is Bernoulli - either a certain capacity is achieved or not. In our application if a transmitted packet is successfully received, then we assume a return equal to the capacity of the communication option used, otherwise the return is zero. Therefore, assuming a BRP is in theory more suitable than NRP. However, as it will be shown in the results, assuming a NRP has some performance advantages over assuming a BRP.

For a NRP the Gittins' index is equal to:

$$\nu(\bar{\mu}, \bar{\sigma}^2, n, \gamma) = \bar{\mu} + \bar{\sigma}\nu(0, 1, n, \gamma) \tag{5}$$

where $\nu(0, 1, n, \gamma)$ is the Gittins' index for a zero mean, unit variance distributed process.

For a BRP the Gittins' index is equal to [16]:

$$\nu(\alpha, \beta, \gamma, R_y) = R_y \nu(\alpha, \beta, \gamma, 1) \tag{6}$$

where $\nu(\alpha, \beta, \gamma, 1)$ is the Gittins' index for a Bernoulli process with α successes, β failures, with a reward of 1, if successful. R_y is the reward received when option y is successful. A challenge of the BRP is that the Gittins' index is not defined when either α or β is zero. This is challenging because there are cases that the probability of success of the BRP is practically either one or zero. For example, beamforming using QPSK with $1/8$ convolutional code in medium to high SNR levels and spatial multiplexing using 256 QAM in medium to low SNR levels, respectively. For this reason we make the practical assumption that when either α or β are zero we calculate $\nu(\alpha{+}1, \beta{+}1, \gamma, R_y)$ instead. This assumption allows us to get an estimate of the index that otherwise we would not be able to achieve in a reasonable amount of trials.

A downside of the Gittins' indices is that is not trivial to estimate them. On the other hand, for most practical purposes the indices tabulated in [11] are sufficient. Finally, like the ϵ-greedy strategy, we limited the choice of y by using (3).

E. Mean and Variance Estimation

For the NRP, we need to have an estimate of the mean and the variance of each of W_y. For their estimation, we adopt the method described in [10]. Subsequently, the mean can be recursively estimated using:

$$\bar{\mu}_y^n = \begin{cases} \frac{N_y^n - 1}{N_y^n}\bar{\mu}_y^{n-1} + \frac{1}{N_y^n}W_y^n & \text{If } x_i^n = 1 \\ \bar{\mu}_y^{n-1} & \text{Otherwise} \end{cases} \tag{7}$$

The variance can be similarly estimated used:

$$\hat{\sigma}_y^{2,n} = \begin{cases} \frac{N_y^n - 2}{N_y^n - 1}\hat{\sigma}_y^{2,n-1} + \frac{1}{N_y^n}\left(W_y^n - \bar{\mu}_y^{n-1}\right)^2 & \text{If } x_y^n = 1 \\ & \text{and } N_y^n \geq 2, \\ \hat{\sigma}_y^{2,n-1} & \text{If } x_y^n = 0 \end{cases} \tag{8}$$

N_y^n can be updated using:

$$N_y^n = N_y^{n-1} + x_y^n \tag{9}$$

The variance of $\bar{\mu}_y^n$ is given by:

$$\bar{\sigma}_y^{2,n} = \frac{1}{N_y^n}\hat{\sigma}_y^{2,n} \tag{10}$$

If N_y^n is 0 or 1, then $\hat{\sigma}_y^{2,n} = \infty$. That means we don't know anything about the distribution. However, assumptions can be made. In this work we initialize the index to the maximum potential return (capacity) of each option until $N_y^n > 1$.

In large problems it is hard to estimate the variance, therefore, we can use a single population variance for the initial steps:

$$\hat{\sigma}_y^{2,n} = \frac{n-2}{n-1}\hat{\sigma}_y^{2,n-1} + \frac{1}{n}\sum_{y\in\mathcal{Y}} x_y^n \left(W_y^n - \bar{\mu}_y^{n-1}\right)^2 \tag{11}$$

which is updated after every trial. The variance of $\bar{\mu}_y^n$ is given by:

$$\bar{\sigma}_y^{2,n} = \frac{1}{N_y^n}\hat{\sigma}^{2,n} \tag{12}$$

III. Test Setup

We evaluated the proposed methods by implementing the system using the MATLAB simulation software package.

A. Configuration

Three MIMO Techniques were considered: Beamforming, Transmit Diversity using a STBC, and Spatial Multiplexing using V-BLAST. The MIMO system was assumed to have four transmit and four receive antennas (4×4). The CE had the choice of the following modulation schemes: QPSK, 8-PSK, 16, 32, 64, 128, and 256 QAM. Furthermore, it could vary the coding rate of a convolutional codec with a constraint length $K = 8$ [17]. The available coding rates were: 1, 7/8, 3/4, 2/3, 1/2, 1/4, 1/6, and 1/8. Not all modulation/coding combinations were allowed. The modulation/coding combinations where chosen such that the combined distance metric [17] and spectral efficiency monotonically decreased and increased respectively. The final combinations had 22 different spectral efficiencies from 0.25 bits/symbol (QPSK, $1/8$ codec) to 8 bits/symbol (256-QAM, uncoded).

B. Methods Tested

We tested the Gittins' index for both NRP and BRP for γ equal to[2] 0.5, 0.7, and 0.99. The $\epsilon-$greedy strategy was evaluated with ϵ equal to 0.01, 0.1. Values around 0.1 are commonly used.

The methods were tested in an SNR range between 5 and 50 dB at 5 dB intervals at a maximum pairwise antenna correlation $\rho = 0.1$. We also tested the case of $\rho = 0.5$. In the latter case we only considered $\gamma = 0.7$ and $\epsilon = 0.1$. A high correlation, ρ, between the antenna elements negatively affects the use of spatial multiplexing. The reader is reminded that spatial multiplexing exploits the availability of multiple spatial modes which are reduced to one as $\rho \rightarrow 1$.

It may be noted that the different channel metrics such as SNR and ρ represent different sets of working options. Higher SNR levels have more options available and a low ρ value allows the use of more spatial multiplexing options. The availability of working options will affect the total return as options that do not work will adversely affect the return (if they are explored).

TABLE I
Average Total Return Over Optimal Total Return

	Number of Trials Performed					
	50	50	50	500	500	500
	Discount Factor, γ					
Method	.5	.7	.99	.5	.7	.99
Gittins' Index, NRP	.73	.73, .70[1]	.72	.93	.93, .93[1]	.94
Gittins' Index, BRP	.60	.59, .65[1]	.56	.89	.89, .87[1]	.85
	Exploration Parameter, ϵ					
Method	.01	.1	.2	.01	.1	.2
ϵ-greedy strategy	.53	.65, .50[1]	.70	.74	.87, .87[1]	.87

[1] Max. pairwise antenna correlation, ρ, equal to .5, .1 otherwise

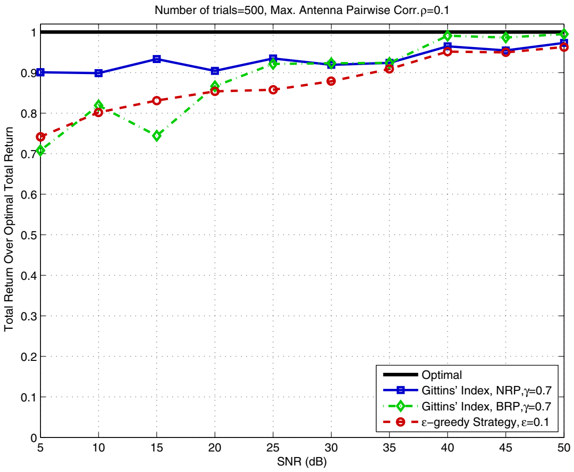

Fig. 1. Total Return vs. SNR, 500 Trials

C. Evaluation Metrics

In the results to follow we will compare the total return and the average instantaneous return to the optimal respective return. The total return is the sum of all returns for a number of trials. The average instantaneous return is the average return experienced after a specific number of trials. The optimal return for each case is the best possible return for the underlying channel conditions. The optimal returns were estimated by employing a brute force search over all available options and using 400 trials *per option*. As a comparison, each method (Gittins' & ϵ-greedy) was evaluated after running up to 500 trials *in total*.

IV. Results

Table I presents the average total return over the optimal total return, averaged over all the SNR levels, for all the methods and parameters tested. There are two sets, the first estimated at 50 trials and the second estimated at 500 trials. The results for 50 trials represent performance with a relatively

[2] Gittins' [11] provides tables for γ equal to 0.5, 0.6, 0.7, 0.8, 0.9 and 0.99 for both NRP and BRP

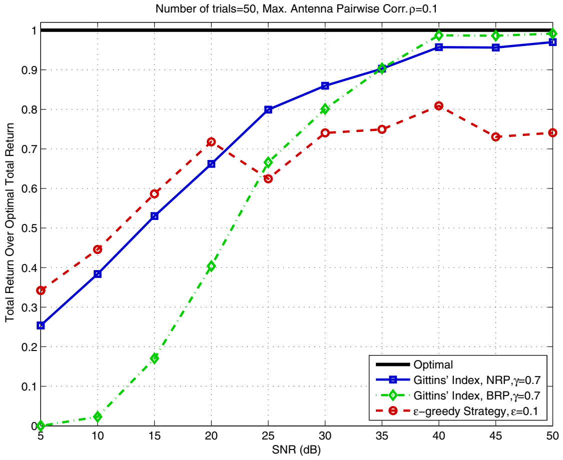

Fig. 2. Total Return vs. SNR, 50 Trials

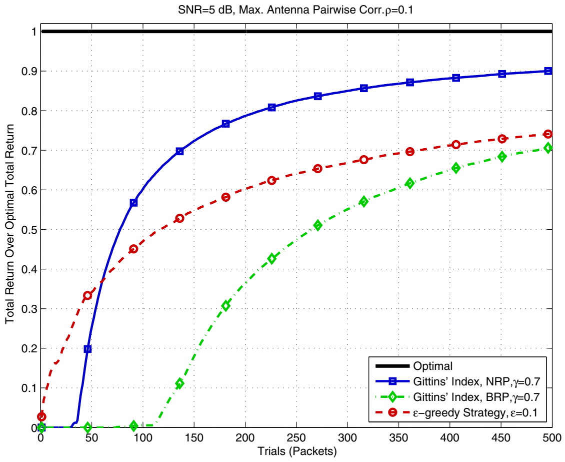

Fig. 3. Total Return vs. Trials, SNR = 5 dB

short time frame and the results for 500 trials for a longer time frame. At 50 trials, both Gittin's index methods have a slightly reduced performance when a higher valued discount factor is used. This is because there is a higher focus on exploration. The same can be observed for the BRP at 500 trials - the NRP has practically no fluctuations. Also, we can observe that after 500 trials the returns are higher compared to 50 trials. Results show that the NRP has the highest overall returns. The reason why NRP performs better vs. the BRP is explained in the following paragraphs, when the results of Figure 1 are discussed. Still commenting on the results of Table I, the $\epsilon-$greedy strategy showed reduced performance with lower values of ϵ. In this case the reduced exploration hurt the returns in both cases.

Most of the results were obtained assuming $\rho = 0.1$. In addition, there is a small set for $\rho = 0.5$ (Section III-B). In the latter case, the results show that the Gittins' NRP is still superior and performance at 50 trials is slightly reduced for NRP and the $\epsilon-$greedy strategy, and slightly improved for the BRP. With 500 trials, the results were nearly the same.

Figure 1 plots the total achieved reward over the optimal reward value by running the Gittins' index methods for $\gamma = 0.7$ and the $\epsilon-$greedy strategy for $\epsilon = 0.1$. The results of Figure 1 show: first, that all the methods have degraded performance at low SNRs. This can be attributed to the fact that many options simply do not work and as a result, the exploration cost is higher. Even though the options used are limited by (3), not all non-performing options are eliminated because the limit given by (3) is not tight since the options used are suboptimal compared to the limit.

Second, the Gittins' index using a BRP was found to perform poorly at low to medium SNR levels. This is possibly due to the fact that this reward process is not defined for the cases where either α or β is zero. Therefore, in order to be able to estimate the index, when either α or β is zero instead of calculating $\nu_y(\alpha, \beta, \gamma, R_y)$, we calculated $\nu_y(\alpha+1, \beta+1, \gamma, R_y)$ instead. This caused the index to be higher for a low number of trials and causing the exploration of those options until enough samples were collected. This disadvantage is more pronounced at the low SNR levels where some of the methods have a extremely low probability of working and requiring an extremely large amount of trials to get a non zero α. The same applies to β for the higher SNRs; however, it is unlikely that the latter is causing any performance degradation. On the other hand, in the NRP case, if no successes are found, the $\bar{\mu}$ in (5) is going to be zero, and the other main contribution factor will be the variance which at the initial stages it is assumed to be the same (11) for all y. Therefore, in this case the current estimate of $\bar{\mu}$ will carry most of the decision weight until enough samples are collected for the individual variance calculation and for $\nu(0, 1, n, \gamma)$ in (5) to be a more decisive factor.

Third, the Gittins' index BRP has a reduced performance at a SNR=15 dB. This is because, instead of the method focusing/settling on the optimal choice of beamforming with 128 QAM uncoded with an expected return of $7 \times 0.98 = 6.86$ bps/Hz, where 7 is the spectral efficiency and 0.98 the probability of a successful packet, it settles on using VBLAST with uncoded QPSK with an expected return of $8 \times 0.7 = 5.6$ bps/Hz. The reason that the latter is preferred is because it has a higher potential return ($8 > 7$ bps/Hz) which makes it more desirable for exploration.

And fourth, the observations in the low SNR region of Figure 1 can be also seen in Figure 3 which shows the progression of the total reward in terms of the optimal reward. Results show that the ϵ-greedy strategy has the best performance for the first 50 trials, followed by the Gittins' index with NRP and BRP. The latter had the worst performance for the reasons explained above. As the number of trials increases, the Gittins' index with NRP outperforms the other two methods.

Figure 2 is the same type as Figure 1 but with 50 trials instead of 500. In this case the ϵ-greedy strategy seems to have the best performance up to an SNR=20dB. Potentially, one could use the ϵ-greedy strategy when the goal is to perform best in a time horizon of less than 50 trials.

A look at the instantaneous return is provided by Figures 4 & 5 for an SNR equal to 25 and 50 dB respectively. It may be observed from both figures that after 150 trials all the

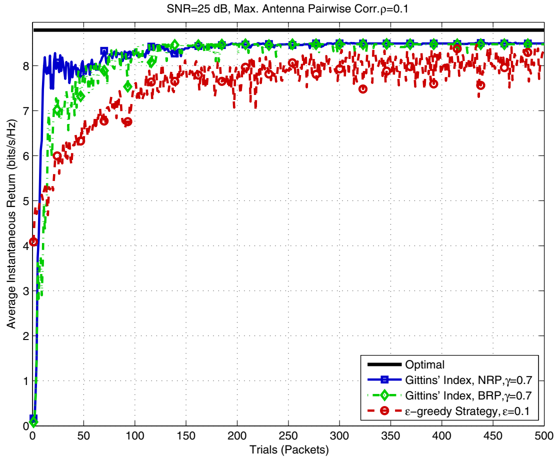

Fig. 4. Average Instantaneous Return Vs Trials, SNR = 25 dB

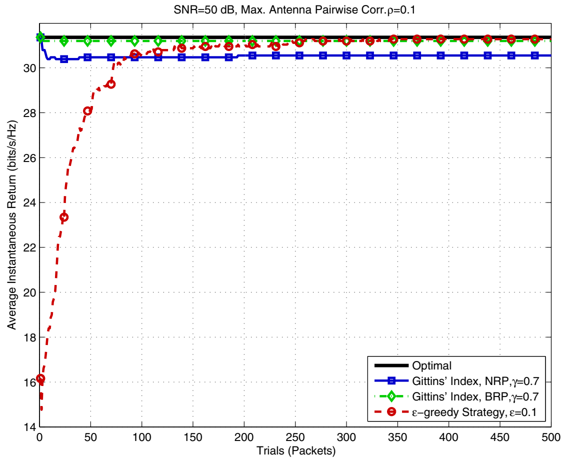

Fig. 5. Average Instantaneous Return vs. Trials, SNR = 50 dB

methods achieve a return very close to the return achieved after 500 trials. In addition, at a medium SNR (Figure 4) the two Gittins' index methods perform the same after 100 trials. On the contrary, at a high SNR level (Figure 5) the Gittins' index with a BRP outperforms the Gittins' index with a NRP. The ϵ-greedy strategy also outperforms the latter after 100 trials and it also settles on the optimal method. The reason behind this is that the Gittins' index with a NRP is using VBLAST with both 128 QAM and 256 QAM (both uncoded) with a return of 28 and 32 bps/Hz respectively. As a result, the average instantaneous reward is 30.55 vs. the optimal of 31.36 bps/Hz

Finally, it may be noted that the results obtained by all the methods tested are always subject to the parameters used and are also dependent on the number of the available options and the underlying conditions. For example, the $\epsilon-$greedy strategy is known to suffer when the number of options is very large [10]. On the other hand, we have seen that the Gittins' indices with the NRP, tends to perform better most of the time verifying the *long-term* optimality of the Gittins' indices.

V. Conclusions

In this paper we defined the exploration vs. exploitation problem in the context of a Cognitive Engine trying to learn (exploring) while providing optimal performance (exploiting) at the same time. This problem can be classically addressed in the terms of the multi-armed bandit problem which can be solved suboptimally by the ϵ-greedy method and optimally by the use of the Gittins' dynamic allocation indices. Even though our test scenario was a Bernoulli reward process, it was found that using the Gittins' index of a normal reward process yielded better results, especially in the low SNR regions. In addition, the ϵ-greedy method was found to work well when the number of trials is small (short-term performance) and SNR is poor. Nevertheless, the Gittins' indices method should be generally preferred as it is more consistent than the ϵ-greedy strategy across different scenarios.

References

[1] H. I. Volos, C. I. Phelps, and R. M. Buehrer, "Initial Design of a Cognitive Engine for MIMO Systems," in *SDR Forum Technical Conference*, Nov 2007.

[2] ——, "Physical Layer Cognitive Engine For Multi-Antenna Systems," in *IEEE Military Communications Conference*, Nov. 2008.

[3] J. Mitola, III, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Ph.D. dissertation, KTH, Stockholm , Sweden, May 2000.

[4] C. Clancy, J. Hecker, E. Stuntebeck, and T. O'Shea, "Applications of Machine Learning to Cognitive Radio Networks," *IEEE Wireless Communications*, vol. 14, no. 4, pp. 47–52, August 2007.

[5] T. W. Rondeau, "Application of Artificial Intelligence to Wireless Communications," Ph.D. dissertation, Virginia Tech, 2007.

[6] C. J. Rieser, "Biologically Inspired Cognitive Radio Engine Model Utilizing Distributed Genetic Algorithms for Secure and Robust Wireless Communications and Networking," Ph.D. dissertation, Virginia Tech, 2004.

[7] T. R. Newman, B. A. Barker, A. M. Wyglinski, A. Agah, J. B. Evans, and G. J. Minden, " Cognitive engine implementation for wireless multicarrier transceivers," *Wiley Journal on Wireless Communications and Mobile Computing*, vol. 7, no. 9, pp. 1129–1142, 2007.

[8] Z. Zhao, S. Xu, S. Zheng, and J. Shang, " Cognitive radio adaptation using particle swarm optimization," *Wiley Journal on Wireless Communications and Mobile Computing*, 2008, published online.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, March 1998.

[10] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2007.

[11] J. C. Gittins, *Multi-armed Bandit Allocation Indices*. Wiley, Chichester, NY, 1989.

[12] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation and competition," *IEEE/ACM Trans. on Networking*, Oct. 2007, submitted.

[13] J. Gittins and D. Jones, "A Dynamic Allocation Index for the Sequential Design of Experiments," *Progress in Statistics*, pp. 241–266, 1974.

[14] I. M. Sonin, "A generalized Gittins index for a Markov chain and its recursive calculation," *Statistics & Probability Letters*, vol. 78, no. 12, pp. 1526–1533, September 2008.

[15] D. Gesbert, M. Shafi, D. shan Shiu, P. Smith, and A. Naguib, "From theory to practice: an overview of MIMO space-time coded wireless systems," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 3, pp. 281–302, Apr 2003.

[16] D. Acuña and P. Schrater, "Bayesian Modeling of Human Sequential Decision-Making on the Multi-Armed Bandit Problem," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, V. Sloutsky, B. Love, and K. McRae, Eds. Washington, DC: Cognitive Science Society, 2008.

[17] J. Proakis, *Digital Communications*, 4th ed. New York: McGraw-Hill, 2001.