Journal of Machine Learning Research 1 (2015) xx-xx Submitted xx/xx; Published xx/xx

Explore no more: Simple and tight high-probability bounds for non-stochastic bandits

Editor:

Abstract

This work addresses the problem of regret minimization in non-stochastic multi-armed bandit problems, focusing on performance guarantees that hold with high probability. A widely accepted common belief is that achieving such strong guarantees requires the learner to explicitly devote several rounds for exploring actions—even if many of the actions are obviously suboptimal. In this paper, we show that it is possible to prove high-probability regret bounds without this undesirable exploration component. Our result relies on a simple and intuitive loss-estimation strategy called *Implicit eXploration* (IX) that allows a very clean analysis that leads to the best known constant factors in the bounds. We also demonstrate the robustness of IX in a simple experiment that shows a significant improvement over previous methods.

Keywords: multi-armed bandits, exploration-exploitation, high-probability bounds, implicit exploration

1. Introduction

Consider the problem of regret minimization in non-stochastic multi-armed bandits, as considered in the classic paper of Auer, Cesa-Bianchi, Freund, and Schapire (2002). This problem can be formalized as a repeated game between a *learner* and an *environment* (sometimes called the *adversary*). In each round $t = 1, 2, \dots, T$, the two players interact as follows: The learner picks an *arm* (also called an *action*) $I_t \in [K] = \{1, 2, \dots, K\}$ and the environment selects a loss function $\ell_t : [K] \to [0, 1]$, where the loss associated with arm $i \in [K]$ is denoted as $\ell_{t,i}$. Then, the learner suffers and observes the loss ℓ_{t,I_t}. Based on solely on these observations, the goal of the learner is to pick its actions so that the *regret* grows as slowly as possible:

$$R_T = \sum_{t=1}^{T} \ell_{t,I_t} - \min_{i \in [K]} \sum_{t=1}^{T} \ell_{t,i} \to \min .$$

We say that the environment is *oblivious* if it selects the sequence of loss vectors irrespective of the past actions taken by the learner, and *adaptive* (or *non-oblivious*) if ℓ_t is chosen based on the past actions $I_{t-1}, \dots, I_1$. An equivalent formulation of the multi-armed bandit game uses the concept of *rewards* (also called *gains* or *payoffs*) instead of losses: in this version, the adversary chooses the sequence of *reward functions* (r_t) with $r_{t,i}$ denoting the reward given to the learner for choosing action i in round t. We will refer to the above two formulations as the *loss game* and the *reward game*, respectively.

Most of the existing literature on non-stochastic bandits is concerned with bounding the *pseudo-regret* (or *weak regret*) defined as

$$\widehat{R}_T = \sum_{t=1}^{T} \ell_{t,I_t} - \min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right],$$

where the expectation integrates over the randomness injected by the learner. Proving bounds on the actual regret that hold with high probability is considered to be a significantly harder task that can be achieved by serious changes made to the learning algorithms and much more complicated analyses. One particular common belief is that in order to guarantee high-confidence performance guarantees, the learner cannot avoid repeatedly sampling arms from a uniform distribution, typically $\Omega(\sqrt{KT})$ times. It is easy to see that such *explicit exploration* can impact the practical performance of the learning algorithm in a very negative way if there are many arms with high losses: even if the base learning algorithm learns quickly to focus on good arms, explicit exploration still forces the regret to grow at a steady rate. As a result, algorithms with high-probability performance guarantees tend to perform poorly even in very simple problems.

This creates quite an uncomfortable situation: while theory suggests that these high-confidence algorithms should be more reliable than standard algorithms with guarantees holding in expectation, practice shows the exact opposite. For example, Seldin et al. (2012) conduct some experiments on a simple 2-armed bandit problem where Exp3 beats Exp3.P by a spectacular margin.

In this paper, we dispel the myth that explicit exploration is necessary for obtaining high-probability bounds. One component that we preserve of the Exp3.P-recipe is the *biased estimation of losses*, although our bias is of a much more delicate nature, and arguably more elegant than previous approaches. In particular, we adopt the *implicit exploration* (IX) strategy first proposed by Kocák, Neu, Valko, and Munos (2014) and also used by Neu (2015) for other purposes. Regarding the Exp3 algorithm endowed with IX-style loss estimates, we prove a high-probability regret bound comparable to those of Exp3.P and the PolyINF algorithm of Audibert and Bubeck (2009, 2010)—without the need to mix in the uniform distribution for exploration purposes.

Notably, our analysis directly works in the loss game which is again a departure from the mainstream. Indeed, while many of the most important results concerning non-stochastic bandits were first proven for the reward game, more recent advances revealed that studying the loss game usually leads to much cleaner theoretical analyses and tighter performance guarantees (see, e.g., Bubeck and Cesa-Bianchi, 2012). However, nearly all lecture notes, monographs and surveys (including Bubeck and Cesa-Bianchi, 2012) still present the analysis of Exp3.P for the reward game to avoid technical complications, even when all other chapters in the respective works considers loss games.

2. Explicit and implicit exploration for stabilizing loss estimation

One key element in all principled algorithms for non-stochastic bandits is constructing importance-weighted loss/reward estimates of the form

$$\widehat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbb{1}_{\{I_t=i\}} \qquad \text{and} \qquad \widehat{r}_{t,i} = \frac{r_{t,i}}{p_{t,i}} \mathbb{1}_{\{I_t=i\}}$$

where $p_{t,i} = \mathbb{P}[I_t = i | \mathcal{F}_{t-1}]$ is the probability that the learner picks action i in round t, conditioned on the observation history $\mathcal{F}_{t-1}$ of the learner up to the beginning of round t. These estimates are unbiased for all i with $p_{t,i} > 0$ in the sense that $\mathbb{E}\left[\widehat{\ell}_{t,i}\right] = \ell_{t,i}$ for all such i. While the above loss estimates are sufficient for proving bounds on the regret that hold in expectation (as done by, e.g., Bubeck and Cesa-Bianchi, 2012), proving bounds that hold with high probability requires a little more work. In particular, the Exp3.P algorithm of Auer et al. (2002) uses the *biased* reward-estimates

$$\widetilde{r}_{t,i} = \widehat{r}_{t,i} + \frac{\beta}{p_{t,i}} \tag{1}$$

for an appropriately chosen $\beta > 0$. An alternative for these estimates was proposed by Audibert and Bubeck (2010), who use

$$\widetilde{r}_{t,i} = -\frac{1}{\beta}\log\left(1 - \beta\widehat{r}_{t,i}\right). \tag{2}$$

The above loss/reward estimates are then usually used as inputs for a black-box online learning algorithm to produce the distributions $(\boldsymbol{p}_t)$. In particular, at the beginning of round $t > 1$, Exp3.P computes the weights

$$w_{t,i} = \exp\left(\eta \sum_{s=1}^{t-1} \widetilde{r}_{s,i}\right)$$

for all i and some positive learning-rate parameter $\eta > 0$ and then samples I_t according to the distribution

$$p_{t,i} = (1-\gamma)\frac{w_{t,i}}{\sum_{j=1}^{K} w_{t,j}} + \frac{\gamma}{K},$$

where $\gamma > 0$ is the exploration parameter. The argument for this explicit exploration is that it keeps the variance of the above loss estimates under control, thus enabling the use of (more or less) standard concentration results[1].

In particular, the simplest known analysis of Exp3.P due to Bartlett et al. (2008) (see also Beygelzimer et al., 2011) relies on an application of Freedman's inequality to show that the inequality

$$\sum_{t=1}^{T}(r_{t,i} - \widehat{r}_{t,i}) \le \frac{\gamma}{K}\sum_{t=1}^{T}\frac{\ell_{t,i}^2}{p_{t,i}} + (e-2)\frac{K\log(K/\delta)}{\gamma}$$

holds simultaneously for all i with probability at least $1-\delta$. This in particular implies that

$$\sum_{t=1}^{T}(r_{t,i} - \widetilde{r}_{t,i}) \le (e-2)\frac{K\log(K/\delta)}{\gamma}$$

holds for the loss estimates defined as in Equation (1), when $\beta \ge \gamma/K$. A similar property can be shown to hold for the estimates of Equation (2). In other words, this shows that the estimates $\widetilde{r}_{t,i}$ are (in a sense) upper-confidence bounds for the true rewards $r_{t,i}$, which is an essential result for proving the high-probability performance guarantees of Exp3.P.

1. Explicit exploration is believed to be inevitable for proving bounds in the reward game for various other reasons, too—see Bubeck and Cesa-Bianchi (2012) for a discussion.

Implicit exploration. In the current paper, we propose to use the loss estimates defined as

$$\widetilde{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma \ell_{t,i}} \mathbb{1}_{\{I_t = i\}}, \tag{3}$$

for all i and an appropriately chosen $\gamma > 0$. Loss estimates of a similar form were first used by Kocák et al. (2014)—following them, we refer to this technique as Implicit eXploration (IX). A seemingly minor difference between IX as defined by Kocák et al. (2014) and the estimates defined above is that we replace their denominator $p_{t,i} + \gamma$ by $p_{t,i} + \gamma \ell_{t,i}$.

Notice that IX as defined above achieves a similar variance-reducing effect as the one achieved by explicit exploration. Furthermore, similarly to the reward-estimates defined in Equations (1) and (2), the IX estimates are also *biased* estimators of the respective losses. A careful investigation reveals further interesting connections between the estimates (1) and (2). In particular, first observe that

$$\widetilde{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma \ell_{t,i}} \mathbb{1}_{\{I_t = i\}} = \frac{\ell_{t,i} \left(\mathbb{1}_{\{I_t = i\}} + \gamma \ell_{t,i} \right)}{p_{t,i} + \gamma \ell_{t,i}} - \gamma \frac{\ell_{t,i}^2}{p_{t,i} + \gamma \ell_{t,i}}.$$

The two terms on the right-hand side of the above equality correspond to an *unbiased* estimate of $\ell_{t,i}$ and a bias term corresponding to γ times the variance of the first term. This allows a direct application of Freedman's inequality to prove

$$\sum_{t=1}^{T} \left(\widetilde{\ell}_{t,i} - \ell_{t,i} \right) \leq (e-2) \frac{\log(K/\delta)}{\gamma},$$

similarly to the bound concerning the reward-estimates (1). Note however that the bias introduced by IX can be much smaller than the bias terms used in (1) when the losses tend to be small.

An even closer look reveals interesting connections to the estimates (2), too. To reveal these connections, observe that

$$\widehat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma \ell_{t,i}} \mathbb{1}_{\{I_t = i\}} = \frac{1}{2\gamma} \cdot \frac{2\gamma \ell_{t,i} / p_{t,i}}{1 + \gamma \ell_{t,i} / p_{t,i}} \mathbb{1}_{\{I_t = i\}} \leq \frac{1}{2\gamma} \cdot \log \left(1 + 2\gamma \widehat{\ell}_{t,i} \right), \tag{4}$$

where the last step follows from the elementary inequality $\frac{z}{1+z/2} \leq \log(1+z)$ that holds for all $z \geq 0$. Notice that the above expression bears a striking similarity to the reward-estimate (2).

This similarity enables us to build on the elegant technique of Audibert and Bubeck (2010) and prove the following lemma that provides an improved constant factor over the technique relying on Freedman's inequality.

Lemma 1 *With probability at least* $1 - \delta$*,*

$$\sum_{t=1}^{T} \left(\widetilde{\ell}_{t,i} - \ell_{t,i} \right) \leq \frac{\log(K/\delta)}{2\gamma}.$$

simultaneously holds for all $i \in [K]$*.*

The proof of the lemma is presented in Section 4. Notably, the proof is much more elementary than that of Bartlett et al. (2008) and Beygelzimer et al. (2011), since it does not rely on any advanced results from probability theory.

3. A simple high-probability bound

To demonstrate the power of implicit exploration, we prove an improved performance guarantees for the variant of Exp3 (Auer et al., 2002) that uses the IX loss estimates (3). This algorithm chooses action i with probability proportional to

$$p_{t,i} \propto w_{t,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{s,i}\right),$$

without mixing any explicit exploration term into the distribution. We prove the following performance guarantee concerning this algorithm, called Exp3-IX. Notably, this theorem provides the best known constant factor of $2\sqrt{2}$ in the leading term, improving on the best known factor of 5.15 due to Bubeck and Cesa-Bianchi (2012).

Theorem 2 *With probability at least $1-\delta$, the regret of* Exp3-IX *satisfies*

$$R_T \leq \frac{\log K}{\eta} + \frac{\log\left(K/\delta\right)}{2\gamma} + \left(\frac{\eta}{2} + \gamma\right) KT + \left(\frac{\eta}{2} + \gamma\right) \frac{K \log\left(K/\delta\right)}{2\gamma}.$$

In particular, setting $\eta = 2\gamma = \sqrt{\frac{2\log K}{KT}}$, the above bound becomes

$$R_T \leq 2\sqrt{2KT \log K} + \sqrt{\frac{KT}{\log K}} \log\left(1/\delta\right)$$

Proof Following the now-standard analysis of Exp3 in the loss game (Bubeck and Cesa-Bianchi, 2012), we can obtain the bound

$$\sum_{t=1}^{T} \left(\sum_{i=1}^{K} p_{t,i} \widehat{\ell}_{t,i} - \widehat{\ell}_{t,j} \right) \leq \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_{t,i} \widehat{\ell}_{t,i}^2$$

for any fixed j. Now observe that

$$\begin{aligned}
\sum_{i=1}^{K} p_{t,i} \widehat{\ell}_{t,i} &= \sum_{i=1}^{K} p_{t,i} \frac{\ell_{t,i}}{p_{t,i} + \gamma \ell_{t,i}} \mathbb{1}_{\{I_t = i\}} \\
&= \sum_{i=1}^{K} \mathbb{1}_{\{I_t = i\}} \frac{\ell_{t,i}\left(p_{t,i} + \gamma \ell_{t,i}\right)}{p_{t,i} + \gamma \ell_{t,i}} - \gamma \sum_{i=1}^{K} \mathbb{1}_{\{I_t = i\}} \frac{\ell_{t,i}^2}{p_{t,i} + \gamma \ell_{t,i}} \\
&= \ell_{t,I_t} - \gamma \sum_{i=1}^{K} \ell_{t,i} \cdot \widehat{\ell}_{t,i}.
\end{aligned}$$

Similarly,

$$\sum_{i=1}^{K} p_{t,i} \widehat{\ell}_{t,i}^2 \leq \sum_{i=1}^{K} \ell_{t,i} \cdot \widehat{\ell}_{t,i}.$$

Combining these results with Lemma 1, the bound can be rewritten as

$$\begin{aligned}\sum_{t=1}^{T}\left(\ell_{t,I_t}-\ell_{t,j}\right) \leq & \frac{\log K}{\eta}+\frac{\log\left(K/\delta\right)}{2\gamma}+\left(\frac{\eta}{2}+\gamma\right)\sum_{t=1}^{T}\sum_{i=1}^{K}\ell_{t,i}\cdot\widehat{\ell}_{t,i}\\ \leq & \frac{\log K}{\eta}+\frac{\log\left(K/\delta\right)}{2\gamma}+\left(\frac{\eta}{2}+\gamma\right)\sum_{i=1}^{K}\widehat{L}_{t,i}\\ \leq & \frac{\log K}{\eta}+\frac{\log\left(K/\delta\right)}{2\gamma}+\left(\frac{\eta}{2}+\gamma\right)KT+\left(\frac{\eta}{2}+\gamma\right)\frac{\log\left(K/\delta\right)}{2\gamma}.\end{aligned}$$

The proof is concluded by taking $j=\arg\min_i \sum_{t=1}^{T}\ell_{t,i}$. ■

4. The proof of Lemma 1

Fix any t. For convenience of notation, we will use the notation $\beta=2\gamma$. Using Equation (4), we get that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\beta\widetilde{\ell}_{t,i}\right)\middle|\,\mathcal{F}_{t-1}\right] =& \mathbb{E}\left[\exp\left(\log\left(1+\beta\widehat{\ell}_{t,i}\right)\right)\middle|\,\mathcal{F}_{t-1}\right]\\ \leq & 1+\beta\ell_{t,i}\leq\exp\left(\beta\ell_{t,i}\right),\end{aligned}$$

where we used $\mathbb{E}\left[\widehat{\ell}_{t,i}\middle|\,\mathcal{F}_{t-1}\right]\leq\ell_{t,i}$ that holds by definition of $\widehat{\ell}_{t,i}$, and $1+z\leq e^z$ that holds for all $z\in\mathbb{R}$. Thus, the process $W_t=\exp\Big(\beta\big(\widetilde{L}_{t,i}-L_{t,i}\big)\Big)$ is a supermartingale with respect to $(\mathcal{F}_t)$: $\mathbb{E}\left[W_t|\,\mathcal{F}_{t-1}\right]\leq W_{t-1}$. Observe that, since $W_0=1$, this implies $\mathbb{E}\left[W_t\right]\leq\mathbb{E}\left[W_{t-1}\right]\leq\ldots\leq 1$, and thus by Markov's inequality,

$$\begin{aligned}\mathbb{P}\left[\widetilde{L}_{T,i}>L_{T,i}+\varepsilon\right] &= \mathbb{P}\left[\widetilde{L}_{T,i}-L_{T,i}>\varepsilon\right]\\ &\leq \mathbb{E}\left[\exp\left(\beta\left(\widetilde{L}_{T,i}-L_{T,i}\right)\right)\right]\exp(-\beta\varepsilon)\leq\exp(-\beta\varepsilon)\end{aligned}$$

holds for any $\varepsilon>0$. The statement of the lemma follows from solving for ε and using the union bound.

5. A simple experiment

We conduct a simple experiment to demonstrate the robustness of Exp3-IX as compared to Exp3 and its superior performance as compared to Exp3.P. Our setting is a 10-arm bandit problem where all losses are independent draws of Bernoulli random variables. The mean losses of arms 1 through 8 are $1/2$ and the mean loss of arm 9 is $1/2-\Delta$ for all rounds $t=1,2,\ldots,T$. The mean losses of arm 10 are changing over time: for rounds $t=1,2,\ldots,T/2$, the mean is $1/2+\Delta$, and $1/2-4\Delta$ afterwards. This choice ensures that up to at least round $T/2$, arm 9 is clearly better than other arms, arm 10 starts to outperform it in the second half of the interval, and eventually becomes the leader.

We have evaluated the performance of Exp3, Exp3.P and Exp3-IX in the above learning problem with $T=100,000$ and $\Delta=0.05$. The parameters of Exp3 and Exp3.P were

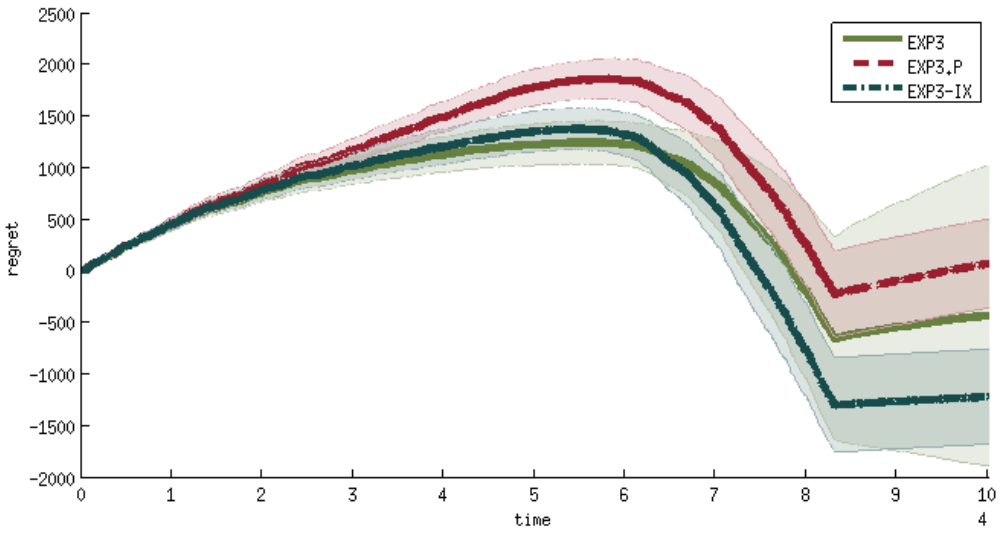

Figure 1: Regret of Exp3, Exp3.P, and Exp3-IX, respectively in the problem described in Section 5.

set as suggested by Bubeck and Cesa-Bianchi (2012), Exp3-IX was tuned according to Theorem 2. Figure 1 shows the empirical means and standard deviations of the regrets of the three algorithms over 100 runs. The results clearly indicate the robust behavior of Exp3-IX. While Exp3 outperforms both algorithms in the first half of the interval (i.e., as long as the losses are i.i.d.), its behavior quickly starts to degrade as the mean losses start to change. As suggested by theory, the performance fluctuations of Exp3.P are much better than those of Exp3, although its mean regret is significantly worse than that of Exp3 in both regimes. We have run the experiment with several settings of Δ and T and experienced essentially the same type of behavior from all three algorithms.

6. Conclusion

We have presented a simple and efficient alternative to explicit exploration for proving high-probability regret bounds in non-stochastic multi-armed bandit problems. Our preliminary experiment suggest that Exp3-IX no longer suffers from poor empirical performance as Exp3.P does, and is also able to outperform Exp3 in certain scenarios. Generalizing the result to more advanced settings such as bandits with expert advice (Auer et al., 2002; McMahan and Streeter, 2009; Beygelzimer et al., 2011), combinatorial semi-bandits (Audibert et al., 2014; Neu and Bartók, 2013) or bandits with side-observations (Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014) is straightforward. An interesting open question left for future research is finding out whether the idea of implicit exploration can help in advancing the state of the art in the more general problem of regret minimization in linear bandits.

References

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From Bandits to Experts: A Tale of Domination and Independence. In *Neural Information Processing Systems*, 2013.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

J. Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39:31–45, 2014.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. ISSN 0097-5397.

Peter L. Bartlett, Varsha Dani, Thomas P. Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *COLT 2008*, pages 335–342, 2008.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS 2011*, 2011.

Sébastien Bubeck and Nicolò Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Now Publishers Inc, 2012.

Tomás Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *NIPS-27*, pages 613–621, 2014.

S Mannor and O Shamir. From Bandits to Experts: On the Value of Side-Observations. In *Neural Information Processing Systems*, 2011.

H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *COLT 2009*, 2009.

G Neu. First-order regret bounds for combinatorial semi-bandits. *Accepted at the 27th Conference on Learning Theory (COLT)*, 2015.

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *ALT 2013*, pages 234–248, 2013.

Yevgeny Seldin, Nicolo Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. Pac-bayes-bernstein inequality for martingales and its application to multiarmed bandits. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, 2012.