Combining Expert Advice in Reactive Environments

Daniela Pucci de Farias PUCCI@MIT.EDU
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Nimrod Megiddo MEGIDDO@ALMADEN.IBM.COM
IBM Almaden Research Center
San Jose, CA 95120, USA

Abstract

"Experts algorithms" constitute a methodology for choosing actions repeatedly, when the rewards depend both on the choice of action and on the unknown current state of the environment. An experts algorithm has access to a set of strategies ("experts"), each of which may recommend which action to choose. The algorithm learns how to combine the recommendations of individual experts so that, in the long run, for any fixed sequence of states of the environment, it does as well as the best expert would have done relative to the same sequence. This methodology may not be suitable for situations where the evolution of states of the environment depends on past chosen actions, as is usually the case, for example, in a repeated non-zero-sum game.

A general exploration-exploitation experts method is presented along with a proper definition of value. The new method is quite different from previously proposed experts algorithms. It represents a shift from the paradigms of regret minimization and myopic optimization to consideration of the long-term effect of a player's actions on the environment. The importance of this shift is demonstrated by the fact that this algorithm is capable of inducing cooperation in the repeated Prisoner's Dilemma game, whereas previous experts algorithms converge to the suboptimal non-cooperative play. The method is shown to asymptotically perform as well as the best available expert. Several variants are analyzed from the viewpoint of the exploration-exploitation tradeoff, including explore-then-exploit, polynomially vanishing exploration, constant-frequency exploration, and constant-size exploration phases. Complexity and performance bounds are proven.

Keywords: Sequential decision making, experts algorithms, reactive environments, exploration-exploitation tradeoffs, complexity and performance bounds.

1. Introduction

Real-world environments require agents to choose actions sequentially. For example, a driver has to choose everyday a route from one point to another, based on past experience and perhaps some current information. In another example, an airline company has to set prices dynamically, also based on past experience and current information. One important difference between these two examples is that the effect of the driver's decision on the future traffic patterns is negligible, whereas prices set by one airline can affect future market prices significantly. In this sense the decisions of the airlines are made in a reactive environment, whereas the driver performs in a non-reactive one. For this reason, the driver's problem is essentially a problem of prediction while the airline's problem has an additional element of control.

In the decision problems we consider, an agent has to repeatedly choose currently feasible actions. The agent then observes a reward, which depends both on the chosen action and the current state of the environment. The state of the environment may depend both on the agent's past choices and on choices made by the environment independent of the agent's current choice. There are various known approaches to sequential decision making under uncertainty. In this paper we focus on the so-called experts algorithm approach. An "expert" (or "oracle") is simply a particular strategy recommending actions based on the past history of the process. An experts algorithm is a method that combines the recommendations of several given "experts" (or "oracles") into another strategy of choosing actions (e.g., Littlestone and Warmuth 1994; Auer et al. 2000; Freund and Schapire 1999). It directs the agent with regard to which expert to follow in the next stage, based on the past history of actions and rewards.

A popular criterion in the design and analysis of experts algorithms is called minimum regret (MR). Regret is defined as the difference between the reward that could have been achieved, given the observed sequence of states of the environment, and what was actually achieved. An expert selection rule is said to minimize regret if it yields an average reward as large as that of any single expert, against any fixed sequence of states of the environment. Indeed, certain experts algorithms, which at each stage choose an expert from a probability distribution that is related to the reward accumulated by the expert prior to that stage, have been shown to minimize regret (Auer et al., 2000; Freund and Schapire, 1999).

It is crucial to note that, since the experts are compared on a sequence-by-sequence basis, the MR criterion ignores the possibility that different experts may induce different sequences of states in the environment. Thus, MR makes sense only under the assumption that the state of the environment evolves independently from the agent's choices. As pointed out in the airline pricing example, this assumption is not satisfied in reactive environments. Ignoring the potential impact of the agent's actions on the environment may lead to substantial loss of performance. This is illustrated with an example involving the Prisoner's Dilemma game.

The Prisoner's Dilemma. In the single-stage Prisoner's Dilemma (PD) game, each player can either cooperate (C) or defect (D). Defecting is better than cooperating regardless of what the opponent does, but it is better for both players if both cooperate than if both defect. Consider the repeated PD. Suppose the row player consults with a set of experts, including the "defecting expert," who recommends defection all the time. Let the strategy of the column player in the repeated game be fixed. In particular, the column player may be very patient and cooperative, willing to wait for the row player to become cooperative, but eventually becoming non-cooperative if the row player does not seem to cooperate. Since defection is a dominant strategy in the stage game, the defecting expert achieves in each step a reward as high as any other expert against any sequence of choices of the column player, so the row player learns with the experts algorithm to defect all the time. This seems to minimize regret, since for any fixed sequence of actions by the column player, constant defection is the best response. However, constant defection is not the best response in the repeated game against many possible strategies of the column player. For instance, the row

player would regret very much using the experts algorithm if he were told later that the column player had been playing a strategy such as Tit-for-Tat.[1]

In this paper, we propose and analyze a new experts method, denoted *Exploration-Exploitation Experts Method* (EEE), which is especially designed for learning in reactive environments. EEE follows experts judiciously, attempting to maximize the long-term average reward. It differs from previous approaches in at least two ways. First, each time an expert is selected, it is followed for multiple time stages rather than a single one. Second, EEE takes into account only the rewards that were actually achieved by an expert in the stages it was followed, rather than the reward that could have been obtained in any stage. EEE enjoys the same appealing simplicity of the previous experts algorithms, yet it leads to a qualitatively different behavior and improved average reward. The effectiveness of EEE is demonstrated by its performance in the repeated PD game, namely, it is capable of identifying the opponent's willingness to cooperate and it induces cooperative behavior.

We provide results about the convergence of several variants of EEE. We develop performance guarantees showing that the method achieves average reward comparable to that achieved by the best expert. We characterize convergence rates that hold both in expected value and with high probability. Many learning algorithms can be interpreted as "exploration-exploitation" methods. Roughly speaking, such algorithms blend choices of exploration, aimed at acquiring knowledge, and exploitation that capitalizes on accumulated knowledge to maximize rewards. In particular, some experts algorithms can be interpreted as alternating between testing all experts and following the ones that achieved best performance in the past. An important aspect of our results is that they provide an explicit characterization of the tradeoff between exploration and exploitation.

Another contribution of this paper is the introduction of a definition for the long-term value of an expert. Appropriate notions of value are required to guide the design and analysis of experts algorithms in reactive environments. In particular, they must capture the reactions of the environment to the expert's actions, as well as the fact that any learning algorithm commits mistakes. We propose a notion of value that satisfies these properties and characterize how fast EEE learns the value of each expert.

The paper is organized as follows. The method is described in section 2. We analyze a variant of the method where experts are followed for "phases" comprising an increasing number of time stages in sections 3 through 6. Convergence rates based on actual expert performance are presented in section 3. In section 4, we present a notion of the long-run value of an expert. This definition gives rise to question of how fast EEE learns the experts' values, which is answered in section 5. In section 6, we analyze and compare several exploration schemes. Finally, in section 7 we analyze a different variant of the method where experts are followed for "phases" with a constant number of time stages.

2. The Exploration-Exploitation Method

The problem we consider in this paper can be described as follows. At each time stage $s = 1, 2, \ldots$, an agent has to choose actions $a_s \in \mathcal{A}$. At the same time the environment

1. The Tit-for-Tat strategy is to play C in the first stage, and later play in every stage whatever the opponent played in the preceding stage.

also "chooses" a state $b_s \in \mathcal{B}$, and the agent receives a reward $R(a_s, b_s)$. The choices of the environment may depend on various factors, including the past choices of the agent.

Let $\mathcal{H}_s$ denote the set of all histories up to stage s, i.e., the set of all sequences of the form $h_s = (a_1, b_1, \ldots, a_{s-1}, b_{s-1})$. Let $\mathcal{H}$ denote the set of all finite histories: $\mathcal{H} = \cup_{s=1}^{\infty} \mathcal{H}_s$. Let $\Delta(\mathcal{A})$ denote the set of probability distributions over $\mathcal{A}$. A *strategy* σ for the agent is a mapping from $\mathcal{H}$ to $\mathcal{A}$. It prescribes a (randomized) action $\sigma(h_s)$ at each time t.

We assume that a finite set $\{1, \ldots, r\}$ of experts is given. Each expert e is uniquely identified with a strategy σ_e. An experts algorithm provides a rule for deciding, at each stage s, which expert should be followed. An intuitive and popular experts method found in the literature is as follows. Denote by $M_e(s-1)$ the average reward achieved by expert e prior to time stage s[2]. Then, a reasonable rule is to follow expert e in stage s with a probability that is proportional to some monotone function of $M_e(s-1)$. In particular, when this probability is proportional to $\exp\{\eta_s M_e(s-1)\}$, for a certain choice of η_s, this algorithm is known to minimize regret (Auer et al., 2000; Freund and Schapire, 1999). Specifically, let b_s $(s = 1, 2, \ldots)$ denote the observed states of the environment up to stage s, and let σ_X denote the strategy induced by the experts algorithm. Then we have (Auer et al., 2000)

$$\sum_{s'=1}^{s} \mathrm{E}[R(a, b_{s'}) : a \sim \sigma_X(h_{s'})] \geq \sup_e \frac{1}{s} \sum_{s'=1}^{s} \mathrm{E}[R(a, b_{s'}) : a \sim \sigma_e(h_{s'})] - O\left(\frac{|\mathcal{A}| \ln r}{s}\right), \quad (1)$$

where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$ and $a \sim \sigma_X(h_{s'})$ indicates that action a is distributed according to $\sigma_X(h_{s'})$. The main deficiency of the regret minimization approach is that it fails to consider the influence of chosen actions of an agent on the future states of the environment — the inequality (1) holds for any *fixed* sequence (b_s) of states, but does not account for the fact that different choices of actions by the agent may induce different state sequences. This subtlety is also missing in the experts algorithm we described above. At each time stage, the selection of expert is based solely on how well various experts have, or could have, done up to that point, given the state sequence. There is no notion of learning how an expert's actions affect the environment. For instance, in the repeated PD game described in the introduction, assuming that the opponent is playing Tit-for-Tat, the algorithm is unable to establish the connection between the opponent's cooperative moves and his own.

We present a new experts method that is especially tailored to deal with reactive environments. The *Exploration-Exploitation Experts* method (EEE) follows chosen experts for multiple stages rather than picking a different expert each stage. A maximal set of consecutive stages during which the same expert is followed is called a *phase*. Phase numbers are denoted by i. The number of phases during which expert e has been followed is denoted by N_e, the total number of stages during which expert e has been followed is denoted by S_e, and the average reward from phases in which expert e has been followed is denoted by M_e. The general method is stated as follows.

2. In different variants of the algorithm and depending on what information is available to the agent, $M_e(s-1)$ could be either an estimate of the average reward based on the reward achieved by expert e in the stages it was followed, or the reward it could have obtained, had it been played in all stages against the same sequence of states of the environment.

- **Exploration.** An exploration phase consists of picking a random expert e (i.e., from the uniform distribution over $\{1, \ldots, r\}$), and following e's recommendations for a certain number of stages depending on the variant of the method.

- **Exploitation.** An exploitation phase consists of picking an expert e with maximum M_e, breaking ties at random, and following e's recommendations for a certain number of stages depending on the variant of the method.

A general Exploration-Exploitation Experts Method:

1. Initialize $M_e = N_e = S_e = 0$ $(e = 1, \ldots, r)$ and $i = 1$.

2. With probability p_i, perform an exploration phase, and with probability $1-p_i$ perform an exploitation phase; denote by e_i the expert chosen to be followed and by n_i the number of stages chosen for the current phase.

3. Follow expert e_i's instructions for the next n_i stages. Increment $N_{e_i} = N_{e_i} + 1$ and update $S_{e_i} = S_{e_i} + n_i$. Denote by $\tilde{R}$ the average reward accumulated during the current phase of n_i stages and update
$$M_{e_i} = M_{e_i} + \frac{n_i}{S_{e_i}}(\tilde{R} - M_{e_i}) \ .$$

4. Increment $i = i + 1$ and go to step 2.

Note that two sets of parameters in the general method must be chosen to specify a particular experts algorithm: the exploration probabilities p_i, and the phase lengths n_i, for $i = 1, 2, \ldots$.

Throughout the paper, s will denote a stage number, and i will denote a phase number. We denote by $M_1(i), \ldots, M_r(i)$ the values of the registers $M_1, \ldots, M_r$, respectively, at the end of phase i. Similarly, we denote by $N_1(i), \ldots, N_r(i)$ the values of the registers $N_1, \ldots, N_r$, respectively, at the end of phase i. Thus, $M_e(i)$ and $N_e(i)$ are, respectively, the average reward accumulated by expert e and the total number of phases this expert was followed on or before phase i. We will also let $M(s)$ and $M(i)$ denote, without confusion, the average reward accumulated by the algorithm in the first s stages or i phases.

In sections 3 through 6, we consider the case where the length of the phase is $n_i = N_{e_i}$. In section 7 we consider the case where $n_i = L$ for a fixed L.

3. Performance Bounds Based on Actual Expert Performance

EEE keeps track of the average reward $M_e(i)$ achieved by each available expert e. This average reward represents an estimate of the value of that expert. In this section, we compare the average reward $M(i)$ achieved by EEE with the averages achieved by the various experts. We present several bounds characterizing the relationship between $M(i)$ and $M_e(i)$. These bounds are valuable in several ways:

- they provide worst-case guarantees about the performance of EEE;

- they provide a starting point for analyzing the behavior of the algorithm under various assumptions about the environment;

- they quantify the relationship between amount of exploration, expressed by the exploration probabilities p_i, and performance loss. In section 5, we present bounds that quantify the relationship between amount of exploration and the rate at which EEE learns the value of each expert. Putting both bounds together allows for an explicit characterization of the tradeoff between exploration and exploitation.

Throughout the paper, we let $Z_{e,j}$ be an indicator variable of the event "phase j is an exploration phase and expert e is followed," i.e., $Z_{e,j} = 1$ if this statement is true, and 0 otherwise. Let $Z_j = \sum_e Z_{e,j}$. Define

$$\bar{Z}_{i_0,i} \equiv \mathrm{E}\left[\sum_{j=i_0+1}^{i} Z_j\right] = \sum_{j=i_0+1}^{i} p_i.$$

Note that $\bar{Z}_{i_0,i}$ denotes the expected number of exploration phases between phases i_0 and i.

The first theorem establishes that EEE has performance comparable to that of the best expert after a finite number of iterations with high probability.

Theorem 1 *For all i_0, i and ϵ such that $\bar{Z}_{i_0,i} \leq \frac{i\epsilon^2}{16\sqrt{r}u^2} - \frac{i_0\epsilon}{8u}$, we have*

$$\Pr\left(M(i) \leq \max_e \min_{i_0+1\leq j\leq i} M_e(j) - \epsilon\right) \leq \exp\left(-\frac{1}{2i}\left(\frac{i\epsilon^2}{16\sqrt{r}u^2} - \frac{i_0\epsilon}{8u} - \bar{Z}_{i_0,i}\right)^2\right).$$

We can also characterize the expected difference between the average reward of EEE and that of the best expert.

Theorem 2 *For all $i_0 \leq i$ and $\epsilon > 0$, we have*

$$\mathrm{E}\left[M(i) - \max_e \min_{i_0+1\leq j\leq i} M_e(i)\right] \geq -\epsilon - u\frac{i_0(i_0+1)}{i\left(\frac{i}{r}+1\right)} - 2u\left(\frac{3u+2\epsilon}{\epsilon}\right)^2 \frac{\bar{Z}_{i_0,i}}{i}.$$

It follows from Theorem 1 that, under certain assumptions on the exploration probabilities, EEE performs at least as well as the expert that did best, asymptotically.

Corollary 3 *If*

$$\lim_{i\to\infty} \frac{\bar{Z}_{0,i}}{i} = 0\ ,$$

then

$$\Pr\left(\liminf_{s\to\infty} M(s) \geq \max_e \liminf_{i\to\infty} M_e(i)\right) = 1\ . \tag{2}$$

Although the claim of Corollary 3 seems very close to regret minimization, there is an essential difference in that we compare the average reward of our algorithm with the average reward *actually achieved* by each expert in the stages when it was played, as opposed to the estimated average reward based on the whole sequence of states of the environment.

In Theorems 1 and 2 the average reward $M(i)$ achieved by EEE until phase i is compared with $\max_e \min_{i_0+1\leq j\leq i} M_e(j)$. Hence an expert is considered 'good' at phase i only if

	L	R
U	v	v
D	0	R

Table 1: Row player rewards for Example 1

its average performance has been consistently good since earlier phases. This may be counterintuitive and leads to a striking difference between the results in this section and more traditional no-regret properties such as (1). Indeed, no-regret analysis usually involves a comparison between average rewards $M(i)$ and $M_e(i)$ experienced by the experts algorithm and each of the experts in the same phase (or stage). A simple counterexample shows that the bound (2) cannot be improved into a guarantee that $M(i)$ will eventually approach $\max_e M_e(i)$.

Example 1 *Consider a repeated game whose row player's payoffs are given in Table 1, with $0 < v < R$. Suppose that there are two experts: AU and AD, corresponding to the pure strategies that always play action U and always play action D, respectively. Consider the following strategy for the opponent, where $0 < \epsilon < R - v$ and $\gamma = \frac{1}{1.1}\sqrt{(R - v - \epsilon)/R}$:*

- *Start by playing L.*
- *Switch from playing L to playing R when $N_D(i) \leq \gamma N_U(i)$ and $M_D(i) < v$.*
- *Switch from playing R to playing L when $M_D(i) \geq v + \epsilon$.*

We will show that $M(i) \leq M_D(i) - \epsilon/2$ infinitely many times.

The first observation is that the opponent alternates infinitely many times between playing L and playing R. Suppose instead that it plays L forever, starting at some phase i_0. Then $M_D(i) < M_U(i) = v$ for all large enough i. In this case, expert D is followed only during exploration phases, and $N_D(i) \leq \gamma N_U(i)$ for some sufficiently large phase number i, with probability one. After this phase, the opponent must switch to playing R, and we conclude that it cannot play L forever. Similarly the opponent does not play R forever, because after a sufficiently long sequence of R, necessarily, $M_D(i) > v + \epsilon$, and the opponent must switch to playing L.

Denote by $I_1 < I_2 < \ldots$ the phases when the opponent switches from L to R and by $J_1 < J_2 < \ldots$ the phases when it switches from R to L. Note that $I_1 < J_1 < I_2 < J_2 < \ldots$. We will show that for all sufficiently large k, $N_D(J_k) \leq N_U(J_k)$. Note that $M_D(J_k) \geq v + \epsilon$ and $M_D(J_k - 1) < v$, from the definition of J_k. The total reward at the end of phase J_k is given by

$$0.5M_D(I_k)N_D(I_k)(N_D(I_k)+1) + 0.5R\left[N_D(J_k)(N_D(J_k)+1) - N_D(I_k)(N_D(I_k)+1)\right].$$

We conclude that $N_D(J_k)$ is the least n such that

$$M_D(I_k)N_D(I_k)(N_D(I_k)+1) + R\left[n(n+1) - N_D(I_k)(N_D(I_k)+1)\right] \geq (v+\epsilon)n(n+1)\ .$$

We first show that $N_D(J_k) \leq \bar{n} = \sqrt{\frac{R}{R-v-\epsilon}}\,(N_D(I_k)+1)$*, as follows.*

$$\begin{aligned}
& \bar{n} = \sqrt{\frac{R}{R-v-\epsilon}}\,(N_D(I_k)+1) \\
\Rightarrow \quad & \bar{n}^2 = \frac{R}{R-v-\epsilon}\,(N_D(I_k)+1)^2 \\
\Rightarrow \quad & \bar{n}(\bar{n}+1) \geq \frac{R}{R-v-\epsilon}\,N_D(I_k)(N_D(I_k)+1) \\
\Rightarrow \quad & (R-v-\epsilon)\bar{n}(\bar{n}+1) \geq R N_D(I_k)(N_D(I_k)+1) \\
\Rightarrow \quad & M_D(I_k)N_D(I_k)(N_D(I_k)+1) + R\,[\bar{n}(\bar{n}+1) - N_D(I_k)(N_D(I_k)+1)] \\
& \quad \geq (v+\epsilon)\bar{n}(\bar{n}+1) \\
\Rightarrow \quad & N_D(J_k) \leq \bar{n}\ .
\end{aligned}$$

We conclude that, for all sufficiently large k*,*

$$\begin{aligned}
N_D(J_k) &\leq \sqrt{\frac{R}{R-v-\epsilon}} \cdot (N_D(I_k)+1) \\
&\leq \sqrt{\frac{R}{R-v-\epsilon}} \cdot (\gamma N_U(I_k)+1) \\
&\leq N_U(I_k)/1.1 + \sqrt{\frac{R}{R-v-\epsilon}} \\
&\leq N_U(J_k)/1.1 + \sqrt{\frac{R}{R-v-\epsilon}} \\
&\leq N_U(J_k).
\end{aligned}$$

The fourth inequality follows from $I_k < J_k$*, so that* $N_U(I_k) \leq N_U(J_k)$*.*

We conclude that, for all large enough k*,* $N_D(J_k) \leq N_U(J_k)$ *and* $M_D(J_k) \geq v+\epsilon = M_U(J_k)+\epsilon$*. Hence*

$$\begin{aligned}
M(J_k) &\leq \frac{M_U(J_k)+M_D(J_k)}{2} \\
&\leq M_D(J_k) - \epsilon/2,
\end{aligned}$$

and $M(i) < M_D(i) - \epsilon/2$ *infinitely many times.*

4. The Value of an Expert

Theorems 1 and 2 and Corollary 3 are statements about the ability of EEE to *exploit* — provided that exploration is not too large, EEE achieves expected reward that is close to the best average reward observed for any expert. Another important aspect is the ability of EEE to explore and learn the potential value of each expert. The following example shows that the bounds in the previous section may be vacuous without proper consideration of the behavior of the average rewards $M_e(i)$.

Example 2 *Suppose that* $p_i = 0$ *for all* i *and* $R(a,b) > 0$ *for all* a *and* b*. It is easy to show that EEE will choose the same expert in every phase. Indeed, let* e_1 *be the expert chosen at*

	H	T
H	-1	1
T	1	-1

Table 2: Row player rewards for Example 3

phase 1; then we have $M(1) = M_{e_1}(1) > M_e(1) = 0$ *for all* $e \neq e_1$*, and expert* e_1 *is selected again at phase 2. We can show by induction that the same holds for every phase:*

$$M(i) = M_{e_1}(i) > M_e(i) = 0 \ \forall i, \ \forall e \neq e_1.$$

In this case we have the apparently stronger (but meaningless) bound on $M(i)$*:*

$$M(i) = \max_e M_e(i) \quad (i = 1, 2, \ldots) \ .$$

The main issue in the previous example is that $M_e(i)$ is not representative of the actual value of each expert e. In order to obtain a more complete understanding of the behavior of EEE, it is necessary to characterize how it explores and learns the value of each expert through the estimates $M_e(i)$. In this and the next section, we will formally define the value of an expert, and provide results characterizing how fast EEE is able to learn those values, as a function of the amount of exploration it performs.

We start with a definition of a "learnable value" of an expert, in reactive environments. In the regret minimization setting, the value concept, which is used for comparing experts, is the average reward that the expert could have achieved against the (fixed) observed sequence of states of the environment. In reactive environments, this definition is not appropriate. A more suitable definition for the value of an expert is the expected average reward it could achieve, if it were followed exclusively in all time stages. However, it is easy to show that it is impossible for a learning algorithm to guarantee, for all reactive environments, a reward that is close to what the best available expert could have achieved, if played exclusively. The following example illustrates this impossibility.

Example 3 (Password Matching Pennies) *In the Matching Pennies (MP) game, the row and column players have to choose either H ("Heads") or T ("Tails"). If the choices match, the row player loses* 1*; otherwise, he wins* 1*. Consider the following* password strategy *for the column player in the repeated MP game:*

Adversary: *Fix a positive integer* s *and a string* $\sigma^s \in \{H, T\}^s$*. In each of the first* s *stages, play the 50:50 mixed strategy. In each of the stages* $s+1, s+2, \ldots$*, if the sequence of choices of the player during the first* s *stages coincided with the string* σ^s*, then play* T*; otherwise, play the 50:50 mixed strategy.*

Suppose that the row player is using an experts algorithm, and each available expert e *corresponds to a strategy of the form:*

Expert: *Fix a string* $\sigma_e \in \{H, T\}^s$*. During the first* s *stages play according to* σ_e*. In each of the stages* $s+1, s+2, \ldots$*, play* H*.*

Suppose that an expert e^* *with* $\sigma_{e^*} = \sigma^s$ *is available. Then, in order for an experts algorithm to achieve at least the reward of* e^**, it needs to precisely follow the string* σ^s

during the first s stages. Of course, without knowing what σ^s is, the algorithm cannot play it with probability one, nor can it learn anything about it during the play.

The password MP example illustrates the need for a refined notion of the value of an expert. An algorithm that attempts to learn what the best expert would achieve if followed exclusively cannot avoid committing fatal "mistakes." As demonstrated by the MP example, in certain environments, any reasonable learning algorithm must commit such fatal mistakes. Hence, such mistakes cannot, in general, be considered necessarily a weakness of the algorithm. A more realistic notion of the value of an expert is desirable for an adequate assessment and comparison of learning algorithms. Bearing this in mind, we introduce the notion of τ-*value* of an expert. The τ-value is defined with respect to the law π for the evolution of states in the environment. π is a mapping from the set of histories $\mathcal{H}$ to a probability distribution over the set of states $\mathcal{B}$.

Definition 4 *Given an expert e and an environment π, denote for any stage s_0, any possible history h_{s_0} at stage s_0 and any number of stages s,*

$$F(s_0, h_{s_0}, s) = \mathbf{E}\left[\ \tfrac{1}{s}\textstyle\sum_{s=s_0+1}^{s_0+s} R(a_e(s), b(s))\ \ :\ \ a_e(s) \sim \sigma_e(h_s),\ b(s) \sim \pi(h_s)\right]$$

and let

$$G(s) = \inf\{F(s_0, h_{s_0}, s)\ \ :\ \ s_0, h_{s_0}\}\ .$$

The τ-value μ_e^τ *of expert e with respect to the environment π is defined as*

$$\mu_e^\tau = \sup_c \inf_s \{G(s) + c/s^\tau\} \tag{3}$$

In words, a value μ is achievable by expert e if the expert can secure an expected average reward during the s stages between stage s_0 and stage $s_0 + s$ that is at least as much as

$$\mathbf{E}\left[\ \tfrac{1}{s}\textstyle\sum_{s=s_0+1}^{s_0+s} R(a_e(s), b(s))\ \ :\ \ a_e(s) \sim \sigma_e(h_s),\ b(s) \sim \pi(h_s)\right] \geq \mu - \frac{c_\tau}{s^\tau},$$

for some constant c_τ, regardless of the history prior to stage s_0. Note that, asymptotically, the expert is guaranteed to achieve at least as much as the τ-value.

In a previous version of this paper (de Farias and Megiddo, 2004), we introduced the notion of *flexibility* as a way of reasoning about the value of an expert and when it can be learned. We can view the flexibility assumption and previous results as special cases of the results of this paper and of the definition of a τ-value. We now introduce the following definition of flexibility, which holds under weaker conditions than the original definition given in (de Farias and Megiddo, 2004):

Definition 5 (Flexibility)

(i) An environment with state evolution law $\pi(s)$ is said to be flexible with respect to expert e *if there exist constants μ_e, $\tau > 0.25$ and c such that for every stage s_0, every possible history h_{s_0} at stage s_0 and any number of stages s,*

$$\left|\ \mathbf{E}\left[\tfrac{1}{s}\textstyle\sum_{s=s_0+1}^{s_0+s} R(a_e(s), b(s)) - \mu_e\ \ :\ \ a_e(s) \sim \sigma_e(h_s),\ b(s) \sim \pi(h_s)\right]\ \right| \leq \frac{c}{s^\tau}\ .$$

(ii) Flexibility with respect to a set of experts is defined as flexibility with respect to every member of the set.

In words, the expected average reward during the s stages between stage s_0 and stage s_0+s converges (as s tends to infinity) to a limit that does not depend on the history of the play prior to stage s_0. The interest in flexibility arises from the fact that, if the environment is flexible with respect to an expert, the τ-value for that expert actually coincides with the expected average reward that could be achieved by the expert, if it had been followed exclusively. Although as we have seen with the password MP example this does not hold in general, certain classes of strategies lead to flexibility.

Example 4 (: Finite Automata) *In the literature on "bounded rationality", agents are often modelled as finite automata. A* probabilistic automaton strategy *(PAS) is specified by a tuple $\mathcal{A} = \langle M, O, A, \sigma, P\rangle$, where $M = \{1, \ldots, m\}$ is the finite set of internal states of the automaton, A is the set of possible actions, $\mathcal{O}$ is the set of possible outcomes, $\sigma_i(a)$ is the probability of choosing action a while in state i ($i = 1, \ldots, m$) and $P^o = (P^o_{ij})$ ($1 \leq i, j \leq m$) is the matrix of state transition probabilities, given an outcome $o \in \mathcal{O}$. Thus, at any time stage, the automaton picks an action from a probability distribution associated with its current state and transitions into a new state, according to a probability distribution which depends on action of the agent. If both the environment and an expert follow PASs, then a Markov chain is induced over the set of pairs of the respective internal states. If this Markov chain has a single class of recurrent states, then the flexibility assumption holds. Note that we do not limit the size of the automata; a larger set of internal states implies a slower convergence of the average rewards, but does not affect the asymptotic results for EEE.*

Example 5 (: Bounded dependence on the history) *The number of possible histories at stage s grows exponentially with s. Thus, it is reasonable to assume that the choice of action would be based not on the exact detail of the history but rather on the empirical distribution of past actions or patterns of actions. If the environment is believed not to be stationary, then discounting previous observations by recency may be sensible. For instance, if the frequency of each state b of the environment is relevant, the agent might condition his choice at stage $s+1$ on the quantities $\tau_b = \sum_{s'=1}^{s} \beta^{s-s'} \delta_{bb_s}$ where $\beta < 1$ and δ is the Kronecker delta. In this case, only actions b_s at stages s that are relatively recent have a significant impact on τ_b. Therefore strategies based on τ_b should exhibit behavior similar to that of bounded history, and lead to flexibility in the same circumstances as the latter.*

5. Performance Bounds Based on Expected Expert Performance

In this section we present a theorem characterizing how fast EEE learns the τ-value of each expert. Combining this result with Theorem 1, we can also derive the rate at which the average reward achieved by EEE approaches the τ-value of the best expert.

Theorem 6 *Denote $\bar{\tau} = \min(\tau, 1)$. For all $\epsilon > 0$ and i, if*

$$\frac{4r}{3}\left(\frac{4c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}} \leq \bar{Z}_{0,i} \ ,$$

then

$$\Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \epsilon\right) \leq \frac{33u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2 \bar{Z}_{0,i}}{43u^2 r}\right).$$

Corollary 7 *For all $\epsilon > 0$, i_0 and i, if*

$$1. \quad \frac{4r}{3}\left(\frac{12c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}} \leq \bar{Z}_{0,i_0} \text{ , and}$$

$$2. \quad \bar{Z}_{i_0,i} \leq \frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} \text{ ,}$$

then

$$\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq \frac{297u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i_0}}{387u^2 r}\right) + \exp\left(-\frac{1}{2i}\left(\frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} - \bar{Z}_{i_0,i}\right)^2\right). \tag{4}$$

Corollary 7 explicitly quantifies the tradeoff between exploration and exploitation. In particular, we would like to choose exploration probabilities p_j such that $\bar{Z}_{0,i_0}$ is large enough to make the first term in the bound small, and $\bar{Z}_{i_0,i}$ is as small as possible. In Section 6, we analyze several exploration schemes and their effect on the convergence rate of EEE.

We can also derive from Theorems 1 and 6 asymptotic guarantees for EEE.

Corollary 8
If

$$\lim_{i\to\infty} \bar{Z}_{0,i} = \infty \text{ ,}$$

then

$$\Pr\left(\liminf_{i\to\infty} M_e(i) \geq \mu_e^\tau\right) = 1 \text{ .}$$

The following is an immediate result from Corollaries 3 and 8:

Corollary 9
If

$$\lim_{i\to\infty} \bar{Z}_{0,i} = \infty$$

and

$$\lim_{i\to\infty} \frac{\bar{Z}_{0,i}}{i} = 0 \text{ ,}$$

then

$$\Pr\left(\liminf_{i\to\infty} M(i) \geq \max_e \mu_e^\tau\right) = 1 \text{ .}$$

Note that all results presented thus far are stated in terms of phases of the algorithm. Since the ratio between the number of stages in phase i and the total number of phases up to phase i decreases to zero at a rate of at least $1/i$, comparisons between average rewards achieved at the end of each phase can easily be extended to average rewards at any time stage.

Example 6 (: Repeated Prisoner's Dilemma revisited) *Consider playing the repeated PD game against an opponent who plays Tit-for-Tat, and suppose that "Always defect" (AD) and "Always cooperate" (AC) are in the set of experts. Thus, AC induces cooperation in every stage and has a greater τ-value than AD, which induces defection in every stage of the game except for the first one. Indeed, the τ-value of AC corresponds to the highest average reward achievable by any strategy against Tit-for-Tat, and Corollary 9 implies that EEE achieves this reward and induces cooperation. By contrast, as mentioned in the introduction, in order to minimize regret, the standard experts algorithm must play D in almost every stage of the game, and therefore achieves a lower reward.*

6. Exploration Schemes

The results of the previous sections hold under generic choices of the exploration probabilities p_i. In particular, as long as there is infinite exploration and the fraction of exploration phases converges to zero, EEE is able to learn the value of each expert and approach the performance of the best expert. However, different exploration schemes satisfying these properties may lead to substantially different behavior for EEE. In this section, we analyze and compare several exploration schemes from the standpoint of speed of convergence and adaptability.

We apply Corollary 7 to derive an explicit measure for the speed of convergence of EEE as follows. We fix tolerance parameters ϵ and δ and consider the number of phases i required to ensure that:

$$\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq \beta. \tag{5}$$

We use the upper bound given in Corollary 7 to compare exploration schemes. Indeed, define

$$U(i_0, i) = \frac{297u^2}{\epsilon^2} \exp\left(-\frac{\epsilon^2 \bar{Z}_{0,i_0}}{387u^2 r}\right) + \exp\left(-\frac{1}{2i}\left[\max\left(\frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} - \bar{Z}_{i_0,i}, 0\right)\right]^2\right).$$

Then, provided that the conditions of Corollary 7 are satisfied, we have $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i)$. For clarity of exposition, we will focus on the case of $\tau \geq 0.5$. Lower values of τ lead to different convergence rates that can be determined via similar analysis.

6.1 Explore-then-Exploit

We first consider an exploration scheme that minimizes $U(i_0, i)$. In this scheme, all exploration takes place before any exploitation. Indeed, according to expression (4), for any fixed number of iterations i, it is optimal to let $\bar{Z}_{0,i_0} = i_0$ (i.e., $p_j = 1$ for all $j \leq i_0$) and $\bar{Z}_{i_0,i} = 0$ (i.e., $p_j = 0$ for all $j > i_0$).

Theorem 10 *In the explore-than-exploit scheme, for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$ satisfies*

$$i = \Omega\left(\frac{u^3 r\sqrt{r}}{\epsilon^3} \log \frac{u^2}{\epsilon^2 \beta}\right)$$

and

$$i = O\left(\max\left[\frac{u^3 r\sqrt{r}}{\epsilon^3}\log\frac{u^2}{\epsilon^2\beta}, \frac{u^4 r}{\epsilon^4}\log\frac{1}{\beta}\right]\right).$$

Note that the number of phases grows polynomially in $1/\epsilon$, u, and r. More precisely, it is on the order of $O(r^{1.5})$, where r is the total number of experts.

The main drawback of explore-then-exploit is its inability to adapt to changes in the environment — since all exploration occurs first, any change that occurs after exploration has ended cannot be learned. Moreover, the choice of the last exploration phase i_0 depends on parameters of the problem that may not be observable. Finally, it requires fixing i, β and ϵ a priori, and can only achieve optimality within these tolerance parameters.

6.2 Polynomially Decreasing Exploration

In (de Farias and Megiddo, 2004), we have provided asymptotic results equivalent to Corollaries 3 and 9 when $p_j = 1/j$. With this choice, the total number of phases required to satisfy $U(i_0, i) \leq \beta$ grows exponentially in $1/\epsilon$, u, and r.

Theorem 11 *If $p_j = 1/j$, for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$ satisfies*

$$i = \Omega\left(\exp\left(\frac{387u^2 r}{\epsilon^2}\log\frac{297u^2}{\epsilon^2\beta} - 1\right)\right).$$

An alternative scheme, leading to polynomial complexity, can be developed by choosing $p_j = j^{-\alpha}$, for some $\alpha \in (0, 1)$.

Theorem 12 *If $p_j = 1/j^\alpha$ and $\alpha < 1$, for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$ satisfies*

$$i = \Omega\left(\max\left[\frac{u^{\frac{3-\alpha}{1-\alpha}} r^{\frac{3-\alpha}{2(1-\alpha)}}}{\epsilon^{\frac{3-\alpha}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}, \frac{u^{\frac{2}{\alpha}} r^{\frac{1}{2\alpha}}}{\epsilon^{\frac{2}{\alpha}}}\right]\right),$$

and

$$i = O\left(\max\left[\frac{u^{\frac{3-\alpha}{1-\alpha}} r^{\frac{3-\alpha}{2(1-\alpha)}}}{\epsilon^{\frac{3-\alpha}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}, \frac{u^{\frac{2}{\alpha}} r^{\frac{1}{2\alpha}}}{\epsilon^{\frac{2}{\alpha}}}, \frac{ru^4}{\epsilon^4}\log\frac{1}{\beta}\right]\right).$$

6.3 Constant-Rate Exploration

The previous exploration schemes have the property that the frequency of exploration vanishes as the number of phases grows. This property is required in order to achieve the asymptotic optimality results described in Corollaries 3 and 9. However, it also makes EEE increasingly slower in tracking changes in the environment. An alternative approach is to use a constant frequency $\eta \in (0, 1)$ of exploration, i.e., $p_j = \eta$. Constant-rate exploration does not satisfy the conditions of Corollaries 3 and 9. However, for any given tolerance level ϵ, we can choose η so that

$$\Pr\left(\liminf_{i\to\infty} M(i) \geq \max_e \mu_e^\tau - \epsilon\right) = 1\ .$$

Theorem 13 *Suppose that*

$$p_j = \frac{\epsilon^2}{16\sqrt{r}u^2} \qquad (j = 1, 2, \ldots).$$

Then for all $\tau \geq 0.5$ *the smallest number of phases* i *such that* $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$ *satisfies*

$$i = O\left(\frac{r^2u^5}{\epsilon^5} \log \frac{u^2}{\epsilon^2\beta}\right).$$

Constant-rate exploration yields complexity results only slightly worse than the explore-then-exploit scheme. We can also compare it with the polynomially decreasing exploration scheme. If α is chosen to minimize the dependence on u or ϵ, we have

$$i = \Omega\left(\frac{u^{4.56}r^{2.28}}{\epsilon^{4.56}}\right).$$

If α is chosen to minimize the dependence on the number of experts r, we have

$$i = \Omega\left(\frac{u^{7.46}r^{1.86}}{\epsilon^{7.46}}\right).$$

Hence polynomially decreasing exploration does not offer significant improvement in convergence rate over constant-rate exploration.

7. Constant Phase Lengths

So far, we have only considered versions of EEE where the number of stages per phase increases linearly as a function of the number of phases during which the same expert has been followed previously. This growth was used to ensure that, as long as the environment exhibits some regularity, that regularity is captured by the algorithm. For instance, if the environment exhibits cyclic behavior, then EEE correctly learns the long-term value of each expert, regardless of the lengths of the cycles. However, for practical purposes, it may be necessary to slow down the growth of phase lengths in order to achieve good performance in reasonable time. In this section, we consider the possibility of a constant number L of stages in each phase. Following the same steps that we took to prove Theorems 1, 2 and 6, we can derive the following results:

Theorem 14 *Suppose EEE is implemented with phases of fixed length* L*. Then for all* i_0*,* i *and* ϵ *such that* $\bar{Z}_{i_0,i} \leq \frac{i\epsilon^2}{8u^2} - \frac{i_0\epsilon}{4u}$*, we have*

$$\Pr\left(M(i) \leq \max_e \min_{i_0+1 \leq j \leq i} M_e(j) - \epsilon\right) \leq \exp\left(-\frac{1}{2i}\left(\frac{i\epsilon^2}{8u^2} - \frac{i_0\epsilon}{4u} - \bar{Z}_{i_0,i}\right)^2\right).$$

We can also characterize the expected difference between the average reward of EEE and that of the best expert.

Theorem 15 *If EEE is implemented with phases of fixed length L, then for all $i_0 \leq i$ and $\epsilon > 0$,*

$$\mathrm{E}\left[M(i) - \max_e \min_{i_0+1 \leq j \leq i} M_e(i)\right] \geq -\epsilon - u\frac{i_0}{i} - \frac{2u^2}{\epsilon}\frac{\bar{Z}_{i_0,i}}{i}.$$

Theorem 16 *If EEE is implemented with phases of fixed length L, then for all $\epsilon > 0$,*

$$\Pr\left(\inf_{j \geq i} M_e(j) < \mu_e^\tau - \frac{c_\tau}{L^\tau} - \epsilon\right) \leq \frac{3u^2}{\epsilon^2} \exp\left(-\frac{\bar{Z}_{0,i_0}\epsilon^2}{8ru^2}\right).$$

An important qualitative difference between fixed-length phases and increasing-length ones is the absence of the number of experts r in the bound given in Theorem 14. This implies that, in the explore-then-exploit or constant-rate exploration schemes, the algorithm requires a number of phases that grows only linearly with r to ensure that

$$\Pr(M(i) \leq \max_e M_e^\tau - c/L^\tau - \epsilon) \leq \beta \ .$$

Note, however, that we cannot ensure performance better than $\max_e \mu_e^\tau - c_\tau/L^\tau$.

Appendix A. Proof of Results for Linearly Increasing Phases

A.1 Preliminary Analysis

We first define the following random variables with respect to a given pair $i_0 < i$:

$$V = \max_e \min_{i_0+1 \leq j \leq i} M_e(j),$$

$$\mathcal{E}_1 = \{e : \max_{i_0+1 \leq j \leq i} M_e(j) < V\}, \tag{6}$$

$$\mathcal{E}_2 = \{e : \max_{i_0+1 \leq j \leq i} M_e(j) \geq V, M_e(i) < V - \epsilon\}, \tag{7}$$

and

$$j_e = \begin{cases} i_0 & \text{if } e \in \mathcal{E}_1 \\ \max\{j : j \leq i, M_e(j) \geq V\} & \text{if } e \in \mathcal{E}_2 \ . \end{cases}$$

Throughout the proofs, we also make use of a function $\delta(\cdot)$ over the set of logical propositions defined in the following way:

$$\delta(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{if } x \text{ is false} \ . \end{cases}$$

The first lemma shows that, if the average reward of an expert drops considerably between any two phases i and i', then the expert must have been followed for several phases during this interval.

Lemma 17 *For every $\epsilon > 0$, any expert e and any two phases $i < i'$, if*

$$M_e(i') \leq M_e(i) - \epsilon \ ,$$

then

$$N_e(i') - N_e(i) \geq \frac{N_e(i)\epsilon}{3u} \ .$$

Proof Fix $\epsilon > 0$, e and $i < i'$. For simplicity, denote

$$N = N_e(i),$$

and

$$v = M_e(i).$$

Let $I_0 = i$, and $i < I_1 < I_k < \ldots < I_k = i'$, denote the phases when expert e is followed between phases i and i'. It follows that $N_e(I_j) = N + j$. Since all rewards are nonnegative, we have

$$M_e(I_j) \geq M_e(I_{j-1}) \frac{(N+j-1)(N+j)}{(N+j)(N+j+1)} .$$

A simple induction argument yields

$$M_e(I_k) \geq v \frac{N(N+1)}{(N+k)(N+k+1)} . \tag{8}$$

By hypothesis, we also have

$$v - \epsilon \geq M_e(I_k) . \tag{9}$$

Combining (8) and (9) and rearranging terms, we conclude that k must satisfy

$$k^2 + k\,(2N+1) - \frac{\epsilon}{v-\epsilon} N(N+1) \geq 0 . \tag{10}$$

Let $\bar{k} = \epsilon N/3u$. Then,

$$\begin{aligned}
\bar{k}^2 + \bar{k}(2N+1) \; - \frac{\epsilon}{v-\epsilon} N(N+1) &\leq^{(a)} \bar{k}^2 + \bar{k}(2N+1) - \frac{\epsilon}{u} N(N+1) \\
&= \left(\frac{\epsilon}{u}\right)^2 \frac{N^2}{9} + \frac{\epsilon}{u} \cdot \frac{N(2N+1)}{3} - \frac{\epsilon}{u} N(N+1) \\
&\leq^{(b)} \frac{\epsilon}{u} \cdot N \left(\frac{N}{9} + \frac{2N+1}{3} - (N+1) \right) \leq 0 .
\end{aligned}$$

Inequality (a) follows from $v \leq u$, and inequality (b) follows from $\epsilon \leq u$, which holds without loss of generality since all rewards are between 0 and u . Since the left-hand side of (10) is an increasing function on $k \geq 0$, we conclude that $k \geq \bar{k}$ for k that satisfies (10), and the lemma follows. □

The next lemma establishes a lower bound on the total number of stages played up to phase i.

Lemma 18 *For every phase i, we have*

$$i(i/r+1) \;\leq\; \sum_{e=1}^{r} N_e(i)(N_e(i)+1) \;\leq\; i(i+1) .$$

Proof At each phase i, $N_e(i)$ must satisfy

$$\sum_e N_e(i) = i.$$

Therefore, $\sum_{e=1}^{r} N_e(i)(N_e(i)+1)$ is bounded from below by the value of the following quadratic minimization problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_e x_e(x_e+1) \\ \text{subject to} \quad & \sum_e x_e = i \\ & x \geq 0 \ . \end{aligned}$$

Convexity and symmetry imply that the symmetric solution $x_e = i/r$ (for every e) is optimal. On the other hand,

$$\sum_{e=1}^{r} N_e(i)(N_e(i)+1) \ \leq \ \sum_{e=1}^{r} N_e(i) \left(\sum_{e=1}^{r} N_e(i) + 1 \right) \ = \ i(i+1) \ .$$

□

A.2 Proof of Theorem 1

Theorem 1 *For all i_0, i and ϵ such that $\bar{Z}_{i_0,i} \leq \frac{i\epsilon^2}{16\sqrt{r}u^2} - \frac{i_0\epsilon}{8u}$, we have*

$$\Pr\left(M(i) \leq \max_e \min_{i_0+1 \leq j \leq i} M_e(j) - \epsilon \right) \leq \exp\left(-\frac{1}{2i} \left(\frac{i\epsilon^2}{16\sqrt{r}u^2} - \frac{i_0\epsilon}{8u} - \bar{Z}_{i_0,i} \right)^2 \right).$$

Proof For simplicity, let

$$\gamma = \frac{\epsilon}{\sqrt{r}u} \ .$$

We first develop an upper bound on

$$\Pr\left(\sum_e N_e(i)\delta(M_e(i) < V - \epsilon) \geq \gamma i \right) \ .$$

We have

$$\begin{aligned} \Pr\left(\sum_e N_e(i)\delta(M_e(i) < V - \epsilon) \geq \gamma i \right) \ &\leq \ \Pr\left(\sum_{e \in \mathcal{E}_1 \cup \mathcal{E}_2} N_e(i) \geq \gamma i \right) \\ &= \ \Pr\left(\sum_{e \in \mathcal{E}_1 \cup \mathcal{E}_2} [N_e(j_e) + (N_e(i) - N_e(j_e))] \geq \gamma i \right) \ . \end{aligned}$$

Note that, for all $e \in \mathcal{E}_1 \cup \mathcal{E}_2$, we have

$$N_e(i) - N_e(j_e) = \sum_{j=j_e+1}^{i} Z_{e,j} \ .$$

Moreover, for all $e \in \mathcal{E}_2$, since $M_e(j_e) \geq V$ and $M_e(i) < V - \epsilon$, by Lemma 17 we conclude that

$$N_e(j_e) \leq \frac{3u}{\epsilon} \left[N_e(i) - N_e(j_e) \right] = \frac{3u}{\epsilon} \sum_{j=j_e+1}^{i} Z_{e,j}. \tag{11}$$

Hence

$$\begin{aligned}
&\Pr\left(\sum_{e \in \mathcal{E}_1 \cup \mathcal{E}_2} \left[N_e(j_e) + (N_e(i) - N_e(j_e)) \right] \geq \gamma i \right) \\
&= \Pr\left(\sum_{e \in \mathcal{E}_1} \left[N_e(i_0) + \sum_{j=i_0+1}^{i} Z_{e,j} \right] + \sum_{e \in \mathcal{E}_2} \left[N_e(j_e) + \sum_{j=j_e+1}^{i} Z_{e,j} \right] \geq \gamma i \right) \\
&\leq \Pr\left(i_0 + \sum_{e \in \mathcal{E}_1} \sum_{j=i_0+1}^{i} Z_{e,j} + \left(\frac{3u}{\epsilon} + 1 \right) \sum_{e \in \mathcal{E}_2} \sum_{j=j_e+1}^{i} Z_{e,j} \geq \gamma i \right) \\
&\leq \Pr\left(\frac{3u + \epsilon}{\epsilon} \sum_{e \in \mathcal{E}_1 \cup \mathcal{E}_2} \sum_{j=i_0+1}^{i} Z_{e,j} \geq \gamma i - i_0 \right) \\
&\leq \Pr\left(\sum_{j=i_0+1}^{i} Z_j \geq \frac{\epsilon(\gamma i - i_0)}{3u + \epsilon} \right) \\
&\leq \exp\left(-\frac{1}{2(i - i_0)} \left(\frac{\epsilon(\gamma i - i_0)}{3u + \epsilon} - \bar{Z}_{i_0,i} \right)^2 \right) \ .
\end{aligned}$$

The first inequality follows from $\sum_{e \in \mathcal{E}_1} N_e(i_0) \leq i_0$ and (11). The last step is an application of Hoeffding's inequality (Hoeffding, 1963).

Now, suppose that

$$\sum_e N_e(i) \delta(M_e(i) < V - \epsilon) \leq \gamma i.$$

Then following the same reasoning of Lemma 18, we conclude that

$$\sum_e N_e(i)(N_e(i) + 1) \delta(M_e(i) < V - \epsilon) \leq \gamma i(\gamma i + 1),$$

and we have

$$\begin{aligned}
M(i) &= \frac{\sum_e N_e(i)(N_e(i) + 1) M_e(i)}{\sum_e N_e(i)(N_e(i) + 1)} \\
&\geq V - \epsilon - (V - \epsilon) \frac{\sum_e N_e(i)(N_e(i) + 1) \delta(M_e(i) < V - \epsilon)}{\sum_e N_e(i)(N_e(i) + 1)}
\end{aligned}$$

$$
\begin{aligned}
&\geq V-\epsilon-(V-\epsilon)\frac{\gamma i(\gamma i+1)}{i(i/r+1)}\\
&\geq V-\epsilon-(V-\epsilon)\max(\gamma^2 r,\gamma)\\
&= V-\epsilon-(V-\epsilon)\max\left(\frac{\epsilon^2}{u^2},\frac{\epsilon}{u\sqrt{r}}\right)\\
&= V-\epsilon-(V-\epsilon)\frac{\epsilon}{u}\\
&\geq V-2\epsilon.
\end{aligned}
$$

In the second inequality, we have used the lower bound on $\sum_e N_e(i)(N_e(i)+1)$ given in Lemma 18.

We conclude that

$$
\begin{aligned}
\Pr\left(M(i)\leq V-2\epsilon\right) &\leq \Pr\left(\sum_e N_e(i)\delta(M_e(i)<V-\epsilon)\leq \gamma i\right)\\
&\leq \exp\left(-\frac{1}{2i}\left(\frac{\epsilon(\gamma i-i_0)}{3u+\epsilon}-\bar{Z}_{i_0,i}\right)^2\right).
\end{aligned}
$$

□

A.3 Proof of Theorem 2

Theorem 2 *For all $i_0 \leq i$ and $\epsilon > 0$, we have*

$$
\mathrm{E}\left[M(i)-\max_e \min_{i_0+1\leq j\leq i} M_e(i)\right] \geq -\epsilon - u\frac{i_0(i_0+1)}{i\left(\frac{i}{r}+1\right)} - 2u\left(\frac{3u+2\epsilon}{\epsilon}\right)^2\frac{\bar{Z}_{i_0,i}}{i}.
$$

Proof We have

$$
\begin{aligned}
\mathbf{E}\left[M(i)-\max_e \min_{i_0+1\leq j\leq i} M_e(i)\right] &= \mathbf{E}\left[\frac{\sum_e N_e(i)(N_e(i)+1)M_e(i)}{\sum_e N_e(i)(N_e(i)+1)}-V\right]\\
&\geq \mathbf{E}\left[V-\epsilon-(V-\epsilon)\frac{\sum_e N_e(i)(N_e(i)+1)\delta(M_e(i)<V-\epsilon)}{i\left(\frac{i}{r}+1\right)}-V\right]\\
&\geq -\epsilon-u\frac{\mathbf{E}\left[\sum_e N_e(i)(N_e(i)+1)\delta(M_e(i)<V-\epsilon)\right]}{i\left(\frac{i}{r}+1\right)}. \qquad (12)
\end{aligned}
$$

By the definitions of $\mathcal{E}_1$ and $\mathcal{E}_2$, (see (6) and (7)),

$$
\begin{aligned}
&\mathbf{E}\left[\sum_e N_e(i)(N_e(i)+1)\delta(M_e(i)<V-\epsilon)\right]\\
&\qquad \leq \mathbf{E}\left[\sum_{e\in\mathcal{E}_1\cup\mathcal{E}_2}\left(N_e(j_e)+\sum_{j=j_e+1}^{i} Z_{e,j}\right)\left(N_e(j_e)+\sum_{j=j_e+1}^{i} Z_{e,j}+1\right)\right]. \qquad (13)
\end{aligned}
$$

Note that

$$\mathbf{E}\left[\sum_{e\in\mathcal{E}_1} N_e(j_e)(N_e(j_e)+1)\right] \;=\; \mathbf{E}\left[\sum_{e\in\mathcal{E}_1} N_e(i_0)(N_e(i_0)+1)\right] \;\le\; i_0(i_0+1)\;. \tag{14}$$

The inequality follows from Lemma 18. Moreover,

$$\begin{aligned}
\mathbf{E}\left[\sum_{e\in\mathcal{E}_1} N_e(j_e)\sum_{j=j_e+1}^{i} Z_{e,j}\right] &= \mathbf{E}\left[\sum_{e\in\mathcal{E}_1} N_e(i_0)\sum_{j=i_0+1}^{i} Z_{e,j}\right] \\
&\le \mathbf{E}\left[\sum_{e} N_e(i_0)\sum_{j=i_0+1}^{i} Z_{e,j}\right] \\
&= \mathbf{E}\left[\sum_{e} N_e(i_0)\right]\cdot\mathbf{E}\left[\sum_{j=i_0+1}^{i} Z_{1,j}\right] \\
&\le \frac{i_0}{r}\bar{Z}_{i_0,i}\;.
\end{aligned} \tag{15}$$

For all $e\in\mathcal{E}_2$, by Lemma 17 we have

$$N_e(i) - N_e(j_e) \ge \frac{\epsilon N_e(j_e)}{3u}$$

so that

$$N_e(j_e) \le \frac{3u}{\epsilon}\sum_{j=j_e+1}^{i} Z_{e,j},$$

and

$$\begin{aligned}
&\mathbf{E}\left[\sum_{e\in\mathcal{E}_2}\left(N_e(j_e)+\sum_{j=j_e+1}^{i} Z_{e,j}\right)\left(N_e(j_e)+\sum_{j=j_e+1}^{i} Z_{e,j}+1\right)\right] \\
&\le \mathbf{E}\left[\sum_{e\in\mathcal{E}_2}\left(\frac{3u+\epsilon}{\epsilon}\sum_{j=j_e+1}^{i} Z_{e,j}\right)\left(\frac{3u+\epsilon}{\epsilon}\sum_{j=j_e+1}^{i} Z_{e,j}+1\right)\right]\;.
\end{aligned} \tag{16}$$

Finally, for every expert e,

$$\begin{aligned}
\mathbf{E}\left[\left(\sum_{j=i_0+1}^{i} Z_{e,j}\right)^2\right] &= \sum_{j=i_0+1}^{i}\mathbf{E}\left[Z_{e,j}^2\right] + \sum_{j=i_0+1}^{i}\sum_{k=i_0+1,k\neq j}^{i}\mathbf{E}\left[Z_{e,j}Z_{e,k}\right] \\
&= \sum_{j=i_0+1}^{i}\mathbf{E}\left[Z_{e,j}\right] + \sum_{j=i_0+1}^{i}\mathbf{E}\left[Z_{e,j}\right]\sum_{k=i_0+1,k\neq j}^{i}\mathbf{E}\left[Z_{e,k}\right] \\
&\le \sum_{j=i_0+1}^{i}\mathbf{E}\left[Z_{e,j}\right] + \sum_{j=i_0+1}^{i}\mathbf{E}\left[Z_{e,j}\right]\sum_{k=i_0+1}^{i}\mathbf{E}\left[Z_{e,k}\right] \\
&= \frac{1}{r}\bar{Z}_{i_0,i} + \frac{1}{r^2}\bar{Z}_{i_0,i}^2\;.
\end{aligned} \tag{17}$$

From (13), (14), (15),(16) and (17) we conclude that

$$
\begin{aligned}
\mathbf{E} & \left[\sum_e N_e(i)(N_e(i)+1)\delta(M_e(i) < V-\epsilon)\right] \\
& \leq \mathbf{E}\left[\sum_{e\in\mathcal{E}_1} N_e(i_0)(N_e(i_0+1)\right] + 2\mathbf{E}\left[\sum_{e\in\mathcal{E}_1} N_e(i_0)\sum_{j=i+1}^{i} Z_{e,j}\right] \\
& \quad + \mathbf{E}\left[\sum_{e\in\mathcal{E}_1}\sum_{j=i_0+1}^{i} Z_{e,j}\left(\sum_{j=i_0+1}^{i} Z_{e,j}+1\right)\right] \\
& \quad + \mathbf{E}\left[\sum_{e\in\mathcal{E}_2}\sum_{j=i_0+1}^{i} \frac{3u+\epsilon}{\epsilon} Z_{e,j}\left(\sum_{j=i_0+1}^{i} \frac{3u+\epsilon}{\epsilon} Z_{e,j}+1\right)\right] \\
& \leq i_0(i_0+1) + \frac{2i_0}{r}\bar{Z}_{i_0,i} + \frac{3u+\epsilon}{\epsilon}\cdot\mathbf{E}\left[\sum_e\sum_{j=i_0+1}^{i} Z_{e,j}\right] \\
& \quad + \frac{(3u+\epsilon)^2}{\epsilon^2}\cdot\mathbf{E}\left[\sum_e\left(\sum_{j=i_0+1}^{i} Z_{e,j}\right)^2\right] \\
& \leq i_0(i_0+1) + \frac{2i_0}{r}\bar{Z}_{i_0,i} + \frac{3u+\epsilon}{\epsilon}\bar{Z}_{i_0,i} \\
& \quad + \frac{(3u+\epsilon)^2}{\epsilon^2}\left(\bar{Z}_{i_0,i} + \frac{1}{r}\bar{Z}^2_{i_0,i}\right) \\
& \leq i_0(i_0+1) + \frac{1}{r}\left(\frac{3u+\epsilon}{\epsilon}\right)^2 (i_0+\bar{Z}_{i_0,i})\bar{Z}_{i_0,i} + \left(\frac{3u+2\epsilon}{\epsilon}\right)^2 \bar{Z}_{i_0,i}\ . \qquad (18)
\end{aligned}
$$

It follows from (12) and (18) that

$$
\begin{aligned}
\mathbf{E}\Bigg[M(i) - \max_e \min_{i_0+1\leq j\leq i} M_e(j)\Bigg] & \\
& \geq -\epsilon - u\frac{i_0(i_0+1) + \frac{1}{r}\left(\frac{3u+\epsilon}{\epsilon}\right)^2 (i_0+\bar{Z}_{i_0,i})\bar{Z}_{i_0,i} + \left(\frac{3u+2\epsilon}{\epsilon}\right)^2 \bar{Z}_{i_0,i}}{i\left(\frac{i}{r}+1\right)} \\
& \geq -\epsilon - u\frac{i_0(i_0+1)}{i\left(\frac{i}{r}+1\right)} - u\left(\frac{3u+\epsilon}{\epsilon}\right)^2 \frac{\frac{i}{r}\bar{Z}_{i_0,i}}{i\left(\frac{i}{r}+1\right)} - u\left(\frac{3u+2\epsilon}{\epsilon}\right)^2 \frac{\bar{Z}_{i_0,i}}{i} \\
& \geq -\epsilon - u\frac{i_0(i_0+1)}{i\left(\frac{i}{r}+1\right)} - 2u\left(\frac{3u+2\epsilon}{\epsilon}\right)^2 \frac{\bar{Z}_{i_0,i}}{i}\ .
\end{aligned}
$$

□

A.4 Proof of Corollary 3

Corollary 3 *If* $\frac{\bar{Z}_{0,i}}{i}$ *converges to zero, we have*

$$\Pr\left(\liminf_{s\to\infty} M(s) \geq \max_e \liminf_{i\to\infty} M_e(i)\right) = 1\ . \tag{19}$$

Proof Let $\epsilon > 0$ and $i_0 = i\epsilon/(2\sqrt{r}u)$. From Theorem 1,

$$\begin{aligned}
\Pr\left(M(i) \leq \max_e \min_{i_0+1\leq j\leq i} M_e(j) - 2\epsilon\right) &\leq \exp\left(-\frac{1}{2(i-i_0)}\left(\frac{\epsilon(i\epsilon - i_0\sqrt{r}u)}{3\sqrt{r}u^2 + \sqrt{r}u\epsilon} - \bar{Z}_{i_0,i}\right)^2\right) \\
&\leq \exp\left(-\frac{1}{2(i-i_0)}\left(\frac{\epsilon^2 i}{6\sqrt{r}u^2 + 2\sqrt{r}u\epsilon} - \bar{Z}_{i_0,i}\right)^2\right) \\
&\leq \exp\left(-\frac{i}{2}\left(\frac{\epsilon^2}{6\sqrt{r}u^2 + 2\sqrt{r}u\epsilon} - \frac{\bar{Z}_{i_0,i}}{i}\right)^2\right)\ .
\end{aligned}$$

Since $\lim_{i\to\infty} \bar{Z}_{i_0,i}/i = 0$, it follows that

$$\sum_{i=1}^{\infty} \Pr\left(M(i) \leq \max_e \min_{i_0+1\leq j\leq i} M_e(j) - 2\epsilon\right) < \infty,$$

and by the Borel-Cantelli Lemma (Feller, 1971), with probability 1 there is a (random) phase $I_0 < \infty$ such that

$$M(i) \geq \max_e \min_{i_0+1\leq j\leq i} M_e(j) - 2\epsilon$$

for all $i \geq I_0$. Moreover, with probability 1 there is also a (random) phase $I_1 < \infty$ such that, for all experts e,

$$\min_{i_0+1\leq j\leq i} M_e(j) \geq \liminf_{i'\to\infty} M_e(i') - \epsilon,$$

for all $i \geq I_1$. We conclude that, with probability one,

$$M(i) \geq \max_e \liminf_{i'\to\infty} M_e(i') - 3\epsilon,$$

for all $i \geq \max(I_0, I_1)$. Since ϵ is arbitrary, the corollary follows. □

A.5 Proof of Theorem 6

Theorem 6 *Let* $\bar{\tau} = \min(\tau, 1)$. *For all* $\epsilon > 0$ *and* i *such that*

$$\frac{4r}{3}\left(\frac{4c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}} \leq \bar{Z}_{0,i},$$

we have

$$\Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \epsilon\right) \leq \frac{33u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i}}{43u^2 r}\right).$$

We will start with two auxiliary lemmas.

Lemma 19 *For all $k \leq \bar{Z}_{0,i}/r$, we have*

$$\Pr\left(N_e(i) < k\right) \leq \exp\left[-\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r}\right].$$

Proof Recall the definition of $Z_{e,i}$. Then

$$\mathbf{E}\left[\sum_{j=1}^{i} Z_{e,j}\right] = \frac{\bar{Z}_{0,i}}{r}.$$

Now note that

$$N_e(i) \geq \sum_{j=1}^{i} Z_{e,j}.$$

It follows that

$$\begin{aligned}\Pr\left(N_e(i) < k\right) &\leq \Pr\left(\sum_{j=1}^{i} Z_{e,j} < k\right) \\ &\leq \exp\left[-\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r}\right].\end{aligned}$$

The second inequality is an application of the multiplicative Chernoff bound (Chernoff, 1952) □

Lemma 20 *Denote by $I_1, I_2, \ldots$ the phases during which expert e is followed, and denote $\bar{\tau} = \min(\tau, 1)$. For all $\epsilon > 0$ and $k \geq \left(4c_\tau/(\epsilon(2-\bar{\tau}))\right)^{1/\bar{\tau}}$,*

$$\Pr\left(M_e(I_k) < \mu_e^\tau - \epsilon\right) \leq \exp\left(-\frac{k\epsilon^2}{32u^2}\right). \tag{20}$$

Proof Let h_k denote the history before phase I_k. Let $\tilde{R}_k$ denote the average reward achieved by expert e during phase I_k, and

$$\hat{R}_k = \mathbf{E}\left[\tilde{R}_k \mid h_k\right].$$

Denote

$$S_k = kM_e(I_k) = \sum_{j=1}^{k} \frac{2j}{k+1} \cdot \tilde{R}_j$$

and

$$\hat{S}_k = \sum_{j=1}^{k} \frac{2j}{k+1} \hat{R}_j.$$

Then,

$$\begin{aligned}
\hat{S}_k - k\mu_e^\tau &= \sum_{j=1}^{k} \frac{2j}{k+1} \cdot \mathbf{E}\left[\tilde{R}_j \mid h_j\right] - k\mu_e^\tau \\
&= \sum_{j=1}^{k} \frac{2j}{k+1} \cdot \mathbf{E}\left[\tilde{R}_j - \mu_e^\tau \mid h_j\right] \\
&\geq -\sum_{j=1}^{k} \frac{2j}{k+1} \cdot \frac{c_\tau}{j^\tau} \\
&\geq -\sum_{j=1}^{k} \frac{2j}{k+1} \cdot \frac{c_\tau}{j^{\bar{\tau}}} \\
&\geq -\frac{2c_\tau k^{1-\bar{\tau}}}{2-\bar{\tau}} \\
&\geq -\frac{k\epsilon}{2} \,.
\end{aligned}$$

Thus, for all $\theta > 0$,

$$\begin{aligned}
\mathbf{E}\left[\exp(\theta(-S_k + k\mu_e^\tau))\right] &= \mathbf{E}\left[\exp(\theta(-S_k + \hat{S}_k - \hat{S}_k + k\mu_e^\tau))\right] \\
&\leq \mathbf{E}\left[\exp(\theta(-S_k + \hat{S}_k))\right] \cdot \exp\left(\frac{\theta k\epsilon}{2}\right). \qquad (21)
\end{aligned}$$

We now have

$$\begin{aligned}
\mathbf{E}\Big[\exp(\theta(-S_k+\hat{S}_k))\Big] &= \mathbf{E}\left[\mathbf{E}\left[\exp(\theta(-S_k+\hat{S}_k)) \mid h_k\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\exp\left(\sum_{j=1}^{k} \frac{2j}{k+1}\theta(\hat{R}_j - \tilde{R}_j)\right) \mid h_k\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\exp\left(\frac{2k}{k+1}\theta(\hat{R}_k - \tilde{R}_k)\right) \mid h_k\right] \cdot \exp\left(\sum_{j=1}^{k-1} \frac{2j}{k+1}\theta(\hat{R}_j - \tilde{R}_j)\right)\right] \\
&= \mathbf{E}\left[\Pi_{j=1}^{k}\mathbf{E}\left[\exp\left(\frac{2j}{k+1}\theta(\hat{R}_j - \tilde{R}_j)\right) \mid h_j\right]\right] \,. \qquad (22)
\end{aligned}$$

We have $\mathbf{E}[\hat{R}_j - \tilde{R}_j \mid h_j] = 0$ and $|\hat{R}_j - \tilde{R}_j|2j/(k+1) \leq 2u$, hence (Williams, 1991)

$$\mathbf{E}\left[\exp\left(\frac{2j}{k+1}\theta(\hat{R}_j - \tilde{R}_j)\right) \mid h_j\right] \leq \exp(2u^2\theta^2). \qquad (23)$$

It follows from (22) and (23) that

$$\mathbf{E}\left[\exp(\theta(-S_k + \hat{S}_k))\right] \leq \exp(2u^2\theta^2 k) \,. \qquad (24)$$

We have

$$\exp(\theta k\epsilon)\Pr\left(-S_k + k\mu_e^\tau > k\epsilon\right) \le \mathbf{E}\left[\exp(\theta(-S_k + k\mu_e^\tau))\right] \le \exp\left(2u^2\theta^2 k + \frac{\theta k\epsilon}{2}\right) ,$$

where the last inequality follows from (21) and (24). It follows that

$$\Pr\left(M_e(I_k) < \mu_e^\tau - \epsilon\right) = \Pr\left(-S_k + k\mu_e^\tau > k\epsilon\right) \le \exp\left(-\frac{\theta k\epsilon}{2} + 2u^2\theta^2 k\right) .$$

Minimizing over $\theta > 0$ yields (20). □

We are now poised to prove Theorem 6.

Proof Denote

$$k = \left(1 - \frac{\epsilon}{4u}\right)\frac{\bar{Z}_{0,i}}{r} . \tag{25}$$

Then,

$$k \ge \frac{3\bar{Z}_{0,i}}{4r} \ge \left(\frac{4c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}} . \tag{26}$$

Now we have

$$\begin{aligned}
\Pr\left(\inf_{j\ge i} M_e(j) < \mu_e^\tau - \epsilon\right) &\le \Pr\left(\inf_{j\ge i} M_e(j) < \mu_e^\tau - \epsilon, N_e(i) \ge k\right) + \Pr\left(N_e(i) < k\right) \\
&\le \Pr\left(\inf_{j\ge k} M_e(I_j) < \mu_e^\tau - \epsilon\right) + \Pr\left(N_e(i) < k\right) \\
&\le \sum_{j=k}^{\infty} \Pr\left(M_e(I_j) < \mu_e^\tau - \epsilon\right) + \Pr\left(N_e(i) < k\right) \\
&\le \sum_{j=k}^{\infty} \exp\left(-\frac{j\epsilon^2}{32u^2}\right) + \exp\left[-\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r}\right] \\
&\le \frac{32u^2}{\epsilon^2}\exp\left(-\frac{k\epsilon^2}{32u^2}\right) + \exp\left[-\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r}\right] . \qquad (27)
\end{aligned}$$

The fourth inequality follows from Lemmas 19 and 20, which holds since k satisfies (26).

By the definition of k (25),

$$\begin{aligned}
\frac{k\epsilon^2}{32u^2} &= \left(1 - \frac{\epsilon}{4u}\right)\frac{\bar{Z}_{0,i}}{r}\frac{\epsilon^2}{32u^2} \\
&\ge \frac{3\bar{Z}_{0,i}\epsilon^2}{128u^2 r} \\
&\ge \frac{\bar{Z}_{0,i}\epsilon^2}{43u^2 r} \qquad (28)
\end{aligned}$$

and

$$\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r} = \frac{\bar{Z}_{0,i}\epsilon^2}{32u^2 r}. \tag{29}$$

Combining (27), (28) and (29) yields

$$\begin{aligned}
\Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \epsilon\right) &\leq \frac{32u^2}{\epsilon^2}\exp\left(-\frac{k\epsilon^2}{32u^2}\right) \\
&\quad + \exp\left[-\left(1-\frac{kr}{\bar{Z}_{0,i}}\right)^2\frac{\bar{Z}_{0,i}}{2r}\right] \\
&\leq \frac{32u^2}{\epsilon^2}\exp\left(-\frac{\bar{Z}_{0,i}\epsilon^2}{43u^2r}\right) + \exp\left(-\frac{\bar{Z}_{0,i}\epsilon^2}{32u^2r}\right) \\
&\leq \frac{33u^2}{\epsilon^2}\exp\left(-\frac{\bar{Z}_{0,i}\epsilon^2}{43u^2r}\right) \ .
\end{aligned}$$

□

A.6 Proof of Corollary 7

Corollary 7 *For all $\epsilon > 0$, i_0 and i satisfying:*

$$\begin{aligned}
&1. \quad \frac{4r}{3}\left(\frac{12c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}} \leq \bar{Z}_{0,i_0} \ , \text{ and} \\
&2. \quad \bar{Z}_{i_0,i} \leq \frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} \ ,
\end{aligned}$$

then

$$\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq \frac{297u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i_0}}{387u^2r}\right) + \exp\left(-\frac{1}{2i}\left(\frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} - \bar{Z}_{i_0,i}\right)^2\right).$$

Proof Let e^* correspond to an expert with $\mu_{e^*}^\tau = \max_e \mu_e^\tau$. Now we have

$$\begin{aligned}
\Pr\left(M(i) \leq \mu_{e^*}^\tau - 3\epsilon\right) &\leq \Pr\left(M(i) \leq \mu_{e^*}^\tau - 3\epsilon, \inf_{j\geq i_0} M_{e^*}(j) \geq \mu_{e^*}^\tau - \epsilon\right) \\
&\quad + \Pr\left(\inf_{j\geq i_0} M_{e^*}(j) \leq \mu_{e^*}^\tau - \epsilon\right) \\
&\leq \Pr\left(M(i) \leq \inf_{j\geq i_0} M_{e^*}(j) - 2\epsilon\right) + \Pr\left(\inf_{j\geq i_0} M_{e^*}(j) \leq \mu_{e^*}^\tau - \epsilon\right) \\
&\leq \Pr\left(M(i) \leq \max_e \min_{i_0\leq j\leq i} M_e(j) - 2\epsilon\right) + \Pr\left(\inf_{j\geq i_0} M_{e^*}(j) \leq \mu_{e^*}^\tau - \epsilon\right) \ .
\end{aligned}$$

The result then follows from application of Theorems 1 and 6. □

A.7 Proof of Corollary 8

Corollary 8 *If* $\lim_{i\to\infty} \bar{Z}_{0,i} = \infty$, *we have*

$$\Pr\left(\liminf_{i\to\infty} M_e(i) \geq \mu_e^\tau\right) = 1.$$

Proof For any $\epsilon > 0$ and i ,

$$\Pr\left(\liminf_{j\to\infty} M_e(j) < \mu_e^\tau - \epsilon\right) \leq \Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \epsilon\right) \leq \frac{33u^2}{\epsilon^2} \exp\left(-\frac{\bar{Z}_{0,i}\epsilon^2}{43u^2 r}\right) \ .$$

Letting i tend to infinity, for every $\epsilon > 0$,

$$\Pr\left(\liminf_{j\to\infty} M_e(j) < \mu_e^\tau - \epsilon\right) = 0 \ ,$$

hence,

$$\Pr\left(\liminf_{j\to\infty} M_e(j) < \mu_e^\tau\right) = 0 \ .$$

□

Appendix B. Proof of Results on Exploration Schemes

The following lemma provides basic bounds on the values of i_0 and i required to ensure that $U(i_0, i) \leq \beta$.

Lemma 21

(i) If

$$\bar{Z}_{0,i_0} \leq \frac{297u^2 r}{\epsilon^2} \log \frac{387u^2}{\epsilon^2 \beta} \tag{30}$$

or

$$i \leq \left(\frac{i_0\epsilon}{12u} + \bar{Z}_{i_0,i}\right) \frac{36\sqrt{r}u^2}{\epsilon^2} \ , \tag{31}$$

then $U(i_0, i) > \beta$.

(ii) If

$$\bar{Z}_{0,i_0} \geq \max\left[\frac{387u^2 r}{\epsilon^2} \log \frac{594u^2}{\epsilon^2\beta}, \frac{4r}{3}\left(\frac{12c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}}\right] \tag{32}$$

and

$$i \geq \max\left[\left(\frac{i_0\epsilon}{12u} + \bar{Z}_{i_0,i}\right) \frac{72\sqrt{r}u^2}{\epsilon^2}, \frac{10368ru^4}{\epsilon^4} \log \frac{2}{\beta}\right] \ , \tag{33}$$

then $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$.

Proof Recall that

$$U(i_0,i) = \frac{297u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i_0}}{387u^2r}\right) + \exp\left(-\frac{1}{2i}\left[\max\left(\frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} - \bar{Z}_{i_0,i}, 0\right)\right]^2\right).$$

If (30) holds, we have

$$\begin{aligned} U(i_0,i) &\geq \frac{297u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i_0}}{387u^2r}\right) \\ &\geq \beta. \end{aligned}$$

If (31) holds, we have

$$\begin{aligned} U(i_0,i) &\geq \exp\left(-\frac{1}{2i}\left[\max\left(\frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} - \bar{Z}_{i_0,i}, 0\right)\right]^2\right) \\ &\geq 1. \end{aligned}$$

Now suppose that (32) holds. Then

$$\frac{297u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i_0}}{387u^2r}\right) \leq \frac{\beta}{2}.$$

Moreover, if (33) holds, then

$$\exp\left(-\frac{1}{2i}\left[\max\left(\frac{i\epsilon^2}{36\sqrt{r}u^2} - \frac{i_0\epsilon}{12u} - \bar{Z}_{i_0,i}, 0\right)\right]^2\right) \leq \frac{\beta}{2}.$$

We conclude that $U(i_0,i) \leq \beta$. Note that, under (32) and (33), the conditions of Corollary 7 are satisfied. It follows that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0,i) \leq \beta$. □

Theorem 10 *In the explore-than-exploit scheme, for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0,i) \leq \beta$ satisfies*

$$i = \Omega\left(\frac{u^3r\sqrt{r}}{\epsilon^3}\log\frac{u^2}{\epsilon^2\beta}\right)$$

and

$$i = O\left(\max\left[\frac{u^3r\sqrt{r}}{\epsilon^3}\log\frac{u^2}{\epsilon^2\beta}, \frac{u^4r}{\epsilon^4}\log\frac{1}{\beta}\right]\right).$$

Proof The lower bound follows from (30) and (31). The upper bound follows from (32) and (33). □

Theorem 11 *If $p_j = 1/j$, for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0,i) \leq \beta$ satisfies*

$$i = \Omega\left(\exp\left(\frac{387u^2r}{\epsilon^2}\log\frac{297u^2}{\epsilon^2\beta} - 1\right)\right).$$

Proof When $p_j = 1/j$, we have

$$\bar{Z}_{0,i_0} \leq \log(i_0) + 1 \ .$$

It follows from (30) that i_0 must satisfy

$$i_0 > \exp\left(\frac{387u^2 r}{\epsilon^2}\log\frac{297u^2}{\epsilon^2\beta} - 1\right).$$

Since $i \geq i_0$, the result follows. □

Theorem 12 *If $p_j = 1/j^\alpha$ and $\alpha < 1$, for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$ satisfies*

$$i = \Omega\left(\max\left[\frac{u^{\frac{3-\alpha}{1-\alpha}} r^{\frac{3-\alpha}{2(1-\alpha)}}}{\epsilon^{\frac{3-\alpha}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}, \frac{u^{\frac{2}{\alpha}} r^{\frac{1}{2\alpha}}}{\epsilon^{\frac{2}{\alpha}}}\right]\right),$$

and

$$i = O\left(\max\left[\frac{u^{\frac{3-\alpha}{1-\alpha}} r^{\frac{3-\alpha}{2(1-\alpha)}}}{\epsilon^{\frac{3-\alpha}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}, \frac{u^{\frac{2}{\alpha}} r^{\frac{1}{2\alpha}}}{\epsilon^{\frac{2}{\alpha}}}, \frac{ru^4}{\epsilon^4}\log\frac{1}{\beta}\right]\right).$$

Proof For all i we have

$$\begin{aligned}
\frac{(i+1)^{1-\alpha}}{1-\alpha} - 1 &= \int_1^{i+1} \frac{1}{j^\alpha} dj \\
&\leq \bar{Z}_{0,i} \\
&\leq \int_1^{i} \frac{1}{j^\alpha} dj + 1 \\
&\leq \frac{i^{1-\alpha}}{1-\alpha}.
\end{aligned}$$

From (32) and (33), we have the following sufficient conditions on i_0 and i to ensure that $U(i_0, i) \leq \beta$:

$$\frac{(i_0+1)^{1-\alpha}}{1-\alpha} - 1 \geq \max\left[\frac{387u^2 r}{\epsilon^2}\log\frac{594u^2}{\epsilon^2\beta}, \frac{4r}{3}\left(\frac{12c_\tau}{\epsilon(2-\bar{\tau})}\right)^{1/\bar{\tau}}\right]$$

and

$$i \geq \max\left[\left(\frac{i_0\epsilon}{12u} + \frac{i^{1-\alpha}}{1-\alpha}\right)\frac{72\sqrt{r}u^2}{\epsilon^2}, \frac{10368ru^4}{\epsilon^4}\log\frac{2}{\beta}\right].$$

We conclude that

$$i_0 = O\left(\frac{u^{\frac{2}{1-\alpha}} r^{\frac{1}{1-\alpha}}}{\epsilon^{\frac{2}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}\right).$$

Furthermore, a sufficient value of i is given by

$$i = \max\left(2\frac{i_0\epsilon}{12u}\frac{72\sqrt{r}u^2}{\epsilon^2}, 2\frac{i^{1-\alpha}}{1-\alpha}\frac{72\sqrt{r}u^2}{\epsilon^2}, \frac{10368ru^4}{\epsilon^4}\log\frac{2}{\beta}\right)$$

We conclude that the smallest number of phases that guarantees that $U(i_0, i) \leq \beta$ is on the order of

$$i = O\left(\max\left[\frac{u^{\frac{3-\alpha}{1-\alpha}}r^{\frac{3-\alpha}{2(1-\alpha)}}}{\epsilon^{\frac{3-\alpha}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}, \frac{u^{\frac{2}{\alpha}}r^{\frac{1}{2\alpha}}}{\epsilon^{\frac{2}{\alpha}}}, \frac{128ru^4}{\epsilon^4}\log\frac{1}{\beta}\right]\right).$$

We can use similar steps and bounds (30) and (31) to conclude that, in order to have $U(i_0, i) \leq \beta$, i must also satisfy

$$i = \Omega\left(\max\left[\frac{u^{\frac{3-\alpha}{1-\alpha}}r^{\frac{3-\alpha}{2(1-\alpha)}}}{\epsilon^{\frac{3-\alpha}{1-\alpha}}}\left(\log\frac{u^2}{\epsilon^2\beta}\right)^{\frac{1}{1-\alpha}}, \frac{u^{\frac{2}{\alpha}}r^{\frac{1}{2\alpha}}}{\epsilon^{\frac{2}{\alpha}}}\right]\right).$$

□

Theorem 13 *Suppose that*

$$p_j = \frac{\epsilon^2}{16\sqrt{r}u^2} \qquad (j = 1, 2, \ldots).$$

Then for all $\tau \geq 0.5$ the smallest number of phases i such that $\Pr\left(M(i) \leq \max_e \mu_e^\tau - \epsilon\right) \leq U(i_0, i) \leq \beta$ satisfies

$$i = O\left(\frac{r^2u^5}{\epsilon^5}\log\frac{u^2}{\epsilon^2\beta}\right).$$

Proof The result follows directly from (30) and (31). □

Appendix C. Proof of Results for Fixed Length Phases

Lemma 22 *For any expert e, let $i' > i$ and suppose that $M_e(i') \leq M_e(i) - \epsilon$. Then we have*

$$N_e(i') - N_e(i) \geq \frac{N_e(i)\epsilon}{u}.$$

Proof We have

$$\begin{aligned} M_e(i)N_e(i) &\leq M_e(i')N_e(i') \\ &\leq (M_e(i) - \epsilon)N_e(i'). \end{aligned}$$

It follows that

$$\begin{aligned} N_e(i') - N_e(i) &\geq \frac{N_e(i)\epsilon}{M_e(i) - \epsilon} \\ &\geq \frac{N_e(i)\epsilon}{u}. \end{aligned}$$

□

Theorem 14 *Suppose EEE is implemented with phases of fixed length L. Then for all i_0, i and ϵ such that $\bar{Z}_{i_0,i} \leq \frac{i\epsilon^2}{8u^2} - \frac{i_0\epsilon}{4u}$, we have*

$$\Pr\left(M(i) \leq \max_e \min_{i_0+1\leq j\leq i} M_e(j) - \epsilon\right) \leq \exp\left(-\frac{1}{2i}\left(\frac{i\epsilon^2}{8u^2} - \frac{i_0\epsilon}{4u} - \bar{Z}_{i_0,i}\right)^2\right).$$

Proof We follow the same steps as in the proof of Theorem 1. For simplicity, let

$$\gamma = \frac{\epsilon}{u}\ .$$

Recall the definitions of $\mathcal{E}_1$ and $\mathcal{E}_2$. Note that, for all $e \in \mathcal{E}_1 \cup \mathcal{E}_2$, we have

$$N_e(i) - N_e(j_e) = \sum_{j=j_e+1}^{i} Z_{e,j}\ .$$

Moreover, for all $e \in \mathcal{E}_2$, since $M_e(j_e) \geq V$ and $M_e(i) < v - \epsilon$, applying Lemma 22 we conclude that

$$N_e(j_e) \leq \frac{u}{\epsilon}\left[N_e(i) - N_e(j_e)\right] = \frac{u}{\epsilon}\sum_{j=j_e+1}^{i} Z_{e,j}. \tag{34}$$

Following the same steps used in Theorem 1, we conclude that

$$\Pr\left(\sum_e N_e(i)\delta(M_e(i) < V - \epsilon) \leq \gamma i\right) \leq \exp\left(-\frac{1}{2(i-i_0)}\left(\frac{\epsilon(\gamma i - i_0)}{u+\epsilon} - \bar{Z}_{i_0,i}\right)^2\right).$$

Now suppose that

$$\sum_e N_e(i)\delta(M_e(i) < V - \epsilon) \leq \gamma i.$$

Then we have

$$\begin{aligned} M(i) &= \frac{\sum_e N_e(i)M_e(i)}{\sum_e N_e(i)} \\ &\geq V - \epsilon - (V-\epsilon)\frac{\sum_e N_e(i)\delta(M_e(i) < V-\epsilon)}{\sum_e N_e(i)} \\ &\geq V - \epsilon - (V-\epsilon)\frac{\gamma i}{i} \\ &\geq V - 2\epsilon. \end{aligned}$$

We conclude that

$$\begin{aligned} \Pr\left(M(i) \leq V - 2\epsilon\right) &\leq \Pr\left(\sum_e N_e(i)\delta(M_e(i) < V-\epsilon) \leq \gamma i\right) \\ &\leq \exp\left(-\frac{1}{2(i-i_0)}\left(\frac{\epsilon(\gamma i - i_0)}{u+\epsilon} - \bar{Z}_{i_0,i}\right)^2\right) \\ &\leq \exp\left(-\frac{1}{2(i-i_0)}\left(\frac{\epsilon(\gamma i - i_0)}{2u} - \bar{Z}_{i_0,i}\right)^2\right). \end{aligned}$$

□

Theorem 15 *Suppose EEE is implemented with phases of fixed length L. Then for all $i_0 \leq i$ and $\epsilon > 0$, we have*

$$\mathrm{E}\left[M(i) - \max_e \min_{i_0+1 \leq j \leq i} M_e(i)\right] \geq -\epsilon - u\frac{i_0}{i} - \frac{2u^2}{\epsilon}\frac{\bar{Z}_{i_0,i}}{i}.$$

Proof We have

$$\begin{aligned}
\mathrm{E}\left[M(i) - \max_e \min_{i_0+1 \leq j \leq i} M_e(i)\right] &= \mathrm{E}\left[\frac{\sum_e N_e(i)M_e(i)}{\sum_e N_e(i)} - V\right] \\
&\geq \mathrm{E}\left[V - \epsilon - (V-\epsilon)\frac{\sum_e N_e(i)\delta(M_e(i) < V - \epsilon)}{i} - V\right] \\
&\geq -\epsilon - u\frac{\mathrm{E}\left[\sum_e N_e(i)\delta(M_e(i) < V - \epsilon)\right]}{i}. \qquad (35)
\end{aligned}$$

Recall the definitions of $\mathcal{E}_1$ (6) and $\mathcal{E}_2$ (7). Then we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_e N_e(i)\delta(M_e(i) < V - \epsilon)\right] \\
&\leq \mathrm{E}\left[\sum_{e\in\mathcal{E}_1}\left(N_e(i_0) + \sum_{j=i_0+1}^{i} Z_{e,j}\right) + \sum_{\mathcal{E}_2}\left(N_e(j_e) + \sum_{j=j_e+1}^{i} Z_{e,j}\right)\right] \\
&\leq i_0 + \mathrm{E}\left[\left(\frac{u}{\epsilon}+1\right)\sum_{j=i_0+1}^{i} Z_j\right] \\
&\leq i_0 + \frac{2u}{\epsilon}\bar{Z}_{i_0,i}. \qquad (36)
\end{aligned}$$

The theorem follows from (35) and (36). □

Theorem 16 *Suppose EEE is implemented with phases of fixed length L. For all $\epsilon > 0$, we have*

$$\Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \frac{c_\tau}{L^\tau} - \epsilon\right) \leq \frac{2L^2u^2}{\epsilon^2}\exp\left(-\frac{\epsilon^2\bar{Z}_{0,i}}{4L^2u^2r}\right).$$

Proof Let $I_1, I_2, \ldots$ denote the phase numbers when expert e is followed. Note that

$$M_e(I_k) = \frac{1}{k}\sum_{j=1}^{k}\tilde{R}_j.$$

Let $\hat{R}_k = \mathrm{E}\left[\tilde{R}_k | h_k\right]$. Let

$$S_e(I_k) = \sum_{j=1}^{k}(\tilde{R}_j - \hat{R}_j).$$

Note that $\mathrm{E}\left[\tilde{R}_j - \hat{R}_j | h_j\right] = 0$ and $|\tilde{R}_j - \hat{R}_j| \leq u$. It follows from Hoeffding's inequality (Hoeffding, 1963) that

$$\Pr\left(S_e(I_k) \leq -\epsilon k\right) \leq \exp\left(-\frac{2\epsilon^2}{u^2}\right).$$

Note that

$$\hat{R}_k \geq \mu_e^\tau - \frac{c_\tau}{L^\tau}.$$

We conclude that, for all k,

$$\begin{aligned}
\Pr\left(M_e(I_k) \leq \mu_e^\tau - \frac{c_\tau}{L^\tau} - \epsilon\right) &= \Pr\left(\sum_{j=1}^{k}\left[\tilde{R}_j - \left(\mu_e^\tau - \frac{c_\tau}{L^\tau}\right)\right] \leq -\epsilon k\right) \\
&\leq \Pr\left(S_e(I_k) \leq -\epsilon k\right) \\
&\leq \exp\left(-\frac{2\epsilon^2}{u^2}\right).
\end{aligned}$$

We now follow the same steps as in the proof of Theorem 6. Let

$$k = \left(1 - \frac{\epsilon}{2u}\right)\frac{\bar{Z}_{0,i}}{r}.$$

Note that

$$k \geq \frac{\bar{Z}_{0,i}}{2r}.$$

Now we have

$$\begin{aligned}
\Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \epsilon\right) &\leq \Pr\left(\inf_{j\geq i} M_e(j) < \mu_e^\tau - \epsilon, N_e(i) \geq k\right) + \Pr\left(N_e(i) < k\right) \\
&\leq \Pr\left(\inf_{j\geq k} M_e(I_j) < \mu_e^\tau - \epsilon\right) + \Pr\left(N_e(i) < k\right) \\
&\leq \sum_{j=k}^{\infty} \Pr\left(M_e(I_j) < \mu_e^\tau - \epsilon\right) + \Pr\left(N_e(i) < k\right) \\
&\leq \sum_{j=k}^{\infty} \exp\left(-\frac{j\epsilon^2}{2u^2}\right) + \exp\left[-\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r}\right] \\
&\leq \frac{2u^2}{\epsilon^2}\exp\left(-\frac{k\epsilon^2}{2u^2}\right) + \exp\left[-\left(1 - \frac{kr}{\bar{Z}_{0,i}}\right)^2 \frac{\bar{Z}_{0,i}}{2r}\right] \\
&\leq \frac{2u^2}{\epsilon^2}\exp\left(-\frac{\bar{Z}_{0,i_0}\epsilon^2}{4ru^2}\right) + \exp\left[-\frac{\bar{Z}_{0,i}\epsilon^2}{8ru^2}\right] \\
&\leq \frac{3u^2}{\epsilon^2}\exp\left(-\frac{\bar{Z}_{0,i_0}\epsilon^2}{8ru^2}\right).
\end{aligned}$$

The sixth inequality follows from $k = \left(1 - \frac{\epsilon}{2u}\right)\frac{\bar{Z}_{0,i}}{r} \geq \frac{\bar{Z}_{0,i}}{2r}$. □

References

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Electronic Colloquium on Computational Complexity*, 7(68), 2000.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

D. de Farias and N. Megiddo. How to combine expert (or novice) advice when actions impact the environment. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

W Feller. *Probability Theory and its Applications*. John Wiley & Sons, 1971.

D. Foster and R. Vohra. Regret and the on-line decision problem. *Games and Economic Behavior*, 29:7–35, 1999.

Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

D. Fudenberg and D.K. Levine. *The Theory of Learning in Games*. The MIT Press, Cambridge, MA, 1997.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, UK, 1991.