Active Learning in Multi-Armed Bandits

András Antos[1], Varun Grover[2], and Csaba Szepesvári[1,2]

[1] Computer and Automation Research Institute
of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary
antos@szit.bme.hu
[2] Department of Computing Science
University of Alberta, Edmonton T6G 2E8, Canada
{vgrover,szepesva}@cs.ualberta.ca

Abstract. In this paper we consider the problem of actively learning the mean values of distributions associated with a finite number of options (arms). The algorithms can select which option to generate the next sample from in order to produce estimates with equally good precision for all the distributions. When an algorithm uses sample means to estimate the unknown values then the optimal solution, assuming full knowledge of the distributions, is to sample each option proportional to its variance. In this paper we propose an incremental algorithm that asymptotically achieves the same loss as an optimal rule. We prove that the excess loss suffered by this algorithm, apart from logarithmic factors, scales as $n^{-3/2}$, which we conjecture to be the optimal rate. The performance of the algorithm is illustrated in a simple problem.

1 Introduction

Consider the problem of production quality assurance in a factory equipped with a number of machines that produce products of different quality. The quality can be monitored by inspecting the products produced: An inspection of a product is modeled as a random number say between zero and one, one meaning the best, zero the poorest quality. The outcome will depend on random effects influencing the production and how the inspection was done, but the main assumption is that the mean of this random variable characterizes the maintenance state of the machine. Due to the randomness of the inspection results, multiple measurements are necessary to control the precision of the quality estimates. We are interested in keeping the precision of the estimates equal across the machines. If the inspection of a product is expensive (as is the case when inspection requires the destruction of the product) then to keep the cost low, it is logical to inspect machines that produce products of highly varying inspection results more frequently. The problem is then to decide about exactly how frequently the quality of each machine should be checked by inspecting a product produced on it. The loss is measured by taking the largest of the mean-squared errors of the estimates produced for the machines.

The basic problem is to estimate unknown quantities corresponding to a finite number of options by sequentially drawing random variables from distributions associated with the options so as to keep the estimation error across

all the options the same. Active learning problems involve estimating unknown parameters by selectively and adaptively sampling from the input space. Hence, this problem can be seen as an instance of *active learning*. The problem is also similar to *multi-armed bandit problems* [7, 2] in that only one option (arm) can be probed at any time. However, the performance criterion is different from that used in bandits where the observed values are treated as rewards and performance during learning is what matters. Nevertheless, we will see that the exploration-exploitation dilemma which characterizes classical bandit problems will still play a role here. Because of this connection we call this problem the *max-loss value-estimation problem in multi-armed bandits*.

The formal description of this problem is as follows: We are interested in estimating the expected values (μ_k) of some distributions ($\mathcal{D}_k$), each associated with an option (or arm). If K is the number of options then $1 \leq k \leq K$. For any k, the decision maker can draw independent samples $\{X_{kt}\}_t$ from $\mathcal{D}_k$. The sample X_{kt} is observed when a sample is requested from option k the t^{th} time. (These samples correspond to the outcomes of inspections in the previous example). The samples are drawn sequentially: Given the information collected up to trial n the decision maker can decide which option to choose next. At any time n, the decision maker keeps an estimate, $\hat{\mu}_{kn}$, of the mean of $\mathcal{D}_k$. The error of estimate k is measured with the expected squared error:

$$L_{kn} = \mathbb{E}\left[(\hat{\mu}_{kn} - \mu_k)^2\right].$$

The overall loss is measured by the worst-case loss over the K options:

$$L_n = \max_{1 \leq k \leq K} L_{kn}.$$

This expresses the desire that all estimates are equally important. The goal of the decision maker is to make this loss as small as possible.

For the sake of simplicity assume that the estimates $\hat{\mu}_{kn}$ are produced by computing the sample means of the respective options:

$$\hat{\mu}_{kn} = \frac{1}{T_{kn}} \sum_{t=1}^{T_{kn}} X_{kt},$$

where T_{kn} denotes the number of times a sample was requested from option k.

Consider the non-sequential version of the problem, i.e., the problem of choosing $T_{1n}, \ldots, T_{Kn}$ such that $T_{1n} + \ldots + T_{Kn} = n$ so as to minimize the loss. Let us assume for a moment full knowledge of the distributions, so there is no value in making this choice data dependent. Due to the independence of samples

$$L_{kn} = \frac{\sigma_k^2}{T_{kn}},$$

where $\sigma_k^2 = \mathrm{Var}\left[X_{k1}\right]$. For simplicity assume that $\sigma_k^2 > 0$ holds for all k. It is not hard to see then that the minimizer of $L_n = \max_k L_{kn}$ is the allocation $\{T_{kn}^*\}_k$

that makes all the losses L_{kn} (approximately) equal, hence (apart from rounding issues)

$$T^*_{kn} = n\frac{\sigma_k^2}{\Sigma^2} = \lambda_k n.$$

Here $\Sigma^2 = \sum_{j=1}^K \sigma_j^2$ is the sum of the variances and

$$\lambda_k = \frac{\sigma_k^2}{\Sigma^2}.$$

The corresponding loss is

$$L^*_n = \frac{\Sigma^2}{n}.$$

The optimal allocation is easy to extend to the case when some options have zero variance. Clearly, it is both necessary and sufficient to make a single observation on such options. The case when all variances are zero (i.e., $\Sigma^2 = 0$) is uninteresting, hence we will assume from now on that $\Sigma^2 > 0$.

We expect a good sequential algorithm $\mathcal{A}$ to achieve a loss $L_n = L_n(\mathcal{A})$ close to the loss L^*_n. We will therefore look into the excess loss

$$R_n(\mathcal{A}) = L_n(\mathcal{A}) - L^*_n.$$

Since the loss of option k can only decrease if we request a new sample from $\mathcal{D}_k$, one simple idea is to request the next sample from option k whose estimated loss, $\hat{\sigma}^2_{kn}/T_{kn}$, is the largest amongst all estimated losses. Here $\hat{\sigma}^2_{kn}$ is an estimate of the variance of the k^{th} option based on the history. The problem with this approach is that the variance might be underestimated in which case the option will not be selected for a long time, which prevents refining the estimated variance, ultimately resulting in a large excess loss. Thus we face a problem similar to the exploration-exploitation dilemma in bandit problems where a greedy policy might incur a large loss if the payoff of the optimal option is underestimated. One simple remedy is to make sure that the estimated variances converge to their true values. This can be ensured if the algorithm is forced to select all the options indefinitely in the limit, which is often called the method of forced selections in the bandit literature. One way to implement this idea is to introduce phases of increasing length. Then in each phase the algorithm could choose all options exactly once at the beginning, while in the rest of the phase it can sample all the options k proportionally to their respective variance estimates computed at the beginning of the phase. The problem then becomes to select the appropriate phase lengths to make sure that the proportion of forced selections diminishes at an appropriate rate with an increasing horizon n. (An algorithm along these lines have been described and analyzed by [5] in the context of stratified sampling. We shall discuss this further in Section 5.) While the introduction of phases allows a direct control of the proportion of forced selections, the algorithm is not incremental and is somewhat cumbersome to implement.

In this paper we propose and study an alternative algorithm that implements forced selections but remains completely incremental. The idea is to select the

option with the largest estimated loss except if some of the options is seriously under-sampled, in which case the under-sampled option is selected. It turns out that a good definition for an option being under-sampled is $T_{kn} \le c\sqrt{n}$ with some constant $c > 0$. (The algorithm will be formally stated in the next section.) We will show that the excess loss of this algorithm decreases with n as $\tilde{O}(n^{-3/2})$.[3]

2 Algorithm

The formal description of the algorithm, that we call GAFS-MAX (greedy allocation with forced selections for max-norm value estimation), is as follows:

Algorithm GAFS-MAX
In the first K trials choose each arm once
Set $T_{k,K+1} = 1$ $(1 \le k \le K)$, $n = K + 1$
At time n do:
Compute $\hat{\sigma}^2_{kn} = \frac{1}{T_{kn}} \sum_{t=1}^{T_{kn}} X_{kt}^2 - \left(\frac{1}{T_{kn}} \sum_{t=1}^{T_{kn}} X_{kt} \right)^2$
Let $\hat{\lambda}_{kn} = \hat{\sigma}^2_{kn} / (\sum_{j=1}^{K} \hat{\sigma}^2_{jn})$ if $\sum_{j=1}^{K} \hat{\sigma}^2_{jn} \neq 0$,
otherwise let $\hat{\lambda}_{kn} = 1/K$.
Let $U_n = \{ k : T_{kn} < \sqrt{n} + 1 \}$
Let

$$I_{n+1} = \begin{cases} \min U_n, & \text{if } U_n \neq \emptyset \\ \operatorname{argmax}_{1 \le k \le K} \frac{\hat{\lambda}_{kn}}{T_{kn}}, & \text{otherwise,} \end{cases}$$

where in the second case ties are broken in an arbitrary, but systematic manner.
Choose option I_{n+1}, let $T_{k,n+1} = T_{k,n} + \mathbb{I}\{ I_{n+1} = k \}$
Observe the feedback $X_{I_{n+1}, T_{I_{n+1},n+1}}$.

Of course, the variance estimates can be computed incrementally. Further, it is actually not necessary to compute $\hat{\lambda}_{kn}$ because in the computation of the arm index $\hat{\lambda}_{kn}$ can be replaced by $\hat{\sigma}^2_{kn}$ without effecting the choices.

3 Main Results

The main result (Theorem 3) for GAFS-MAX is a bound of the form $L_n \le L_n^* + \tilde{O}(n^{-3/2})$. We also prove high probability bounds on $T_{nk}/n - \lambda_k$ (Theorem 1). The proof is somewhat involved, hence we start with an outline: Clearly, the rate of growth of T_{kn} controls the rate of convergence of $\hat{\lambda}_{kn}$ to λ_k. In particular, we will show that given $T_{kn} \ge f(n)$ it follows that $\hat{\lambda}_{kn}$ converges to λ_k at a rate of $O(1/f(n)^{1/2})$ (Lemma 2). The second major tool is a result (Lemma 3) that shows how a faster rate for $\hat{\lambda}_{kn}$ transforms into better bounds on T_{kn}. The actual proof is then started by observing that due to the forced selections $T_{kn} \ge \sqrt{n}$.

[3] A nonnegative sequence (a_n) is said to be $\tilde{O}(f(n))$, where $f : \mathbb{N} \to \mathbb{R}^+$, if $a_n \le C f(n) \log(n)$ with a suitable constant $C > 0$.

The proof is developed through a series of Lemmata. First, we state Hoeffding's inequality in a form that suits the best our needs:

Lemma 1 (Hoeffding's inequality, [6]). *Let Z_t be a sequence of zero-mean, i.i.d. random variables, where $a \leq Z_t \leq b$, $a < b$ reals. Then, for any $0 < \delta \leq 1$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t \geq \sqrt{\frac{1}{2}\frac{(b-a)^2}{n}\log(1/\delta)}\right) \leq \delta.$$

Let

$$\Delta(R^2, n, \delta) = R\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Let $\mu_k^{(2)} = \mathbb{E}\left[X_{kt}^2\right]$, R_k be the size of the range of the random variables $\{X_{kt}\}_t$ (i.e., $|\mathrm{supp}(X_{kt})| \leq R_k$), S_k be the size of the range of the random variables $\{X_{kt}^2\}_t$, and B_k be the size of the range of the random variables $\{|X_{kt}|\}_t$. Note that $B_k \leq R_k$ and $S_k \leq B_k^2$. Let

$$A_\delta = \left(\bigcap_{1\leq k\leq K, n\geq 1}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} X_{kt}^2 - \mu_k^{(2)}\right| \leq \Delta(S_k^2, n, \delta_n)\right\}\right)\bigcap$$
$$\bigcap_{1\leq k\leq K, n\geq 1}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} X_{kt} - \mu_k\right| \leq \Delta(R_k^2, n, \delta_n)\right\},$$

where $\delta_n = \delta/(4K(n(n+1))$. Note that δ_n is chosen such that $\sum_{k=1}^{K}\sum_{n=1}^{\infty}\delta_n = \delta/4$. Hence, we observe that by Hoeffding's inequality

$$\mathbb{P}(A_\delta) \geq 1 - \delta.$$

The sets $\{A_\delta\}_\delta$ will play a key role in the proof: Many of the statements will be proved on these set.

Our first result connects a lower bound on T_{kn} to the rate of convergence of $\hat{\lambda}_{kn}$. Let $a_k = |\mu_k| + B_k$, $b_k = S_k + a_k R_k$, and $a_k' = \sigma_k^4/(4b_k^2)$.

Lemma 2. *Fix $0 < \delta \leq 1$ and $n_0 > 0$, and assume that for $n \geq n_0$, $1 \leq k \leq K$, $T_{kn} \geq f(n) \geq 2$ holds on A_δ, where $f(n) \to \infty$. Then there exists constants $N_0 \geq n_0$ and $c > 0$ such that for any $n \geq N_0$, $1 \leq k \leq K$, on A_δ*

$$\left|\hat{\lambda}_{kn} - \lambda_k\right| \leq c\sqrt{\frac{\log(\delta_n^{-1})}{f(n)}} \tag{1}$$

holds. In particular, $c = \sqrt{2}(b_k + \lambda_k\sum_{j=1}^{K} b_j)/\Sigma^2 \leq 5\sqrt{2}(B_k^2 + \sum_{j=1}^{K} B_j^2)/\Sigma^2$.

If $f(n) = bn^p$ ($p > 0$) then $N_0 = \max(n_0, n_1)$, where n_1 is a number such that for $n \geq n_1$

$$\log n \leq \frac{ba_k'}{p}\, n^p - \frac{1 + \log\left(\frac{4K}{\delta}\right) + 2\log b}{2p}. \tag{2}$$

Proof. First, we develop a bound on $|\hat{\sigma}_{kn}^2 - \sigma_k^2|$. Let $\hat{\mu}_{kn}^{(2)} = 1/T_{kn} \sum_{t=1}^{T_{kn}} X_{kt}^2$ and $\hat{\mu}_{kn} = 1/T_{kn} \sum_{t=1}^{T_{kn}} X_{kt}$. Consider any element of A_δ. Then by the definition of A_δ, $|1/m \sum_{t=1}^{m} X_{kt}^2 - \mu_k^{(2)}| \leq \Delta(S_k^2, m, \delta_m)$ holds simultaneously for any $m \geq 1$. Hence, for $n \geq n_0$ it also holds that

$$\left| \frac{1}{T_{kn}} \sum_{t=1}^{T_{kn}} X_{kt}^2 - \mu_k^{(2)} \right| \leq \Delta(S_k^2, T_{kn}, \delta_{T_{kn}}) \leq \Delta(S_k^2, f(n), \delta_{f(n)}),$$

where we have used that $\log(x(x+1)/\delta)/x$ is monotonically decreasing when $x \geq 2$ and $T_{kn} \geq f(n) \geq 2$. Similarly, we get that

$$\left| \frac{1}{T_{kn}} \sum_{t=1}^{T_{kn}} X_{kt} - \mu_k \right| \leq \Delta(R_k^2, f(n), \delta_{f(n)}).$$

Using $\hat{\sigma}_{kn}^2 = \hat{\mu}_{kn}^{(2)} - \hat{\mu}_{kn}^2$ and $\sigma_k^2 = \mathbb{E}\left[X_{kt}^2\right] - (\mathbb{E}\left[X_{kt}\right])^2 = \mu_k^{(2)} - \mu_k^2$, we get

$$\begin{aligned} \left|\hat{\sigma}_{kn}^2 - \sigma_k^2\right| &\leq \left|\hat{\mu}_{kn}^{(2)} - \mu_k^{(2)}\right| + \left|\hat{\mu}_{kn}^2 - \mu_k^2\right| \\ &\leq \Delta(S_k^2, f(n), \delta_{f(n)}) + \Delta(R_k^2, f(n), \delta_{f(n)})(|\mu_k| + B_k), \end{aligned} \tag{3}$$

where we used $|a^2 - b^2| \leq |a - b|\,(|a| + |b|)$.

Denote the right-hand side of (3) by $\Delta_k(n, \delta)$. Now, let us develop a lower bound on $\hat{\lambda}_{kn}$ in terms of λ_k. Then, for $n \geq n_0$,

$$\hat{\lambda}_{kn} = \frac{\hat{\sigma}_{kn}^2}{\sum_{j=1}^{K} \hat{\sigma}_{jn}^2} \geq \frac{\sigma_k^2 - \Delta_k(n, \delta)}{\Sigma^2 + \sum_{j=1}^{K} \Delta_j(n, \delta)} \geq \lambda_k \left(1 - \frac{\sum_{j=1}^{K} \Delta_j(n, \delta)}{\Sigma^2} \right) - \frac{\Delta_k(n, \delta)}{\Sigma^2},$$

where we used $1/(1+x) \geq 1 - x$ that holds for $x > -1$.

An upper bound can be obtained analogously: For $n \geq n_0$, if

$$\Sigma^2 \geq 2 \sum_{j=1}^{K} \Delta_j(n, \delta) \tag{4}$$

then

$$\hat{\lambda}_{kn} = \frac{\hat{\sigma}_{kn}^2}{\sum_{j=1}^{K} \hat{\sigma}_{jn}^2} \leq \frac{\sigma_k^2 + \Delta_k(n, \delta)}{\Sigma^2 - \sum_{j=1}^{K} \Delta_j(n, \delta)} \leq \lambda_k \left(1 + 2\frac{\sum_{j=1}^{K} \Delta_j(n, \delta)}{\Sigma^2} \right) + 2\frac{\Delta_k(n, \delta)}{\Sigma^2},$$

where we used $1/(1-x) \leq 1 + 2x$ that holds for $0 \leq x \leq 1/2$. This constraint follows from (4), that is implied if n is big enough so that

$$2\Delta_j(n, \delta) \leq \sigma_j^2, \quad 1 \leq j \leq K. \tag{5}$$

The upper and lower bounds above together give

$$|\hat{\lambda}_{kn} - \lambda_k| \le \frac{2}{\Sigma^2} \left(\lambda_k \sum_{j=1}^{K} \Delta_j(n,\delta) + \Delta_k(n,\delta) \right).$$

Noting that $\Delta_j(n,\delta)$ equals to

$$(S_j + R_j(|\mu_j| + B_j)) \sqrt{\frac{\log(\delta_{f(n)}^{-1})}{2f(n)}} = b_j \sqrt{\frac{\log(\delta_{f(n)}^{-1})}{2f(n)}}, \tag{6}$$

where $b_j = S_j + R_j(|\mu_j| + B_j)$, we get

$$|\hat{\lambda}_{kn} - \lambda_k| \le \frac{\sqrt{2}}{\Sigma^2} \left(\lambda_k \sum_{j=1}^{K} b_j + b_k \right) \sqrt{\frac{\log(\delta_{f(n)}^{-1})}{f(n)}}.$$

Since $f(n) \le T_{kn} \le n$, $\delta_{f(n)}^{-1}$ can be upper bounded by δ_n^{-1} leading to (1).

At last, to satisfy (5), by (6), it suffices if

$$f(n) \ge \frac{2b_j^2}{\sigma_j^4} \log(\delta_{f(n)}^{-1}) = \frac{2b_j^2}{\sigma_j^4} (\log(f(n)(f(n)+1)) + \log(4K/\delta))$$

that is guaranteed by $f(n) \to \infty$ for n large enough.

If $f(n) = bn^p$ then $bn^p \ge \frac{2b_j^2}{\sigma_j^4}(2p\log n + 2\log b + 1 + \log(4K/\delta))$ will ensure that. Reordering this gives (2). □

Now we show how a rate of convergence result for $\hat{\lambda}_{kn}$ can be turned into bounds on $T_{kn}/n - \lambda_k$. Let $\lambda_{\min} = \min_{1 \le j \le K} \lambda_j$. In what follows, unless otherwise stated, we will assume that $\lambda_{\min} > 0$.

Lemma 3. *Fix $0 < \delta \le 1$ and $n_0 > 0$. Assume that $f(n) \le n$ such that $f(n)/n^2$ is monotone decreasing, and consider an event such that*

$$|\hat{\lambda}_{kn} - \lambda_k| \le c\sqrt{\log(\delta_n^{-1})/f(n)}, \qquad 1 \le k \le K \tag{7}$$

holds with some $c \ge 1$, for all $n \ge n_0$. Let

$$H(n,\delta) = c \left(1 + \frac{2}{\lambda_{\min}} \right) n \sqrt{\frac{\log(\delta_n^{-1})}{f(n)}}.$$

Then the following inequalities hold for $n \ge n_0$ and $1 \le k \le K$:

$$T_{kn} \le n\lambda_k + \max(n_0, 1 + H(n,\delta)),$$
$$T_{kn} \ge n\lambda_k - (K-1)\max(n_0, 1 + H(n,\delta)).$$

Proof. By definition $T_{k,n+1} = T_{kn} + \mathbb{I}\{ I_{n+1} = k \}$. Let $E_{kn} = T_{kn} - n\lambda_k$. Note that

$$\sum_{k=1}^{K} E_{kn} = 0 \tag{8}$$

holds for any $n \geq 1$. Notice that the desired result can be stated as bounds on E_{kn}. Hence, our goal now is to study E_{kn}. If b_{jn} is an upper bound for E_{jn} $(1 \leq j \leq K)$ then from (8) we get the lower bound $E_{kn} = -\sum_{j \neq k} E_{jn} \geq -\sum_{j \neq k} b_{jn} \geq -(K-1)\max_j b_{jn}$. Hence, we target upper bounds on $\{E_{kn}\}_k$.

From the definition of E_{kn} and T_{kn} we get

$$E_{k,n+1} = E_{k,n} - \lambda_k + \mathbb{I}\{ I_{n+1} = k \}.$$

By the definition of the algorithm

$$\mathbb{I}\{ I_{n+1} = k \} \leq \mathbb{I}\left\{ T_{kn} \leq \lceil \sqrt{n} \rceil \text{ or } k = \operatorname*{argmin}_{1 \leq j \leq K} \frac{T_{jn}}{\hat{\lambda}_{jn}} \right\},$$

with the understanding that $c/0 = +\infty$. Assume now that k is an index where $\{\frac{T_{jn}}{\hat{\lambda}_{jn}}\}_j$ takes its minimum, that is,

$$\frac{T_{kn}}{\hat{\lambda}_{kn}} \leq \min_j \frac{T_{jn}}{\hat{\lambda}_{jn}}.$$

Using $T_{jn} = E_{jn} + n\lambda_j$ and reordering the terms gives

$$E_{kn} + n\lambda_k \leq \hat{\lambda}_{kn} \min_j \frac{E_{jn} + n\lambda_j}{\hat{\lambda}_{jn}} \leq \hat{\lambda}_{kn} \left(\min_j \frac{E_{jn}}{\hat{\lambda}_{jn}} + n \max_j \frac{\lambda_j}{\hat{\lambda}_{jn}} \right).$$

By (8), there exists an index j such that $E_{jn} \leq 0$. Since $\hat{\lambda}_{jn} \geq 0$ for any j, it holds that $\min_j \frac{E_{jn}}{\hat{\lambda}_{jn}} \leq 0$. Hence, $E_{kn} + n\lambda_k \leq n\hat{\lambda}_{kn} \max_j \frac{\lambda_j}{\hat{\lambda}_{jn}}$. Using (7) and $1/(1-x) = 1 + x/(1-x) \leq 1 + 2x$, which holds for $x \leq 1/2$, provided that $n \geq n_0$, we get

$$\frac{\lambda_j}{\hat{\lambda}_{jn}} \leq \frac{\lambda_j}{\lambda_j - c\sqrt{\log(\delta_n^{-1})/f(n)}} \leq 1 + \frac{2c}{\lambda_j} \sqrt{\frac{\log(\delta_n^{-1})}{f(n)}}.$$

Using $\hat{\lambda}_{kn} \leq 1$ and (7) again,

$$E_{kn} \leq n(\hat{\lambda}_{kn} - \lambda_k) + \frac{2cn}{\lambda_{\min}} \sqrt{\frac{\log(\delta_n^{-1})}{f(n)}} \leq c \left(1 + \frac{2}{\lambda_{\min}} \right) n \sqrt{\frac{\log(\delta_n^{-1})}{f(n)}}.$$

Note that the right-hand side is $H(n, \delta)$. Hence,

$$\mathbb{I}\{ I_{n+1} = k \} \leq \mathbb{I}\left\{ T_{kn} \leq \lceil \sqrt{n} \rceil \text{ or } E_{kn} \leq H(n, \delta) \right\}.$$

Assume now that $T_{kn} \leq \lceil \sqrt{n} \rceil$. We want to show that in this case $E_{kn} \leq H(n,\delta)$. By the definition of E_{kn}, from $T_{kn} \leq \lceil \sqrt{n} \rceil$ it follows that $E_{kn} = T_{kn} - n\lambda_k \leq \lceil \sqrt{n} \rceil \leq \sqrt{2n}$. Hence, $E_{kn} \leq H(n,\delta)$ follows if $\sqrt{2n} \leq H(n,\delta)$. In particular, this follows from the bounds on c, $\lambda_{\min}$, $f(n)$, and δ. Therefore

$$\mathbb{I}\{\, I_{n+1} = k \,\} \leq \mathbb{I}\{\, E_{kn} \leq H(n,\delta) \,\}.$$

We need the following technical lemma:

Lemma 4. *Let $0 \leq \lambda \leq 1$. Consider the sequences $E_n, \tilde{E}_n, I_n, \tilde{I}_n$ ($n \geq 1$) where $I_n, \tilde{I}_n \in 0,1$, $E_{n+1} = E_n + I_n - \lambda$, $\tilde{E}_{n+1} = \tilde{E}_n + \tilde{I}_n - \lambda$, $\tilde{E}_1 = E_1$ and assume that $I_n \leq \tilde{I}_n$ holds whenever $E_n = \tilde{E}_n$. Then $E_n \leq \tilde{E}_n$ holds for $n \geq 1$.*

Due to the lack of space we only sketch the proof of Lemma 4. The idea is that $P_n = \tilde{E}_n - E_n$ can only take on integer values and step 0 or 1. Then $P_n \geq 0$, $n \geq 1$ follows since $P_1 = 0$ and when in $P_n = 0$ then $P_{n+1} \geq 0$.

Now, returning to the proof of Lemma 3, define $\tilde{E}_{kn}$ by

$$\begin{aligned} \tilde{E}_{k,n+1} &= \tilde{E}_{k,n} - \lambda_k + \mathbb{I}\left\{\, \tilde{E}_{kn} \leq H(n,\delta) \,\right\}, \quad n \geq n_0, \\ \tilde{E}_{k,n_0} &= E_{k,n_0}. \end{aligned}$$

The conditions of Lemma 4 are clearly satisfied from index n_0. Consequently $E_{k,n} \leq \tilde{E}_{k,n}$ holds for any $n \geq n_0$. Further, since $H(n,\delta)$ is monotone increasing in n, $\tilde{E}_{k,n} \leq \max(E_{k,n_0}, 1 + H(n,\delta)) \leq \max(n_0, 1 + H(n,\delta))$, finishing the upper-bound. □

Using the previous result we are now in the position to prove a linear lower bound on T_{kn}:

Lemma 5. *Let $0 < \delta \leq 1$ arbitrary. Then there exists an integer N_1 such that for any $n \geq N_1$, $T_{kn} \geq n\lambda_k/2$ holds on A_δ.*

In particular,

$$N_1 = \max\left(\frac{2(K-1)}{\lambda_{\min}} \max(3, N_0), D_2^2 \left[\log D_2^2 + \frac{1}{2} \left(\log\left(\frac{4K}{\delta} \right) + 1 \right) \right]^2 \right), \quad (9)$$

where $N_0 = \max\left(K^2, (1/a'_k)^2 \left[\log((1/a'_k)^2) + (1 + \log(4K/\delta)) \right]^2 \right)$ and $D_2 = 4(9c(K-1))^2/\lambda_{\min}^4$.

For the proof we need the following technical lemma that quantifies the point when for $a > 0$ the function $at^{1/2} + b$ overtakes $\log t$.

Lemma 6. *Let $q(t) = at^{1/2} + b$, $\ell(t) = \log t$, where $a > 0$. Then for any $t \geq (2/a)^2 \left[\log((2/a)^2) - b \right]^2$, $q(t) \geq \ell(t)$.*

The proof of this lemma is elementary and is hence omitted.

Proof (Lemma 5). Due to the forced selection of the options built into the algorithm, $T_{kn} \geq \sqrt{n}$ holds for $n \geq K^2$. Hence, we can apply Lemma 2 with $f(n) = n^{1/2}$. By Lemma 6, n_1 defined by (2) can be chosen to be

$$(1/a'_k)^2 \left[\log((1/a'_k)^2) + (1 + \log(4K/\delta))\right]^2 .$$

Hence, for $n \geq N_0 = \max(K^2, n_1)$ and $c > 0$ as defined in Lemma 2, we get,

$$\left|\hat{\lambda}_{kn} - \lambda_k\right| \leq c\sqrt{\frac{\log(\delta_n^{-1})}{n^{1/2}}}. \tag{10}$$

Possibly replacing c with $\max(c, 1)$, we can assume that $c \geq 1$. By Lemma 3, for $n \geq \max(N_0, 1/\lambda_{\min})$, $T_{kn} \geq n\lambda_k - (K-1)\max(N_0, 1 + H(n,\delta))$, and $H(n,\delta) = D_1 n^{3/4}\sqrt{\log(\delta_n^{-1})}$, where $D_1 = c\left(1 + \frac{2}{\lambda_{\min}}\right) \leq 3c/\lambda_{\min}$. Hence, $T_{kn} \geq n\lambda_k/2$ by the time when $n \geq 2N_0(K-1)/\lambda_{\min}$ and $n \geq 2(K-1)(1 + H(n,\delta))/\lambda_{\min}$. Lemma 6 and some tedious calculations then show that these two constrained are satisfied when $n \geq N_1$, where N_1 is defined as in equation (9). □

With the help of this result we can get better bounds on T_{kn}, resulting in our first main result:

Theorem 1. *Let $0 < \delta \leq 1$ be arbitrary. Then there exists an integer N_2 and a positive real number D_3 such that for any $n \geq N_2$,*

$$-(K-1)\frac{\max(N_2, 1 + G(n,\delta))}{n} \leq \frac{T_{kn}}{n} - \lambda_k \leq \frac{\max(N_2, 1 + G(n,\delta))}{n}$$

holds on A_δ, where

$$G(n,\delta) = D_3\sqrt{n\log(\delta_n^{-1})}. \tag{11}$$

Here $D_3 \leq 3\sqrt{2}\, c/\lambda_{\min}^{3/2}$,

$$N_2 = \max\left(N_1, \left(\frac{4}{\lambda_{\min} a'_k}\right)\left[\log\left(\frac{2}{\lambda_{\min} a'_k}\right) + \frac{1}{2} + \frac{1}{2}\log\left(\frac{4K}{\delta}\right)\right]\right),$$

where N_1 is defined in Lemma 5.

The theorem shows that asymptotically the GAFS-MAX algorithm behaves the same way as an optimal allocation rule that knows the variances. It also shows that the deviation of the proportion of choices of any option from the optimal value decays as $\tilde{O}(1/\sqrt{n})$.

For the proof we need the counterpart of Lemma 6 for linear functions:

Lemma 7. *Let $q(t) = at + b$, $\ell(t) = \log t$, where $a > 0$. Then for any $t \geq (2/a)(\log((1/a)) - b)$, $q(t) \geq \ell(t)$.*

Proof (Theorem 1). The proof is almost identical to that of Lemma 5. The difference is that now we start with a better lower bound on T_{kn}. In particular, by Lemma 5 $T_{kn} \geq n\lambda_k/2$ holds whenever $n \geq N_1$. By Lemma 2, for some $N_2 \geq N_1$, $c \geq 1$,

$$\left|\hat{\lambda}_{kn} - \lambda_k\right| \leq \frac{c}{\lambda_k^{1/2}} \sqrt{\frac{\log(\delta_n^{-1})}{n}} \tag{12}$$

holds for all $n \geq N_2$. In particular, solving (2) for n_1 with $f(n) = n\lambda_k/2$ and Lemma 7 give that

$$N_2 = \max\left(N_1, \frac{4}{\lambda_{\min} a'_k}\left[\log\left(\frac{2}{\lambda_{\min} a'_k}\right) + \frac{1}{2} + \frac{1}{2}\log\left(\frac{4K}{\delta}\right)\right]\right)$$

will suffice. By Lemma 3, for $n \geq \max(N_2, \lambda_{\min}^{-1}) = N_2$,

$$\begin{aligned} T_{kn} &\leq n\lambda_k + \max(N_2, 1 + G(n,\delta)), \quad \text{and} \\ T_{kn} &\geq n\lambda_k - (K-1)\max(N_2, 1 + G(n,\delta)), \end{aligned}$$

where $G(n,\delta)$ is given by (11), and $D_3 = \sqrt{\frac{c}{\lambda_{\min}}}\, c\left(1 + \frac{2}{\lambda_{\min}}\right)$. □

This result yields a bound on the expected value of $\mathbb{E}[T_{kn}]$:

Theorem 2. *Let N_2' be such that $N_2 \leq N_2' \log^2(4K/\delta)$ holds for any $\delta > 0$, where N_2 is defined in Theorem 1. Then, there exists and index N_3 that depends only on N_2', D_3 and K, such for any $n \geq N_3$,*

$$\mathbb{E}[T_{kn}] \leq n\lambda_k + D_3\sqrt{n(1 + \log(4Kn(n+1)))} + 2. \tag{13}$$

Proof. First note that N_2' exists and $N_2 \leq N_2' \log^2(\delta_n^{-1})$ holds for any $n \geq 2$. Fix $0 < \delta \leq 1$. If $n \geq N_2^2/(D_3^2 \log(\delta_n^{-1}))$, then $1 + G(n,\delta) \geq N_2$, thus it follows from Theorem 1 that for $n \geq \max(N_2, N_2^2/(D_3^2\log(\delta_n^{-1})))$,

$$\mathbb{P}\left(\frac{T_{kn} - n\lambda_k - 1}{D_3 n^{1/2}} > \sqrt{\log(\delta_n^{-1})}\right) \leq \delta$$

where we used $\mathbb{P}(A_\delta) \geq 1 - \delta$. Let $Z = (T_{kn} - n\lambda_k - 1)/(D_3 n^{1/2})$ and $\epsilon = \sqrt{\log(\delta_n^{-1})}$. The above inequality is equivalent to

$$\mathbb{P}(Z > \varepsilon) \leq 4Kn(n+1)\, e^{-\varepsilon^2}.$$

By the constraints that connect n and δ, this inequality holds for any pair (n, ε) that satisfy

$$n \geq \max(N_2' \log^2(\delta_n^{-1}), {N_2'}^2 \log^3(\delta_n^{-1})/D_3^2) = \max(N_2'\varepsilon^4, {N_2'}^2\varepsilon^6/D_3^2),$$

that is, for any (n, ε) such that

$$\varepsilon \leq \min((n/N_2')^{1/4}, (nD_3^2/{N_2'}^2)^{1/6}).$$

Also, since $Z \leq n^{1/2}/D_3$ is always true, $\mathbb{P}(Z > \varepsilon) = 0$ holds for $\varepsilon \geq n^{1/2}/D_3$. We need the following technical lemma, a variant of which can be found, e.g., as Exercise 12.1 in [4]:

Lemma 8. *If* $\mathbb{P}(Z > \varepsilon) \le C \exp(-c\varepsilon^2)$ *for any* $\varepsilon \le a$, $a > 0$, *and* $\mathbb{P}(Z > \varepsilon) = 0$ *for any* $\varepsilon \ge b\,(\ge a)$, *then*

$$\mathbb{E}[Z] \le \sqrt{(1 + \log C)/c + Cb^2 e^{-ca^2}}. \tag{14}$$

Due to the lack of space the proof is omitted.

Applying Lemma 8 with $a = \min((n/N_2')^{1/4}, (nD_3^2/{N_2'}^2)^{1/6})$, and $b = n^{1/2}/D_3$, $C = 4Kn(n+1)$, $c = 1$,

$$\mathbb{E}[Z] \le \sqrt{1 + \log(4Kn(n+1)) + 4Kn^2(n+1)e^{-\min((n/N_2')^{1/2}, (nD_3^2/{N_2'}^2)^{1/3})}/D_3^2}.$$

Equation (13) then follows by straightforward algebra.

In order to develop a bound on the loss $L_{n,k}$ we need Wald's (second) identity:

Lemma 9 (Wald's Identity, Theorem 13.2.14 of [1]). *Let* $\{\mathcal{F}_t\}_t$ *be a filtration and let* Y_t *be an* $\mathcal{F}_t$*-adapted sequence of i.i.d. random variables. Assume that* $\mathcal{F}_t$ *and* $\sigma(\{Y_s : s \ge t+1\})$ *are independent and* T *is a stopping time w.r.t.* $\mathcal{F}_t$ *with a finite expected value:* $\mathbb{E}[T] < +\infty$. *Consider the partial sums* $S_n = Y_1 + \ldots + Y_n$, $n \ge 1$. *If* $\mathbb{E}[Y_1^2] < +\infty$ *then*

$$\mathbb{E}\left[(S_T - T\mathbb{E}[Y_1])^2\right] = \mathrm{Var}[Y_1]\,\mathbb{E}[T]. \tag{15}$$

The following theorem is the main result of the paper:

Theorem 3. *Fix* k, $n \ge N_2$, *where* N_2 *is as in Theorem 1. Then*

$$L_n \le L_n^* + \tilde{O}(n^{-3/2}).$$

Proof. Let $S_{kn} = \sum_{t=1}^n X_{kt}$, $\hat{L}_{kn} = (S_{k,T_{kn}} - T_{kn}\mu_k)/T_{kn}$, $G'(n,\delta) = (K-1)\max(N_2, 1 + G(n,\delta))$ and

$$G''(n) = D_3\sqrt{n(1 + \log(4Kn(n+1)))} + 2.$$

Note that by Theorem 1,

$$\mathbb{P}(T_{kn} \le n\lambda_k - G'(n,\delta)) \le P(n,\delta) \triangleq \mathbb{I}\{n < N_2\} + \mathbb{I}\{n \ge N_2\}\,\delta \tag{16}$$

holds for any $n \ge 1$ and $0 < \delta \le 1$. Then, for any $0 < \delta \le 1$,

$$\begin{aligned}
L_{kn} &= \mathbb{E}\left[\hat{L}_{kn}^2\right] \\
&= \mathbb{E}\left[\hat{L}_{kn}^2 \mathbb{I}\{T_{kn} > n\lambda_k - G'(n,\delta)\}\right] + \mathbb{E}\left[\hat{L}_{kn}^2 \mathbb{I}\{T_{kn} \le n\lambda_k - G'(n,\delta)\}\right] \\
&\le \frac{\mathbb{E}\left[(S_{k,T_{kn}} - T_{kn}\mu_k)^2\right]}{(n\lambda_k - G'(n,\delta))^2} + R^2\,\mathbb{P}(T_{kn} \le n\lambda_k - G'(n,\delta)) \\
&= \frac{\sigma_k^2 \mathbb{E}[T_{kn}]}{(n\lambda_k - G'(n,\delta))^2} + R^2\,\mathbb{P}(T_{kn} \le n\lambda_k - G'(n,\delta)) \qquad \text{(by Lemma 9)} \\
&= \frac{\sigma_k^2 \mathbb{E}[T_{kn}]}{(n\lambda_k - G'(n,\delta))^2} + R^2\,P(n,\delta) \qquad \text{(by (16))} \\
&\le \frac{\sigma_k^2 (n\lambda_k + G''(n))}{(n\lambda_k - G'(n,\delta))^2} + R^2\,P(n,\delta) \qquad \text{(by 13)} \\
&= \frac{\sigma_k^2}{n\lambda_k} \frac{1}{(1 - G'(n,\delta)/(n\lambda_k))^2} + \frac{\sigma_k^2 G''(n)}{(n\lambda_k - G'(n,\delta))^2} + R^2\,P(n,\delta).
\end{aligned}$$

Now choose $\delta = n^{-3/2}$. Then, for n sufficiently large, $G'(n, n^{-3/2})/(n\lambda_k) \leq 1/2$. Further, since $N_2 \leq N_2' \log(4K/\delta)$, for n sufficiently large $\mathbb{I}\{\, n < N_2 \,\} \leq \mathbb{I}\left\{\, n < N_2' \log(4Kn^{3/2}) \,\right\} = 0$ and thus $P(n, \delta) = \delta$.

Therefore, for n sufficiently large, using $1/(1-x) \leq 1 + 2x$ ($|x| \leq 1/2$) we get,

$$L_{kn} \leq \frac{\sigma_k^2}{n\lambda_k} \left(1 + 2\frac{G'(n, n^{-3/2})}{n\lambda_k}\right)^2 + \frac{\sigma_k^2 G''(n)}{(n\lambda_k - G'(n, n^{-3/2}))^2} + R^2\, n^{-3/2},$$

which gives

$$L_{kn} \leq \frac{\sigma_k^2}{n\lambda_k} + \tilde{O}(n^{-3/2}) = \frac{\Sigma^2}{n} + \tilde{O}(n^{-3/2}) = L_n^* + \tilde{O}(n^{-3/2}).$$

Taking the maximum with respect to k yields the desired result. □

With a little extra work the case when for some options $\lambda_k = 0$ can also be handled and we can get identical bounds. Due to the lack of space this is not considered here.

4 Illustration

In addition to theory, empirical experiments show that our method indeed performs better than the non-adaptive solution. Further, our experiments verified that the allocation strategy found by our algorithm converges to the optimal allocation strategy at the rate predicted by the theory.

Here we illustrate the behavior of these algorithms in a simple problem with $K = 2$, with the random responses modeled as Bernoulli random variables for each of the options. In order to estimate the expected squared loss between the true mean and the estimated mean we repeat the experiment 100,000 times, then take the average. The error bars shown on the graphs show the standard deviations of these averages. The algorithms compared are GAFS-MAX (the algorithm studied here), GFSP-MAX (the algorithm described in the introduction that works in phases) and "UNIF", the uniform allocation rule. In order for an adaptive algorithm to have any advantage the two options have to have different variances. For this purpose we chose $p_1 = 0.8$, $p_2 = 0.9$ so that $\lambda_1 = 0.64$ and $\lambda_2 = 0.36$.

Figure 1 shows the rescaled excess loss, $n^{3/2}(L_n - L_n^*)$, for the three algorithms. We see that the rescaled excess losses of the adaptive algorithms stay bounded, as predicted by the theory, while the rescaled loss of the uniform sampling strategy grows as $\sqrt{n}$. It is remarkable that the limit of the rescaled loss seems to be a small number, showing the efficiency of the algorithm. Note that this example shows that the uniform allocation initially performs better than the adaptive rules. This is because the adaptive algorithms need to get a good estimate of the statistics before they can start exploiting.

Figure 2 shows and the rescaled allocation ratio deviations, $\sqrt{n}(T_{kn}/n - \lambda_k)$, for $k = 1$. Again, as predicted by the theory, the rescaled deviations stay bounded

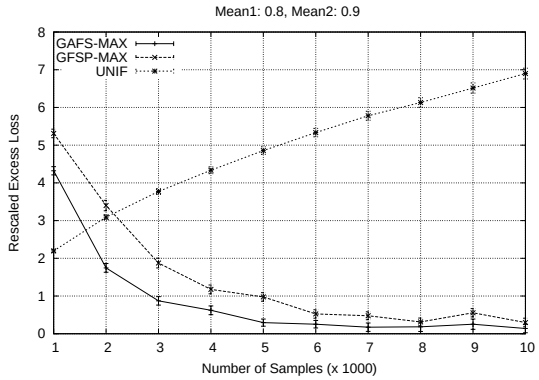

Fig. 1. The rescaled excess loss against the number of samples.

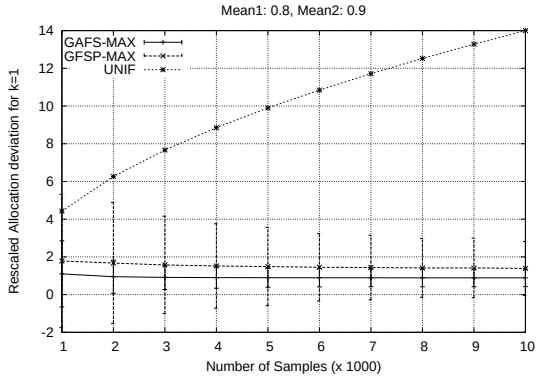

Fig. 2. The rescaled allocation deviation for $k = 1$ against the number of samples.

for the adaptive algorithms, while, due to mismatch of the allocation ratios, grows as $\sqrt{n}$ for the uniform sampling method. In this case the incremental method (GFSP-MAX) performs better than the algorithm that works in phases (GAFS-MAX), although their performance is quite similar.

5 Related Work

This work is closely related to active learning in a regression setting (e.g., [3]). Interestingly, in the by now rather extensive active learning literature to the best of our knowledge no one looked into the problem of learning in a situation where the noise in the dependent variable varies in space, i.e., under *heteroscedastic noise*. Although the rate of convergence of a method that pays attention to heteroscedasticity will not be better than that of the one that does not, the finite-time performance can be improved greatly by such adaptive algorithms. This has been demonstrated convincingly in the related problem of actively deciding about the proportions of samples to be used in stratified sampling [5].

Interestingly, this application is very closely related to the problem studied here. The only difference is that the loss is measured by taking the weighted sum of the losses of the individual prediction errors with some fix set of weights that sum to one. With obvious changes, the algorithm presented here can be modified to work in this setting and the analysis carries through with almost no changes. The algorithm studied in [5] is the phase-based algorithm. The results in this paper are weak consistency results, i.e., no rate of convergence is derived. In fact, the only condition the authors pose on the proportion of forced selections is that this proportion should go to zero such that the total number of forced selections for any option goes to infinity.

6 Conclusions and Future Work

When finite sample performance is important, one may exploit heteroscedasticity to allocate more samples to parts of the input space where the variance is larger. In this paper we designed an algorithm for such a situation and showed that the excess loss of this algorithm compared with that of an optimal rule, that knows the variances, decays as $\tilde{O}(n^{-3/2})$. We conjecture that the optimal minimax rate is in fact $O(n^{-3/2})$. Our analysis can probably be improved. In particular, the dependence of our constants on $\lambda_{\min}^{-1}$ can probably be improved by a great extent.

Although in this paper we have not considered the full non-parametric regression problem, we plan to extend the algorithm and the analysis to such problems. We also plan to apply the technique to stratified sampling.

Acknowledgements

This research was funded in part by the National Science and Engineering Research Council (NSERC), iCore and the Alberta Ingenuity Fund and by the Hungarian Academy of Sciences (Bolyai Fellowship for András Antos).

References

1. K.B. Athreya and S.N. Lahiri. *Measure Theory and Probability Theory*. Springer, 2006.
2. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
3. R. Castro, R. Willett, and R.D. Nowak. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems 18 (NIPS-05)*, 2005.
4. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer-Verlag New York, 1996.
5. P. Etore and B. Jourdain. Adaptive optimal allocation in stratified sampling methods, 2007. http://www.citebase.org/abstract?id=oai:arXiv.org:0711.4514.
6. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
7. T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.