Report on *Correlated bandits or: How to minimize mean-squared error online* (Boda and Prashanth, 2019)

1 Abstract

This report is a high level summary, review, and extension of Correlated bandits or: How to minimize mean-squared error online (Boda and Prashanth, 2019).

2 Introduction

In this paper, Boda and Prashanth ("the authors") consider a multi-armed bandit problem where the arms are correlated (sub-) Gaussian random variables, and the objective is to identify the single arm which best approximates the full set of arms. They quantify this by introducing a Mean Squared Error objective to evaluate how well an arm approximates all others. They present a formal environment for the game in which an agent pulls two arms at a time, receiving a sample from the induced bivariate distribution on those two arms. This agent has a fixed sampling budget (a fixed number of pulls allowed) and is judged on its probability of error of identifying the best arm within this budget, according to the Mean Squared Error (MSE) criteria. They then present an algorithm to solve this problem with greater correctness than a naive uniform sampling algorithm. Their algorithm is an adaptation of the Successive Rejects algorithm, typically used in the traditional highest mean bandit problem, but also used in correlated bandit environments with different objectives [Liu and Bubeck, 2014], here modified to operate on pairs of arms. They present an upper bound on the probability of misidentifying the best arm given a total number of trials, which improves upon the upper bound obtained from a naive uniform sampling algorithm.

While most of the formulations are well defined and analysis is rigorous, the paper has some inconsistencies in the definition of the Successive Rejects Algorithm. Namely, it includes a partition of the input number of samples that yields a total number of samples in the algorithm that is much lower than the budget, and a fix in this partition can increase the number of samples made, thus improving estimated statistics, while remaining under the budget. See the Main Theorems/Algorithms section for this analysis.

The experiments that the authors present demonstrate the superiority of their Successive Rejects algorithm in the case where the distribution of arms is optimal for its success; however, we show in a further experiment that other cases of the distribution of arms results in a worse advantage, and even a slight disadvantage, of the Successive Rejects Algorithm compared to the uniform sampling strategy. See the Experiments section for this analysis.

3 Related Work

The authors initially list applications of estimating correlated sources of data using one or a few sources of the data. They then look at formulations and analysis of bandit problems.

[Kaufman et. al., 2016] among others analyze multi-armed bandit models where the objective is to find the arm with the highest expected value. They discuss two settings for best-arm identification:

1. *Fixed Budget:* In this setting, the agent has a fixed number of samples it may make from the bandit, where the samples must be split across the arms according to some strategy. After making some number of samples less than or equal to the given budget, it must output an estimate of the best arm. The goal here is to find a strategy that has an optimal probability of correctly idenfitying the best arm, given this budget.
2. *Fixed Confidence:* In this setting, the number of samples an agent may take is not upper bounded, and the agent must output an estimate of the best arm that has a probability of correctness within some fixed error bound. The goal then is to find a strategy that uses the fewest number of samples (or lowest expected value of the number of samples, in the case of an adaptive strategy) to obtain an estimate within this given error bound.

[Audibert et. al., 2010] considers the *Fixed Budget* setting and proposes the Successive Rejects Algorithm. In this algorithm, the budget is partitioned into rounds, and during each round an arm is successively eliminated according to some estimate of an objective, and the remaining arm is output. [Audibert et. al., 2010] use the objective of finding the arm with the highest reward (highest expected value), so each round eliminates the arm with the lowest sample mean so far.

These settings can both apply to multi-agent bandit problems where the objective is not to find the arm with the highest expected value as the rest of [Kaufman et. al., 2016] discusses, but rather to find some arm or subset of arms with some other statistical property. Here, the authors consider the case of the *Fixed Budget* setting, where the best approximating arm must found within a given number of samples, in the sense of optimizing the Mean Squared Error.

Other papers such as [Gupta et. al., 2019], [Pandey et. al., 2007] analyze multi-armed bandits with dependences between arms, but most prior work considers only the objective of finding the arm with highest expected value, unlike the objective here.

[Liu and Bubeck, 2014] have done the work most similar to that of the authors. They consider a similar setting as the authors, with a multi-armed bandit in which the K arms are correlated Gaussian random variables. The objective of the agent in this environment is, given an input $0 < h < K$, to find the subset $S^* \subset [K], |S| = h$ of the arms that has the most mutual correlation. This contrasts the work of Boda and Prashanth here where [Liu and Bubeck, 2014] seek a subset of arms that may or may not have high correlation with arms outside of the subset, whereas Boda and Prashanth seek a single arm that has high correlation with all other arms of the bandit. Despite this difference in objective, the work of Boda and Prashanth follows from the work of [Liu and Bubeck, 2014], in that the authors work is primarily an adaptation the analysis, proof techniques, and algorithms of [Liu and Bubeck, 2014]. Therefore, while the authors' contributions of formalizing a new multi-armed bandit environment/objective with corresponding algorithms and accuracy bounds are novel, it follows much more directly from [Liu and Bubeck, 2014] than mentioned in [Boda and Prashanth, 2019].

[Liu and Bubeck, 2014] consider both the *Fixed Budget* and *Fixed Confidence* settings; towards the *Fixed Budget* setting, they propose a Successive Rejects Algorithm *SR-C*, and it is this algorithm that Boda and Prashanth adapt for their own bandit objective. (It can be noted that some variation of the Successive Elimination Algorithm in [Liu and Bubeck, 2014] for the *Fixed Confidence* may potentially be adapted similarly for the objective in [Boda and Prashanth, 2019]; however, all further discussion pertains to the Successive Rejects Algorithm and the *Fixed Budget* setting).

[Liu and Bubeck, 2014] assume unit variance for all arms; they then measure their objective with their *suboptimality ratio*:

$$\mathcal{D}_\Sigma(A, B) = \frac{\sum_{(j,\ell)\in B^2\setminus(A\cap B)^2, j\neq\ell}(1-\sigma_{j,\ell})}{\sum_{(j,\ell)\in A^2\setminus(A\cap B)^2, j\neq\ell}(1-\sigma_{j,\ell})}$$

where $\sigma_{j,l}$ is the covariance (and thus correlation, assuming unit variance) between arm j and arm l, and $A, B \subseteq [K]$. The MSE objective function in [Boda and Prashanth, 2019] is similar to a special case of the suboptimality ratio, and therefore the Successive Rejects Algorithm that arises from the MSE objective resembles that of the Successive Rejects Algorithm in [Liu and Bubeck, 2014]. This can be seen by considering arm i and letting B be the set of all arms while A is the set of all arms excluding arm i; so the numerator of the expression $\mathcal{D}_\Sigma(A, B)$ becomes a sum over all pairs of arms that include i, and the denominator becomes a sum over the null set, which for the sake of definedness we take to be 1 by convention. Finally, if we weight each term of the sum by the variance of arm j (since the authors in [Liu and Bubeck, 2014] take all variances to be unity and this paper

considers general variances), and if we square the correlation terms (the authors in [Liu and Bubeck, 2014] assume positive correlations so squaring enforces positivity), this yields the MSE expression. Therefore, [Boda and Prashanth, 2019] is nearly a special case of [Liu and Bubeck, 2014] when considering the sets A and B to be defined appropriately.

4 Preliminaries

In their paper, [Boda and Prashanth, 2019] consider bandits whose arms are random variables drawn from a multivariate Gaussian distribution. The mean of this distribution is assumed to be 0 to simplify analysis (in general, estimators for the mean can be factored into estimates of MSE without significantly affecting accuracy bounds).

An interaction with this bandit in the environment presented in [Boda and Prashanth, 2019] is a sample of a pair of arms, (X_i, X_j), drawn from the bivariate distribution induced by the multivariate distribution of the bandit in dimensions i and j, $i, j \in [K] = 1, 2, ..., K$. In the sample (X_i, X_j), order matters in that the first sample is decidedly from arm i and the second from arm j, but the order also does not matter, in the sense that the pair is sampled simultaneously, and the order of i and j only matters as far as any agent interacting with the bandit "knowing" which sample came from which arm. A typical sequence of n samples made by some agent from this bandit (aka arm pulls) may look like $\{(X_{i_1,1}, X_{j_1,1}), (X_{i_2,2}, X_{j_2,2}), ..., (X_{i_n,n}, X_{j_n,n})\}$ where $i_t, j_t \in [K], i_t \neq j_t \forall t = 1, ..., n$, and i_t, j_t are chosen by the agent according to some strategy, possibly independently for each time step, and possibly depending on previous observed samples. Given a fixed budget n, an agent may make up to n samples, then must output an arm, and the agent is judged on the probability of this output to be the arm i^* with the smallest MSE among all arms of the bandit.

The MSE for arm $i \in [K] = 1, 2, ..., K$ is

$$\mathcal{E}_i = \sum_{j=1}^{K} \mathbb{E}\left[\left(X_j - \mathbb{E}[X_j|X_i]\right)^2\right] = \sum_{j \neq i} \sigma_j^2(1 - \rho_{ij}^2).$$

and $i^* = \text{argmax}_i \mathcal{E}_i$. We assume such i^* is unique.

To estimate the MSE for arm i, the following typical estimators are used:

$$\hat{\rho}_{ij} = \frac{1}{\hat{\sigma}_i^2 \hat{\sigma}_j^2} \frac{1}{n} \sum_{t=1}^{n} X_{i,t} X_{j,t}$$
$$\hat{\sigma}_i^2 = \frac{1}{n} \sum_{t=1}^{n} X_{i,t}^2$$
$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_{t=1}^{n} X_{j,t}^2$$

for $j \in [K] \setminus \{i\}$ and $\{(X_{i,t}, X_{j,t}),\ t = 1, \ldots, n\}$ is some subsequence of a sequence of samples taken from the bandit, such that all pairs in the subsequence are pairs of samples from arm i and arm j (again, regardless of order).

Then the MSE for arm i is estimated by

$$\hat{\mathcal{E}}_i = \sum_{j \in [K] \setminus \{i\}} \hat{\sigma_j}^2 (1 - \hat{\rho}_{ij}^2)$$

It should be emphasized that in computing some term $\hat{\sigma_j}^2(1 - \hat{\rho}_{ij}^2)$ of the sum for some $j \neq i$, the estimators $\hat{\sigma}_j$ and $\hat{\rho}_i j$ are computed only from the subsequence of pairs of samples that are from the bivariate distribution of arm i and arm j; instances of arm i or arm j that occur as elements of a pair with some other arm not equal to i or j are excluded from these estimates; in other words, the term only uses estimators computed only from pairs of samples in which one element of the pair is arm i and the other is arm j.

The authors write the correlation estimator $\hat{p}_{ij}$ in another form to resemble the estimator in [Liu and Bubeck, 2014]. However, the estimator in [Liu and Bubeck, 2014] assumes unit variance, and although the authors in [Boda and Prashanth, 2019] attempt to normalize variance estimates, they in fact normalize by the estimates themselves. This is both why their estimator reduces to the typical estimator and why it does not share the properties that they claim it shares with the estimator in [Liu and Bubeck, 2014], which it would if they were to normalize by (unknown) *true* variances.

Next, analogously to the typical multi-armed bandit case, define the differences $\Delta_i = \mathcal{E}_i - \mathcal{E}_{i^*} > 0$ for $i \neq i^*$. Using these values, the following terms arise in analysis of the accuracy of the algorithms of this paper:

$$H_2 = \max_i \frac{i}{\Delta^2_{(i)}},$$
$$\overline{H} = \sum_i \frac{1}{\Delta_i^2}.$$

We index the differences Δ_i for arm i, and $\Delta_{(i)}$ for the $i'th$ best difference, ie $\Delta_{(2)} \leq \Delta_{(3)} \leq \ldots \leq \Delta_{(K)}$ (and $\Delta_{(1)} := \Delta_{(2)}$ to exclude any 0 terms).

The relationship with these values in complexities of the algorithms is most apparent in the statements and proofs of the complexities, but intuitively, $\bar{H}$ is greater for bandits whose arms have a small range of MSE values, which we shall see in Section 4 diminish the advantage of the Successive Rejects Algorithm. Meanwhile, H_2 intuitively measures how far the MSE of the i'th best arm is from that of the optimal arm, weighted higher for arms further away. This arises in the analysis of the Successive Rejects Algorithm because arms that are more suboptimal in relation to other arms are rejected earlier on (when respective MSE estimators are appropriately concentrated), so a more suboptimal arm having a closer MSE value to that of i^* is worse for accuracy – again, illustrated in the Section 4.

5 Experiments

For all of the following experiments (unless otherwise noted), the arms are assumed to be random variables with a multivariate Gaussian distribution, with mean 0 and covariances specified. We also use positive covariances for the sake of exposition and clarity, noting that the estimated terms in the MSE objective are squared, so only the absolute values of the underlying statistics affect the optimality (or lack thereof) of an arm.

Experiment 0

First, we do a simple experiment to illustrate the differences between the two algorithms. We consider a 4 armed bandit and use the covariance matrix

$$\Sigma_0 = \begin{bmatrix} 1 & 0.6 & 0.2 & 0.1 \\ 0.6 & 1 & 0.1 & 0.1 \\ 0.2 & 0.1 & 1 & 0 \\ 0.1 & 0.1 & 0 & 1 \end{bmatrix}$$

where the entry $(\Sigma_0)_{ii}$ is the variance of arm i and the entry $(\Sigma_0)_{ij}$ is the covariance between arms i and j. Since the diagonals are all 1, we can interpret the off diagonal entries of Σ_0 as correlations; thus, by reading off values from Σ_0 we see that arms 3 and 4 have correlations < 0.5 with all other arms, and arm 1 has correlation 0.2 with arm 3 while arm 2 only has correlation 0.1 with arm 3, so arm 1 is the optimal arm that best informs about the remaining arms.

Since we know the underlying statistics of the distribution of the arms, we can compute the actual MSE of each arm, giving (in order of arm):

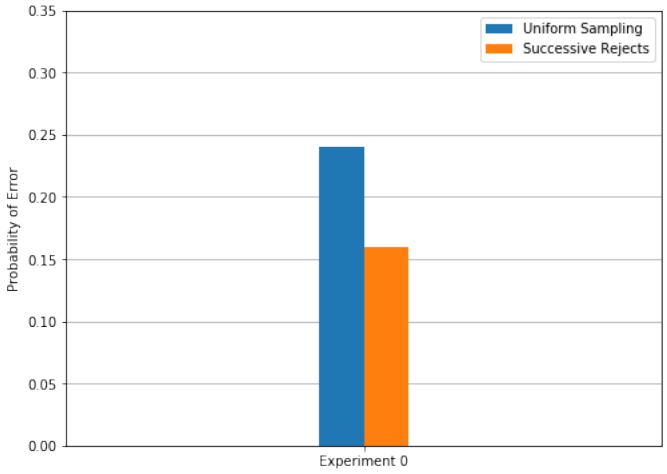
$$[2.59, 2.62, 2.95, 2.98]$$

Again, we clearly see that arm 1 is optimal. Furthermore, the MSE values of arm 1 and arm 2 are relatively close, indicating that a relatively high number of samples may be needed to distinguish their MSE values during estimation.

The Successive Rejects Algorithm proceeds as follows, with a budget of $50,000$ samples, with the number of samples taken per pair listed for each phase:

```
Phase 1: 4412 samples per pair
    MSE Estimates: [2.60899347 2.60139873 2.93319304 2.97910962]
    Rejected arm:  4
Phase 2: 1470 samples per pair
    MSE Estimates: [2.64747031 2.63043893 2.92916114 2.98060918]
    Rejected arm:  3
Phase 3: 2941 samples per pair
    MSE Estimates: [2.62030655 2.65433491 1.92476152 2.00599332]
    Rejected arm:  2
Best arm: 1
```

The total number of samples taken is $\binom{4}{2}*4412+\binom{4}{2}*1470+\left(\binom{4}{2}-\binom{2}{2}\right)*2941 = 49,997 < 50000$. Note that the MSE estimates for arms 3 and 4 are update even after they are eliminated, because the pairs of arms $(1,3)$ and $(1,4)$ remain active until termination and sampling these pairs gives information about the MSE values of arms 3 and 4. Also note that after the first phase, the algorithm cannot well distinguish between the MSE values of arms 1 and 2, but is able to update their MSE estimates with further samples by rejecting suboptimal arms, and thus identifies the optimal arm correctly. The advantage of the rejection strategy is summarized in the following figure, which compares the uniform sampling strategy with the Successive Rejects Algorithm, each given a sample budget of $50,000$ samples and repeated for 100 trials:



Experiments 1-3

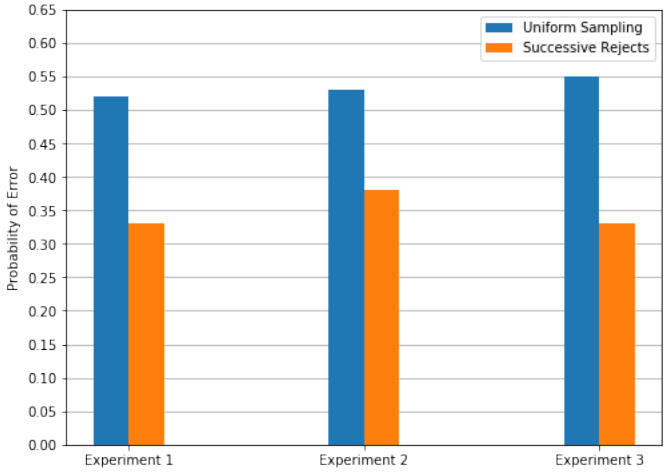
Next, we attempt to replicate the results in [Boda and Prashanth, 2019]. We use the same settings, where the bandits have arms that are jointly Gaussian with the covariance matrices covariance matrices $\Sigma_1, \Sigma_2, \Sigma_3$ and the corresponding number of arms ($K_1 = 29, K_2 = 35, K_3 = 34$) for experiments $1, 2, 3$ respectively.

$$\Sigma_1 = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{25\times 25} \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Tr}_{31\times 31} \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} 1 & 0.5 & 0.45 & 0.5 & \mathbf{0} \\ 0.5 & 1 & 0.45 & 0.4 & \mathbf{0} \\ 0.45 & 0.45 & 1 & 0.4 & \mathbf{0} \\ 0.5 & 0.4 & 0.4 & 1 & \mathbf{0} \\ & & \mathbf{0} & & \mathbf{I}_{30\times 30} \end{bmatrix},$$

where $\mathbf{M}_1 = [1\ 0.9\ 0.9\ 0.9;\quad 0.9\ 1\ 0.85\ 0.85;$
$0.9\ 0.85\ 1\ 0.85;\quad 0.9\ 0.85\ 0.85\ 1]$, and $\mathbf{Tr}_{31\times 31}$ is a tridiagonal matrix with ones along the main diagonal, 0.2 in the diagonals above and below and zeros elsewhere.

We compare average accuracies of the uniform sampling strategy and the Successive Rejects Algorithm over 100 trials, where the sample budget for each experiment i is approximately $\bar{H}_i * 32^2 * 4$. For experiment 1, this corresponds roughly to a budget of $419,000$ samples; for experiment 2, $424,000$ samples; and for experiment 3, $1,200,000$ samples. The authors state that they use a sample budget approximately equal to $\frac{\bar{H}}{32^2}$; however, this gives a sample budget of less than 1 sample for each experiment, so we assume this is a typo and use this new sample budget to obtain results resembling the original experiments.



We see that in each case, the Successive Rejects Algorithm gets a probability of misidentification that is much lower than that of the uniform sampling strategy. While the results for the first two experiments resemble those in [Boda and Prashanth, 2019], we see that the third experiment, despite having a proportional sample budget, has much higher error than the authors' experiment.

Now we discuss the conditions and limitations of these experiments. First, observe that $\Sigma_1, \Sigma_2, \Sigma_3$ are all covariance matrices, and furthermore the diagonals are all 1 so each defines a set of arms each with unit variance, and therefore each nondiagonal entry is a correlation value between -1 and 1.

Next, observe that each experiment has arm 1 the most optimal arm (because its correlations with the other arms are overall the highest). Furthermore, each covariance matrix defines a cluster of arms that is highly correlated among themselves, while the remainder of arms are either loosely correlated among themselves or uncorrelated, while the two groups of arms have no correlation. This is an optimal setting for the Successive Rejects Algorithm to outperform the uniform sampling strategy on the same sampling budget, because the Successive Rejects Algorithm can reject the uncorrelated/loosely correlated arms in beginning phases and then use the remainder of the sample budget to distinguish between the highly correlated arms, while the uniform sampling strategy "wastes" samples on arms that are clearly unlikely to be optimal.

To show that this setting presented in [Boda and Prashanth, 2019] is optimal for Successive Rejects, we present alternative experiments.

Experiment 4

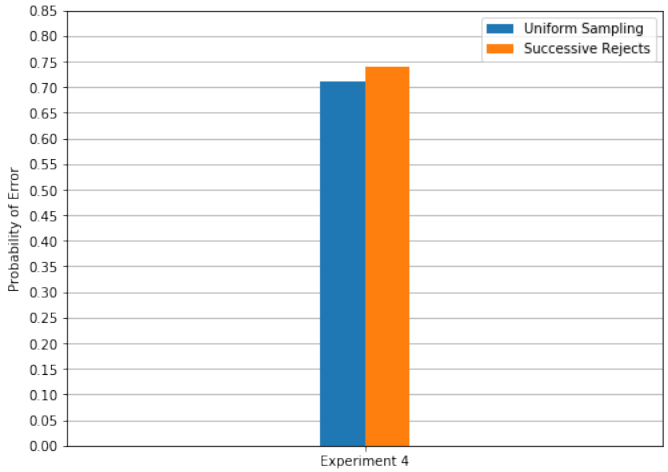
Experiment 4 examines a setting with a bandit whose 30 arms are again random variables with a multivariate Gaussian distribution. Again, there are two groups of arms; the first group of 15 arms has low pairwise correlations, and the second group of 15 arms has no pairwise correlations. There is also no correlation between the two groups of arms. The actual MSE values for the arms are

approximately (in order of arms):

$$[28.38, 28.40, 28.42, 28.45, 28.47,\\ 28.49, 28.52, 28.54, 28.56, 28.58,\\ 28.60, 28.62, 28.64, 28.66, 28.68,\\ 29, 29, ..., 29]$$

Arm 1 has the optimal MSE value, and the first 15 arms all have very close MSE values. In this setting, although arm 1 is optimal, it is barely more optimal (has MSE value barely greater) than other arms. The inclusion of a group of arms with 0 correlation with the first group of arms helps to squash the distribution of MSE values in the first group, bringing all values closer together and making identification of the optimal arm more difficult.

The uniform sampling strategy and the Successive Rejects Algorithm are each run with a sample budget of approximately $1,460,000$, and the experiment is repeated for 100 trials, giving the follow errors of prediction:



The Successive Rejects Algorithm fails to out-predict the uniform sampling strategy because in the first few phases, a relatively small amount of the sample budget is allocated to all pairs of arms, so the concentration bounds on the MSE estimates are wide, and there is a higher probability of rejecting the optimal arm. In the uniform sampling strategy, however, more samples are taken before any decisions are made, and this allows the algorithm to match or beat the Successive Rejects Algorithm.

This experiment highlights where the Successive Rejects Algorithm is at least as weak as the uniform sampling strategy – when a bandit has a lot of arms with a small range of MSE values.

6 Main Theorems/Algorithms (Supplement)

6.1 Uniform Sampling

The advantage of the Successive Rejects Algorithm is analyzed in comparison to the naive Uniform Sampling algorithm.

Let K be the number of arms of the bandit and n the total sampling budget of the agent. Assume $K \geq 2$ and $\binom{K}{2} \leq n$ (so that all pairs of arms pay be pulled at least once).

Uniform Sampling Algorithm

Sample each pair of arms $\lfloor n/\binom{K}{2} \rfloor$ times. Compute the MSE estimates of each arm from these samples. Then return the arm $a \in [K]$ with the lowest MSE estimate.

Theorem: For uniform sampling, the probability of error in identifying the optimal arm is

$$\mathbb{P}\left(\hat{A}_n \neq i^*\right) \leq 84K^2 \exp\left(-\frac{nl^2\Delta_{(1)}^2}{cK^7}\right),$$

where c is a universal constant and $l = \min_i \sigma_i^2$.

This is proved from the concentration bound for the MSE estimator.

6.2 Successive Rejects

Now we introduce the Successive Rejects algorithm. Again, let K be the number of arms of the bandit and n the total sampling budget of the agent, and assume $K \geq 2$ and $\binom{K}{2} \leq n$. We partition the number of trials in the following way. Define the normalizing value $C(K)$ as

$$C(K) := \frac{K-1}{2} + \sum_{j=1}^{K-2} \frac{j}{K-j} \leq K \log K.$$

and define the values n_k, $k = 1, ..., K-1$, as:

$$\begin{aligned} n_0 &= 0, \\ n_k &= \left\lceil \frac{n - \binom{K}{2}}{C(K)\,(K-(k-1))} \right\rceil \\ &\text{For } k = 1, ..., K-1 \end{aligned}$$

Additionally, let the set $B_1 := [K]$ be the set of all arms of the bandit, and let $A_1 := \{(i_1, i_2) \,|\, i_1, i_2 \in [K], i_1 \neq i_2\}$ be the set of all unordered pairs of arms.

The Successive Rejects algorithm presented here slightly differs from the algorithm presented in [Boda and Prashanth, 2019]. This difference is noted below the presentation of the algorithm. Our altered version is the following:

Successive Rejects Algorithm

Define A_1, B_1 and n_k for $k = 0, ..., K-1$ as mentioned above.

Repeat for Phase $k = 1, \ldots, K-1$:

1. Sample each pair in A_k $(n_k - n_{k-1})$ number of times. Estimate the MSEs using these samples along with samples from previous phases, and find the worst arm a_k among the active arms in B_k, ie. the arm with highest estimated MSE.
2. Set $B_{k+1} = B_k \setminus a_k$ and $A_{k+1} = A_k \setminus \{(a_k, a_1), (a_k, a_2), \ldots, (a_k, a_{k-1})\}$, where $B_k^c = \{a_1, \ldots, a_{k-1}\}$ is the set of arms that are out of contention by the end of phase $k-1$. In other words, obtain A_{k+1} by removing all pairs of arms $\{a_k, a_i\}$ such that a_k is the arm rejected in this phase and a_i is an arm that has been rejected in a previous phase.

After Phase $K-1$, the remaining set B_K contains a single arm a_K ($K-1$ arms have been rejected). Return this arm.

Note that after arm a_k is rejected, the set A_{k+1} may have pairs of arms for which one element of the pair is the arm a_k. This occurs when the other element of the pair is an arm $a_i, i > k$ that has not been rejected yet.

Now we show that the total number of samples made by this algorithm is within the given sampling budget. Following this algorithm, at the beginning of the k'th phase, $k-1$ arms will have been rejected, giving $\binom{k-1}{2}$ pairs that are not sampled in the k'th phase, and $\binom{K}{2} - \binom{k-1}{2}$ pairs that are sampled in the k'th phase. Thus, the total number of samples is:

$$\left[\binom{K}{2} - \binom{0}{2}\right](n_1 - n_0) + \left[\binom{K}{2} - \binom{1}{2}\right](n_2 - n_1)$$
$$+ \left[\binom{K}{2} - \binom{2}{2}\right](n_3 - n_2) + ... + \left[\binom{K}{2} - \binom{k-1}{2}\right](n_k - n_{k-1}) \tag{1}$$
$$= \sum_{k=1}^{K-1} \left[\binom{K}{2} - \binom{k-1}{2}\right](n_k - n_{k-1}) \tag{2}$$
$$= \left[\binom{K}{2} - \binom{K-1}{2}\right] n_{K-1} - \left[\binom{K}{2} - \binom{K-1}{2}\right] n_{K-1} + \sum_{k=1}^{K-1} \left[\binom{K}{2} - \binom{k-1}{2}\right](n_k - n_{k-1}) \tag{3}$$
$$= \left[\binom{K}{2} - \binom{K-1}{2}\right] n_{K-1} + \sum_{k=1}^{K-1} \left(\left[\binom{K}{2} - \binom{k-1}{2}\right] n_k - \left[\binom{K}{2} - \binom{k}{2}\right] n_k\right) \tag{4}$$
$$= \left[\binom{K}{2} - \binom{K-1}{2}\right] n_{K-1} + \sum_{k=1}^{K-1} \left[\binom{k}{2} - \binom{k-1}{2}\right] n_k \tag{5}$$
$$= \left[\frac{K(K-1)}{2} - \frac{(K-1)(K-2)}{2}\right] n_{K-1} + \sum_{k=1}^{K-1} \left[\frac{k(k-1)}{2} - \frac{(k-1)(k-2)}{2}\right] n_k \tag{6}$$
$$= (K-1)n_{K-1} + \sum_{k=1}^{K-1} (k-1)n_k \tag{7}$$
$$\tag{8}$$

where (4) follows from (3) by distributing the coefficient in the summation across terms n_k, collecting like terms, and noting that $n_0 = 0$. We use the convention that $\binom{0}{2} = \binom{1}{2} = 0$.

Then

$$\begin{aligned}
&= (K-1)n_{K-1} + \sum_{k=1}^{K-1} (k-1)n_k \\
&\leq \left(\frac{n - \binom{K}{2}}{C(K)}\right) \left(\frac{K-1}{K-(K-1-1)} + \sum_{k=1}^{K-1} \frac{k-1}{K-(k-1)}\right) + K - 1 \\
&= \left(\frac{n - \binom{K}{2}}{\frac{K-1}{2} + \sum_{j=1}^{K-2} \frac{j}{K-j}}\right) \left(\frac{K-1}{2} + \sum_{k=1}^{K-2} \frac{k}{K-k}\right) + K - 1 \\
&= \left(n - \binom{K}{2}\right) + K - 1 \\
&= n - (K-1)\frac{K}{2} + (K-1) \\
&\leq n
\end{aligned}$$

where in the first inequality we add the term $K-1$ to account for the ceiling operations on each of $n_1, ..., n_{K-1}$, and in the last inequality we use $K \geq 2$.

Note that with these definitions, in some cases this algorithm may give $n_{k+1} = n_k$ in which case phase $k+1$ has $n_{k+1} - n_k = 0$ samples; in this case, arm a_{k+1} is rejected immediately after arm a_k is rejected in phase k, without taking more samples in phase $k+1$.

The authors in [Boda and Prashanth, 2019] define the Successive Rejects algorithm on only $K-2$ phases, using a sampling schedule based only on the values $n_k, k = 1, ..., K-2$. To reject $K-1$ arms in $K-2$ phases, they reject two arms in the first phase. However, defining the algorithm on $K-1$ phases, with $n_k, k = 1, ..., K-1$, gives a cleaner algorithm that rejects only one arm in

each phase for all phases $1, ..., K-1$, gives a more accurate algorithm that uses more samples to determine the second-worst arm (in phase 2, which the authors originally combine with phase 1), and gives an algorithm that still has a total number of samples less than the sample budget, as seen by the above computations. This new version of their Successive Rejects algorithm also more closely aligns with the Successive Rejects algorithm presented in [Liu and Bubeck, 2014] in terms of the schedule of rejection.

The authors also present the Successive Rejects algorithm with inconsistent notation/definitions for the sets A_k and B_k. This is cleared up in the above algorithm.

Theorem: The probability of error in identifying the best arm of SR in a sample budget of n satisfies

$$\mathbb{P}(a_K \neq i^*) \leq 84K^3 \exp\left(-\frac{l^2}{cK^5}\frac{\left(n-\binom{K}{2}\right)}{C(K)H_2}\right),$$

where c is a universal constant and $l = \min_i \sigma_i^2$.

The proof of this theorem uses an analysis very similar to that in [Audibert et. al., 2010] which computes the probability of the event that the optimal arm is rejected in phase k and can be found in [Boda and Prashanth, 2019].

From the above algorithms we get that a target accuracy in identifying the optimal arm requires $O(\frac{K^7}{\Delta_2})$ samples in Uniform Sampling but only $O(K^6\bar{H})$ in the Successive Rejects algorithm.

We can see intuitively that the Successive Rejects algorithm is more accurate than Uniform Sampling in cases where arms have a large deviation in MSE. If there are arms that are "obviously" not optimal in the sense that they have MSE much worse than the optimal arm, it would take relatively few samples to determine such suboptimality (relative to if all MSE's were close together), as estimates of MSE concentrate around the true value for increasing numbers of samples. However, Uniform Sampling splits the sampling budget over all $\binom{K}{2}$ pairs of arms, so it "wastes" samples on such suboptimal arms, and has a smaller portion of the sampling budget to distinguish the optimal arm from nearly optimal arms. However, the Successive Rejects algorithm rules out suboptimal arms early on (where *more* suboptimal arms are ruled out with higher probability of correctness), and thus has a larger portion of the sampling budget to distinguish between the top-k optimal arms.

The converse of this is also true. In cases where all arms have MSE in a small range and the optimal arm is not as distinguishable from the least optimal arm, the Successive Rejects Algorithm will incorrectly eliminate the optimal arm during some phase where the number of samples so far does not tighten the concentration of the MSE estimates enough to distinguish the optimal arm from suboptimal arms with high probability.

This is seen quantitatively in Section 4.

References

J.Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 41–53, 2010.

VP Boda and Prashanth L.A. Correlated Bandits or: How to Minimize Mean-Squared Error Online. *ICML Poster Session*, 2019.

Samarth Gupta and Shreyas Chaudhari and Gauri Joshi and Osman Yagan. Multi-Armed Bandits with Correlated Arms. *Arxiv*. 2019.

E. Kaufmann, O. Cappe, and A. Garivier. On the complexity ofbest arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015

C. Y. Liu and S. Bubeck. Most correlated arms identification. In *Conference on Learning Theory*, pages 623–637,2014.

Sandeep Pandey and Deepayan Chakrabarti and Deepak Agarwal. Multi-Armed Bandit Problems with Dependent Arms. *Association for Computing Machinery*, 2007.

Appendix: Code for Experiments