Data Source Adaptive Online Learning under Heteroscedastic Noise

Amith Bhat Hosadurga Anand ABHAT69@UIC.EDU
University of Illinois at Chicago, USA

Aadirupa Saha AADIRUPA@UIC.EDU
University of Illinois at Chicago, USA

Thomas Kleine Buening THOMAS.KLEINEBUENING@AI.ETHZ.CH
ETH AI Center, Switzerland

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California, USA

Abstract

In this paper, we address the standard K-armed multi-armed bandit (MAB) with M heterogeneous data sources, each exhibiting unknown and distinct noise variances, σ_j^2. We propose SOAR (*Source-Optimistic Adaptive Regret Minimization*), a novel algorithm that adaptively balances exploration and exploitation by jointly constructing upper confidence bounds for arm rewards and lower confidence bounds for data source variances. Our theoretical analysis establishes that SOAR achieves a regret bound of $\tilde{O}\left(\sigma^{*2} \sum_{i=2}^{K} \frac{1}{\Delta_i}\right)$, along with a preprocessing cost that depends only on the problem parameters $\{\sigma_j\}_{j=1}^{M}$, K, and grows at most logarithmically with the horizon T; where σ^{*2} is the minimum source variance, and Δ_i denotes the suboptimality-gap of the i-th arm reward. The $\tilde{O}(.)$ notation hides the polylogarithmic factors in these problem parameters. This near-optimal performance underscores SOAR's effectiveness in dynamically managing heteroscedastic noise without incurring significant overhead. We believe this work opens a new direction for adaptively leveraging multiple heterogeneous data sources, extending beyond traditional bandit frameworks.

1. Introduction

Online learning with multi-armed bandits (MABs) becomes especially challenging when data is drawn from multiple heterogeneous sources with unknown and distinct noise levels. Existing approaches fall short: multi-fidelity bandits focus on explicit cost-accuracy trade-offs, while variance-adaptive bandits adapt within a single stream, neither addressing multiple parallel sources. Intuitive baselines, such as uniform sampling or two-phase identify-then-exploit, fail in worst-case scenarios, highlighting a clear gap. A detailed discussion of these motivations, related work, and baseline limitations is provided in Appendix A.

To address this challenge, we propose SOAR (Source-Optimistic Adaptive Regret Minimization, Algorithm 2), a framework that combines Upper Confidence Bounds (UCB) on arm rewards with Lower Confidence Bounds (LCB) on source variances. SOAR adaptively filters out high-variance sources while balancing exploration across both arms and sources. We establish that SOAR attains regret nearly matching that of an oracle with prior knowledge of the best source. *Our primary contributions are:* (i) formalizing the heterogeneous multi-source bandit problem in Section 2; (ii) introducing an LCB–UCB framework for joint source–arm selection in Section 3, with the associated concentration lemmas stated and derived in Appendix C.1 and Appendix C.2, and the core algorithm presented in Section 4 and Algorithm 2; and (iii) proving regret guarantees comparable to variance-aware single-source MABs in Theorem 2 and comparing these bounds to the proposed baselines in Remark 3.

2. Problem Formulation: Source Adaptive MAB

In this section, we formalize the heterogeneous multi-source bandit setting by specifying the arms, sources, reward model, and regret formulation.

Useful Notation. Let $\mathbb{R}_+$ and $\mathbb{N}$ denote the set of positive reals and positive integers, respectively. Let $[n] := \{1, 2, \ldots n\}$, for any $n \in \mathbb{N}$.

Consider a set of K arms $[K] := \{1, \ldots, K\}$ and M data sources $[M] := \{1, \ldots, M\}$. Each arm $i \in [K]$ is associated with an unknown mean reward $\mu_i \in \mathbb{R}$, while each data source $j \in [M]$ is associated with an unknown standard deviation $\sigma_j \in \mathbb{R}_+$.

At each round $t \in [T]$, the learner selects an arm $i_t \in [K]$ and a source $j_t \in [M]$, then observes a reward $X_t = \mu_{i_t} + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{D}(j_t)$ and $\mathcal{D}(j_t)$ is a mean-zero distribution with variance σ_j^2. We further assume bounded means, i.e., $\mu_i \in [0, \bar{\mu}]$, and a bound on the noise $\mathcal{D}(j_t) \in [-\bar{\eta}, \bar{\eta}]$.

Let $i^* := \arg\max_{i \in [K]} \mu_i$ denote the optimal arm (without loss of generality, assume $i^* = 1$) and $j^* := \arg\min_{j \in [M]} \sigma_j$ denote the optimal source. We let $\mu^* := \mu_{i^*}$ and $\sigma^* := \sigma_{j^*}$ denote, respectively, the largest mean reward across arms and the smallest standard deviation across sources. For the purpose of this work, we assume $\sigma^* = 1$. The learner's objective is to minimize the expected regret, defined as: $\text{Reg}_T := \mathbf{E}\left[\sum_{t=1}^{T} \left(\mu^* - \mu_{i_t}\right)\right]$. Although the regret bound does not explicitly depend on the choice of sources, their selection plays a crucial indirect role. This is discussed in greater detail in Appendix B.

3. Warm-Up: Parameter Estimation and Confidence Bounds

We begin by introducing the estimators of mean and variance parameters, together with the confidence bounds that form the foundation of the algorithm in Section 4 and its regret analysis.

Notation. For any $t \in [T]$, let
$n_t(i) := \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}, \quad m_t(j) := \sum_{s=1}^{t} \mathbb{1}\{j_s = j\}, \quad n_t(i,j) := \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}\mathbb{1}\{j_s = j\}$
denote the number of times arm i, source j, and the pair (i, j) have been selected up to time t, respectively. The empirical mean reward of arm i and the empirical variance of source j are given by

$$\hat{\mu}_t(i) := \frac{1}{n_t(i)} \sum_{s \le t : i_s = i} X_s, \qquad \hat{\sigma}_t^2(j) := \frac{1}{m_t(j)} \sum_{s \le t : j_s = j} \left(X_s - \hat{\mu}_t(i_s)\right)^2 \tag{1}$$

i.e., the average reward of arm i up to time t, and the average squared deviation of rewards collected from source j relative to their empirical arm means.

3.1. Estimating the Confidence Bounds

For each arm $i \in [K]$ and each source $j \in [M]$, we define the Upper Confidence Bound (UCB) on the empirical mean reward $\hat{\mu}_t(i)$ and the Lower Confidence Bound (LCB) on the variance estimate

$\hat{\sigma}_t^2(j)$

$$\mathrm{UCB}_t^{\mu}(i) := \hat{\mu}_t(i) + \frac{2\sqrt{2\log(3KT/\delta)\,\sum_{j=1}^{M} n_t(i,j)\hat{\sigma}_t^2(j)}}{n_t(i)} \tag{2}$$

$$\mathrm{LCB}_t^{\sigma}(j) := \hat{\sigma}_t^2(j) - 2\bar{\eta}\hat{\sigma}_j\sqrt{\frac{4\log(12MT/\delta)}{m_t(j)}} \tag{3}$$

The mean reward concentration lemma, which supports the derivation of the UCB estimate, is stated and proven for our problem in Appendix C.1. Similarly, the reward variance concentration lemma, used to derive the LCB estimate, is presented in Appendix C.2. The variance sandwiching corollary, which provides a high-probability link between the true variance and its empirical estimate, is given in Corollary 9. Additional corollaries that support the construction of the above estimates are provided in Corollary 10 and Corollary 11. Finally, a detailed discussion and interpretation of the UCB and LCB estimates is given in Appendix C.3.

4. Algorithm: Source-Optimistic Adaptive Regret minimization (SOAR)

In this section, we present our proposed algorithm, SOAR, outlining its preprocessing and adaptive LCB–UCB design, and establish its regret guarantees in comparison to natural baselines.
Notations. We define three additional quantities that will be used in the algorithm's regret analysis. For each arm $i \in [K]$, let $\Delta_i := \mu^* - \mu_i$ denote the suboptimality gap. For each source $j \in [M]$, let $\Delta_j^{\sigma} := \sigma_j - \sigma^*$ and $\Delta_j^{\sigma^2} := \sigma_j^2 - \sigma^{*2}$ denote its excess standard deviation and excess variance relative to the most reliable source, respectively.

Some Baselines and Limitations. We highlight two natural baselines for the heterogeneous multi-source bandit problem. The first, *Baseline-1: Uniform Source MAB*, selects sources uniformly at random, leading to regret that scales with the average variance and failing in *Worst-Case instance 1 (WC-1)*, where *the variance of multiple, similar, high variance sources dominates performance.* The second, *Baseline-2: Two-Phase (Source-then-Arm) MAB*, first attempts to identify the lowest-variance source by fixing a single arm and running a best-arm identification procedure over the sources, before applying a standard MAB on the identified "best source". This strategy fails in *Worst-Case instance 2 (WC-2)*, where *all sources have nearly identical variances, resulting in unnecessary exploration cost.* Detailed analyses for these baselines are provided in Appendix D.1, and their shortcomings are discussed more technically in Remark 3, motivating the design of our proposed algorithm, SOAR, which jointly balances arm and source exploration.

4.1. Proposed Algorithm: SOAR

We now present our proposed algorithm, SOAR, which consists of two key components: (i) a preprocessing phase that eliminates "bad" or high variance sources, and (ii) an adaptive LCB–UCB procedure for joint source–arm selection.

Preprocessing to Remove "Bad" Sources: PREPROCESS is a preprocessing subroutine that eliminates high-variance sources before applying SOAR. Each source is queried a fixed number of times on the same arm, variance confidence intervals (LCB/UCB) are constructed, and a source is removed if its LCB exceeds the minimum UCB across all sources. Intuitively, sources with variance much larger than the minimum (e.g., $\sigma_j^2 > c\sigma^{*2}, c \in \mathbb{R}^+$) only add noise without reducing regret.

Setting $c = 9$ along with our assumption that $\sigma_* = 1$ ensures all retained sources satisfy $\sigma_j^2 - \sigma_*^2 < 8$. This preprocessing simplifies the variance landscape, leading to tighter regret bounds and more efficient exploration. Supporting concentration inequalities are detailed in the Appendix D.2. We do state one key result from our analysis.

Algorithm 1 PREPROCESS: Variance-Based Source Pruning

```
1: Input: Arm set: $[K]$, Feedback Sources: $[M]$, Confidence parameter: $\delta \in (0, 1)$, Runtime
   budget: $n \in \mathbb{N}_+$.
2: Init: $S_{\mathcal{G}} \leftarrow [M]$. Fix any arm $i_0 \in [K]$
3: for $j \in [M]$ do
4:     Query source $j$ for Arm-$i_0$, $n$ times
5: end for
6: Compute $\text{LCB}_n^\sigma(j)$ and $\text{UCB}_n^\sigma(j)$ for all $j \in [M]$ using Equation (8) and Equation (7)
7: $m \leftarrow \arg\min_{j \in [M]} \text{UCB}_n^\sigma(j)$
8: for $j \in [M]$ such that $\text{LCB}_n^\sigma(j) > \text{UCB}_n^\sigma(m)$ do
9:     $S_{\mathcal{G}} \leftarrow S_{\mathcal{G}} \setminus \{j\}$   // Eliminate "high" variance source
10: end for
11: Return $S_{\mathcal{G}}$      // Pruned set of sources
```

Theorem 1 (Stopping Condition of PREPROCESS) *Consider any $\delta \in (0, 1)$. If* PREPROCESS *is run with runtime budget n, where $n > 16\bar{\eta}^4 \log(12M/\delta)$, then any source $j \in [M]$ with variance $\sigma_j^2 > 9\sigma^{*2}$ will be eliminated with probability $(1 - \delta/3)$.*

Proof The complete proof of Theorem 1 is provided in Appendix D.2.3, with the variance concentration result used in the analysis defined and proven in Appendix D.2. ■

Main Ideas of SOAR: Algorithm 2 takes as input the number of arms K, number of sources M, a confidence parameter $\delta \in (0, 1)$, and an exploration parameter $\tau \in \mathbb{N}$ specifying the initial number of queries per source. In the initialization phase, each source is queried τ times with arms chosen uniformly at random, yielding initial estimates of mean rewards $\hat{\mu}(i)$ and source variances $\hat{\sigma}(j)$ along with their confidence bounds. The exploration parameter τ ensures that each source is sampled sufficiently often in the initialization phase to construct meaningful confidence estimates.

From round $t = \tilde{M}\tau + 1$ onward, the algorithm adaptively selects the arm i_t with the largest upper confidence bound on mean reward and the source j_t with the smallest lower confidence bound on variance, then queries the pair (i_t, j_t), updates counts, and recomputes estimates and bounds. Intuitively, this UCB–LCB mechanism balances optimistic exploration of arms with cautious, variance-aware selection of sources, guiding the learner toward rewarding arms and low-noise feedback. Our analysis shows that this strategy yields regret nearly matching that of an oracle with access to the optimal low-variance source, highlighting the effectiveness of our simultaneous exploration–exploitation design. A more detailed look of Algorithm 2 can be found in Appendix D.2.4.

Theorem 2 (Main Result: Regret Analysis of SOAR) *For any choice of preprocessing budget $n \geq 16 \cdot \bar{\eta}^4 \log\left(\frac{12M}{\delta}\right)$, initial-exploration $\tau = \max\left\{\frac{288M^2K^2}{\bar{\eta}^2} \frac{(\log(3KT/\delta))^2}{\log(12MT/\delta)}, 32\bar{\eta}^2 \log(12MT/\delta)\right\}$ and $\sigma^* = 1$, the regret of* SOAR *(Algorithm 2) can be bounded by $\tilde{O}\left(M\bar{\mu}(n + \tau) + \sigma^{*2} \sum_{i=2}^{K} \frac{1}{\Delta_i}\right)$*

Algorithm 2 Source-Optimistic Adaptive Regret Minimization (SOAR)

```
**Input:** Arm set: $[K]$, Feedback Sources: $[M]$, Confidence parameter: $\delta \in (0,1)$, Exploration parameter: $\tau \in \mathbb{N}_+$, Preprocessing budget: $n \in \mathbb{N}_+$.
$S_\mathcal{G} \leftarrow$ PREPROCESS$([M],[K],\delta/3,n)$ // $S_\mathcal{G}$ is set of pruned sources.
$\tilde{M}$ = $|S_\mathcal{G}|$
**Initial Exploration:** For each $j \in [\tilde{M}]$, query $j$ for $\tau$ rounds on any arm $i \in [K]$ uniformly at random
At $t = \tilde{M}\tau$, compute $n_t(i) = \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}$, $\hat{\mu}_t(i)$, $\mathrm{UCB}_t^\mu(i)$ for all $i \in [K]$, and $m_t(j) = \sum_{s=1}^{t} \mathbb{1}\{j_s = j\}$, $\hat{\sigma}_t(j)$, $\mathrm{LCB}_t^\sigma(j)$ and $n_t(i,j) = \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}\mathbb{1}\{j_s = j\}$ for all $j \in [S_\mathcal{G}]$ as defined in Equation (1), Equation (2) and Equation (3).
**for** $t = \tilde{M}\tau + 1, \ldots, T$ **do**
    $i_t \leftarrow \arg\max_{i\in[K]} \mathrm{UCB}_{t-1}^\mu(i)$, $j_t \leftarrow \arg\min_{j\in[M]} \mathrm{LCB}_{t-1}^\sigma(j)$ as defined in Equation (2) and Equation (3).
    Pull arm $i_t$ using source $j_t$ and receive reward $X_t$
    Update counts: $n_t(i) = \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}, m_t(j) = \sum_{s=1}^{t} \mathbb{1}\{j_s = j\}$ and $n_t(i,j) = \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}\mathbb{1}\{j_s = j\}$.
    Update mean and variance estimators: $\hat{\mu}_t(i)$, $\hat{\sigma}_t(j)$ as defined in Equation (1).
    Update bounds: $\mathrm{UCB}_t^\mu(i)$, $\mathrm{LCB}_t^\sigma(j)$ as defined in Equation (2) and Equation (3).
**end for**
```

with high probability $(1-\delta)$. *SOAR can also be shown to yield an instance-independent (worst-case) regret bound of* $O\left(M\bar{\mu}(n+\tau) + \sigma^*\sqrt{KT\log(KT/\delta)}\right)$

Notably, our regret analysis shows only a negligible dependence on source variances, a consequence of the LCB–UCB selection mechanism that swiftly prioritizes lower-variance sources while maintaining aggressive reward exploration. The tight confidence bounds derived in Section 3.1 are central to this result, enabling our *simultaneous exploration–exploitation* strategy to achieve regret nearly matching that of a standard MAB with privileged access to the optimal low-variance source σ^*. This highlights the strength of our adaptive design and its effectiveness in minimizing regret.

Proof The detailed proof of Theorem 2 is deferred to the Appendix D.2.5. ■

Remark 3 (Improved Regret Bound and Comparison to Baselines) *Comparing the regret guarantee in Theorem 2 with the baselines in Appendix D.1, we find that SOAR improves over both Baseline-1 (Uniform Source MAB) and Baseline-2 (Two-Phase MAB) under the worst-case instances **WC-1** and **WC-2**.*

Recalling the regret guarantee from Theorem 2, we have initial exploration $\tau = \tilde{O}\left(\max\left\{\bar{\eta}^2, \frac{M^2K^2}{\bar{\eta}^2}\right\}\right)$

In the regime where $M^2K^2 \gtrsim \bar{\eta}^4$*, the term* $M^2K^2/\bar{\eta}^2$ *dominates the maximum, therefore* $\tau = \tilde{O}(M^2K^2/\bar{\eta}^2)$.

Consequently, SOAR incurs an instance-dependent regret of $\tilde{O}\left(M\bar{\mu}\left(\bar{\eta}^4 + \frac{M^2K^2}{\bar{\eta}^2}\right) + \sigma^{*2}\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)$.

*In contrast, Baseline-1 (Uniform Source MAB) suffers, under **WC-1**, an instance-dependent regret of* $\tilde{O}\left(\sigma_{\max}^2 \sum_{i\neq i^*}\frac{\log(MKT)}{\Delta_i^\mu}\right)$*, and Baseline-2 (Two-Phase MAB) incurs* $\tilde{O}\left(\sum_{j\neq j^*}\frac{\bar{\mu}}{(\Delta_j^{\sigma^2})^2} + \sigma^{*2}\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)$

whose first term can blow up under **WC-2** *when the variance gaps* $\{\Delta_j^{\sigma^2}\}$ *are very small. Moreover,* SOAR *attains a worst-case (instance-independent) regret of* $\tilde{O}\left(M\bar{\mu}\left(\bar{\eta}^4 + \frac{M^2K^2}{\bar{\eta}^2}\right) + \sigma^*\sqrt{KT}\right)$ *, compared to Baseline-1 (Uniform Source MAB), which under* **WC-1** *has worst-case regret* $\tilde{O}\left(\sigma_{\max}\sqrt{KT}\right)$ *, and Baseline-2 (Two-Phase MAB), which has a worst-case regret of* $\tilde{O}\left(\sum_{j \neq j^*} \frac{\bar{\mu}}{(\Delta_j^{\sigma^2})^2} + \sigma^*\sqrt{KT}\right)$ *.*

Unlike Baseline-1 (Uniform Source MAB), whose regret scales with the average variance and can be dominated by multiple , similar noisy sources, SOAR *adaptively limits the effect of high-variance sources. Similarly, Baseline-2 (Two-Phase MAB) incurs unnecessary regret in regimes where source variances are nearly identical, as it dedicates a full phase to distinguishing them. By balancing arm and source exploration,* SOAR *avoids this inefficiency. A detailed analysis of these baselines is found in Appendix D.1. Corresponding plots verifying these claims are presented in Section 5.*

5. Experiments

In this section, we compare the performance of SOAR against the proposed baselines under the canonical worst-case instances **WC-1** and **WC-2**. Arm rewards are modeled as Gaussian random variables with fixed means and source-dependent variances; shaded regions in the plots denote 95% confidence intervals. In all setups, we observe an initial linear growth in regret, scaling with M, which arises from the preprocessing phase combined with the initial exploration rounds.

Additional experiments illustrating how the regret of SOAR scales with the number of arms K and the number of sources M, along with a more detailed description of the experimental setup, are provided in Appendix E.

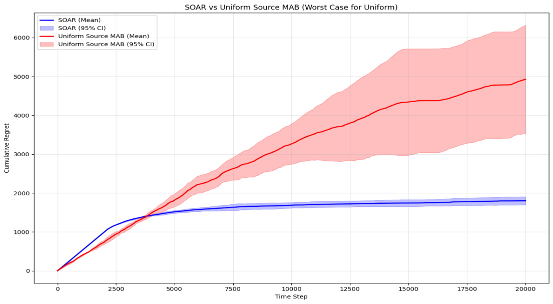

Figure 1: SOAR vs Baseline 1: WC1

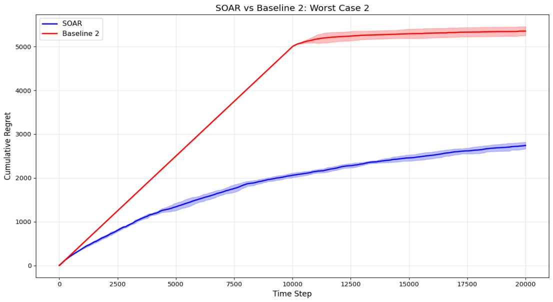

Figure 2: SOAR vs Baseline 2: WC2

6. Discussion

We studied the problem of online learning with multiple data sources under heteroscedastic noise, aiming to minimize regret through adaptive source and arm selection. Our proposed algorithm, SOAR, simultaneously explores and exploits both the data sources (with varying unknown variances) and the reward-generating arms, significantly improving over conventional baselines. Promising future directions include extending this framework to contextual bandits and reinforcement learning, and designing algorithms for non-stationary settings where source quality evolves over time.

References

[1] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics (AISTATS)*, pages 99–107, 2013.

[2] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. See Corollary 2.11.

[4] Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

[5] Victor Gabillon, Alessandro Lazaric, and Mohammad Ghavamzadeh. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3212–3220, 2012.

[6] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4105–4114, 2018.

[7] Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. *Advances in neural information processing systems*, 29, 2016.

[8] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[9] Matthias Poloczek, Jian Wang, and Peter I Frazier. Multi-information source optimization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4288–4298, 2017.

[10] Daniel Raban. Statistics 210b lecture 6 notes: Theorem 1.1 (freedman's inequality). Online lecture notes. Accessed via Pillowmath repository.

[11] Jialin Song, Yuxin Chen, and Yisong Yue. A general framework for multi-fidelity bayesian optimization with gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3158–3167. PMLR, 2019.

[12] Sho Takeno, Shou Tsutsui, Ryohei Nishihara, Masashi Onishi, Makoto Yamada, and Hayaru Shouno. Multi-fidelity bayesian optimization with max-value entropy search and its parallelization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9334–9345, 2020.

[13] Wenxin Zhou. Math 281c: Mathematical statistics — lecture 4, theorem 1.2. https://mathweb.ucsd.edu/~xip024/Teaching/Math281C_Spring2020/lect4.pdf, 2020.

Supplementary: Data Source Adaptive Online Learning under Heteroscedastic Noise

Appendix A. Problem Motivation, Baselines and Related Work

Problem Motivation: Online learning, particularly multi-armed bandit (MAB) problems, has proven foundational for decision-making in uncertain and sequential environments. A crucial yet understudied scenario within this paradigm arises when learners can access multiple heterogeneous data sources, each characterized by unknown and distinct levels of noise. Such problems are prevalent in numerous real-world applications, including clinical trials (where different hospitals or laboratories provide data with varying degrees of reliability), recommender systems (with user segments differing in the quality of their feedback), and online advertising platforms (where different advertising channels generate outcomes with heterogeneous variance levels). Successfully managing these multiple data sources while simultaneously identifying the most rewarding actions (arms) is critical for minimizing cumulative regret and maximizing overall system performance. The motivation for studying this problem lies in its practical significance and inherent complexity. Decision-makers frequently face trade-offs between cheaper yet noisy data sources and more expensive yet reliable ones.

This scenario naturally aligns with the literature on multi-fidelity learning and variance-adaptive bandits. Multi-fidelity bandits (e.g., Kandasamy et al., 2016 [7], Song et al., 2019 [11]) typically involve leveraging low-cost, low-accuracy models to accelerate learning with expensive, high-accuracy sources. While related, our setting notably differs in that we focus explicitly on adaptive exploration across multiple sources characterized by differing noise variances rather than explicit cost-accuracy trade-offs. Variance-adaptive bandits (see, e.g., Audibert et al., 2009 [2], Cowan and Katehakis, 2015 [4]) adapt the exploration rate based on observed variance within a single data-generating process. However, these existing approaches typically do not address scenarios involving multiple parallel sources with heterogeneous noise profiles, underscoring a clear gap in the literature.

Baselines: To bridge this gap, two intuitive baseline strategies naturally emerge but, unfortunately, fail dramatically in key worst-case scenarios. The first baseline selects data sources uniformly at random, effectively averaging the variance across all sources. In practice, this leads to severe performance degradation when all the sources, barring the optimal source exhibit extremely high variance—referred to as Worst-Case Instance 1 (WC-1). Here, the average variance becomes dominated by the variance of the high-variance sources, which dramatically inflates the regret. The second baseline employs a two-phase approach, initially identifying the source with the lowest variance through a dedicated exploration phase before transitioning to standard MAB strategies. However, this method becomes problematic in instances where all sources exhibit nearly identical variances—Worst-Case Instance 2 (WC-2). In WC-2, the initial exploration phase becomes redundant and costly, incurring significant regret without practical gain since distinguishing among similar-variance sources is unnecessary.

Motivating `SOAR`: To address these shortcomings, we propose `SOAR` (Source-Optimistic Adaptive Regret Minimization), a novel adaptive algorithm that jointly balances exploration and exploitation across both data sources and reward arms. SOAR begins with a preprocessing phase to eliminate high-variance sources, then employs an innovative combination of upper confidence bounds (UCB) on arm rewards and lower confidence bounds (LCB) on source variances. By dynamically adapting to observed feedback, SOAR gravitates toward reliable, low-noise sources while simultaneously identifying and exploiting the most rewarding arms. Our theoretical analysis demonstrates that SOAR attains a near-optimal regret bound of $\tilde{O}\Big(M\bar{\mu}(n+\tau)+\sum_{i=2}^{K}\frac{1}{\Delta_i}\Big)$ approaching the performance of an oracle algorithm with privileged knowledge of the best variance source. This remarkable performance is attributed to our carefully crafted confidence-bound mechanism, which enables adaptive management of heterogeneous noise with negligible regret overhead. We anticipate that this work will open up exciting new avenues in adaptive online learning, inspiring future research directions such as handling arm-dependent variances, incorporating explicit costs associated with querying various data sources, and developing algorithms that are robust to non-stationary conditions where source qualities evolve over time.

Related Work. Multi-fidelity bandits address scenarios where learners can query information at different levels of fidelity, often trading off cost and accuracy. For example, Kandasamy et al., 2016 [7] introduced the idea of leveraging inexpensive, low-fidelity evaluations to accelerate optimization with expensive, high-fidelity sources, while Song et al., 2019 [11] extended this framework to generalized black-box optimization. Further developments explored Gaussian process models [9] and batched multi-fidelity queries [12].

Variance-adaptive bandits study settings where exploration policies explicitly account for the variance of observed rewards, with a rich literature focused on reward-dependent variance scaling. Audibert et al., 2009 [2] proposed UCB-V, which adjusts exploration using empirical variance; Gabillon et al., 2012 [5], and Cowan and Katehakis, 2018 [4] further analyzed the benefits of variance-based sampling in best-arm identification. Howard et al., 2018 [6] proposed empirical Bernstein inequalities to tighten exploration bounds. Agrawal and Goyal, 2013 [1] explored variance-aware Thompson Sampling. While these works significantly improve learning under variable noise, they assume a single stream of data per arm and are not designed to reason about multiple parallel, noisy sources with arm-independent structure.

Our setting integrates and generalizes both paradigms: it learns from multiple independent sources with unknown variances while simultaneously identifying optimal reward arms. This joint optimization—across sources and arms—calls for new algorithmic and analytical tools, which we develop in SOAR. Our results demonstrate that a carefully designed LCB-UCB framework can match the performance of idealized single-source MABs, while gracefully adapting to complex heteroscedastic environments.

Appendix B. On the Role of Data Sources in Regret

While the expected regret expression

$$\mathrm{Reg}_T := \mathbf{E}\left[\sum_{t=1}^{T}\left(\mu^* - \mu_{i_t}\right)\right]$$

does not explicitly depend on the selected data sources j_t at each time step t, the choice of sources has a crucial indirect effect. Failing to quickly identify and prioritize low-variance sources can result in noisier observations, delaying accurate estimation of arm rewards and thereby increasing the difficulty of identifying the optimal arm, which in turn inflates overall regret.

Furthermore, regret minimization in this setting cannot be adequately addressed by a simple two-step approach—first identifying the lowest-variance source and then applying a standard multi-armed bandit algorithm exclusively on that source. As we discuss in detail in our discussions on the baselines in Appendix D.1, such a strategy may lead to suboptimal regret in general.

Appendix C. Supplementary for Parameter Estimates and Confidence Bounds

Remark 4 (On the Number of Sources After Preprocessing) *After the preprocessing phase, the number of surviving sources is denoted by $\tilde{M}$, where $\tilde{M} \leq M$. Since our concentration and regret guarantees are stated as upper bounds, it is sufficient to replace $\tilde{M}$ by M in the analysis. This simplification allows us to present cleaner expressions without loss of generality.*

C.1. Mean Reward Concentration

Lemma 5 (Mean-Reward Concentration) *Consider any $\delta \in (0, 1)$. For any time step $t \in [T], i \in [K]$, and any realization of $n_t(i) \geq \bar{\eta}^2 \log(KT/\delta)$, with probability at least $1 - \frac{\delta}{3}$:*

$$|\mu_i - \hat{\mu}_t(i)| \leq \frac{2\sqrt{\log(3KT/\delta) \sum_{j=1}^{M} n_t(i,j)\sigma_j^2}}{n_t(i)}$$

Proof Suppose $T_t(i) = \{s \leq t : i_s = i\}$ denotes the set of time steps at which arm i was selected up to time t. Consider the sequence of random variables defined by

$$D_{t,i} := X_{t,i} - \mu_i, \quad \text{for all } t \in T_t(i).$$

To establish a concentration bound, the key observation is that $\{D_{t,i}\}_{t \in T_t(i)}$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}_{t-1} = \sigma\left(\{X_s, i_s, j_s\}_{s=1}^{t-1}, i_t = i\right)$, generated by all arm selections and observations prior to round t. Precisely note that each $D_{t,i}$ is integrable, adapted to $\mathcal{F}_t$, and satisfies $\mathbf{E}[D_{t,i} \mid \mathcal{F}_{t-1}] = \mathbf{E}[X_{t,i} - \mu_i \mid \mathcal{F}_{t-1}] = \mu_i - \mu_i = 0$ almost surely. Thus, by construction, the sequence satisfies the martingale difference property, enabling us to leverage standard martingale concentration inequalities.

Theorem 6 (Freedman's inequality) *[10] Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence such that*

1. $\mathbb{E}[D_k|\mathcal{F}_{k-1}] = 0$
2. $D_k \leq b$ *a.s.*

Then for all $\lambda \in (0, 1/b)$ and $\delta \in (0, 1)$,

$$\mathbb{P}\left(\sum_{t=1}^{T} D_t \leq \lambda \sum_{t=1}^{T} \mathbb{E}[D_k^2|\mathcal{F}_{k-1}] + \frac{\log(1/\delta)}{\lambda}\right) \geq 1 - \delta.$$

From our problem formulation (Section 2) it is evident that $D_{t,i}$ is bounded by $\bar{\eta}$. Hence $b = \bar{\eta}$. Similarly $\mathbb{E}[D_{t,i}^2|\mathcal{F}_{t-1}] = \sigma_{j_t}^2$. Applying Freedman's inequality to $D_{t,i}$ we get

$$\mathbb{P}\left(\sum_{t\in T_i(t)} D_{t,i} \leq \lambda \sum_{t\in T_i(t)} \mathbb{E}[D_{t,i}^2|\mathcal{F}_{t-1}] + \frac{\log(3/\delta)}{\lambda}\right) \geq 1 - \frac{\delta}{3},$$

$$\mathbb{P}\left(\sum_{t\in T_i(t)} D_{t,i} \leq \lambda \sum_{j=1}^{n} n_t(i,j)\cdot\sigma_j^2 + \frac{\log(3/\delta)}{\lambda}\right) \geq 1 - \frac{\delta}{3},$$

Optimizing over λ to find the least upper bound for $D_{t,i}$ gives us

$$\lambda^* = \sqrt{\frac{\log(3/\delta)}{\sum_{j=1}^{n} n_t(i,j)\cdot\sigma_j^2}},$$

$$\implies \mathbb{P}\left(\sum_{t\in T_i(t)} D_{t,i} \leq 2\sqrt{\log\left(\frac{3}{\delta}\right)\sum_{j=1}^{n} n_t(i,j)\sigma_j^2}\right) \geq 1 - \frac{\delta}{3},$$

A union bound across K arms and T time-steps gives us

$$\implies \mathbb{P}\left(\sum_{t\in T_i(t)} D_{t,i} \leq 2\sqrt{\log\left(\frac{3KT}{\delta}\right)\sum_{j=1}^{n} n_t(i,j)\sigma_j^2}\right) \geq 1 - \frac{\delta}{3},$$

Dividing both sides by $n_t(i)$ we get our mean concentration bound w.p $1 - \frac{\delta}{3}$

$$\left|\hat{\mu}_t(i) - \mu_i\right| \leq \frac{2\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_t(i,j)\sigma_j^2}}{n_t(i)},$$

Additionally we need to make sure $n_t(i)$ is large enough so $\lambda^* < \frac{1}{\bar{\eta}}$

$$\lambda^* = \sqrt{\frac{\log(3/\delta)}{\sum_{j=1}^{n} n_t(i,j)\cdot\sigma_j^2}} \leq \frac{1}{\bar{\eta}},$$

$$n_t(i) \geq \bar{\eta}^2\log(3/\delta) \;\; [\text{Using } \sigma^* = 1].$$

A union bound over all timesteps T and arms K gives us, $n_t(i) \geq \bar{\eta}^2\log(3KT/\delta)$. ■

C.2. Reward Variance Concentration

Lemma 7 (Reward Variance Concentration) *For any $j \in [M], t \in [T]$ and realization of $m_t(j) > \frac{32M^2K^2\left(\log(3KT/\delta)\right)^2}{\bar{\eta}^2\log(12MT/\delta)}$ with probability at least $1 - \frac{\delta}{3}$ and $\epsilon' < 27$ (where ϵ' is a parameter in Bernstein's Inequality [3][13]) we have:*

$$\left|\hat{\sigma}_t^2(j) - \sigma_j^2\right| \;\leq\; 2\bar{\eta}\sigma_j\sqrt{\frac{2\log(12MT/\delta)}{m_t(j)}}.$$

Proof Suppose $\mathcal{T}_t(j) = \{s \le t : j_s = j\}$ denotes the set of time steps at which the source j was selected up to time $t \in [T]$.

$$\begin{aligned}
\left|\hat{\sigma}_t^2(j) - \sigma_j^2\right| &= \left|\frac{1}{m_t(j)} \sum_{i=1}^{K} \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}\mathbb{1}\{j_t = j\}\big(X_s - \hat{\mu}_t(i)\big)^2 - \sigma_j^2\right| \\
&= \left|\frac{1}{m_t(j)} \sum_{i=1}^{K} \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}\mathbb{1}\{j_t = j\}\big(X_s - \mu_i + \mu_i - \hat{\mu}_t(i)\big)^2 - \sigma_j^2\right| \\
&\le \left|\frac{2}{m_t(j)} \sum_{i=1}^{K} \sum_{s=1}^{t} \mathbb{1}\{i_s = i\}\mathbb{1}\{j_t = j\}\big(X_s - \mu_i\big)^2 + \frac{2}{m_t(j)} \sum_{i=1}^{K} n_t(i,j)(\mu_i - \hat{\mu}_t(i))^2 - \sigma_j^2\right|, \\
&\le \underbrace{2\left|\frac{1}{m_t(j)} \sum_{s \in \mathcal{T}_t(j)} \big(X_s - \mu_{i_s}\big)^2 - \sigma_j^2\right|}_{I} + \underbrace{\left|\frac{2}{m_t(j)} \sum_{i=1}^{K} n_t(i,j)(\mu_i - \hat{\mu}_t(i))^2\right|}_{II},
\end{aligned}$$

where the second last inequality applied $(a+b)^2 \le 2(a^2+b^2)$ and the last inequality uses that $|a+b| \le |a| + |b|$, $\forall a, b \in \mathbb{R}$. The remaining task is to find a tight upper bound for each of the above terms, I and II, separately.

Bounding Term I using Bernstein's Inequality Consider $Z_s = (X_s - \mu_{i_s})^2$, where recall $X_s = \mu_{i_s} + \varepsilon_j$ s.t. $\varepsilon \sim \mathcal{D}(j)$ is a random noise draw from an (unknown) underlying noise distribution $\mathcal{D}(j)$ of the selected data-source $j \in [M]$. Hence $\left|Z_s\right| \le \bar{\eta}^2$. Now

$$V = \sum_{s \in \mathcal{T}_t(j)} \mathbb{E}[Z_s^2] \le \sum_{s \in \mathcal{T}_t(j)} \bar{\eta}^2 E[Z_s] \le m_t(j)\bar{\eta}^2\sigma_j^2.$$

Theorem 8 (Bernstein's Inequality) *[3, Cor. 2.11]; [13, Theorem 1.2] Suppose $Z_1, \ldots, Z_n$ are independent random variables with finite variances, and suppose that*

$$\max_{1 \le k \le n} |Z_k| \le b$$

almost surely for some constant $b > 0$. Let

$$V = \sum_{k=1}^{n} \mathbb{E}[Z_k^2].$$

Then, for every $t \ge 0$,

$$\Pr\left(\sum_{k=1}^{n} (Z_k - \mathbb{E}Z_k) \ge t\right) \;\le\; \exp\left(-\frac{t^2}{2\big(V + \frac{1}{3}bt\big)}\right),$$

and

$$\Pr\left(\sum_{k=1}^{n} (Z_k - \mathbb{E}Z_k) \le -t\right) \;\le\; \exp\left(-\frac{t^2}{2\big(V + \frac{1}{3}bt\big)}\right).$$

Applying Bernstein's to Z_s we get:

$$\Pr\left(\Big|\sum_{s\in\mathcal{T}_t(j)}\Big((X_s-\mu_{i_s})^2-\sigma_j^2\Big)\Big|\geq\epsilon\right)\ \leq\ 2\exp\left(-\frac{\epsilon^2}{2\big(m_t(j)\bar{\eta}^2\sigma_j^2+\frac{1}{3}\bar{\eta}^2\epsilon\big)}\right),$$

$$\equiv\Pr\left(\Big|\frac{1}{m_t(j)}\sum_{s\in\mathcal{T}_t(j)}(X_s-\mu_{i_s})^2-\sigma_j^2\Big|\geq\frac{\epsilon}{m_t(j)}\right)\ \leq\ 2\exp\left(-\frac{\epsilon^2}{2\big(m_t(j)\bar{\eta}^2\sigma_j^2+\frac{1}{3}\bar{\eta}^2\epsilon\big)}\right),$$

$$\implies\ \Pr\left(\Big|\frac{1}{m_t(j)}\sum_{s\in\mathcal{T}_t(j)}(X_s-\mu_{i_s})^2-\sigma_j^2\Big|\geq\epsilon'\right)\ \leq\ 2\exp\left(-\frac{m_t(j)\epsilon'^2}{2\big(\bar{\eta}^2\sigma_j^2+\frac{1}{3}\bar{\eta}^2\epsilon'\big)}\right).\quad\left[\epsilon'=\frac{\epsilon}{m_t(j)}\right]$$

Choose ϵ' such that $\dfrac{\sigma_j^2}{\epsilon'^2}>\dfrac{1}{3\epsilon}\implies\epsilon'<3\sigma_j^2$

Then if $\dfrac{\delta}{6}\geq2\exp\left(-\dfrac{m_t(j)\epsilon'^2}{2\big(\bar{\eta}^2\sigma_j^2\big)}\right)\implies\epsilon'\leq\bar{\eta}\sigma_j\sqrt{\dfrac{2\log(12/\delta)}{m_t(j)}}$

Taking a union bound over all sources M and all steps T we have with probability at least $1-\dfrac{\delta}{6}$

$$\Big|\frac{1}{m_t(j)}\sum_{s\in\mathcal{T}_t(j)}(X_s-\mu_{i_s})^2-\sigma_j^2\Big|\leq\bar{\eta}\sigma_j\sqrt{\frac{2\log(12MT/\delta)}{m_t(j)}}$$

Bounding Term II using the Mean concentration derived in Appendix C.1

$$(\hat{\mu}_t(i)-\mu_i)^2\leq\frac{4\log(3KT/\delta)\sum_{j=1}^M n_t(i,j)\sigma_j^2\cdot}{n_t(i)^2}$$

This further implies with probability at least $\left(1-\dfrac{\delta}{6}\right)$:

$$\begin{aligned}\frac{2}{m_t(j)}\sum_{i=1}^K n_t(i,j)(\mu_i-\hat{\mu}_t(i))^2&\leq\frac{8}{m_t(j)}\sum_{i=1}^K\frac{n_t(i,j)}{n_t(i)^2}\Big(\sum_{j=1}^M n_t(i,j)\sigma_j^2\Big)\log(3KT/\delta),\\&\leq\frac{8}{m_t(j)}\sum_{i=1}^K\frac{n_t(i,j)}{n_t(i)}\sum_{j=1}^M\sigma_j^2\log(3KT/\delta),\quad\left[\text{since }\frac{n_t(i,j)}{n_t(i)}\leq1\right]\\&=\frac{8}{m_t(j)}\sum_{j=1}^M\sigma_j^2\log(3KT/\delta)\sum_{i=1}^K\frac{n_t(i,j)}{n_t(i)}\leq\frac{8K\log(3KT/\delta)}{m_t(j)}\sum_{j=1}^M\sigma_j^2.\end{aligned}$$

Putting it all together: We have with probability $1-\dfrac{\delta}{3}$ and $\epsilon'<3\sigma_j^2$

$$\left|\hat{\sigma}_t^2(j)-\sigma_j^2\right|\leq\bar{\eta}\sigma_j\sqrt{\frac{2\log(12MT/\delta)}{m_t(j)}}+\frac{8K\log(3KT/\delta)}{m_t(j)}\sum_{j=1}^M\sigma_j^2,$$

$$\le \underbrace{\bar{\eta}\sigma_j\sqrt{\frac{2\log(12MT/\delta)}{m_t(j)}}}_{P} + \underbrace{\frac{8MK\log(3KT/\delta)}{m_t(j)}\sigma_{j_{max}}^2}_{Q}.$$

Pick $m_t(j)$ such that $P > Q$ then bounding $P + Q$ by $2P$

$$P > Q \implies \bar{\eta}\sigma_j\sqrt{\frac{2\log(12MT/\delta)}{m_t(j)}} > \frac{8MK\log(3KT/\delta)}{m_t(j)}\sigma_{j_{max}}^2,$$

$$\implies m_t(j) > \frac{288M^2K^2}{\bar{\eta}^2}\frac{(\log(3KT/\delta))^2}{\log(12MT/\delta)}.$$

Since the preprocessing step discards all sources with variance greater than 9, and given that the optimal source has variance $\sigma_*^2 = 1$, it follows with high probability that the maximum variance among the retained sources satisfies $\sigma_{j_{\max}}^2 \le 9$. Substituting this into our bound yields a bound of order $\tilde{O}\left(\frac{M^2K^2}{\bar{\eta}^2}\right)$ where $\tilde{O}(\cdot)$ suppresses logarithmic factors.

Hence we have with probability $1 - \frac{\delta}{6}$, $\epsilon' < 27$ and $m_t(j) \ge \frac{288M^2K^2\big(\log(3KT/\delta)\big)^2}{\bar{\eta}^2\log(12MT/\delta)}$ that

$$\left|\hat{\sigma}_t^2(j) - \sigma_j^2\right| \le 2\bar{\eta}\sigma_j\sqrt{\frac{2\log(12MT/\delta)}{m_t(j)}}.$$

■

C.2.1. BOUNDING VARIANCE AND VARIANCE ESTIMATES USING OUR VARIANCE CONCENTRATION

Corollary 9 (Variance Sandwiching) *Consider any $\delta \in (0,1)$. For any time step $t \in [T]$ and source $j \in [M]$, choosing $m_t(j) \ge 32\bar{\eta}^2\log(12MT/\delta)$, with probability at least $1 - \frac{\delta}{3}$ we have*

$$\sigma_j^2 \le 2\hat{\sigma}_t^2(j) \le 3\sigma_j^2.$$

Proof Choosing $m_t(j)$ such that $m_t(j) = \frac{32\bar{\eta}^2\log(12MT/\delta)}{\sigma_j^2}$

Our variance concentration now becomes

$$\left|\hat{\sigma}_t^2(j) - \sigma_j^2\right| \le \frac{\sigma_j^2}{2},$$

$$\frac{\sigma_j^2}{2} \le \hat{\sigma}_t^2(j) \le \frac{3\sigma_j^2}{2},$$

$$\sigma_j^2 \le 2\hat{\sigma}_t^2(j) \le 3\sigma_j^2.$$

Now since we know that $\sigma^* = 1$, $\frac{32\bar{\eta}^2\log(12MT/\delta)}{\sigma_j^2}$ can be further upper bounded by $32\bar{\eta}^2\log(12MT/\delta)$. ■

Corollary 10 (Reward UCB) *Consider any $\delta \in (0,1)$. At any time step $t \in [T]$ and arm $i \in [K]$, with probability at least $1 - \frac{\delta}{3}$:*

$$\mu_i \leq \mathrm{UCB}_t^{\mu}(i)$$

Proof The proof of the above result directly follows from the mean reward concentration defined in Lemma 5, our definition of $\mathrm{UCB}_t^{\mu}(i)$ from Section 3.1 and an application of Corollary 9 ■

Corollary 11 (Variance UCB) *Consider any $\delta \in (0,1)$. At any time step $t \in [T]$ and source $j \in [M]$, with probability at least $1 - \frac{\delta}{3}$:*

$$\sigma_j \geq \mathrm{LCB}_t^{\sigma}(j)$$

Proof The proof of the above result directly follows from the mean reward concentration defined in Lemma 7, our definition of $\mathrm{LCB}_t^{\mu}(i)$ from Section 3.1 and an application of Corollary 9 ■

C.3. Confidence Bounds for Mean and Variance Estimators

In this subsection of the appendix, we discuss the confidence bounds used in our algorithm for both the mean reward estimates and the variance of the data sources. These bounds guide the arm and source selection rules in Algorithm 2.

Upper Confidence Bound on Mean Reward. For each arm $i \in [K]$, we define an upper confidence bound (UCB) on its empirical mean reward $\hat{\mu}_t(i)$ as:

$$\mathrm{UCB}_t^{\mu}(i) \coloneqq \hat{\mu}_t(i) + \frac{2\sqrt{2\log(3KT/\delta)\sum_{j=1}^{M} n_t(i,j)\bar{\sigma}_t^2(j)}}{n_t(i)}, \tag{4}$$

where $n_t(i,j)$ denotes the number of samples observed for arm i from source j up to time t, and $n_t(i) = \sum_{j=1}^{M} n_t(i,j)$ is the total number of samples for arm i.

Here, $\bar{\sigma}_t^2(j)$ represents an upper confidence estimate on the variance of source j (to be defined below). Intuitively, this bound reflects a weighted measure of uncertainty in the estimated mean, where the contribution of each source is scaled by the number of times it has been used to sample arm i and the confidence in its variance. Higher variance or fewer samples leads to a wider confidence interval.

Upper Confidence Bound on Source Variance. For each data source $j \in [M]$, we estimate its variance using the empirical estimator $\hat{\sigma}_t^2(j)$, and construct the following upper confidence bound:

$$\bar{\sigma}_t^2(j) = \mathrm{UCB}_t^{\sigma}(j) \coloneqq \hat{\sigma}_t^2(j) + 2\bar{\eta}\hat{\sigma}_j\sqrt{\frac{4\log(12MT/\delta)}{m_t(j)}} \tag{5}$$

where $m_t(j)$ is the total number of times source j has been selected up to time t.

This UCB accounts for both the sampling variability within source j and the interaction between arms and sources, capturing how the variance of other sources (via the nested sum) impacts our confidence in source j's reliability. As is intuitive, the bound grows with the data source variance and shrinks with more observations. One can similarly define a lower confidence bound of the variance estimator given by:

Lower Confidence Bound on Source Variance. We also define a lower confidence bound (LCB) on the variance of each source:

$$\underline{\sigma}_t^2(j) = \mathrm{LCB}_t^\sigma(j) := \hat{\sigma}_t^2(j) - 2\bar{\eta}\hat{\sigma}_j\sqrt{\frac{4\log(12MT/\delta)}{m_t(j)}} \tag{6}$$

Appendix D. Supplementary Material on the SOAR Algorithm

D.1. Some Standard Baselines and Limitations

We now discuss two natural baseline algorithms and highlight their limitations, motivating the design of our simultaneous source-arm exploration strategy.

Baseline-1. Uniform Source MAB. This baseline selects a data source uniformly at random at each round t and uses the corresponding observations to update reward estimates and apply standard multi-armed bandit (MAB) routines. Under this uniform selection strategy, each arm $i \in [K]$ experiences "*an effective variance*" equal to the average variance across all sources: $\tilde{\sigma}^2 := \frac{1}{M}\sum_{j=1}^{M}\sigma_j^2$.

Regret Analysis of Baseline-1: Following standard upper confidence bound (UCB)-based analyses for MAB algorithms [8], the regret for this uniform-source MAB algorithm can be bounded as $O\left(\tilde{\sigma}\sqrt{KT}\right)$, or, more specifically, using instance-dependent regret bounds as $O\left(\tilde{\sigma}^2\sum_{i\neq i^*}\frac{\log(MKT)}{\Delta_i^\mu}\right)$. However, this uniform averaging approach can be gravely suboptimal in certain instances. Consider a worst-case scenario (call this *worst-case instance 1 (WC-1)*) where all sources except the optimal source have a similar, significantly high variance say σ_{max}^2.
In this case, the effective variance $\tilde{\sigma}^2$ is dominated by the multiple high variance sources: $\tilde{\sigma}^2 = \frac{(M-1)\sigma_{\text{max}}^2}{M} \approx \sigma_{\text{max}}^2$. Consequently, the instance dependent regret bound deteriorates to $O\left(\sigma_{\text{max}}^2\sum_{i\neq i^*}\frac{\log(MKT)}{\Delta_i^\mu}\right)$ while the instance independent bound becomes $O(\sigma_{\text{max}}\sqrt{KT})$. Thus, the learner incurs undesirably high regret, as the estimation accuracy is effectively governed by the extremely noisy non-optimal sources.

Baseline-2. Two-Phase (Source-then-Arm) MAB. In this baseline, the learner first attempts to identify the data source with the smallest variance by fixing a single arm and running a best-arm identification (BAI) algorithm across the sources. After this initial phase, the learner commits exclusively to the identified "good" variance source and runs a standard MAB algorithm on the K arms, querying arm rewards solely from this selected source.

Regret Analysis of Baseline-2: To analyze the regret incurred by this two-phase strategy, we note that the first phase (source identification) incurs a regret of approximately $O\left(\sum_{j\neq j^*}\frac{\bar{\mu}\log(MKT)}{\max\{\epsilon^2,(\Delta_j^{\sigma^2})^2\}}\right)$, where $\epsilon \in \mathbb{R}_+$ is a user-specified error tolerance parameter determining the desired precision (PAC-optimality) in identifying the minimal-variance source. Importantly, during each round of the first phase, the algorithm potentially incurs $O(1)$ regret (which in turn is bounded by $\bar{\mu}$) since it continues to query a fixed arm and does not make progress in identifying the optimal reward arm $i^* \in [K]$.

In the second phase (reward-arm identification and exploitation), the algorithm's regret can be bounded as $O\left((\sigma^* + \epsilon)^2 \sum_{i \neq i^*} \frac{\log(MKT)}{(\Delta_i^{\mu})}\right)$, leading to an overall regret (summing two phases) of:

$$O\left(\sum_{j \neq j^*} \frac{\bar{\mu}\log(MKT)}{\max\{\epsilon^2, (\Delta_j^{\sigma^2})^2\}} + (\sigma^* + \epsilon)^2 \sum_{i \neq i^*} \frac{\log(MKT)}{\Delta_i^{\mu}}\right).$$

We remark that ϵ is a tunable parameter whose choice significantly impacts the performance of this approach. However, it is hard for the learner to optimize the value of ϵ without the knowledge of the variance gaps $\{\Delta_j^{\sigma^2}\}_{j \in [M]}$.

Consider now the worst-case scenario for this two-phase algorithm (call this *worst-case instance 2 (WC-2)*). This occurs when all data sources have roughly identical variances :

$$\sigma_1^2 \approx \sigma_2^2 \approx \cdots \approx \sigma_M^2.$$

In this regime all variance gaps satisfy $\Delta_j^{\sigma^2} \approx 0$, so setting $\epsilon = 0$ is particularly disastrous: the factor $1/\max\{\epsilon^2, (\Delta_j^{\sigma^2})^2\}$ in our regret bound effectively blows up, and the algorithm is forced to keep sampling all sources in an attempt to detect non-existent variance differences, leading to an arbitrarily poor bound.

In such scenarios, the learner could have selected any data source (or even randomly selected sources at each round, as in the uniform source baseline described above), since there is no practical benefit in spending effort to distinguish among equally good sources. Yet, the two-phase algorithm will unnecessarily incur substantial regret in the initial source-identification phase while trying to pinpoint the optimal (lowest-variance) source, accumulating a high regret of roughly $O\left(\sum_{j \neq j^*} \frac{\bar{\mu}\log(MKT)}{\max\{\epsilon^2, (\Delta_j^{\sigma^2})^2\}}\right)$.

Here, note that $\Delta_j^{\sigma^2} \approx 0$ for all sources j, which exacerbates the incurred regret. Moreover, since the learner does not initially know the minimal source variance σ^*, choosing an appropriate ϵ to balance this trade-off and minimize regret becomes challenging, leading to additional practical issues.

This dependence on ϵ becomes clearer when we look at two extreme choices. Suppose the learner naively fixes $\epsilon = c_1 > 0$ independently of the instance, and consider a WC-2 instance in which all source variances are nearly identical, so that $\Delta_j^{\sigma^2} \approx 0$ for all $j \neq j^*$. Then, for each such source, $\max\{\epsilon^2, (\Delta_j^{\sigma^2})^2\} \approx \epsilon^2 = c_1^2$, and the corresponding contribution to the regret satisfies

$$\sum_{j \neq j^*} \frac{\bar{\mu}\log(MKT)}{\max\{\epsilon^2, (\Delta_j^{\sigma^2})^2\}} \approx \sum_{j \neq j^*} \frac{\bar{\mu}\log(MKT)}{c_1^2},$$

which can be made arbitrarily large if c_1 is very small.
Conversely, if the learner picks ϵ overly large, then the term

$$(\sigma^* + \epsilon)^2 \sum_{i \neq i^*} \frac{\log(MKT)}{\Delta_i^{\mu}},$$

is dominated by ϵ^2, and the regret again scales poorly. Thus any fixed, instance-independent choice $\epsilon = c_1$ can be far from optimal across different problem instances.

These analyses reveal crucial limitations in both approaches, highlighting that an optimal strategy must *simultaneously* balance exploration and exploitation across data sources and reward arms. A well-designed algorithm must dynamically and adaptively explore low-variance sources, while simultaneously identifying the best-performing arms as quickly as possible to minimize overall regret.

D.2. Analysis of the Preprocessing Algorithm

In this subsection, we derive the concentration bounds that underpin the preprocessing algorithm. Specifically, we establish probabilistic guarantees on the accuracy of variance estimates obtained from a finite number of queries to each source. These results ensure that the constructed lower and upper confidence bounds ($\mathrm{LCB}_n^\sigma(j)$, $\mathrm{UCB}_n^\sigma(j)$) concentrate tightly around the true variance of each source with high probability.

Furthermore, we provide the conditions under which the algorithm successfully eliminates high-variance sources. In particular, we show that if a source satisfies $\sigma_j^2/\sigma_*^2 > c$ for some constant $c > 1$, then, with probability at least $\left(1 - \frac{\delta}{3}\right)$, it will be removed during preprocessing. For clarity of exposition, we set $c = 9$ in the main text, though our analysis extends to any fixed constant. These results formally justify the effectiveness of the preprocessing phase in reducing the set of candidate sources to those with variance close to the optimal one.

We begin by stating the variance concentration for preprocessing which is at the core of our analysis.

D.2.1. VARIANCE CONCENTRATION FOR PREPROCESSING

Theorem 12 (Preprocessing Variance Concentration.) *Given $\epsilon'' < \min\left\{6\sigma_j^2, \frac{18\sigma_j^4}{\bar{\eta}^2}\right\}$ (where ϵ'' is a parameter in Bernstein's Inequality [3][13]) and $\delta \in (0,1)$, for all sources $j \in [M]$ we have, with probability at least $1 - \dfrac{\delta}{3}$,*

$$\left|\hat{\sigma}_n^2(j) - \sigma_j^2\right| \;\le\; 8\bar{\eta}\,\sigma_j\sqrt{\frac{\log(4M/\delta)}{n}}.$$

Proof

$$\begin{aligned}
\left|\hat{\sigma}_n^2(j) - \sigma_j^2\right| &= \left|\frac{1}{n}\sum_{k=1}^{n}\big(X_k - \hat{\mu}_n(i)\big)^2 - \sigma_j^2\right| \\
&= \left|\frac{1}{n}\sum_{k=1}^{n}\big(X_k - \mu_i + \mu_i - \hat{\mu}_n(i)\big)^2 - \sigma_j^2\right| \\
&\le \left|\frac{2}{n}\sum_{k=1}^{n}\big(X_k - \mu_i\big)^2 + \frac{2}{n}\sum_{k=1}^{n}(\mu_i - \hat{\mu}_n(i))^2 - \sigma_j^2\right|, \\
&\le \underbrace{\left|\frac{2}{n}\sum_{k=1}^{n}\big(X_k - \mu_i\big)^2 - \sigma_j^2\right|}_{\text{(I)}\le\epsilon/2} + \underbrace{\left|\,\frac{2}{n}\sum_{k=1}^{n}(\mu_i - \hat{\mu}_n(i))^2\,\right|}_{\text{(II)}\le\epsilon/2}
\end{aligned}$$

Bounding Term (I) : To bound Term I we will use Bernstein's Inequality.

Theorem 8 (Bernstein's Inequality) *[3, Cor. 2.11]; [13, Theorem 1.2] Suppose $Z_1, \dots, Z_n$ are independent random variables with finite variances, and suppose that*

$$\max_{1\le k\le n} |Z_k| \le b$$

almost surely for some constant $b > 0$. Let

$$V = \sum_{k=1}^{n} \mathbb{E}[Z_k^2].$$

Then, for every $t \ge 0$,

$$\Pr\left(\sum_{k=1}^{n}(Z_k - \mathbb{E}Z_k) \ge t\right) \;\le\; \exp\left(-\frac{t^2}{2\left(V + \frac{1}{3}bt\right)}\right),$$

and

$$\Pr\left(\sum_{k=1}^{n}(Z_k - \mathbb{E}Z_k) \le -t\right) \;\le\; \exp\left(-\frac{t^2}{2\left(V + \frac{1}{3}bt\right)}\right).$$

Fix a source $j \in [M]$ and arm $i \in [K]$. Consider $Z_k = (X_k - \mu_i)^2$, where $X_k = \mu_i + \varepsilon$ and $\varepsilon \sim \mathcal{D}(j)$ is drawn from an (unknown) underlying noise distribution $\mathcal{D}(j)$ of the selected data-source $j \in [M]$.From our problem setting (Section 2), it is evident $\Big|Z_k\Big| \le \bar{\eta}^2$. Now

$$V = \sum_{k=1}^{n} \mathbb{E}[Z_k^2] \le \sum_{k=1}^{n} \bar{\eta}^2 E[Z_k] \le n\bar{\eta}^2\sigma_j^2$$

Applying Bernstein's, we have w.p $\left(1 - \dfrac{\delta}{6}\right)$:

$$\begin{aligned}
\Pr\left(\Big|\sum_{k=1}^{n}\Big((X_k - \mu_i)^2 - \sigma_j^2\Big)\Big| \ge \epsilon\right) \;&\le\; 2\exp\left(-\frac{\epsilon^2}{2\left(n\bar{\eta}^2\sigma_j^2 + \frac{1}{3}\bar{\eta}^2\epsilon\right)}\right), \\
\equiv \Pr\left(\Big|\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_i)^2 - \sigma_j^2\Big| \ge \frac{\epsilon}{n}\right) \;&\le\; 2\exp\left(-\frac{\epsilon^2}{2\left(n\bar{\eta}^2\sigma_j^2 + \frac{1}{3}\bar{\eta}^2\epsilon\right)}\right), \\
\implies \Pr\left(\Big|\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_i)^2 - \sigma_j^2\Big| \ge \epsilon'\right) \;&\le\; 2\exp\left(-\frac{n^2\epsilon'^2}{2\left(n\bar{\eta}^2\sigma_j^2 + \frac{1}{3}\bar{\eta}^2 n\epsilon'\right)}\right), \quad \left[\epsilon' = \frac{\epsilon}{n}\right]
\end{aligned}$$

Replace ϵ' by $\epsilon''/2$

$$\Pr\left(\Big|\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_i)^2 - \sigma_j^2\Big| \ge \frac{\epsilon''}{2}\right) \;\le\; 2\exp\left(-\frac{n\epsilon''^2}{8\left(\bar{\eta}^2\sigma_j^2 + \frac{1}{6}\bar{\eta}^2\epsilon\right)}\right).$$

Choose ϵ'' such that $\dfrac{\sigma_j^2}{\epsilon''^2} > \dfrac{1}{6\epsilon}$ i.e $\epsilon'' < 6\sigma_j^2$

Hence $\dfrac{\delta}{6} \geq 2\exp\left(-\dfrac{n\epsilon''^2}{16\bar{\eta}^2\sigma_j^2}\right) \implies \epsilon'' \leq 4\bar{\eta}\sigma_j\sqrt{\dfrac{\log(12/\delta)}{n}}$.

Thus with $\epsilon'' < 6\sigma_j^2$, and a union bound over all sources $j \in [M]$ we have with probability $\left(1 - \dfrac{\delta}{6}\right)$

$$\Big|\frac{2}{n}\sum_{k=1}^{n}(X_k - \mu_i)^2 - \sigma_j^2\Big| \leq 4\bar{\eta}\sigma_j\sqrt{\frac{\log(12M/\delta)}{n}}.$$

Bounding Term (II): Consider $D_k = (X_k - \mu_i)^2$, where $X_k = \mu_i + \varepsilon$ and $\varepsilon \sim \mathcal{D}(j)$ is drawn from an (unknown) underlying noise distribution $\mathcal{D}(j)$ of the selected data-source $j \in [M]$.From our problem setting (Section 2), it is evident $\Big|D_k\Big| \leq \bar{\eta}$.
Additionally conditional variance $V = \sum_{k=1}^{n} E[D_k^2] = n \cdot \sigma_j^2$
Applying Bernstein's we get with probability $\delta/6$:

$$\Pr\left(\frac{1}{n}\Big|\sum_{k=1}^{n}\Big(X_k - \mu_i\Big)\Big| \geq \epsilon'/2\right) \;\leq\; 2\exp\left(-\frac{n\epsilon'^2}{8\big(\sigma_j^2 + \frac{1}{6}\bar{\eta}\epsilon'\big)}\right), \quad \Big[\epsilon' = \frac{\epsilon}{n}\Big]$$

$$\implies \Pr\Big(\Big|\hat{\mu}_n(i) - \mu_i\Big| \geq \epsilon'/2\Big) \;\leq\; 2\exp\left(-\frac{n\epsilon'^2}{8\big(\sigma_j^2 + \frac{1}{6}\bar{\eta}\epsilon'\big)}\right),$$

$$\implies \Pr\left(\Big|\hat{\mu}_n(i) - \mu_i\Big|^2 \geq \epsilon'^2/4\right) \;\leq\; 2\exp\left(-\frac{n\epsilon'^2}{8\big(\sigma_j^2 + \frac{1}{6}\bar{\eta}\epsilon'\big)}\right),$$

Choose $\epsilon'' = \epsilon'^2/2 \implies \epsilon' = \sqrt{2\epsilon''}$

$$\Pr\left(\Big|\hat{\mu}_n(i) - \mu_i\Big|^2 \geq \epsilon''/2\right) \;\leq\; 2\exp\left(-\frac{n\epsilon''}{4\big(\sigma_j^2 + \frac{\sqrt{2}}{6}\bar{\eta}\sqrt{\epsilon''}\big)}\right).$$

Choosing ϵ such that $\dfrac{\sigma_j^2}{\epsilon} > \dfrac{\sqrt{2}\bar{\eta}}{6\sqrt{\epsilon}} \implies \epsilon < \dfrac{18\sigma_j^4}{\bar{\eta}^2}$

Additionally we have $\epsilon'' \leq \dfrac{8}{n}\sigma_j^2\log(12/\delta)$

Thus with $\epsilon'' < \dfrac{18\sigma_j^4}{\bar{\eta}^2}$ and a union bound over all sources $j \in [M]$ we have with probability $\left(1 - \dfrac{\delta}{6}\right)$

$$\Pr\left(\Big|\hat{\mu}_n(i) - \mu_i\Big|^2 \leq \frac{4}{n}\sigma_j^2\log(12M/\delta)\right).$$

Putting it all together:

$$\begin{aligned}\Big|\hat{\sigma}_n^2(j) - \sigma_j^2\Big| \leq& \Big|\frac{2}{n}\sum_{k=1}^{n}(X_k - \mu_i)^2 - \sigma_j^2\Big| + \Big|\frac{2}{n}\sum_{k=1}^{n}(\mu_i - \hat{\mu}_n(i))^2\Big| \\ \leq& \underbrace{4\bar{\eta}\sigma_j\sqrt{\frac{\log(12M/\delta)}{n}}}_{C} + \underbrace{\frac{8}{n}\sigma_j^2\log(12M/\delta)}_{D}\end{aligned}$$

We can choose the number of time each source is queried (i.e n) such that $C > D$ and then bound the expression $C + D$ by $2C$. This is achieved when

$$n > \frac{4\sigma_j^2 \log(12M/\delta)}{\bar{\eta}^2}$$

Now from our problem setup $\sigma_j^2 \leq \bar{\eta}^2$. Therefore the greatest lower bound for n is $n > 4\log(12M/\delta)$ which is almost always true for any valid M. Additionally this condition on n can be absorbed by our stopping condition which is $16\bar{\eta}^4 \log(12M/\delta)$ for $\bar{\eta} \geq 1$. Hence we have with probability $1 - \frac{\delta}{3}$:

$$\left|\hat{\sigma}_n^2(j) - \sigma_j^2\right| \leq 8\bar{\eta}\sigma_j \sqrt{\frac{\log(12M/\delta)}{n}}$$

■

D.2.2. CONFIDENCE BOUNDS ON THE VARIANCE ESTIMATOR FOR PREPROCESS:

From the above concentration we can derive the Upper Confidence and Lower Confidence Bound on the Source Variance during PREPROCESS.

Upper Confidence Bound on Source Variance during PREPROCESS: For each source $j \in [M]$, we estimate its variance during PREPROCESS with runtime budget n using the empirical estimator $\hat{\sigma}_n^2(j)$ and construct the following upper confidence bound:

$$\bar{\sigma}_n^2(j) = \mathrm{UCB}_n^\sigma(j) := \hat{\sigma}_n^2(j) + 8\bar{\eta}^2 \sqrt{\frac{\log(12M/\delta)}{n}} \tag{7}$$

Lower Confidence Bound on Source Variance during PREPROCESS: We also construct a Lower Confidence Bound (LCB) on the variance of each source during PREPROCESS.

$$\underline{\sigma}_n^2(j) = \mathrm{LCB}_n^\sigma(j) := \max\left\{\hat{\sigma}_n^2(j) - 8\bar{\eta}^2 \sqrt{\frac{\log(12M/\delta)}{n}}, 0\right\} \tag{8}$$

D.2.3. DERIVING THE ELIMINATION CONDITION

Theorem 1 (Stopping Condition of PREPROCESS) *Consider any $\delta \in (0,1)$. If* PREPROCESS *is run with runtime budget n, where $n > 16\bar{\eta}^4 \log(12M/\delta)$, then any source $j \in [M]$ with variance $\sigma_j^2 > 9\sigma^{*2}$ will be eliminated with probability* $(1 - \delta/3)$.

Proof For the purposes of this proof we will be using the variance concentration defined above in Theorem 12.
For the preprocessing algorithm to proceed normally, we need to find n such that,

$$\mathrm{LCB}_n^\sigma(\sigma_j^2) \leq \mathrm{UCB}_n^\sigma(\sigma^{*2})$$
$$\hat{\sigma}_n^2(j) - \mathrm{conf}_j^\sigma(n) \leq \hat{\sigma}_n^2(j^*) + \mathrm{conf}_{j*}^\sigma(n)$$
$$\sigma_j^2 - 2\mathrm{conf}_j^\sigma(n) \leq \hat{\sigma}_n^2(j) - \mathrm{conf}_j^\sigma(n) \leq \hat{\sigma}_n^2(j^*) + \mathrm{conf}_{j*}^\sigma(n) \leq \sigma^2(j^*) + 2\mathrm{conf}_{j*}^\sigma(n)$$

$$\begin{aligned}
\sigma_j^2 - 2\text{conf}_j^\sigma(n) &\le \sigma^2(j^*) + 2\text{conf}_{j*}^\sigma(n) \\
\sigma_j^2 - \sigma^2(j^*) &\le 2\text{conf}_j^\sigma(n) + 2\text{conf}_{j*}^\sigma(n) \le 4\text{conf}_j^\sigma(n) \\
\Delta^{\sigma^2} &\le 4\text{conf}_j^\sigma(n) \\
\Delta^{\sigma^2} &\le 32\bar{\eta}\sigma_j\sqrt{\frac{\log(12M/\delta)}{n}}
\end{aligned}$$

We know that for any $j \in [M]$, we eliminate arm if $\sigma_j^2 - \sigma_*^2 > 8 \implies \Delta^{\sigma_j^2} > 8$.

$$\begin{aligned}
8 &\le 32\bar{\eta}\sigma_j\sqrt{\frac{\log(12M/\delta)}{n}} \\
64 &\le \frac{1024\bar{\eta}^2\sigma_j^2\log(12M/\delta)}{n} \\
n &\le \frac{1024}{64}\bar{\eta}^2\sigma_j^2\log(12M/\delta)
\end{aligned}$$

Hence for elimination of sources we need

$$n > 16 \cdot \bar{\eta}^2\sigma_j^2 \log\left(\frac{12M}{\delta}\right) \implies n > 16 \cdot \bar{\eta}^4 \log\left(\frac{12M}{\delta}\right)$$

■

D.2.4. PRIMER ON ALGORITHM 2

Algorithm 2 takes as input the number of arms K, number of data sources M, and two tunable parameters: a confidence parameter $\delta \in (0, 1)$, controlling the confidence interval widths, and an exploration parameter $\tau \in \mathbb{N}$, determining the initial number of queries per source. Post preprocessing, the algorithm begins with an initial exploration phase: each of the $\tilde{M}$ data sources is queried exactly τ times, with arms chosen uniformly at random. At the conclusion of this phase (i.e., at round $t = \tilde{M}\tau$), we compute initial empirical estimates: the mean reward estimates $\hat{\mu}_t(i)$ for each arm $i \in [K]$ and the variance estimates $\hat{\sigma}_t(j)$ for each source $j \in [\tilde{M}]$, along with their corresponding confidence bounds $\text{UCB}^\mu_{\tilde{M}\tau}(i)$ and $\text{LCB}^\sigma_{\tilde{M}\tau}(j)$, as defined previously in Section 3.

After the initialization phase, from rounds $t = \tilde{M}\tau + 1$ onwards, the algorithm executes the following adaptive procedure at each round: it selects the arm i_t with the largest current upper confidence bound on mean reward, i.e., $i_t = \arg\max_{i\in[K]} \text{UCB}^\mu_{t-1}(i)$, and simultaneously selects the data source j_t with the smallest current lower confidence bound on variance, i.e., $j_t = \arg\min_{j\in[\tilde{M}]} \text{LCB}^\sigma_t(j)$. Once an arm-source pair (i_t, j_t) is chosen, the learner queries arm i_t via source j_t and receives a reward X_t. Subsequently, the algorithm updates the counts: $n_t(i)$, representing the total number of selections of arm i, $m_t(j)$ representing the total number of selections of source j and $n_t(i, j)$ representing the total number of selections of the arm-source pair (i, j), upto time t. Using these updated counts, the algorithm recalculates empirical estimates $\hat{\mu}_t(i)$ and $\hat{\sigma}_t(j)$ as well as their corresponding upper and lower confidence bounds ($\text{UCB}^\mu_t(i)$ and $\text{LCB}^\sigma_t(j)$) as defined in Equations (4) and (5).

Intuitively, employing a UCB approach for arm selection promotes optimistic exploration to rapidly identify rewarding arms, whereas the LCB approach for variance selection ensures cautious, risk-averse choice of data sources, guiding the learner towards consistently low-noise feedback.

Perhaps most notably, our regret analysis reveals a surprisingly negligible dependence on the source variances that arises from our carefully crafted LCB-UCB selection mechanism, which adaptively and swiftly prioritizes lower-variance sources while maintaining aggressive reward exploration. The tightness of the derived confidence bounds also plays a crucial role in the regret analysis, enabling our *simultaneous exploration and exploitation* approach to achieve performance nearly matching that of a hypothetical standard MAB algorithm with privileged access to the optimal (lowest-variance) data source σ^*. This striking result underscores the strength of our proposed adaptive strategy and its effective regret-minimization capability.

D.2.5. SUPPLEMENTARY FOR SECTION 4.1

Theorem 2 (Main Result: Regret Analysis of SOAR) *For any choice of preprocessing budget $n \geq 16 \cdot \bar{\eta}^4 \log\left(\frac{12M}{\delta}\right)$, initial-exploration $\tau = \max\left\{\frac{288M^2K^2}{\bar{\eta}^2}\frac{(\log(3KT/\delta))^2}{\log(12MT/\delta)}, 32\bar{\eta}^2\log(12MT/\delta)\right\}$ and $\sigma^* = 1$, the regret of SOAR (Algorithm 2) can be bounded by $\tilde{O}\left(M\bar{\mu}(n+\tau) + \sigma^{*2}\sum_{i=2}^{K}\frac{1}{\Delta_i}\right)$ with high probability $(1-\delta)$. SOAR can also be shown to yield an instance-independent (worst-case) regret bound of $O\left(M\bar{\mu}(n+\tau) + \sigma^*\sqrt{KT\log(KT/\delta)}\right)$*

Proof We begin by recalling an earlier remark

Remark 13 (On the Number of Sources After Preprocessing) *After the preprocessing phase, the number of surviving sources is denoted by $\tilde{M}$, where $\tilde{M} \leq M$. Since our concentration and regret guarantees are stated as upper bounds, it is sufficient to replace $\tilde{M}$ by M in the analysis. This simplification allows us to present cleaner expressions without loss of generality.*

Bounding $n_t(i)$ Note that at any time t, arm-i can not be selected if $\mathrm{UCB}_t^{\mu}(i) \leq \mathrm{UCB}_t^{\mu}(i^*)$. So arm-$i$ only gets selected at time t if:

$$\mu_{i^*} \leq \mathrm{UCB}_t^{\mu}(i^*) \leq \mathrm{UCB}_t^{\mu}(i) = \hat{\mu}_t(i) + \frac{2\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_t(i,j)\sigma_j^2}}{n_t(i)},$$

$$\mu_{i^*} \leq \mu_t(i) + \frac{4\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_t(i,j)\sigma_j^2}}{n_t(i)},$$

$$\implies n_t(i) \leq \frac{4\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_t(i,j)\sigma_j^2}}{\Delta_i}.$$

In Step 2, we made use of our mean reward concentration defined in Lemma 5.

The total regret bound Total Regret = Regret from PREPROCESS + Regret from initial exploration τ+ Regret from running SOAR Algorithm

The Regret from our preprocessing algorithm given is $\sum_{j=1}^{M} n(\mu_* - \mu_1)_j \leq Mn\bar{\mu}$
The Regret from initial exploration is $\leq \tilde{M}\tau\bar{\mu} \leq M\tau\bar{\mu}$

The Regret in $T - Mn - M\tau$ rounds of Running SOAR:

$$\text{Reg}_T = \sum_{i=2}^{K} n_T(i)\Delta_i \le \sum_{i=2}^{K} 4\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_T(i,j)\sigma_j^2},$$

$$\text{Reg}_T \le \sum_{i=2}^{K} 12\sigma^*\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_T(i,j)} \quad \left[\sigma_j^2 \le 9\sigma^{*2} \text{ w.h.p}\right]$$

$$\le 12\sigma^*\sqrt{\log(3KT/\delta)}\sum_{i=2}^{K}\sqrt{\sum_{j=1}^{M} n_T(i,j)}.$$

$$\implies \frac{\text{Reg}_T}{c} \le \sigma^* \sum_{i=2}^{K}\sqrt{\sum_{j=1}^{M} n_T(i,j)} \quad \left[c = 12\sqrt{\log(3KT/\delta)}\right]$$

$$= \sigma^* \sum_{i=2}^{K} \frac{\sqrt{\sum_{j=1}^{M} n_T(i,j)\ c\sigma^*}}{\sqrt{\Delta_i}} \frac{\sqrt{\Delta_i}}{\sqrt{c\sigma^*}}$$

$$\le \sigma^* \sqrt{\sum_{i=2}^{K}\frac{c\sigma^*}{\Delta_i}}\sqrt{\sum_{i=2}^{K}\sum_{j=1}^{M}\frac{n_T(i,j)\Delta_i}{c\sigma^*}}$$

$$\le \frac{1}{2}\left[\sum_{i=2}^{K}\frac{c\sigma^{*2}}{\Delta_i} + \sum_{i=2}^{K}\sum_{j=1}^{M}\frac{n_T(i,j)\Delta_i}{c}\right]$$

$$= \sum_{i=2}^{K}\frac{c\sigma^{*2}}{2\Delta_i} + \frac{\text{Reg}_T}{2c}.$$

The final SOAR regret bound will thus be:

$$\text{Reg}_T \le 2c\left(\sum_{i=2}^{K}\frac{c\sigma^{*2}}{2\Delta_i}\right),$$

$$\text{Reg}_T \le 144\log(3KT/\delta)\left(\sum_{i=2}^{K}\frac{\sigma^{*2}}{\Delta_i}\right).$$

Therefore total Regret, whp $1-\delta$, is bounded by:

$$\text{Reg}_\text{T} \le M\bar{\mu}(n+\tau) + 144\log(3KT/\delta)\sum_{i=2}^{K}\frac{\sigma^{*2}}{\Delta_i}$$

■

This proves the first part of Theorem 2, which yields an instance-dependent guarantee of SOAR. Finally, to see the worst case (instance independent) regret analysis of SOAR, note that the regret in $T - M\tau - Mn_p$ rounds of SOAR can be alternatively bounded as:

$$\text{Reg}_T = \sum_{i=2}^{K} n_T(i)\Delta_i \le \sum_{i=2}^{K} 4\sqrt{\log(3KT/\delta)\sum_{j=1}^{M} n_t(i,j)\sigma_j^2}$$

$$
\begin{aligned}
&\leq 4\sigma_{\max}\sqrt{K\log(3KT/\delta)}\sqrt{\sum_{i=2}^{K}\sum_{j=1}^{M} n_t(i,j)} \;\; \text{(applying Cauchy's Schwarz)} \\
&\leq 12\sigma^*\sqrt{K\log(3KT/\delta)}\sqrt{\sum_{i=2}^{K}\sum_{j=1}^{M} n_t(i,j)} \;\; \text{(noting } \sigma_{\max} \leq 3\sigma^* \text{ w.h.p in after } Mn \text{ rounds)} \\
&= 12\sigma^*\sqrt{KT\log(3KT/\delta)},
\end{aligned}
$$

proving the final claim of Theorem 2. This concludes our regret analysis of SOAR. ■

Appendix E. Experimental Setup and Extended Results

In this section, we describe the experimental setup and present our baseline comparison results along with the results obtained by varying the number of arms K and the number of sources M.

E.1. Variation in the Number of Arms and Sources

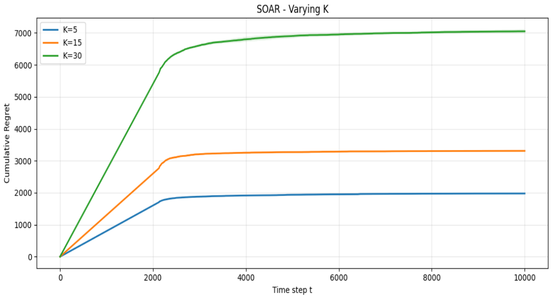

Figure 3: Regret of SOAR with varying number of arms $K \in \{5, 15, 30\}$

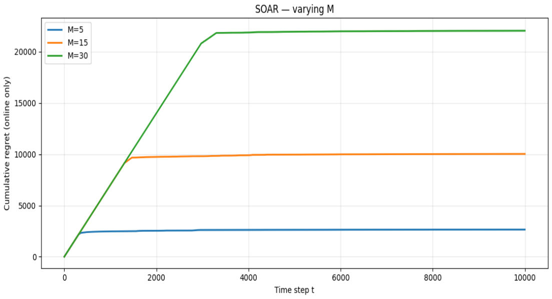

Figure 4: Regret of SOAR with varying number of sources $M \in \{5, 15, 30\}$

We evaluate SOAR on synthetic multi-source bandit tasks under two complementary setups. Arm rewards are modeled as Gaussian random variables with fixed means and source-dependent variances. Arm means are either drawn uniformly from $[1, 10]$ (rounded to one decimal place) or fixed as $\mu = [1, 5, 8, 6, 4]$ as described below. Source variances are specified directly or sampled uniformly from $[1, 3]$ (rounded to one decimal place). All runs are executed for a time horizon of $T = 10{,}000$. In both setups, we observe an initial linear growth in regret, which stems from the preprocessing phase combined with the initial exploration rounds.

Varying K We fix the number of sources to $M = 3$ with variances $\sigma = [5, 1, 10]$, and vary the number of arms $K \in \{5, 15, 30\}$. For each configuration, arm means are drawn uniformly from $[1, 10]$ (rounded to one decimal place). This setup isolates how SOAR scales with the number of arms while holding source variability constant.

Varying M We fix the arms to $\mu = [1, 5, 8, 6, 4]$ and vary the number of sources $M \in \{5, 15, 30\}$. The source variances are sampled independently from $[1, 3]$ (rounded to one decimal place). This setup isolates how `SOAR` scales with the number of sources while holding the arm structure fixed. The initial growth in regret clearly reflects the expected scaling with M, consistent with our theoretical analysis.

E.2. Baseline Comparison Results

For both baselines, we set arm means in the range $[0, 1]$ and source variances in the range $[1, 10]$.

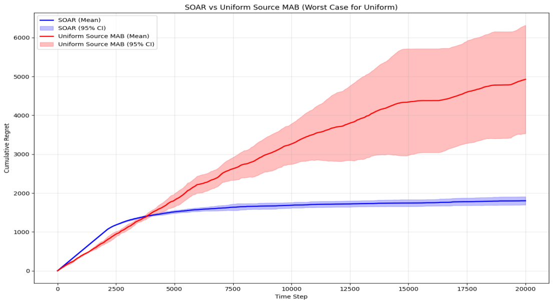

Figure 5: `SOAR` vs Baseline 1: WC1

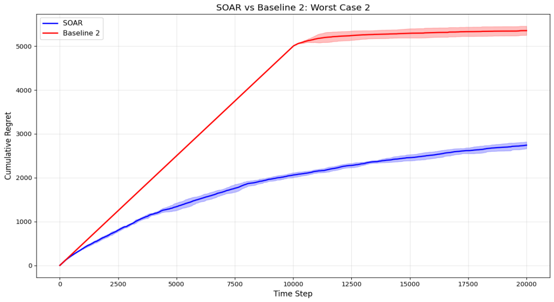

Figure 6: `SOAR` vs Baseline 2: WC2

Baseline 1: Uniform Source MAB For the uniform-source baseline (WC-1: $K = 5$, $M = 3$), we construct instances with multiple high-variance sources alongside a low-variance source, creating a stark variance disparity. In this setting, `SOAR` initially incurs a higher cost due to its exploration phase, but then stabilizes and achieves substantially lower regret by adaptively prioritizing the low-variance source. In contrast, the uniform baseline continues to suffer from repeatedly sampling the high-variance sources throughout the horizon.

Baseline 2: Two Phase MAB For the two-phase baseline (WC-2: $K = 10$, $M = 8$), we instead consider sources with gradually increasing variances, introducing only incremental differences across sources. Here, `SOAR` effectively handles the fine-grained variance differences by relying on continuous confidence bounds for adaptive source selection, whereas the two-phase baseline incurs significant regret due to its rigid elimination phase. Both scenarios are evaluated over $T = 20{,}000$ rounds.