Stochastic Bandits on Multimodal Distributions: Metrics & Regret Bounds

Adit Jain
CollinearAI

Sujay Bhatt
JPMorgan AI Research

Alec Koppel
Johns Hopkins APL

Abstract

Multimodal reward distributions naturally arise in real-world applications such as targeted recommendations to heterogeneous sub-populations and selective unit-level interventions. These settings challenge standard mean or risk-based bandit approaches, requiring metrics that quantify the merit of mixture parameters without prior mode knowledge. We consider the bandit setting where the reward associated with an arm is sampled from a finite mixture of Gaussians, which is strictly more general than the unimodal setting. We consider ranking arms using functions of the mixture parameters and propose methods to minimize the cumulative regret with respect to the induced ranking. We show that the achievable pseudo-regret has a lower bound of the order $\Omega(\mathsf{T}^{1/2})$ and propose an explore and exploit based on expectation maximization (ETE-EM) algorithm which achieves a regret of $\widetilde{\mathsf{O}}(\mathsf{T}^{2/3})$. Further, we show that the modification of Thompson sampling (TS-EM) achieves a Bayes regret of $\widetilde{\mathsf{O}}(\mathsf{T}^{1/2})$. Experiments validate our approach in practice, where we benchmark against both algorithms designed for sub-Gaussian bandits and naive clustering-based extensions of empirical CDF methods, showing our approach achieves consistently lower regret across choice of metrics.

1 INTRODUCTION

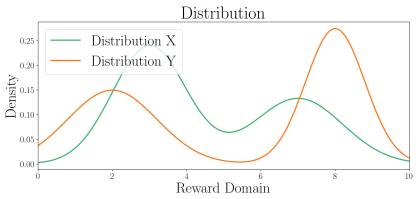
Multi-armed bandits (Lattimore and Szepesvári, 2020) are a framework for sequential decision-making with incrementally revealed rewards that find application in personalized medicine (Tewari and Murphy, 2017), online advertising (Li et al., 2010a), and product testing (Kaufmann et al., 2014; Xiong et al., 2024). In stochastic bandits, a learner chooses between N arms at each round and receives a reward sampled from the reward distribution of the arm. Typically, performance is quantified in terms of the difference between accumulated reward of causally selecting actions and a best-in-hindsight defined in terms of the maximum *expected* reward. There are many algorithms proposed to minimize this difference, referred to as stochastic regret; however, in many preference modelling problems, pulling the arm with the best mean is not the desiderata, especially when reward distribution has a richer structure. For example, in situations where the reward distribution convolves multiple input signals, such as (i) a demand curve containing samples from distinct sub-populations (Allenby et al., 1998); (ii) a source signal being sampled in the presence of interference (Poor, 2013); (iii) portfolio selection problems when market indicators contain distinct micro and macro-effects (Luxenberg and Boyd, 2024), optimizing for the mean is suboptimal in many problems.

In lieu of this objective, previous work has looked the definition of regret to incorporate risk (Sani et al., 2012), or considered general empirical distributions and metrics over them to define regret (Chan et al., 2013), but doing so is computationally costly, require tuning hyperparameters and fail to model particular shape characteristics of the reward distribution since they are still aggregate metrics of the entire distribution. This necessitates new metrics that preserve distributional structure while enabling flexible risk definitions.

To close this gap, we develop machinery for the case where the reward distribution for each arm is a mixture of Gaussians (MoG). Doing so reveals subtleties that do not arise in unimodal settings (which it is a relaxation of), either for mean or risk-sensitive settings.

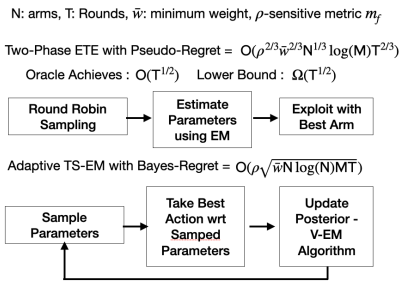
Defining an ordering over mixture families presents questions that are absent in the unimodal case (see

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

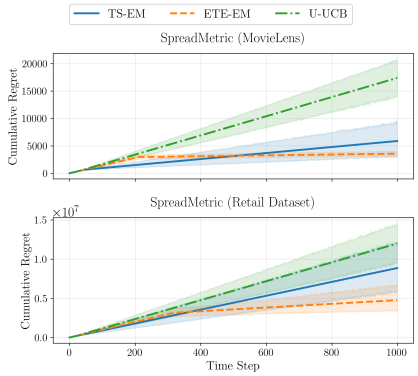


	Distribution	Mean	Sharpe Ratio	CVaR
Aggregate	X	4.59	**2.05**	**1.14**
	Y	**5.30**	1.68	-0.05
Mode 1	X	**2.99**	**2.99**	**0.92**
	Y	2.00	1.66	-0.47
Mode 2	X	6.99	5.82	4.52
	Y	**7.99**	**9.98**	**6.35**

(a) Misalignment of Aggregate and Per-Mode Metrics for Mixture Rewards



(b) Algorithms and Theory for Regret Minimization of General Metrics



(c) Empirical Benchmarking on Synthetic and Semi-Synthetic Tasks

Figure 1: (left) Aggregate metrics are not sufficient for modelling preferences over multimodal distributions: The example illustrates different metrics for two bimodal reward distributions for the entire distribution and for each mode. The ordering of the arms with respect to the aggregate metrics is not preserved with respect to the metrics of the individual modes. This motivates studying generalized metrics that characterize preference structures over multimodal distributions. (middle) We further propose algorithms to minimize the resulting regret with respect to these metrics and show theoretical guarantees. (right) We numerically benchmark against unimodal algorithms which can be used to model preferences and show that our algorithms perform better across a range of metrics.

Fig. 1): a learner may be interested in selecting an arm with respect to a metric other than the mean, which is too coarse. For example, the mean or risk of the multimodal distribution may be different from the mode with maximal mean (risk), and therefore ignores distributional microstructure. Instead, by defining metrics in terms of mixture parameters, we provide more granular ranking criteria that are attuned to the mixture family, otherwise not possible with empirical distribution-based techniques (Mathews and Schmidler, 2024).

Motivating application. Consider the problem of *subpopulation targeting.* When a demographic or behavioural segment for content recommendation is chosen, rarely is the whole response curve observed. Rather, only one person is observed at a time and each individuals return (click-through, enrollment probability, etc.) might be well-modelled by a unimodal distribution, but the segment itself is a mixture, and therefore parameterizing via a unimodal distribution would be mis-specified. In the case of movierecommendation, the audience may naturally split into user "modes": enthusiasts, who prefer niche titles, and casual viewers, which prefer mainstream. If one wants bestsubgroup clickthrough rate, i.e., how exciting a film can be for its most receptive subgroup, regardless of how it does on average, then a traditional meanbased bandit would ignore these modes.Other applications include: (i) *A/B Testing:* In A/B testing, multi-armed bandits are used to evaluate whether introducing a new feature outperforms a baseline on a population (Qin et al., 2023). However, if multiple sub-groups of participants are present in the experiment, the participation of different sub-groups can vary over time, and therefore, the reward distribution quantifying performance assessment is better modeled as multimodal. (ii) *Asset Selection for Liquidation:* When assets have to be sequentially selected to invest in, the reward distribution is multimodal if the market regime is comprised of distinct causal effects, e.g., macro events vs. media driven interest (Huo and Fu, 2017). (iii) *LLM Routing and Selection:* When a platform must sequentially choose among candidate LLMs to serve user queries (Bouneffouf and Feraud, 2025), the observed reward (e.g., user satisfaction, response quality) is drawn from a mixture distribution whose components correspond to heterogeneous user subpopulations—domain experts versus novices, or enthusiasts versus casual users—each responding differently to the same LLM. Standard mean-based routing collapses this subpopulation structure. Our mixture-parameter metrics enable finer-grained objectives: max-mean selects the LLM that excels for its best-served subgroup, while min-CVaR ensures no subgroup is poorly served. Recent work on reward model selection for RLHF (Nguyen et al., 2025) and prompt optimization (Shi et al., 2024) operates in precisely this mixture-reward regime.

Related Literature and Novelty. Previous work in risk-aware bandits has looked at alternative ways of ranking arms (Khurshid et al., 2024; Sani et al.,

2012), specifically looking at ranking arms by their conditional value at risk at a user-specified risk level and using a regularized square Sharpe ratio. Similarly, there has been work done in extending the mean-based regret minimization framework using stochastic dominance (Chen et al., 2016). However, these methods model (risk) preferences of the user but do so for the entire distribution, whereas our work is focused on modeling preferences across modes that a user might have, which are more granular and can avoid being overly conservative (Guegan et al., 2017). Cassel et al. (2018) look at a more general framework to rank bandit arms based on risk metrics and use the empirical CDF to compute the best arm. On the other hand, we look at a mixture of Gaussians where exact estimation of parameters is possible, however it is difficult to write the functional form required in (Cassel et al., 2018). We also note that our per-mode metric framework has a natural interpretation in algorithmic fairness: by defining metrics on individual mixture components, one can ensure that the selected arm performs well for each subpopulation rather than only on aggregate, aligning with fairness desiderata across heterogeneous groups.

Further, by formulating sensitivity to the effect of parameter variations, we extend ideas beyond general metric-based bandits (Feng et al., 2024). More recently, Tatl et al. (2025) have proposed looking at a notion of alternative regret measures defined with respect to a mixture of arms. Their approach optimizes the mixture coefficients over different arms to optimize for the preference-based metrics. Whereas in our parametric framework, the mixture coefficients of the distribution are unknown, and then we optimize preference-based metrics defined on the distribution parameters. We note that our setting is distinct from contextual bandits (Li et al., 2010a): in contextual bandits the context (e.g., subpopulation identity) is observed before the action is chosen, whereas in our formulation the context is randomly sampled and unknown to the agent at decision time. The learner only observes a scalar reward drawn from the mixture, never the identity of the component that generated it. A detailed discussion including contextual and restless bandits is in Appendix A.

Our **main contributions are as follows:**

• We propose a generalized metric to order arms which have a multimodal distribution by using a function of distribution parameters (Section 2). We provide examples of such metrics to model preferences which are otherwise intractable by unimodal metrics.

• To minimize the regret with respect to a chosen metric, we look at modifications of two popular algorithms: Explore then Exploit (ETE) and Thompson Sampling (TS) using the expectation maximization (EM) algorithm. For the frequentist variant we show that the lower bound on the pseudo regret is $\Omega(\mathsf{T}^{1/2})$ and propose an oracle which can achieve it. However such an oracle has conditions which current estimations techniques do not follow, and we show that the ETE-EM achieves a regret of $\widetilde{\mathsf{O}}(\mathsf{T}^{2/3})$. We show that the TS-EM algorithm (for the Bayesian variant of the problem) is able to achieve a Bayes-Regret of $\widetilde{\mathsf{O}}(\mathsf{T}^{1/2})$.

• We present numerical simulations in Section 4 on synthetic and semi-synthetic datasets on a variety of different metrics to show how the proposed metrics capture different learner preferences naturally arising when the reward distribution is multimodal.

Notation. We denote vectors with bold small letters, scalars with plain letters, and sets with curly capital letters. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ and $\mathbb{E}$ to denote the expectation. $\widetilde{\mathsf{O}}(\cdot)$ notation hides logarithmic factors in T, M. $\mathcal{N}(\mu, \nu)$ denotes a Gaussian with mean μ and variance ν. $\|\cdot\|$ denotes the ℓ_1-norm.

2 ARM ORDERING FOR MOG REWARDS

One can order the arms based on either some function of the parameters of the distribution (similar to mean for unimodal) which we refer to as a parameteric metric or in terms of the expectation of a function with respect to the distribution which we refer to as non-parameteric metrics, (discussed in Appendix C.1) of ranking bandit arms with multimodal reward distribution along with the associated regret, which we aim to minimize.

Consider a multi-armed bandit setting with N bandit arms. Each arm $i \in [N]$ has a distinct reward distribution which is mixture of M Gaussian distributions with parameters of the M unimodal distribution as $((\mu_m^i, \nu_m^i, w_m^i))_{m=1}^M$ denoted collectively by $\theta^i \in \Theta$. That is, the reward distribution of arm i is given by $\sum_m w_m^i \mathcal{N}(\mu_m^i, \nu_m^i)$, where $w_m^i \in [0, 1]$ and $\sum_{m=1}^M w_m^i = 1$ for all $i \in [N]$. We assume that the number of mixture components is known to the learner and fixed across arms. On the other hand, the parameters of reward distribution are unknown. This is a non-standard specification that generalizes the standard unimodal Gaussian setting, which we refer to as a mixture of Gaussians (MoG) (we also refer to this as a multimodal distribution). The arm selected at time k is given by $a_k \in [N]$ and the received reward by r_k.

The following is assumed for the mixture distribution,

(A1) (*Bounded Variance*) The mixture variances are bounded by 1, $\nu_m^i \leq 1 \ \forall m \in [M],\ i \in [N]$.

(A2) (*Minimum Separation*) For $\min_m^i w_m \geq \bar{w} >$

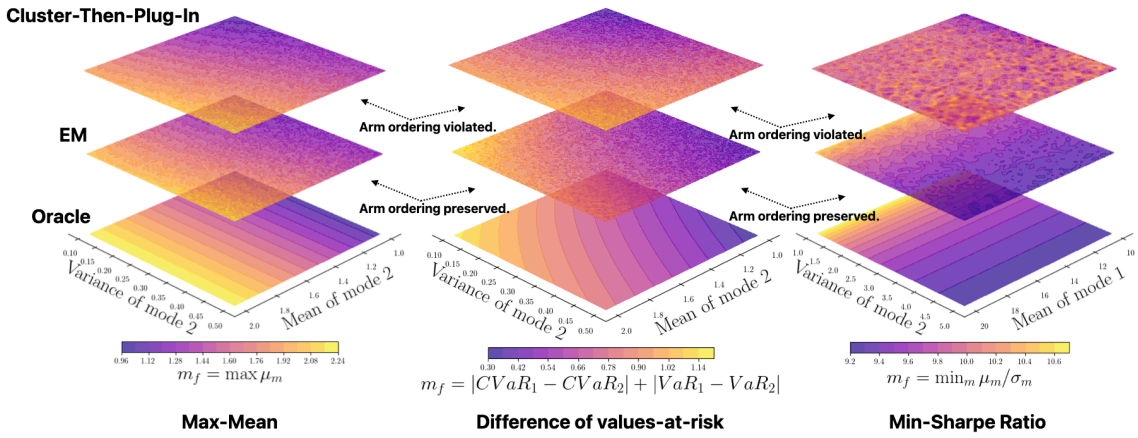


Figure 2: Illustration of preference alignment with multimodal distributions: We visualize three different metrics for a bimodal distribution from Eg. 2.3-2.4, the contour layer shows the outcome of different parameter estimation methods. The EM-based approach consistently recovers the underlying distribution parameters such that preferences are well-aligned with the true structure. In contrast, the clustering-based plug-in estimation approach yields preferences that are slightly off (left), misaligned (middle), or noisy (right). Therefore, our method enables modelling a broader class of preference structures compared to standard plug-in techniques.

0, $\min_m^i \nu_m \geq \bar{\nu} > 0$, where $\bar{w}$ and $\bar{\nu}$ are strictly positive lower bounds on the minimum component weight and variance respectively, then $\min_{m,l} |\mu_m^i - \mu_l^i| = \Omega(\sqrt{\frac{1}{\bar{w}\bar{\nu}}}), \forall i \in [N]$. Note that in Assumption (A1), the variance of the components is upper bounded by 1, and the results still hold up to constants; however, we let the variance be bounded by 1 for notational convenience. Assumption (A2) is referred to as the separability of the mixture components in the literature and ensures that the overlap between the components is controlled.

We analyze parametric maps that convert MoG parameters into scalar values, establishing a total ordering (ranking) over the arms. The ordering aligns naturally with mean-based regret minimization (where arms are ranked by their means). However, without total order, conventional regret requires identifying a single "best arm" a definition that breaks in multimodal settings.

To resolve this, define a function m_f for an instance of the MoG distribution from the parameter space to the real numbers: $m_f : \Theta \to \mathbb{R}$. Hence, the m_f function naturally imposes a total order on the arm space (we consider two arms as equally good if the m_f function is the same to avoid corner cases, and break ties arbitrarily). Denote $m_f^i = m_f(\theta^i)$ as the m_f of arm i. Therefore, m_f imposes an ordering on the arms, which permits us to define regret with respect to the best arm. We restrict focus to the case where one tries to select the arm with the best m_f and denote $m_f^* = \max m_f^i$.We also define the metric sensitivity, which we use in our theoretical results.

Definition 2.1. (Metric Sensitivity) A metric m_f is (ϵ, ρ)-sensitive if $\|\theta' - \theta\| \leq \epsilon$ implies $|m_f(\theta') - m_f(\theta)| \leq \rho\epsilon$, where $\epsilon, \rho > 0$ are positive scalars.

We refer to a metric as $\rho-$sensitive if the metric is sensitive for all $\epsilon > 0$. The metric sensitivity defined above is similar to Lipschitz continuity used when analyzing bandits on metric spaces (Kleinberg et al., 2019; Xue et al., 2024), where the condition implies that the change in the payoff function is controlled across arms (which lie on a metric space). In our case, the arms are defined by the parameters of the multimodal reward distribution; however, the key difference is that the parameters are unknown to the learner apriori.

We now present a few exemplar metrics (with more metrics presented in Section 4). Note that the mean is contained within this problem class with $m_f(\theta) = \sum w_m \mu_m$. Thus, the subsequent results generalize standard mean-based regret minimization.

Example 2.2. *Max-Mean:* Maximum of the means of modes of a multimodal distribution, $\mathsf{mm}(\theta) = \max_{m \in [M]} \mu_m$. For a bimodal distribution, this metric captures the learner's preference in choosing the arm that performs the best in a higher average reward regime, and it is $(\epsilon, 1)-$sensitive for all $\epsilon > 0$. It is useful in model selection with multimodal data, where the subgroup with the highest mean is more informative (Liang et al., 2023), and in patient care, where bed allocation for people with more acuity can improve health outcomes (Noyes, 2008).

Example 2.3. *Weighted Sharpe Ratio* A linear combination of the Sharpe ratio of the different modes of

the MoG. $\mathsf{wsr}(\theta) = \sum_{m\in[M]} \beta_m \frac{\mu_m}{\sqrt{\nu_m}}$, where the modes are ordered by their means and β_m are weighing factors decided by the learner and can be equal to w_m. This metric quantifies a way to weigh the different risk preferences for the different modes. Analyzing modal components separately reduces estimation error compared to treating multimodal returns as a single distribution (Liu et al., 2012). In bimodal markets (bull or bear), weighting Sharpe ratios by regime duration provides better risk-adjusted returns (Krimm et al., 2012). Further, for a given MoG where the variance and means lie between $\mu_m \in [\underline{\mu}, \bar{\mu}]$ and $\nu_m \in [\underline{\nu}, \bar{\nu}]\ \forall m$, the weighted Sharpe ratio is $(\underline{\nu}, \frac{\bar{\mu}}{2\underline{\nu}^2})$-sensitive.

Example 2.4. *Minimum Conditional Value-at-Risk (CVaR)* For a Gaussian $\mathcal{N}(\mu, \nu)$, the CVaR can be written exactly as $\mathsf{CVaR}_\alpha(\mu, \nu) = -\mu + \sqrt{\nu}\frac{\psi(\varphi(\alpha))}{\alpha}$, where ψ and φ are the p.d.f. and c.d.f. of the standard normal distribution, respectively. Therefore a risk-seeking designer can design a metric which computes the minimum CVaR (across modes - assuming each mode to be a different distribution) for each arm, $\mathsf{m} - \mathsf{CVaR}_\alpha(\theta) = \min_{m\in[M]} \mathsf{CVaR}_\alpha(\mu_m, \nu_m)$. For $\epsilon < 1$, this metric is $(\epsilon, \frac{2}{\sqrt{\epsilon}})$-sensitive. Using minimum CVaR among subgroups leads to a more conservative approach in risk management, which is beneficial if data is drawn from mixture families (Guegan et al., 2017).

Figure 2 shows how the ordering imposed by preference-based metrics defined on the parametric space is not preserved by clustering and then plug-in estimation. However, EM (estimation technique used by us) does preserve the ordering over arms. Doing so, however, invalidates standard techniques for reasoning about uncertainty based on concentration bounds of sub-Gaussian random variables, which we resolve through the introduction of statistical guarantees specific to the estimation procedure used in the recovery of mixture parameters using the EM algorithm.

3 MINIMIZING REGRET FOR MOG METRICS

3.1 Pseudo-Regret Minimization

In this section, we aim to minimize (pseudo-)regret for the metric m_f with respect to the best arm. The expected regret with respect to m_f for a selected arm sequence $a_1, \ldots, a_\mathsf{T} \in [N]$ for T time steps is given by,

$$\mathsf{R}_\mathsf{T}\left((\theta_i)_{i=1}^N, m_f, \pi\right) = \mathbb{E}\left\{\sum_{k=1}^{\mathsf{T}} (m_f^* - m_f^{a_k})\right\}, \quad (1)$$

where the expectation is taken with respect to the randomness in the algorithm.

We first derive a lower bound on the achievable regret and propose an oracle which can achieve it using a upper confidence bound approach. However in lieu of such an oracle we propose a explore then exploit (ETE) algorithm which achieves a slightly worse-off $\mathsf{O}(T^{2/3})$ regret without the restrictive conditions of the oracle. To derive the lower bound we define a policy,

Definition 3.1. (Policy) A policy π is a mapping from history of actions and rewards, $\mathcal{H}_k = (a_1, r_1, \ldots, a_{k-1}, r_{k-1})$ to the action space, which is the set of arms $\{1, \ldots, N\}$.

The learner's policy is parameterized by the estimate of the metric calculated using the estimates of the mixture components of the Gaussian mixture. Further we need the following assumption,

(**B1**) There exists a mode $m \in [M]$, such that if two parameter vectors θ and θ', differ only in mean coordinate μ_m, $\mu_m - \mu'_m > \Delta$ and $\mu_l = \mu'_l, l \neq m$, then $m_f(\theta) - m_f(\theta') \geq \kappa\Delta$ for some $\kappa > 0$.

Assumption B1 is a very mild assumption, and is much weaker than a reverse Lipschitz-like condition. Specifically, it only requires that the metric be sensitive in atleast one mean of the multimodal distribution. All the metrics presented in this paper except the modal entropy satisfy this assumption. We now state the lower bound the regret of any policy (Def. 3.1).

Theorem 3.2. *Consider the N armed bandit problem where the arm i have a mixture of M Gaussian distribution with each arm having distribution parameters $(\mu_m^i, 1, \frac{1}{M})$ for the m-th mixture. Consider the regret minimization with respect to a metric m_f which is satisfies (B1) for some $\kappa > 0$. Let $\mathsf{T} > N > 1$ be the number of learning rounds. Then for any policy π there exists a sequence of mean vectors $(\mu^i)_{i=1}^N$ for the arm distribution parameters such that the regret is,*

$$\boxed{\mathsf{R}_\mathsf{T}\left((\theta_i)_{i=1}^N, m_f, \pi\right) = \Omega\left(\kappa\sqrt{(N-1)\mathsf{T}M}\right)}$$

The above theorem establishes that the regret of any algorithm is at least $\Omega(\sqrt{\mathsf{T}})$. The result generalizes the lower bound of the mean-based regret minimization of unimodal bandits (Theorem 15.2 of Lattimore and Szepesvári (2020)) and recovers it when variances are 1 and the number of mixtures is $M = 1$. We conjecture that regret is achievable using an EM-based adaptive algorithm; however, one requires estimation guarantees free of large and unknown constants. To be precise let us define an Mixture-Parameter Confidence Oracle (MPCO) which lets us construct adaptive algorithms,

Definition 3.3 (Anytime MPCO)**.** Fix an arm i. For every round k and failure level $\delta \in (0, 1)$ the oracle takes the cumulative sample set $\mathcal{S}_i(k) = \{r_i^1, , r_i^n\}, n =$

$n_i(k)$, and returns a pair $(\hat{\theta}_i(k), c_i(k,\delta))$ such that,

$$\mathbb{P}\left\{\forall j \le k : \|\hat{\theta}_i(j) - \theta_i\|_2 \le c_a(j,\delta)\right\} \le 1-\delta$$

Here $c_i(\cdot)$ is the radius which satisfies the following, (a. shrinkage) $c_i(k,\delta)$ is non-increasing in k. (b. square root rate) $c_i(k,\delta) = \mathsf{O}\big(\sqrt{\frac{\log(1/\delta)}{n_i(k)}}\big)$. (c. computability) $(\hat{\theta}_i, c_i)$ are computable in $\mathsf{poly}(n_i(k), M)$ time.

Using such an MCPO, for a ρ-sensitive metric m_f (Def. 2.1 of the paper) one can form per-round upper confidence bounds on the parameter estimates, $\mathrm{UCB}_i(k) = m_f(\hat{\theta}_i(k)) + \rho c_i(k,\delta)$ and plug them into a standard UCB (Bubeck and Nicolò, 2012) to obtain anytime regret, closing the gap to the lower bound.

However such oracles are non-trivial for three primary reasons, (a) Confidence sets must be permutation-invariant; this couples the coordinates of θ in a way absent from unimodal means. (b) EM iterations are only locally contractive under separability and good initialisation and turning the one-shot guarantees into uniform-in-t bounds is open (c) Each observation is drawn from a latent mode, so standard martingale methods (Hoeffding, Bernstein) fail. For M = 1 the oracle reduces to the classical *Mean-Variance Oracle*: $(\hat{\theta}_i(k) = \hat{\mu}(t), c_i(k,\delta) = \sqrt{(2\log(2/\delta))/n_i(k)})$.

Algorithm 1 Explore-then-exploit for regret minimization of m_f metric in multimodal bandits (ETE-EM)

1: *Input: N arms $\mathcal{A}$, Number of rounds T, Metric Sensitivity ρ, Metric m_f*
2: Set exploration period $\mathsf{T}_e = (\rho\mathsf{T}/\bar{w}^3 N)^{2/3}$
3: **for** i in $[N]$ **do**
4: **for** k in $1,\dots,\mathsf{T}_e$ **do**
5: Observe reward $r_k \sim r(i)$
6: Store reward $\mathcal{S}^{i_k} = \mathcal{S}^{i_k} \cup \{r_k\}$
7: **end for**
8: Estimate parameters $\hat{\theta}_i$ using EM algorithm (Kwon and Caramanis, 2020) with $\mathcal{S}^{i_k}$.
9: **end for**
10: **for** k in $\mathsf{T}_e + 1,\dots,\mathsf{T}$ **do**
11: Choose $a_k = \arg\max_i m_f(\hat{\theta}_i)$
12: **end for**

We now develop an explore-then-exploit (ETE) algorithm which is used extensively in, e.g., adversarial bandits (Lattimore and Szepesvári, 2020), sparse bandits (Hao et al., 2020) and contextual bandits (Li et al., 2010b) and does not require the conditions of the MCPO. The ETE algorithm is simple and efficient, with only one estimation being performed per arm during the entire run.

We first detail the algorithm and the estimation procedure used. Then, we prove an upper bound on the regret achieved by the algorithm. Our ETE method (Algorithm 1) has two phases. The input to the algorithm is the number of arms N, the total experimentation period T, the metric m_f and its sensitivity ρ. We explore for T_e rounds and sample rewards from each arm for T_e rounds (Line 4-7). We estimate the parameters for each arm using the collected rewards (denoted by set $\mathcal{S}^i$ for arm i) using EM estimation (Line 8). For the remaining $\mathsf{T} - N\mathsf{T}_e$, we play the arm with the maximal empirical m_f metric. These metrics are plugged-in estimates using the parameter estimates (Line 11).

Upper Bound on Regret: To prove the upper bound on the regret of Algorithm 1 we use the result we use a restated version of Theorem 7 from Kwon and Caramanis (2020) for the one-dimensional case (Lemma F.1). This requires the the following additional assumption about the initialization of EM **(A3)**,

$$\forall m \in [M], |\tilde{\mu_m}^{(0)} - \mu_m| \le \frac{1}{16}\min_{m\neq l}|\mu_m - \mu_l|,$$
$$|\tilde{w_m}^{(0)} - w_m| \le \frac{w_m}{2},\ |\tilde{\nu_m}^{(0)} - \nu_m| \le \frac{\nu_m}{2}.$$

Assumption (A3) is a strong condition, but it is needed to ensure that the EM algorithm does not get stuck in local maxima of the log-likelihood function. Further, we show in our numerical experiments that the EM algorithm is fairly robust to initialization (we initialize it randomly). We state the upper bound on the regret achieved by ETE (Alg. 1), under the condition that the regret metric is (ϵ,ρ)-sensitive (Def. 2.1).

Theorem 3.4. *Given N arms in a multi-armed bandits where each arm has a mixture of M Gaussians as a reward distribution each of which satisfy assumption (A1-A3) and let $\bar{w} = \min_{m\in[M],i\in[N]} w_m^i$. Let Θ be the parameter space, and metric $m_f : \Theta \to \mathbb{R}$ be $\rho-$sensitive and regret be defined as* (1) *for T rounds. Then the expected regret of Algorithm 1 with an exploration period $\mathsf{T}_e = (\rho\mathsf{T}/\bar{w}^3 N)^{2/3}$ using a sample splitting EM algorithm for estimation of parameters is,*

$$\boxed{\mathsf{R}_\mathsf{T}\left((\theta_i)_{i=1}^N, m_f, \pi\right) = \tilde{\mathsf{O}}\left(\bar{w}^{-2} N^{1/3}\rho^{2/3}\mathsf{T}^{2/3}\log(M)\right)}$$

Proof is in Appendix F.3. Thus, Algorithm 1 achieves a sublinear regret of the order $\tilde{\mathsf{O}}(\mathsf{T}^{2/3})$, which increases in the number of arms N, as expected and reduces to the same order as bounds known for mean-based regret minimization using Explore then Commit algorithm in sub-Gaussian bandits Lattimore and Szepesvári (2020). The dependence on the number of mixture components is $\log M$ and the regret grows as $\bar{w}^{-2}$ in the minimum mixture weight, reflecting the increased estimation difficulty when components have small weights.

3.2 Bayesian Regret Minimization

One can consider the Bayesian version of the problem where the parameter of problem instance is sampled from a Bayesian prior γ. In this version of the problem one aims to minimize the Bayesian regret given by,

$$\mathsf{BR}_{\mathsf{T}}(\gamma) = \mathbb{E}_{\theta\sim\gamma}\mathbb{E}\left\{\sum_{k=1}^{\mathsf{T}} m_f^*(\theta) - m_f^{a_k}(\theta)\right\} \quad (2)$$

We summarize our algorithm in 2 which is the extension of Thompson Sampling to the multimodal regime where a posterior is updated when new reward samples are collected. At each timestep a parameter set is sampled from the posterior (Line 4) and one selects the arm with the maximum m_f calculated on these sampled parameter set (Line 5). The posterior for each arm is chosen as Dirichlet for the mixture weights and Normal-Inverse-Gamma for the mean and variance. The posterior update (Line 7) in practice can be done using a standard variational inference procedure for Gaussian mixture models (Blei and Jordan, 2006; Pedregosa et al., 2011).

Algorithm 2 Thompson Sampling for Bayes-Regret Minimization in multimodal bandits (TS-EM)

1: *Input* N arms $\mathcal{A}$, Number of rounds T, Metric m_f
2: Initialize Prior ξ
3: **for** $k = 1, \ldots, \mathsf{T}$ **do**
4: $\quad\hat{\theta} \sim \xi(\cdot|\mathcal{S})$
5: $\quad$Choose $a_k = \arg\max_i m_f(\hat{\theta}_i)$
6: $\quad$Sample reward $r_k \sim \theta_i$ and store $\mathcal{S}^{i_k} = \mathcal{S}^{i_k} \cup \{r_k\}$
7: $\quad$Update posterior ξ using Variational-EM
8: **end for**

We now present our first main result which shows a $O(\sqrt{T})$ optimal Bayesian regret bound for Algorithm 2 and the proof can be found in Appendix F.1.

Theorem 3.5. *Consider a multi-armed bandit problem with N arms, where each arm $i \in [N]$ has a reward distribution that is a mixture of M Gaussians with parameters $\theta^i = \{(\mu_m^i, \nu_m^i, w_m^i)\}_{m=1}^M$. Let the parameters be drawn from a regular conjugate prior distribution γ. Assume (A1), (A2) and that the metric m_f is (ϵ, ρ)-sensitive. Then the Bayesian regret of (2) of Thompson Sampling (Algorithm 2) after T rounds satisfies:*

$$\boxed{\mathsf{BR}_{\mathsf{T}}(\gamma) = \mathsf{O}\left(\rho\sqrt{2NM\bar{w}\cdot\mathsf{T}\log N}\right)}$$

4 EXPERIMENTAL RESULTS

Synthetic Experiments. We synthesise $N = 10$ arms whose rewards follow a mixture of $M = 4$ Gaussians. Component means are sampled i.i.d. from the interval $[0, 5]$; each component has a variance of $\mathsf{Unif}(0.5, 2)$, and mixture weights are drawn from a symmetric Dirichlet prior. Each algorithm is run for $\mathsf{T} = 2000$ rounds and repeated 20 times and we report averages with bootstrapped confidence intervals at 95% confidence level. We now present more metrics designed for a mixture of Gaussians (MoG) distribution, which are both more informative than the mean and interpretable than other existing approaches for the multimodal case.

Example 4.1. *Spread:* This metric calculates the difference between the maximum and minimum means among the components of GMM. It quantifies the range or spread of the component means within a single arm's distribution and is computed as $\max_m \mu_m - \min_m \mu_m$. This metric is $(\epsilon, 2)$-sensitive for all $\epsilon > 0$.

Example 4.2. *Modal Entropy:* This metric computes the entropy of the mixture weights associated with the components of the GMM. It measures the diversity or uncertainty in the contribution of each mode to the overall distribution. Higher entropy indicates that the weights are more evenly distributed among the components and the entropy is computed as, $-\sum_m w_m \log(w_m)$.

Example 4.3. *Mode Separation:* This metric calculates the maximum absolute difference between the means of adjacent components after sorting them. This metric indicates how distinct or well-separated the modes of the distribution are and is computed as $\max_{m,l} |\mu_m - \mu_l|$. This metric is also $(\epsilon, 2)-$sensitive for all $\epsilon > 0$.

Although there are no direct algorithms for the proposed metrics, we compare our Algorithm 1 (ETE-EM) and Algorithm 2 (TS-EM) against the following algorithms: (a) **CVaR-UCB** Xu et al. (2018), (b) **Sharpe Ratio-UCB** Khurshid et al. (2025), (c) **MeanUCB** (the standard UCB algorithm that optimizes for the mean, i.e., Algorithm 3, Chapter 7 of Lattimore and Szepesvári (2020); we label it MeanUCB to distinguish it from our mixture-metric algorithms), (d) **U-UCB**, which is an extension of Cassel et al. (2018): We first cluster the points using K-Means (we set K=4) algorithm and then use empirically estimate the parameters for each mode using an Empirical Distribution Performance Measures approach. The exploration length for ETE is fixed to $\mathsf{T}_e = (\frac{\rho\mathsf{T}}{N})^{2/3}$. as suggested by Theorem 3.4.

Insight 1: While some algorithms perform well on certain metrics (because the preferences align), ETE-EM and TS-EM obtains sub-linear regret across metrics and performs better than alternatives. Figure 3

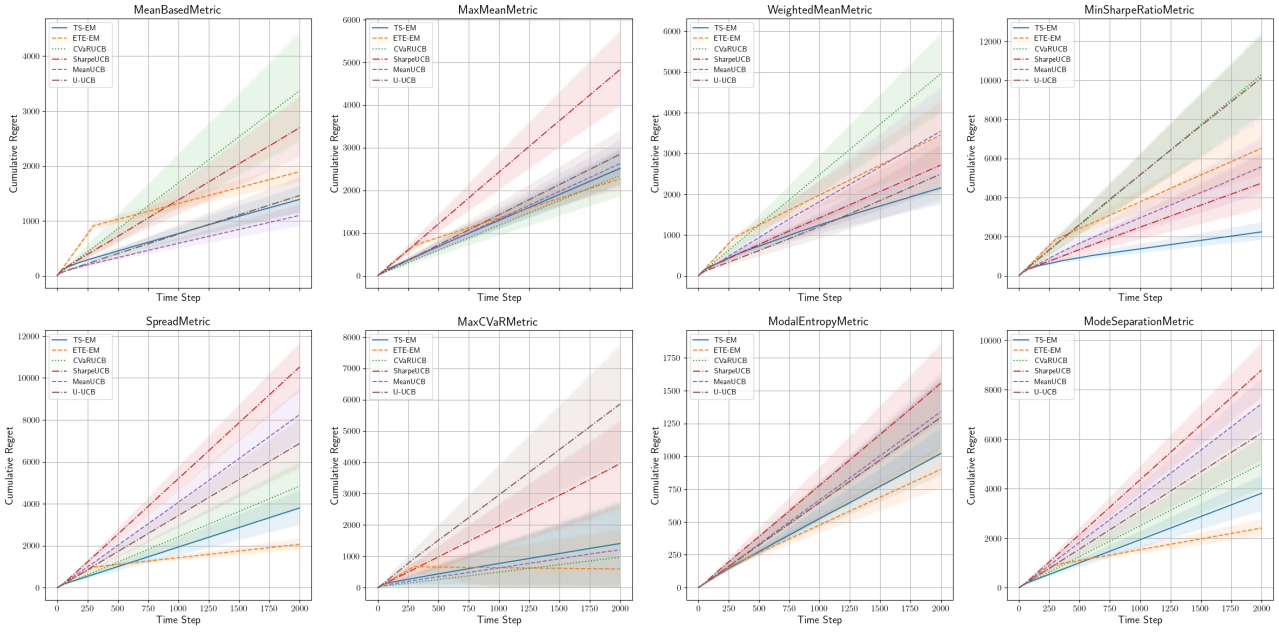


Figure 3: Expected cumulative regret for a multimodal bandit instance benchmarked on different parameteric metrics (See Section 4 for definitions) against standard risk-neutral and sensitive techniques designed for unimodal preferences. The instance has $N = 10$ arms, each with $M = 4$ Gaussian-mixture components with 500 random initializations. The horizon is $\mathsf{T} = 2000$ rounds and the ETE exploration budget is $\mathsf{T}_e = (\mathsf{T}/N)^{2/3}$.

shows cumulative regret versus time. All curves grow sub-linearly, consistent with the $\widetilde{O}(T^{2/3})$ guarantee established in Section 3. *U-UCB* has the closest performance however lacks a) theoretical guarantees in the multimodal regime and b) has more variance.

Case Study on Semi-Synthetic Data. In addition to our (A) *recommendation for subpopulations* which was discussed in the introduction, we illustrate another case-study upon which we base our experiments on. (B) *Selecting stores for executive field visits:* Weekly sales in retail outlets often bounce between two regimesordinary trade and promotion/holiday surges, plus occasional slumps from stock-outs or local events. The distribution of "value of a visit" is thus naturally bimodal or trimodal: a low-mode week is perfect for diagnosing operational issues, while a high-mode week is ideal for harvesting best practices. An executive with a handful of trips wants either extreme, not an average week. Yet if stores are ranked by expected sales (the mean) there is little spread and the variance treats the high-mode upside and low-mode downside symmetrically, even though the visits objective is asymmetric.

We show results on applications (A) and (B) using the MovieLens Harper and Konstan (2015) and Rossmann Store Sales (store) Knauer and Cukierski (2015). We detail the construction of the semi-synthetic dataset in Appendix E. For the MovieLens dataset, we use a mixture of linear reward models to generate a mixture of $M = 4$ Gaussians. For the store dataset, We consider weekly aggregated sales for each stores - 72% of which are multimodal (Average number of modes estimated by GMM-BIC is 2.42). For both datasets, we generate a total of 200 arms (movies and stores, respectively) and sample $N = 10$ arms randomly for each randomized instance (the results are averaged over 100 runs). The bandit instance is run for $\mathsf{T} = 1000$ rounds. We benchmark different instances of our ETE algorithm, assuming different numbers of components, to study how our methods perform under model misspecification. Results are summarized in Figure 4. We restrict the semi-synthetic benchmarks to methods that do not require contextual side-information; extending to contextual baselines (e.g., cold-start recommendation methods (Li et al., 2010a)) is a direction for future work.

Insight 2: Our algorithms achieve a better cumulative regret than *U-UCB* for different values of exploration periods for clustering. The max-mean metric in the MovieLens picks movies can be seen as doubling down on a movie's appeal within its most enthusiastic audience, and our algorithm is robust to the choice of a number of mixture components. For the retail dataset, the max-mean aligns with the order imposed by the mean, and therefore, the U-UCB method performs better. However, our algorithms give a sub-linear regret

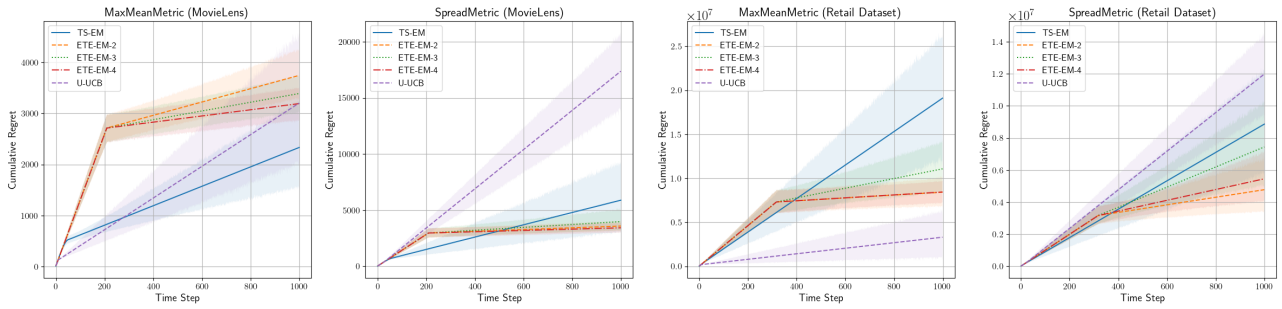


Figure 4: Expected cumulative regret on bandit instances using parameters approximated from rewards for decision making problems on public datasets, MovieLens and Rossmann Store Sales. ETE-EM and TS-EM obtains a sub-linear regret even with model-mispecification and outperforms U-UCB (Cassel et al. (2018)), except on Max-Mean for the Retail dataset where mean has the same ordering.

as well. The SpreadMetric maximises the spread of the distribution - an executive is interested in selecting stores with the highest difference so that they can intervene (or movies which can polarize more), and all instances of our algorithms perform reasonably well.

Conclusion and Limitations. We considered stochastic bandits with a multimodal reward distribution and developed metrics and techniques that model preferences that arise due to the multimodal structure. We proposed a variant of the thompson sampling algorithm and explore and exploit algorithm for minimizing regret with respect to metrics defined on the parameters of the multimodal distribution and show that each achieves sublinear regret. Future directions include mixture modeling the contextual setting, with linkages to kernelized bandits, and resolving the case where the number of mixture components is unknown. Regarding misspecification of M: if M is overestimated, the extra components collapse to near-zero weight under EM and sublinear regret is preserved provided the true components remain identifiable; if M is underestimated, the metric becomes biased, but our semi-synthetic experiments (Figure 4) demonstrate empirical robustness to moderate misspecification. The framework also extends to other parametric mixture families (e.g., Bernoulli, Poisson) where robust estimation procedures exist, with the EM step replaced by the corresponding family-specific estimator. Further, while employing multimodal reward distributions with known components, we fail to account for Markovian mode-switching inherent to restless bandits.

Future Work. Several directions remain open. First, there is a gap between the achievable regret lower bound ($\Omega(\mathsf{T}^{1/2})$) and the regret achieved by ETE-EM ($\widetilde{\mathsf{O}}(\mathsf{T}^{2/3})$) for the parametric case; closing this gap requires tighter anytime confidence sets for EM estimates. For $\bar{w}$, a better estimation bound for EM can reduce the dependence; our regret bounds are order optimal in T and N for non-adaptive algorithms, matching the order of explore-then-exploit algorithms for unimodal bandits. Second, our theoretical analysis assumes a known number of mixture components; when the number of components is unknown, Bayesian inference with a prior on the number of components is a natural approach, though only asymptotic guarantees currently exist for such techniques (Miller and Harrison, 2018). We note that Gaussian mixture models possess universal approximation capabilities and can approximate any continuous density on a compact subset of $\mathbb{R}$ arbitrarily well in the ℓ_1 norm given sufficiently many components. Third, the initialization requirement (A3) can be relaxed when the mixture likelihood is log-concave (implying a unique global maximum) or when the true distribution lies within the parametric family estimated by EM. In practice, multiple randomized restarts selecting the solution with the highest likelihood, or domain-specific heuristics such as k-means++ and spectral methods, produce initializations within the desirable basin of attraction with high probability; our experiments confirm robustness to random initialization. Fourth, the Gaussians considered are univariate, although the algorithms proposed here are extensible to a multivariate setting. Doing so causes scalability issues as most robust estimation guarantees are either exponential in the dimension or have an expensive polynomial dependency (Moitra and Valiant, 2010). Fifth, the metrics presented in this paper are not comprehensive; incorporating prior knowledge about the parameters through Thompson sampling with variational inference of MoG is a natural extension, though obtaining finite-sample guarantees for variational inference remains nontrivial. Finally, extending to contextual and non-stationary settings with Markovian mode-switching are natural next steps.

REFERENCES

Acharya, J., Jafarpour, A., Orlitsky, A., and Suresh, A. T. (2014). Near-optimal-sample estimators for spherical gaussian mixtures. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 13951403, Cambridge, MA, USA. MIT Press.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR.

Allenby, G. M., Arora, N., and Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, 35(3):384–389.

Arbas, J., Ashtiani, H., and Liaw, C. (2023). Polynomial time and private learning of unbounded gaussian mixture models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121 – 143.

Bouneffouf, D. and Feraud, R. (2025). Survey: Multi-armed bandits meet large language models. *arXiv preprint arXiv:2505.13355*.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. In *Séminaire de Probabilités de Strasbourg*, volume 13 of *Lecture Notes in Mathematics*, pages 342–363. Springer.

Bubeck, S. and Nicolò, C.-B. (2012).

Cassel, A., Mannor, S., and Zeevi, A. (2018). A General Approach to Multi-Armed Bandits Under Risk Criteria. In *Proceedings of the 31st Conference On Learning Theory*, pages 1295–1306. PMLR. ISSN: 2640-3498.

Chan, S.-O., Diakonikolas, I., Servedio, R. A., and Sun, X. (2013). Efficient density estimation via piecewise polynomial approximation.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016). Combinatorial multi-armed bandit with general reward functions. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA. PMLR.

Dasgupta, S. (1999). Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039)*, pages 634–644.

Daskalakis, C. and Kamath, G. (2014). Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1183–1213, Barcelona, Spain. PMLR.

Devroye, L., Mehrabian, A., and Reddad, T. (2023). The total variation distance between high-dimensional gaussians with the same mean.

Dwivedi, R., Khamaru, K., Wainwright, M. J., Jordan, M. I., et al. (2018). Theoretical guarantees for em under misspecified gaussian mixture models. *Advances in Neural Information Processing Systems*, 31.

Feng, Y., Huang, Z., and Wang, T. (2024). Lipschitz Bandits With Batched Feedback. *IEEE Transactions on Information Theory*, 70(3):2154–2176.

Ge, R., Huang, Q., and Kakade, S. M. (2015). Learning mixtures of gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 761770, New York, NY, USA. Association for Computing Machinery.

Guegan, D., Hassani, B., and Li, K. (2017). Impact of multimodality of distributions on var and es calculations. halshs-01491990.

Hao, B., Lattimore, T., and Wang, M. (2020). High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763.

Hardt, M. and Price, E. (2015). Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 753760, New York, NY, USA. Association for Computing Machinery.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):1–19.

Hershey, J. R. and Olsen, P. A. (2007). Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–317–IV–320.

Hong, J., Kveton, B., Zaheer, M., Ghavamzadeh, M., and Boutilier, C. (2022). Thompson sampling with a mixture prior. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7565–7586. PMLR.

Hong, Y., Xie, Q., Chen, Y., and Wang, W. (2023). Restless bandits with average reward: Breaking the uniform global attractor assumption. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 12810–12844. Curran Associates, Inc.

Huo, X. and Fu, F. (2017). Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4(11):171377.

Kaufmann, E., Cappé, O., and Garivier, A. (2014). On the complexity of a/b testing. In *Conference on Learning Theory*, pages 461–481. PMLR.

Khurshid, S., Abdulla, M. S., and Ghatak, G. (2024). Optimizing sharpe ratio: Risk-adjusted decision-making in multi-armed bandits.

Khurshid, S., Abdulla, M. S., and Ghatak, G. (2025). Optimizing sharpe ratio: risk-adjusted decision-making in multi-armed bandits. *Mach. Learn.*, 114(2).

Kleinberg, R., Slivkins, A., and Upfal, E. (2019). Bandits and experts in metric spaces. *J. ACM*, 66(4).

Knauer, F. and Cukierski, W. (2015). Rossmann store sales. `https://kaggle.com/competitions/rossmann-store-sales`. Kaggle.

Krimm, S., Scholz, H., and Wilkens, M. (2012). The sharpe ratio's market climate bias: Theoretical and empirical evidence from us equity mutual funds. *Journal of Asset Management*, 13:227–242.

Kwon, J. and Caramanis, C. (2020). The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2425–2487. PMLR.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press, 1 edition.

Li, J. and Barron, A. (1999). Mixture density estimation. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Li, J. and Schmidt, L. (2017). Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1302–1382. PMLR.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010a). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010b). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661670, New York, NY, USA. Association for Computing Machinery.

Liang, P. P., Cheng, Y., Fan, X., Ling, C. K., Nie, S., Chen, R. J., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R., and Morency, L.-P. (2023). Quantifying & modeling multimodal interactions: An information decomposition framework. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Liu, Y., Rekkas, M., and Wong, A. (2012). Inference for the sharpe ratio using a likelihood-based approach. *Journal of Probability and Statistics*, 2012(1):878561.

Luxenberg, E. and Boyd, S. (2024). Portfolio construction with gaussian mixture returns and exponential utility via convex optimization. *Optimization and Engineering*, 25(1):555–574.

Mathews, J. and Schmidler, S. C. (2024). Finite sample complexity of sequential Monte Carlo estimators on multimodal target distributions. *The Annals of Applied Probability*, 34(1B):1199 – 1223.

McDiarmid, C. (1989). *On the method of bounded differences*, page 148188. London Mathematical Society Lecture Note Series. Cambridge University Press.

McDiarmid, C. et al. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356. PMID: 29983475.

Moitra, A. and Valiant, G. (2010). Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102.

Najafi, A., Motahari, S. A., and Rabiee, H. R. (2020). Reliable clustering of bernoulli mixture models.

Nguyen, D., Prasad, A., Stengel-Eskin, E., and Bansal, M. (2025). LASeR: Learning to adaptively select reward models with multi-armed bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Noyes, C. W. (2008). *Analysis and Optimization of the Emergency Department at Beth Israel Deaconess Medical Center via Simulation*. PhD thesis, Massachusetts Institute of Technology.

Park, J. and Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Poor, H. (2013). *An Introduction to Signal Detection and Estimation*. Springer Texts in Electrical Engineering. Springer New York.

Qin, C., Ariu, K., and Proutiere, A. (2023). On universally optimal algorithms for a/b testing. *Advances in Neural Information Processing Systems (NeurIPS)*.

SABER, H. and Maillard, O.-A. (2024). Bandits with multimodal structure. *Reinforcement Learning Journal*, 5:2400–2439.

Salgia, S. (2023). Provably and practically efficient neural contextual bandits. In *International Conference on Machine Learning*, pages 29800–29844. PMLR.

Sani, A., Lazaric, A., and Munos, R. (2012). Risk-aversion in multi-armed bandits. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Shi, C., Yang, K., Chen, Z., Li, J., Yang, J., and Shen, C. (2024). Efficient prompt optimization through the lens of best arm identification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tatl, M., Mukherjee, A., A., P. L., Shanmugam, K., and Tajer, A. (2025). Preference-centric bandits: Optimality of mixtures and regret-efficient algorithms.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pages 495–517.

Urteaga, I. and Wiggins, C. H. (2022). Nonparametric gaussian mixture models for the multi-armed bandit.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 654–663.

Wang, Y., Chen, B., and Simchi-Levi, D. (2021). Multimodal dynamic pricing. *Management Science*, 67(10):6136–6152.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298.

Wu, Q., Wang, H., Li, Y., and Wang, H. (2019). Dynamic ensemble of contextual bandits to satisfy users' changing interests. In *The World Wide Web Conference*, WWW '19, page 20802090, New York, NY, USA. Association for Computing Machinery.

Xiong, Z., Camilleri, R., Fazel, M., Jain, L., and Jamieson, K. (2024). A/b testing and best-arm identification for linear bandits with robustness to non-stationarity. In *International Conference on Artificial Intelligence and Statistics*, pages 1585–1593. PMLR.

Xu, J., Haskell, W. B., and Ye, Z. (2018). Index-based policy for risk-averse multi-armed bandit. *arXiv preprint arXiv:1809.05385*.

Xu, K., Bastani, H., Goel, S., and Bastani, O. (2024). Stochastic bandits with relu neural networks. *arXiv preprint arXiv:2405.07331*.

Xue, B., Cheng, J., Liu, F., Wang, Y., and Zhang, Q. (2024). Multiobjective lipschitz bandits under lexicographic ordering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16238–16246.

Zhu, Z. and Van Roy, B. (2023). Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3636–3646.

CHECKLIST

We follow the AISTATS reproducibility checklist below. For each item, the answer is Yes / No / Not Applicable, with a brief justification or pointer to the relevant section.

1. For all models and algorithms presented, check if you include:

(a) *A clear description of the mathematical setting, assumptions, algorithm, and/or model.*
[Yes] See Section 2 (mixture-of-Gaussians bandit setting and metric definitions), Section 3 (Algorithms 1 and 2), and assumptions (A1)–(A3).

(b) *An analysis of the properties and complexity (time, space, sample size) of any algorithm.*
[Yes] Regret bounds are stated in Theorems 3.2, 3.4, and 3.5; sample complexity for EM estimation is given in Lemma F.1.

(c) *(Optional) Anonymized source code, with specification of all dependencies, including external libraries.*
[No] Source code is not released with this submission. The algorithms (ETE-EM and TS-EM) are specified in pseudocode in Algorithms 1 and 2 and rely only on standard EM / variational-EM implementations available in scikit-learn.

2. For any theoretical claim, check if you include:

(a) *Statements of the full set of assumptions of all theoretical results.*
[Yes] Assumptions (A1), (A2), (A3), and (B1) are stated in Sections 2 and 3 prior to each result.

(b) *Complete proofs of all theoretical results.*
[Yes] Full proofs of Theorems 3.2, 3.4, 3.5, and C.4 appear in Appendix F.

(c) *Clear explanations of any assumptions.*
[Yes] The role and tightness of each assumption is discussed where introduced (see Section 2 for separability and Section 3 for the EM-initialization condition).

3. For all figures and tables that present empirical results, check if you include:

(a) *The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).*
[No] Code is not released with this submission. The datasets used (MovieLens and Rossmann Store Sales) are publicly available, and the experimental protocol is specified in detail in Section 4 and Appendix E to enable independent re-implementation.

(b) *All the training details (e.g., data splits, hyperparameters, how they were chosen).*
[Yes] Exploration budgets, mixture-component counts, and per-arm sample sizes are reported in Section 4 and Appendix E (E.2, E.3).

(c) *A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).*
[Yes] Cumulative regret is defined in (1)–(3); shaded bands in Figures 3, 4, 6, 7 show bootstrapped 95% confidence intervals over 20–500 random seeds.

(d) *A description of the computing infrastructure used (e.g., type of GPUs, internal cluster, or cloud provider).*
[Yes] Experiments run on commodity CPU hardware; no GPUs or specialized infrastructure are required, as detailed in Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) *Citations of the creator if your work uses existing assets.*
[Yes] MovieLens (Harper and Konstan, 2015) and Rossmann Store Sales (Knauer and Cukierski, 2015) are cited in Section 4.

(b) *The license information of the assets, if applicable.*
[Yes] MovieLens is released for non-commercial research use; the Rossmann dataset is released under the Kaggle competition rules. Licenses are reproduced in Appendix E.

(c) *New assets either in the supplemental material or as a URL, if applicable.*
[Not Applicable] No new datasets are released; only derived bandit instances built from public sources.

(d) *Information about consent from data providers/curators.*
[Not Applicable] All datasets are publicly released by their creators with consent for research use.

(e) *Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.*
[Not Applicable] The datasets contain no personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) *The full text of instructions given to participants and screenshots.*
[Not Applicable] No crowdsourcing or human subjects were used.

(b) *Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.*
[Not Applicable] No human-subjects research was conducted.

(c) *The estimated hourly wage paid to participants and the total amount spent on participant compensation.*
[Not Applicable] No participants were compensated.

Appendix & Supplementary Materials

CONTENTS

A RELATED WORK AND NOVELTY

Two bandit frameworks related to the multimodal setting are *Contextual Bandits* and *Restless Bandits*.

Contextual bandits extend the multi-armed bandit framework by incorporating contextual information to make more personalized decisions (Li et al., 2010b). This approach has been successfully applied in various domains, including content recommendation and dynamic pricing. To derive theoretical guarantees, the reward is assumed to be a linear function of the context, and the arm vector Chu et al. (2011) and recently Wu et al. (2019) uses an ensemble of contextual bandit models to deal with dynamic reward distributions. If the relationship between reward and context is nonlinear, then kernel Salgia (2023); Valko et al. (2013) or neural supervised learners Xu et al. (2024); Zhu and Van Roy (2023) may be embedded to fit to the average reward distribution; however, deriving regret bounds may require inner-loop training and retraining which is not well-characterized by the theory. Restless bandits extend traditional multi-armed bandits by modeling arms as Markov decision processes with state transitions occurring even when inactive, enabling applications in resource allocation and recommendation systems(Whittle, 1988). Most theoretical guarantees are asymptotic, relying on restrictive assumptions like the uniform global attractor property or indexability(Hong et al., 2023), whereas our guarantees our finite-time bounds on the expected regret. Further, neither of these approaches enables the exploitation of the multimodal structure of reward distribution.

The different modes or mixture components can be considered as different i.i.d. contexts which have sampling probabilities equal to the weights of the mixture components. However, there are two key differences from a standard contextual bandit setting (Li et al., 2010a): a) Each context has a different reward distribution, and the learner is unaware of the context that a particular reward is sampled from; b) Usually, the contexts are given by a context vector with an additional constraint (e.g., linearity (Agrawal and Goyal, 2013)). However, we consider a finite number of contexts, each with a Gaussian reward distribution. For ease of analysis and simplicity of presentation, we consider the reward distribution of each arm as a mixture of Gaussians; however, other mixtures can also be considered. For distributions like the multivariate Bernoulli distribution, where robust sample efficient estimation procedure are known (Najafi et al., 2020) one can extend the methods presented in this paper.

Recently SABER and Maillard (2024) have looked at structured bandits where they look at multimodal structure among the bandit's arms, generalizing the notion of Lipschitz bandits Feng et al. (2024). They show regret bounds on the mean-based regret framework where arms can be locally optimal (with respect to their neighbors) but do not look at multimodal reward distributions (like the MoG) and subsequent metrics, which is the focus of this paper.

Parameter Estimation in Multimodal Distributions: Numerous methods have been proposed to estimate the parameters of a mixture model from samples; we summarize the most popular ones below, along with standard assumptions required to obtain statistical guarantees. Some of the methods present closeness of the distribution with the estimated parameters and the true parameters in terms of the total variation (TV) distance. However, one can use results such as Arbas et al. (2023); Devroye et al. (2023) to obtain the corresponding closeness in model parameters.

1. The expectation maximization (EM) algorithm (Balakrishnan et al., 2017; Dwivedi et al., 2018) iteratively optimizes the log-likelihood function and the corresponding parameters to obtain a locally optimal maximum

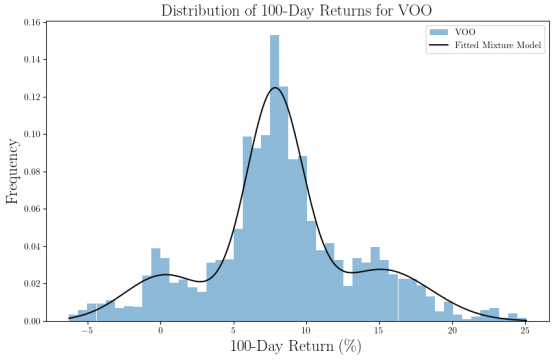


Figure 5: Realworld motivation for bandits with mixture distributions: Multimodal distribution of 100-day returns for VOO (a financial stock), showing the observed histogram of returns (blue bars) and a fitted mixture model (black curve) capturing the multimodal structure.

likelihood estimate. Although traditionally, the asymptotic convergence of EM has been studied, its sample complexity has been recently established in a few forms: (Balakrishnan et al., 2017) presents sample-based probably approximately correct (PAC) guarantees for a mixture of 2 Gaussians which state given a close enough initialization and a high enough SNR ratio, then the estimation error in the TV distance is $\mathsf{O}(\frac{1}{\sqrt{n}})$ with high probability. Further, recently Kwon and Caramanis (2020) showed that a sample-splitting variant converges to the true parameters with rate $\mathsf{O}(\frac{1}{\sqrt{\mathsf{T}}})$ with high probability under a separability assumption.

2. Acharya et al. (2014) presents a method based on Scheffe's estimator, which for a one-dimension mixture model needs at least $\tilde{\mathsf{O}}(\frac{\log(\frac{1}{\epsilon})}{\epsilon^2})$ samples to achieve an estimation error of ϵ with high probability in the TV distance under the assumption of separability of the mixture components. Daskalakis and Kamath (2014); Li and Schmidt (2017) provides similar bounds using tournament-based hypothesis selection and systems of polynomial inequalities-based algorithms, respectively.

3. Dasgupta (1999) presents a random projection based estimator which achieves an ϵ estimation guarantee of the mean parameter with $\mathsf{O}(\exp(\log^2 \frac{1}{\delta\epsilon}))$ with probability $1-\delta$.

There have been other works as well, e.g. Moitra and Valiant (2010) proposes a method of moment-based proposes a polynomial algorithm. Hardt and Price (2015) provides tight bounds for a mixture of two Gaussians, and Ge et al. (2015) exploits the structure in high-dimensional space to provide exact guarantees. *Shortcoming of these methods:* Note that these methods don't provide exact concentration bounds free of unknown constants for the Gaussian mixture model. Further, as mentioned in the system model, we assume that the number of mixture components is fixed and known to the learner. This is a restrictive assumption, and future work can look at relaxation for the same. The primary challenge is that current statistical methods for estimating the parameters when the number of mixture components is unknown rely on Bayesian inference with a prior on the number of mixture components, and only asymptotic guarantees yet exist for such techniques (Miller and Harrison, 2018).

A.1 Other work at intersection of GMMs and Bandits

Hong et al. (2022) studies the problem when the environment is sampled from a mixture distribution, however the rewards are still sampled from a sub Gaussian distribution and the objective is still minimizing the regret with respect to the arm with the highest mean. Similarly, Urteaga and Wiggins (2022) study the setting of reward model uncertainity and use GMMs to model the uncertainity. Both of the approaches are orthogonal to the current paper, however can be combined with our methods.

A.2 Application in Recommendation

We concentrate on a movierecommendation platform whose audience naturally splits into two user "moods":enthusiasts, who love niche titles, and casual viewers, who prefer mainstream fare. Our single met-

ric of interest is the bestsubgroup clickthrough ratein other words, how exciting a film can be for its most receptive mood, regardless of how it does on average. A traditional meanbased bandit chooses movies that make the whole audience moderately happy, so blockbusters dominate. Our method, by first using a brief exploration phase to separate the two moods and then ranking movies by the higher of their two moodspecific click rates, quickly uncovers more niche movies that are favourites of enthusiasts. Note that standard mean is a special case of our broader class of metrics as discussed in the paper.

To give the reviewer a concrete illustration, we focus on a simple movie-recommendation platform. Each movie i is shown to one user per round; the user either clicks (or gives a high rating) or not. We model the reward distribution for movie i as a mixture of two latent moods:

$$r_i \;\sim\; w_i^{(1)}\,\mathcal{N}\big(\mu_i^{(1)},\,\sigma^2\big) \;+\; (1-w_i^{(1)})\,\mathcal{N}\big(\mu_i^{(2)},\,\sigma^2\big),$$

where

1. $\mu_i^{(1)}$ = average click-through rate (CTR) among **Enthusiasts**,
2. $\mu_i^{(2)}$ = average CTR among **Casuals**,
3. $w_i^{(1)}$ = fraction of Enthusiasts for movie i, and σ^2 is shared (unspecified) variance.

$$\textbf{Max-Mean:}\quad m_{\max}(\theta_i) \;=\; \max\big\{\mu_i^{(1)},\,\mu_i^{(2)}\big\}.$$

This captures a desire to surface movies that, *in their best-serving subgroup*, perform maximallye.g. a niche cult hit that excites a small but passionate audience.

1. Traditional mean-UCB would rank movies by the overall mean $\bar{\mu}_i = w_i^{(1)}\mu_i^{(1)} + (1-w_i^{(1)})\mu_i^{(2)}$. Items with broad but shallow appeal can dominate, while true fan favorites may be under-recommended.
2. Our EM-based Max-Mean bandit first allocates T_e rounds to explore each movie, fits a 2-component GMM by EM, then in all remaining rounds picks the movie with the highest estimated $m_{\max}$.

A.3 Summary of Technical Novelty

1. Our lowerbound rederives the $\sqrt{T}$ rate under hiddenmode mixture rewards, where the KL term depends on latent parameters that are unobservablean aspect not covered by (Lattimore and Szepesvári, 2020, Theorem 15.2).
2. In the subGaussian setting the two test instances differ only in means; in our construction they differ in a single mixture component, forcing a new KL evaluation that reveals an unavoidable mixturedependent constant.
3. Standard ETC proofs assume highprobability concentration for empirical means; such bounds do not exist for EM estimates. Lemma F.1 allows us to use a finitesample biasdecay guarantee for EM and allows us to set the ρsensitive exploration budget.
4. Combining this guarantee with ETC yields the first $O(T^{2/3})$ regret for mixture metrics; prior ETC analyses cannot be "plugged in" because they have no handle on EM error.

These additions demonstrate that, while the proof template is classical, extending it to multimodal, latentvariable bandits requires nontrivial new ingredients.

B LIMITATIONS

B.1 Number of Components Robustness

While our theoretical analysis assumes a known number of Gaussian mixture components, we emphasize that our algorithm's core mechanisms especially the sampling strategy and reward estimation do not rely heavily on the specific parametric form of the underlying distributions. Empirically, our method performs well even under modest misspecification in the number of components (2 v/s 4) when compared to unimodal based methods.

To the best of our knowledge, there are no strong theoretical results (equivalent to GMM used in the paper) for a broader class of mixture classes (e.g. mixture of exponentials or Dirichlet process mixture models). Further, it is well-established that Gaussian Mixture Models (GMMs) possess universal approximation capabilities. In particular, GMMs can approximate any continuous probability density function on a compact subset of $\mathbb{R}^d$ arbitrarily well in the ℓ_1 norm, provided a sufficiently large number of components Li and Barron (1999); Park and Sandberg (1991).

B.2 Order Optimality of Parameters in Regret Gaurantees

For non-adaptive algorithms our regret bounds are indeed order optimal in T and N, since for T and N they match the order of order optimal explore than exploit algorithms. Similarly since ρ captures the sensitivity in the metric estimation with respect to the error in parameter estimation, it shares the same order as T. Since ρ acts as a sensitivity constant for error in the parameter estimation it's difficult to get rid of a dependence similar to T. We select the exploration period with respect to these parameters to optimize for optimality given the EM estimation procedure (and the associated statistical gaurantees). For $\bar{w}$ a better estimation bound for EM can reduce the dependence.

B.3 Initialization Requirements

The requirement can be relaxed in scenarios where the true underlying distribution lies within the parametric family estimated by the EM algorithm, or when the mixture likelihood function is log-concave, implying a unique global maximum. In such cases, the EM algorithm exhibits improved global convergence behavior, and strict initialization constraints may be unnecessary. However, when the likelihood surface is non-convex as is common in finite mixture models poor initialization can lead to convergence to suboptimal local modes.

In practice, one approach is to use multiple randomized restarts and select the solution with the highest likelihood, which increases the probability (with high confidence $1-\delta$) that the initialization is within an ϵ-neighborhood of the optimal basin of attraction. Alternatively, domain-specific heuristics (e.g., k-means++ initialization for Gaussian mixtures, spectral methods, or method-of-moments estimates) can be used to produce informed initializations that fall within this desirable region with high probability.

Other Limitations. First, while employing multimodal reward distributions with known components, it fails to account for Markovian mode-switching inherent to restless bandits. Second, a theoretical gap exists between the established lower bound and the achieved regret in parametric cases. The metrics presented in this paper offer more choices to the learner in modeling the required decision-making criteria; however, they are not comprehensive. The parametric metrics considered are sufficiently sensitive to perturbation in the distribution parameters, and the nonparametric case only covers the expectation of a suitably chosen function under the multimodal distribution. Finally, there is a gap in the achievable regret ($\Omega(\mathsf{T}^{1/2})$) and the regret achieved by our explore-then-exploit algorithm ($\widetilde{\mathsf{O}}(\mathsf{T}^{2/3})$) for the parametric case.

C NONPARAMETERIC METRICS

We also discuss *Regret Minimization for Non-Parametric Metrics* here, whence we introduce a nonparametric arm-ranking via expectation over a bounded function chosen by the learner, allowing the incorporation of inductive bias. The UCB-type algorithm in Section achieves $\widetilde{\mathsf{O}}(\mathsf{T}^{1/2})$ regret. Our nonparametric approach allows user-defined preference modeling over reward modes and can be seen as extension of Cassel et al. (2018). Here the nonparametric aspect refers to the fact that the metrics are no longer explicit functions of the parameters of the mixture family. Consider a continuous function $f : \mathbb{R} \to [0, U]$. We consider ranking the arms using the

Algorithm 3 UCB-f for regret minimization with respect to expectation of function f

Input: N arms $\mathcal{A}$, Number of rounds T, Confidence Parameter δ
for i in $[N]$ **do**
  for k in $[W]$ **do**
    Store reward $\mathcal{S}^{i_k} = \mathcal{S}^{i_k} \cup \{r_k\}$ *(Warm Start)*
  end for
end for
for k in $\mathsf{T} - W, \ldots, \mathsf{T}$ **do**
  Choose $a_k = \arg\max_{i\in[N]} \mathsf{UCB}_i^f(\mathcal{S}^i, \delta)$ (computed using (4))
  Observe reward $r_k \sim r(a_k)$
  Store reward $\mathcal{S}^{i_k} = \mathcal{S}^{i_k} \cup \{r_k\}$
end for

expectation of the function f under the arm distribution. Therefore, for arm i, compute the metric by $\mathbb{E}_i\{f\}$ where the expectation is w.r.t. the arm reward distribution which is multimodal Gaussian with parameters θ_i.

Denote the best arm as $i^* = \arg\max_{i\in[N]} \mathbb{E}_{\theta_i}\{f\}$. Define then the regret with respect to this metric:

$$\mathsf{R}_\mathsf{T}\big((\theta_i)_{i=1}^N, m_f, \pi\big) = \mathbb{E}\left\{\sum_k \mathbb{E}_i\{f\} - \mathbb{E}_{i_k}\{f\}\right\}, \tag{3}$$

where the outer expectation is taken with respect to the randomness in the algorithm. Some examples of nonparametric metrics of this form are given next.

Example C.1. *Domain Weighted Mean:* Consider the case where the designer has some domain knowledge of the rewards and wishes to optimize the average with respect to modes that lie in the interval $[a, b]$. One can consider a metric to be the expectation of the function $f = \frac{1}{1+\exp(-\xi(x+a))} - \frac{1}{1+\exp(-\xi(x+b))}$, under the multimodal distribution, $\mathbb{E}_\theta\{f\}$. Such a metric can be used in the dynamic pricing of a product with multimodal demand (Wang et al., 2021), where the demand in a certain interval $[a, b]$ is likely to maximize profits.

Example C.2. *Power Utility Function:* One could construct functions attributing different types of risk-taking behavior over different domains. One example of such a function is $f(x) = \frac{x^\alpha - a^\alpha}{b^\alpha - a^\alpha}\mathbb{1}_{[a,b]}(x) + \exp(-(x-b)^2)(\mathbb{1}_{[b,\infty)}(x))) + (\exp(|x-a|) - 1)(\mathbb{1}_{(-\infty,a)}(x))$, which attributes a risk-sensitive behavior over the domain $[a, b]$ and can model risk-aversion for $\alpha < 1$ and risk-seeking for $\alpha > 1$.

Example C.3. *Cluster-Based Harmonic Mean:* Weighted sums of certain harmonic functions could model preference around specific points. For e.g., $f(x) = \sum_{i\in[B]} \frac{1}{(x-c_i)^{\frac{4}{5}}}$ where $c_1, \ldots, c_B$ are B points of interest. This function is not bounded and therefore the theoretical arguments do not hold, however as our analysis show our algorithm works well empirically with it. Alternatively, an indicator function-based approach may be used.

Rational functions like $\frac{1}{1+x^2}$ can help model the behavior of returns normalized by volatility, and hyperbolic functions like sinh can help model leverage effects where downside moves might have different impacts than upside moves.

These functions can be used when the reward distribution is unimodal distributions as well; however, we are concerned with the case when the learner has some knowledge about where the modes would lie and seeks to exploit that information in modeling decisions. For example, a financial analyst might have projections on the different magnitudes of returns of an asset.

C.1 Regret in Nonparametric Metrics

We now propose an algorithm to minimize the nonparametric regret [cf. 3] with respect to the ranking defined by the expectation of the function f. The advantage of considering this nonparametric method is that one can make use of concentration inequalities (specifically McDiarmid's inequality (McDiarmid et al., 1989)) to obtain adaptive algorithms (like the UCB variant presented in Algorithm 3) which obtain a better theoretical guarantee than the explore than exploit presented in the previous section.

The main advantage of our algorithm for the multimodal setting is that it offers a choice to the learner to model preferences over different modes given some domain knowledge about them. The primary disadvantage

in this setting is that functions like CVaR might not be expressed a expectation of a function under the reward distribution and therefore approaches like empirical distribution performance measures (Cassel et al., 2018) can be tailored to a multimodal distribution.

C.2 Upper Confidence Bound-f and Regret Bound

We aim to minimize the nonparametric regret [cf. (3)] with respect to the best arm ranked by the expected value of the function f. We propose a UCB-type algorithm, UCB-f, summarized in Algorithm 3. First, we sample each arm a few times $W \ll \mathsf{T}/N$ (Line 2-5) to give a warm start to the algorithm and have informative confidence intervals. Then at each round, we compute UCB for each arm i as,

$$\mathsf{UCB}_i^f(\mathcal{S}^i, \delta) = \hat{f}(\mathcal{S}^i) + \sqrt{\frac{2\log(1/\delta)}{|\mathcal{S}^i|}}, \tag{4}$$

where $\hat{f}(\mathcal{S}^i) = \frac{1}{|\mathcal{S}^i|}\sum_{k=1}^{|\mathcal{S}^i|} f(\mathcal{S}_k^i)$ is the empirical average of the function f and $\mathcal{S}^i$ are the previous reward samples for arm i.

In each round, we pull the arm with highest UCB and store the reward sample (Lines 8-10). Next we present the regret associated with Alg. 3

Theorem C.4. *Consider the expected regret minimization problem on N bandit arms in T rounds. Consider a continuous function $f : \mathbb{R} \to [0, U]$, with best arm $i^* = \arg\max \mathbb{E}_i f$ and the corresponding regret [cf. (3)] Then the expected regret of the UCB-f algorithm (Algorithm 3) with a confidence parameter $\delta = 1/\mathsf{T}^2$ is upper bounded as,*

$$\mathsf{R}_\mathsf{T}\left((\theta_i)_{i=1}^N, m_f, \pi\right) = \tilde{\mathsf{O}}(\sqrt{NU\mathsf{T}}).$$

D FUTURE WORK

The Gaussians considered are univariate, although the algorithms proposed here are extensible to a multivariate setting. Doing so causes scalability issues with respect to the input dimension as most robust estimation guarantees are either exponential in the dimension or have an expensive polynomial dependency (Moitra and Valiant, 2010) and is an interesting problem for future work. Another method of incorporating prior knowledge about the parameters is the framework of Bayesian bandits and performing Thompson sampling. However, obtaining finite-sample guarantees and subsequent regret bounds for variational inference of MoG is a nontrivial extension.

E EXPERIMENTAL DETAILS AND ADDITIONAL EXPERIMENTS

This section outlines the methodology used to conduct the experiments, including data preparation, semi-synthetic data generation, bandit algorithm implementations, and evaluation metrics.

E.1 Datasets and Preprocessing

Two primary sources of data were utilized to create the semi-synthetic experimental environment.

E.1.1 MovieLens Dataset

- The MovieLens dataset was used, specifically focusing on user ratings.
- Ratings from a pool of highly active users were selected.
- A subset of movies rated by a common group of these users was identified.
- Movie descriptions for these common movies were fetched using an external API.
- Sentence embeddings for these movie descriptions were generated using a pre-trained transformer model (‘all-MiniLM-L6-v2‘).

- Principal Component Analysis (PCA) was applied to reduce the dimensionality of the movie embeddings to 50 components.
- The dataset was subsequently split into training and testing sets.

E.1.2 Online Retail Dataset

- An online retail dataset containing transactional data was loaded.
- Initial exploratory analysis was performed to investigate the multimodality of product sales.
- Weekly sales data for a sample of stores/products was aggregated.
- This aggregated sales data formed the basis for arm reward distributions in experiments involving this dataset.

E.2 Semi-Synthetic Data Generation for Bandit Arms

The core of the experiments relies on semi-synthetic data generation, where real-world data characteristics are used to define the reward distributions of the bandit arms.

E.2.1 MovieLens-based Arms

- **User Profile Vectors:** For each common user, a linear regression model was trained. The features (X) were the sentence embeddings of the training movies, and the target (y) was the user's rating for those movies, with small Gaussian noise $\mathcal{N}(0, 0.2)$ added. The learned regression coefficients for each user served as their "user vector."
- **Arm Reward Generation:**
  - Each arm in the bandit instance corresponded to a movie from the testing set.
  - The reward for pulling an arm (selecting a movie) was generated multimodally. A user was chosen from the common users based on a set of weights (e.g., uniform weights).
  - The reward was then calculated as the dot product of the chosen user's vector and the movie's embedding vector, with additional Gaussian noise $\mathcal{N}(0, 0.2)$ added.
  - This process was repeated a large number of times (e.g., 3000) for each test movie (arm) to create a pool of potential rewards.
- **Bandit Instance Creation:** For each experimental trial:
  - A set number of arms (e.g., 8) were selected by randomly choosing from the generated pool of potential rewards.
  - For each selected arm's sample rewards, a Gaussian Mixture Model (GMM) with a predefined number of components (e.g., 4) was fitted using the Expectation-Maximization (EM) algorithm.
  - The means, variances (diagonal covariances), and weights from the fitted GMM were used to define a multimodal bandit arm. The means were scaled by a constant factor (e.g., 10).
  - A bandit instance was composed of these multimodal arms.

E.2.2 Retail Dataset-based Arms

- Preprocessed weekly sales data, representing different stores or products, served as the base samples for arms.
- Similar to the MovieLens setup, for each trial:
  - A set number of arms (e.g., 8) were selected by randomly choosing from the sales data.
  - For each selected arm's sample data, a GMM with a predefined number of components (e.g., 2) and diagonal covariances was fitted.
  - The means, variances, and weights from this GMM defined a multimodal bandit arm.
  - A bandit instance was created with these arms.

E.3 Bandit Algorithms Implemented

The experimental framework included the following bandit algorithms.

- **ExploreThenCommitEM (ETC-EM):**
  - Implemented an Explore-Then-Commit strategy.
  - *Exploration Phase:* Pulled each arm a fixed number of times (e.g., 50 explore rounds) sequentially.
  - *Parameter Estimation:* After exploration, GMM parameters were estimated for each arm using the EM algorithm, fitting a specified number of Gaussian components (e.g., 2, 3, 4, or 5 components).
  - *Commitment Phase:* Committed to the arm with the highest score based on a chosen evaluation metric and the estimated GMM parameters. Variants of this algorithm differed in the number of components fitted during estimation.
- **U-UCB Algorithm:**
  - Implemented the U-UCB algorithm.
  - The selection rule was $\pi_t = \arg\max_i [U(\hat{F}_i(t-1)) + \phi^{-1}((\alpha \log t)/\tau_i(t-1))]$.
  - $U(\hat{F}_i(t-1))$ is a utility function applied to the empirical CDF of rewards from arm i. Various utility functions were implemented (e.g., based on max mean, min Sharpe ratio, spread), which used K-Means clustering to estimate GMM-like parameters from samples to compute the utility.
  - $\phi^{-1}(x) = \max\{2b \cdot (x/a)^{1/2}, 2b \cdot (x/a)^{q/2}\}$.
  - Standard parameters were used for a, b, q, and α (e.g., $a = 1.0, b = 1.0, q = 1.0, \alpha = 0.05$).

E.4 Metrics for Arm Evaluation and Regret Calculation

Several metrics were defined to evaluate the desirability of an arm based on its GMM parameters. These metrics were also used to calculate regret.

- **Implemented Metrics:**
  - *MaxMeanMetric:* Returns the maximum of the GMM component means: $\max(\mu_j)$.
  - *MinSharpeRatioMetric:* Computes the Sharpe ratio (μ/σ) for each component and returns the minimum (or its negative for maximization). Specifically, it calculates $\min_j(|\mu_j|/\sigma_j)$.
  - *SpreadMetric:* Computes the difference between the maximum and minimum GMM component means: $\max(\mu_j) - \min(\mu_j)$.
- **Cumulative Regret Calculation:**
  - The optimal arm was determined based on the true GMM parameters and the specific metric being used.
  - The optimal expected reward was the score of this optimal arm according to the metric.
  - For each arm pulled at time step t, its expected reward (score based on true parameters and the metric) was determined.
  - Instantaneous regret at step t was: E[optimal reward under metric] $-$ E[pulled arm's reward under metric].
  - Cumulative regret was the sum of instantaneous regrets over the time horizon.

This detailed setup allowed for the evaluation of different bandit strategies in handling arms with multimodal reward distributions, benchmarked against various preference metrics.

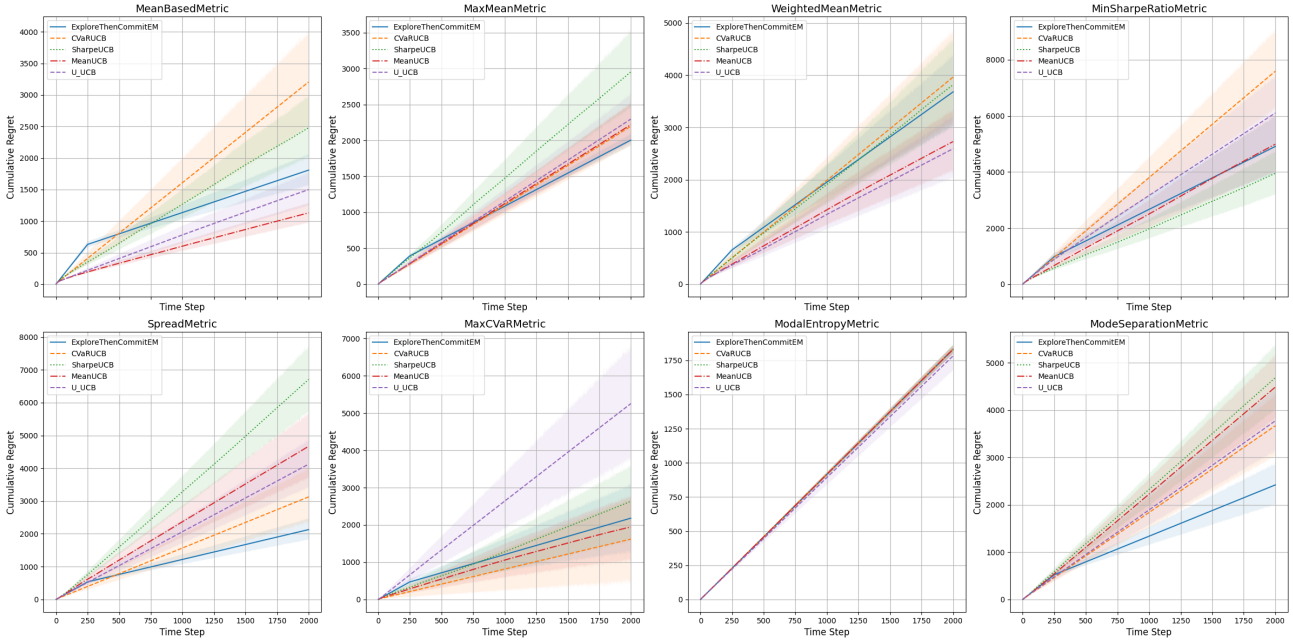


Figure 6: Effect of Number of Components: Same bandit instance as the main paper except with 8 components. Our ETE-EM algorithm still performs competitively on most metrics (except the modal-entropy metric, which does not necessarily have an upper bound on regret).

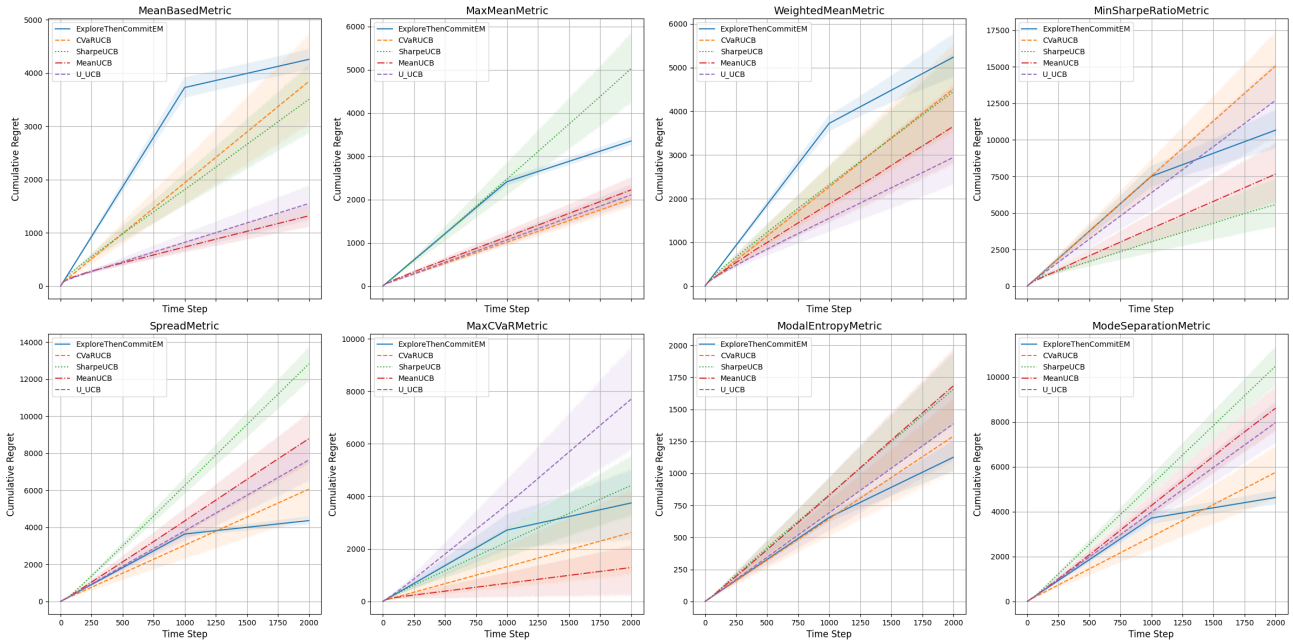


Figure 7: Effect of Number of Components: Same bandit instance as the main paper except with 20 arms. Our ETE-EM algorithm still performs competitively on most metrics (except the modal-entropy metric, which does not necessarily have an upper bound on regret).

E.5 Additional Experiments

F PROOFS

F.1 Proof of Theorem 3.5

Proof. Represent the instantaneous m_f-regret by,

$$\Delta_k := \mathbb{E}\left\{m_f^* - m_f^{a_k} \mid \mathcal{H}_{k-1}\right\},$$

the information gain by,

$$\mathcal{I}_k := \mathsf{MI}(a^\star; (a_k, r_k) \mid \mathcal{H}_{k-1}),$$

where MI is the conditional mutual information between the optimal arm, and the selected arm and reward.

$$\mathbb{P}(a^\star = i \mid \mathcal{H}_{k-1}) = \mathbb{P}(a_k = i \mid \mathcal{H}_{k-1}) \ \forall i \in [N]$$

which we can derive using,

$$\begin{aligned}&\mathbb{P}(a^\star = i \mid \mathcal{H}_{k-1}) = \mathbb{P}(a_\star(\theta) = i \mid \mathcal{H}_{k-1})\\ &= \mathbb{P}(a_\star(\hat{\theta}) = i \mid \mathcal{H}_{k-1}) = \mathbb{P}(a^k = i \mid \mathcal{H}_{k-1}) \ \forall i \in [N]\end{aligned}$$

We show that the information ratio given by $\Gamma := \frac{\Delta_k^2}{\mathbb{E}\{\mathcal{I}_k \mid \mathcal{H}_{k-1}\}}$ is bounded by $2NM\bar{w}\bar{\nu}\rho^2$.

Step 1: Bounding the Instantaneous Regret. Let $p_i = \mathbb{P}(a_k = i \mid \mathcal{H}_{k-1})$ and $\Delta_i = m_f^* - m_f^i$. By metric sensitivity (Definition 2.1):

$$\Delta_k = \sum_{i=1}^N p_i \Delta_i \leq \rho \sum_{i=1}^N p_i \|\theta^* - \theta^i\|$$

Step 2: Lower Bound on Information Gain. Under the separation assumption (A2) with $\delta_{\text{sep}} = \Omega\left(\sqrt{\frac{1}{\bar{w}\bar{\nu}}}\right)$, the information gain satisfies:

$$\begin{aligned}\mathbb{E}\{\mathcal{I}_k \mid \mathcal{H}_{k-1}\} &\geq \sum_{i=1}^N p_i \cdot D_{\text{KL}}(\mathbb{P}(r|a^*, \mathcal{H}_{k-1}) \| \mathbb{P}(r|i, \mathcal{H}_{k-1}))\\ &\geq \frac{1}{M}\sum_{i=1}^N p_i \cdot \min_{m\in[M]} \frac{(\mu_m^* - \mu_m^i)^2}{2\nu_m^i} \geq \frac{1}{2M\bar{\nu}} \sum_{i=1}^N p_i \cdot \min_{m\in[M]} (\mu_m^* - \mu_m^i)^2\end{aligned}$$

Step 3: Relating Parameter Distance to Metric Difference. By metric sensitivity, for arms differing in at least one mixture component:

$$\Delta_i^2 \leq \rho^2 \|\theta^* - \theta^i\|^2 \leq \rho^2 \bar{w}^{-1} \sum_{m=1}^M w_m (\mu_m^* - \mu_m^i)^2$$

Step 4: Bounding the Information Ratio. Using Cauchy-Schwarz inequality on the numerator:

$$\Delta_k^2 = \left(\sum_{i=1}^N p_i \Delta_i\right)^2 \leq N \sum_{i=1}^N p_i \Delta_i^2$$

Therefore:

$$\Gamma = \frac{\Delta_k^2}{\mathbb{E}\{\mathcal{I}_k \mid \mathcal{H}_{k-1}\}} \leq \frac{N\sum_{i=1}^N p_i \Delta_i^2}{\frac{1}{2M\bar{\nu}} \sum_{i=1}^N p_i \cdot \|\theta^* - \theta^i\|_{\min}^2} \leq 2NM\bar{w}\bar{\nu}\rho^2$$

Step 5: Deriving the Bayesian Regret Bound. The Bayesian regret is:

$$\mathsf{BR}_{\mathsf{T}}(\gamma) = \mathbb{E}_{\theta\sim\gamma}\left[\sum_{k=1}^{\mathsf{T}} \Delta_k\right]$$

Using the information ratio bound and the fact that $\sum_{k=1}^{\mathsf{T}} \mathcal{I}_k \leq H(a^*) \leq \log N$:

$$\mathsf{BR}_{\mathsf{T}}(\gamma) \leq \mathbb{E}\left[\sum_{k=1}^{\mathsf{T}} \sqrt{\Gamma \cdot \mathcal{I}_k}\right] \leq \sqrt{\mathsf{T}\cdot\Gamma}\cdot\sqrt{\sum_{k=1}^{\mathsf{T}}\mathbb{E}[\mathcal{I}_k]} \leq \sqrt{\mathsf{T}\cdot 2NM\bar{w}\bar{\nu}\rho^2\cdot\log N}$$

Therefore, the Bayesian regret for Thompson Sampling with mixture distributions satisfies:

$$\mathsf{BR}_{\mathsf{T}}(\gamma) = \mathsf{O}\left(\rho\sqrt{2NM\bar{w}\bar{\nu}\cdot\mathsf{T}\log N}\right)$$

□

F.2 Proof of Theorem 3.2

Proof Outline: The proof follows closely the proof of the lower bound for standard unimodal bandits (Theorem 15.2 of (Lattimore and Szepesvári, 2020)) – see Appendix F.2 for full proof. The main difference is in the regret decomposition step, where we use the metric sensitivity to get the difference in terms of the parameters, and in bounding the KL divergence for a pair of distinct mixture of Gaussians where we use the bound from (Hershey and Olsen, 2007). The proof establishes an $\Omega(\sqrt{NM\mathsf{T}})$ lower bound for multimodal Gaussian bandits by: (1) constructing two instances differing by Δ in the mean of one of the modes of an arm, (2) decomposing regret based on arm-pull probabilities, (3) applying the Bretagnolle-Huber inequality (Bretagnolle and Huber, 1979) to link regret with KL divergence, (4) bounding divergence via mixture properties and arm-pull expectations, and (5) optimizing Δ to balance terms and derive the final bound. This reveals the inherent complexity of learning in mixture environments scales with both arm count (N) and mixture components (M).

Proof. The proof works in five steps similar to the proof of Theorem 15.2 of Lattimore and Szepesvári (2020). The idea is to create and construct two instances such that any bandit algorithm will have regret greater than the lower bound on at least one of the two bandit instances.

1. Instance Construction: We construct two bandit instances. In case of unimodal bandits, these instances differ in the mean of one of the arms. However we make use of the following assumption,

Without loss of generality, we consider a generalization where the instances differ in mean of the first one of the modes of the arms. The first instance has the first arm with one mode with mean $\Delta < \epsilon$ and other modes at mean 0. Each mode has variance 1 and weight $\frac{1}{M}$. All the other arms have all their modes at mean 0. All the modes have a variance of 1. We can write the mean vectors of the arms as,

$$\mu = \left(\begin{pmatrix}\Delta\\ \vdots \\ 0\end{pmatrix}, \begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \ldots, \begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}\right).$$

where without loss of generality, the mean is non-zero for one of the coordinate for which the metric is sensitive according to B1. Note that this also implies that arm 1 is the best arm since the metric is atleast $\kappa\Delta$ better than any other arm.

Further now consider the following arm which has been played the least under the policy π,

$$j = \arg\min_i \mathbb{E}_\pi\left\{T_i(\mathsf{T})\right\}.$$

The second bandit instance is created as follows,

$$\mu' = \left(\begin{pmatrix}\Delta\\ \vdots \\ 0\end{pmatrix}, \ldots \underbrace{\begin{pmatrix}2\Delta\\ \vdots \\ 0\end{pmatrix}}_{\text{i-th arm}}, \ldots, \begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}\right).$$

This can be seen as an extension to the instance considered for the unimodal Gaussian case, however the upper bound on the KL divergence increases due to multiple unimodal distribution that we need to compare with. Here the i-th arm is better than any other arm.

2. Regret Decomposition: Let T_1 denote the number of pulls of arm 1. For both the bandit instances one can use regret decomposition to show the following,

$$\mathsf{R}_\mathsf{T}(\pi,\mu) \geq \mathbb{P}(T_1(\mathsf{T}) \leq \mathsf{T}/2)\frac{\kappa\Delta\mathsf{T}}{2}, \quad \mathsf{R}_\mathsf{T}(\pi,\mu') > \mathbb{P}(T_1(\mathsf{T}) > \mathsf{T}/2)\frac{\kappa\Delta\mathsf{T}}{2}$$

Let us show this for one of the terms to better see why κ comes out due to B1. By decomposing the regret on the event $T_1 \leq T/2$,

$$\mathsf{R}_\mathsf{T}(\pi,\mu) \geq \mathbb{P}(T_1(\mathsf{T}) \leq \mathsf{T}/2)\mathbb{E}[\sum_k m_f^* - m_f^{i_k}|T_1(\mathsf{T}) \leq \mathsf{T}/2]$$

Since the parameters between arm 1 (which is the best arm) and any other arm (played for atleast $T/2$ differ by Δ the difference in the metric m_f is atleast $\kappa\Delta$ (by assumption B1) and therefore $\mathbb{E}[\sum_k m_f^* - m_f^{i_k}|T_1(\mathsf{T}) \leq \mathsf{T}/2] \geq \kappa\Delta T/2$,

$$\mathsf{R}_\mathsf{T}(\pi,\mu) \geq \mathbb{P}(T_1(\mathsf{T}) \leq \mathsf{T}/2)\kappa\Delta\frac{T}{2}$$

Along similar lines for the bandit instance μ', the best arm is the $i-th$ arm and the difference is Δ between the arms at the same coordinate and hence,

$$\mathsf{R}_\mathsf{T}(\pi,\mu') \geq \mathbb{P}(T_1(\mathsf{T}) \geq \mathsf{T}/2)\kappa\Delta\frac{T}{2}$$

3. Applying Bretagnolle-Huber Inequality (Bretagnolle and Huber, 1979) ($\mathbb{P}(T_1(\mathsf{T}) \geq \mathsf{T}/2) + \mathbb{P}(T_1(\mathsf{T}) \leq \mathsf{T}/2) \geq \exp(-D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}))$): Then applying BretagnolleHuber inequality one gets,

$$\mathsf{R}_\mathsf{T}(\pi,\mu) + \mathsf{R}_\mathsf{T}(\pi,\mu') \geq \frac{\kappa\Delta\mathsf{T}}{4}\exp(-D(\mathbb{P}_\mu, \mathbb{P}_{\mu'})), \tag{5}$$

where $D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}))$ is the KL divergence.

4. KL Divergence Bound: The final step of the proof is to bound the KL divergence $D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}))$. For this we use the divergence decomposition (Lemma 15.1 from Lattimore and Szepesvári (2020)) and the upper bound on KL divergence for mixture models (Eq.9 from Hershey and Olsen (2007)). Further we use $\sum_i \mathbb{E}[N_i] = \mathsf{T}$ to obtain $\mathbb{E}[N_i] \leq \frac{\mathsf{T}}{N-1}$ We obtain the following bound,

$$D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) \leq \frac{\mathsf{T}}{N-1}2\sum\frac{(M-1)}{M^2}\frac{4\Delta^2}{2} \leq \frac{4\mathsf{T}\Delta^2}{(N-1)(M)}$$

5. Final Regret Bound: We choose $\Delta = \sqrt{\frac{M(N-1)}{8\mathsf{T}}}$, $\exp(-1/2) \geq \frac{16}{27}$ and substitute the bound in (5) to obtain the following,

$$\max(\mathsf{R}_\mathsf{T}(\pi,\mu), \mathsf{R}_\mathsf{T}(\pi,\mu')) \geq \frac{\mathsf{R}_\mathsf{T}(\pi,\mu) + \mathsf{R}_\mathsf{T}(\pi,\mu')}{2} \geq \frac{\kappa\sqrt{M(N-1)\mathsf{T}}}{27\sqrt{2}}.$$

In our lower bound construction, we assume for simplicity that each latent component has uniform weight $1/M$. This assumption is not essential to the argument and can be relaxed to allow arbitrary mixture weights bounded below by a constant $w_{\min} > 0$. Specifically, in Step 1 of the construction, the role of uniformity is to simplify the KL divergence computation in Step 4 (lines 862865), where only a single component's mean differs between the mixture distributions. If instead that component has weight $w_{m^\star}$, the KL divergence bound becomes $\frac{4Tw_{m^\star}\Delta^2}{N-1}$. By adjusting $\Delta = \sqrt{(N-1)/(8Tw_{m^\star})}$, we recover the same constant bound on the KL divergence, leading to a regret lower bound of $\Omega\big(\kappa\sqrt{(N-1)T/w_{\min}}\big)$. Thus, the result stated in Theorem 3.3 holds under a strictly weaker condition: it suffices that each component appears with non-zero weight, a relaxation that aligns with standard practice in the bandit literature.

Further, we clarify that assumptions (A1) and (A2) are not needed for the lower bound argument. Assumption (A1), which restricts the number of modes M and arms K, is not invoked in the construction or any part of the regret decomposition. Similarly, assumption (A2) which ensures that the mixture means μ_m lie in the interior of the simplex, is not used in the information-theoretic argument of Appendix F.1. The lower bound holds for arbitrary mixtures as long as there is one informative mode whose weight is bounded below and whose mean distinguishes the optimal arm. For simplicity and alignment with symmetric hard instances often used in the literature, we construct the instance with uniform weights and symmetric arms, but the argument remains valid in the more general non-uniform setting. □

F.3 Proof of Theorem 3.4

Proof Outline: The proof follows in three steps. First, the regret term is decomposed into two phases: exploration and exploitation. In the exploration phase, the regret is bounded by the exploration length, $\mathsf{O}(N\mathsf{T}_e)$. Further, the regret term for the exploitation phase is decomposed into two parts based on the success of the estimation procedure. The failure event has a small probability and when the estimation succeeds, the regret is of the order $\mathsf{O}(\rho\mathsf{T}/\mathsf{T}_e^{1/2})$. Finally, the exploration period is set to balance both the terms such that regret of the order $\mathsf{T}^{2/3}$. The main difference in our analysis is how metric sensitivity and the MoG estimation guarantees of EM (Lemma F.1) are employed.

Proof. The proof first splits the time horizon into an exploration phase (to estimate arm performances) and an exploitation phase (to leverage the estimated best arm). During exploration, regret scales linearly with the phase length. In exploitation, regret depends on whether the best arm is correctly identified: it vanishes under successful estimation but incurs a penalty proportional to the misidentification probability. Concentration inequalities bound this probability as a function of the exploration budget, and balancing exploration length against the estimation error trade-off optimizes the overall bound, yielding sublinear regret.

The pseudo-regret for the algorithm can be decomposed as,

$$\mathbb{E}[\mathsf{R}] = \mathbb{E}\left[\sum_{k=1}^{\mathsf{T}} m_f(\theta_{i_k}) - m_f(\theta_{i_*})\right] \leq \mathbb{E}\left[\sum_{k=1}^{\mathsf{T}_e} m_f(\theta_{i_k}) - m_f(\theta_{i_*})\right] + \mathbb{E}\left[\sum_{k=\mathsf{T}_e+1}^{\mathsf{T}} m_f(\theta_{i_k}) - m_f(\theta_{i_*})\right]$$

Let the event $\mathcal{E}$ is the event that the estimated maximum-f arm, $\hat{a} = \arg\max_{i,m} \hat{\mu_m^i}$ is the best possible arm,

$$\mathcal{E} := \bigcap_{i\in\mathcal{A}} \bigcap_{m} \{\mu_m^i \leq \mu_m^{i^*}\}. \tag{6}$$

Then,

$$\mathsf{R} \leq N\mathsf{T}_e\xi_1 + \mathsf{T}P(\mathcal{E})\epsilon_1 + \mathsf{T}P(\mathcal{E}^{\mathsf{C}})\xi_1$$

where ξ_1 is a constant and ϵ_1 is the error of estimation of the best arm. Note that by sensitivity, $\epsilon_1 \leq \rho\epsilon$, where ϵ is an error in parameter estimation.

Lemma F.1. *Suppose a mixture of M one-dimensional Gaussians has parameters $\theta = \{(\mu_m, \nu_m, w_m), m \in [M]\}$. Denote the iterates of the EM algorithm on iteration t by $\tilde{\theta}^{(t)} = \{(\tilde{\mu}_m^{(t)}, \tilde{\nu}_m^{(t)}, \tilde{w}_m^{(t)}), m \in [M]\}$. Assume (A1-A3) holds. Then with $n = \mathsf{O}(\frac{\log^2(\frac{M^2L}{\delta})}{\epsilon^2})$ samples, the sample splitting EM algorithm converges in the parameter estimates, $|\tilde{\mu_m}^{(L)} - \mu_m| \leq \epsilon$, $|\tilde{w_m}^{(L)} - w_m| \leq \epsilon$ and $|\tilde{\nu_m}^{(0)} - \nu_m| \leq \epsilon$ with prob. at least $1 - \delta - \frac{L}{M^{30}n^{\xi_1}}$ for some constant $\xi_1 > 0$.*

We now use Lemma F.1 and plug it in (6) and set the $\delta = 1/T^2$, therefore obtaining,

$$\mathbb{E}[\mathsf{R}] \leq N\mathsf{T}_e\xi_1 + \mathsf{T}P(\mathcal{E})\rho\epsilon + o(1)$$
$$\mathbb{E}[\mathsf{R}] \leq N\mathsf{T}_e\xi_1 + \mathsf{T}\rho\mathsf{O}\left(\frac{\bar{w}\log(M\mathsf{T})}{\sqrt{\mathsf{T}_e}}\right) + o(1)$$

Set $\mathsf{T}_e = N^{-2/3}\rho^{2/3}\bar{w}^{2/3}\mathsf{T}^{2/3}$ And we obtain regret,

$$\mathsf{R}_{\mathsf{T}}\left((\theta_i)_{i=1}^N, m_f, \pi\right) = \tilde{\mathsf{O}}(N^{1/3}\rho^{2/3}\bar{w}^{2/3}\mathsf{T}^{2/3}\log(M)).$$

□

F.4 Proof of Theorem C.4

The proof decomposes the expected regret into contributions from each arm, weighted by their suboptimality gaps and expected pulls. A good event is defined where the upper confidence bounds (UCBs) of suboptimal arms never exceed the optimal arms true mean, ensuring these arms are only pulled finitely often. We use McDiarmid's inequality (McDiarmid et al., 1989) to bound the probability of violating this event, while parameter tuning ensures the UCBs remain valid with high probability. By optimizing exploration parameters (e.g., confidence intervals and failure probabilities), the regret splits into a logarithmic term for suboptimal arm pulls and a residual error term, culminating in an overall bound scaling with $\sqrt{\mathsf{T}N\log\mathsf{T}}$.

Proof. Since the warm start period is much less than the total time period, for ease of analysis we do not account for the regret terms (and they can be absorbed into the constant error in the end).

First, we decompose the expected regret as a sum,

$$\mathsf{R}_k = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left\{N_i(k)\right\},$$

where $N_i(k)$ is the number of pulls of arm i till time k. Without loss of generality, let arm 1 be the optimal arm.

Good Event: Analogous to the proof of the UCB algorithm (Lattimore and Szepesvári, 2020), define the following event wherein the UCB never underestimates the best arm and the UCB of the arm i at time index t_a is lesser than the metric for the best arm.

$$G_i = \{\mathbb{E}_1 f < \min_{k\in[\mathsf{T}]} \mathsf{UCB}_1^f(\mathcal{S}_k^i)\} \cap \{\mathsf{UCB}_i^f(\mathcal{S}_{t_a}^i) < \mathbb{E}_1 f\}$$

where t_a is a suitably chosen time period for arm i.

We first show that $N_i(k) \le t_a$ if G_i occurs. This can be shown using contradiction: assume $N_i(k) > t_a$. Then there must exist a round $k \in [\mathsf{T}]$ where $N_i(k-1) = t_a$ and $a_k = i$. Therefore by definition of G_i, definition of $\mathsf{UCB}_i^f(\mathcal{S}_{k-1}^i)$ we have,

$$\mathsf{UCB}_i^f(\mathcal{S}_{k-1}^i) = \hat{f}(\mathcal{S}_{k-1}^i) + \sqrt{\frac{2\log(1/\delta)}{|\mathcal{S}_{k-1}^i|}} \qquad = \hat{f}(\mathcal{S}_{k-1}^i) + \sqrt{\frac{2\log(1/\delta)}{t_a}} + \mathsf{UCB}_1^f(\mathcal{S}_{k-1}^1),$$

where the last two inequality follows from the definition of G_i. which is a contradiction to the assumption arm i was chosen.

Further, the probability complement of the event G_i can be bounded using a union bound,

$$\mathbb{P}(G_i^c) \le \mathbb{P}(\left\{\mathbb{E}_1 f \ge \min_k \mathsf{UCB}_1^f(\mathcal{S}_k^i)\right\}) + \mathbb{P}(\left\{\mathbb{E}_1 f \le \mathsf{UCB}_i^f(\mathcal{S}_k^i)\right\}) \tag{7}$$

We first bound $\mathbb{P}(\left\{\mathbb{E}_1 f \ge \min_k \mathsf{UCB}_1^f(\mathcal{S}_k^i)\right\})$ using the union bound,

$$\begin{aligned}
&\mathbb{P}(\left\{\mathbb{E}_1 f \ge \min_k \mathsf{UCB}_1^f(\mathcal{S}_k^i)\right\}) \\
&\le \mathbb{P}(\left\{\mathbb{E}_1 f \ge \min_k \hat{f}(\mathcal{S}_k^i) + \sqrt{\frac{2\log(1/\delta)}{k}}\right\}) \le \sum_k \mathbb{P}(\mathbb{E}_1 f \ge \hat{f}(\mathcal{S}_k^i) + \sqrt{\frac{2\log(1/\delta)}{k}}) \le \mathsf{T}\delta
\end{aligned}$$

Next to bound $\mathbb{P}(\left\{\mathbb{E}_1 f \le \mathsf{UCB}_i^f(\mathcal{S}_k^i)\right\})$, consider t_a is chosen such that,

$$\Delta_i - \sqrt{\frac{2\log(1/\delta)}{U t_a}} \ge c\Delta_i$$

where Δ_i is the suboptimality gap and for $c \in (0,1)$ which is chosen later.

Using McDiarmid's Inequality(McDiarmid, 1989): For functions f with bounded differences ξ, independent random variables r_k, and ϵ:

$$\mathbb{P}(|\mathbb{E}_i f(r) - \frac{1}{k}\sum_{r\in\mathcal{S}_k^i} r| \geq \epsilon) \leq \exp(-\frac{|\mathcal{S}_k^i|\epsilon^2}{2\xi^2}) \implies \mathbb{P}(|\mathbb{E}_i f(r) - \hat{f}(\mathcal{S}_k^i)| > \epsilon) \leq \exp(-\frac{|\mathcal{S}_k^i|\epsilon^2}{2\xi^2})$$

Therefore we can write,

$$\mathbb{P}(\Big\{\mathbb{E}_1 f \leq \mathsf{UCB}_i^f(\mathcal{S}_k^i)\Big\}) \leq \exp(-\frac{-t_a^2 c^2 \Delta_i^2}{2U^2})$$

since the bounded difference for f is U.

Now substituting in 7, $\mathbb{P}(G_i^C) \leq \mathsf{T}\delta + \exp(-\frac{-t_a^2 c^2 \Delta_i^2}{2U^2})$

Now we decompose the expectation of the number of pulls for arm i with respect to the event G_i,

$$\mathbb{E}\left\{N_i(k)\right\} = \mathbb{P}(G_i)\mathbb{E}\left\{N_i(k)|G_i\right\} + \mathbb{P}(G_i^C)\mathbb{E}\left\{N_i(k)|G_i^C\right\} \leq t_a + \mathbb{P}(G_i^C)\mathsf{T}$$

Substituting the bound on $\mathbb{P}(G_i^C)$,

$$\mathbb{E}\left\{N_i(k)\right\} \leq t_a + \mathsf{T}\delta + \mathsf{T}\exp(-\frac{-t_a^2 c^2 \Delta_i^2}{2U^2})$$

Choose $t_a = \lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2 U} \rceil$ and $c = 1/2$ and $\delta = 1/\mathsf{T}^2$ then we may rewrite the previous inequality as

$$\mathbb{E}\left\{N_i(k)\right\} \leq 3 + \frac{16U\log(\mathsf{T})}{\Delta_i^2} \tag{8}$$

$$\mathsf{R}_\mathsf{T} = \sum_1^N \Delta_i \mathbb{E}\{N_i(\mathsf{T})\} = \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}\{N_i(\mathsf{T})\} + \sum_{i:\Delta_i>\Delta} \Delta_i \mathbb{E}\{N_i(\mathsf{T})\}$$

Further substituting equation (8),

$$\mathsf{R}_\mathsf{T} \leq \mathsf{T}\Delta + \sum_{i:\Delta_i>\Delta}\left(3\Delta_i + \frac{16U\log(\mathsf{T})}{\Delta_i}\right) \leq \mathsf{T}\Delta + \frac{16UN\log(\mathsf{T})}{\Delta} + 3\sum_i \Delta_i$$

Optimizing this bound by using, $\Delta = 4\sqrt{\frac{UN\log(\mathsf{T})}{\mathsf{T}}}$

$$\mathsf{R}_\mathsf{T} \leq 8\sqrt{\mathsf{T}UN\log(\mathsf{T})} + 3\sum_i \Delta_i$$

Note that the dependency on $\sum_i \Delta_i$ since we need to pull each arm at least once. □