Risk-sensitive Bandits: Arm Mixture Optimality and Regret-efficient Algorithms

Meltem Tatlı* RPI
Arpan Mukherjee* RPI
Prashanth L.A. IIT Madras
Karthikeyan Shanmugam Google Deepmind India
Ali Tajer RPI

Abstract

This paper introduces a general framework for risk-sensitive bandits that integrates the notions of risk-sensitive objectives by adopting a rich class of *distortion riskmetrics*. The introduced framework subsumes the various existing risk-sensitive models. An important and hitherto unknown observation is that for a wide range of riskmetrics, the optimal bandit policy involves selecting a *mixture* of arms. This is in sharp contrast to the convention in the multi-arm bandit algorithms that there is generally a *solitary* arm that maximizes the utility, whether purely reward-centric or risk-sensitive. This creates a major departure from the principles for designing bandit algorithms since there are uncountable mixture possibilities. The contributions of the paper are as follows: (i) it formalizes a general framework for risk-sensitive bandits, (ii) identifies standard risk-sensitive bandit models for which solitary arm selections is not optimal, (iii) and designs regret-efficient algorithms whose sampling strategies can accurately track optimal arm mixtures (when mixture is optimal) or the solitary arms (when solitary is optimal). The algorithms are shown to achieve a regret that scales according to $O((\log T/T)^{\nu})$, where T is the horizon, and $\nu > 0$ is a riskmetric-specific constant.

1 MOTIVATION & OVERVIEW

The canonical objective of stochastic multi-armed bandits is designing a sequence of experiments to identify the arm with the largest expected reward. This is a *utilitarian* view that lacks a mechanism for risk management. For instance, designing a financial portfolio based on maximizing the expected financial return is susceptible to incurring heavy financial losses due to tail events. Recovering from such losses translates into increased utility regret due to the losses and the upcoming missed opportunities.

Leveraging the recent advances in risk analysis, we provide a versatile risk-sensitive bandit framework that replaces the utilitarian reward functions with proper risk-sensitive counterparts. Specifically, we adopt the notion of *distortion riskmetrics* (DRs) (Wang et al., 2020), which encompass a broad range of widely-used *risk* and *deviation* measures. For a given function $h : [0,1] \to [0,1]$ that satisfies $h(0) = 0$, and a given probability measure with the cumulative distribution function (CDF) $\mathbb{Q}$, the distortion riskmetric U_h is defined as the *signed* Choquet integral

$$U_h(\mathbb{Q}) = \int_{-\infty}^{0} \Big(h\big(1-\mathbb{Q}(x)\big) - h(1)\Big)\,\mathrm{d}x + \int_{0}^{\infty} h\big(1-\mathbb{Q}(x)\big)\,\mathrm{d}x. \quad (1)$$

The function h, referred to as the *distortion function*, enables high versatility in specifying various risk measures and unifies various notions of risk, variability, and preference. Most such notions can be classified based on the monotonicity of the distortion function h as follows.

- **Monotone DRs:** There exists extensive literature on a subclass of DRs in which h is monotonically increasing and $h(1) = 1$. These include L-functionals in statistics (Huber and Ronchetti, 2009), Yaari's dual utilities in decision theory (Yaari, 1987), distorted premium principles in insurance (Denneberg, 1994), and distortion risk measures in finance (Kusuoka, 2001). Some widely used examples of monotone distortion riskmetrics are value-at-risk (VaR), expected shortfall (ES) (Artzner et al., 1999), and conditional VaR (CVaR) (Rockafellar and Uryasev, 2000).

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

- **Non-monotone DRs:** Unlike their monotone counterparts and despite their significance, the non-monotone DRs are far less investigated. These include different measures of variability, such as deviation measures (Rockafellar et al., 2006), mean-median deviation, inter-quantile range, Wang's right-tail deviation, inter-expected shortfall, Gini deviation (Wang et al., 2020), cumulative Tsallis past entropy (Zuo and Yin, 2024); preference measures such as Gini shortfall (Furman et al., 2017); and rank-based decision-making in decision theory (Quiggin, 1982).

Risk-sensitive bandits. Canonical bandit algorithms aim to maximize the expected cumulative reward. Risk-sensitive bandit algorithms, in contrast, seek to strike a balance between risk and reward, often formalized through proper risk measures. Notable recent advances on risk-sensitive bandits include adopting VaR and CVaR as risk measures (Gopalan et al., 2017; L.A. and Bhat, 2022; Chang and Tan, 2022; Cassel et al., 2018; Baudry et al., 2021; Tamkin et al., 2019; Tan and Weng, 2022; Liang and Luo, 2023).

Besides balancing the risk-reward dichotomy, there also exist other studies on risk-sensitive best arm identification (Kagrecha et al., 2019; L. A. et al., 2020) and contextual bandits (Huang et al., 2021).

The majority of the existing studies on risk-sensitive bandits, despite their distinct emphases and approaches, fall in the category of *monotone* DRs, with the only exception of mean-variance risk studied in (Chang and Tan, 2022; Cassel et al., 2018; Sani et al., 2013; Vakili and Zhao, 2016). The non-monotone DRs, while actively studied in the adjacent fields, e.g., reinforcement learning, remain unexplored for bandits. Some recent studies of non-monotone DRs in reinforcement learning include Gini deviation (Luo et al., 2023), and inter-ES (Han et al., 2022).

Contributions. The contributions are three-fold. We (i) design a risk-sensitive bandit framework that unifies all the monotone and non-monotone DRs; (ii) introduce novel *fixed-horizon* risk-sensitive algorithms; and (iii) establish that our algorithms are regret-efficient. Due to the high versatility of DRs in accommodating a significant range of risk measures, our framework subsumes most existing studies in risk-sensitive bandits through proper choices of the distortion function h. Some special cases are presented in Table 1. Our framework provides regret analysis for non-monotone DRs, which the existing literature cannot address, and recovers the regret guarantees for some of the existing monotone DRs. Finally, we note that by setting $h(u) = u$, the DR $U_h(\mathbb{Q})$ simplifies to the expected value of $\mathbb{Q}$, representing the standard risk-neutral objective.

Key observations. A hitherto unknown observation is that for certain DRs, the optimal policy is a *mixture* policy, i.e., there is no single optimal arm and the optimal strategy should be mixing arm selections according to carefully designed mixing coefficients. This contrasts the existing utility-centric or risk-sensitive bandits in which the optimal policy involves selecting a *solitary* arm (Cassel et al., 2018).

More specifically, we show that contrary to the convention of considering the optimal solution involving a *solitary* arm, the optimal solution of a non-monotone DR might involve a *mixture* of arms. In the latter case, we show that the risk of choosing any single arm is uniformly dominated by that of sampling based on a proper mixture of the arms. This guides a significantly different way of designing bandit algorithms, an integral part of which will be estimating mixing coefficients and designing arm selection strategies that accurately track the optimal mixture over time.

Technical challenge of learning optimal mixtures. Learning the optimal mixtures is a problem fundamentally distinct from identifying a solitary optimal arm. Searching for a solitary arm is guided by sequentially identifying and selecting an arm that optimizes a desired metric (e.g., upper confidence bound). In contrast, a mixture policy introduces two new challenges to the arm-selection policy. The one pertains to estimating K continuous-valued mixing coefficients and determining the optimal mixture from these estimates. The second dimension is the need for an efficient algorithm to track the optimal mixture. This necessitates balancing arm selections over time so that their mixture, in aggregate, conforms to the optimal one.

Addressing the first challenge (estimating mixture coefficients) requires estimating arms' CDFs over time, since the mixture coefficients depend on knowing the DRs associated with different arms, which are functions of the arms' CDFs. For the second challenge (tracking the optimal mixture), we design an arm-selection policy using the mixing estimates that deviates from the counterpart strategies with a solitary optimal arm.

Organization. The DR-centric risk-sensitive bandit framework is presented in Section 2. We present two algorithms in Section 3 based on the explore-then-commit (ETC) and upper-confidence bound (UCB) principles. The key processes in these algorithms are routines for learning the optimal mixtures. This new addition renders the designs and, especially, analyses of these algorithms significantly distinct from the standard ETC- and UCB-type analyses. The regret analyses are presented in Section 4 and the empirical evaluations are discussed in Section 5. The proofs are relegated to the appendices.

Table 1: Regret bounds of ETC-type ($\mathfrak{R}^{\mathrm{E}}_{\boldsymbol{\nu}}(T)$) and UCB-type ($\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T)$) algorithms, where $\varpi(T) \triangleq \sqrt{\frac{\log T}{T}}$.

Distortion Riskmetrics [a b c]	$h(u)$	parameter	$\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T)$	$\mathfrak{R}^{\mathrm{E}}_{\boldsymbol{\nu}}(T)$
Risk-neutral Mean Value	u		$O(\varpi^{1/2}(T))$	$O(\varpi^{2}(T))$
Dual Power	$1-(1-u)^s$	$s \in [2,+\infty)$	$O(\varpi^{1/2}(T))$	$O(\varpi^{2}(T))$
Quadratic	$(1+s)u - su^2$	$s \in [0,1]$	$O(\varpi^{1/2}(T))$	$O(\varpi^{2}(T))$
CVaR$_\alpha$ [d]	$\min\left\{\frac{u}{1-\alpha}, 1\right\}$		$O(\varpi^{1/2}(T))$	$O(\varpi^{2}(T))$
PHT	u^s	$s = 1/2$	$O(\varpi^{1/5}(T))$	$O(\varpi(T))$
Mean-Median Deviation	$\min\{u, 1-u\}$		$O(\varpi^{2\kappa}(T))$	$O(\varpi^{1/\beta}(T^\gamma))$
Inter-ES Range	$\min\left\{\frac{u}{1-\alpha}, 1\right\} + \min\left\{\frac{\alpha-u}{1-\alpha}, 0\right\}$	$\alpha = 1/2$	$O(\varpi^{2\kappa}(T))$	$O(\varpi^{1/\beta}(T^\gamma))$
Wang's Right-Tail Deviation	$\sqrt{u} - u$		$O(\varpi^{2\kappa}(T))$	$O(\varpi^{1/2\beta}(T^\gamma))$
Gini Deviation	$u(1-u)$		$O(\varpi^{4\kappa}(T))$	$O(\varpi^{2/\beta}(T^\gamma))$

[a] Summary of results for the K-arm Bernoulli bandit model. The more general results are presented in Section 4.
[b] In non-shaded rows, solitary arms are optimal. In the shaded rows, mixtures of arms are optimal.
[c] $\kappa \in (0, \frac{1}{2})$ and $\gamma \in (0,1)$ are constants associated with the distortion function h and are specified in Section 4.
[d] Arm means $< 1-\alpha$ for $\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T)$.

2 DR-CENTRIC FRAMEWORK

Bandit model. Consider a K-armed unstructured stochastic bandit. Each arm $i \in [K] \triangleq \{1, \cdots, K\}$ is endowed with a probability space $(\Omega, \mathcal{F}, \mathbb{F}_i)$, where $\mathcal{F}$ is the σ-algebra on $\Omega \subseteq \mathbb{R}_+$ and $\mathbb{F}_i$ is an *unknown* probability measure[1]. Accordingly, define $\mathbb{F} \triangleq \{\mathbb{F}_i : i \in [K]\}$. At time $t \in \mathbb{N}$, a policy π selects an arm $A_t \in [K]$ and the arm generates a stochastic sample X_t distributed according to $\mathbb{F}_{A_t}$. Denote the sequence of actions, observations, and the σ-algebra that policy π generates up to time $t \in \mathbb{N}$ by $\mathcal{X}^\pi_t \triangleq (X_1, \cdots, X_t)$, $\mathcal{A}^\pi_t \triangleq (A_1, \cdots, A_t)$, and $\mathcal{H}^\pi_t \triangleq \sigma(A_1, X_1, \cdots, A_{t-1}, X_{t-1})$.

Corresponding to any bandit instance $\boldsymbol{\nu}$ and policy π, $\mathbb{P}^\pi_{\boldsymbol{\nu}}$ denotes the push-forward measure on $\mathcal{H}^\pi_t$, and $\mathbb{E}^\pi_{\boldsymbol{\nu}}$ denotes the associated expectation. The sequence of independent and identically distributed (i.i.d.) rewards generated by arm $i \in [K]$ up to time $t \in \mathbb{N}$ is denoted by $\mathcal{X}_t(i) \triangleq \{X_t : A_t = i\}$. We define $\tau^\pi_t(i) \triangleq |\mathcal{X}_t(i)|$ as the number of times that policy π selects arm $i \in [K]$ up to time t.

Distortion riskmetric. We have a significant departure from the objective of evaluating the cumulative reward collected over time, and instead focus on evaluating the risk associated with the decisions made over time. We consider a generic distortion function $h : [0,1] \to [0,1]$, which is not necessarily monotonic. The distortion riskmetric associated with h is denoted by U_h and is defined in (1). Corresponding to any given policy π at time t, the overall risk associated with the sequence of arm selections $(A_1, \ldots, A_t)$ is given by

$$U_h\left(\sum_{s=1}^t \sum_{i\in[K]} \frac{\mathbb{1}\{A_s = i\}}{t}\, \mathbb{F}_i\right) = U_h\left(\sum_{i\in[K]} \frac{\tau^\pi_t(i)}{t}\, \mathbb{F}_i\right). \tag{2}$$

We highlight that the DR U_h depends on the full descriptions of the arms' statistical models (CDFs).

[1] We focus on positive-valued random variables. Extension to include negative values is straightforward.

Oracle policy. An oracle policy can accurately identify the optimal sequence of arm selections $\{A_t : t \in \mathbb{N}\}$. Given the structure in (2), designing an optimal oracle policy is equivalent to determining the optimal mixing of the CDFs. For a bandit instance $\boldsymbol{\nu} \triangleq (\mathbb{F}_1, \cdots, \mathbb{F}_K)$, the vector of optimal mixture coefficients is denoted by

$$\boldsymbol{\alpha}^\star_{\boldsymbol{\nu}} \in \operatorname*{arg\,max}_{\boldsymbol{\alpha}\in\Delta^{K-1}} U_h\left(\sum_{i\in[K]} \alpha(i)\, \mathbb{F}_i\right), \tag{3}$$

where Δ^{K-1} is a K-dimensional simplex. When clear from the context, we use the shorthand $\boldsymbol{\alpha}^\star$ for $\boldsymbol{\alpha}^\star_{\boldsymbol{\nu}}$. We note that the oracle considered in (3) is omniscient, i.e., it is aware of all the distributions $\mathbb{F} = \{\mathbb{F}_i : i \in [K]\}$ and the mixing policy $\boldsymbol{\alpha}^\star_{\boldsymbol{\nu}}$ that optimizes the DR. Throughout the rest of the paper, for any given vector $\boldsymbol{\alpha}$ and set of CDFs $\mathbb{F} = \{\mathbb{F}_i : i \in [K]\}$, we define

$$V(\boldsymbol{\alpha}, \mathbb{F}) \triangleq U_h\left(\sum_{i\in[K]} \alpha(i)\, \mathbb{F}_i\right). \tag{4}$$

Next, we introduce a motivating example for designing mixture policies. Consider the widely-used Gini deviation. We show that the optimal policy is a mixture. The distortion function for Gini deviation is $h(u) = u(1-u)$ for $u \in [0,1]$ (Wang et al., 2020).

Lemma 1 (Gini Deviation) *Consider a two-arm Bernoulli bandit model. For a given $p \in [0,1]$, the arms' distributions are* Bern(p) *and* Bern$(1-p)$. *For distortion function $h(u) = u(1-u)$, we have*

$$\sup_{\boldsymbol{\alpha}\in\Delta^{K-1}} V(\boldsymbol{\alpha}, \mathbb{F}) = U_h(0.5(\mathbb{F}_1 + \mathbb{F}_2)) > \max_{i\in\{1,2\}} U_h(F_i).$$

Hence, **(i) the optimal value of this DR is achieved when samples come from the mixture distribution $\frac{1}{2}(\mathbb{F}_1 + \mathbb{F}_2)$, and (ii) the DR associated with this mixture is strictly larger than those associated with the individual arms.**

Mixture-centric Objective. Motivated by the observation that, generally, for a given distortion function h, the optimal DR might be achieved by a *mixture* of arm distributions, we provide a general mixture-centric framework in which bandit policies can take mixture forms. This also subsumes solitary policies as a special case when the mixing mass is placed on one arm.

By the oracle policy's definition $\boldsymbol{\alpha}^\star_{\boldsymbol{\nu}}$ in (3), for a bandit instance $\boldsymbol{\nu} \triangleq (\mathbb{F}_1, \cdots, \mathbb{F}_K)$, our objective is to design a bandit algorithm with minimal regret with respect to the DR achieved by the oracle policy. Hence, for a given policy π and horizon T, we define the regret as the gap between the DR achieved by the oracle policy and the *average* DR achieved by π, i.e.,

$$\mathfrak{R}^\pi_{\boldsymbol{\nu}}(T) \triangleq V(\boldsymbol{\alpha}^\star, \mathbb{F}) - \mathbb{E}^\pi_{\boldsymbol{\nu}}\Big[V\Big(\frac{1}{T}\boldsymbol{\tau}^\pi_T, \mathbb{F}\Big)\Big] . \tag{5}$$

Assumptions. For each arm $i \in [K]$, let $\mathcal{P}^1(\Omega)$ denote the set of all probability measures with finite first moment on Ω. We assume the maximum value the DRs can take is bounded, i.e., $U_h \le B$, where B is a positive constant. Additionally, assume that the distributions $\mathbb{F}_i$ are 1−sub-Gaussian and belong to the metric space $(\mathcal{P}^1(\Omega), \|\cdot\|_{\mathrm{W}})$, where $\|\mathbb{G} - \mathbb{S}\|_{\mathrm{W}}$ denotes the 1−Wasserstein distance between distributions $\mathbb{G}$ and $\mathbb{S}$. We define W as the maximum ratio between Wasserstein and total variation distances between any two mixture of arms, i.e.,

$$W \triangleq \max_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}\in\Delta^K} \frac{1}{\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_1}\Big\|\sum_i \alpha_i\mathbb{F}_i - \sum_j \beta_j\mathbb{F}_j\Big\|_{\mathrm{W}} .$$

In Theorem 6 in Appendix C, we show that W is bounded. We denote the convex hull of the set of distributions $\{\mathbb{F}_i : i \in [K]\}$ by Ξ, and assume U_h satisfies the following notions of continuity.

Definition 1 (Hölder continuity) *The DR U_h is Hölder continuous with exponent $q \in (0, 1]$, if for all distributions $\mathbb{G}_1, \mathbb{G}_2 \in (\Xi, W_1)$, there exists a finite $\mathcal{L}_{\mathrm{H}} \in \mathbb{R}_+$ such that*

$$U_h(\mathbb{G}_1) - U_h(\mathbb{G}_2) \le \mathcal{L}_{\mathrm{H}} \|\mathbb{G}_1 - \mathbb{G}_2\|^q_{\mathrm{W}} . \tag{6}$$

Definition 2 (Mixture Hölder continuity) *The DR U_h is Hölder continuous with exponent $r \in \mathbb{R}_+$ if for a set of distributions $\{\mathbb{F}_i : i \in [K]\}$ with the convex hull Ξ and a set of mixing coefficients $\{\alpha^\star(i) : i \in [K]\}$ for $\mathbb{F}^\star \triangleq \sum_{i\in[K]} \alpha^\star(i)\mathbb{F}_i$ and any distribution $\mathbb{G} \in (\Xi, W_1)$, there exists a finite $\mathcal{L}_{\mathrm{MH}} \in \mathbb{R}_+$ such that*

$$U_h(\mathbb{F}^\star) - U_h(\mathbb{G}) \le \mathcal{L}_{\mathrm{MH}} \|\mathbb{F}^\star - \mathbb{G}\|^r_{\mathrm{W}} . \tag{7}$$

Finally, we define $\mathcal{L} \triangleq \max\{\mathcal{L}_{\mathrm{H}}, \mathcal{L}_{\mathrm{MH}}\}$ as a unified Hölder constant. Many of the widely used DRs are Hölder continuous. In Table 2 in Appendix A.1, we present a list of riskmetrics, and specify their associated Hölder exponents q and r.

3 DR-CENTRIC ALGORITHMS

Overview. We present a mixture-centric algorithm designed to select and sample the arms based on a mixture policy. A mixture policy consists of the following two sub-routines that will be integral parts of its algorithm design.

- **Estimate mixing coefficients:** The first routine forms estimates of the mixing coefficients. These estimates are updated over time based on the collected empirical DR U_h and are expected to get refined and eventually track the optimal mixture $\boldsymbol{\alpha}^\star$ over time.
- **Track mixtures:** The second routine translates the mixing coefficient estimates into arm selection decisions. These decisions have to balance two concurrent objectives. First, there is a need to improve the estimates of the mixing coefficients. Second, the arm selection rules need to ensure that, in aggregate, their selections track the mixing estimates. This necessitates an explicit mixture tracking rule that is regret-efficient.

We provide two ETC- and UCB-type algorithms, each with its performance or viability (model information) advantages. Furthermore, we provide a variation of the UCB-type algorithm to address some computational challenges of the base design. To this end, we provide a few definitions and notations that will be used to describe and analyze the algorithms. We use the shorthand $\pi \in \{\mathrm{E}, \mathrm{U}, \mathrm{C}\}$ to refer to the ETC-type, UCB-type, and the computationally efficient UCB-type policies, respectively.

An important routine in these algorithms is estimating the arms' CDFs accurately. All three algorithms form estimates of arms' CDFs. We denote the empirical estimate of $\mathbb{F}_i$ at time t using policy π by

$$\mathbb{F}^\pi_{i,t}(x) \triangleq \frac{1}{\tau^\pi_t(i)} \sum_{s\in[t]:A_s=i} \mathbb{1}\{X_s \le x\} . \tag{8}$$

The following lemma provides a concentration bound on the empirical CDFs in the 1−Wasserstein metric.

Lemma 2 *For any policy π and arm $y \in \mathbb{R}_+$ we have*

$$\mathbb{P}^\pi_{\boldsymbol{\nu}}\Big(\big\|\mathbb{F}^\pi_{i,t} - \mathbb{F}_i\big\|_{\mathrm{W}} > y\Big) \le 2\exp\left(-\frac{\tau^\pi_t(i)}{256\mathrm{e}}\Big(y - \frac{512}{\sqrt{\tau^\pi_t(i)}}\Big)^2\right) . \tag{9}$$

Proof: See (L.A. and Bhat, 2022, Lemma 8). ■

3.1 Discrete Mixture Coefficients

The optimal mixing coefficients $\boldsymbol{\alpha}^\star$ can take arbitrary irrational values. Sampling frequencies, on the other hand, will always be rational values. Hence, estimating

$\boldsymbol{\alpha}^\star$ beyond a certain accuracy level is not beneficial as it cannot be implemented. We define ε to specify the desired level of fidelity for each mixing coefficient. Accordingly, we specify Δ_ε^{K-1} by uniformly discretizing each dimension of Δ^{K-1} with the ε discrete level. The discretization schemes are slightly different for the ETC- and UCB-type algorithms and the specifics are provided in Sections 3.2 and 3.3. We define $\mathbf{a}^\star$ as the counterpart of the $\boldsymbol{\alpha}^\star$ in the discrete simplex, i.e.,

$$\mathbf{a}^\star \triangleq \arg\max_{\mathbf{a}\in\Delta_\varepsilon^{K-1}} V(\mathbf{a},\mathbb{F}) \ . \tag{10}$$

Finally, we define the *minimum sub-optimality gap* with respect to the discretization level ε as

$$\Delta_{\min}(\varepsilon) \triangleq \min_{\mathbf{a}\in\Delta_\varepsilon^{K-1}:\mathbf{a}\neq\mathbf{a}^\star} \{V(\mathbf{a}^\star,\mathbb{F}) - V(\mathbf{a},\mathbb{F})\} \ . \tag{11}$$

3.2 Risk-sensitive ETC for Mixtures

We propose the **R**isk-**S**ensitive **ETC** for **M**ixtures (RS-ETC-M) algorithm, following the ETC principles, albeit with important deviations needed to accommodate an implementation of mixture policies.

The RS-ETC-M algorithm consists of an initial *exploration* phase during which the arms are sampled uniformly for a fixed interval to form high-fidelity estimates of the arms' CDFs (in contrast to canonical ETC that estimates arms' mean values). The duration of this phase depends on the minimum sub-optimality gap. Subsequently, in the next phase, the ETC algorithm *commits* to a fixed policy, which is a mixture of arms with pre-fixed mixing coefficients. The mixing coefficients are selected to maximize the estimate of the DR. We show that with a high probability, they are equal to $\mathbf{a}^\star$. The main processes of this algorithm are explained next and its pseudocode is presented in Algorithm 1.

Discretization. We specify Δ_ε^{K-1} by uniformly discretizing the each coordinate of Δ^{K-1} into intervals of length ε, i.e.[2],

$$\Delta_\varepsilon^{K-1} \triangleq \{\varepsilon\mathbf{n} : \mathbf{n}\in\{\mathbb{N}\cup\{0\}\}^K \ , \mathbf{1}^\top\cdot\varepsilon\mathbf{n}=1\} \ . \tag{12}$$

Explore (estimate mixing coefficients). The purpose of this phase is to form high-confidence estimates of the empirical arm CDFs. We specify a time instant $N(\varepsilon)$ that determines the duration of the exploration phase, and that is the instance by which we have confident-enough CDF estimates. The arms are selected uniformly, each $\lceil \frac{1}{K}N(\varepsilon)\rceil$ times. For a DR with a distortion function that has Hölder continuity exponent q and constant $\mathcal{L}$, the time instant $N(\varepsilon)$ becomes a function of q, $\mathcal{L}$, the horizon T, and the minimum sub-optimality gap $\Delta_{\min}(\varepsilon)$ as specified below.

$$N(\varepsilon) \triangleq 256Ke\left(\frac{2K\mathcal{L}}{\Delta_{\min}(\varepsilon)}\right)^{\frac{2}{q}} \times\left[\frac{32}{\sqrt{e}}+\log^{\frac{1}{2}}\left(2KT^2\left(\varepsilon^{-(K-1)}+1\right)\right)\right]^2 \ . \tag{13}$$

This implies that $N(\varepsilon)$ scales as $O(\log T)$. We also define $M(\varepsilon)\triangleq\frac{N(\varepsilon)}{\log T}$, which based on (13) scales as $O(1)$. Next, using the collected samples during exploration, the RS-ETC-M algorithm constructs the empirical estimates of arms' CDFs, as specified in (8).

Choosing the mixing coefficient: At time instant $N(\varepsilon)$, the RS-ETC-M algorithm identifies the discrete mixture coefficients that maximize the DR using the empirical CDFs. These coefficients are denoted by $\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}$, where

$$\mathbf{a}^{\mathrm{E}}_t \in \arg\max_{\mathbf{a}\in\Delta_\varepsilon^{K-1}} U_h\left(\sum_{i\in[K]} a(i)\,\mathbb{F}^{\mathrm{E}}_{i,t}\right) . \tag{14}$$

Commit (track mixtures). For the remaining sampling instants $t\in[N(\varepsilon),T]$, the RS-ETC-M algorithm commits to selecting the arms such that their selection frequencies are as close as possible to $\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}$. Until time instant $N(\varepsilon)$ the algorithm samples all arms uniformly. Uniform sampling results in some arms being sampled more than what the policy $\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}$ dictates and hence, these arms will not be sampled again. Having over-sampled arms implies that some arms are under-sampled. In that case, these arms would be sampled such that the number of times they are chosen converges to the mixing coefficient $\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}$.

To formalize how to track the mixture, let S be the set of first $K-1$ arms (or any desired set of arms). For each arm $i\in S$, the algorithm calculates the required number of samples, which is $T\times a_{N(\varepsilon)}(i)$ for a horizon T. The sampling procedure proceeds as follows:

1. if $Ta_{N(\varepsilon)}(i) > \lceil\frac{N(\varepsilon)}{K}\rceil$ (insufficient exploration), then, the arm i is sampled according to $Ta_{N(\varepsilon)}(i)$ before the algorithm moves on to the next arm.

2. if $Ta_{N(\varepsilon)}(i)\leq\lceil\frac{N(\varepsilon)}{K}\rceil$ (sufficient exploration), then, the arm i is skipped.

The remaining sampling budget is allocated to arm K.

Discussion. An important advantage of the RS-ETC-M algorithm is its computational simplicity. The algorithm involves uniform arms exploration for a finite interval followed by committing to a mixture estimate based on the data in the exploration phase. In Section 3.3, we will discuss that the relative performances of RS-ETC-M versus the UCB-type counterpart depends on the choice of DR, and neither has a uniform regret advantage over the other.

[2]When $\frac{1}{\varepsilon}$ is not an integer, we make up for the deficit/excess of the weights in the last coordinate.

Algorithm 1 RS-ETC-M

Input: Minimum gap $\Delta_{\min}(\varepsilon)$, horizon T
Sample each arm $\lceil N(\varepsilon)/K \rceil$ times and obtain observation sequences $\mathcal{X}_{\lceil N(\varepsilon)/K \rceil}(1), \cdots, \mathcal{X}_{\lceil N(\varepsilon)/K \rceil}(K)$
Initialize: $\tau_K^{\mathrm{E}}(i) = \lceil N(\varepsilon)/K \rceil; \forall\ i \in [K]$, empirical arm CDFs $\mathbb{F}^{\mathrm{E}}_{1,\lceil N(\varepsilon)/K \rceil}, \cdots, \mathbb{F}^{\mathrm{E}}_{K,\lceil N(\varepsilon)/K \rceil}$
for $t = K\lceil N(\varepsilon)/K \rceil + 1, \cdots, T$ **do**
  Select an arm A_t via (3.2) and obtain reward X_t
  Update the empirical CDF $\mathbb{F}^{\mathrm{E}}_{A_t,t}$ according to (8)
end for

Despite its computational simplicity and its better regret guarantee for some DRs, the RS-ETC-M algorithm has the crucial bottleneck of relying on the instance-specific gap information (through $N(\varepsilon)$), which may not always be available – an issue that is addressed by the UCB-type algorithms.

3.3 Risk-sensitive UCB for Mixtures

In this section, we present the **R**isk-**S**ensitive **UCB** for **M**ixtures (RS-UCB-M) algorithm, which does not require the information on the sub-optimality gap $N(\varepsilon)$. The salient features of this algorithm are (i) a distribution estimation routine for forming high-confidence estimates for arms' CDFs and subsequently mixture coefficients; and (ii) a sampling rule based on an *under-sampling* criteria to ensure that arm selections track the mixture coefficients. The pseudocode of RS-UCB-M is presented in Algorithm 2.

Discretization. Similar to RS-ETC-M, we uniformly discretize each coordinate of Δ^{K-1} into intervals of length ε with the distinction that for RS-UCB-M we use the the intervals mid-points, i.e.,

$$\Delta_{\varepsilon}^{K-1} \triangleq \{\varepsilon \mathbf{n} : \mathbf{n} \in \Big\{\mathbb{N} \cup \{0\}\Big\}^K\ , \mathbf{1}^\top \cdot \varepsilon\big(\mathbf{n} + \frac{1}{2}\big) = 1\Big\}\ .$$

Based on the concentration of CDF estimates (9), at time t and given the empirical CDFs $\{\mathbb{F}^{\mathrm{U}}_{i,t} : i \in [K]\}$, for arm $i \in [K]$ we define the *distribution confidence space* as the collection of all the distributions that are within a bounded $1-$Wasserstein distance of $\mathbb{F}^{\mathrm{U}}_{i,t}$. Specifically, for each $i \in [K]$, we define

$$\mathcal{C}_t(i) \triangleq \left\{\eta \in \Omega : \left\|\mathbb{F}^{\mathrm{U}}_{i,t} - \eta\right\|_{\mathrm{W}} \leq 16\ \frac{\sqrt{2\mathrm{e}\log T + 32}}{\sqrt{\tau_t^{\mathrm{U}}(i)}}\right\}. \tag{15}$$

Next, we also need to estimate the optimal mixing coefficients. For this purpose, we apply the UCB principle. This, in turn, requires that we compute the following *optimistic* estimates for the mixing coefficients:

$$\mathbf{a}_t^{\mathrm{U}} \in \arg\max_{\mathbf{a} \in \Delta_{\varepsilon}^{K-1}}\ \max_{\eta_i \in \mathcal{C}_t(i), \forall i \in [K]}\ U_h\Big(\sum_{i \in [K]} a(i)\ \eta_i\Big)\ . \tag{16}$$

Track mixtures. Once we have estimates of the optimal mixing coefficients, i.e., $\mathbf{a}_t^{\mathrm{U}}$, we design an arm selection rule that translates the mixing coefficients and CDF estimates to arm selection choices. Designing such a rule requires addressing a few technical challenges. First, the optimal mixing coefficients might not be unique. Let us denote them by $\{\psi_\ell : \ell \in [L]\}$. It is critical to ensure that we *consistently* track only one of these optimal choices over time. The reason is that if we track multiple mixtures, in aggregate, we will be tracking a mixture of $\{\psi_\ell : \ell \in [L]\}$, which is not necessarily optimal. Secondly, in the initial sampling rounds, the estimates $\boldsymbol{\alpha}_t^{\mathrm{U}}$ are relatively inaccurate, and tracking them leads to highly sub-optimal decisions. Finally, we need a rule for translating the estimated mixtures to arm selections. Next, we discuss how we address these issues.

Tracking a single mixture. To ensure tracking only one optimal mixture, at time t, the RS-UCB-M algorithm checks if the coefficient from the previous step, i.e., $\mathbf{a}_{t-1}^{\mathrm{U}}$ also maximizes (16). If it does, then $\mathbf{a}_{t-1}^{\mathrm{U}}$ is chosen as a candidate optimistic estimate $\mathbf{a}_t^{\mathrm{U}}$. Otherwise, any random candidate solving (16) is chosen.

Initial rounds: To circumvent estimation inaccuracies in the initial rounds, we introduce a *short* forced exploration phase for each arm. Specifically, for $\rho \in (0,1)$, for the first $K\lceil \rho T\epsilon/4 \rceil$ rounds, we explore the arms in a round-robin fashion to initiate the algorithm with sufficiently accurate estimates of the arm CDFs.

Decision rules: We specify a rule that converts mixtures to arm selections. When all the arms are sufficiently explored, motivated by the effective approaches to best arm identification (Garivier and Kaufmann, 2016; Jourdan et al., 2022; Agrawal et al., 2020; Mukherjee and Tajer, 2024) we sample the most *under-sampled* arm. An arm is considered under-sampled if it has been sampled less frequently than the rate indicated by the estimate $\mathbf{a}_t^{\mathrm{U}}$, and has the largest gap between its current fraction and its estimated fraction $a_t^{\mathrm{U}}(i)$. At time $t \geq K\lceil \rho T\varepsilon/4 \rceil$, the arm selection rule is specified by

$$A_{t+1} \triangleq \arg\max_{i \in [K]} \{t a_t^{\mathrm{U}}(i) - \tau_t^{\mathrm{U}}(i)\}\ . \tag{17}$$

Algorithm 2 RS-UCB-M

Input: Exploration rate ρ, horizon T
Sample each arm $N(\rho,\varepsilon) \triangleq \lceil \rho T\varepsilon/4 \rceil$ times and obtain observation sequences $\mathcal{X}_{KN(\rho,\varepsilon)}(1), \cdots, \mathcal{X}_{KN(\rho,\varepsilon)}(K)$
Initialize: $\tau^{\mathrm{U}}_{KN(\rho,\varepsilon)}(i) = N(\rho,\varepsilon)\ \ \forall i \in [K]$, emprical arm CDFs $\mathbb{F}^{\mathrm{U}}_{1,KN(\rho,\varepsilon)}, \cdots, \mathbb{F}^{\mathrm{U}}_{K,KN(\rho,\varepsilon)}$, confidence sets $\mathcal{C}_{KN(\rho,\varepsilon)}(1), \cdots \mathcal{C}_{KN(\rho,\varepsilon)}(K)$ according to (15)
for $t = KN(\rho,\varepsilon) + 1, \cdots, T$ **do**
  Select an arm A_t via (17) and obtain reward X_t
  Update the empirical CDF $\mathbb{F}^{\mathrm{U}}_{A_t,t}$ according to (8)
  Update the confidence set $\mathcal{C}_t(A_t)$ according to (15)
  Compute the optimistic estimate $\mathbf{a}_t^{\mathrm{U}}$ according to (16)
end for

3.4 Computationally Efficient CE-UCB-M

In the RS-UCB-M algorithm, determining the mixing coefficients $\mathbf{a}_t^{\mathrm{U}}$ via (16) involves extremization over a class of distribution functions, and it is computationally expensive. To circumvent this, we present the **C**omputationally-**E**fficient risk-sensitive **UCB** for **M**ixtures (CE-UCB-M) algorithm as a computationally tractable modification of RS-UCB-M. In CE-UCB-M, instead of solving (16), for any given set of CDF estimates $\{\mathbb{F}_{i,t}^{\mathrm{U}} : i \in [K]\}$ and mixing vector $\mathbf{a}$, we define

$$\mathrm{UCB}_t(\mathbf{a}) \triangleq U_h\Big(\sum_{i\in[K]} a(i)\,\mathbb{F}_{i,t}^{\mathrm{C}}\Big) + \mathcal{L}\sum_{i\in[K]}\left(a(i)\cdot 16\,\frac{\sqrt{2\mathrm{e}\log T}+32}{\sqrt{\tau_t^{\mathrm{U}}(i)}}\right)^q,$$

where q and $\mathcal{L}$ are the Hölder parameters of the underlying distortion function h. We specify estimates of the mixing coefficients as

$$\mathbf{a}_t^{\mathrm{C}} \in \operatorname*{arg\,max}_{\mathbf{a}\in\Delta_\varepsilon^{K-1}} \mathrm{UCB}_t(\mathbf{a})\,. \tag{18}$$

The remainder of the algorithm (the tracking block) follows the same steps as in RS-CS-UCB-M. The CE-UCB-M procedure is summarized in Algorithm 3 in Appendix G.1.

Discussion. Unlike RS-ETC-M, RS-UCB-M and the CE-UCB-M are independent of instance-dependent parameters. The explicit exploration phase involves a hyperparameter ρ, which must be bounded away from 0. The performance of RS-ETC-M and RS-UCB-M depends on the choice of DR, with neither uniformly dominating the other. In Section 4 we show that RS-ETC-M has better regret guarantees for DRs favoring a solitary arm policy. In contrast, for the DRs that have a mixture optimal policy, the performance advantage depends on the DR. For instance, for Wang's Right-tail deviation, RS-ETC-M is better, whereas for Gini deviation RS-UCB-M outperforms RS-ETC-M.

4 REGRET ANALYSIS

In this section, we characterize regret guarantees for the three algorithms presented in the previous section. We provide a decomposition for the regret to two terms, where one term accounts for the discretization inaccuracies and one accounts for the scaling behavior in terms of T, which we refer to as the *discrete regret*. Based on this decomposition, we present a regret bound in terms of ε for any desired discretization level. Subsequently, we also provide an ε-independent regret in which ε is chosen carefully to achieve the best performance subject to algorithmic constraints.

An important observation is the following contrast between the regrets characterized for the mixture algorithms and their canonical ETC and UCB counterparts: the canonical ETC and UCB algorithms generally exhibit the same regret, even though ETC requires access to instance-dependent parameters. In the mixtures setting, however, access to instance-dependent parameters yields better regret guarantees for the ETC-type algorithm (i.e., RS-ETC-M).

For a given discretization level ε and a bandit instance $\boldsymbol{\nu} \triangleq (\mathbb{F}_1, \cdots, \mathbb{F}_K)$, for policy π, we decompose the regret defined in (5) into a discretization error component and a discrete regret component as follows:

$$\mathfrak{R}_{\boldsymbol{\nu}}^{\pi}(T) = \Delta(\varepsilon) + \bar{\mathfrak{R}}_{\boldsymbol{\nu}}^{\pi}(T)\,, \tag{19}$$

where based on the definition of $\mathbf{a}^\star$ in (10) we have

$$\Delta(\varepsilon) \triangleq V(\boldsymbol{\alpha}^\star, \mathbb{F}) - V(\mathbf{a}^\star, \mathbb{F})\,, \tag{20}$$

$$\bar{\mathfrak{R}}_{\boldsymbol{\nu}}^{\pi}(T) \triangleq V(\mathbf{a}^\star, \mathbb{F}) - \mathbb{E}_{\boldsymbol{\nu}}^{\pi}\Big[V\Big(\frac{1}{T}\boldsymbol{\tau}_T^{\pi}, \mathbb{F}\Big)\Big]\,. \tag{21}$$

In the decomposition in (19), $\Delta(\varepsilon)$ accounts for the discretization error and $\bar{\mathfrak{R}}_{\boldsymbol{\nu}}^{\pi}(T)$ represents the *discrete* regret. We begin by presenting the regret guarantees for the RS-ETC-M algorithm.

Theorem 1 (RS-ETC-M – ε-dependent) *For any $\varepsilon \in \mathbb{R}_+$ and distortion function h with Hölder exponent q, for all $T > N(\varepsilon)$, RS-ETC-M's regret is upper bounded as*

$$\mathfrak{R}_{\boldsymbol{\nu}}^{\mathrm{E}}(T) \le (\mathcal{L}K + W^{-q})\left(3WM(\varepsilon)\,\frac{\log T}{T}\right)^q + \Delta(\varepsilon)\,.$$

Choosing ε. Theorem 1 is valid for any $\varepsilon \in \mathbb{R}_+$, allowing freedom to appropriately choose the desired accuracy for the mixing coefficients. Observe that $\Delta(\varepsilon)$ in proportional to ε, and $N(\varepsilon)$ is inversely proportional to ε. Hence, while arbitrarily diminishing ε decreases the discretization error, it may violate the condition that $T > N(\varepsilon)$. We chose ε to be small enough (small discretization error), while conforming to the condition $T > N(\varepsilon)$ in Theorem 1 (feasibility). The minimum feasible ε depends on q, r, and its connection to the sub-optimality gap as follows. Let us define

$$\beta \triangleq \lim_{\varepsilon\to 0} \frac{\log \Delta_{\min}(\varepsilon)}{\log \varepsilon}\,, \tag{22}$$

which quantifies how fast the discretization error $\Delta_{\min}(\varepsilon)$ diminishes as ε tends to 0. Appendix A.1 provides characterizes β values for some DRs. Let $\varepsilon = \Theta((K^{2+2/q}T^{-\gamma}\log T)^{q/2\beta})$ where we set $\gamma \triangleq 2\beta/(2\beta + r)$, which leads to the following regret bound.

Theorem 2 (RS-ETC-M – ε-independent) *Under the conditions of Theorem 1, when β exists, the minimum feasible regret RS-ETC-M satisfies is*

$$\mathfrak{R}_{\boldsymbol{\nu}}^{\mathrm{E}}(T) \le O\left(\left[K^{c_{\mathrm{E}}}\cdot\frac{\log T}{T^\gamma}\right]^{\frac{qr}{2\beta}}\right), \tag{23}$$

where $c_{\mathrm{E}} \triangleq 2(1+\frac{\beta+1}{q})$ and $\gamma = \frac{2\beta}{2\beta+r}$.

Next, we present the regret guarantees for the RS-UCB-M and CE-UCB-M algorithms. An important observation is that these algorithms yield weaker *discrete* regret guarantees compared to RS-ETC-M. The regret degradations are the expense of not knowing the instance-dependent gaps. For stating the theorem, we define an instance-dependent finite time instant $T(\varepsilon)$.

$$T_0(\varepsilon) \triangleq \inf\{t \in \mathbb{N} : \forall s \geq t \ , \ s \in \mathcal{Q}\} \ , \tag{24}$$

where we have defined

$$\mathcal{Q} \triangleq \left\{ s : \frac{\sqrt{2\mathrm{e}\log s}+32}{\sqrt{\frac{p}{4}s\varepsilon}} \leq \frac{1}{16}\left(\frac{\Delta_{\min}(\varepsilon)}{2K\mathcal{L}}\right)^{\frac{1}{q}} \right\} \ . \tag{25}$$

Accordingly, we define

$$T(\varepsilon) \triangleq \tfrac{2}{\varepsilon}\Big(T_0(\varepsilon)-1\Big) \ . \tag{26}$$

The next theorem presents a ε-dependent regret for the RS-UCB-M and CE-UCB-M algorithms.

Theorem 3 (RS/CE-UCB-M – ε-dependent) *For any $\varepsilon \in \mathbb{R}_+$, and distortion function h with Hölder exponent q, for all $T > \max\{\mathrm{e}^K, T(\varepsilon)\}$, for $\pi \in \{\mathrm{U}, \mathrm{C}\}$ we have*

$$\mathfrak{R}^{\pi}_{\boldsymbol{\nu}}(T) \leq [B + \mathcal{L}(W^q+1)]\left[\frac{64}{\sqrt{\varepsilon\rho T}}\Big(\sqrt{2\mathrm{e}\log T}+32\Big)\right]^q + \Delta(\varepsilon) \ . \tag{27}$$

Proof sketch. The proof differs significantly from the standard UCB-type analyses. For vanilla UCB, the general proof uses the fact that selecting any suboptimal arm more than $O(\log T)$ times is unlikely, and hence, the overall regret is bounded by $O(\frac{1}{T}K\log T)$. However, in our analyses, we have the new dimension of estimating the mixing coefficients and need these estimates to converge to an optimal choice. Such convergence requires that all arms to be sampled at a rate linear in T (unless an arm's mixing coefficient is 0). Hence, selecting any arm $O(\log T)$ times is insufficient.

The other key difference pertains to finding a bound on the mixing coefficient errors (estimation and convergence). Characterizing this bound hinges on two key steps: (i) the convergence of the UCB estimates ($\mathbf{a}^{\mathrm{U}}_t$ for RS-UCB-M and $\mathbf{a}^{\mathrm{C}}_t$ for CE-UCB-M) to the discrete optimal solution $\mathbf{a}^\star$ in probability, and (ii) a sublinear regret incurred in the process of tracking the mixing coefficient estimates using under-sampling. The first step is analyzed in Appendix F.3, where we show that the probability of error for the RS-UCB-M and CE-UCB-M algorithms in identifying the discrete optimal mixture is upper bounded by $T((\frac{1}{T^2}+1)^K-1)$. The second step is analyzed in Lemma 15 in Appendix F.3, in which we show that the regret incurred by the tracking block of the RS-UCB-M and CE-UCB-M algorithms is of the order $O(K/T)$. The regret upper bound is a combination of the regrets in these results.

Choosing ε. Similarly to RS-ETC-M, we choose an ε that ensures $T > T(\varepsilon)$ for the bound in Theorem 3. The choice

$$\varepsilon = \Theta\left(\left(K^{\frac{2}{q}}\frac{\log T}{T}\right)^{\kappa}\right) \ , \quad \text{where } \kappa \triangleq \left(\frac{2\beta}{q}+2\right)^{-1} \ .$$

Theorem 4 (RS/CE-UCB-M – ε-independent) *Under the conditions of Theorem 3, when β exists and $r \leq \beta + \frac{q}{2}$ the regret upper bound for $\pi \in \{\mathrm{U}, \mathrm{C}\}$ is*

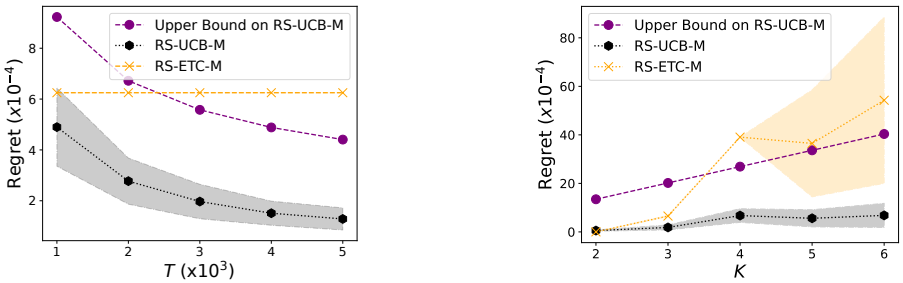
$$\mathfrak{R}^{\pi}_{\boldsymbol{\nu}}(T) \leq O\left(K^{c_{\mathrm{U}}}\left[\frac{\log T}{T}\right]^{r\kappa}\right) \ , \tag{28}$$

where $c_{\mathrm{U}} \triangleq \max\{r(1+\frac{2\kappa}{q}), 1-\kappa\}$.

In the case of Gini deviation, we observe that for discrete regret, the RS-ETC-M algorithm achieves $O(K/T)$ regret while RS-UCB-M achieves $O(K/\sqrt{T})$ regret, ignoring polylogarithmic factors. However, when we optimize the algorithms for the discretization level ε, we observe an order-wise improvement in the regret of the RS-UCB-M and CE-UCB-M algorithms.

Regret bounds for important cases. We specialize our general results to a number of widely-used DRs in Table 1 and Table 2 (Appendix A.2). Table 1 provides the regret bounds and Table 2 specifies the Hölder continuity constants q, r, and β. We remark that the ε-dependent regret bounds depend only on the Hölder continuity constant q (Theorems 1 and 3). On the other hand, the ε-independent regret bounds, will additionally depend on the mixture Hölder continuity constant r and β (Theorems 2 and 4). It is worth noting that, for RS-ETC-M algorithm, the constant β will be relevant only for the DR choices with mixture policies. The reason is that when we have a solitary arm policy, we can discretize the simplex into unit vectors along each coordinate. In this case $\Delta_{\min}(\varepsilon) = O(1)$, which does not scale with ε. Consequently, for RS-ETC-M and for solitary arms, the order of the regret is the same as the discrete regret.

Relevance to the existing literature. In Section A.2 we provide a thorough discussion on the relevance of our general results to those of the existing risk-sensitive literature. In summary, the existing literature has studied only the case of monotone DRs, and solitary arm policies including Liang and Luo (2023) and L.A. and Bhat (2022), which proposes Wasserstein distance based UCB algorithms. We present a detailed comparison of regret bounds, which depend on whether the underlying distributions are bounded or sub-Gaussian, in Table 3 (Appendix A.2). For the well-investigated CVaR, the existing literature establishes a regret bound of $O(\log T/T)$ for bounded support (Baudry et al., 2021) and $O(\sqrt{\log T/T})$ for sub-Gaussian distributions (L.A. and Bhat, 2022). Under the same sub-Gaussian assumption, RS-ETC-M improves the regret bound to $O(\log T/T)$.



(a) Regret versus time horizon T.

(b) Regret versus number of arms K.

Figure 1: Regret of the algorithms for different parameters

5 EMPIRICAL EVALUATIONS

We provide empirical evaluations of the RS-ETC-M and RS-UCB-M algorithms and compare them with the following baseline algorithm: a uniform sampling strategy that selects each arm T/K times, where T denotes the horizon. We also note that RS-ETC-M assumes knowing instance-specific gap information, while RS-UCB-M does not. The experiments are conducted for the DR Gini deviation, whose distortion function is $h(p) = p(1-p)$. We focus on Bernoulli bandits with mean vectors $\mathbf{p} \in [0,1]^K$, in which case, the DR has a closed-form expression for the case of mixtures, given by $V(\boldsymbol{\alpha}, \mathbb{F}) = \langle \boldsymbol{\alpha}, \mathbf{p} \rangle (1 - \langle \boldsymbol{\alpha}, \mathbf{p} \rangle)$. All the experiments are averaged over 10^3 trials. [3]

We empirically investigate regret's variations with varying levels of the horizon T and number of arms K as well as other properties of the algorithms, including their sensitivity to the exploration rate ρ and instance-specific gap parameters. Regret evaluations are presented in this section, and analyzing the properties is presented in Appendix H.

Regret versus horizon. Our first experiment considers a 2-armed bandit instance $\boldsymbol{\nu} \triangleq \big(\text{Bern}(0.4), \text{Bern}(0.9)\big)$. For this instance, note that the optimal mixing coefficient is $\boldsymbol{\alpha}^\star = [0.8, 0.2]^\top$. For implementing the RS-UCB-M algorithm, we have set the discretization level to $\varepsilon = \sqrt{(K \log T)/T}$, and explore each arm $\frac{T}{10}$ times. The results are presented in Figure 1a and demonstrate the regret variations of RS-UCB-M and RS-ETC-M versus T.

For RS-UCB-M, we present the range of the results in the shaded region with their average specified by the dashed curve. We do the same for RS-UCB-M, but its range is notably narrow and invisible on the curve. We observe that RS-UCB-M uniformly outperforms RS-ETC-M. This is expected as the exploration phase of the RS-ETC-M algorithm explores an arm more frequently than the associated estimated mixing coefficient recommends. Additionally, Figure 2a (Appendix H) demonstrates that RS-UCB-M outperforms uniform sampling.

Regret versus number of arms. Next, we investigate the variations of RS-UCB-M's regret in the number of arms K. We choose $K \in \{2,3,4,5,6\}$. For fair comparisons, we create the instances so that they have uniform gaps between the arm means for every instance. For both RS-UCB-M and RS-ETC-M, we set the discretization level to $\varepsilon = \sqrt{(K \log T)/T}$; set the horizon to $T = 3 \times 10^5$; and explore every arm $\frac{T}{20}$ times for RS-UCB-M algorithm. Figure 1b shows that, on average, the RS-UCB-M algorithm outperforms the RS-ETC-M, especially as the number of arms increases.

6 CONCLUDING REMARKS

We have provided a novel approach to viewing and analyzing stochastic bandits in which the decision-makers are expected to be conscious of the decision risks involved. This contrasts with the conventional risk-neutral approaches, which aim to optimize the average reward in a fully utilitarian way. Adopting a rich class of distortion riskmetrics, we observed that many of deviation measures and measures of variability are optimized by a mixture distribution over the arms. This is in sharp contrast to the commonly adopted premise that there exists a solitary best arm – a premise shared by the existing literature on risk-sensitive bandit algorithms, too. Designing regret-efficient algorithms for such mixtures poses various technical and design challenges, mainly pertinent to identifying and tracking the optimal mixing coefficients of the arms. Based on the UCB and ETC principles, we have designed two sets of bandit algorithms and established regret results for a broad spectrum of commonly used distortion riskmetrics. A potential future direction is finding a general lower-bound for distortion riskmetrics in this setting, which is in general uninvestigated for risk-sensitive bandits.

[3] https://github.com/MeltemTatli/Risk-sensitive-Bandits-Arm-Mixture-Optimality.git

Acknowledgements

A portion of this work was done when Prashanth L. A. was at IIT Bombay. The work of Meltem Tatlı, Arpan Mukherjee, and Ali Tajer was supported in part by the U.S. National Science Foundation under Grants ECCS-193310 and DMS-2319996, and in part by the Rensselaer-IBM Future of Computing Research Collaboration (FCRC).

References

Qiuqi Wang, Ruodu Wang, and Yunran Wei. Distortion riskmetrics on general spaces. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 50(3):827–851, May 2020.

Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, New Jersey, 2nd edition, 2009.

Menahem E Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.

Dieter Denneberg. *Non-additive Measure and Integral*. Springer Science & Business Media, 1994.

Shigeo Kusuoka. On law invariant coherent risk measures. *Advances in Mathematical Economics*, 3:83–95, 2001.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

R. Tyrrell Rockafellar, Stanislav Uryasev, and Michael Zabarankin. Generalized deviation in risk analysis. *Finance and Stochastics*, 10:51–74, 2006.

Baishuai Zuo and Chuancun Yin. Worst-cases of distortion riskmetrics and weighted entropy with partial information. *arXiv:2405.19075v1*, 2024.

Eliezer Furman, Ruodu Wang, and Ricardas Zitikis. Gini-type measures of risk and variability: Gini shortfall, capital allocation and heavy-tailed risks. *Journal of Banking and Finance*, 83:70–84, 2017.

John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343, 1982.

Aditya Gopalan, Prashanth L. A., Michael Fu, and Steve Marcus. Weighted bandits or: How bandits learn distorted values that are not expected. In *Proc. AAAI Conference on Artificial Intelligence*, San Francisco, CA, February 2017.

Prashanth L.A. and Sanjay P. Bhat. A Wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23(238):1–61, 2022.

Joel Q. L. Chang and Vincent Y. F. Tan. A unifying theory of thompson sampling for continuous risk-averse bandits. In *Proc. AAAI Conference on Artificial Intelligence*, virtual, February 2022.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proc. Conference on Learning Theory*, Stockholm, Sweden, July 2018.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware CVaR bandits. In *Proc. International Conference on Machine Learning*, virtual, July 2021.

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for cvar bandits. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2019.

Chenmien Tan and Paul Weng. CVaR-regret bounds for multi-armed bandits. In *Proc. Asian Conference on Machine Learning*, India, December 2022.

Hao Liang and Zhi-Quan Luo. A distribution optimization framework for confidence bounds of risk measures. In *Proc. Machine Learning Research*, Hawaii, July 2023.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019.

Prashanth L. A., Krishna Jagannathan, and Ravi Kumar Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *Proc. International Conference on Machine Learning*, virtual, July 2020.

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. In *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *arXiv:1301.1936*, 2013.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Yudong Luo, Guiliang Liu, Pascal Poupart, and Yangchen Pan. An alternative to variance: Gini deviation for risk-averse policy gradient. *arXiv:2307.08873*, 2023.

Xia Han, Ruodu Wang, and Xun Yu Zhou. Choquet regularization for reinforcement learning. *arXiv:2208.08497*, 2022.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proc. Conference on Learning Theory*, New York, NY, June 2016.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

Shubhada Agrawal, Sandeep Juneja, and Peter Glynn. Optimal δ-correct best-arm selection for heavy-tailed distributions. In *Proc. International Conference on Algorithmic Learning Theory*, San Diego, CA, February 2020.

Arpan Mukherjee and Ali Tajer. Best arm identification in stochastic bandits: Beyond $\beta-$ optimality. *IEEE Transactions on Information Theory*, 2024.

Nithia Vijayan and Prashanth L.A. Policy gradient methods for distortion risk measures. *arXiv 2107.04422*, 2021.

Kevin Dowd and David Blake. After VaR: The theory, estimation and insurance applications of quantile-based risk measures. *The Journal of Risk and Insurance*, 73(2), May 2006.

Bruce L. Jones and Ricardas Zitikis. Empirical estimation of risk measures and related quantities. *North American Actuarial Journal*, 7(4):44–54, October 2003.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Checklist

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. Please see 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes. Please see 3 on more discussion on algorithms.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes. All of our theorems were included in section 4.
   (b) Complete proofs of all theoretical results. Yes. Proof overview is given in 4. Due to the space constraints, the proofs are postponed to the Appendix B, C, D, E, G, and F.
   (c) Clear explanations of any assumptions. Yes. We have included our assumptions in section 2.
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). Yes. We provided the code as an URL.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable.
   (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. We have provided the details in Section 5 and Appendix H.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes. We have provided the details in Appendix H.
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. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.
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
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Risk-sensitive Bandits: Arm Mixture Optimality and Regret-efficient Algorithms

Supplementary Materials

Table of Contents

Organization of the Supplementary Material

The supplementary material consists of eight parts grouped in Appendices A-H. We begin by presenting the Hölder parameters used in the theorems for DRs listed in Table 1, along with a comparison to the risk-sensitive bandit literature (Appendix A). The proof of Lemma 1 is presented in Appendix B and the proof of the finiteness of the Wasserstein-related constant W is presented in Appendix C. Subsequently, in Appendix D, we provide several auxiliary lemmas that are used for the proofs of regret results presented in Theorems 1-4 and the subsequent appendices. Next, we have Appendices E-G to present the CE-UCB-M pseudo-algorithm (Appendix G), as well as the performance guarantees for the RS-ETC-M (Appendix E), RS-UCB-M (Appendix F), and CE-UCB-M (Appendix G) algorithms. Finally, we provide additional empirical results in Appendix H.

A Specializing the Results to Distortion Riskmetrics

In this section, we present the parameters used in Theorems 1-4 for some of the widely-used DRs and compare our results with known regret guarantees in the literature.

A.1 Distortion Riskmetric Examples and Parameters

Hölder exponents: In this section, we present the Hölder exponents q, r for the DRs discussed in Table 1. These constants are reported in Table 2. For the Hölder exponents, we note that the expressions of q and r depend on the bandit model. Furthermore, based on the definitions of Hölder continuity and mixture Hölder continuity provided in (6) and (7), respectively, any given Hölder continuity exponent r is also a mixture Hölder continuity exponent, that is given q, $r = q$ is always a valid choice for r. We characterize the Hölder exponents and constants for general for Bernoulli bandits in Lemmas 5–11 in Appendix D. Table 2 summarizes these results.

Gap constant β: We also analyze the gap constant β, which we have defined as

$$\beta = \lim_{\varepsilon \to 0} \frac{\log \Delta_{\min}(\varepsilon)}{\log \varepsilon} \ , \tag{29}$$

which plays an important role in characterizing ε-independent regret guarantees. We characterize β for DRs with strictly increasing distortion functions (risk-neutral mean value, dual power, quadratic, and CVaR under a condition) in Lemma 13. Table 2 lists these β values.

Table 2: Hölder continuity parameters and distortion functions of the DRs reported in Table 1.

Distortion Riskmetrics	q [a]	r [b]	β [c] [d]
Risk-Neutral Mean Value	1	1	1
Dual Power (Vijayan and L.A., 2021)	1	1	1
Quadratic (Vijayan and L.A., 2021)	1	1	1
CVaR (Dowd and Blake, 2006)	1 [e]	1	1 [f]
PHT Measure ($s \in (0,1)$) (Jones and Zitikis, 2003)	s	s	1
Mean-median deviation (Jones and Zitikis, 2003) (Wang et al., 2020)	1	1	–
Inter-ES Range $\text{IER}_{\alpha=1/2}$ (Wang et al., 2020)	1	1	–
Wang's Right-Tail Deviation (Jones and Zitikis, 2003)	1/2	1	–
Gini Deviation	1	2	–

[a] We report q for K-arm Bernoulli bandit distributions except for CVaR and PHT Measure.
[b] We report r for K-arm Bernoulli bandit distributions.
[c] We report β for K-arm Bernoulli bandit distributions.
[d] The non-shaded DRs have strictly increasing DRs and their β values are characterized analytically.
[e] L.A. and Bhat (2022) reports $q = 1$ for sub-gaussian distributions.
[f] When the arms' mean values are smaller than $1 - \alpha$. This assumption is only needed for the analysis of β. Hence, it is needed for only the regret of RS-UCB-M.

A.2 Comparison with Risk-Sensitive Bandit Literature

In this subsection, we discuss two broad categories of the existing studies. First, we discuss the existing studies focused on risk-sensitive bandits and establish their connection to our framework and regret results. Secondly, we discuss the empirical distribution performance measures (EDPM) framework (Cassel et al., 2018).

Frameworks subsumed by the DR-centric framework. As discussed earlier, the DR-centric framework subsumes several existing frameworks for risk-sensitive bandits. we note that the existing literature has considered only monotone DRs and they include CVaR and distortion risk measures (DRM), which we discuss next.

CVaR: For CVaR, the best-known regret is $O(\log(T)/T)$ under the assumption of bounded support (Baudry et al., 2021; Cassel et al., 2018; Tamkin et al., 2019). When the assumption is loosened to sub-Gaussian, the regret weakens to $O(\sqrt{\log(T)/T})$ for (L.A. and Bhat, 2022). Compared to these known regret bounds, for sub-Gaussian models, RS-ETC-M achieves a regret bound of $O(\log(T)/T)$ while for Bernoulli model RS-UCB-M achieves a regret bound of $O\left((\log(T)/T)\right)^{1/4}$. The reason that RS-ETC-M outperforms RS-UCB-M is that it assumes knowing the minimum suboptimality gap $\Delta_{\min}$.

DRMs: For the more general class of DRMs, under the assumption of bounded support, the study in (Chang and Tan, 2022; Gopalan et al., 2017) reports a regret of $O(\log(T)/T)$. For the sub-Gaussian support, (L.A. and Bhat, 2022) reports a regret bound of $O(\sqrt{\log(T)/T})$. In the sub-Gaussian setting, the regret bound of RS-ETC-M depends on Hölder exponent q, and hence, for the DRMs with $q = 1$, we achieve the same regret as in CVaR. For $q < 1$, which is the case for PHT with $s = 1/2$, we report a regret of $O(\sqrt{\log(T)/T})$ for RS-ETC-M and under Bernoulli model we report $O((\log(T)/T)^{1/5})$ for RS-UCB-M. All these results are summarized in Table 3.

EDPM Framework. The EDPM framework introduced in (Cassel et al., 2018) is, in principle, more general than the DR-centric framework, and it includes risk measures that cannot be modeled by DRs, e.g., variance and Sharpe Ratio. Nevertheless, the analysis of the EDPM is focused entirely on settings in which the optimal policy is necessarily a solitary arm policy. Such focus precludes risk measures with optimal mixture policies, e.g., Gini deviation and Wang's Right-tail deviation.

Table 3: Comparison of the existing risk-sensitive bandit studies where.

Work	**DR**	**Distribution Assumptions**	**Regret** ($\varpi(T) \triangleq \sqrt{\frac{\log T}{T}}$)
Baudry et al. (2021)	CVaR	Bounded	$O(\varpi^2(T))$
Tamkin et al. (2019)	CVaR	Bounded	$O(\varpi^2(T))$ [a]
Cassel et al. (2018)	CVaR	Bounded	$O(\varpi^2(T))$ [b]
Chang and Tan (2022)	DRM	Bounded	$O(\varpi^2(T))$
Gopalan et al. (2017)	DRM	Bounded	$O(\varpi^2(T))$
Liang and Luo (2023)	DRM	Bounded	$O(\varpi^2(T))$
L.A. and Bhat (2022)	DRM	Sub-Gaussian	$O(\varpi(T))$
RS-ETC-M (This work)	DRM	Sub-Gaussian	$O(\varpi^{2q}(T))$ [c]
RS-UCB-M (This work)	DRM	Sub-Gaussian	$O(\varpi^{2r\kappa}(T))$

[a] This is a problem-dependent bound.
[b] This bound holds under some assumptions on the arm distribution densities.
[c] Assuming that the optimal solution is a solitary arm.

B Proof of Lemma 1 (Gini Deviation)

In this section, we prove Lemma 1, which states that the utility-maximizing solution for Gini deviation is a mixture of arm CDFs. For any values $p_1, p_2 \in (0,1)$, consider two Bernoulli distributions $\mathsf{Bern}(p_1)$ and $\mathsf{Bern}(p_2)$ with CDFs $\mathbb{F}_1$ and $\mathbb{F}_2$, respectively. It can be readily verified that for the distribution $\mathbb{F} = \mathsf{Bern}(p)$ for any $p \in [0,1]$, we have

$$U_h(\mathbb{F}) = \int_0^\infty h(1-\mathbb{F}(x))\,\mathrm{d}x \tag{30}$$

$$= h(p)\ , \tag{31}$$

which implies that $U_h(\mathbb{F}_1) = h(p_1)$ and $U_h(\mathbb{F}_2) = h(p_2)$. Owing to concavity of the distortion function h, the maximizer

$$p^\star \triangleq \arg\max_{p\in[0,1]} h(p) \tag{32}$$

is unique. Furthermore, due to the function being non-monotone, $p^\star$ cannot lie at the boundaries, i.e., at 0 or at 1. Hence, $p^\star \in (0,1)$. Let us choose the mean values of the arms such that $p_1 < p^\star$ and $p_2 > p^\star$. With these choices, there exists $\lambda \in (0,1)$ such that

$$p^\star = \lambda p_1 + (1-\lambda)p_2\ . \tag{33}$$

Accordingly, define the mixture distribution

$$\mathbb{F}^\star \triangleq \lambda\mathbb{F}_1 + (1-\lambda)\mathbb{F}_2\ . \tag{34}$$

Subsequently, we have

$$U_h(\mathbb{F}^\star) = h(p^\star) > \max\{h(p_1), h(p_2)\} = \max\{U_h(\mathbb{F}_1), U_h(\mathbb{F}_2)\}\ . \tag{35}$$

This indicates that there exists a mixture of F_1 and F_2 whose DR value dominates those of F_1 and F_2.

C Finiteness of W for sub-Gaussian Random Variables

We characterize an upper bound on the parameter W defined in Section 2 and show that it is finite. Recall the definition of W:

$$W \triangleq \max_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}\in\Delta^K} \frac{1}{\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_1}\Big\|\sum_i \alpha_i\mathbb{F}_i - \sum_j \beta_j\mathbb{F}_j\Big\|_{\mathrm{W}}\ . \tag{36}$$

In order to show the finiteness of W for 1-sub-Gaussian random variables, we leverage the Kantorvich-Rubinstein duality of the 1-Wasserstein measure, which is stated below.

Theorem 5 (Villani (2009)) *Let $\mathcal{L}^1(\Omega)$ denote the space of probability measures supported on Ω with finite first moment. For any $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{L}^1(\Omega)$, we have*

$$\|\mathbb{P}_1 - \mathbb{P}_2\|_{\mathrm{W}} = \sup_{\|f\|_L\le 1}\left\{\int_\Omega f\,\mathrm{d}\mathbb{P}_1 - \int_\Omega f\,\mathrm{d}\mathbb{P}_2\right\}\ , \tag{37}$$

where $\|f\|_L \le 1$ denotes the space of all 1$-$ Lipschitz functions $f:\mathbb{R}\to\mathbb{R}$.

Based on the characterization of the 1-Wasserstein distance in Theorem 5, we next provide an upper bound on W for sub-Gaussian random variables.

Theorem 6 (Upper bound on W) *Let $\{\mathbb{F}_i : i \in [K]\}$ be probability measures on $\Omega \subseteq \mathbb{R}$ that are 1-sub-Gaussian and define W as in (36). We have*

$$W \ \le\ \sqrt{2\pi}\ . \tag{38}$$

Proof: Let $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Delta^{K-1}$ denote two distinct probability mass functions on $[K]$. We have

$$\Big\|\sum_{i\in[K]}\Big(\alpha(i)-\beta(i)\Big)\mathbb{F}_i\Big\|_{\mathrm{W}} \ =\ \sup_{\|f\|_L\le 1}\sum_{i\in[K]}\Big(\alpha(i)-\beta(i)\Big)\mathbb{E}_{\mathbb{F}_i}\Big[f(X)\Big] \tag{39}$$

$$\le\ \sup_{\|f\|_L\le 1}\Big|\sum_{i\in[K]}\Big(\alpha(i)-\beta(i)\Big)\Big(\mathbb{E}_{\mathbb{F}_i}\Big[f(X)-f(0)\Big]+f(0)\Big)\Big| \tag{40}$$

$$= \sup_{\|f\|_L \le 1} \left| \sum_{i\in[K]} \Big(\alpha(i) - \beta(i)\Big)\Big(\mathbb{E}_{\mathbb{F}_i}\Big[f(X) - f(0)\Big]\Big)\right| \tag{41}$$

$$\le \sup_{\|f\|_L \le 1} \sum_{i\in[K]} \left|\Big(\alpha(i) - \beta(i)\Big)\mathbb{E}_{\mathbb{F}_i}\Big[f(X) - f(0)\Big]\right| \tag{42}$$

$$\le \sum_{i\in[K]} \Big|\alpha(i) - \beta(i)\Big| \cdot \sup_{\|f\|_L \le 1} \Big|\mathbb{E}_{\mathbb{F}_i}[f(X) - f(0)]\Big| \tag{43}$$

$$\le \sum_{i\in[K]} \Big|\alpha(i) - \beta(i)\Big| \cdot \mathbb{E}_{\mathbb{F}_i}\Big[|X|\Big] \ , \tag{44}$$

where,

- the equality in (39) follows from Theorem 5;
- the transition (39)-(40) holds since we take the absolute value;
- the transition (40)-(41) follows from the fact that $\sum_{i\in[K]} \alpha(i) = \sum_{i\in[K]} \beta(i) = 1$;
- the transition (41)-(42) follows from triangle inequality;
- the transition (42)-(43) follows from the fact that for any two functions f_1 and f_2, we have $\sup_x\{f_1(x) + f_2(x)\} \le \sup_x f_1(x) + \sup_x f_2(x)$;
- and the transition (43)-(44) follows from 1-Lipschitzness of f.

For sub-Gaussian variables, $\mathbb{E}[|X|]$ is bounded in terms of the sub-Gaussian parameter. Since all distributions are $1-$sub- Gaussian, we have

$$E_{\mathbb{F}_i}[|X|] = \int_0^{+\infty} \mathbb{P}(|x| > u)du \tag{45}$$

$$\le \int_0^{+\infty} 2\exp -t^2/2dt \tag{46}$$

$$= \int_{-\infty}^{+\infty} \exp -t^2/2dt \tag{47}$$

$$= \sqrt{2\pi} \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp -t^2/2dt}_{=1} \tag{48}$$

$$= \sqrt{2\pi} \tag{49}$$

which implies that

$$\Big\| \sum_{i\in[K]} \Big(\alpha(i) - \beta(i)\Big)\mathbb{F}_i \Big\|_{\mathrm{W}} \overset{(44)}{\le} \sqrt{2\pi} \cdot \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 \ . \tag{50}$$

Hence, from (50), we have that

$$W \ = \ \max_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}\in\Delta^K} \frac{1}{\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1} \Big\| \sum_i \alpha_i \mathbb{F}_i - \sum_j \beta_j \mathbb{F}_j \Big\|_{\mathrm{W}} \ \le \ \sqrt{2\pi} \ . \tag{51}$$

■

D Auxiliary Lemmas

D.1 Useful Properties

In this section, we present some auxiliary lemmas that will be used in the proofs of Theorems 1, 2, 3, and 4. For the proofs, we use equivalent characterizations of the $1-$Wasserstein metric, which is provided in the lemma below. For the proof, the reader is referred to Lemma 2 (L.A. and Bhat, 2022).

Lemma 3 *Consider random variables X and Y with CDFs F_X and F_Y, respectively. Then,*

$$\|F_X - F_Y\|_{\mathrm{W}} = \sup_{\|f\|_L \le 1} \Big|\mathbb{E}(f(X) - \mathbb{E}(f(Y))\Big| = \int_{-\infty}^{\infty} |F_X(s) - F_Y(s)|\mathrm{d}s = \int_0^1 |F_X^{-1}(\beta) - F_Y^{-1}(\beta)|\mathrm{d}\beta \ . \tag{52}$$

Lemma 4 (Concave Distortion Functions) *For the DRs of the form stated in (1), if the distortion function $h:[0,1]\mapsto[0,1]$ is concave, then the DR evaluated for a mixture distribution over the arms is concave in the mixing coefficient.*

Proof: For any $\boldsymbol{\alpha}\in\Delta^{K-1}$, we have

$$\begin{aligned} U_h\left(\sum_{i\in[K]}\alpha(i)\mathbb{F}_i\right) &= \int_0^\infty h\left(1-\sum_{i\in[K]}\alpha(i)\mathbb{F}_i(x)\right)\mathrm{d}x && (53)\\ &= \int_0^\infty h\left(\sum_{i\in[K]}\alpha(i)\Big(1-\mathbb{F}_i(x)\Big)\right)\mathrm{d}x && (54)\\ &\geq \int_0^\infty \sum_{i\in[K]}\alpha(i)h\Big(1-\mathbb{F}_i(x)\Big)\,\mathrm{d}x && (55)\\ &= \sum_{i\in[K]}\alpha(i)\int_0^\infty h\Big(1-\mathbb{F}_i(x)\Big)\,\mathrm{d}x && (56)\\ &= \sum_{i\in[K]}\alpha(i)U_h(\mathbb{F}_i)\ , && (57)\end{aligned}$$

where,

- the inequality in (55) follows from the concavity of the distortion function w;
- and the equality in (56) follows from the Fubini-Tonelli's theorem.

■

D.2 Hölder Exponents and Constants

Lemma 5 (Gini Deviation Hölder Constants) *Consider K Bernoulli distributions $\{\mathsf{Bern}(p(i)):i\in[K]\}$ with CDFs $\{\mathbb{F}_i:i\in[K]\}$. Consider the optimal mixture $\mathbb{F}^\star=\sum_{i=1}^K\alpha^\star(i)\mathbb{F}_i$, and for a given $\boldsymbol{\alpha}\in\Delta^{K-1}$, consider the mixture $\mathbb{G}=\sum_{i=1}^K\alpha(i)\mathbb{F}_i$. For the Gini deviation DR, i.e., $h(u)=u(1-u)$, we have the following properties.*

1. *The Hölder continuity exponent is $q=1$.*
2. *The Hölder mixture exponent is $r=1$ if $\max_{i\in[K]}p(i)<0.5$ or $\min_{i\in[K]}p(i)>0.5$, and otherwise it is $r=2$.*
3. *The Hölder constant is $\mathcal{L}=\max\{\mathcal{L}_{\mathrm{MH}},\mathcal{L}_{\mathrm{H}}\}=1$.*

Proof: Consider the mixture distributions $\mathbb{F}$ and $\mathbb{G}$ which have Bernoulli distributions with parameters

$$p_{\mathbb{F}}\triangleq\sum_{i=1}^K\alpha_{\mathbb{F}}(i)p(i)\ ,\qquad\text{and}\qquad p_{\mathbb{G}}=\sum_{i=1}^K\alpha_{\mathbb{G}}(i)p(i)\ , \tag{58}$$

respectively. For the Gini DR, it can be easily verified that

$$U_h\left(\mathbb{F}\right)\ =\ p_{\mathbb{F}}(1-p_{\mathbb{F}})\ ,\qquad\text{and}\qquad U_h\left(\mathbb{G}\right)\ =\ p_{\mathbb{G}}(1-p_{\mathbb{G}})\ . \tag{59}$$

Hence, we have

$$\begin{aligned} U_h\left(\mathbb{F}\right)-U_h\left(\mathbb{G}\right) &\overset{(59)}{=} p_{\mathbb{F}}(1-p_{\mathbb{F}})-p_{\mathbb{G}}(1-p_{\mathbb{G}}) && (60)\\ &= (p_{\mathbb{F}}-p_{\mathbb{G}})(1-p_{\mathbb{F}}-p_{\mathbb{G}}) && (61)\\ &\leq |p_{\mathbb{F}}-p_{\mathbb{G}}||1-p_{\mathbb{F}}-p_{\mathbb{G}}| && (62)\\ &\leq |p_{\mathbb{F}}-p_{\mathbb{G}}| && (63)\\ &= \|\mathbb{F}-\mathbb{G}\|_{\mathrm{W}}\ , && (64)\end{aligned}$$

where, (63) follows from the fact that $p_{\mathbb{F}}\leq 1$ and $p_{\mathbb{G}}\leq 1$ and hence, $|1-p_{\mathbb{F}}-p_{\mathbb{G}}|\leq 1$. Hence, for the DR considered, the Hölder constant is $\mathcal{L}_{\mathrm{H}}=1$ and the exponent is $q=1$.

Let us denote the optimal mixture as $\mathbb{F}^\star$ with parameter $p^\star\triangleq\sum_{i=1}^K\alpha^\star(i)p(i)$, based on which $U_h(\mathbb{F}^\star)=p^\star(1-p^\star)$. Without any constraints on $p^\star$, this term will be maximized at $p^\star=\frac{1}{2}$. However, $p^\star$ is constrained to be a mixture of $\{p(i):i\in[K]\}$. This means that $p^\star=\frac{1}{2}$ is not viable when all the p_i's are either larger than $1/2$ or smaller than $1/2$. Hence, depending on the values of $\{p(i):i\in[K]\}$, we analyze two separate cases.

1. **Case 1:** $\min_{i\in[K]} p(i) > 0.5$ **or** $\max_{i\in[K]} p(i) < 0.5$**:** In either of these cases, for the Gini deviation DR and a K-armed Bernoulli bandit instance, it is easy to see that the optimal solution is a solitary arm. Specifically, this arm is given by

$$a_{\min} \in \arg\min_{i\in[K]} p(i) \qquad \text{if} \quad \min_{i\in[K]} p(i) > 0.5 \ , \tag{65}$$

or

$$a_{\max} \in \arg\max_{i\in[K]} p(i) \qquad \text{if} \quad \max_{i\in[K]} p(i) < 0.5 \ . \tag{66}$$

We will characterize the Hölder exponent q for the case that the optimal arm is $a_{\max}$, and the analysis for $a_{\min}$ follows similarly. For $\mathbb{F}^\star$ and $\mathbb{G}$ we have

$$\begin{aligned} U_h(\mathbb{F}^\star) - U_h(\mathbb{G}) &= a_{\max}(1-a_{\max}) - p_{\mathbb{G}}(1-p_{\mathbb{G}}) && (67)\\ &\le |a_{\max} - p_{\mathbb{G}}| && (68)\\ &= \|\mathbb{F}^\star - \mathbb{G}\|_{\mathrm{W}} \ , && (69) \end{aligned}$$

where (68) follows from the exact steps as the transitions (60)-(64). This indicates that $q = 1$ and $\mathcal{L}_{\mathrm{H}} = 1$.

2. **Case 2:** $\exists i \in [K]$ **such that** $p(i) \le \frac{1}{2}$ **and** $\exists i \in [K]$ **such that** $p(i) \ge \frac{1}{2}$**:** In this case, it can be readily verified that $p^\star = 1/2$ is a viable solution, in which case $U_h(\mathbb{F}^\star) = 1/4$. Accordingly, we have

$$\begin{aligned} U_h\left(\mathbb{F}^\star\right) - U_h\left(\mathbb{G}\right) &= \frac{1}{4} - p_{\mathbb{G}}(1-p_{\mathbb{G}}) && (70)\\ &= \left(\frac{1}{2} - p_{\mathbb{G}}\right)^2 && (71)\\ &= \|\mathbb{F}^\star - \mathbb{G}\|_{\mathrm{W}}^2 \ . && (72) \end{aligned}$$

This shows that $\mathcal{L}_{\mathrm{MH}} = 1$ and $r = 2$ in this case.

Hence, in summary $\mathcal{L} = 1$, $q = 1$, and $r = 1$. ■

Lemma 6 (PHT Measure Hölder Constants) *Consider two distinct CDFs $\mathbb{F}$ and $\mathbb{G}$ supported on $[0, \tau]$. For the PHT measure, i.e., $h(u) = u^s$ for some $s \in (0, 1)$, we have the following properties.*

1. *The Hölder continuity exponent is $q = s$.*
2. *The Hölder mixture exponent is $r = s$.*
3. *The Hölder constant is $\mathcal{L} = \max\{\mathcal{L}_{\mathrm{MH}}, \mathcal{L}_{\mathrm{H}}\} = 1$.*

Proof: We have

$$\begin{aligned} U_h(\mathbb{F}) - U_h(\mathbb{G}) &= \int_0^\tau (1-\mathbb{F}(x))^s - (1-\mathbb{G}(x))^s \mathrm{d}x\\ &\le \int_0^\tau |\mathbb{F}(x) - \mathbb{G}(x)|^s \mathrm{d}x\\ &\le \left[\int_0^\tau |\mathbb{F}(x) - \mathbb{G}(x)| \mathrm{d}x\right]^s\\ &\le \|\mathbb{F} - \mathbb{G}\|_{\mathrm{W}}^s \, , \end{aligned}$$

where we used Jensen's inequality for the penultimate inequality and Lemma 3 for the final inequality. This indicates that $q = s$ and $\mathcal{L}_{\mathrm{H}} = 1$. We remark that based on the definitions of Hölder continuity and mixture Hölder continuity provided in (6) and (7), respectively, any given Hölder continuity exponent r is also a mixture Hölder continuity exponent, that is given q, $r = q$ is always a valid choice for r. Hence, we have $r = q = s$ and $\mathcal{L} = \mathcal{L}_{\mathrm{H}} = \mathcal{L}_{\mathrm{MH}} = 1$. ■

Lemma 7 (Wang's Right-Tail Deviation Hölder Constants) *Consider K Bernoulli distributions $\{\mathsf{Bern}(p_i) : i \in [K]\}$ with CDFs $\{\mathbb{F}_i : i \in [K]\}$. Consider the optimal mixture $\mathbb{F}^\star = \sum_{i=1}^K \alpha^\star(i)\mathbb{F}_i$, and for given $\boldsymbol{\alpha}_{\mathbb{F}}, \boldsymbol{\alpha}_{\mathbb{G}} \in \Delta^{K-1}$, consider the mixtures $\mathbb{F} = \sum_{i=1}^K \alpha_{\mathbb{F}}(i)\mathbb{F}_i$ and $\mathbb{G} = \sum_{i=1}^K \alpha_{\mathbb{G}}(i)\mathbb{F}_i$. For the Wang's right-tail deviation, i.e., $h(u) = \sqrt{u} - u$, we have the following properties.*

1. *The Hölder continuity exponent is $q = 1/2$.*
2. *The Hölder mixture exponent is $r = 1$ if $\max_{i\in[K]} p(i) > 0.25$ and $\min_{i\in[K]} p(i) < 0.25$, and otherwise it is $r = 1/2$.*

3. *The Hölder constant is* $\mathcal{L} = \max\{\mathcal{L}_{\rm MH}, \mathcal{L}_{\rm H}\} = 1$.

Proof: Note that that the mixture distributions $\mathbb{F}$ and $\mathbb{G}$ have Bernoulli distributions with parameters $\sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)p(i)$ and $\sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)p(i)$, respectively. Let us define

$$p_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)p(i) \ , \qquad \text{and} \qquad p_{\mathbb{G}} \triangleq \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)p(i) \ . \tag{73}$$

It can be easily verified that

$$U_h\left(\mathbb{F}\right) \ = \ \sqrt{p_{\mathbb{F}}} - p_{\mathbb{F}} \ , \qquad \text{and} \qquad U_h\left(\mathbb{G}\right) \ = \ \sqrt{p_{\mathbb{G}}} - p_{\mathbb{G}} \ . \tag{74}$$

Note that for any $a, b \in [0, +\infty)$ we have

$$\left|\sqrt{a} - \sqrt{b}\right| \leq \sqrt{|a-b|} \ . \tag{75}$$

Hence, we obtain

$$\begin{aligned}
U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(74)}{=} \sqrt{p_{\mathbb{F}}}(1-\sqrt{p_{\mathbb{F}}}) - \sqrt{p_{\mathbb{G}}}(1-\sqrt{p_{\mathbb{G}}}) && (76)\\
&= (\sqrt{p_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}})(1-\sqrt{p_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}}) && (77)\\
&\leq |\sqrt{p_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}}||1-\sqrt{p_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}}| && (78)\\
&\leq |\sqrt{p_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}}| && (79)\\
&\overset{(75)}{\leq} |p_{\mathbb{F}} - p_{\mathbb{G}}|^{1/2} && (80)\\
&= \|\mathbb{F} - \mathbb{G}\|_{\rm W}^{1/2} \ , && (81)
\end{aligned}$$

where, (79) follows from the fact that $p_{\mathbb{F}} \leq 1$ and $p_{\mathbb{G}} \leq 1$ and hence, $|1 - \sqrt{p_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}}| \leq 1$. Hence, we have $\mathcal{L}_{\rm H} = 1$ and $q = 1/2$. Next, we characterize the mixture Hölder exponent. For this purpose, we denote the parameter for the optimal mixture by $p^{\star}_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha^{\star}(i)p(i)$.

1. **Case 1:** $\max_{i\in[K]} p(i) > 0.25$ **and** $\min_{i\in[K]} p(i) < 0.25$**:** In this case, it can be readily verified that $p^{\star}_{\mathbb{F}} = \frac{1}{4}$ and $U_h(\mathbb{F}^{\star}) = \frac{1}{4}$. Note that

$$\begin{aligned}
U_h\left(\mathbb{F}^{\star}\right) - U_h\left(\mathbb{G}\right) &= \frac{1}{4} - \sqrt{p_{\mathbb{G}}} + p_{\mathbb{G}} && (82)\\
&= \left(\frac{1}{2} - \sqrt{p_{\mathbb{G}}}\right)^2 && (83)\\
&= \left(\sqrt{p^{\star}_{\mathbb{F}}} - \sqrt{p_{\mathbb{G}}}\right)^2 && (84)\\
&\overset{(75)}{\leq} |p^{\star}_{\mathbb{F}} - p_{\mathbb{G}}| && (85)\\
&= \|\mathbb{F}^{\star} - \mathbb{G}\|_{\rm W} \ , && (86)
\end{aligned}$$

   Hence, from (82), we can conclude that $\mathcal{L}_{\rm MH} = 1$ and $r = 1$.

2. **Case 2:** $\max_{i\in[K]} p(i) < 0.25$ **or** $\min_{i\in[K]} p(i) > 0.25$**:** In this case, the optimal policy is a solitary arm, in which case $r = q$, i.e., $r = 1/2$ and $\mathcal{L}_{\rm MH} = 1$.

Hence, in summary $\mathcal{L} = 1$, $q = 1$, and $r = 1/2$ or $r = 1$ as specified.

■

Lemma 8 (Mean-median Deviation Hölder Constants) *Consider K Bernoulli distributions $\{\mathsf{Bern}(p_i) : i \in [K]\}$ with CDFs $\{\mathbb{F}_i : i \in [K]\}$. Consider the optimal mixture $\mathbb{F}^{\star} = \sum_{i=1}^{K} \alpha^{\star}(i)\mathbb{F}_i$, and for given $\boldsymbol{\alpha}_{\mathbb{F}}, \boldsymbol{\alpha}_{\mathbb{G}} \in \Delta^{K-1}$, consider the mixtures $\mathbb{F} = \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)\mathbb{F}_i$ and $\mathbb{G} = \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)\mathbb{F}_i$. For the mean-median deviation, i.e., $h(u) = \min\{u, (1-u)\}$, we have the following properties.*

1. *The Hölder continuity exponent is $q = 1$.*
2. *The Hölder mixture exponent is $r = 1$.*
3. *The Hölder constant is $\mathcal{L} = \max\{\mathcal{L}_{\rm MH}, \mathcal{L}_{\rm H}\} = 1$.*

Proof: Let us define

$$p_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)p(i) \ , \qquad \text{and} \qquad p_{\mathbb{G}} = \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)p(i) \ . \tag{87}$$

It can be easily verified that

$$U_h\left(\mathbb{F}\right) \ = \ \min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ , \qquad \text{and} \qquad U_h\left(\mathbb{G}\right) \ = \ \min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} \ . \tag{88}$$

1. **Case 1:** When $p_{\mathbb{F}} > 1/2,\ p_{\mathbb{G}} > 1/2$ or $p_{\mathbb{F}} < 1/2,\ p_{\mathbb{G}} < 1/2$ we have

$$\begin{aligned} U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(88)}{=} \min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ - \min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (89)\\ &\leq \ |p_{\mathbb{F}} - p_{\mathbb{G}}| \ . && (90)\end{aligned}$$

2. **Case 2:** When $p_{\mathbb{F}} < 1/2,\ p_{\mathbb{G}} > 1/2$ we have

$$\begin{aligned} U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(88)}{=} \min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ - \min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (91)\\ &= \ p_{\mathbb{F}} - 1 + p_{\mathbb{G}} && (92)\\ &\leq \ |p_{\mathbb{F}} - p_{\mathbb{G}}| \ . && (93)\end{aligned}$$

3. **Case 3:** When $p_{\mathbb{F}} > 1/2,\ p_{\mathbb{G}} < 1/2$ we have

$$\begin{aligned} U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(88)}{=} \min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ - \min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (94)\\ &= \ 1 - p_{\mathbb{F}} - p_{\mathbb{G}} && (95)\\ &\leq \ |p_{\mathbb{F}} - p_{\mathbb{G}}| \ . && (96)\end{aligned}$$

Hence, overall

$$U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) \leq |p_{\mathbb{F}} - p_{\mathbb{G}}| = \|\mathbb{F} - \mathbb{G}\|_{\mathrm{W}} \ . \tag{97}$$

This indicates that $\mathcal{L}_{\mathrm{H}} = 1$ and $q = 1$. Next, we characterize the mixture Hölder exponent r. For this purpose, let us denote the parameter for the optimal mixture by $p^{\star}_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha^{\star}(i)p(i)$.

1. **Case 1:** $\max_{i\in[K]} p(i) > 0.5$ **and** $\min_{i\in[K]} p(i) < 0.5$**:** It can be readily verified that for $p^{\star} = \frac{1}{2}$ and $U_h(p^{\star}) = 1/2$. Note that

$$\begin{aligned} U_h\left(\mathbb{F}^{\star}\right) - U_h\left(\mathbb{G}\right) &= \frac{1}{2} - \min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (98)\\ &= \left|\frac{1}{2} - p_{\mathbb{G}}\right| && (99)\\ &= \|\mathbb{F} - \mathbb{G}\|_{\mathrm{W}} \ , && (100)\end{aligned}$$

2. **Case 2:** $\max_{i\in[K]} p(i) < 0.5$ **or** $\min_{i\in[K]} p(i) > 0.5$**:** In this case, the optimal solution is a solitary arm, in which case $r = q$, i.e., $r = 1$ and $\mathcal{L}_{\mathrm{MH}} = 1$.

Hence, in summary $\mathcal{L} = 1$, $q = 1$, and $r = 1$.

■

Lemma 9 (Inter-ES Hölder Constants) *Consider K Bernoulli distributions $\{\mathsf{Bern}(p_i) : i \in [K]\}$ with CDFs $\{\mathbb{F}_i : i \in [K]\}$. Consider the optimal mixture $\mathbb{F}^{\star} = \sum_{i=1}^{K} \alpha^{\star}(i)\mathbb{F}_i$, and for given $\boldsymbol{\alpha}_{\mathbb{F}}, \boldsymbol{\alpha}_{\mathbb{G}} \in \Delta^{K-1}$, consider the mixtures $\mathbb{F} = \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)\mathbb{F}_i$ and $\mathbb{G} = \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)\mathbb{F}_i$. For the Inter-ES range with $\alpha = 0.5$, denoted by $\mathrm{IER}_{0.5}$, i.e.,*

$$h(u) = \min\left\{\frac{u}{1-\alpha}, 1\right\} + \min\left\{\frac{\alpha - u}{1-\alpha}, 0\right\} \ ,$$

we have the following properties.

1. *The Hölder continuity exponent is $q = 1$.*
2. *The Hölder mixture exponent is $r = 1$.*
3. *The Hölder constant is $\mathcal{L} = \max\{\mathcal{L}_{\mathrm{MH}}, \mathcal{L}_{\mathrm{H}}\} = 2$.*

Proof: The distortion function for inter-ES range for $\alpha = 1/2$, denoted by $\mathrm{IER}_{0.5}$, is

$$\begin{aligned} h(u) &= \min\{2u, 1\} + \min\{1 - 2u, 0\} && (101)\\ &= 2\min\{u, 1-u\} \ . && (102)\end{aligned}$$

Let us define

$$p_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)p(i) \ , \qquad \text{and} \qquad p_{\mathbb{G}} = \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)p(i) \ . \tag{103}$$

It can be easily verified that

$$U_h\left(\mathbb{F}\right) \ = \ 2\min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ , \qquad \text{and} \qquad U_h\left(\mathbb{G}\right) \ = \ 2\min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} \ . \tag{104}$$

1. **Case 1:** For $p_{\mathbb{F}} > 1/2,\ p_{\mathbb{G}} > 1/2$ or $p_{\mathbb{F}} < 1/2,\ p_{\mathbb{G}} < 1/2$ we have

$$\begin{aligned} U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(88)}{=} \ 2\min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ - 2\min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (105)\\ &\leq \ 2|p_{\mathbb{F}} - p_{\mathbb{G}}| \ . && (106)\end{aligned}$$

2. **Case 2:** For $p_{\mathbb{F}} < 1/2,\ p_{\mathbb{G}} > 1/2$,

$$\begin{aligned} U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(88)}{=} \ 2\min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ - 2\min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (107)\\ &= \ 2(p_{\mathbb{F}} - 1 + p_{\mathbb{G}}) && (108)\\ &\leq \ 2|p_{\mathbb{F}} - p_{\mathbb{G}}| \ . && (109)\end{aligned}$$

3. **Case 3:** For $p_{\mathbb{F}} > 1/2,\ p_{\mathbb{G}} < 1/2$,

$$\begin{aligned} U_h\left(\mathbb{F}\right) - U_h\left(\mathbb{G}\right) &\overset{(88)}{=} \ 2\min\{p_{\mathbb{F}}, (1-p_{\mathbb{F}})\} \ - 2\min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\} && (110)\\ &= \ 2(1 - p_{\mathbb{F}} - p_{\mathbb{G}}) && (111)\\ &\leq \ 2|p_{\mathbb{F}} - p_{\mathbb{G}}| \ . && (112)\end{aligned}$$

Hence, overall, we have

$$|p_{\mathbb{F}} - p_{\mathbb{G}}| = 2\left\|\mathbb{F} - \mathbb{G}\right\|_{\mathrm{W}} \ . \tag{113}$$

This indicates that $\mathcal{L}_{\mathrm{H}} = 2$ and $q = 1$. Next, we characterize the mixture Hölder exponent. For this purpose, let us denote the parameter for the optimal mixture by $p^{\star}_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha^{\star}(i)p(i)$.

1. **Case 1:** $\max_{i\in[K]} p(i) > 0.5$ **and** $\min_{i\in[K]} p(i) < 0.5$**:** It can be readily verified that for $p^{\star} = \frac{1}{2}$ and $U_h(p^{\star}) = 1$. Note that

$$\begin{aligned} U_h\left(\mathbb{F}^{\star}\right) - U_h\left(\mathbb{G}\right) &= 2(\frac{1}{2} - \min\{p_{\mathbb{G}}, (1-p_{\mathbb{G}})\}) && (114)\\ &= 2\left|\frac{1}{2} - p_{\mathbb{G}}\right| && (115)\\ &= 2\left\|\mathbb{F} - \mathbb{G}\right\|_{\mathrm{W}} \ , && (116)\end{aligned}$$

2. **Case 2:** $\max_{i\in[K]} p(i) < 0.5$ **or** $\min_{i\in[K]} p(i) > 0.5$**:** In this case, the optimal solution is a solitary arm, in which case $r = q$, i.e., $r = 1$ and $\mathcal{L}_{\mathrm{MH}} = 2$.

Hence, in summary $\mathcal{L} = 2$, $q = 1$, and $r = 1$.

■

Lemma 10 (Dual Power Hölder Constants) *Consider K Bernoulli distributions $\{\mathsf{Bern}(p_i) : i \in [K]\}$ with CDFs $\{\mathbb{F}_i \ : \ i \in [K]\}$. Consider the optimal mixture $\mathbb{F}^{\star} = \sum_{i=1}^{K} \alpha^{\star}(i)\mathbb{F}_i$, and for given $\boldsymbol{\alpha}_{\mathbb{F}}, \boldsymbol{\alpha}_{\mathbb{G}} \in \Delta^{K-1}$, consider the mixtures $\mathbb{F} = \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)\mathbb{F}_i$ and $\mathbb{G} = \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)\mathbb{F}_i$. For dual power with the parameter $s \geq 2$, i.e., $h(u) = 1 - (1-u)^s$, we have the following properties.*

1. *The Hölder continuity exponent is $q = 1$.*
2. *The Hölder mixture exponent is $r = 1$.*
3. *The Hölder constant is $\mathcal{L} = \max\{\mathcal{L}_{\mathrm{MH}}, \mathcal{L}_{\mathrm{H}}\} = s$.*

Proof: Let us define

$$p_{\mathbb{F}} \triangleq \sum_{i=1}^{K} \alpha_{\mathbb{F}}(i)p(i) \ , \qquad \text{and} \qquad p_{\mathbb{G}} = \sum_{i=1}^{K} \alpha_{\mathbb{G}}(i)p(i) \ . \tag{117}$$

It can be easily verified that

$$U_h\left(\mathbb{F}\right) \ = \ 1 - (1-p_{\mathbb{F}})^s, \qquad \text{and} \qquad U_h\left(\mathbb{G}\right) \ = \ 1 - (1-p_{\mathbb{G}})^s \ . \tag{118}$$

Let us define $p_{\mathrm{M}} \in (\min\{p_{\mathbb{F}}, p_{\mathbb{G}}\}, \max\{p_{\mathbb{F}}, p_{\mathbb{G}}\})$. We have

$$\begin{aligned} U_h(\mathbb{F}) - U_h(\mathbb{G}) &\overset{(118)}{=} 1-(1-p_{\mathbb{F}})^s - (1-(1-p_{\mathbb{G}})^s) && (119)\\ &= (1-p_{\mathbb{G}})^s - (1-p_{\mathbb{F}})^s && (120)\\ &= (p_{\mathbb{F}} - p_{\mathbb{G}}) \cdot s \cdot (1-p_{\mathrm{M}})^{s-1} && (121)\\ &\le |p_{\mathbb{F}} - p_{\mathbb{G}}| \cdot s \cdot (1-p_{\mathrm{M}})^{s-1} && (122)\\ &\le |p_{\mathbb{F}} - p_{\mathbb{G}}| \cdot s && (123)\\ &= s \left\|\mathbb{F} - \mathbb{G}\right\|_{\mathrm{W}} && (124) \end{aligned}$$

where

- (121) follows from observing that the distortion function is continuous and then applying the mean-value theorem,
- (123) follows from the facts that $s \ge 2$ and $p_{\mathrm{M}} \le 1$.

(124) indicates that $q = 1$ and $\mathcal{L}_{\mathrm{H}} = s$. With a similar argument to Lemma 6, $r = q$, and $\mathcal{L}_{\mathrm{MH}} = \mathcal{L}_{\mathrm{H}}$ which means $\mathcal{L} = s$. ■

Lemma 11 (Quadratic Hölder Constants) *Consider K Bernoulli distributions $\{\mathsf{Bern}(p_i) : i \in [K]\}$ with CDFs $\{\mathbb{F}_i : i \in [K]\}$. Consider the optimal mixture $\mathbb{F}^\star = \sum_{i=1}^K \alpha^\star(i)\mathbb{F}_i$, and for given $\boldsymbol{\alpha}_{\mathbb{F}}, \boldsymbol{\alpha}_{\mathbb{G}} \in \Delta^{K-1}$, consider the mixtures $\mathbb{F} = \sum_{i=1}^K \alpha_{\mathbb{F}}(i)\mathbb{F}_i$ and $\mathbb{G} = \sum_{i=1}^K \alpha_{\mathbb{G}}(i)\mathbb{F}_i$. For quadratic with the parameter $s \in [0,1]$, i.e., $h(u) = (1+s)u - su^2$, we have the following properties.*

1. *The Hölder continuity exponent is $q = 1$.*
2. *The Hölder mixture exponent is $r = 1$.*
3. *The Hölder constant is $\mathcal{L} = \max\{\mathcal{L}_{\mathrm{MH}}, \mathcal{L}_{\mathrm{H}}\} = 1 + s$.*

Proof: Let us define

$$p_{\mathbb{F}} \triangleq \sum_{i=1}^K \alpha_{\mathbb{F}}(i)p(i)\ , \qquad \text{and} \qquad p_{\mathbb{G}} = \sum_{i=1}^K \alpha_{\mathbb{G}}(i)p(i)\ . \tag{125}$$

It can be easily verified that

$$U_h(\mathbb{F}) \ = \ (1+s)p_{\mathbb{F}} - sp_{\mathbb{F}}^2, \qquad \text{and} \qquad U_h(\mathbb{G}) \ = \ (1+s)p_{\mathbb{G}} - sp_{\mathbb{G}}^2\ . \tag{126}$$

Then, we have

$$\begin{aligned} U_h(\mathbb{F}) - U_h(\mathbb{G}) &\overset{(126)}{=} (1+s)p_{\mathbb{F}} - sp_{\mathbb{F}}^2 - ((1+s)p_{\mathbb{G}} - sp_{\mathbb{G}}^2) && (127)\\ &= (1+s)(p_{\mathbb{F}} - p_{\mathbb{G}}) - s(p_{\mathbb{F}} - p_{\mathbb{G}})(p_{\mathbb{F}} + p_{\mathbb{G}}) && (128)\\ &= (p_{\mathbb{F}} - p_{\mathbb{G}})((1+s) - s(p_{\mathbb{F}} + p_{\mathbb{G}})) && (129)\\ &\le (1+s)|p_{\mathbb{F}} - p_{\mathbb{G}}| && (130)\\ &= (1+s)\left\|\mathbb{F} - \mathbb{G}\right\|_{\mathrm{W}}\ . && (131) \end{aligned}$$

(131) implies that $q = 1$ and $\mathcal{L}_{\mathrm{H}} = (1+s)$. Similar to Lemma 6, here we choose $r = q$ and $\mathcal{L}_{\mathrm{MH}} = \mathcal{L}_{\mathrm{H}}$, and hence, $\mathcal{L} = (1+s)$.

■

D.3 Algorithms' Properties

Lemma 12 (Discretization Error) *The discretization error $\Delta(\varepsilon)$ is upper bounded as*

$$\Delta(\varepsilon) \le \mathcal{L}(KW)^r \left(\frac{\varepsilon}{2}\right)^r\ . \tag{132}$$

Proof: Let us define $\bar{\mathbf{a}}$ as the discrete mixing coefficient that has the least L_1 distance to the optimal coefficient $\boldsymbol{\alpha}^\star$, i.e.,

$$\bar{\mathbf{a}} \ \in \ \underset{\mathbf{a} \in \Delta_\varepsilon^{K-1}}{\arg\min} \left\|\boldsymbol{\alpha}^\star - \mathbf{a}\right\|_1\ . \tag{133}$$

Accordingly, we have

$$\Delta(\varepsilon) = \mathbb{E}^{\pi}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]}\alpha^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)\right] \tag{134}$$

$$\leq U_h\left(\sum_{i\in[K]}\alpha^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]}\bar{a}(i)\mathbb{F}_i\right) \tag{135}$$

$$\leq \mathcal{L}\|\sum_{i\in[K]}\Big(\alpha^\star(i)-\bar{a}(i)\Big)\mathbb{F}_i\|^r_{\mathrm{W}} \tag{136}$$

$$\leq \mathcal{L}\,\|\boldsymbol{\alpha}^\star - \bar{\mathbf{a}}\|_1^r\, W^r \tag{137}$$

$$\leq \mathcal{L}K^r\left(\frac{\varepsilon}{2}\right)^r W^r, \tag{138}$$

where,

- (136) follows from Definition (2);
- (137) follows from the definition of W in (36);
- and (138) follows from the fact that $\bar{\boldsymbol{\alpha}}$ may lie at most $\varepsilon/2$ away from the optimal coefficient $\boldsymbol{\alpha}^\star$ along each coordinate.

■

Lemma 13 *Consider a K-arm Bernoulli bandit instance with mean values* $\mathbf{p} = (p(1)\cdots p(K))$. *For a DR with concave and strictly monotone distortion function, we have* $\beta = 1$.

Proof: Let $\mathbf{a}'$ denote the second best discrete optimal solution, i.e,

$$\mathbf{a}' \triangleq \underset{\mathbf{a}\in\Delta_\varepsilon^{K-1}:\mathbf{a}\neq\mathbf{a}^\star}{\arg\max}\; U_h\left(\sum_{i\in[K]}a(i)\mathbb{F}_i\right)\ . \tag{139}$$

For a Bernoulli bandit instance, we have the following simplification.

$$U_h\left(\sum_{i\in[K]}a(i)\mathbb{F}_i\right) = h\left(\sum_{i\in[K]}a(i)p_i\right)\ . \tag{140}$$

We order the mean values from smallest to largest, i.e., $p(1) < p(2)\cdots < p(K)$. Since the distortion function is strictly increasing, the optimal solution is a solitary arm, i.e,

$$\max_{\alpha\in\Delta^{K-1}} h\left(\sum_{i\in[K]}\alpha(i)p_i\right) = h(p_K)\ . \tag{141}$$

Equivalently, we have

$$\alpha^\star(i) = \begin{cases}1, & \text{if } i = K\\ 0, & \text{if } i\neq K\end{cases}\ . \tag{142}$$

Furthermore, $\forall\mathbf{a}\in\Delta_\varepsilon^{K-1}$ we have

$$h\left(\sum_{i\in[K]}a(i)p_i\right) < h(p_K)\ . \tag{143}$$

It can be readily verified that the best and the second best discrete mixing coefficients $\mathbf{a}^\star$ and $\mathbf{a}'$ are obtained as follows.

$$a^\star(i) = \begin{cases}1-(K-1)\varepsilon/2, & \text{if } i = K\\ \varepsilon/2, & \text{if } i\neq K\end{cases}\ , \tag{144}$$

and,

$$a'(i) = \begin{cases} 1-(K+1)\varepsilon/2, & \text{if } i = K \\ 3\varepsilon/2, & \text{if } i = K-1 \\ \varepsilon/2, & \text{if } i \neq K-1, K \end{cases} . \tag{145}$$

Note that (144) and (145) are obtained by first choosing the convex combination (in terms of the probabilities $\{p(i) : i \in [K]\}$) with the maximal value, and subsequently leveraging the monotonicity of the distortion function h. Next, we have

$$\Delta_{\min}(\varepsilon) = U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]} a'(i)\mathbb{F}_i\right) \tag{146}$$

$$= h\left(\sum_{i\in[K]} a^\star(i)p(i)\right) - h\left(\sum_{i\in[K]} a'(i)p(i)\right) \tag{147}$$

$$\geq h'\left(\sum_{i\in[K]} a^\star(i)p(i)\right)\left(\sum_{i\in[K]} (a^\star(i)-a'(i))p(i)\right) \tag{148}$$

$$= h'\left(\sum_{i\in[K]} a^\star(i)p(i)\right)(\varepsilon(p(K)-p(K-1))) \tag{149}$$

$$\geq h'(p(K))(p(K)-p(K-1))\cdot\varepsilon\ , \tag{150}$$

where

- (148) follows from concavity of h;
- (149) follows from the definitions of the best and second best discrete mixing coefficients;
- and (150) follows from the fact that leveraging concavity, we have $h'(a) \geq h'(b)$ for $a \leq b$, together with (143).

For a strictly increasing function, the derivative $\forall p \in (0,1)$, $h'(p) \neq 0$. From (150), we can conclude that for DRs with strictly increasing and concave distortion functions $\Delta_{\min}(\varepsilon) = \Omega(\varepsilon)$, and hence $\beta = 1$. ■

E Risk-sensitive Explore Then Commit for Mixtures (RS-ETC-M) Algorithm

In this section, we provide the proofs of Theorems 1 and 2, which characterize the upper bound on the algorithm's regret.

E.1 Regret Decomposition

Throughout this subsection, we prove Theorem 1. We use the decomposition (19) to provide a regret bound on the discrete regret $\bar{\mathfrak{R}}^{\mathrm{E}}_{\boldsymbol{\nu}}(T)$. Let us define the *set* of discrete optimal mixtures of the DR U_h as

$$\mathrm{OPT}_\varepsilon \triangleq \underset{\mathbf{a}\in\Delta_\varepsilon^{K-1}}{\arg\max}\ U_h\left(\sum_{i\in[K]} a(i)\mathbb{F}_i\right)\ , \tag{151}$$

and the *set* of optimistic mixtures computed from the estimated CDFs at time instant t as

$$\widehat{\mathrm{OPT}}_{\varepsilon,t} \triangleq \underset{\mathbf{a}\in\Delta_\varepsilon^{K-1}}{\arg\max}\, U_h\left(\sum_{i\in[K]} a(i)\mathbb{F}^{\mathrm{E}}_{i,t}\right)\ . \tag{152}$$

In the regret decomposition we have provided in (19), the discretization error $\Delta(\varepsilon)$ is upper bounded in Lemma 12. Hence, for regret analysis, we focus on the discrete regret. We can decompose the discrete regret into three main parts as follows.

$$\bar{\mathfrak{R}}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) = \underbrace{\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\right)\right)\mathbb{1}\{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)} \in \mathrm{OPT}_\varepsilon\}\right]}_{\triangleq A_1(T)}$$

$$+\underbrace{\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[\left(U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\right)\right)\mathbb{1}\{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\}\right]}_{\triangleq A_2(T)}$$

$$+\underbrace{\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}\frac{\tau^{\mathrm{E}}_t(i)}{T}\mathbb{F}_i\right)\right]}_{\triangleq A_3(T)}. \tag{153}$$

It can be readily verified that $A_1(T)=0$. The term $A_2(T)$ captures the *mixing coefficient estimation error* when the RS-ETC-M algorithm generates an incorrect mixing coefficient at the end of its exploration phase. Finally, the term $A_3(T)$ captures the *sampling estimation error*, i.e., the error in matching the arm selection fractions to the estimated mixing coefficient at the end of the exploration phase. Next, we provide an upper bound for $A_2(T)$ and $A_3(T)$.

E.2 Upper Bound on the RS-ETC-M Mixing Coefficient Estimation Error $A_2(T)$

Expanding the mixing coefficient estimation error term $A_2(T)$, we obtain

$$A_2(T)=\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[\left(U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\right)\right)\mathbb{1}\{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\}\right] \tag{154}$$

$$=\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\right)\;\middle|\;\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\right]\times\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\right). \tag{155}$$

Furthermore, owing to the fact that the DR is bounded above by B, we have

$$A_2(T)\le B\cdot\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\right). \tag{156}$$

Next, we bound the probability of forming an incorrect estimate of the mixing coefficients. For this, for any t, we define the following events.

$$\mathcal{E}_{1,t}(x)\;\triangleq\;\left\{\left|U_h\Big(\sum_{i\in[K]}a^\star(i)\mathbb{F}^{\mathrm{E}}_{i,t}\Big)-U_h\Big(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\Big)\right|\le x\right\}, \tag{157}$$

$$\mathcal{E}_{2,t}(x)\;\triangleq\;\left\{\left|U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}^{\mathrm{E}}_{i,t}\Big)-U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)\right|\le x\right\}, \tag{158}$$

$$\text{and,}\quad\mathcal{E}_t(x)\;\triangleq\;\mathcal{E}_1(x,t)\bigcap\mathcal{E}_2(x,t). \tag{159}$$

Note that for any $\mathbf{a}^\star\in\mathrm{OPT}_\varepsilon$, we have

$$\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\Big(\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\Big) \tag{160}$$

$$\le\;\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(U_h\Big(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\Big)\ge U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)+\Delta_{\min}(\varepsilon)\right) \tag{161}$$

$$=\;\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(U_h\Big(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\Big)\ge U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)+\Delta_{\min}(\varepsilon)\;\middle|\;\mathcal{E}_{N(\varepsilon)}\Big(\frac{1}{2}\Delta_{\min}(\varepsilon)\Big)\right)\times\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\mathcal{E}_{N(\varepsilon)}\Big(\frac{1}{2}\Delta_{\min}(\varepsilon)\Big)\right) \tag{162}$$

$$+\,\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(U_h\Big(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\Big)\ge U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)+\Delta_{\min}(\varepsilon)\;\middle|\;\overline{\mathcal{E}}_{N(\varepsilon)}\Big(\frac{1}{2}\Delta_{\min}(\varepsilon)\Big)\right)$$

$$\times \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right) . \tag{163}$$

Note that the first term, i.e., (162) is upper bounded by

$$\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(U_h\Big(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\Big) \geq U_h\Big(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big) + \Delta_{\min}(\varepsilon) \;\Big|\; \mathcal{E}_{N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right) \tag{164}$$

$$\leq \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\Bigg(U_h\Big(\sum_{i\in[K]} a^\star(i)\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}\Big) + \frac{1}{2}\Delta_{\min}(\varepsilon) \geq U_h\Big(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}\Big) + \frac{1}{2}\Delta_{\min}(\varepsilon) \;\Big|\; \mathcal{E}_{N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\Bigg) \tag{165}$$

$$= 0 , \tag{166}$$

where

(i) (165) holds due to the conditioning on the event $\mathcal{E}_{N(\varepsilon)}(\Delta_{\min}(\varepsilon)/2)$, and,

(ii) (166) holds since $\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)} \in \widehat{\mathrm{OPT}}_{\varepsilon,N}$.

The second term, i.e., the term in (163) is upper bounded by

$$\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right) \;\leq\; \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{1,N}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right) + \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{2,N}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right) . \tag{167}$$

Expanding each term, we have

$$\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{1,N}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right) = \mathbb{P}\left(\left|U_h\Big(\sum_{i\in[K]} a^\star(i)\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}\Big) - U_h\Big(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\Big)\right| > \frac{1}{2}\Delta_{\min}(\varepsilon)\right) \tag{168}$$

$$\overset{(6)}{\leq} \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\mathcal{L}\sum_{i\in[K]}\left\|a^\star(i)^\star(\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)} - \mathbb{F}_i)\right\|^q_{\mathrm{W}} > \frac{1}{2}\Delta_{\min}(\varepsilon)\right) \tag{169}$$

$$\leq \sum_{i\in[K]} \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\mathcal{L}\underbrace{\left(a^\star(i)\right)^q}_{\leq 1}\left\|\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)} - \mathbb{F}_i\right\|^q_{\mathrm{W}} > \frac{1}{2K}\Delta_{\min}(\varepsilon)\right) \tag{170}$$

$$\leq \sum_{i\in[K]} \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\left\|\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)} - \mathbb{F}_i\right\|_{\mathrm{W}} > \left(\frac{1}{2K\mathcal{L}}\Delta_{\min}(\varepsilon)\right)^{\frac{1}{q}}\right) \tag{171}$$

$$\overset{(9)}{\leq} \sum_{i\in[K]} 2\exp\left(-\frac{\tau^{\mathrm{E}}_{N(\varepsilon)}(i)}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q} - \frac{512}{\sqrt{\tau^{\mathrm{E}}_{N(\varepsilon)}}}\right)^2\right) \tag{172}$$

$$= \sum_{i\in[K]} 2\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q} - \frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right) \tag{173}$$

$$= 2K\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q} - \frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right) . \tag{174}$$

Furthermore, we have

$$\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{2,N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)$$

$$= \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\left|U_h\Big(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}\Big) - U_h\Big(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)\right| > \frac{1}{2}\Delta_{\min}(\varepsilon)\right) \tag{175}$$

$$\leq \mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\bigcup_{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\in\widehat{\mathrm{OPT}}_{\varepsilon,N(\varepsilon)}}\left|U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}\Big)-U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)\right|>\frac{1}{2}\Delta_{\min}(\varepsilon)\right) \tag{176}$$

$$\leq \sum_{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\in\widehat{\mathrm{OPT}}_{\varepsilon,N(\varepsilon)}}\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\left|U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}\Big)-U_h\Big(\sum_{i\in[K]}a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\Big)\right|>\frac{1}{2}\Delta_{\min}(\varepsilon)\right) \tag{177}$$

$$\overset{(6)}{\leq}\sum_{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\in\widehat{\mathrm{OPT}}_{\varepsilon,N(\varepsilon)}}\sum_{i\in[K]}\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\left\|\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}-\mathbb{F}_i\right\|_{\mathrm{W}}>\left(\frac{1}{2K\mathcal{L}\underbrace{\left(a^{\mathrm{E}}_{N(\varepsilon)}(i)\right)^q}_{\leq 1}}\Delta_{\min}(\varepsilon)\right)^{\frac{1}{q}}\right) \tag{178}$$

$$\leq \sum_{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\in\widehat{\mathrm{OPT}}_{\varepsilon,N(\varepsilon)}}\sum_{i\in[K]}\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\left\|\mathbb{F}^{\mathrm{E}}_{i,N(\varepsilon)}-\mathbb{F}_i\right\|_{\mathrm{W}}>\left(\frac{1}{2K\mathcal{L}}\Delta_{\min}(\varepsilon)\right)^{\frac{1}{q}}\right) \tag{179}$$

$$\overset{(9)}{\leq}\sum_{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\in\widehat{\mathrm{OPT}}_{\varepsilon,N(\varepsilon)}}\sum_{i\in[K]}\exp\left(-\frac{\tau^{\mathrm{E}}_{N(\varepsilon)}(i)}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q}-\frac{512}{\sqrt{\tau^{\mathrm{E}}_{N(\varepsilon)}}}\right)^2\right) \tag{180}$$

$$\leq \sum_{\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\in\widehat{\mathrm{OPT}}_{\varepsilon,N(\varepsilon)}}2K\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q}-\frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right) \tag{181}$$

$$\leq 2K\left(\frac{1}{\varepsilon}\right)^{K-1}\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q}-\frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right), \tag{182}$$

where (182) follows from the fact that the total number of discrete $\mathbf{a}$ values may not exceed $\left(\frac{1}{\varepsilon}\right)^{K-1}$. Combining (174) and (182), we obtain

$$\begin{aligned}
&\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{1,N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)+\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{2,N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)\\
&\quad\leq 2K\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q}-\frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right)\\
&\qquad+\left(\frac{1}{\varepsilon}\right)^{K-1}2K\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q}-\frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right) \qquad (183)\\
&\quad=2K\exp\left(-\frac{\frac{N(\varepsilon)}{K}}{256e}\left(\left(\frac{1}{2K\mathcal{L}(a^\star(i))^q}\Delta_{\min}(\varepsilon)\right)^{1/q}-\frac{512}{\sqrt{\frac{N(\varepsilon)}{K}}}\right)^2\right)\cdot\left(\left(\frac{1}{\varepsilon}\right)^{K-1}+1\right). \qquad (184)
\end{aligned}$$

At this step, for some $\delta\in(0,1)$, choosing $N(\varepsilon)$ as

$$N(\varepsilon)\triangleq 256K\mathrm{e}\left(\frac{2K\mathcal{L}}{\Delta_{\min}(\varepsilon)}\right)^{\frac{2}{q}}\left[\frac{32}{\sqrt{\mathrm{e}}}+\log^{\frac{1}{2}}\Big(2K\delta\big(\varepsilon^{-(K-1)}+1\big)\Big)\right]^2, \tag{185}$$

and leveraging (184), it can be readily verified that

$$\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{1,N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)+\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\overline{\mathcal{E}}_{2,N(\varepsilon)}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)\ \leq\ \delta\,, \tag{186}$$

which implies that $\mathbb{P}^{\mathrm{E}}_{\boldsymbol{\nu}}\left(\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)}\notin\mathrm{OPT}_\varepsilon\right)\leq\delta$. Hence, we have

$$A_2(T)\overset{(156)}{\leq}B\cdot\delta\,. \tag{187}$$

E.3 Upper Bound on the RS-ETC-M Sampling Estimation Error $A_3(T)$

Next, we turn to analyzing $A_3(T)$, which captures the sampling estimation error. We have

$$A_3(T) = \mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{E}}_T(i)}{T}\mathbb{F}_i\right)\right] \tag{188}$$

$$\leq \mathcal{L}\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[\left\|\mathbf{a}^{\mathrm{E}}_{N(\varepsilon)} - \frac{\boldsymbol{\tau}^{\mathrm{E}}_T}{T}\right\|_1^q W^q\right] \tag{189}$$

$$\leq \mathcal{L}KW^q\mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}}\left[\max_{i\in[K]}\left|a^{\mathrm{E}}_{N(\varepsilon)}(i) - \frac{\tau^{\mathrm{E}}_t(i)}{T}\right|^q\right], \tag{190}$$

where (189) follows from the Hölder continuity of the utility stated in (6) and recalling the definition of W stated in (36). We require to upper bound the estimation error due to the policy's sampling proportions. Note that as a result of the "commitment" process of the ETC algorithm stated in (3.2), for the first $K-1$ arms, the RS-ETC-M commits its arms selection such that it is pulled $\max\{0, \lfloor Ta^{\mathrm{E}}_{N(\varepsilon)}(i)\rfloor - N(\varepsilon)/K\}$ times post exploration, and it allocates the remaining rounds to arm K. Let us define the set

$$\mathcal{S}' \triangleq \left\{i \in [K] : a^{\mathrm{E}}_{N(\varepsilon)}(i) < N(\varepsilon)/KT\right\}. \tag{191}$$

Accordingly, for any arm $i \notin \mathcal{S}'$, if $i \neq K$, we have the following bound.

$$\left|\frac{\tau^{\mathrm{E}}_t(i)}{T} - a^{\mathrm{E}}_{N(\varepsilon)}(i)\right| \overset{(191)}{=} a^{\mathrm{E}}_{N(\varepsilon)}(i) - \frac{\tau^{\mathrm{E}}_t(i)}{T} \tag{192}$$

$$\overset{(3.2)}{=} a^{\mathrm{E}}_{N(\varepsilon)}(i) - \frac{\lfloor Ta^{\mathrm{E}}_{N(\varepsilon)}(i)\rfloor}{T} \tag{193}$$

$$\leq a^{\mathrm{E}}_{N(\varepsilon)}(i) - \frac{Ta^{\mathrm{E}}_{N(\varepsilon)}(i) - 1}{T} \tag{194}$$

$$< \frac{1}{T}. \tag{195}$$

Alternatively, if $K \notin \mathcal{S}'$, we have

$$\left|\frac{\tau^{\mathrm{E}}_T(K)}{T} - a^{\mathrm{E}}_{N(\varepsilon)}(K)\right| \overset{(191)}{=} a^{\mathrm{E}}_{N(\varepsilon)}(K) - \frac{\tau^{\mathrm{E}}_T(K)}{T} \tag{196}$$

$$\overset{(3.2)}{=} a^{\mathrm{E}}_{N(\varepsilon)}(K) - \frac{T - \sum_{i\neq K}\tau^{\mathrm{E}}_t(i)}{T} \tag{197}$$

$$= a^{\mathrm{E}}_{N(\varepsilon)}(K) - 1 - \frac{1}{T}\left(\sum_{i\in\mathcal{S}':i\neq K}\tau^{\mathrm{E}}_t(i) + \sum_{i\notin\mathcal{S}':i\neq K}\tau^{\mathrm{E}}_t(i)\right) \tag{198}$$

$$= a^{\mathrm{E}}_{N(\varepsilon)}(K) - 1 - \frac{1}{T}\left(\frac{N(\varepsilon)}{K}|\mathcal{S}'| + \sum_{i\notin\mathcal{S}':i\neq K}\lfloor Ta^{\mathrm{E}}_{N(\varepsilon)}(i)\rfloor\right) \tag{199}$$

$$\leq a^{\mathrm{E}}_{N(\varepsilon)}(K) - 1 - \frac{1}{T}\left(\frac{N(\varepsilon)}{K}|\mathcal{S}'| + \sum_{i\notin\mathcal{S}':i\neq K}Ta^{\mathrm{E}}_{N(\varepsilon)}(i)\right) \tag{200}$$

$$= \sum_{i\notin\mathcal{S}'} a^{\mathrm{E}}_{N(\varepsilon)}(i) - 1 + \frac{N(\varepsilon)|\mathcal{S}'|}{KT} \tag{201}$$

$$= \frac{N(\varepsilon)|\mathcal{S}'|}{KT} - \sum_{i\in\mathcal{S}'} a^{\mathrm{E}}_{N(\varepsilon)}(i) \tag{202}$$

$$\leq \sum_{i\in\mathcal{S}'}\frac{N(\varepsilon)}{KT} + \frac{N(\varepsilon)|\mathcal{S}'|}{KT} \tag{203}$$

$$= \frac{2N(\varepsilon)|\mathcal{S}'|}{KT} \tag{204}$$

$$\leq \frac{2N(\varepsilon)}{T} , \tag{205}$$

where,

- (199) follows from the fact that for every $i \in \mathcal{S}'$, since these arms have already been over-explored, they are not going to get sampled further by the RS-ETC-M algorithm;

- and (203) follows from the definition of the set $\mathcal{S}$, which dictates that for every arm $i \in \mathcal{S}'$, we have $a^{\mathrm{E}}_{N(\varepsilon)}(i) < N(\varepsilon)/KT$.

Finally, for every $i \in \mathcal{S}'$, we have

$$\left| \frac{\tau^{\mathrm{E}}_T(i)}{T} - a^{\mathrm{E}}_{N(\varepsilon)}(i) \right| \overset{(191),(3.2)}{=} \left| \frac{N(\varepsilon)}{KT} - a^{\mathrm{E}}_{N(\varepsilon)}(i) \right| \tag{206}$$

$$\overset{(191)}{=} \frac{N(\varepsilon)}{KT} - a^{\mathrm{E}}_{N(\varepsilon)}(i) \tag{207}$$

$$< \frac{N(\varepsilon)}{KT} . \tag{208}$$

Hence, combining (195), (205), and (208), we conclude that for any $i \in [K]$, we have

$$\left| \frac{\tau^{\mathrm{E}}_T(i)}{T} - a^{\mathrm{E}}_{N(\varepsilon)}(i) \right| \leq \frac{2N(\varepsilon)}{T} . \tag{209}$$

Leveraging (209), we can now upper bounded $A_3(T)$ as follows.

$$A_3(T) = \mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}} \left[U_h \left(\sum_{i\in[K]} a^{\mathrm{E}}_{N(\varepsilon)}(i) \mathbb{F}_i \right) - U_h \left(\sum_{i\in[K]} \frac{\tau^{\mathrm{E}}_t(i)}{T} \mathbb{F}_i \right) \right] \tag{210}$$

$$\leq \mathcal{L} K W^q \mathbb{E}^{\mathrm{E}}_{\boldsymbol{\nu}} \left[\max_{i\in[K]} \left| a^{\mathrm{E}}_{N(\varepsilon)}(i) - \frac{\tau^{\mathrm{E}}_t(i)}{T} \right|^q \right] \tag{211}$$

$$\overset{(209)}{\leq} \mathcal{L} K W^q \left(\frac{2N(\varepsilon)}{T} \right)^q . \tag{212}$$

E.4 Proof of Theorem 1

We can upper bound the discretized regret leveraging the upper bounds on $A_1(T)$, $A_2(T)$ in (187), and $A_3(T)$ in (212) and considering $\delta = 1/T^2$. We obtain

$$\bar{\mathfrak{R}}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) \overset{(153)}{=} A_1(T) + A_2(T) + A_3(T) \tag{213}$$

$$\leq \frac{B}{T^2} + \mathcal{L} K \left(3W \frac{N(\varepsilon)}{T} \right)^q \tag{214}$$

$$\leq (\mathcal{L} K + W^{-q}) \left(3W \frac{N(\varepsilon)}{T} \right)^q , \tag{215}$$

where (215) follows from the fact that B/T^2 is upper bounded by $(N(\varepsilon)/T)^q$. Finally, combining all the terms $A_1(T)$, $A_2(T)$ in (187), $A_3(T)$ in (212), and the discretization error $\Delta(\varepsilon)$ in (12), we obtain

$$\mathfrak{R}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) \overset{(19)}{=} \Delta(\varepsilon) + \bar{\mathfrak{R}}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) \tag{216}$$

$$\leq \Delta(\varepsilon) + (\mathcal{L} K + W^{-q}) \left(\frac{2N(\varepsilon)}{T} \right)^q \tag{217}$$

$$= \Delta(\varepsilon) + (\mathcal{L} K + W^{-q}) \left(3W M(\varepsilon) \frac{\log T}{T} \right)^q , \tag{218}$$

where (218) follows by defining $M(\varepsilon) \triangleq N(\varepsilon)/\log T$.

E.5 Proof of Theorem 2

Let us set $\varepsilon = \Theta((K^{2+2/q}T^{-\gamma}\log T)^{q/2\beta})$ for $\gamma \in (0,1)$. Assuming $\Delta_{\min}(\varepsilon) = \Omega(\varepsilon^{\beta})$, from (11), the exploration horizon is $N(\varepsilon) = \Theta(T^{\gamma})$. Hence, for discrete regret, from (218) we have

$$\bar{\mathfrak{R}}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) = O(KT^{(\gamma-1)q}) \ . \tag{219}$$

Also, from Lemma (12), we have

$$\Delta(\varepsilon) \leq \mathcal{L}(KW)^r \left(\frac{\varepsilon}{2}\right)^r \ . \tag{220}$$

As a result, $\Delta(\varepsilon) = O(\varepsilon^r)$. Hence, by our choice of ε, we obtain

$$\Delta(\varepsilon) = O\left(K^{r+\frac{r(q+1)}{\beta}}T^{-\frac{rq\gamma}{2\beta}}(\log T)^{\frac{rq}{2\beta}}\right). \tag{221}$$

From (216), (219), and (221) the regret of the RS-ETC-M algorithm is upper bounded by

$$\mathfrak{R}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) = O\left(\max\left\{K^{r+\frac{r(q+1)}{\beta}}T^{-\frac{rq\gamma}{2\beta}}(\log T)^{\frac{rq}{2\beta}} \ , \ K^qT^{(\gamma-1)q}\right\}\right) \ . \tag{222}$$

Finally, setting $\gamma = \frac{2\beta}{2\beta+r}$, the regret of the RS-ETC-M algorithm is upper bounded by

$$\mathfrak{R}^{\mathrm{E}}_{\boldsymbol{\nu}}(T) \ \leq O\Big(\Big[K^{2(q+\beta+1)}\ T^{-\frac{q}{1+r/2\beta}}\ (\log T)^q\Big]^{\frac{r}{2\beta}}\Big) \ . \tag{223}$$

F Risk-sensitive Upper-Confidence Bound (RS-UCB-M) Algorithm

In this section, we provide the proofs of Theorems 3 and 4, which characterize the upper bound on the algorithm's regret.

F.1 Regret Decomposition

In (19), the regret is decomposed into a discretization error component, and the discrete regret. Lemma 12 shows an upper bound for the discretization error. In this section, we provide an upper bound on discrete regret $\bar{\mathfrak{R}}^{\mathrm{U}}_{\boldsymbol{\nu}}(T)$ of Algorithm 2 as follows. Recall that we have defined $\tau^{\mathrm{U}}_t(i)$ as the number of times RS-UCB-M selects arm $i \in [K]$ up to time t. We have

$$\bar{\mathfrak{R}}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) = U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]}\frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\right] \tag{224}$$

$$\leq \underbrace{\sum_{\mathcal{S}\subseteq[K]:\mathcal{S}\neq\emptyset}\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]}\frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\right)\mathbb{1}\{\mathbb{F}_i \notin \mathcal{C}_T(i) : i\in\mathcal{S}\}\right]}_{\triangleq B_1(T)} \tag{225}$$

$$+ \underbrace{\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]}\frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\right)\mathbb{1}\{\mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\}\right]}_{\triangleq B_2(T)} \ . \tag{226}$$

Expanding $B_1(T)$, we have

$$B_1(T) = \sum_{\mathcal{S}\subseteq[K]:\mathcal{S}\neq\emptyset}\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbb{F}_i \notin \mathcal{C}_T(i) : i\in\mathcal{S}\Big)$$
$$\times \quad \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i=1}^{K}\frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\,\middle|\,\mathbb{F}_i \notin \mathcal{C}_T(i) : i\in\mathcal{S}\right] \ . \tag{227}$$

Leveraging the fact that

$$\sum_{i=1}^{K} \binom{K}{i} x^i = (x+1)^K - 1 \ , \tag{228}$$

along with the large deviation bound on the confidence sequences from Lemma 2, i.e., $\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}(\mathbb{F}_i \notin \mathcal{C}_T(i)) \leq 1/T^2$ for every $i \in [K]$, we have

$$\sum_{\mathcal{S}\subseteq[K]:\mathcal{S}\neq\emptyset} \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbb{F}_i \notin \mathcal{C}_T(i) : i \in \mathcal{S}\Big) = \left(\frac{2}{T^2}+1\right)^K - 1 \ . \tag{229}$$

Furthermore, owing to the fact that $U_h(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i) \leq B$ for any $\boldsymbol{\alpha} \in \Delta^{K-1}$, we have the following upper bound on $B_1(T)$.

$$B_1(T) \ \leq \ B\left(\left(\frac{2}{T^2}+1\right)^K - 1\right) \ . \tag{230}$$

Next, we will upper bound $B_2(T)$. For this purpose, we begin by defining the upper confidence bound of a parameter $\boldsymbol{\alpha} \in \Delta^{K-1}$ at time $t \in \mathbb{N}$ as

$$\mathrm{UCB}_t(\boldsymbol{\alpha}) \ \triangleq \ \max_{\eta_1\in\mathcal{C}_t(1),\cdots,\mathcal{C}_t(K)} \ U\left(\sum_{i\in[K]} \alpha(i)\eta_i\right) \ . \tag{231}$$

Furthermore, let us define the optimistic CDF estimates which maximize the upper confidence bound for every arm $i \in [K]$ as

$$\widetilde{\mathbb{F}}_{i,t} \ \triangleq \ \operatorname*{arg\,max}_{\eta_i\in\mathcal{C}_t(i)} \ \max_{\eta_j\in\mathcal{C}_t(j):j\neq i} \mathrm{UCB}_t(\mathbf{a}^{\mathrm{U}}_t) \ . \tag{232}$$

Expanding $B_2(T)$, we have

$$B_2(T) = \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} \bar{\alpha}^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i) \ \forall i \in [K]\right] \tag{233}$$

$$\overset{(231)}{\leq} \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\mathrm{UCB}_T(\mathbf{a}^\star) - U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i) \ \forall i \in [K]\right] \tag{234}$$

$$\overset{(16)}{\leq} \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\mathrm{UCB}_T(\mathbf{a}^{\mathrm{U}}_T) - U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i) \ \forall i \in [K]\right] \tag{235}$$

$$= \underbrace{\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\mathrm{UCB}_T(\mathbf{a}^{\mathrm{U}}_T) - U_h\left(\sum_{i\in[K]} a^{\mathrm{U}}_T(i)\mathbb{F}_i\right)\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i) \ \forall i \in [K]\right]}_{\triangleq B_{21}(T)}$$

$$+ \underbrace{\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} a^{\mathrm{U}}_T(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{U}}_T(i)}{T}\mathbb{F}_i\right)\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i) \ \forall i \in [K]\right]}_{B_{22}(T)} \ . \tag{236}$$

Note that the term $B_{21}(T)$ captures the *CDF estimation error* incurred by the UCB algorithm, and the term $B_{22}(T)$ reflects the *sampling estimation error* incurred by the UCB algorithm. Subsequently, we analyze each of these components.

F.2 Upper Bound on the CDF Estimation Error

Expanding $B_{21}(T)$ we obtain

$$B_{21}(T) \leq \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} a^{\mathrm{U}}_T(i)\widetilde{\mathbb{F}}_{i,T}\right) - U_h\left(\sum_{i\in[K]} a^{\mathrm{U}}_T(i)\mathbb{F}_i\right)\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\right] \tag{237}$$

$$\overset{(7)}{\leq} \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\mathcal{L}\sum_{i\in[K]}\left(a^{\mathrm{U}}_T(i)\right)^q \left\|\widetilde{\mathbb{F}}_{i,T} - \mathbb{F}_i\right\|^q_{\mathrm{W}}\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\right] \tag{238}$$

$$\overset{(15)}{\leq} \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\mathcal{L}(32)^q\Big(\sqrt{2\mathrm{e}\log T}+32\Big)^q \sum_{i\in[K]}\left(a^{\mathrm{U}}_T(i)\right)^q\left(\frac{1}{\tau^{\mathrm{U}}_T(i)}\right)^{\frac{q}{2}}\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\right] \tag{239}$$

$$\leq \mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\mathcal{L}\left(32\Big(\sqrt{2\mathrm{e}\log T}+32\Big)\right)^q \cdot \sum_{i\in[K]}\left(\frac{1}{\tau^{\mathrm{U}}_T(i)}\right)^{\frac{q}{2}}\middle|\, \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\right] \tag{240}$$

$$\leq \mathcal{L}\left(32\Big(\sqrt{2\mathrm{e}\log T}+32\Big)\right)^q \cdot \sum_{i\in[K]}\left(\frac{1}{\frac{\rho}{4}T\varepsilon}\right)^{\frac{q}{2}} \tag{241}$$

$$= \frac{\mathcal{L}K}{T}\left(\frac{32}{\sqrt{\frac{\rho}{4}}}\right)^q T^{1-\frac{q}{2}}\left(\frac{\Big(\sqrt{2\mathrm{e}\log T}+32\Big)}{\sqrt{\varepsilon}}\right)^q\ , \tag{242}$$

where,

- (240) follows from the fact that $a^{\mathrm{U}}_T(i) \leq 1$ for every $i \in [K]$;
- and (241) follows from the explicit exploration of each arm $i \in [K]$ in Algorithm 3.

F.3 Upper Bound on the Sampling Estimation Error

Finally, we will bound $B_{22}(T)$. The key idea in upper bounding the sampling estimation error is to show that after a certain instant, the RS-UCB-M algorithm outputs an estimate $\boldsymbol{\alpha}^{\mathrm{U}}_t$ of the mixing coefficients that belongs to the set of discrete optimal mixtures with a very high probability. Subsequently, we show that the undersampling routine of the RS-UCB-M algorithm ensures that once it has identified a correct optimal mixture, it navigates its sampling proportions to match the estimated coefficient. Let us denote the vector of arm sampling fractions by $\boldsymbol{\tau}^{\mathrm{U}}_T \triangleq [\tau^{\mathrm{U}}_T(1), \cdots, \tau^{\mathrm{U}}_T(K)]^\top$. For the first step, note that

$$B_{22}(T) \overset{(6)}{\leq} \mathcal{L}\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\left\|\sum_{i\in[K]}\left(a^{\mathrm{U}}_T(i) - \frac{\tau^{\mathrm{U}}_T(i)}{T}\right)\mathbb{F}_i\right\|^q_{\mathrm{W}}\ \middle|\ \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\right] \tag{243}$$

$$\overset{(36)}{\leq} \mathcal{L}\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[W^q \cdot \left\|\mathbf{a}^{\mathrm{U}}_T - \frac{1}{T}\boldsymbol{\tau}^{\mathrm{U}}_T\right\|^q_1\ \middle|\ \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall i\in[K]\right]\ . \tag{244}$$

In order to bound $B_{22}(T)$, we will leverage a low probability event (which we will call $\mathcal{E}_{3,T}$), and expand (244) by conditioning on this event. In order to define the low probability event, let us first lay down a few definitions. First, let us define the event

$$\mathcal{E}_{0,t} \triangleq \Big\{\mathbb{F}_i \in \mathcal{C}_t(i) \quad \forall i\in[K]\Big\}\ . \tag{245}$$

Let us denote the *set* of discrete optimal mixtures of the DR U_h by

$$\mathrm{OPT}_\varepsilon \triangleq \operatorname*{arg\,max}_{\mathbf{a}\in\Delta^{K-1}_\varepsilon} U_h\left(\sum_{i\in[K]} a(i)\mathbb{F}_i\right)\ , \tag{246}$$

and the *set* of optimistic mixtures at each instant t by

$$\widetilde{\mathrm{OPT}}_{\varepsilon,t} \triangleq \operatorname*{arg\,max}_{\mathbf{a}\in\Delta^{K-1}_\varepsilon}\ \max_{\eta_i\in\mathcal{C}_t(i)\ \forall i\in[K]} U_h\left(\sum_{i\in[K]} a(i)\eta_i\right)\ . \tag{247}$$

Furthermore, for any $\mathbf{a}^\star \in \mathrm{OPT}_\varepsilon$ and $\widetilde{\boldsymbol{\alpha}}_t \in \widetilde{\mathrm{OPT}}_{\varepsilon,t}$, let us define the events

$$\mathcal{E}_{1,t}(x) \triangleq \left\{ \left| U_h\left(\sum_{i\in[K]} a^\star(i)\widetilde{\mathbb{F}}_{i,t}\right) - U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right)\right| < x \right\} , \tag{248}$$

$$\mathcal{E}_{2,t}(x) \triangleq \left\{ \left| U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\widetilde{\mathbb{F}}_{i,t}\right) - U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\mathbb{F}_i\right)\right| < x \right\} , \tag{249}$$

and

$$\mathcal{E}_t(x) \triangleq \mathcal{E}_{1,t}(x) \bigcap \mathcal{E}_{2,t}(x) \ . \tag{250}$$

Note that

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\widetilde{\mathrm{OPT}}_{\varepsilon,t} \neq \mathrm{OPT}_\varepsilon\Big) \ = \ \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\exists \mathbf{a} \in \widetilde{\mathrm{OPT}}_{\varepsilon,t} : \mathbf{a} \notin \mathrm{OPT}_\varepsilon\Big) \ . \tag{251}$$

Let us denote the mixing coefficients that is contained in $\widetilde{\mathrm{OPT}}_{\varepsilon,t}$ and yet not in OPT_ε by $\mathbf{a}_t^{\mathrm{U}}$. Accordingly, we have

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_\varepsilon\Big) = \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_\varepsilon \mid \bar{\mathcal{E}}_{0,t}\Big)\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}(\bar{\mathcal{E}}_{0,t}) + \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big)\mathbb{P}(\mathcal{E}_{0,t}) \tag{252}$$

$$\leq \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}(\bar{\mathcal{E}}_{0,t}) + \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big) \tag{253}$$

$$\overset{(229)}{\leq} \left(\left(\frac{2}{T^2}+1\right)^K - 1\right) + \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big) \ . \tag{254}$$

Next, note that

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big)$$

$$= \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) > U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\mathbb{F}_i\right) + \Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t}\right) \tag{255}$$

$$= \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) > U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\mathbb{F}_i\right) + \Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t},\ \mathcal{E}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)$$
$$\times\ \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\mathcal{E}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \tag{256}$$

$$+\ \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) > U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\mathbb{F}_i\right) + \Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t},\ \bar{\mathcal{E}}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)$$
$$\times\ \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \tag{257}$$

$$\leq \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) > U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\mathbb{F}_i\right) + \Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t},\ \mathcal{E}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)$$
$$+\ \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \tag{258}$$

$$\leq \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(U_h\left(\sum_{i\in[K]} a^\star(i)\widetilde{\mathbb{F}}_{i,t}\right) > U_h\left(\sum_{i\in[K]} a_t^{\mathrm{U}}(i)\widetilde{\mathbb{F}}_{i,t}\right)\middle|\mathcal{E}_{0,t},\ \bar{\mathcal{E}}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\right)$$
$$+\ \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \tag{259}$$

$$= \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \tag{260}$$

$$\leq \mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_{1,t}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right)+\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_{2,t}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right)\ , \tag{261}$$

where

- (259) follows from the definition of the event $\mathcal{E}_t(\Delta_{\min}(\frac{1}{2}\varepsilon))$;
- (260) follows from the definitions of $\mathbf{a}^\star$ (which is an optimizer in the set OPT_ε) and $\boldsymbol{\alpha}^{\mathrm{U}}_t$ (which is an optimizer in the set $\widetilde{\mathrm{OPT}}_{\varepsilon,t}$);
- and (261) follows from a union bound.

Next, we will find upper bounds on the two probability terms in (261). Note that for any $t > K\lceil\rho T\varepsilon/4\rceil$ we have

$$\begin{aligned}
&\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_{1,t}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right)\\
&\quad=\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\left|U_h\left(\sum_{i\in[K]}a^\star(i)\widetilde{\mathbb{F}}_{i,t}\right)-U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)\right|\geq\frac{1}{2}\Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t}\right) &(262)\\
&\quad\overset{(6)}{\leq}\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\sum_{i\in[K]}(a^\star(i))^q\left\|\widetilde{\mathbb{F}}_{i,t}-\mathbb{F}_i\right\|^q_{\mathrm{W}}\geq\frac{1}{2\mathcal{L}}\Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t}\right) &(263)\\
&\quad\overset{(15)}{=}\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\sum_{i\in[K]}\left(a^\star(i)\right)^q\left\|\widetilde{\mathbb{F}}_{i,t}-\mathbb{F}_i\right\|^q_{\mathrm{W}}\middle|\mathcal{E}_{0,t}\right) &(264)\\
&\quad\overset{(15)}{\leq}\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\sum_{i\in[K]}\underbrace{\left(a^\star(i)\right)^q}_{\leq 1}\left(16\sqrt{\frac{1}{\tau^{\mathrm{U}}_T}(i)}\cdot\left(\sqrt{2\mathrm{e}\log T}+32\right)\right)^q>\frac{1}{2\mathcal{L}}\Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t}\right) &(265)\\
&\quad\leq\sum_{i\in[K]}\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\left(\sqrt{\frac{1}{\tau^{\mathrm{U}}_T}(i)}\cdot\left(\sqrt{2\mathrm{e}\log T}+32\right)\right)^q>\frac{1}{2K\mathcal{L}(16)^q}\Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t}\right) &(266)\\
&\quad\leq\sum_{i\in[K]}\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\left(\frac{\left(\sqrt{2\mathrm{e}\log T}+32\right)}{\sqrt{\frac{\rho}{4}T\varepsilon}}\right)^q>\frac{1}{2K\mathcal{L}(16)^q}\Delta_{\min}(\varepsilon)\middle|\mathcal{E}_{0,t}\right), &(267)
\end{aligned}$$

where the transition (266)-(267) follows from the explicit exploration phase of the RS-UCB-M algorithm. Now, let us define a time instant $T_0(\varepsilon)$ as follows.

$$T_0(\varepsilon)\ \triangleq\ \inf\left\{t\in\mathbb{N}:\left(\frac{\left(\sqrt{2\mathrm{e}\log s}+32\right)}{\sqrt{\frac{\rho}{4}s\varepsilon}}\right)\leq\frac{1}{16}\left(\frac{\Delta_{\min}(\varepsilon)}{2K\mathcal{L}}\right)^{\frac{1}{q}}\quad\forall s\geq t\right\}\ . \tag{268}$$

Hence, $\forall t\geq T_0(\varepsilon)$ we have

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_{1,t}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right)\ =\ 0\ . \tag{269}$$

Using a similar line of arguments, we may readily show that for all $t\geq T_0(\varepsilon)$,

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\bar{\mathcal{E}}_{2,t}\left(\frac{1}{2}\Delta_{\min}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right)\ =\ 0\ . \tag{270}$$

Combining (269) and (270) we infer that for all $t\geq T_0(\varepsilon)$,

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\mathbf{a}^{\mathrm{U}}_t\notin\mathrm{OPT}_\varepsilon\mid\mathcal{E}_{0,t}\right)\ =\ 0\ , \tag{271}$$

which implies that

$$\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\left(\widetilde{\mathrm{OPT}}_{\varepsilon,t}\neq\mathrm{OPT}_\varepsilon\mid\mathcal{E}_{0,t}\right)\ =\ 0\ . \tag{272}$$

We are now ready to define the low probability event $\mathcal{E}_{3,T}$. Let us define

$$\mathcal{E}_{3,T}\triangleq\left\{\exists t\in[T_0(\varepsilon),T]:\widetilde{\mathrm{OPT}}_{\varepsilon,t}\neq\mathrm{OPT}_\varepsilon\right\}. \tag{273}$$

Accordingly, we have the following lemma.

Lemma 14 *For event $\mathcal{E}_{3,T}$, we have*

$$\mathbb{P}_{\boldsymbol{\nu}}^{\mathrm{U}}\left(\mathcal{E}_{3,T}\right) \leq T\left(\left(\frac{2}{T^2}+1\right)^{K}-1\right) \ . \tag{274}$$

Proof: Note that

$$\mathbb{P}_{\boldsymbol{\nu}}^{\mathrm{U}}\Big(\exists t > T_0(\varepsilon) : \widetilde{\mathrm{OPT}}_{\varepsilon,t} \neq \mathrm{OPT}_{\varepsilon}\Big) \leq \sum_{t=T_0(\varepsilon)+1}^{T} \mathbb{P}\left(\widetilde{\mathrm{OPT}}_{\varepsilon,t} \neq \mathrm{OPT}_{\varepsilon}\right) \tag{275}$$

$$\overset{(254)}{\leq} \sum_{t=T_0(\varepsilon)+1}^{T}\left(\left(\frac{2}{T^2}+1\right)^{K}-1\right) + \mathbb{P}_{\boldsymbol{\nu}}^{\mathrm{U}}\Big(\mathbf{a}_t^{\mathrm{U}} \notin \mathrm{OPT}_{\varepsilon} \mid \mathcal{E}_{o,t}\Big) \tag{276}$$

$$\overset{(271)}{\leq} T\left(\left(\frac{2}{T^2}+1\right)^{K}-1\right) \ . \tag{277}$$

■

First, thanks to the discretization, all arms are sampled at least once, according to under-sampling. Let us assume that after the explicit exploration, an arm is never under sampled.

$$\tau_t(i) = \frac{\rho}{4} T\epsilon \tag{278}$$

If an arm is not under sampled that would mean

$$\tau_t(i) \geq t a_t^U(i) \ . \tag{279}$$

$$\frac{\rho}{4} T\epsilon t^{-1} \geq a_t^U(i) \tag{280}$$

$$\geq \frac{\varepsilon}{2} \tag{281}$$

where

- (280) follows from (278) and (279);
- and (281) follows from the definition of discretization.

$$\frac{\rho}{2} \geq \frac{t}{T} \ . \tag{282}$$

For $t > \frac{\rho T}{2}$, this inequality does not hold. Therefore, if an arm is not sampled after the explicit exploration, it becomes under sampled at the latest at the time instant $\frac{\rho T}{2}$.

Next, we will show that under the event $\bar{\mathcal{E}}_{3,T}$, i.e., when the RS-UCB-M algorithm only selects a mixture distribution from the set of optimal mixtures, the RS-UCB-M arm selection routine, using the *under-sampling* procedure, eventually converges to an optimal mixture distribution. This is captured in the following lemma. Prior to stating the lemma, note that under the event $\bar{\mathcal{E}}_{3,T}$, the set of estimated mixing coefficients is the same in each iteration between $T_0(\varepsilon)$ and T. Hence, the RS-UCB-M algorithm aims to track *one* of these optimal mixing coefficients, which we denote by $\mathbf{a}^\star \in \mathrm{OPT}_{\varepsilon}$.

Lemma 15 *Under the event $\bar{\mathcal{E}}_{3,T}$, there exists a time instant $T(\varepsilon) < +\infty$, and $\mathbf{a}^\star \in \mathrm{OPT}_{\varepsilon}$ such that we have*

$$\left|\frac{\tau_t^{\mathrm{U}}(i)}{t} - a^\star(i)\right| < \frac{K}{t} \quad \forall t \geq T(\varepsilon) \ . \tag{283}$$

Proof: We begin by by defining a set of *over-sampled* arms as follows. We define the set $\mathcal{O}_t$ as

$$\mathcal{O}_t \triangleq \left\{ i \in [K] : \frac{\tau_t^{\mathrm{U}}(i)}{t} > a^\star(i) + \frac{1}{t} \right\} \ . \tag{284}$$

We will leverage Lemma 16, in which we show that there exists a finite time instant, which we call $T(\varepsilon)$, such

that for all $t > T(\varepsilon)$ the set of over-sampled arms is *empty*. Specifically, leveraging Lemma 16 we obtain

$$\frac{\tau_t^{\mathrm{U}}(i)}{t} - a^\star(i) \;\leq\; \frac{1}{t} \quad \forall t \geq T(\varepsilon)\;, \tag{285}$$

which also implies that

$$\frac{\tau_t^{\mathrm{U}}(i)}{t} - a^\star(i) \;<\; \frac{K}{t} \quad \forall t \geq T(\varepsilon)\;. \tag{286}$$

Next, let us assume that there exists some $j \in [K]$ such that

$$\frac{\tau_t^{\mathrm{U}}(j)}{t} < a^\star(j) - \frac{K}{t}\;. \tag{287}$$

In this case, we have

$$\sum_{i\in[K]} \frac{\tau_t^{\mathrm{U}}(i)}{t} = \sum_{i\neq j} \frac{\tau_t^{\mathrm{U}}(i)}{t} + \frac{\tau_t^{\mathrm{U}}(j)}{t} \tag{288}$$

$$\overset{(286)}{\leq} \sum_{i\neq j} a^\star(i) + \frac{K-1}{t} + a^\star(j) - \frac{K}{t} \tag{289}$$

$$= 1 - \frac{1}{t}\;, \tag{290}$$

which is a contradiction, since we should have

$$\sum_{i\in[K]} \frac{\tau_t^{\mathrm{U}}(i)}{t} \;=\; 1\;. \tag{291}$$

Hence, we can conclude that

$$\left|\frac{\tau_t^{\mathrm{U}}(i)}{t} - a^\star(i)\right| < \frac{K}{t} \quad \forall t \geq T(\varepsilon)\;. \tag{292}$$

■

Lemma 16 *Under the event $\bar{\mathcal{E}}_{3,T}$, there exists a finite time $T(\varepsilon) < +\infty$ such that for every $t > T(\varepsilon)$, we have $\mathcal{O}_t = \emptyset$.*

Proof: For any time $t > T_0(\varepsilon)$, and some arm $j \notin \mathcal{O}_t$, we have

$$\frac{\tau_t^{\mathrm{U}}(j)}{t} \;\leq\; a^\star(j) + \frac{1}{t}\;. \tag{293}$$

Based on this we have the following two regimes of the arm selection fraction for arm j.

Case (a): ($\tau_t^{\mathrm{U}}(j)/t \leq \bar{\alpha}_j^\star$). Since the RS-UCB-M algorithm only samples under-sampled arms, it is probable that the arm j is sampled at time $t+1$. In that case, we would have

$$\frac{\tau_{t+1}^{\mathrm{UCB}}(j)}{t+1} \;\leq\; \frac{\tau_t^{\mathrm{U}}(j)+1}{t+1} \tag{294}$$

$$=\; \frac{\tau_t^{\mathrm{U}}(j)}{t+1} + \frac{1}{t+1} \tag{295}$$

$$\leq\; \frac{a^\star(j)t}{t+1} + \frac{1}{t+1} \tag{296}$$

$$\leq\; a^\star(j) + \frac{1}{t+1}\;, \tag{297}$$

where (296) follows from the fact that $\tau_t^{\mathrm{U}}(j) \leq t\bar{\alpha}_j^\star$. Clearly, we see that the arm j does not enter the set of over-sampled arms $\mathcal{O}_t$ in the subsequent round $t+1$.

Case (b): ($\bar{\alpha}_j^\star \leq \tau_t^{\mathrm{U}}(j)/t \leq \bar{\alpha}_j^\star + 1/t$). In this case, there always exists at least one arm $k \in [K]$ which satisfies that $\tau_t^{\mathrm{U}}(k)/t \leq a_k^\star$. This implies that the arm j is not sampled by the RS-UCB-M arm selection rule, since it is not the most under-sampled arm. In this case, we have

$$\frac{\tau_{t+1}^{\mathrm{UCB}}(j)}{t+1} \;=\; \frac{\tau_t^{\mathrm{U}}(j)}{t+1} \tag{298}$$

$$= \frac{\tau_t^{\mathrm{U}}(j)}{t}\frac{t}{t+1} \tag{299}$$

$$\leq \left(a^\star(j)+\frac{1}{t}\right)\frac{t}{t+1} \tag{300}$$

$$= a^\star(j)\frac{t}{t+1}+\frac{1}{t+1} \tag{301}$$

$$\leq a^\star(j)+\frac{1}{t+1}\ , \tag{302}$$

where (300) from the definition of case (b). Again, we have concluded that $j \notin \mathcal{O}_{t+1}$. Combining (297) and (302), we conclude that none of the arms which are not already contained in the set $\mathcal{O}_t$ ever enters this set. Hence, all we are left to show is the existence of the time instant after which the set $\mathcal{O}_t$ becomes an empty set. Evidently, since the RS-UCB-M algorithm never samples from the set of over-sampled arms $\mathcal{O}_t$, we will derive an upper bound m such that after $t > T_0(\varepsilon)+m$, all arms leave the set $\mathcal{O}_t$. Notably, for any arm $i \in \mathcal{O}_t$, it holds that

$$\frac{\tau_t^{\mathrm{U}}(i)}{t} \ > \ a^\star(i)+\frac{1}{t}\ . \tag{303}$$

Let m be such that

$$\frac{\tau^{\mathrm{U}}_{T_0(\varepsilon)+m}(i)}{T_0(\varepsilon)+m} \ \leq \ a^\star(i)+\frac{1}{T_0(\varepsilon)+m}\ , \tag{304}$$

which implies that the arm i has left the set $\mathcal{O}_{T(\varepsilon)+m}$. Furthermore, since arm i is never sampled between times instants $T(\varepsilon)$ and $T(\varepsilon)+m$ (as it belongs to the over-sampled set), (304) can be equivalently written as

$$\frac{\tau^{\mathrm{U}}_{T_0(\varepsilon)}(i)}{T_0(\varepsilon)+m} \ \leq \ a^\star(i)+\frac{1}{T_0(\varepsilon)+m}\ . \tag{305}$$

Next, noting that for any arm $i \in [K]$, we have the upper bound $\tau_t^{\mathrm{U}}(i) \leq t$ on the number of times i is chosen up to time t, and that $a^\star(i) \geq \varepsilon/2$ for every $i \in [K]$, we have the following choice for m.

$$m \ = \ \left(\frac{2}{\varepsilon}-1\right)T_0(\varepsilon)-\frac{2}{\varepsilon}\ . \tag{306}$$

The proof is completed by defining

$$T(\varepsilon) \triangleq T_0(\varepsilon)+m\ . \tag{307}$$

■

Next, from (244), we have

$$B_{22}(T) \ \leq \ \mathcal{L}W^q\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\sum_{i\in[K]}\left|a_T^{\mathrm{U}}(i)-\frac{\tau_T(i)}{T}\right|^q \ \middle|\ \mathbb{F}_i \in \mathcal{C}_T(i)\ \forall\, i\in[K]\right] \tag{308}$$

$$\leq \ \frac{1}{\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathbb{F}_i \in \mathcal{C}_T(i)\ \forall\, i\in[K]\Big)}\mathcal{L}W^q\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\sum_{i\in[K]}\left|a_T^{\mathrm{U}}(i)-\frac{\tau_T^{\mathrm{U}}(i)}{T}\right|^q\right] \tag{309}$$

$$\leq \ \frac{1}{2-\left(\frac{1}{T^2}+1\right)^K}\mathcal{L}W^q\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\sum_{i\in[K]}\left|a_T^{\mathrm{U}}(i)-\frac{\tau_T^{\mathrm{U}}(i)}{T}\right|^q\right]\ , \tag{310}$$

where,

(i) (309) follows from the fact that $\mathbb{E}[A|B] = (\mathbb{E}[A]-\mathbb{E}[A|\bar{B}]\cdot\mathbb{P}(\bar{B}))/\mathbb{P}(B) \leq \mathbb{E}[A]/\mathbb{P}(B)$, and,

(ii) (310) follows from (229).

Furthermore, for all $T \geq 3$ and $K \geq 2$, it can be readily verified that $2-(2/T^2+1)^K > 1/2$, which implies that

$$B_{22}(T) \overset{(310)}{\leq} 2\mathcal{L}W^q\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\sum_{i\in[K]}\left|a_T^{\mathrm{U}}-\frac{\tau_T^{\mathrm{U}}(i)}{T}\right|^q\right] \tag{311}$$

$$= 2\mathcal{L}W^q\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\sum_{i\in[K]}\left|a^{\mathrm{U}}_T - \frac{\tau^{\mathrm{U}}_T(i)}{T}\right|^q \,\middle|\, \mathcal{E}_{3,T}\right]\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\mathcal{E}_{3,T}\Big) + 2\mathcal{L}W^q\mathbb{E}^{\mathrm{U}}_{\boldsymbol{\nu}}\left[\sum_{i\in[K]}\left|a^{\mathrm{U}}_T - \frac{\tau^{\mathrm{U}}_T(i)}{T}\right|^q \,\middle|\, \bar{\mathcal{E}}_{3,T}\right]\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}\Big(\bar{\mathcal{E}}_{3,T}\Big) \tag{312}$$

$$\leq 2\mathcal{L}KW^q\left(\mathbb{P}^{\mathrm{U}}_{\boldsymbol{\nu}}(\mathcal{E}_{3,T}) + K\left(\frac{K}{T}\right)^q\right) \tag{313}$$

$$\leq 2\mathcal{L}KW^q\left(T\left(\left(\frac{2}{T^2}+1\right)^K - 1\right) + \left(\frac{K}{T}\right)^q\right)\,, \tag{314}$$

where (313) follows from Lemma 16 and (314) follows from Lemma 14.

F.4 Proof of Theorem 3 for RS-UCB-M

Finally, we add up all the terms to find an upper bound on the discrete regret. We have

$$\bar{\mathfrak{R}}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) \;=\; B_1(T) + B_{21}(T) + B_{22}(T) \tag{315}$$

$$\leq\; \frac{1}{T}\Bigg[BT\left(\left(\frac{2}{T^2}+1\right)^K - 1\right) \tag{316}$$

$$+ \mathcal{L}K\left(\frac{32}{\sqrt{\rho}}\right)^q T^{1-\frac{q}{2}}\left(\sqrt{2\mathrm{e}\log T}+32\right)^q$$

$$+ 2\mathcal{L}W^q\left(KT^2\left(\left(\frac{2}{T^2}+1\right)^K - 1\right) + K^{1+q}T^{1-q}\right)\Bigg]\,. \tag{317}$$

Leveraging the upper bounds on discretization error and the discrete regret, we state an upper bound on the regret of the RS-UCB-M algorithm.

$$\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) = \Delta(\varepsilon) + \bar{\mathfrak{R}}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) \tag{318}$$

$$\leq \Delta(\varepsilon) + \frac{1}{T}\Bigg[BT\left(\left(\frac{2}{T^2}+1\right)^K - 1\right)$$

$$+ 2\mathcal{L}W^q\left(KT^2\left(\left(\frac{2}{T^2}+1\right)^K - 1\right) + K^{1+q}T^{1-q}\right)$$

$$+ \underbrace{\mathcal{L}K\left(\frac{64}{\sqrt{\rho\varepsilon}}\right)^q T^{1-\frac{q}{2}}\left(\sqrt{2\mathrm{e}\log T}+32\right)^q\Bigg]}_{B_3(T)}\,. \tag{319}$$

$$\tag{320}$$

$B_3(T)$ is the dominating term for any $T > \mathrm{e}^K$. Hence, we can simplify the upper bound as follows.

$$\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) \leq \Delta(\varepsilon) + (B + \mathcal{L}(W^q+1))\Big(64\varepsilon^{-\frac{1}{2}}\rho^{-\frac{1}{2}}T^{-\frac{1}{2}}\left(\sqrt{2\mathrm{e}\log T}+32\right)\Big)^q\,. \tag{321}$$

F.5 Proof of Theorem 4 for RS-UCB-M

Let us set

$$\varepsilon = \Theta\left(\left(K^{\frac{2}{q}}\log T/T\right)^{\kappa}\right)\,, \tag{322}$$

where $\kappa = \frac{1}{\frac{2\beta}{q}+2}$. Assuming $\Delta_{\min}(\varepsilon) = \Omega(\varepsilon^{\beta})$, it can be readily verified that this choice of the discretization level ε satisfies the condition in (307). Hence, for discrete regret, from (315) we have

$$\bar{\mathfrak{R}}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) = O\left((K^{\frac{2}{q}}\log T/T)^{(1-\kappa)\frac{q}{2}}\right)\,. \tag{323}$$

Furthermore, from Lemma (12), we have

$$\Delta(\varepsilon) \leq \mathcal{L}(KW)^r\left(\frac{\varepsilon}{2}\right)^r\,. \tag{324}$$

Hence, $\Delta(\varepsilon) = O(\varepsilon^r)$. Consequently, we have

$$\Delta(\varepsilon) = O\left(K^{r\left(1+\frac{2\kappa}{q}\right)} (\log T/T)^{r\kappa}\right). \tag{325}$$

From (323) and (325) the regret of the RS-UCB-M is bounded from above by

$$\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) = O\left(\max\left\{K^{r\left(1+\frac{2\kappa}{q}\right)} (\log T/T)^{r\kappa} \ , \ (K^{\frac{2}{q}} \log T/T)^{(1-\kappa)\frac{q}{2}}\right\}\right) \tag{326}$$

$$= O\left(\max\left\{K^{r\left(1+\frac{2\kappa}{q}\right)}, K^{1-\kappa}\right\} \max\left\{ (\log T/T)^{r\kappa} \ , \ (\log T/T)^{(1-\kappa)\frac{q}{2}}\right\}\right) \ . \tag{327}$$

When $r \leq \beta + q/2$, this bound becomes

$$\mathfrak{R}^{\mathrm{U}}_{\boldsymbol{\nu}}(T) \leq O\left(\max\left\{K^{r\left(1+\frac{2\kappa}{q}\right)}, K^{1-\kappa}\right\} (\log T/T)^{r\kappa}\right) \ . \tag{328}$$

G CE-UCB-M Algorithm and Its Performance

Finally, in this section, we provide the pseudo-code of the CE-UCB-M algorithm and the proof of Theorem 3 which characterize the upper bound on the algorithm's regret.

G.1 CE-UCB-M Algorithm

The detailed steps of the RS-UCB-M algorithm with the sampling rule specified as in (17) are presented in Algorithm 3.

Algorithm 3 Risk-Sensitive Upper Confidence Bound for Mixtures (CE-UCB-M)

1: **Input:** Exploration rate ρ, horizon T
2: Sample each arm $\lceil \rho T\varepsilon/4 \rceil$ times and obtain observation sequences $\mathcal{X}_{\lceil \rho T\varepsilon/4 \rceil}(1), \cdots, \mathcal{X}_{\lceil \rho T\varepsilon/4 \rceil}(K)$
3: **Initialize:** $\tau^{\mathrm{C}}_{K\lceil \rho T\varepsilon/4 \rceil}(i) = \lceil \rho T\varepsilon/4 \rceil \ \forall \ i \in [K]$, arm CDFs $\mathbb{F}^{\mathrm{C}}_{1,\lceil K\rho T\varepsilon/4 \rceil}(1), \cdots, \mathbb{F}^{\mathrm{C}}_{K,K\lceil \rho T\varepsilon/4 \rceil}$, confidence sets $\mathcal{C}_{K\lceil \rho T\varepsilon/4 \rceil}(1), \cdots \mathcal{C}_{K\lceil \rho T\varepsilon/4 \rceil}(K)$ according to (15)
4: **for** $t = K\lceil \rho T\varepsilon/4 \rceil + 1, \cdots, T$ **do**
5: Select an arm A_t specified by (17) and obtain reward X_t
6: Update the empirical CDF $\mathbb{F}^{\mathrm{C}}_{A_t,t}$ according to (8)
7: Compute the optimistic estimate $\mathbf{a}^{\mathrm{C}}_t$ according to (18)
8: **end for**

G.2 Proof of Theorem 3 for CE-UCB-M

The analysis of the CE-UCB-M algorithm closely follows that of the RS-UCB-M algorithm with some minute differences. We will briefly state the steps in the RS-UCB-M analysis, and in the process, highlight the key distinctions in the analysis of the CE-UCB-M algorithm. Henceforth, we will use $\pi = \mathrm{C}$ to denote the policy CE-UCB-M. Similarly to the RS-UCB-M analysis and (19), we decompose the regret into a discretization error component $\Delta(\varepsilon)$, and the discrete regret $\bar{\mathfrak{R}}^{\mathrm{C}}_{\boldsymbol{\nu}}(T)$. Leveraging Lemma 12, it may be readily verified that

$$\Delta(\varepsilon) \leq \mathcal{L}(KW)^r \left(\frac{\varepsilon}{2}\right)^r \ . \tag{329}$$

Next, we digress to upper bounding the discrete regret $\bar{\mathbb{R}}^{\mathrm{C}}_{\boldsymbol{\nu}}(T)$. We have

$$\bar{\mathfrak{R}}^{\mathrm{C}}_{\boldsymbol{\nu}}(T) = U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - \mathbb{E}^{\mathrm{C}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{C}}_t(i)}{T}\mathbb{F}_i\right)\right] \tag{330}$$

$$\leq \underbrace{\sum_{\mathcal{S}\subseteq[K]:\mathcal{S}\neq\emptyset} \mathbb{E}^{\mathrm{C}}_{\boldsymbol{\nu}}\left[U_h\left(\sum_{i\in[K]} a^\star(i)\mathbb{F}_i\right) - U_h\left(\sum_{i\in[K]} \frac{\tau^{\mathrm{C}}_t(i)}{T}\mathbb{F}_i\right) \mathbb{1}\{\mathbb{F}_i \notin \mathcal{C}_T(i) : i \in \mathcal{S}\}\right]}_{\triangleq C_1(T)} \tag{331}$$

$$+\underbrace{\mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\left[U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\mathbb{1}\{\mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\}\right]}_{\triangleq C_2(T)}, \tag{332}$$

where $\mathcal{C}_t(i)$ has the same definition as in (15) for every $i\in[K]$. Furthermore, we have already upper bounded the term $A_1(T)$ in Appendix F (equation (230)), which is given by

$$C_1(T)\ \leq\ B\left(\left(\frac{1}{T^2}+1\right)^K-1\right)\ . \tag{333}$$

We now resort to upper bounding the term $C_2(T)$. Note that

$$\begin{aligned}
C_2(T)\ &=\ \mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\left[\left(U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\right)\mathbb{1}\{\mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\}\right] &(334)\\
&\leq\ \mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\left[U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\ \middle|\ \mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\right] &(335)\\
&\leq \mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\Bigg[U_h\left(\sum_{i\in[K]}a^\star(i)\mathbb{F}_{i,T}^{\mathrm{C}}\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\\
&\qquad+\mathcal{L}\sum_{i\in[K]}a^\star(i)^q\left(16\,\frac{\sqrt{2\mathrm{e}\log T}+32}{\sqrt{\tau_t^{\mathrm{C}}(i)}}\right)^q\ \Bigg|\ \mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\Bigg] &(336)\\
&\leq \mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\Bigg[U_h\left(\sum_{i\in[K]}a_T^{\mathrm{C}}(i)\mathbb{F}_{i,T}^{\mathrm{C}}\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\\
&\qquad+\mathcal{L}\sum_{i\in[K]}a_T^{\mathrm{C}}(i)^q\left(16\,\frac{\sqrt{2\mathrm{e}\log T}+32}{\sqrt{\tau_t^{\mathrm{C}}(i)}}\right)^q\ \Bigg|\ \mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\Bigg] &(337)\\
&\leq \underbrace{\mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\left[U_h\left(\sum_{i\in[K]}a_T^{\mathrm{C}}(i)\mathbb{F}_{i,T}^{\mathrm{C}}\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\ \middle|\ \mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\right]}_{\triangleq\, C_3(T)}\\
&\qquad+\mathcal{L}K\left(32\,\frac{\sqrt{2\mathrm{e}\log T}+32}{\sqrt{\rho T\varepsilon}}\right)^q\ , &(338)
\end{aligned}$$

where,

- (336) follows from Hölder defined in Definition 1 and the conditioning on the fact that $\mathbb{F}_i\in\mathcal{C}_T(i)$ for every $i\in[K]$;
- (337) follows from the upper confidence bound in (18);
- and (338) follows from the explicit exploration phase of the RS-UCB-M algorithm.

Furthermore, note that $C_3(T)$ can be expanded as

$$\begin{aligned}
C_3(T)\ &=\ \mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\Bigg[U_h\left(\sum_{i\in[K]}a_T^{\mathrm{C}}(i)\mathbb{F}_{i,T}^{\mathrm{C}}\right)-U_h\left(\sum_{i\in[K]}\alpha_T(i)\mathbb{F}_i\right)\\
&\qquad+U_h\left(\sum_{i\in[K]}a_T^{\mathrm{C}}(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\ \Bigg|\ \mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\Bigg] &(339)\\
&\leq\ \underbrace{\mathbb{E}_{\boldsymbol{\nu}}^{\mathrm{C}}\left[U_h\left(\sum_{i\in[K]}a_T^{\mathrm{C}}(i)\mathbb{F}_i\right)-U_h\left(\sum_{i\in[K]}\frac{\tau_t^{\mathrm{C}}(i)}{T}\mathbb{F}_i\right)\ \middle|\ \mathbb{F}_i\in\mathcal{C}_T(i)\ \forall i\in[K]\right]}_{C_4(T)}
\end{aligned}$$

$$+ \mathcal{L}K \left(32 \, \frac{\sqrt{2e \log T} + 32}{\sqrt{\rho T \varepsilon}} \right)^q . \tag{340}$$

Finally, note that the term $C_4(T)$ is similar to the term $B_{22}(T)$ in the RS-UCB-M analysis in Appendix F, and can be handled in the exact same way. Recall the finite time instant $T(\varepsilon) = T_0(\varepsilon) + m$, where m has been defined in (306) and $T_0(\varepsilon)$ has been defined in (268). For all $T > T(\varepsilon)$, we have the following bound on $C_4(T)$.

$$C_4(T) \; \leq \; 2\mathcal{L}KW^q \left(T \left(\left(\frac{1}{T^2} + 1 \right)^K - 1 \right) + \left(\frac{K}{T} \right)^q \right) . \tag{341}$$

Aggregating $C_1(T) - -C_4(T)$, we have that for all $T > T(\varepsilon)$,

$$\begin{aligned} \mathfrak{R}^{\mathrm{C}}_{\boldsymbol{\nu}}(T) \; \leq \; \Delta(\varepsilon) + \frac{1}{T} \Bigg[& \underbrace{2T^{1-q/2}\mathcal{L}K \left(32 \, \frac{\sqrt{2e \log T} + 32}{\sqrt{\rho \varepsilon}} \right)^q}_{\triangleq \, C_5(T)} \\ & + 2\mathcal{L}KW^q \left(T^2 \left(\left(\frac{1}{T^2} + 1 \right)^K - 1 \right) + K^q T^{1-q} \right) \\ & + BT \left(\left(\frac{1}{T^2} + 1 \right)^K - 1 \right) \Bigg] . \end{aligned} \tag{342}$$

Furthermore, similar to the RS-UCB-M analysis, it can be readily verified that $C_5(T)$ is the dominating term for any $T > \mathrm{e}^K$. Hence, we can simplify the upper bound as follows.

$$\mathfrak{R}^{\mathrm{C}}_{\boldsymbol{\nu}}(T) \; \leq \; \Delta(\varepsilon) + 5\mathcal{L}KW \left(32 \varepsilon^{-\frac{1}{2}} \rho^{-\frac{1}{2}} T^{-\frac{1}{2}} \left(\sqrt{2e \log T} + 32 \right) \right)^q \tag{343}$$

where (343) follows from $W > 1$ for sub-Gaussian distributions as shown in Appendix C and $q \in (0, 1]$.

G.3 Proof of Theorem 4 for CE-UCB-M

Let us set

$$\varepsilon = \Theta \left(\left(K^{\frac{2}{q}} \log T / T \right)^{\kappa} \right) , \tag{344}$$

where $\kappa = \frac{1}{\frac{2\beta}{q} + 2}$. Assuming $\Delta_{\min}(\varepsilon) = \Omega(\varepsilon^{\beta})$, it can be readily verified that this choice of the discretization level ε satisfies the condition in (268). Hence, for discrete regret, from (342) we have

$$\bar{\mathfrak{R}}^{\mathrm{C}}_{\boldsymbol{\nu}}(T) = O \left((K^{\frac{2}{q}} \log T / T)^{(1-\kappa)\frac{q}{2}} \right) . \tag{345}$$

Furthermore, from Lemma (12), we have

$$\Delta(\varepsilon) \leq \mathcal{L}(KW)^r \left(\frac{\varepsilon}{2} \right)^r . \tag{346}$$

Hence, $\Delta(\varepsilon) = O(\varepsilon^r)$. Consequently, we have

$$\Delta(\varepsilon) = O \left(K^{r\left(1 + \frac{2\kappa}{q}\right)} \left(\log T / T \right)^{r\kappa} \right) . \tag{347}$$

From (345) and (347) the regret of the CE-UCB-M is bounded from above by

$$\mathfrak{R}^{\mathrm{C}}_{\boldsymbol{\nu}}(T) = O \left(\max \left\{ K^{r\left(1 + \frac{2\kappa}{q}\right)} \left(\log T / T \right)^{r\kappa} \; , \; (K^{\frac{2}{q}} \log T / T)^{(1-\kappa)\frac{q}{2}} \right\} \right) \tag{348}$$

$$= O \left(\max \left\{ K^{r\left(1 + \frac{2\kappa}{q}\right)}, K^{1-\kappa} \right\} \max \left\{ \left(\log T / T \right)^{r\kappa} \; , \; (\log T / T)^{(1-\kappa)\frac{q}{2}} \right\} \right) . \tag{349}$$

When $r \leq \beta + q/2$, this bound becomes

$$\mathfrak{R}^{\mathrm{C}}_{\boldsymbol{\nu}}(T) \leq O \left(\max \left\{ K^{r\left(1 + \frac{2\kappa}{q}\right)}, K^{1-\kappa} \right\} (\log T / T)^{r\kappa} \right) . \tag{350}$$

H Additional Experiments

In this section, we provide figures mentioned in the main paper, additional details of the experiments, some additional comparisons, and some details about computing resources.

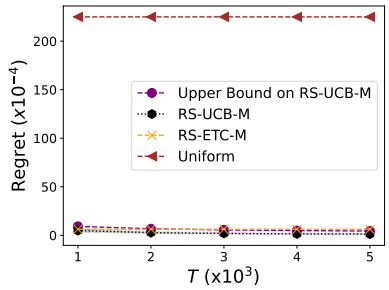
Computing Resources. All experiments are conducted on Mac Mini 2023 equipped with 24 Gigabytes of RAM, and 2 CPU cores have been used for each experiment.

General Setting. We have experimented using Gini deviation. For a Bernoulli distributed $K-$arm bandit instance, we have run 1000 independent trials for each configuration.

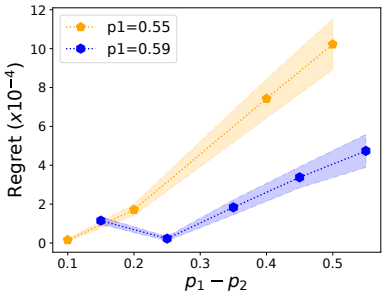
Regret versus horizon. The configuration of the experiment considered in Figure 2a is the following. We chose the number of arms as $K = 2$, the exploration coefficient as $\rho = 0.1$, the arm distributions as Bern(0.4) and Bern(0.9). Figure 2a shows that uniform sampling is not regret-efficient in the premise of mixtures.

Regret versus gaps. In Figure 2b, for the $2-$armed bandit instance, we examine the regret of RS-UCB-M for varying mean values of the arms. Specifically, we set the mean of the first arm as $p_1 = 0.55$ and vary the mean values of the second arm p_2. The exploration coefficient is set to $\rho = 0.1$ and time horizon is set to $T = 10^5$. From 2b, we observe that with increasing difference between the means of the two arms, the regret increases. Additionally, we run a similar experiment for $p_1 = 0.59$ which demonstrates similar results.

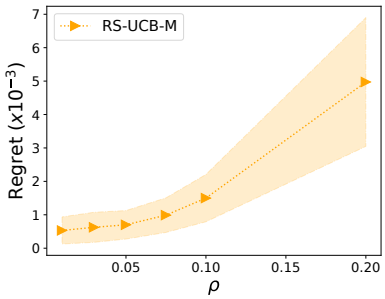
Regret versus exploration coefficient. In Figure 2c, we show the behaviour of the regret of the RS-UCB-M algorithm for increasing values of the exploration coefficient ρ. We fix the horizon at $T = 7.5 \cdot 10^4$ and the number of arms is set to $K = 3$. We observe that when the exploration coefficient ρ increases, there is an increase in the regret.



(a) Regret for algorithms RS-UCB-M, RS-ETC-M and Uniform



(b) Regret of RS-UCB-M algorithm for different values of Bernoulli pmf differences



(c) Regret of RS-UCB-M algorithm for different values of exploration coefficient ρ

Figure 2: Regrets of algorithms for different settings.