A Unifying Theory of Thompson Sampling for Continuous Risk-Averse Bandits

Joel Q. L. Chang¹, Vincent Y. F. Tan¹, ²

¹Department of Mathematics, National University of Singapore
²Department of Electrical and Computer Engineering, National University of Singapore
joel.chang@u.nus.edu, vtan@nus.edu.sg

Abstract

This paper unifies the design and the analysis of risk-averse Thompson sampling algorithms for the multi-armed bandit problem for a class of risk functionals $\rho$ that are continuous and dominant. We prove generalised concentration bounds for these continuous and dominant risk functionals and show that a wide class of popular risk functionals belong to this class. Using our newly developed analytical toolkits, we analyse the algorithm $\rho$-MTS (for multinomial distributions) and prove that they admit asymptotically optimal regret bounds of risk-averse algorithms under the CVaR, proportional hazard, and other ubiquitous risk measures. More generally, we prove the asymptotic optimality of $\rho$-MTS for Bernoulli distributions for a class of risk measures known as empirical distribution performance measures (EDPMs); this includes the well-known mean-variance. Numerical simulations show that the regret bounds incurred by our algorithms are reasonably tight vis-à-vis algorithm-independent lower bounds.

Introduction

Consider a $K$-armed multi-armed bandit (MAB) with unknown distributions $\nu = (\nu_k)_{k \in [K]}$ called arms and a time horizon $n$. At each time step $t \in [n]$, a learner chooses an arm $A_t \in [K]$ and obtains a random reward $X_{A_t}$ from the corresponding distribution $\nu_{A_t}$. In the vanilla MAB setting, the learner aims to maximise their expected total reward after $n$ selections, requiring a strategic balance of exploration and exploitation of the arms. Much work has been developed in this field for L/UCB-based algorithms, and in recent developments, more Thompson sampling-based algorithms have been designed and proven to attain the theoretical asymptotic lower bounds that outperform their L/UCB-based counterparts. However, many real-world settings include the presence of risk, which precludes the adoption of the mean-maximisation objective. Risk-averse bandits address this issue for bandit models by replacing the expected value by some measure of risk.

Recent work has incorporated risk into the analysis, with different works working with different risk measures that satisfy various properties. In the existing literature, the more popular risk measures being considered are mean-variance (Sani, Lazaric, and Munos 2012; Zhu and Tan 2020) and conditional value-at-risk (CVaR) (Tamkin et al. 2019; Kha-jonchotpanya, Xue, and Rujeerapaiboon 2021; Baudry et al. 2021; Chang, Zhu, and Tan 2021). In particular, CVaR is an instance of a general class of risk functionals, called coherent risk functionals (Artzner et al. 1999). Huang et al. (2021) observed that for nonnegative rewards, coherent risk functionals are subsumed in broader class of functionals called distorted risk functionals. Common distorted risk functionals, such as the expected value and CVaR, satisfy theoretically convenient continuity properties.

However, not much work has been done to unify these various risk-averse algorithms to elucidate the common machinery that underlie them. In this paper, we provide one way to unify these risk-averse Thompson sampling algorithms through continuous and dominant risk functionals, which we denote by $\rho$. We design two Thompson sampling-based algorithms—$\rho$-MTS and $\rho$-NPTS—to solve the modified MABs, provably achieving asymptotic optimality for $\rho$-MTS under a variety of risk functionals and empirically doing so for $\rho$-NPTS. Therefore, we unify much of the progress made in analysing Thompson sampling-based solutions to risk-averse MABs.

Related Work

Thompson (1933) proposed the first Bayesian algorithm for MABs known as Thompson sampling. Lai and Robbins (1985) proved a lower bound on the regret for any instance-dependent bandit algorithm for the vanilla MAB. Kaufmann, Korda, and Munos (2012); Agrawal and Goyal (2012) analysed the Thompson sampling algorithm to solve the $K$-armed MAB for Bernoulli and Gaussian reward distributions respectively, and proved the asymptotic optimality in the Bernoulli setting relative to the lower bound given by Lai and Robbins (1985). Granmo (2008) proposed the Bayesian learning automaton that is self-correcting and converges to only pulling the optimal arm with probability 1. Riou and Honda (2020) designed and proved the asymptotic optimality of Thompson sampling on bandits which firstly follow multinomial distributions, followed by general bandits that are bounded in $[0, 1]$ by discretising $[0, 1]$ and using suitable approximations on each sub-interval.

Many variants of the MAB which factor risk have been considered. One popular risk measure is mean-variance. Sani, Lazaric, and Munos (2012) proposed the first U/LCB-
based algorithm called MV-UCB to solve the mean-variance MAB problem. Vakili and Zhao (2015) tightened the regret analysis of MV-UCB, establishing the order optimality of MV-UCB. Zhu and Tan (2020) designed and analysed the first risk-averse mean-variance bandits based on Thompson sampling which follow Gaussian distributions, providing novel tail upper bounds and a unifying framework to consider Thompson samples with various means and variances. Du et al. (2021) further generalised this problem, considering continuous mean-covariance linear bandits, which specialises into the stochastic mean-variance MAB in the 1-dimensional setting.

Another popular risk measure is Conditional Value-at-Risk (abbreviated as CVaR), Galichet, Sebag, and Teytaud (2013) designed the LUCB-based Multi-Armed Risk-Aware Bandit (MARAB) algorithm to solve the CVaR MAB problem. Chang, Zhu, and Tan (2021) and Baudry et al. (2021) contemporaneously designed and analysed Thompson sampling algorithms for the risk measure CVaR. The former proved near-asymptotically optimal regret bounds for Gaussian bandits, and the latter proved asymptotically optimal regret bounds for rewards in [0, 1] by judiciously analysing the compact spaces induced by CVaR and designing and proving new concentration bounds.

Other generalised frameworks of risk functionals have also been studied. Wang (1996) studied distorted risk functionals that generalised the expectation and CVaR, characterising the risk functionals by their distortion functions that are non-decreasing on [0, 1]. Cassel, Mannor, and Zeevi (2018) analysed empirical distribution performance measures (EDPMs), which are by definition continuous on the (Banach) space of bounded random variables under the uniform norm. In Table 1 therein, these EDPMs provide the interface for many instances of other popular risk functionals, such as second moment, entropic risk, and Sharpe ratio. Lee, Park, and Shin (2020) studied risk-sensitive learning schemes by rejuvenating the notion of optimized certainty equivalents (OCE), which subsumes common risk functionals like expectation, entropic risk, mean-variance, and CVaR. Huang et al. (2021) defined Lipschitz risk functionals which subsume many of these common risk measures under suitable smoothness assumptions; these include variance, mean-variance, distorted risk functionals, and Cummulative Prospect Theory-inspired (CPT) risk functionals.

Contributions

• We present the key properties that any continuous and dominant risk functional (Definition 2) ρ possesses that are then exploited in the regret analysis of the Thompson sampling algorithms. This provides the theoretical underpinnings for our proposed Thompson sampling-based algorithms to solve any ρ-MAB problem.

• We state and prove new upper and lower tail bounds for ρ on multinomial distributions, generalising and unifying the underlying theory for the upper and lower bounds obtained in Riou and Honda (2020) and Baudry et al. (2021). These new tail bounds generalise the risk functional beyond expected value (Riou and Honda 2020) and CVaR (Baudry et al. 2021), to apply to continuous and dominant risk functionals.

• We also design two Thompson sampling-based algorithms: ρ-MTS for bandits on multinomial distributions and ρ-NPTS for bandits on distributions whose rewards are bounded in any compact subset C ⊆ R. We show that for many continuous and dominant risk functionals ρ, ρ-MTS is asymptotically optimal. Setting ρ to common risk measures, we recover asymptotically optimal algorithms for the respective ρ-MAB problems (Riou and Honda 2020; Zhu and Tan 2020; Baudry et al. 2021), and significantly improve on the regret bounds for Bernoulli-MVTS in Zhu and Tan (2020); see Remark 3.

Preliminaries

Let \( \mathbb{N} \) be the set of positive integers. For any \( M \in \mathbb{N} \), define \( [M] = \{1, \ldots, M\} \) and \( [M]_0 = [M] \cup \{0\} \). For any \( M \in \mathbb{N} \), denote the \( M \)-probability simplex as \( \Delta^M = \{ p \in [0,1]^{M+1} : \sum_{i \in [M]_0} p_i = 1 \} \). For any \( p, q \in \Delta^M \), we denote the \( \ell_\infty \) distance between them as

\[
d_\infty(p,q) := \max_{i \in [M]_0} |p_i - q_i|.
\]

Before formally stating the problem, we need to introduce some measure-theoretic and topological notions which will be essential in the analysis.

Fix a compact subset \( C \subseteq \mathbb{R} \). Then \( (C,|\cdot|) \) is a separable metric space with Borel \( \sigma \)-algebra denoted by \( \mathcal{B}(C) \), constituting the measurable space \( (C,\mathcal{B}(C)) \). For each \( c \in C \), let \( \delta_c := \{ e \in \cdot \} \) denote the Dirac measure at \( e \).

Let \( \mathcal{P} \) denote the collection of probability measures on \( (C,\mathcal{B}(C)) \). Each \( \mu \in \mathcal{P} \) admits a cumulative distribution function (CDF) \( F_\mu = \mu((-\infty, \cdot]) : C \to \{0,1\} \). Hence, on \( \mathcal{P} \), we can define the Kolmogorov-Smirnov metric

\[
D_\infty : (\mu, \eta) \mapsto \sup_{t \in C} |F_\mu(t) - F_\eta(t)|.
\]

We can also define the Lévy-Prokhorov metric

\[
D_L : (\mu, \eta) \mapsto \inf \{ \epsilon > 0 : F_\mu(x - \epsilon) - \epsilon \leq F_\eta(x) \leq F_\mu(x + \epsilon) + \epsilon, \forall x \in \mathbb{R} \}
\]

on \( \mathcal{P} \). Thus, \( (\mathcal{P},d) \) is a metric space in either metric \( d \in \{D_\infty, D_L\} \). For any \( \mu, \eta \in \mathcal{P} \), let \( \text{KL}(\mu, \eta) := \int_C \log(d\mu/d\eta) \ d\mu \) denote the relative entropy or Kullback-Leibler (KL) divergence between \( \mu \) and \( \eta \).

We will now provide three examples of compact metric subspaces \( (C,d) \) of \( (\mathcal{P},d) \) which will utilise in our algorithms and lemmas therein.

Example 1 \((\mathcal{P}_S, D_\infty)\). We first consider \( (\mathcal{P}_S, D_\infty) \)—the set of probability mass functions on a finite alphabet \( S = \{s_0, \ldots, s_M\} \subset C \) under the \( D_\infty \) metric. For each \( p \in \Delta^M \), define \( \mu_p = \sum_{i=0}^M p_i \delta_{s_i} \), and \( \mathcal{D}_S : \Delta^M \to \mathcal{P} \) by \( p \mapsto \mu_p \). Then \( \mathcal{D}_S \) is an imbedding into \( \mathcal{P} \) due to the inequality \( d_\infty(p,q) \leq 2D_\infty(\mathcal{D}_S(p), \mathcal{D}_S(q)) \leq 2Md_\infty(p,q) \). This implies that \( (C,d) := (\mathcal{D}_S(\Delta^M), D_\infty) \) is a compact metric space. For brevity, we denote \( \mathcal{P}_S := \mathcal{D}_S(\Delta^M) \).
In this section, we define continuous risk functionals, which are the risk measures of interest in our Thompson sampling algorithms. We demonstrate that when \( \rho \) is continuous and dominant (see Definition 4), its corresponding \( \rho \)-MTS and \( \rho \)-NPTS algorithms achieve the asymptotically optimal regret bound, the former provably and the latter empirically.

**Definition 2** (Continuous Risk Functional). Let \( \mathcal{P} \) be equipped with the metric \( d \). A risk functional \( \rho \) is said to be **continuous at \( \mu \in \mathcal{P} \)** if for any \( \varepsilon > 0 \), there exists \( \delta > 0 \), which may depend on \( \mu \in \mathcal{P} \), such that

\[
d(\mu, \eta) < \delta \Rightarrow |\rho(\mu) - \rho(\eta)| < \varepsilon.
\]

We say that \( \rho \) is **continuous on a subset \( \mathcal{Q} \subseteq \mathcal{P} \)** if it is continuous at every \( \mu \in \mathcal{Q} \). We say that \( \rho \) is **uniformly continuous on \( \mathcal{Q} \)** if for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) that does not depend on \( \mu \in \mathcal{Q} \), such that (1) holds.

We also remark that for any compact metric subspace \((\mathcal{C}, d)\) of \((\mathcal{P}, d)\) and continuous risk functional \( \rho, \rho|_{\mathcal{C}} \) is uniformly continuous on \((\mathcal{C}, d)\).

Let \((\mathcal{C}, d)\) be any of the three compact metric spaces \((\mathcal{P}_S, D_{\infty}), (\mathcal{P}_D, D_{\infty}), (\mathcal{P}_C(B), D_{\infty})\). For any risk functional \( \rho : \mathcal{P} \to \mathbb{R} \), define

\[
\kappa_{\inf}^\rho(\mu, r) := \inf_{\eta \in \mathcal{C}} \{ \mathbb{K}(\mu, \eta) : \rho(\eta) \geq r \}.
\]

In the case \((\mathcal{C}, d) = (\mathcal{P}_S, D_{\infty})\) for some fixed alphabet \( S \), define \( \sigma_{\rho, S} := \rho \circ \mathcal{D}_S : \Delta^M \to \mathbb{R} \). We note that \( \sigma_{\rho, S} \) is continuous on \( \Delta^M \) if \( \rho \) is continuous on \((\mathcal{P}_S, D_{\infty})\).

By Lemma 18 in Riou and Honda (2020) \( \rho \) being continuous on \((\mathcal{P}_D, D_{\infty})\) implies its continuity on \((\mathcal{P}_C(B), D_{\infty})\), and \( \rho \) being continuous on \((\mathcal{P}_C(B), D_{\infty})\) implies its continuity on \((\mathcal{P}_C(B), D_{\infty})\). This conclusion is consistent with that in Baudry et al. (2021) whose B-CVTS algorithm assumes that the reward distributions are continuous.

**Tail Upper Bound** Risk functionals \( \rho \) that are continuous on \( \mathcal{P}_S \) satisfy a generalization of the tail upper bound developed by Riou and Honda (2020).

**Lemma 1.** Let \( \rho : \mathcal{P} \to \mathbb{R} \) be a risk functional continuous on \((\mathcal{P}_S, D_{\infty})\) for some finite alphabet \( S \) of size \( M + 1 \), and \( r \in \mathbb{R} \). Fix \( \alpha \in \mathbb{N}^{M+1} \), \( n = \sum_{i=0}^{M} \alpha_i \), and \( p = \alpha/n \). Then for any random variable \( L \sim \text{Dir}(\alpha) \),

\[
\mathbb{P}(\sigma_{\rho, S}(L) \geq r) \leq C_1 n^{M/2} \exp(-n \kappa_{\inf}^\rho(\mathbb{D}_S(p), r));
\]

and

\[
\mathbb{P}(\sigma_{\rho, S}(L) \leq r) \leq C_1 n^{M/2} \exp(-n \kappa_{\inf}^\rho(\mathbb{D}_S(p), r)),
\]

where \( C_1 := \Gamma(M + 1)^{-1}(2^\alpha)^{-M/2} e^{1/12} \).

We remark that Lemma 1 generalises the upper bound in Riou and Honda (2020, Lemma 13) to risk functionals that are “sufficiently continuous”.

**Proposition 1.** Let \( \rho : \mathcal{P} \to \mathbb{R} \) be a continuous risk functional. Then the mapping \( \kappa_{\inf}^\rho : \mathcal{P} \times \rho(\mathcal{C}) \to \mathbb{R} \) is lower-semicontinuous in both of its arguments.

**Examples of Continuous Risk Functionals** We provide numerous examples of risk functionals that satisfy the proposed notion of continuity.

---

**Example 2** \((\mathcal{P}, D_{\infty})\). By Halmos (1959, Theorem 1.12), \( \mathcal{P} \) is a compact set in the topology of weak convergence, which is metrized by the Lévy-Prokhorov metric \( D_{\infty} \) on \( \mathcal{P} \). This implies that \((\mathcal{P}, D_{\infty})\) is a compact metric space. Furthermore, by Posner (1975), KL(\( \cdot, \cdot \)) is jointly lower-semicontinuous in both arguments.

**Example 3** \((\mathcal{P}_c(B), D_{\infty})\). This is the set of probability measures whose CDFs have continuous derivatives that are uniformly bounded by \( B \), i.e., \( \mathcal{P}_c(B) := \{ \mu \in \mathcal{P} : F_\mu' \text{ is cts on } C \text{ and sup}_{c \in C} |F_\mu'(c)| \leq B \}. \) By the Arzelà-Ascoli Theorem, \((\mathcal{P}_c(B), D_{\infty})\) is compact and thus as topological spaces \((\mathcal{P}_c(B), D_{\infty}) = (\mathcal{P}_c(B), D_{\infty}) \) is a compact metric space.

---

**Paper Outline**

In the following, we first define continuous and dominant risk functionals, and state some essential properties and crucial concentration bounds that guarantee the asymptotic optimality guarantee for \( \rho \)-MTS. We also provide examples of many popular risk functionals that satisfy the proposed notion of a continuous and dominant risk functional. We then formally define the risk-averse \( \rho \)-MAB problem, and design two Thompson sampling-based algorithms \( \rho \)-MTS and \( \rho \)-NPTS to solve this problem. Finally, we state our derived regret bound for \( \rho \)-MTS and provide a proof outline of the key ideas involved therein, thus demonstrating the asymptotic optimality of \( \rho \)-MTS. This significantly generalises existing work on Thompson sampling for MABs with finite alphabets to many popular risk functionals used in practice.

**Continuous Risk Functionals**

In this section, we define continuous risk functionals, which are the risk measures of interest in our Thompson sampling algorithms.
Definition 3 (Distorted Risk Functional). Let $C = [0, D]$ and $X$ be a $C$-valued random variable sampled from a probability measure $\mu \in \mathcal{P}$ and CDF $F_\mu$ its corresponding CDF. A conventional risk functional is said to be a distorted risk functional (Wang 1996; Huang et al. 2021) if there exists a distortion function $g$ whose domain is $\mathbb{R}$ and CDF $F$ satisfies $g(0) = 0$ and $g(1) = 1$ such that
\[
    \varrho_g(X) = \int_0^D g(1 - F_\mu(t)) \, dt.
\]  
We append the subscript $g$ to $g$ and write $\varrho_g$ to emphasise the distorted function $g$ associated with $\rho$. By definition, distorted risk functionals are law-invariant. By Remark 1, we can write $\varrho_g(\mu) \equiv \varrho_g(X)$ thereafter and consider distorted risk functionals $\varrho_g$ whose domain is $\mathcal{P}$.

Proposition 2. Suppose $g$ is continuous on $[0, 1]$. Then the distorted risk functional $\varrho_g : \mathcal{P} \to \mathbb{R}$ is continuous on $(\mathcal{P}, D_\infty)$. Consequently, $\varrho_g$ is continuous on $(\mathcal{P}_\epsilon^{(B)}, D_k)$.

Example 4. Table 1 lists some commonly used distorted risk functionals and the properties that they satisfy.

Corollary 1. On the space of rewards in $C$, the risk functionals expected value, CVaR$_\alpha$, proportional hazard, and Lookback as defined in Table 1 are continuous on $(\mathcal{P}, D_\infty)$.

Similar arguments can be used to show that the Cumulative Prospect Theory-Inspired (CPT) functionals (Huang et al. 2021), are also continuous on $(\mathcal{P}, D_\infty)$. Nevertheless, we remark that VaR$_\alpha$ (last row of Table 1) is not continuous on $(\mathcal{P}, D_\infty)$, and thus, does not necessarily enjoy the regret bounds from $\rho$-MTS.

Example 5. Table 2 lists some commonly used EDPMs, their distortion functions, and the (convexity) properties that they satisfy. By their formulation in Cassel, Mannor, and Zeevi (2018), they are all continuous on $(\mathcal{P}, D_\infty)$.

Remark 2. We observe that for scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and continuous risk functionals $\rho_1, \ldots, \rho_n$ on $(\mathcal{P}, d)$, the linear combination $\sum_{i=1}^n \lambda_i \rho_i$ is a continuous risk functional on $(\mathcal{P}, d)$. Furthermore, for any continuous function $\phi : \mathbb{R} \to \mathbb{R}$ and continuous risk functional $\rho$, the composition $\phi \circ \rho$ is also a continuous risk functional. This allows us to consider many combinations of risk functionals.

Example 6 (Continuity of Linear Combinations). For instance, consider the risk functionals MV$_\gamma$, CVaR$_\alpha$, Prop$_p$, LB$_q$ for fixed parameters $\gamma > 0$, $\alpha \in [0, 1)$, $p \in (0, 1)$, $q \in (0, 1)$. By Example 5 and Corollary 1, these risk functionals are continuous on $(\mathcal{P}, D_\infty)$, and the risk functionals $\rho_1 := \text{MV}_\gamma + \text{CVaR}_\alpha$ and $\rho_2 := \text{Prop}_p + \text{LB}_q$ are continuous on $(\mathcal{P}, D_\infty)$, and consequently, are continuous on $(\mathcal{P}_\epsilon^{(B)}, D_k)$. Thus, innumerable risk functionals can be synthesised (as will be done in the section on numerical experiments) and our Thompson sampling-based algorithms are
Definition 4. We show that dominant risk functionals by Riou and Honda (2020) and Baudry et al. (2021). Let $\mathcal{P} \rightarrow \mathbb{R}$ be a risk functional, $\mathcal{S}$ a finite set with size $M + 1$, and denote $\rho_\mathcal{S} := \rho|_{\mathcal{P}\mathcal{S}}$ and $\mathcal{T}_{\rho,\mathcal{S}}(r) := \rho_\mathcal{S}^{-1}([r, \infty))$. For any $p \in \Delta^M$, define $\mathcal{T}_{p,\mathcal{S},p}^{(1)} := \mathcal{T}_{p,\mathcal{S}}(\sigma_{p,\mathcal{S}}(p))$. For any $\mathcal{I} \subseteq [M]_0$, define

$$\mathcal{T}_{\rho,\mathcal{S},p}^{(2)}(\mathcal{I}) := \left\{ \mathcal{D}_\mathcal{S}(q) \in \mathcal{P}\mathcal{S} : q_i \in \left[\{0, p_1\} \text{ for } i \in \mathcal{I} \right] \cup \{0, 1\} \text{ for } i \notin \mathcal{I} \right\}.$$ 

We have $\mathcal{T}_{\rho,\mathcal{S},p}^{(2)}(\emptyset) = \mathcal{T}_{\rho,\mathcal{S},p}^{(2)}([M]_0) = (\mathcal{D}_\mathcal{S}(p))$.

Definition 4. We say that a risk functional $\rho : \mathcal{P} \rightarrow \mathbb{R}$ is dominant if for any finite alphabet $\mathcal{S}$ and $r \in \mathbb{R}$, there exists $\mu_* = \mathcal{D}_\mathcal{S}(p_*)$ with

$$\mathcal{K}^{p}_{\text{inf}}(\mu, r) = \mathcal{K}(\mu, \mu_*)$$

(3)

and $\mathcal{I} \subseteq [M]_0, \mathcal{I} \neq \emptyset, [M]_0$, such that

$$\mathcal{T}_{p,\mathcal{S},p}^{(1)} \supseteq \mathcal{T}_{\rho,\mathcal{S},p}^{(2)}(\mathcal{I}).$$

We remark that if $\rho$ is continuous on $\mathcal{P}\mathcal{S}$, then it satisfies (3) by the Extreme Value Theorem.

We illustrate the property of $\rho$ being dominant in Figure 1. The property states that there exists $p_* \in \Delta^M$ together with some region $\mathcal{A} \subseteq \Delta^M$ such that for all $q \in \mathcal{A}$, $\sigma_{p,\mathcal{S}}(q) \geq \sigma_{p,\mathcal{S}}(p_*)$. The bullet point represents $\mathcal{D}_\mathcal{S}(p_*)$, and the shaded region represents $\mathcal{T}_{\rho,\mathcal{S},p}^{(2)}(\{0, 2\})$. The various types of risk functionals are classified in Figure 3 in the supplementary material.

Tail Lower Bound We show that dominant risk functionals $\rho$ satisfy a generalization of the tail lower bound developed by Riou and Honda (2020) and Baudry et al. (2021).

Lemma 2. Fix a finite set $\mathcal{S}$ with size $M + 1$ and $r \in \mathbb{R}$. Let $\rho : \mathcal{P} \rightarrow \mathbb{R}$ be a dominant risk functional such that $[\mathcal{I}] = M$ (see Definition 4). Fix $\alpha \in \mathbb{N}^{M+1}$, $n = \sum_{i=0}^{M+1} \alpha_i$, and $p = \alpha/n$. For $L \sim \text{Dir}(\alpha)$ and large $n$,

$$\mathbb{P}(\sigma_{p,\mathcal{S}}(L) \geq r) \geq C_2 n^{-M+1} \exp(-n \mathcal{K}^{p}_{\text{inf}}(\mathcal{D}_\mathcal{S}(p), r)),$$

where $C_2 := \sqrt{2\pi} (M/2.13)^{M/2}$.

We remark that Lemma 2 generalises the lower bound in Baudry et al. (2021, Lemma 2) to dominant risk functionals.

Examples of Dominant Risk Functionals We provide numerous examples of risk functionals that satisfy the proposed notion of dominance.

Proposition 3. Let $\rho : \mathcal{P} \rightarrow \mathbb{R}$ be a risk functional satisfying (3). Suppose $\rho$ satisfies one of the following two properties: (a) $\rho$ is a distorted risk functional; (b) for any finite alphabet $\mathcal{S}$ with size $M + 1$, $\sigma_{p,\mathcal{S}} = \mathcal{M}_w$ for some convex $g : \mathbb{R}^{M+1} \rightarrow \mathbb{R}$ with first-order partial derivatives. Then $\rho$ is dominant.

Corollary 2. The risk functionals in Table 1 that are continuous (indicated by $\vee$) and those in Table 2 that are convex (indicated by $\wedge$) are continuous and dominant.

Problem Formulation Given a risk functional $\rho$ on a compact metric subspace $(\mathcal{C}, d) \subset (\mathcal{P}, d)$ of probability measures and $K$ arms with probability measures $(\nu_k)_{k \in [K]} \subset \mathcal{C}$, the learner’s objective is to choose the optimal arm $k^* := \arg \max_{k \in [K]} \rho(\nu_k)$ as many times as possible. All other arms $k \neq k^*$ are called suboptimal. Here we adopt the convention that the arm with higher $\rho(\nu_k)$ offers a higher reward. To adopt the cost perspective, consider the negation of the reward, and the objective as choosing the minimum $\rho(\nu_k)$ over all $k \in [K]$.

Akin to Tammkin et al. (2019), Baudry et al. (2021), and Chang, Zhu, and Tan (2021), we assess the performance of an algorithm $\pi$ using $\rho$, defined at time $n$, by the $\rho$-risk regret

$$\mathcal{R}_\rho^\pi(\pi, n) = \mathbb{E}_\nu \left[ \sum_{t=1}^{n} \left( \max_{k \in [K]} \rho(\nu_{k,t}) - \rho(\nu_{A_t}) \right) \right]$$

$$= \mathbb{E}_\nu \left[ \sum_{t=1}^{n} \Delta_{k^*}^t \right] = \sum_{k=1}^{K} \mathbb{E}_\nu[T_k(n)] \Delta_{k^*}^t,$$

where $\Delta_{k}^t := \rho(\nu_{k,t}) - \rho(\nu_{k^*,t})$ is the difference between the expected reward of arm $k$ and that of the optimal arm $k^*$, and $T_k(n) = \sum_{t=1}^{n} 1(A_t = k)$ is the number of pulls of arm $k$ up to and including time $n$.

Lower Bound We establish an instance-dependent lower bound on the regret incurred by any consistent policy $\pi$, that is,

$$\lim_{n \rightarrow \infty} \mathcal{R}_\rho^\pi(\pi, n)/n^a = 0$$

for any $a > 0$.

Theorem 1. Let $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_K$ be a set of bandit models $\nu = (\nu_1, \ldots, \nu_K)$ where each $\nu_k$ belongs to the class of distributions $\mathcal{Q}_k$. Let $\pi$ be any consistent policy. Suppose without loss of generality that $1$ is the optimal arm, i.e. $T_1^p = \max_{k \in [K]} T_1^p$. For any $\nu \in \mathcal{Q}$ and suboptimal arm $k$,

$$\liminf_{n \rightarrow \infty} \frac{1}{\log n} \mathbb{E}_\nu[T_k(n)] \geq \frac{1}{K \mathcal{K}^{p}_{\text{inf}}(\nu_{k}, T_1^p)}.$$ 

The proof follows that of Baudry et al. (2021) by replacing (CVaR, $c^*$) therein by $(\rho, r_1^p)$, who in turn adapted the proof in Garivier, Ménard, and Stoltz (2019) for their lower bound on the CVaR regret on consistent policies, and thus we relegate it to the supplementary material for brevity.
Algorithm 1: $\rho$-MTS

1: **Input:** Continuous risk functional $\rho$, horizon $n$, support $S = \{s_0, s_1, \ldots, s_M\}$.
2: Set $\alpha_k^m := 1$ for $k \in [K]$, $m \in [M]$, denote $\alpha_k = (\alpha_k^0, \alpha_k^1, \ldots, \alpha_k^M)$.
3: for $t \in [n]$ do
4: for $k \in [K]$ do
5: Sample $L_k^t \sim \text{Dir}(\alpha_k)$. 
6: Compute $r_k^t, \Delta_k^t = \rho(\mathcal{D}_S(L_k^t))$.
7: end for
8: if $t \in [K]$ then
9: Choose action $A_t = t$. 
10: else
11: Choose action $A_t = \arg\max_{k \in [K]} r_k^t$.
12: end if
13: Observe reward $X_{A_t}$. 
14: Increment $a_k^m$ by $\mathbb{I}(X_{A_t} = s_m)$, $m \in [M]$.
15: end for

The $\rho$-MTS and $\rho$-NPTS Algorithms

We design two Thompson sampling-based algorithms, which follow in the spirit of Riou and Honda (2020) and Baudry et al. (2021), called $\rho$-Multinomial-TS ($\rho$-MTS) (resp. $\rho$-Nonparametric-TS ($\rho$-NPTS)), where each $v_k$ follows a multinomial distribution (resp. distribution with bounded support).

$p$-Multinomial-TS ($\rho$-MTS)

Denote the Dirichlet distribution of parameters $\alpha = (\alpha^0, \alpha^1, \ldots, \alpha^M)$ by $\text{Dir}(\alpha)$ with density function

$$f_{\text{Dir}(\alpha)}(x) = \frac{\Gamma(\sum_{i=1}^n n_i \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n x_i^{\alpha_i-1},$$

where $x \in \Delta^M$. The first algorithm, $\rho$-MTS, generalises the index policy in Baudry et al. (2021) from CVaR$_{\alpha}$ to $\rho$.

The conjugate of the multinomial distribution is precisely the Dirichlet distribution. Hence, we generate samples from the Dirichlet distribution, and demonstrate that $\rho$-MTS is optimal in the case where for each $k \in [K]$, $v_k$ follows a multinomial distribution with support $S = \{s_0, s_1, \ldots, s_M\}$ regarded as a subset of $C$, $|S| = M + 1$, $s_0 < s_1 < \ldots < s_M$ without loss of generality, and probability vector $p_k \in \Delta^M$.

In particular, for each $k \in [K]$, we initialise arm $k$ with a distribution of $\text{Dir}(1^M)$, the uniform distribution over $\Delta^M$, where for any $d \in \mathbb{N}$, we denoted $1^d := (1, \ldots, 1) \in \mathbb{R}^d$.

After $t$ rounds, the posterior distribution of arm $k$ is given by $\text{Dir}(1 + T_k^0(t), \ldots, 1 + T_k^M(t))$, where $T_k^d(t)$ denotes the number of times arm $k$ was chosen and reward $s_d$ was received until time $t$. Let $v_k := \mathcal{D}_S(p_k)$ denote the distribution of arm $k$, where $p_k = (p_k^0, p_k^1, \ldots, p_k^M) \in \Delta^M$.

$p$-Nonparametric-TS ($\rho$-NPTS)

To generalise to the bandit setting where the $K$ arms have general distributions with supports in $C \subseteq [0, 1]$, we propose the $\rho$-NPTS algorithm. Unlike $\rho$-MTS that samples for each $k \in [K]$ a probability distribution over a fixed support $\{s_0, s_1, \ldots, s_M\} \subseteq C$, $\rho$-NPTS samples for each $k \in [K]$ a probability vector $L_k^t \sim \text{Dir}(\alpha_k)$ over $(1, X_1^k, \ldots, X_{N_k}^k)$, where $N_k$ is the number of times arm $k$ has been pulled so far. Thus, the support of the sample distribution for $\rho$-NPTS depends on the observed reward, and is not technically a posterior sample with respect to some fixed prior distribution.

The regret guarantee for $\rho$-NPTS is given by

$$R_{\rho}^n(\rho, M, n) = \sum_{k: \Delta_k^t > 0} \frac{\Delta_k^t \log n}{k \inf r_k^t} + o(\log n),$$

where $r_k^t = \rho(v_k)$ for each $k \in [K]$, and $r_k^t = \max_{k \in [K]} r_k^t$ without loss of generality.

Remark 3. When $\rho = \mathbb{E}[:]$ and $\rho = \text{CVaR}_\alpha$, $\rho$ satisfies $|\mathcal{I}| = M$, and we recover the asymptotically optimal algorithms for multinomial distributions in Riou and Honda (2020) and Baudry et al. (2021) respectively. Furthermore, in the setting where $M = 1$, $\rho = \text{MV}_\gamma$, satisfies $|\mathcal{I}| = M$, recovering the Bernoulli-MVTS algorithm proposed by Zhu and Tan (2020). Theorem 2 improves their analysis significantly. First, we replace the term $(2 \min \{p_1 - p_1\}^2(1 - \gamma - p_1 - p_1)^2)^{-1}$ with the exact pre-constant $K_{\inf}(v_k, r_1)^{-1}$ in the log term. Second, we show this attains the asymptotic lower bound in Theorem 1. In general, the distorted risk functionals indicated by $\sqrt{\mathcal{C}}$ in Table 1 are continuous and dominant for any $S$, and EPDMs in Table 2 indicated by $\checkmark$ are continuous and dominant for $|S| = 2 =: M + 1$ (implying $|\mathcal{I}| = 1 = M$), admitting asymptotically optimal $\rho$-MTS algorithms.

Regret Analysis of $\rho$-MTS

In this section we present our regret guarantee for $\rho$-MTS, and show that it matches the lower bound in Theorem 1 and thus is asymptotically optimal.
Proof Outline for Theorem 2. Fix \( \varepsilon_1, \varepsilon_2 > 0 \) and define the two events \( \mathcal{E}_1 := \{ r_k^* \geq r_i - \varepsilon_1, D_\infty(\hat{\sigma}_k(t), \sigma_k) \leq \varepsilon_2 \} \) and \( \mathcal{E}_2 := \mathcal{E}_1^c \), where \( (\hat{\sigma}_k(t), \sigma_k) = (\mathcal{D}_S(\hat{\sigma}_k(t)), \mathcal{D}_S(\sigma_k)) \). We upper bound \( \mathbb{E}[T_k(n)] \) by

\[
\mathbb{E}[T_k(n)] \leq \left[ \sum_{i=1}^n \mathbb{I}(A_i = k, \mathcal{E}_1) \right] + \left[ \sum_{i=1}^n \mathbb{I}(A_i = k, \mathcal{E}_2) \right] + O(1),
\]

where \( \varepsilon_3 \in (0, \min\{\varepsilon_1, \varepsilon_2\}) \) by Lemmas 3 and 4, which are stated below and proven in the supplementary material. Taking \( \varepsilon_1 \to 0^+, \varepsilon_2 \to 0^+ \),

\[
\mathcal{R}_n^\rho(\rho, n) \leq \sum_{k: \Delta_k > 0} \frac{\Delta_k \log n}{K_{n}\inf(\nu_k, r_k^i)} + o(\log n).
\]

as desired. \( \square \)

Lemma 3. Suppose \( \rho \) is continuous \((\mathcal{P}_S, D_\infty)\). For small \( \varepsilon_1, \varepsilon_2 > 0, \varepsilon_3 \in (0, \max\{\varepsilon_1, \varepsilon_2\}) \),

\[
A \leq \frac{\log n}{K_{n}\inf(\nu_k, r_k^i)} - \varepsilon_3/2 + O(1) \quad \text{as} \quad n \to \infty.
\]

Lemma 4. Suppose \( \rho \) satisfies the conditions in Theorem 2. For small enough \( \varepsilon > 0, B \leq O(1) \) as \( n \to \infty \).

These lemmas, which are proved in the supplementary material, arise from the upper bound for continuous risk functionals (Lemma 1) and the lower bound for dominant risk functionals (Lemma 2). These concentration bounds generalise the conclusions of Riou and Honda (2020) and Baudry et al. (2021) to continuous and dominant risk functionals, canonical examples include \( \mathbb{E}[\cdot] \) and \( \text{CVaR}_\alpha \).

Numerical Experiments

We verify our theory via numerical experiments on \( \rho \)-NPTS for new risk measures that are linear combinations of existing ones. These risk measures illustrate the generality and versatility of the theory developed.

We consider a 3-arm bandit instance \((i.e., K = 3)\) with a horizon of \( n = 5,000 \) time steps and over 50 experiments, where the arms 1, 2, 3 follow probability distributions Beta(1, 3), Beta(3, 1), Beta(3, 1) respectively. In particular, we have the means of each arm \( i \) to equal \( i/4 \) for \( i = 1, 2, 3 \). Define the risk functionals \( \rho_1 := \text{MV}_{0.5} + \text{CVaR}_{0.95} \) and \( \rho_2 := \text{Prop}_{0.7} + \text{LB}_{0.6} \) on \((\mathcal{P}_c, D_L)\), where we set \((\gamma, \alpha, p, q) = (0.5, 0.95, 0.7, 0.6)\) as the parameters for the mean-variance, CVaR, Proportional risk hazard, and Lookback components respectively (see Table 1). By Example 6, \( \rho_j \) for \( j = 1, 2 \) are both continuous on \((\mathcal{P}_c, D_L)\).

In Figure 2, we plot the average empirical performance of \( \rho_j \) respectively in green, together with their error bars denoting 1 standard deviation. In both figures, we also plot the theoretical lower bound \( \underline{\varepsilon}_{\rho_j}(n) := \sum_{k=1}^K (\Delta_k \log n) K_{n}\inf(\nu_k, r_k^i)^\alpha \) (cf. Theorem 1) in red and demonstrate that the regrets incurred by \( \rho_j \)-NPTS are competitive compared to the lower bounds, i.e., \( \mathcal{R}_n^\rho(\rho, n) \approx \underline{\varepsilon}_{\rho_j}(n) \) for \( j = 1, 2 \) and large \( n \). The Java code to reproduce the plots in Figure 2 can be found at tinyurl.com/unifyRhoTs.

Conclusion

We posit the first unifying theory for Thompson sampling algorithms on risk-averse MABs. We designed two Thompson sampling-based algorithms given any continuous and dominant risk functional, and prove for many of them asymptotic optimality in the case of multinomially and Bernoulli distributed bandits. We generalise concentration bounds that utilise the continuity and dominance of the risk functional rather than its other properties. There can be further analysis of \( \rho \)-NPTS, which we believe also solves the \( \rho \)-MAB, for \( \rho \)’s under appropriate conditions, with asymptotic optimality. Further work can also adapt the techniques in Baudry, Saux, and Maillard (2021), who designed asymptotically optimal Dirichlet sampling algorithms for bandits whose rewards are unbounded but satisfy mild light-tailed conditions, to the risk-averse setting.

Acknowledgements

The authors would like to thank the reviewers of AAAI 2022 for their detailed and constructive comments, Qiuyu Zhu for initial discussions, and Emilie Kaufmann and Dorian Baudry for indispensable feedback on an earlier version of the manuscript. This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-018) and Singapore Ministry of Education AcRF Tier 1 grant (R-263-000-E80-114).
References


