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

Joel Q. L. Chang[1], **Vincent Y. F. Tan**[1, 2]

[1]Department of Mathematics, National University of Singapore
[2]Department 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 ρ 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 ρ-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 ρ-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 ν_{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; Khajonchotpanya, 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 fuctionals*. 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 ρ. We design two Thompson sampling-based algorithms—ρ-MTS and ρ-NPTS—to solve the modified MABs, provably achieving asymptotic optimality for ρ-MTS under a variety of risk functionals and empirically doing so for ρ-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 L/UCB-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 generalise 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 Cumulative 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 $C \subset \mathbb{R}$. 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 $[M] = \{1, \dots, M\}$ 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 ℓ_∞ 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 σ-algebra denoted by $\mathfrak{B}(C)$, constituting the measurable space $(C, \mathfrak{B}(C))$. For each $c \in C$, let $\delta_c := \mathbb{I}\{c \in \cdot\}$ denote the Dirac measure at c.

Let $\mathcal{P}$ denote the collection of probability measures on $(C, \mathfrak{B}(C))$. 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*

$$D_{\mathrm{L}} : (\mu, \eta) \mapsto \inf\{\varepsilon > 0 : F_\mu(x - \varepsilon) - \varepsilon \leqslant F_\eta(x) \leqslant F_\mu(x + \varepsilon) + \varepsilon, \forall\, x \in \mathbb{R}\}$$

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

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

Example 1 ($(\mathcal{P}_\mathcal{S}, D_\infty)$)**.** We first consider $(\mathcal{P}_\mathcal{S}, D_\infty)$—the set of probability mass functions on a finite alphabet $\mathcal{S} = \{s_0, \dots, s_M\} \subset C$ under the D_∞ metric. For each $p \in \Delta^M$, define $\mu_p = \sum_{i=0}^M p_i \delta_{s_i}$, and $\mathfrak{D}_\mathcal{S} : \Delta^M \to \mathcal{P}$ by $p \mapsto \mu_p$. Then $\mathfrak{D}_\mathcal{S}$ is an imbedding into $\mathcal{P}$ due to the inequality $d_\infty(p, q) \leqslant 2D_\infty(\mathfrak{D}_\mathcal{S}(p), \mathfrak{D}_\mathcal{S}(q)) \leqslant 2M d_\infty(p, q)$. This implies that $(\mathcal{C}, d) := (\mathfrak{D}_\mathcal{S}(\Delta^M), D_\infty)$ is a compact metric space. For brevity, we denote $\mathcal{P}_\mathcal{S} := \mathfrak{D}_\mathcal{S}(\Delta^M)$.

Example 2 ($(\mathcal{P}, D_{\mathrm{L}})$)**.** 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_{L} on $\mathcal{P}$. This implies that $(\mathcal{P}, D_{\mathrm{L}})$ is a compact metric space. Furthermore, by Posner (1975), $\mathrm{KL}(\cdot,\cdot)$ is jointly lower-semicontinuous in both arguments.

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

Thus, we let $(\mathcal{C}, d)$ denote any compact metric subspace of $(\mathcal{P}, d)$, of which includes $(\mathcal{P}_{\mathcal{S}}, D_\infty)$, $(\mathcal{P}, D_{\mathrm{L}})$, and $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_{\mathrm{L}})$. Since C is closed and bounded, we can assume without loss of generality that $C \subseteq [0,1]$ by rescaling.

Let $\mathcal{L}_\infty$ denote the space of C-valued bounded random variables. In particular, we do not place restrictions on the probability space that each $X \in \mathcal{L}_\infty$ is defined on.

Definition 1. A *risk functional* is an $\mathbb{R}$-valued map $\rho : \mathcal{P} \to \mathbb{R}$ on $\mathcal{P}$. A *conventional risk functional* $\varrho : \mathcal{L}_\infty \to \mathbb{R}$ is an $\mathbb{R}$-valued map on $\mathcal{L}_\infty$.

A conventional risk functional $\varrho : \mathcal{L}_\infty \to \mathbb{R}$ is said to be *law-invariant* (Huang et al. 2021) if for any pair of C-valued random variables $X_i : (\Omega_i, \mathcal{F}_i, \mathbb{P}_i) \to (C, \mathfrak{B}(C))$ with probability measures $\mu_i := \mathbb{P}_i \circ X_i^{-1} \in \mathcal{P}$, $i = 1, 2$,

$$\mu_1 = \mu_2 \Rightarrow \varrho(X_1) = \varrho(X_2).$$

Remark 1. We demonstrate in the first section of the supplementary material that ρ is indeed well-defined. That is, for any random variable X sampled from a probability measure μ and law-invariant conventional risk functional ϱ, we can write $\rho(\mu) = \varrho(X)$ without ambiguity. However, we consider it more useful to assume ρ whose domain is a metric space $(\mathcal{P}, d)$, since we can apply the topological results of $(\mathcal{P}, d)$ in the formulation of our concentration bounds.

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 ρ-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 ρ-MAB problem, and design two Thompson sampling-based algorithms ρ-MTS and ρ-NPTS to solve this problem. Finally, we state our derived regret bound for ρ-MTS and provide a proof outline of the key ideas involved therein, thus demonstrating the asymptotic optimality of ρ-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. We demonstrate that when ρ is continuous and dominant (see Definition 4), its corresponding ρ-MTS and ρ-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 ρ 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

$$d(\mu, \eta) < \delta \Rightarrow |\rho(\mu) - \rho(\eta)| < \varepsilon. \tag{1}$$

We say that ρ is *continuous* on a subset $\mathcal{Q} \subseteq \mathcal{P}$ if it is continuous at every $\mu \in \mathcal{Q}$. We say that ρ 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 $(\mathcal{C}, d) \subseteq (\mathcal{P}, d)$ and continuous risk functional ρ, $\rho|_{\mathcal{C}}$ is uniformly continuous on $(\mathcal{C}, d)$.

Let $(\mathcal{C}, d)$ be any of the three compact metric spaces $(\mathcal{P}_{\mathcal{S}}, D_\infty)$, $(\mathcal{P}, D_{\mathrm{L}})$, $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_{\mathrm{L}})$. For any risk functional $\rho : \mathcal{P} \to \mathbb{R}$, define

$$\mathcal{K}_{\inf}^{\rho}(\mu, r) := \inf_{\eta \in \mathcal{C}} \{\mathrm{KL}(\mu, \eta) : \rho(\eta) \geqslant r\}.$$

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

By Lemma 18 in Riou and Honda (2020) ρ being continuous on $(\mathcal{P}, D_{\mathrm{L}})$ implies its continuity on $(\mathcal{P}, D_\infty)$, and ρ being continuous on $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_\infty)$ implies its continuity on $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_{\mathrm{L}})$. 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 ρ that are continuous on $\mathcal{P}_{\mathcal{S}}$ 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 $(\mathcal{P}_{\mathcal{S}}, D_\infty)$ for some finite alphabet $\mathcal{S}$ 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 $L \sim \mathrm{Dir}(\alpha)$,*

$$\mathbb{P}(\sigma_{\rho,\mathcal{S}}(L) \geqslant r) \leqslant C_1 n^{M/2} \exp(-n\mathcal{K}_{\inf}^{\rho_{\mathcal{S}}}(\mathfrak{D}_{\mathcal{S}}(p), r));$$
$$\mathbb{P}(\sigma_{\rho,\mathcal{S}}(L) \leqslant r) \leqslant C_1 n^{M/2} \exp(-n\mathcal{K}_{\inf}^{\rho_{\mathcal{S}}}(\mathfrak{D}_{\mathcal{S}}(p), r)),$$

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

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 $\mathcal{K}_{\inf}^{\rho} : \mathcal{P} \times \rho(\mathcal{C}) \to \mathbb{R}$ 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.

Distorted risk functional	Definition of $\rho_g(\mu) = \varrho_g(X)$	$g(x)$	Continuity of ρ_g
Expectation ($\mathbb{E}$)	$\mathbb{E}[X]$	x	✓
CVaR (CVaR_α)	$-\frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\gamma(X)\,\mathrm{d}\gamma$	$\min\{x/(1-\alpha), 1\}$	✓
Proportional hazard (Prop_p)	$\int_0^\infty (S_X(t))^p\,\mathrm{d}t$	x^p	✓
Lookback (LB_q)	$\int_0^\infty (S_X(t))^q(1 - q\log S_X(t))\,\mathrm{d}t$	$x^q(1 - q\log x)$	✓
VaR (VaR_α)	$-\inf\{x \in \mathbb{R} : F_X(x) > \alpha\}$	$\mathbb{I}\{x \geqslant 1-\alpha\}$	✗

Table 1: A table of common distorted risk functionals, where $S_X(t) := 1 - F_X(t)$ denotes the *decumulative* distribution function (Wang 1996). The risk functionals indicated by ✓ admit the asymptotically optimal result in Theorem 2.

EDPM	$U : \mathcal{P} \to \mathbb{R}$	Convexity
Expectation	$U^{\mathrm{ave}}(\nu) := \mathbb{E}[X]$	✓
Second moment	$U^{\mathbb{E}^2}(\nu) := \mathbb{E}[X^2]$	✓
Below target semi-variance	$U^{-\mathrm{TSV}_r}(\nu) := -\mathbb{E}[(X-r)^2\mathbb{I}\{X \leqslant r\}]$	✓
Entropic risk	$U^{\mathrm{ent}_\theta}(\nu) := -\frac{1}{\theta}\log(\mathbb{E}[-\theta X])$	✓
Negative variance	$U^{-\sigma^2}(\nu) := -(U^{\mathbb{E}^2}(\nu) - (U^{\mathrm{ave}}(\nu))^2)$	✓
Mean-variance	$U^{\mathrm{MV}_\rho}(\nu) := \gamma U^{\mathrm{ave}}(\nu) + U^{-\sigma^2}(\nu)$	✓
Sharpe ratio	$U^{\mathrm{Sh}_r}(\nu) := \frac{U^{\mathrm{ave}}(\nu)-r}{\sqrt{\varepsilon_\sigma - U^{-\sigma^2}(\nu)}}$	✗
Sortino ratio	$U^{\mathrm{So}_r}(\nu) := \frac{U^{\mathrm{ave}}(\nu)-r}{\sqrt{\varepsilon_\sigma - U^{-\mathrm{TSV}_r}(\nu)}}$	✗

Table 2: A table of EDPMs, where $U : \mathcal{P} \to \mathbb{R}$ characterises each risk functional (Cassel, Mannor, and Zeevi 2018). When $M = 1$ (i.e. in the case of Bernoulli bandits), the non-asymptotic lower bound in Lemma 2 and the asymptotically optimal result in Theorem 2 hold for risk functionals indicated by ✓; see Remark 3.

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_μ 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 non-decreasing function $g : [0,1] \to [0,1]$, called a *distortion function*, satisfying $g(0) = 0$ and $g(1) = 1$ such that

$$\varrho_g(X) = \int_0^D g(1 - F_\mu(t))\,\mathrm{d}t. \tag{2}$$

We append the subscript g to ϱ and write ϱ_g to emphasise the distorted function g associated with ρ. By definition, distorted risk functionals are law-invariant. By Remark 1, we can write $\rho_g(\mu) \equiv \varrho_g(X)$ thereafter and consider distorted risk functionals ρ_g whose domain is $\mathcal{P}$.

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

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_α, 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_α (last row of Table 1) is not continuous on $(\mathcal{P}, D_\infty)$, and thus, does not necessarily enjoy the regret bounds from ρ-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 ρ, 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_γ, CVaR_α, 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 $\rho_1 := \mathrm{MV}_\gamma + \mathrm{CVaR}_\alpha$ and $\rho_2 := \mathrm{Prop}_p + \mathrm{LB}_q$ are continuous on $(\mathcal{P}, D_\infty)$, and consequently, are continuous on $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_{\mathrm{L}})$. Thus, innumerable risk functionals can be synthesised (as will be done in the section on numerical experiments) and our Thompson sampling-based algorithms are

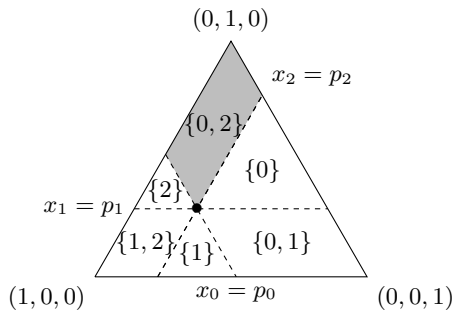

Figure 1: A illustration of a dominant ρ (Definition 4), where x_i's denote the arguments of the Dirichlet distribution in (4).

not only applicable, but also empirically competitive with the theoretical lower bound.

Dominant Risk Functionals

We now introduce the notion of *dominant risk functionals*. These risk functions admit a crucial tail lower bound (stated in Lemma 2 to follow). Let $\rho : \mathcal{P} \to \mathbb{R}$ 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 $\mathcal{T}^{(1)}_{\rho,\mathcal{S},p} := \mathcal{T}_{\rho,\mathcal{S}}(\sigma_{\rho,\mathcal{S}}(p))$. For any $\mathcal{I} \subseteq [M]_0$, define

$$\mathcal{T}^{(2)}_{\mathcal{S},p}(\mathcal{I}) := \left\{ \mathfrak{D}_{\mathcal{S}}(q) \in \mathcal{P}_{\mathcal{S}} : q_i \in \begin{cases} [0, p_i] & \text{for } i \in \mathcal{I} \\ [p_i, 1] & \text{for } i \notin \mathcal{I} \end{cases} \right\}.$$

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

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

$$\mathcal{K}^{\rho}_{\inf}(\mu, r) = \mathrm{KL}(\mu, \mu_*) \tag{3}$$

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

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

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

We illustrate the property of ρ 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 $q \in \mathcal{A}$, $\sigma_{\rho,\mathcal{S}}(q) \geqslant \sigma_{\rho,\mathcal{S}}(p_*)$. The bullet point represents $\mathfrak{D}_{\mathcal{S}}(p_*)$, and the shaded region represents $\mathcal{T}^{(2)}_{\mathcal{S},p}(\{0,2\})$. The various types of risk functionals are classified in Figure 3 in the supplementary material.

Tail Lower Bound We show that dominant risk functionals ρ 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 $\rho : \mathcal{P} \to \mathbb{R}$ be a dominant risk functional such that $|\mathcal{I}| = M$ (see Definition 4). Fix $\alpha \in \mathbb{N}^{M+1}$, $n = \sum_{i=0}^{M} \alpha_i$, and $p = \alpha/n$. For $L \sim \mathrm{Dir}(\alpha)$ and large n,*

$$\mathbb{P}(\sigma_{\rho,\mathcal{S}}(L) \geqslant r) \geqslant C_2 n^{-\frac{M+1}{2}} \exp(-n\mathcal{K}^{\rho_{\mathcal{S}}}_{\inf}(\mathfrak{D}_{\mathcal{S}}(p), r)),$$

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

Corollary 2. *The risk functionals in Table 1 that are continuous (indicated by ✓) and those in Table 2 that are convex (indicated by ✓) are continuous and dominant.*

Problem Formulation

Given a risk functional ρ 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 Tamkin et al. (2019), Baudry et al. (2021), and Chang, Zhu, and Tan (2021), we assess the performance of an algorithm π using ρ, defined at time n, by the *ρ-risk regret*

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

where $\Delta^{\rho}_k := \rho(\nu_{k^*}) - \rho(\nu_k)$ is the difference between the expected reward of arm k and that of the optimal arm k^*, and $T_k(n) = \sum_{t=1}^{n} \mathbb{I}(A_t = k)$ 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 π, that is, $\lim_{n\to\infty} \mathcal{R}^{\rho}_{\nu}(\pi, n)/n^a = 0$ 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 ν_k belongs to the class of distributions $\mathcal{Q}_k$. Let π be any consistent policy. Suppose without loss of generality that 1 is the optimal arm, i.e. $r^{\rho}_1 = \max_{k\in[K]} r^{\rho}_k$. For any $\nu \in \mathcal{Q}$ and suboptimal arm k,*

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu}[T_k(n)]}{\log n} \geqslant \frac{1}{\mathcal{K}^{\rho,\mathcal{Q}_k}_{\inf}(\nu_k, r^{\rho}_1)}.$$

The proof follows that of Baudry et al. (2021) by replacing $(\mathrm{CVaR}_{\alpha}, c^*)$ therein by (ρ, r^{ρ}_1), 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: ρ-MTS

Input: Continuous risk functional ρ, horizon n, support $S = \{s_0, s_1, \dots, s_M\}$.
Set $\alpha_k^m := 1$ for $k \in [K]$, $m \in [M]_0$, denote $\alpha_k = (\alpha_k^0, \alpha_k^1, \dots, \alpha_k^M)$.
for $t \in [n]$ **do**
  for $k \in [K]$ **do**
    Sample $L_k^t \sim \mathrm{Dir}(\alpha_k)$.
    Compute $r_{k,t}^\rho = \rho(\mathfrak{D}_S(L_k^t))$.
  end for
  if $t \in [K]$ **then**
    Choose action $A_t = t$.
  else
    Choose action $A_t = \arg\max_{k\in[K]} r_{k,t}^\rho$.
  end if
  Observe reward X_{A_t}.
  Increment $a_{A_t}^m$ by $\mathbb{I}\{X_{A_t} = s_m\}$, $m \in [M]_0$.
end for

The ρ-MTS and ρ-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 ρ-Multinomial-TS (ρ-MTS) (resp. ρ-Nonparametric-TS (ρ-NPTS)), where each ν_k follows a multinomial distribution (resp. distribution with bounded support).

ρ-Multinomial-TS (ρ-MTS)

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

$$f_{\mathrm{Dir}(\alpha)}(x) = \frac{\Gamma(\sum_{i=1}^n \alpha^i)}{\prod_{i=1}^n \Gamma(\alpha^i)} \prod_{i=1}^n x_i^{\alpha^i - 1}, \tag{4}$$

where $x \in \Delta^M$. The first algorithm, ρ-MTS, generalises the index policy in Baudry et al. (2021) from CVaR_α to ρ. The conjugate of the multinomial distribution is precisely the Dirichlet distribution. Hence, we generate samples from the Dirichet distribution, and demonstrate that ρ-MTS is optimal in the case where for each $k \in [k]$, ν_k follows a multinomial distribution with support $\mathcal{S} = \{s_0, s_1, \dots, s_M\}$ regarded as a subset of C, $|\mathcal{S}| = M + 1$, $s_0 < s_1 < \ldots < s_M$ 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 $\mathrm{Dir}(1^{M+1})$, the uniform distribution over Δ^M, where for any $d \in \mathbb{N}$, we denoted $1^d := (1, \dots, 1) \in \mathbb{R}^d$. After t rounds, the posterior distribution of arm k is given by $\mathrm{Dir}(1 + T_k^0(t), \dots, 1 + T_k^M(t))$, where $T_k^i(t)$ denotes the number of times arm k was chosen and reward s_i was received until time t. Let $\nu_k := \mathfrak{D}_\mathcal{S}(p_k)$ denote the distribution of arm k, where $p_k = (p_k^0, p_k^1, \dots, p_k^M) \in \Delta^M$.

ρ-Nonparametric-TS (ρ-NPTS)

To generalise to the bandit setting where the K arms have general distributions with supports in $C \subseteq [0, 1]$, we propose the ρ-NPTS algorithm. Unlike ρ-MTS that samples for each $k \in [K]$ a probability distribution over a fixed support

Algorithm 2: ρ-NPTS

Input: Continuous risk functional ρ, horizon n, history of the k-th arm $S_k = (1)$, $k \in [K]$.
Set $S_k := (1)$ for $k \in [K]$, $N_k = 1$.
for $t \in [n]$ **do**
  for $k \in [K]$ **do**
    Sample $L_k^t \sim \mathrm{Dir}(1^{N_k})$.
    Compute $r_{k,t}^\rho = \rho(\mathfrak{D}_{S_k}(L_k^t))$.
  end for
  Choose action $A_t = \arg\max_{k\in[K]} r_{k,t}^\rho$.
  Observe reward X_{A_t}.
  Increment N_{A_t} and update $S_{A_t} := (S_{A_t}, X_{A_t})$.
end for

$\{s_0, s_1, \dots, s_M\} \subset C$, ρ-NPTS samples for each $k \in [K]$ a probability vector $L_k^t \sim \mathrm{Dir}(1^{N_k})$ over $(1, X_1^k, \dots, X_{N_k}^k)$, where N_k is the number of times arm k has been pulled so far. Thus, the support of the sampled distribution for ρ-NPTS depends on the observed reward, and is not technically a posterior sample with respect to some fixed prior distribution.

Regret Analysis of ρ-MTS

In this section we present our regret guarantee for ρ-MTS, and show that it matches the lower bound in Theorem 1 and thus is *asymptotically optimal*.

Theorem 2. *Fix a finite alphabet $\mathcal{S} = \{s_0, s_1, \dots, s_M\} \subset C \subseteq [0, 1]$. Let $\nu = (\nu_k)_{k\in[K]} \subset (\mathcal{P}_\mathcal{S}, D_\infty)$ be a K-armed bandit model. Let ρ be a continuous and dominant risk functional on $(\mathcal{P}, D_\infty)$ such that $|\mathcal{I}| = M$ (see Definition 4). Then the regret of ρ-MTS is given by*

$$\mathcal{R}_\nu^\rho(\rho\text{-}MTS, n) \leqslant \sum_{k:\Delta_k^\rho > 0} \frac{\Delta_k^\rho \log n}{\mathcal{K}_{\inf}^\rho(\nu_k, r_1^\rho)} + o(\log n),$$

where $r_k^\rho = \rho(\nu_k)$ for each $k \in [K]$, and $r_1^\rho = \max_{k\in[K]} r_k^\rho$ without loss of generality.

Remark 3. When $\rho = \mathbb{E}[\cdot]$ and $\rho = \mathrm{CVaR}_\alpha$, ρ 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 $M = 1$, $\rho = \mathrm{MV}_\gamma$ 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 $(2\min\{(p_1 - p_i)^2, (1 - \gamma - p_1 - p_i)^2\})^{-1}$ (where $\{p_i\}_{i=1}^K$ are the means of the Bernoulli distributions, and γ is the risk tolerance of the mean-variance) that creates some slackness in their regret bound with the *exact* preconstant $\mathcal{K}_{\inf}^\rho(\nu_k, r_1)^{-1}$ in the log term. Second, we show this attains the asymptotic lower bound in Theorem 1. In general, the distorted risk functionals indicated by ✓ in Table 1 are continuous and dominant for any $\mathcal{S}$, and EPDMs in Table 2 indicated by ✓ are continuous and dominant for $|\mathcal{S}| = 2 =: M + 1$ (implying $|\mathcal{I}| = 1 = M$), admitting asymptotically optimal ρ-MTS algorithms.

Proof Outline for Theorem 2. Fix $\varepsilon_1, \varepsilon_2 > 0$ and define the two events $\mathcal{E}_1 := \{r^\rho_{k,t} \geqslant r^\rho_1 - \varepsilon_1, D_\infty(\widehat{\nu}_k(t), \nu_k) \leqslant \varepsilon_2\}$ and $\mathcal{E}_2 := \mathcal{E}_1^c$, where $(\widehat{\nu}_k(t), \nu_k) = (\mathfrak{D}_\mathcal{S}(\widehat{p}_k(t)), \mathfrak{D}_\mathcal{S}(p_k))$. We upper bound $\mathbb{E}[T_k(n)]$ by

$$\mathbb{E}[T_k(n)] \leqslant \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}(A_t = k, \mathcal{E}_1)\right]}_{A} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}(A_t = k, \mathcal{E}_2)\right]}_{B}$$
$$\leqslant \frac{\log n}{\mathcal{K}^\rho_{\inf}(\nu_k, r^\rho_1) - \varepsilon_3/2} + \mathcal{O}(1),$$

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 $\varepsilon_1 \to 0^+, \varepsilon_2 \to 0^+$,

$$\mathcal{R}^\rho_\nu(\rho\text{-MTS}, n) \leqslant \sum_{k:\Delta^\rho_k > 0} \frac{\Delta^\rho_k \log n}{\mathcal{K}^\rho_{\inf}(\nu_k, r^\rho_1)} + o(\log n).$$

as desired. □

Lemma 3. *Suppose ρ is continuous on $(\mathcal{P}_\mathcal{S}, D_\infty)$. For small $\varepsilon_1, \varepsilon_2 > 0$, $\varepsilon_3 \in (0, \max\{\varepsilon_1, \varepsilon_2\})$,*

$$A \leqslant \frac{\log n}{\mathcal{K}^\rho_{\inf}(\nu_k, r^\rho_1) - \varepsilon_3/2} + \mathcal{O}(1) \quad \textit{as} \quad n \to \infty.$$

Lemma 4. *Suppose ρ satisfies the conditions in Theorem 2. For small enough $\varepsilon > 0$, $B \leqslant \mathcal{O}(1)$ 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 CVaR_α.

Numerical Experiments

We verify our theory via numerical experiments on ρ-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 $\mathrm{Beta}(1,3)$, $\mathrm{Beta}(3,3)$, $\mathrm{Beta}(3,1)$ respectively. In particular, we have the means of each arm i to equal $i/4$ for $i = 1, 2, 3$. Define the risk functionals $\rho_1 := \mathrm{MV}_{0.5} + \mathrm{CVaR}_{0.95}$ and $\rho_2 := \mathrm{Prop}_{0.7} + \mathrm{LB}_{0.6}$ on $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_{\mathrm{L}})$, 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, ρ_j for $j = 1, 2$ are both continuous on $(\mathcal{P}_{\mathrm{c}}^{(B)}, D_{\mathrm{L}})$. In Figure 2, we plot the average empirical performance of ρ_j respectively in green, together with their error bars denoting 1 standard deviation. In both figures, we also plot the theoretical lower bound $\ell_{\rho_j}(n) := \sum_{k=1}^{K} (\Delta_k^{\rho_j} \log n)/\mathcal{K}^{\rho_j}_{\inf}(\nu_k, r_1^{\rho_j})$ (cf. Theorem 1) in red and demonstrate that the regrets incurred by ρ_j-NPTS are competitive compared to the lower bounds, i.e., $\mathcal{R}^{\rho_j}_\nu(\rho_j\text{-NPTS}, n) \approx \ell_{\rho_j}(n)$ for $j = 1, 2$ and large n. The Java code to reproduce the plots in Figure 2 can be found at tinyurl.com/unifyRhoTs.

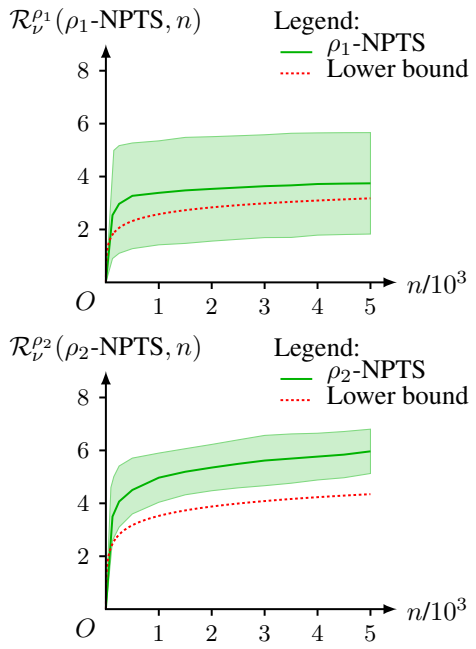

Figure 2: Regrets with risks $\rho_1 = \mathrm{MV}_{0.5} + \mathrm{CVaR}_{0.95}$, $\rho_2 = \mathrm{Prop}_{0.7} + \mathrm{LB}_{0.6}$, and $n = 5,000$ over 50 experiments.

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 ρ-NPTS, which we believe also solves the ρ-MAB, for ρ'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

Agrawal, S.; and Goyal, N. 2012. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In Mannor, S.; Srebro, N.; and Williamson, R. C., eds., *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, 39.1–39.26. Edinburgh, Scotland: JMLR Workshop and Conference Proceedings.

Artzner, P.; Delbaen, F.; Eber, J.-M.; and Heath, D. 1999. Coherent measures of risk. *Mathematical Finance*, 9(3): 203–228.

Baudry, D.; Gautron, R.; Kaufmann, E.; and Maillard, O. 2021. Optimal Thompson Sampling strategies for support-aware CVaR bandits. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 716–726. PMLR.

Baudry, D.; Saux, P.; and Maillard, O.-A. 2021. From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits. arXiv:2111.09724.

Cassel, A.; Mannor, S.; and Zeevi, A. 2018. A General Approach to Multi-Armed Bandits Under Risk Criteria. In Bubeck, S.; Perchet, V.; and Rigollet, P., eds., *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, 1295–1306. PMLR.

Chang, J. Q. L.; Zhu, Q.; and Tan, V. Y. F. 2021. Risk-Constrained Thompson Sampling for CVaR Bandits. arXiv:2011.08046.

Du, Y.; Wang, S.; Fang, Z.; and Huang, L. 2021. Continuous Mean-Covariance Bandits. In *Advances in Neural Information Processing Systems*.

Galichet, N.; Sebag, M.; and Teytaud, O. 2013. Exploration vs Exploitation vs Safety: Risk-Aware Multi-Armed Bandits. In Ong, C. S.; and Ho, T. B., eds., *Proceedings of the 5th Asian Conference on Machine Learning*, volume 29 of *Proceedings of Machine Learning Research*, 245–260. Australian National University, Canberra, Australia: PMLR.

Garivier, A.; Ménard, P.; and Stoltz, G. 2019. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *Mathematics of Operations Research*, 44(2): 377–399.

Granmo, O.-C. 2008. A Bayesian Learning Automaton for Solving Two-Armed Bernoulli Bandit Problems. In *2008 Seventh International Conference on Machine Learning and Applications*, 23–30.

Halmos, P. R. 1959. Review: Nelson Dunford and Jacob T. Schwartz, Linear operators. Part I: General theory. *Bulletin of the American Mathematical Society*, 65(3): 154 – 156.

Huang, A.; Leqi, L.; Lipton, Z. C.; and Azizzadenesheli, K. 2021. Off-Policy Risk Assessment in Contextual Bandits. arXiv:2104.08977.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. arXiv:1205.4217.

Khajonchotpanya, N.; Xue, Y.; and Rujeerapaiboon, N. 2021. A revised approach for risk-averse multi-armed bandits under CVaR criterion. *Operations Research Letters*, 49(4): 465–472.

Lai, T.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22.

Lee, J.; Park, S.; and Shin, J. 2020. Learning Bounds for Risk-sensitive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 13867–13879. Curran Associates, Inc.

Posner, E. 1975. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4): 388–391.

Riou, C.; and Honda, J. 2020. Bandit Algorithms Based on Thompson Sampling for Bounded Reward Distributions. In Kontorovich, A.; and Neu, G., eds., *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, 777–826. PMLR.

Sani, A.; Lazaric, A.; and Munos, R. 2012. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, 3275–3283.

Tamkin, A.; Keramati, R.; Dann, C.; and Brunskill, E. 2019. Distributionally-Aware Exploration for CVaR Bandits. In *Neural Information Processing Systems 2019 Workshop on Safety and Robustness in Decision Making*.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4): 285–294.

Vakili, S.; and Zhao, Q. 2015. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1330–1335.

Wang, S. 1996. Premium Calculation by Transforming the Layer Premium Density. *ASTIN Bulletin*, 26(1): 71–92.

Zhu, Q.; and Tan, V. Y. F. 2020. Thompson Sampling Algorithms for Mean-Variance Bandits. In *International Conference on Machine Learning*, 2645–2654.