

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

Joel Q. L. Chang¹, Vincent Y. F. Tan^{1, 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 simplifies the analysis of risk-averse Thompson sampling algorithms for the multi-armed bandit problem for a generic class of risk functionals ρ that are continuous. Using the contraction principle in the theory of large deviations, we prove novel concentration bounds for these continuous risk functionals. In contrast to existing works in which the bounds depend on the samples themselves, our bounds only depend on the number of samples. This allows us to sidestep significant analytical challenges and unify existing proofs of the regret bounds of existing Thompson sampling-based algorithms. We show that a wide class of risk functionals as well as “nice” functions of them satisfy the continuity condition. Using our newly developed analytical toolkits, we analyse the algorithms ρ -MTS (for multinomial distributions) and ρ -NPTS (for bounded distributions) and prove that they admit asymptotically optimal regret bounds of risk-averse algorithms under the mean-variance, CVaR, and other ubiquitous risk measures, as well as a host of newly synthesized risk measures. Numerical simulations show that our bounds 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 her 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 a specification of a general class of risk functionals, called *coherent risk functionals* (Artzner et al. 1999). Huang et al. (2021) observed that when rewards are nonnegative, coherent risk functionals are subsumed in broader class of functionals called *distortion risk functionals*. Common distortion 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 risk functionals*, which we denote by ρ . We design and analyse two Thompson sampling-based algorithms— ρ -MTS and ρ -NPTS—to solve the modified MABs, achieving asymptotic optimality. 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 risk functionals, 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 the Lipschitz risk functionals which subsumes many of these common risk measures under suitable smoothness assumptions, including variance, mean-variance, distorted risk functionals, and Cumulative Prospect Theory-inspired (CPT) risk functionals.

Contributions

- We explicitly present the key properties that any continuous 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). By the contraction principle in the theory of large deviations, these new tail bounds do not depend on the realisation of the samples $X = (X_1, \dots, X_n)$, which significantly shortens the regret analyses.

- 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 any continuous risk functional ρ , both algorithms are 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).

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) $F_\mu = \mu((-\infty, \cdot]) : C \rightarrow [0, 1]$. 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*

$$D_L : (\mu, \eta) \mapsto \inf\{\varepsilon > 0 :$$

$$F_\mu(x - \varepsilon) - \varepsilon \leq F_\eta(x) \leq F_\mu(x + \varepsilon) + \varepsilon, \forall x \in \mathbb{R}\}$$

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

Example 2 $((\mathcal{P}, D_L))$. 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_L) is a compact metric space. Furthermore, by Posner (1975), $KL(\cdot, \cdot)$ is jointly lower-semicontinuous in both arguments.

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

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

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

A conventional risk functional $\varrho : \mathcal{L}_\infty \rightarrow \mathbb{R}$ is said to be *law-invariant* (Huang et al. 2021) if for any pair of C -valued random variables $X_i : (\Omega_i, \mathcal{F}_i, \mathbb{P}_i) \rightarrow (C, \mathfrak{B}(C))$ 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 risk functionals, and state some essential properties and crucial concentration bounds that guarantee the asymptotic optimality guarantee for ρ -MTS and ρ -NPTS. We also provide examples of many popular risk functionals that satisfy the proposed notion of continuity. Following that, we 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 bounds for ρ -MTS and ρ -NPTS and provide a proof outline of the key ideas involved therein, thus demonstrating the asymptotic optimalities of both algorithms. This significantly expands and generalises existing work on Thompson sampling for MABs with bounded rewards—finite alphabet or continuous—to many popular risk functionals used in practice.

Continuous Risk Functionals

In this section, we will define continuous risk functionals, which are the risk measures of interest in our Thompson sampling algorithms. We demonstrate that when ρ is continuous, its corresponding ρ -MTS and ρ -NPTS algorithms achieve the asymptotically optimal regret bound.

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

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

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

It is straightforward by Lemma 18 in Riou and Honda (2020) that ρ being continuous on (\mathcal{P}, D_L) implies its continuity on (\mathcal{P}, D_∞) , and ρ being continuous on $(\mathcal{P}_c^{(B)}, D_\infty)$ implies its continuity on $(\mathcal{P}_c^{(B)}, D_L)$. This conclusion is consistent with that in Baudry et al. (2021) whose B-CVTS algorithm assumes the rewards of the arm distributions to be continuous.

Example 4 (Continuity of Mean-Variance). Let $\mathbb{E}[\cdot], \mathbb{V}[\cdot]$ denote the risk functionals expectation and variance respectively. By Huang et al. (2021), the risk functionals negative-variance, $-\mathbb{V}[\cdot]$ and mean-variance with parameter $\gamma > 0$, defined by

$$MV_\gamma := \gamma \mathbb{E}[\cdot] - \mathbb{V}[\cdot],$$

are continuous on (\mathcal{P}, D_∞) , and thus are continuous on $(\mathcal{P}_c^{(B)}, D_L)$.

The popular *distorted risk functionals* (Wang 1996; Huang et al. 2021) are continuous under mild assumptions.

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 $g : [0, 1] \rightarrow [0, 1]$, called a *distortion function*, satisfying $g(0) = 0$ and $g(1) = 1$ such that

$$\varrho_g(X) = \int_0^D g(1 - F_\mu(t)) dt. \quad (2)$$

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 1. Suppose g is continuous on $[0, 1]$. Then the distorted risk functional $\rho_g : \mathcal{P} \rightarrow \mathbb{R}$ is continuous on (\mathcal{P}, D_∞) . Consequently, ρ_g is continuous on $(\mathcal{P}_c^{(B)}, D_L)$.

Example 5. Table 1 lists some commonly used distorted risk functionals, their distortion functions, and the properties that they satisfy.

Distorted risk functional	Definition of $\rho_g(\mu) = \varrho_g(X)$	$g(x)$	Continuity of ρ_g
Expectation (\mathbb{E})	$\mathbb{E}[X]$	x	Yes
CVaR (CVaR_α)	$-\frac{1}{\alpha} \int_0^\alpha \text{VaR}_\gamma(X) d\gamma$	$\min\{x/(1-\alpha), 1\}$	Yes
Proportional hazard (Prop_p)	$\int_0^\infty (S_X(t))^p dt$	x^p	Yes
Lookback (LB_q)	$\int_0^\infty (S_X(t))^q (1 - q \log S_X(t)) dt$	$x^q (1 - q \log x)$	Yes
VaR (VaR_α)	$-\inf\{x \in \mathbb{R} : F_X(x) > \alpha\}$	$\mathbb{I}\{x \geq 1 - \alpha\}$	No

Table 1: A table of common distorted risk functionals, where $S_X(t) := 1 - F_X(t)$ denotes the *decumulative* distribution function (Wang 1996).

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

Furthermore, similar arguments can be used to show that the Cumulative Prospect Theory-Inspired (CPT) functionals (Huang et al. 2021), which generalise distorted risk functionals, are also continuous on (\mathcal{P}, D_∞) . Nevertheless, we remark that VaR_α (last row of Table 1) is not continuous on (\mathcal{P}, D_∞) , and thus, does not necessarily enjoy the regret bounds from the ρ -TS algorithms.

Remark 2. We observe that for scalars $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and continuous risk functionals ρ_1, \dots, ρ_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 $\phi : \mathbb{R} \rightarrow \mathbb{R}$ 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 4 and Corollary 1, these risk functionals are continuous on (\mathcal{P}, D_∞) , 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_∞) , and consequently, are continuous on $(\mathcal{P}_c^{(B)}, D_L)$. Thus, innumerable risk functionals can be synthesised (as will be done in the section on numerical experiments) and our Thompson sampling-based algorithms are not only applicable, but also asymptotically optimal.

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 $(\mathcal{S}_S^M, D_\infty)$, (\mathcal{P}, D_L) , $(\mathcal{P}_c^{(B)}, D_L)$. For any risk functional $\rho : \mathcal{P} \rightarrow \mathbb{R}$, define

$$\mathcal{G}_{\inf}^\rho(\mu, r) := \inf_{\eta \in \mathcal{C}} \{\text{KL}(\mu, \eta) : \rho(\mu) \leq r\}, \quad \text{and}$$

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

Novel Concentration Bounds

Next, we include novel concentration bounds which will be needed to prove the near-optimality of the regret bounds of

ρ -MTS and ρ -NPTS. When there is no ambiguity, we will let ρ -TS denote either algorithm.

Lemma 1. *Let (\mathcal{C}, d) be a compact metric space and ρ be a continuous risk functional. Let $\{X_i\}_{i \in [n]}$ denote n i.i.d. rewards sampled from a probability measure $\mu \in \mathcal{C}$, and define $\mathbb{P}_n := \mathbb{P}(\cdot \mid X_1, \dots, X_n)$ for brevity.*

1. *Suppose the metric space $(\mathcal{C}, d) = (\mathcal{S}_S^M, D_\infty)$, $S := \{s_0, s_1, \dots, s_M\} \subseteq \mathcal{C}$. Define $\beta \in \mathbb{N}^{M+1}$ by $\beta_j = \sum_{i=1}^n \mathbb{I}\{X_i = s_j\}$, $\mu = \mathfrak{D}_S(p)$ for some $p \in \Delta^M$, and $L \sim \text{Dir}(\beta)$.*
2. *Suppose $(\mathcal{C}, d) = (\mathcal{P}, D_L)$. Let $S = \{1, X_1, \dots, X_n\}$, $L \sim \text{Dir}(1^{n+1})$.*

Let $\eta = \mathfrak{D}_S(L)$ be a random measure. For any $r \in \rho(\mathcal{C}) \subset \mathbb{R}$, $\delta > 0$, and $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for $n \geq N$,

$$f_{\rho, \mu, \varepsilon}^{n, +}(r + \delta) \leq \mathbb{P}_n(\rho(\eta) \geq r) \leq f_{\rho, \mu, \varepsilon}^{n, -}(r), \quad \text{and}$$

$$g_{\rho, \mu, \varepsilon}^{n, +}(r - \delta) \leq \mathbb{P}_n(\rho(\eta) \leq r) \leq g_{\rho, \mu, \varepsilon}^{n, -}(r)$$

almost surely, where

$$f_{\rho, \mu, \varepsilon}^{n, \pm} := \exp(-n(\mathcal{K}_{\inf}^\rho(\mu, \cdot) \pm \varepsilon)), \quad \text{and}$$

$$g_{\rho, \mu, \varepsilon}^{n, \pm} := \exp(-n(\mathcal{G}_{\inf}^\rho(\mu, \cdot) \pm \varepsilon)).$$

We remark that for Point 2, $1 \in S$ since we initialised ρ -NPTS with an “empirical support” $S = (1)$ for each arm k .

These tail upper and lower bounds generalise the results of Riou and Honda (2020) and Baudry et al. (2021) to the case when the rewards are composed with ρ , and are derived from the theory of large deviations. The proof of Lemma 1 is a consequence of the contraction principle (Dembo and Zeitouni 2009), when applied on the space of random measures that are distributed according to a Dirichlet process (Ganesh and O’Connell 2000). Consequently, we recover relatively simple proofs for theoretically desirable exponential tail bounds which we will use to analyse ρ -TS.

Furthermore, unlike in Riou and Honda (2020) and Baudry et al. (2021), the concentration bounds in Lemma 1 do not depend on the realisations $\{X_i\}_{i \in [n]}$, but only the number of samples and the probability measures they were sampled from. This “independence” allows us to sidestep the discretisation of the samples that reduces the problem to a similar setting to that of ρ -MTS, which vastly shortens the proof of ρ -TS compared to its counterparts in Riou

and Honda (2020); Baudry et al. (2021). We discuss this in greater detail after the proof sketch of ρ -TS (Remark 4) and in the supplementary material.

We will state a corollary of the contraction principle in the theory of large deviations (Dembo and Zeitouni 2009); this is crucial in proving the asymptotic optimality of ρ -TS.

Corollary 2. *Let $\rho : \mathcal{P} \rightarrow \mathbb{R}$ be a continuous risk functional. Then the mapping $\mathcal{K}_{\inf}^\rho : \mathcal{P} \times \rho(\mathcal{C}) \rightarrow \mathbb{R}$ is lower-semicontinuous in its second argument.*

In previous works, the regret bounds of the counterpart algorithms to ρ -MTS and ρ -NPTS were proven to have an asymptotic upper bound involving the term $(\mathcal{K}_{\inf}^\rho(\nu_k, r_1 - \varepsilon))^{-1}$. To remove the $-\varepsilon$ slack term, Riou and Honda (2020) and Baudry et al. (2021) appealed to the continuity of $\mathcal{K}_{\inf}^\rho(\nu_k, \cdot)$. However, by Corollary 2, we note that we only require the *lower-semicontinuity* of $\mathcal{K}_{\inf}^\rho(\nu_k, \cdot)$ to remove the $-\varepsilon$ slack term and derive the asymptotic upper bound involving the term $\mathcal{K}_{\inf}^\rho(\nu_k, r_1)^{-1}$. This significantly extends the asymptotic optimality of the algorithms beyond risk functionals ρ whose corresponding $\mathcal{K}_{\inf}^\rho(\nu_k, \cdot)$ is lower-semicontinuous, but not necessarily continuous. Indeed, by Corollary 2, it suffices for ρ to be continuous in order for $\mathcal{K}_{\inf}^\rho(\nu_k, \cdot)$ to be lower-semicontinuous.

Problem Formulation

Given a continuous 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 $k \in [K]$.

In the spirit of 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

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

where $\Delta_k^\rho := \rho(\nu_{k^*}) - \rho(\nu_k)$ 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, $\lim_{n \rightarrow \infty} \mathcal{R}_\nu^\rho(\pi, n)/n^a = 0$ for any $a > 0$.

Theorem 1. *Let $\mathcal{Q} = \mathcal{Q}_1 \times \dots \times \mathcal{Q}_K$ be a set of bandit models $\nu = (\nu_1, \dots, \nu_K)$ 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. $r_1^\rho = \max_{k \in [K]} r_k^\rho$. Then for any $\nu \in \mathcal{Q}$, for any suboptimal arm k , we have

$$\liminf_{n \rightarrow \infty} \frac{\mathbb{E}_\nu [T_k(n)]}{\log n} \geq \frac{1}{\mathcal{K}_{\inf}^{\rho, \mathcal{Q}_k}(\nu_k, r_1^\rho)}.$$

The proof follows that of Baudry et al. (2021) by replacing $(\text{CVaR}_\alpha, c^*)$ therein by (ρ, r_1^ρ) , 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.

The ρ -MTS and ρ -NPTS Algorithms

In this paper, we design and analyse 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. bounded distribution).

ρ -Multinomial-TS (ρ -MTS)

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

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

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 Dirichlet 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 $S = (s_0, s_1, \dots, s_M)$ regarded as a subset of C , $|S| = M + 1$, $s_0 < s_1 < \dots < s_M$ 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 $\text{Dir}(1^{M+1})$, 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 $\text{Dir}(1 + T_k^0(t), \dots, 1 + T_k^M(t))$, where $T_k^i(t)$ denotes the number of times arm k was chosen and reward s_i was received until time t . Let $\nu_k := \mathcal{D}_S(p_k)$ 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 $\{s_0, s_1, \dots, s_M\} \subset C$, ρ -NPTS samples for each $k \in [K]$ a probability vector $L_k^t \sim \text{Dir}(1^{N_k})$ 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. Nevertheless, the probability measures $\mathcal{D}_{S_k}(L_k^t)$ are

Algorithm 1: ρ -MTS

```

1: Input: Continuous risk functional  $\rho$ , horizon  $n$ , support
    $S = \{s_0, s_1, \dots, s_M\}$ .
2: Set  $\alpha_k^m := 1$  for  $k \in [K]$ ,  $m \in [M]_0$ , denote  $\alpha_k =$ 
    $(\alpha_k^0, \alpha_k^1, \dots, \alpha_k^M)$ .
3: for  $t \in [n]$  do
4:   for  $k \in [K]$  do
5:     Sample  $L_k^t \sim \text{Dir}(\alpha_k)$ .
6:     Compute  $r_{k,t}^\rho = \rho(\mathfrak{D}_S(L_k^t))$ .
7:   end for
8:   if  $t \in [K]$  then
9:     Choose action  $A_t = t$ .
10:  else
11:    Choose action  $A_t = \arg \max_{k \in [K]} r_{k,t}^\rho$ .
12:  end if
13:  Observe reward  $X_{A_t}$ .
14:  Increment  $a_{A_t}^m$  by  $\mathbb{I}\{X_{A_t} = s_m\}$ ,  $m \in [M]_0$ .
15: end for

```

Algorithm 2: ρ -NPTS

```

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

```

still distributed according to a Dirichlet process, and we can still obtain exponential tail bounds on the respective conditional probabilities; see Lemma 1.

Regret Analyses of ρ -MTS and ρ -NPTS

In this section we present our regret guarantees for ρ -MTS and ρ -NPTS, and show that they both match the lower bound in Theorem 1 and thus are *asymptotically optimal*. We will let ρ -TS denote ρ -MTS, ρ -NPTS in the settings $(\mathcal{C}, d) = (\mathcal{S}_S^M, D_\infty)$, (\mathcal{P}, D_L) respectively.

Theorem 2. *Let $\nu = (\nu_k)_{k \in [K]} \subset (\mathcal{C}, d)$ be a bandit model with K arms with common compact support $C \subseteq [0, 1]$. In the case $(\mathcal{C}, d) = (\mathcal{S}_S^M, D_\infty)$, let $S = \{s_0, s_1, \dots, s_M\} \subset C$ be the common support. Let ρ be a continuous risk functional on (\mathcal{C}, d) . Then the regret of ρ -TS is given by*

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

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. Replacing the setting (\mathcal{P}, D_L) with $(\mathcal{P}_c^{(B)}, D_L)$ does not change the conclusion.

Remark 3. We remark that in the settings $\rho = \mathbb{E}[\cdot]$ and $\rho = \text{CVaR}_\alpha$, we recover the asymptotically optimal algorithms in Riou and Honda (2020) and Baudry et al. (2021) respectively. Furthermore, in the setting $\rho = \text{MV}_\gamma$, and $M = 1$ in Theorem 2 for ρ -MTS, we recover the Bernoulli-MVTS (or B-MVTS) algorithm in Zhu and Tan (2020). We improve their analyses in two significant ways. 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) that creates some slackness in their regret bound with the *exact* pre-constant $\mathcal{K}_{\inf}^\rho(\nu_k, r_1)^{-1}$ in the log term. Second, we show this attains the theoretical asymptotic lower bound (Theorem 1). Finally, the EDPMs in Cassel, Mannor, and Zeevi (2018) are continuous on (\mathcal{P}, D_∞) by definition, and thus continuous on $(\mathcal{P}_c^{(B)}, D_L)$. Hence, the plethora of risk measures discussed therein, such as the second moment, the entropic risk, and the Sharpe ratio, admit corresponding asymptotically optimal ρ -TS algorithms that improve on all existing risk-averse MAB regret minimization algorithms.

Proof Outline for Theorem 2. Let (\mathcal{C}, d) be either of the compact metric spaces $(\mathcal{S}_S^M, D_\infty)$ or (\mathcal{P}, D_L) . Fix $\varepsilon > 0$ and define the two events

$$\mathcal{E}_1 := \{r_{k,t}^\rho \geq r_1^\rho - \varepsilon\} \quad \text{and} \quad \mathcal{E}_2 := \{r_{k,t}^\rho < r_1^\rho - \varepsilon\},$$

where $(\widehat{\nu}_k(t), \nu_k) = (\mathfrak{D}_S(\widehat{p}_k(t)), \mathfrak{D}_S(p_k))$ in the setting $(\mathcal{C}, d) = (\mathcal{S}_S^M, D_\infty)$. It suffices to upper bound $\mathbb{E}[T_k(n)]$ by partitioning into events \mathcal{E}_1 and \mathcal{E}_2 , namely,

$$\begin{aligned} \mathbb{E}[T_k(n)] &\leq \underbrace{\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}(A_t = k, \mathcal{E}_1)\right]}_A + \underbrace{\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}(A_t = k, \mathcal{E}_2)\right]}_B \\ &\leq \frac{\log n}{(1 - \varepsilon)\mathcal{K}_{\inf}^\rho(\nu_k, r_1^\rho - \varepsilon)} + O(1), \end{aligned}$$

by Lemmas 2 and 3 in either setting $(\mathcal{S}_S^M, D_\infty)$ or (\mathcal{P}, D_L) , which are stated below and proven in the supplementary material. Taking $\varepsilon \rightarrow 0^+$, and using the lower semi-continuity of \mathcal{K}_{\inf}^ρ in its second argument (Corollary 2) which yields $\mathcal{K}_{\inf}^\rho(\nu_k, r_1^\rho - \varepsilon) \geq \mathcal{K}_{\inf}^\rho(\nu_k, r_1^\rho)(1 + o_\varepsilon(1))$,

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

as desired. \square

Lemma 2. *Suppose ρ is continuous on (\mathcal{C}, d) . For sufficiently small $\varepsilon > 0$ and sufficiently large n ,*

$$A \leq \frac{\log n}{(1 - \varepsilon)\mathcal{K}_{\inf}^\rho(\nu_k, r_1^\rho - \varepsilon)} + 1.$$

Lemma 3. *Suppose ρ is continuous on (\mathcal{C}, d) . For sufficiently small $\varepsilon > 0$, $B \leq O(1)$.*

These lemmas, which are proved in the supplementary material, arise from the novel concentration bounds for continuous risk functionals ρ stated in Lemma 1. These concentration bounds generalise the conclusions of Riou and

Honda (2020) and Baudry et al. (2021) to continuous risk functionals, canonical examples include $\mathbb{E}[\cdot]$ and CVaR_α . Furthermore, by Remark 2, we can generate other risk functionals that are continuous on their respective metric spaces, and hence, admit asymptotically optimal ρ -TS algorithms.

Remark 4. We vastly simplify the proof of the upper bound on the regret of ρ -TS, since unlike in previous work, our concentration bounds depend only on the **number** of samples drawn up to time n , and which probability measures they are drawn from, rather than on the **empirical distribution** which requires partitioning of its plausible values.

To illustrate this point, let the rewards $\{X_i\}_{i \in [n]}$ be drawn from a certain probability measure μ . Let $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ denote the empirical measure derived from the samples, and $\mathbb{P}_n := \mathbb{P}(\cdot \mid X_1, \dots, X_n)$. The concentration bounds in previous works (e.g., Corollary 16 in Riou and Honda (2020) and Appendix E in Baudry et al. (2021)) take the form

$$\mathbb{P}_n(\rho(\eta) \leq r) \leq O(\exp(-n\mathcal{G}_{\text{inf}}^\rho(\hat{\mu}_n, r))) \quad \text{a.s.}$$

while the concentration bound in Lemma 1 takes the form

$$\mathbb{P}_n(\rho(\eta) \leq r) \leq \exp(-n(\mathcal{G}_{\text{inf}}^\rho(\mu, r) - \varepsilon)) \quad \text{a.s.}$$

In Lemma 1, the upper bound depends only on μ and not on $\hat{\mu}_n$, sidestepping the need to partition $\rho(\hat{\mu}_n)$ into various cases; see Appendix B.2 in Riou and Honda (2020) for example. Consequently, we are able to sidestep the technically challenging discretisation arguments for ρ -NPTS; see Appendix C.2 in Baudry et al. (2021) for example. Furthermore, the upper bound in Lemma 1 does not require knowledge of the closed form of ρ , unlike previous works (e.g. Appendices F and G in Riou and Honda (2020) and Appendices D and E in Baudry et al. (2021)), which widens its applicability in the analysis of general ρ -TS algorithms. These advantages greatly shorten the proof of Theorem 2 into a clearer and more elegant one. These points are further elaborated on at the end of the supplementary material.

Numerical Experiments

We verify our theory via numerical experiments on ρ -NPTS for new risk measures that are linear combinations of existing ones. Even though these risk measures may not be widely used at this point in time, they 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 $\text{Beta}(1, 3)$, $\text{Beta}(3, 3)$, $\text{Beta}(3, 1)$ 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 $(\mathcal{P}_c^{(B)}, D_L)$, 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 Look-back components respectively (see Table 1). By Example 6, ρ_j for $j = 1, 2$ are both continuous on $(\mathcal{P}_c^{(B)}, D_L)$. In Figure 1, 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 $\ell_{\rho_j}(n) := \sum_{k=1}^K (\Delta_k^{\rho_j} \log n) / \mathcal{K}_{\text{inf}}^{\rho_j}(\nu_k, r_1^{\rho_j})$

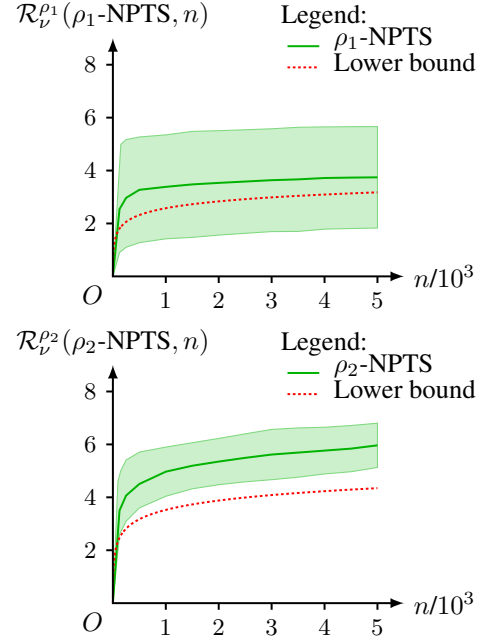


Figure 1: Regrets with risks $\rho_1 = \text{MV}_{0.5} + \text{CVaR}_{0.95}$, $\rho_2 = \text{Prop}_{0.7} + \text{LB}_{0.6}$, and $n = 5000$ over 50 experiments.

(cf. Theorem 1) in red and demonstrate that the regrets incurred by ρ_j -NPTS are competitive compared to the lower bounds, i.e., $\mathcal{R}_n^{\rho_j}(\rho_j\text{-NPTS}, n) \approx \ell_{\rho_j}(n)$ for $j = 1, 2$ and large n . The Java code to reproduce the plots in Figure 1 can be found at this Github link.

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 risk functional, and prove their asymptotic optimality. We proved new concentration bounds that utilise the continuity of the risk functional rather than its other properties. There can be further exploration of Thompson sampling algorithms for non-continuous risk functionals, and exploring sufficient conditions to extend the theory of Thompson sampling algorithms for risk-averse MABs. 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 and Qiuyu Zhu for initial discussions. 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.
- Dembo, A.; and Zeitouni, O. 2009. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg. ISBN 9783642033117.
- 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.
- Ganesh, A. J.; and O’Connell, N. 2000. A large-deviation principle for Dirichlet posteriors. *Bernoulli*, 1021–1034.
- 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.