Approximate Optimality and the Risk/Reward Tradeoff in a Class of Bandit Problems[*]

Zengjing Chen Larry G. Epstein Guodong Zhang

November 29, 2023

Abstract

This paper studies a sequential decision problem where payoff distributions are known and where the riskiness of payoffs matters. Equivalently, it studies sequential choice from a repeated set of independent lotteries. The decision-maker is assumed to pursue strategies that are approximately optimal for large horizons. By exploiting the tractability afforded by asymptotics, conditions are derived characterizing when specialization in one action or lottery throughout is asymptotically optimal and when optimality requires intertemporal diversification. The key is the constancy or variability of risk attitude, that is, the decision-maker's risk/reward tradeoff. The main technical tool is a new central limit theorem.

Keywords: sequential decision problem, multi-armed bandit, risk/reward tradeoff, large-horizon approximations, central limit theorem, semivariance, asymptotics, repeated gambles, diversification

1 Introduction

We study the following sequential choice problem. There are K arms (or actions), each yielding a random payoff. Payoff distributions are independent across arms and identical and independent for a given arm across distinct trials. At each stage $i = 1, 2, ..., n$, the decision-maker (DM) must choose one arm, knowing both the realized payoffs from previous choices and the distribution of the payoff for each arm. She chooses a strategy ex ante specifying future contingent choices. This is a special case of a bandit problem, whence the usage of 'arm' rather than 'action.' Alternatively, the decision problem can be viewed as choice of a dynamic strategy when facing a repeated set of (single-stage) gambles or lotteries, where each lottery is repeated independently. Thus it is an instance of dynamic risk management.

Because we are interested in varying horizons, we define a strategy for an infinite horizon, and then use its truncation for any given finite horizon. Refer to a strategy as *asymptotically optimal* if the expected utility it implies in the limit as horizon

[*] Chen is at School of Mathematics, Shandong University, zjchen@sdu.edu.cn, Epstein is at Department of Economics, McGill University, larry.epstein@mcgill.ca, and Zhang is at School of Mathematics, Shandong University, zhang_gd@mail.sdu.edu.cn. Chen gratefully acknowledges the support of the National Key R&D Program of China (grant No. ZR2019ZD41), and the Taishan Scholars Project. Zhang gratefully acknowledges the support of Shandong Provincial Natural Science Foundation, China (grant No. ZR2022QA063 and ZR2021MA098).

$n \to \infty$ is at least as large as that implied by any other strategy; or equivalently, if it is *approximately optimal for large horizons*. We study large-horizon approximations to the value (indirect utility) of the sequential choice problem and also corresponding asymptotically optimal strategies. Our focus is on the derivation of analytical (as opposed to computational) results, particularly with regard to the effect of risk non-neutrality. For example, we demonstrate that (non)constancy of risk attitude, suitably measured, determines whether specialization in a single arm throughout or diversification across time is asymptotically optimal.

Consider three concrete settings that fit our model well. *Gambling*: A gambler chooses sequentially which of several given slot machines to play. *News site*: Each visitor to a site decides whether to click depending on the news header presented to her. The website (DM) chooses the header (arm) with clicks being the payoffs. Users are drawn independently from a fixed distribution. *Ad selection*: A website (DM) displays an ad (arm) for each visitor, who is an i.i.d. draw as above. If she clicks, the payoff to the website is a predetermined price, depending on the ad and paid by the advertiser. Importantly for the fit with our model, in all three settings payoffs are realized quickly after an arm is chosen, and plausibly a large number of trials occur in a relatively short period of time.[1]

We have two related reasons for studying asymptotics. First, from the modeler's perspective, it promotes tractability and the derivation of analytical results. Bandit problems are notoriously difficult to solve analytically, as opposed to numerically, in the presence of nonindifference to risk. A second reason for studying asymptotics is that tractability may be a concern also for the decision-maker within the model who faces an extremely complicated large-horizon optimization problem. In such circumstances, she may seek a strategy that is approximately optimal if her horizon is sufficiently long. The presumption that a large-horizon heuristic can alleviate cognitive limitations is supported by two features of our results: (i) asymptotic optimality depends on payoff distributions and the values they induce *only through their means and variances* (Theorem 1), that is, *DM need not know more about the distributions*; and (ii) also by the relative simplicity of the explicit asymptotically optimal strategies in some cases (Theorem 3).

The focus on asymptotics leads to other noteworthy features of our analysis. First, unsurprisingly, it leads to our exploiting limit theorems, most notably a central limit theorem (CLT). The classical CLT considers a sequence (X_i) of identically and independently distributed random variables, hence having a fixed mean and variance, which assumptions are adequate for evaluation of the repeated play of a single arm, and hence also for addressing the once-and-for-all choice between arms. However, in the more economically relevant case of sequential choice, we must evaluate strategies which permit switching between arms, and hence also between payoff distributions, at any stage. Accordingly, in our key technical result, CLT (Proposition 6), means and variances of (X_i) can vary with i subject only to the restriction that they lie in a fixed set.

The role played by limit theorems is reflected also in our specification of the utility index u. We adopt a form of multiattribute utility theory (Keeney and Raiffa 1993), whereby two attributes of random payoff streams are assumed to be impor-

[1] Daily life provides other repeated choice problems, for example, which transportation mode or route to use to get to work, though the longer time interval between choices suggests a poorer fit with the model.

tant. Accordingly, $u : \mathbb{R}^2 \longrightarrow \mathbb{R}$ has two arguments, namely the sample average and the $\sqrt{n}$-weighted average of deviations from conditional means, exactly the statistics whose limiting distributions are the focus in the LLN (law of large numbers) and CLT respectively. The function u itself is restricted only by technical conditions. Nevertheless, the resulting model is both tractable and also flexible enough to accommodate interesting special cases. As an example of the diversity of cases accommodated, one is a form of mean-variance for our sequential setting, and another essentially replaces variance by semivariance. The differing implications of these two specifications illustrate one message that the paper is intended to convey: the mean-variance model exhibits constant risk attitude and accordingly predicts specialization in one arm, that is, time-diversification is not important in sufficiently large horizons, while risk attitude varies endogenously in the mean-semivariance model which therefore predicts specialization only for some but not all parameter values.

The paper proceeds as follows. Related literature is discussed next. The model and main results follow in Section 2. Most proofs are provided in the Appendix, which also contains our CLT. Proofs of some details are collected in the Online Appendix.

1.1 Related literature

Decision-making in the presence of repeated gambles has been studied in several papers. We mention some that help to locate this paper in the context of this literature. In McCardle and Winkler (1992), a coin with uncertain bias is tossed repeatedly. The decision-maker observes the outcomes of all tosses, updates her beliefs accordingly, and chooses sequentially how much to bet on heads at each history.Full Bayesian rationality is assumed. The authors argue that some of the model's predictions about willingness to bet are unintuitive and they attribute this to the assumption that future betting opportunities are fully anticipated and incorporated via optimization. Accordingly, they suggest the need for simplifying heuristics that still accommodate some, but not all, "grand world" considerations, though no specific heuristics are proposed. We share the broad view that dynamic decision problems under uncertainty are exceedingly complex and propose, for our setting, the simplification consisting of approximate optimality for large horizons. In Gollier (1996), a single lottery (with known distribution) is repeated independently, and the decision-maker accepts or rejects the lottery at each stage. Choice is determined by maximization of the expected utility of terminal wealth. His paper and ours address different questions. Gollier focuses on how the option to gamble in the future affects the willingness to gamble today, while we are focussed on behavior in the remote future because it describes approximately optimal behavior for sufficiently long horizons. Another difference is that his setting with a riskless option can be viewed as the special case of our setting where there are two lotteries at each stage and where one is degenerate (attaches probability 1 to the outcome 0). The assumption that there is only one risky asset (or lottery) and one riskless is common in finance. However, it is restrictive and it is not clear if and how Gollier's analysis would extend. Both options being risky poses significant technical complications for the modeler and cognitive challenges for the decision-maker within the model.

Samuelson (1963) identified as fallacious the reliance on the law of large numbers as justifying acceptance of any sufficiently long sequence of repetitions of a positive mean bet even if the single bet is rejected, and suggested that it indicated undue attention to the variance associated with multiple repetitions. While we do not address Samuelson's fallacy here, we see in it the hint that there could be a role for the other major limit theorem, the CLT, in the broader study of risk-taking given repeated gambles. In that sense, this paper is inspired by Samuelson (1963). Related is the literature examining the effect on financial risk-taking of horizon length (e.g. of age in a life-cycle portfolio context), for example, whether a longer horizon promotes risk-taking because it offers a greater possibility to smooth out risks over time (see, for example, Samuelson (1989) and Gollier and Zeckhauser (2002)). We differ from this literature in (at least) two respects. First, we model behavior in the long-horizon limit; we do not study the effect of differing horizon length on risk-taking. A second critical difference is that while in the finance literature, assets are divisible and can be combined into portfolios at any stage, the lotteries available to our decision-maker are indivisible and only one can be chosen at any stage. Consequently, portfolio diversification is excluded herein while diversification over time is feasible and a focus.[2]

Approximate optimality for long horizons has been studied in finance in the context of portfolio turnpike theorems (see, for example, Huberman and Ross (1983) and the references therein). This literature studies the conditions under which wealth-independent (hence "constant") portfolios are approximately optimal for sufficiently long horizons. Accordingly, an important factor is the relation between wealth and the benefit from diversifying across assets at any given time. In contrast, in our model at any instant the decision-maker can choose a single lottery from the given finite set, and thus only diversification over time is feasible. In addition, we study approximate optimality without imposing any form of constancy; for example, Theorem 2(v) illustrates the case of an asymptotically optimal strategy that is not constant across time (that is, the gamble chosen varies with time).

All of the papers cited above assume maximization of the expected utility of terminal wealth. As outlined earlier, we model payoffs and utility differently. There are precedents for "nonstandard" utility specifications in the context of repeated gambles; for example, alternatives to expected utility theory are either adopted or advocated by Chew and Epstein (1988), Benartzi and Thaler (1999) and Lopes (1996).

The other major connection is to the bandit literature since our decision problem is the special case of a multi-armed bandit problem where payoff distributions are known and hence need not be learned. Most of the literature (see Berry and Fristedt (1985) and Slivkins (2022) for textbook-like treatments) assumes a finite horizon and that choices are driven by expected total rewards, that is, risk neutrality. Studies that explicitly address risk attitudes include Sani, Lazaric and Munos (2013), Zimin, Ibsen-Jensen and Chatterjee (2014), Vakili and Zhao (2016), and Cassel, Manor and Zeevi (2021). They assume regret minimization rather than expected utility maximization, and focus on computational algorithms rather than on qualitative theoretical results. Further, they are motivated by the nature of learning about unknown payoff distributions, and thus by the exploration/exploitation tradeoff, while we assume known distributions and focus instead on the risk/reward tradeoff. Though it is important to understand both tradeoffs and their interactions, as an initial step we focus on only one in this paper, that being the tradeoff for which there exists very limited theoretical analysis. Theorem 3 gives analytical results on the latter tradeoff by exploiting the advantages of large-horizon approximations.

[2] When both kinds of diversification are feasible, Samuelson (1989,1997) argues that time-diversification is inferior. Here we explore whether time-diversification is useful for long horizon planning in settings where portfolio diversification is not feasible.

In a more technical vein, our CLT connects this paper to the literature on nonlinear CLTs, that is CLTs where the expectations operator is nonlinear, for example, because of the multiplicity of priors and where expectation is defined by the infimum (or supremum) of expectations as one varies over all priors. The infimum is typically motivated, as in the maxmin model (Gilboa and Schmeidler 1989), by robustness to ambiguity or model uncertainty. The nonlinear CLTs in Peng (2007, 2019) and Fang et al (2019) are motivated in this way (see Peng (2019, Thm 2.4.8), for example). They do not make a connection to Bayesian sequential decision-making, nor is such a connection apparent in their work. In contrast, the decision-maker in our model is Bayesian and does not perceive ambiguity. Nevertheless, a set of probability measures arises (implicitly) from the multiplicity of arms and strategies, and a supremum applies because of utility maximization over the set of strategies, or equivalently, over the probability measures they induce. Chen, Epstein and Zhang (2023) introduced the use of a nonlinear CLT to model Bayesian decision-makers. It differs from the present paper both technically and in its economic focus as explained following the statement of our CLT (Proposition 6).

2 The Model

2.1 Preliminaries

Let $(\Omega, \mathcal{F}, P)$ be the probability space on which all subsequent random variables are defined. The random variables X_k, $1 \leq k \leq K$, represent the random rewards from the K arms, and $\{X_{k,n} : n \geq 1\}$ denote their independent and identically distributed copies. We assume that each X_k has a finite mean and variance, denoted by

$$\mu_k := E_P[X_k],\ \sigma_k^2 := Var_P\left[X_k\right], \quad 1 \leq k \leq K. \tag{1}$$

The largest and smallest means and variances are given by

$$\begin{aligned} &\overline{\mu} = \max\{\mu_1, \cdots, \mu_K\}, \ \ \underline{\mu} = \min\{\mu_1, \cdots, \mu_K\}, \\ &\overline{\sigma}^2 = \max\{\sigma_1^2, \cdots, \sigma_K^2\}, \ \underline{\sigma}^2 = \min\{\sigma_1^2, \cdots, \sigma_K^2\}. \end{aligned} \tag{2}$$

The set of mean-variance pairs is

$$\mathcal{A} = \{\left(\mu_k, \sigma_k^2\right) : 1 \leq k \leq K\}. \tag{3}$$

The convex hull of $\mathcal{A}$ is a convex polygon. Denote by $\mathcal{A}^{ext}$ its set of extreme points.

A *strategy* θ is a sequence of $\{1, \cdots, K\}$-valued random variables, $\theta = (\theta_1, \cdots, \theta_n, \cdots)$. θ selects arm k at round n in states for which $\theta_n = k$. Thus the corresponding reward is Z_n^θ given by

$$Z_n^\theta = X_{k,n} \text{ where } \theta_n = k. \tag{4}$$

The strategy θ is *admissible* if θ_n is $\mathcal{H}_{n-1}^\theta$-measurable for all $n \geq 1$, where

$$\mathcal{H}_{n-1}^\theta = \sigma\{Z_1^\theta, \cdots, Z_{n-1}^\theta, \theta_1, ..., \theta_{n-1}\} \text{ for } n > 1, \text{ and } \mathcal{H}_0^\theta = \{\emptyset, \Omega\}.$$

The information at stage n captured by $\mathcal{H}_{n-1}^\theta$ includes both past choices of arms and the corresponding history of payoffs. Allowing the arm chosen at stage n to depend on past choices permits strategies that alternate stochastically between arms. Given the serial independence of payoffs, there is no learning rationale for conditioning on past payoffs. However, *past payoffs matter in general at any stage because they may influence the attitude towards the risk associated with current and future choices.*

The set of all admissible strategies is Θ. (All strategies considered below will be admissible, even where not specified explicitly.)

2.2 Utility

For each horizon n, we specify the expected utility function U_n used to evaluate strategies θ and the payoff streams that they generate. Let $u : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be the corresponding von-Neumann Morgenstern (vNM) utility index and define U_n by

$$U_n(\theta) = E_P \left[u \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^\theta, \left(\sum_{i=1}^{n} \frac{1}{\sqrt{n}} \left(Z_i^\theta - E_P[Z_i^\theta | \mathcal{H}_{i-1}^\theta] \right) \right) \right) \right]. \tag{5}$$

The two arguments of u correspond to the two attributes or characteristics of a random payoff stream that DM takes into account. The first argument of u is the sample average outcome under strategy θ, and the second, the $\sqrt{n}$-weighted average of deviations from conditional means, represents sample volatility. Observe that the second argument has zero expected value relative to the measure P. Though one might have expected the term (as volatility) to be replaced by its square or by its absolute value, the important point is that its evaluation be nonlinear, and here nonlinearity enters via u. The presence of conditional rather than unconditional means reflects the sequential nature of the setting. With regard to the $\sqrt{n}$-weighting, as is familiar from discussions of the classical LLN and CLT, the scaling by $\frac{1}{n}$ implies that in large samples "too little" weight is given to volatility (e.g. variance) relative to mean. Roughly, as described further at the end of this section, the above specification *models a decision-maker who takes into account both mean and variance even asymptotically.*

Remark: As is familiar, a Savage act (random variable) defined over a state space that is endowed with a probability measure induces a lottery over outcomes. Similarly here, any strategy θ induces, via P, a multistage lottery, from which it follows that θ can be viewed as describing the sequential (or contingent) choice from a set of repeated lotteries.

Admittedly, the specification (5) is ad hoc in the sense of (currently) lacking axiomatic foundations. We propose it because it seems plausible and it delivers novel results. In addition, we are not aware of any other model of preference over random payoff streams of arbitrary finite length that has axiomatic foundations and that has something interesting to say in our context. The special case of (5) where u is additively separable and linear in its second argument (example (u.1) below) can be axiomatized, but imposes a priori that only means matter asymptotically when choosing between arms and hence is too special (Theorem 3(v)). Take the further special case where u is also linear in its first argument but where payoffs are denominated in utils. This is the expected additive utility model (discounting can be added) that is the workhorse model in economics. However, it does not work well in our setting, for example, in the applied contexts in the introduction. We take the underlying payoffs or rewards at each stage to be objective quantities, such as the number of clicks or dollars. In all these cases, the relevant payoff when choosing a strategy is the sum of single stage payoffs, e.g. the total number of clicks, or in more formal terms, stage payoffs are perfect substitutes. However, discounted expected utility with nonlinear stage utility index models them as imperfect substitutes.

Utility has a particularly transparent form when $\theta = \theta^{\mu,\sigma}$ specifies choosing an arm described by the pair $\left(\mu, \sigma^2\right)$ repeatedly regardless of previous outcomes. In this case payoffs are i.i.d. with mean μ and variance σ^2. Thus the conditional expectation appearing in (5) equals μ, and the classical LLN and CLT imply that in the large horizon limit risk is described by the normal distribution $\mathbb{N}\left(0, \sigma^2\right)$ and

$$\lim_{n\to\infty} U_n\left(\theta^{\mu,\sigma}\right) = \int u\left(\mu, \cdot\right) d\mathbb{N}\left(0, \sigma^2\right). \tag{6}$$

Consequently, if $u\left(\mu, \cdot\right)$ is concave, then (asymptotic) risk aversion is indicated in the sense that

$$\lim_{n\to\infty} U_n\left(\theta^{\mu,\sigma}\right) \leq u\left(\mu, 0\right).$$

Here are examples of utility indices u and the implied utility functions U_n that will be referred to again in the sequel.

Example (utility indices)
(u.1) $u\left(x, y\right) = \varphi\left(x\right) + \alpha y$. Then

$$U_n\left(\theta\right) = E_P\left[\varphi\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^\theta\right)\right]$$

(u.2) $u\left(x, y\right) = \varphi\left(\left(1-\alpha\right)x + \alpha y\right)$, where $0 < \alpha \leq 1$. Then

$$U_n\left(\theta\right) = E_P\left[\varphi\left(\left(1-\alpha\right)\frac{1}{n}\sum_{i=1}^{n} Z_i^\theta + \alpha\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i^\theta - E_P[Z_i^\theta|\mathcal{H}_{i-1}^\theta]\right)\right)\right]$$

(u.3) (Mean-variance) $u\left(x, y\right) = x - \alpha y^2$, where $\alpha > 0$. Then

$$\begin{aligned} U_n\left(\theta\right) &= \frac{1}{n}E_P\left[\sum_{i=1}^{n} Z_i^\theta\right] - \alpha\frac{1}{n}Var_P\left[\sum_{i=1}^{n}\left(Z_i^\theta - E_P[Z_i^\theta|\mathcal{H}_{i-1}^\theta]\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(E_P\left[Z_i^\theta\right] - \alpha Var_P\left[Z_i^\theta - E_P[Z_i^\theta|\mathcal{H}_{i-1}^\theta]\right]\right), \end{aligned} \tag{7}$$

which is a form of the classic mean-variance specification for our setting.[3] For any arm with mean-variance pair $\left(\mu, \sigma^2\right)$ that is played repeatedly,

$$U_n\left(\theta^{\mu,\sigma}\right) = \mu - \alpha\sigma^2, \text{ for every } n. \tag{8}$$

(u.4) (Mean-semivariance) $u\left(x, y\right) = x - \alpha y^2 I_{(-\infty,0)}\left(y\right)$. Only negative cumulative deviations from (conditional) means are penalized. Then, given θ and letting $Y = \sum_{i=1}^{n}\left(Z_i^\theta - E_P[Z_i^\theta|\mathcal{H}_{i-1}^\theta]\right)$, $Var_P\left[Y\right]$ in (7) is replaced by the *semivariance* $E_P\left[Y^2 I_{Y<0}\right]$. If $\theta = \theta^{\mu,\sigma}$, then

$$U_n\left(\theta^{\mu,\sigma}\right) \underset{n\to\infty}{\longrightarrow} \mu - \alpha\int_{-\infty}^{0} y^2 d\mathbb{N}\left(0, \sigma^2\right) = \mu - \alpha\sigma^2/2.$$

[3] The second equality follows from the fact that, for $i \neq j$, $Z_i^\theta - E_P[Z_i^\theta|\mathcal{H}_{i-1}^\theta]$ and $Z_j^\theta - E_P[Z_j^\theta|\mathcal{H}_{j-1}^\theta]$ have zero covariance under P.

(u.5) (Shortfall penalty) $u(x,y) = x - \alpha I_{(-\infty,0)}(y)$. Only the existence of a shortfall, and not its size, matters. Then

$$\begin{aligned} U_n\left(\theta^{\mu,\sigma}\right) &= \mu - \alpha P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i^{\theta^{\mu,\sigma}} - E_P[Z_i^{\theta^{\mu,\sigma}}|\mathcal{H}_{i-1}^{\theta^{\mu,\sigma}}]\right) < 0\right) \qquad (9) \\ &\underset{n\to\infty}{\longrightarrow} \mu - \alpha\mathbb{N}_{(0,\sigma^2)}(-\infty,0) = \mu - \alpha/2. \end{aligned}$$

In particular, in the large horizon limit the utility of playing the single arm $\left(\mu,\sigma^2\right)$ repeatedly does not depend on the variance.

Remark: For the last 3 examples, horizon length drops out in the sense that maximizing $U_n(\theta)$ is equivalent to maximizing the modified objective function $U_n'(\theta)$ defined as in (5) except that both $\frac{1}{n}$ and $\frac{1}{\sqrt{n}}$ are deleted.

Our model of utility provides a (local) measure of risk aversion, or alternatively, of the mean-variance tradeoff, assuming that u is suitably differentiable (thus excluding examples (u.4) and (u.5)). Though it is a slight variant of the well-known Arrow-Pratt measure (Pratt, 1964), it might be worthwhile to derive it in our context. Consider a horizon equal to n stages and consider the choice for the last stage contingent on the history represented by (x,y), (partial sums corresponding to the two averages in (5)). Accordingly, DM uses the utility index $u(x+\cdot, y+\cdot)$ to evaluate the next step. Consider her evaluation of using the arm $\left(\epsilon^2\mu, \epsilon^2\sigma^2\right)$ for the final stage, where $\epsilon > 0$ has the effect, when small, of scaling down both the mean and variance of payoffs by ϵ^2. Using a second-order Taylor series approximation of $u(x+\cdot, y+\cdot)$ about $\epsilon = 0$, one obtains the expected utility

$$u(x,y) + \partial_x u(x,y)\frac{\epsilon^2\mu}{n} + \frac{1}{2}\partial_{yy}^2 u(x,y)\frac{\epsilon^2\sigma^2}{n}.$$

Therefore, if we let

$$\mu = \frac{-\frac{1}{2}\partial_{yy}^2 u(x,y)}{\partial_x u(x,y)}\sigma^2, \qquad (10)$$

then we can interpret $-\partial_{yy}^2 u(x,y)/\partial_x u(x,y)$ as approximating *twice the mean-variance ratio consistent with indifference to a small increase in risk.*

Two special cases are revealing. The measure of risk aversion is constant for the mean-variance model:

$$\frac{-\frac{1}{2}\partial_{yy}^2 u(x,y)}{\partial_x u(x,y)} = \alpha \text{ for all } (x,y).$$

(See Theorem 3 and the ensuing discussion for behavioral implications of this constancy.) Second, it is identically equal to 0 for (u.1), indicating risk neutrality in the sense defined by the measure, and this is so regardless of the curvature of φ. More generally, the measure does not involve $\partial_{xx}^2 u(x,y)$, contrary to what might be expected based on the Arrow-Pratt measure in expected utility theory. As an "explanation" for this possibly puzzling feature, we point out that $\partial_{xx}^2 u(x,y)$ would appear in a 2nd-order Taylor series expansion if the added mean-variance pair (or arm) were $\left(\epsilon\mu, \epsilon^2\sigma^2\right)$ instead of $\left(\epsilon^2\mu, \epsilon^2\sigma^2\right)$, and thus it is necessary to understand our choice of scaling.[4] The latter scaling fits and "works in" our model because the scale-invariant

[4] Our scaling may bring to mind the small risks modeled by Brownian motion for which both drift and variance are proportional to the time interval dt (identified here with ϵ^2).

mean-variance ratio matches the key hypothesis embedded in (5), that as n increases and the payoff at each stage is effectively a smaller gamble, neither the mean or the variance dominates.

2.3 Optimization and the value of a set of arms

Given a horizon of length n, DM solves the following optimization problem:

$$V_n \equiv \sup_{\theta \in \Theta} E_P U_n\left(\theta\right). \tag{11}$$

The finite horizon problem is generally not tractable, even when u has the special form (u.1). For reasons of tractability, Bayesian models in the literature typically take φ in (u.1) to be linear, reducing the problem to maximization of expected total rewards, but at the cost of assuming risk neutrality. Instead, we consider large horizons and approximate optimality. Then we can accommodate a much more general class of utility indices.

The first step in developing asymptotics is to define

$$V \equiv \lim_{n \to \infty} V_n. \tag{12}$$

Our first theorem proves that V is well-defined, that is, values have a limit, and more. (Below $||(x,y)||$ denotes the Euclidean norm.)

Theorem 1 *Let $u \in C(\mathbb{R}^2)$ and let payoffs to the K arms conform to (1), with $\underline{\sigma} \geq 0$. Suppose further that there exists $g \geq 1$ such that u satisfies the growth condition $|u(x,y)| \leq c(1 + ||(x,y)||^{g-1})$, and that payoffs satisfy $\sup_{1 \leq k \leq K} E_P[|X_k|^g] < \infty$. Then:*

(i) Values have a limit: $\lim_{n\to\infty} V_n$ exists.

(ii) Only means and variances matter: Consider another set of arms, described by the random payoffs X'_k, $1 \leq k \leq K'$, and denote the corresponding set of mean-variance pairs by $\mathcal{A}'$ and the corresponding values by V'_n and V'. Let the mean-variance pairs $\left(\mu'_k, \sigma'^2_k\right)$ be defined by the obvious counterpart of (1). Then

$$\mathcal{A}' = \mathcal{A} \implies V' = V.$$

Thus we can write

$$V = V\left(\mathcal{A}\ \right) = V\left(\left\{\left(\mu_k, \sigma_k^2\right) : 1 \leq k \leq K\right\}\right).$$

(iii) Extreme arms are enough:

$$V\left(\mathcal{A}\ \right) = V\left(\mathcal{A}^{ext}\right). \tag{13}$$

Remark: The assumption that u is continuous rules out example (u.4). However, because these functions can be approximated by continuous functions, the CLT (Proposition 6) and subsequently the above theorem, can be extended to cover them as well. (See Chen, Epstein and Zhang (2023, section A.3), for a similar extension from continuous functions to indicators.) Similarly for results below. Because the details are standard, we will ignore the discontinuity of (u.4).

The Appendix contains a proof of (i) and also gives two alternative expressions for the limit V. (ii) describes a simplification for the decision-maker afforded by adoption of the infinite-horizon heuristic - *she need only know and take into account the means and variances for each arm.* In addition, it permits identifying an arm with its mean-variance pair; thus we will often refer to a pair $\left(\mu, \sigma^2\right)$ as an arm. (iii) describes a further possible simplification for DM – she need only consider "extreme arms", that is, the extreme points of the convex polygon generated by $\mathcal{A}$. All other arms are redundant. For example, *given two arms* $\left(\mu_1, \sigma_1^2\right)$ *and* $\left(\mu_2, \sigma_2^2\right)$*, then any arm lying on the straight line between them has no value asymptotically even if it moderates large differences in the mean-variance characteristics of the two given arms.* For another implication of (iii), because $\mathcal{A}$ is contained in the rectangle defined by the four pairs on the right, one obtains that

$$V\left(\mathcal{A}\right) \leq V\left(\left\{(\overline{\mu}, \overline{\sigma}^2), (\overline{\mu}, \underline{\sigma}^2), (\underline{\mu}, \overline{\sigma}^2), (\underline{\mu}, \underline{\sigma}^2)\right\}\right).$$

Finally, note that both (ii) and (iii) are *true under weak (nonparametric) assumptions on u, for example, without any assumptions about monotonicity or risk attitudes. Therefore, they accommodate situations that feature targets, aspiration levels, loss aversion, and other deviations from the common assumption of global monotonicity and risk aversion.*

The sufficiency of means and variances might be expected from the classic CLT, and arises here for similar reasons.[5] We turn to intuition for (iii). Consider the evaluation of arm k in the context of making the contingent decision for stage i. If the horizon n is large, then the payoff to arm k contributes little to the averages determining overall utility. Accordingly, a second-order Taylor series expansion provides a good approximation to the incremental benefit from arm k, which expansion, to order $O\left(n^{-1}\right)$, is linear in $\left(\mu_k, \sigma_k^2\right)$. Therefore, the value when maximizing over the K arms (asymptotically) equals that when maximizing over the convex hull of $\mathcal{A}$, or over its set of extreme points $\mathcal{A}^{ext}$, as asserted in (13). *In more economic terms, extreme arms are sufficient because switching suitably between them across stages can, in the infinite-horizon limit, replicate or improve upon the payoff distribution achievable when choosing from the entire set of K arms.*

2.4 Strategies and the risk/reward tradeoff

Turn to strategies. Given the K arms corresponding to $\mathcal{A}$, the strategy θ^* is *asymptotically optimal* if

$$\lim_{n\rightarrow\infty} E_P U_n\left(\theta^*\right) = V\left(\mathcal{A}\right).$$

It follows that θ^* is *approximately optimal* for large horizons in that: for every $\epsilon > 0$, there exists n^* such that

$$\mid U_n\left(\theta^*\right) - V_n \mid < \epsilon \text{ if } n > n^*.$$

[5] This is not to say that the result can be derived from the classical CLT, or that it is in any way "obvious." Its proof is decidedly nontrivial.

Say that $\left(\mu,\sigma^2\right)$ is *feasible* if it lies in $\mathcal{A}$. Theorem 1(iii) states that DM can limit herself to strategies that choose between extreme arms. More can be said under added assumptions on the utility index and what is feasible, as illustrated by the next result.

Theorem 2 *Adopt the assumptions in Theorem 1. If $u(x,y)$ is increasing in x and concave in y, and if $(\overline{\mu},\underline{\sigma}^2)$ is feasible, then: the strategy of always choosing an arm exhibiting $(\overline{\mu},\underline{\sigma}^2)$ is asymptotically optimal, and the corresponding limiting value, defined in (12), is given by*

$$V=\int u\left(\overline{\mu},\cdot\right)d\mathbb{N}\left(0,\underline{\sigma}^2\right).$$

Intuition argues for the choice of $\left(\overline{\mu},\underline{\sigma}^2\right)$ at stage n if there are no later trials remaining, but may seem myopic more generally. Notably, the strategy of always choosing the high-mean/low-variance pair is not in general optimal given a finite horizon (even apart from the fact that arms may not be adequately characterized by mean and variance alone). That it is asymptotically optimal demonstrates a simplifying feature of the long-horizon heuristic. An additional comment is that one can similarly consider three other possible combinations of monotonicity and curvature assumptions for u, where each property is assumed to hold globally. For example, if $u(x,y)$ is decreasing in x and concave (convex) in y, then it is asymptotically optimal to always choose an arm exhibiting $(\underline{\mu},\underline{\sigma}^2)$ $((\underline{\mu},\overline{\sigma}^2))$ if it is feasible.

However, the theorem does not provide any insight into the risk/reward tradeoff that is at the core of decision-making under uncertainty. Under common assumptions about monotonicity and risk aversion, the tradeoff concerns the increase in mean reward needed to compensate the individual for facing an increase in risk (for example, a larger variance). But Theorem 2 assumes that there exists an arm having *both* the largest mean *and* the smallest variance, thus ruling out the need for DM to make such a tradeoff.

Next we investigate asymptotic optimality when the risk/reward tradeoff is integral. For greater clarity, we do so in a canonical setting where there are 2 arms ($K=2$), described by $\left(\mu_1,\sigma_1^2\right)$ and $\left(\mu_2,\sigma_2^2\right)$, and where

$$\mu_1>\mu_2,\ \ \sigma_1>\sigma_2\geq 0. \tag{14}$$

Parts (i) and (ii) of the next theorem describe conditions under which it is asymptotically optimal to *specialize* in one arm, that is, to choose that arm always (at every stage and history). The remaining parts give conditions under which specializing in one arm is not asymptotically optimal (that is, not even approximately optimal for large horizons). Some results are limited to utility specifications in the Example.

Theorem 3 *Adopt the assumptions in Theorem 1 and consider the 2-arm case above. Then, for each of the following specifications of u, the indicated strategy is asymptotically optimal and V denotes the corresponding limiting value defined in (12).*

(i) Let $u:\mathbb{R}^2\longrightarrow\mathbb{R}$ be twice continuously differentiable. Suppose that

$$\partial_x u\left(x,y\right)\left(\mu_1-\mu_2\right)+\tfrac{1}{2}\partial_{yy}^2 u\left(x,y\right)\left(\sigma_1^2-\sigma_2^2\right)\geq 0\ \text{ for all } (x,y)\in\mathbb{R}^2. \tag{15}$$

Then specializing in arm 1 always is asymptotically optimal and, (by (6)), $V = \int u(\mu_1, \cdot)\, d\mathbb{N}\left(0, \sigma_1^2\right)$. If $\partial_x u$ is everywhere positive, then (15) is equivalent to

$$\frac{-\frac{1}{2}\partial_{yy}^2 u(x,y)}{\partial_x u(x,y)} \leq \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2} \text{ for all } (x,y) \in \mathbb{R}^2. \tag{16}$$

When the inequality in (15) is reversed, then it is asymptotically optimal to specialize in arm 2.

(ii) Adopt the conditions on u in (i), and assume that $\partial_x u(x,y) > 0$ for all $(x,y) \in \mathbb{R}^2$. Suppose further that

$$\frac{-\frac{1}{2}\partial_{yy}^2 u}{\partial_x u} = \alpha > 0 \text{ for all } (x,y) \in \mathbb{R}^2. \tag{17}$$

Then specializing in arm 1 (arm 2) is asymptotically optimal if

$$\alpha \leq (\geq) \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2}. \tag{18}$$

Both strategies are asymptotically optimal when there is equality in (18).

(iii) Let $u(x,y) = x - \alpha y^2 I_{(-\infty,0)}(y), \alpha > 0$. Observe that

$$\frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2} < \underline{\alpha} < \overline{\alpha},$$

where the critical values $\underline{\alpha}$ and $\overline{\alpha}$ are given by[6]

$$\underline{\alpha} \equiv \frac{4(\mu_1 - \mu_2)}{3(\sigma_1^2 - \sigma_2^2)}, \overline{\alpha} \equiv \frac{2(\mu_1 - \mu_2)}{\sigma_2(\sigma_1 - \sigma_2)}.$$

If $\alpha \leq \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2}$, then specializing in arm 1 is asymptotically optimal. If $\underline{\alpha} < \alpha$ (respectively $\alpha < \overline{\alpha}$), then specializing in arm 1 (arm 2) is *not* asymptotically optimal, from which it follows that specialization in *either* arm is *not* asymptotically optimal if $\underline{\alpha} < \alpha < \overline{\alpha}$.

(iv) Let $u(x,y) = x - \alpha I_{(-\infty,0)}(y)$, $\alpha > 0$. Let $\sigma_2 > 0$. Then specializing in arm 2 is *not* asymptotically optimal for any α, and, if

$$\underline{\alpha}' \equiv \frac{2(\mu_1 - \mu_2)\sigma_1}{(\sigma_1 - \sigma_2)} < \alpha,$$

then neither is specializing in arm 1.

(v) Let $u(x,y) = \varphi(x) + \alpha y$, $\varphi \in C(\mathbb{R})$ and $\alpha \in \mathbb{R}$. Fix $x^* \in \arg \max_{\mu_1 \leq x \leq \mu_2} \varphi(x)$, and let $\lambda \in [0,1]$ be such that $x^* = \lambda\mu_1 + (1-\lambda)\mu_2$. Denote by ψ_i the number times that arm 1 is chosen in first i stages. Let the strategy θ^* choose arm 1 at stage 1, and also at stage $i+1$, $(i \geq 1)$, if and only if $\frac{\psi_i}{i} \leq \lambda$. Then θ^* is asymptotically optimal and

$$V = \max_{\mu_2 \leq x \leq \mu_1} \varphi(x).$$

Further, specializing in one arm is asymptotically optimal if and only if $\max\{\varphi(\mu_1), \varphi(\mu_2)\} = \max_{\mu_2 \leq x \leq \mu_1} \varphi(x)$.

[6] $\overline{\alpha} = \infty$ if $\sigma_2 = 0$.

Remark: It is straightforward to extend the theorem to an arbitrary set of K arms. For example, in (i), with $\partial_x u$ everywhere positive, specializing in arm j is asymptotically optimal if

$$j \in \arg\max_{k=1,\ldots,K}\{\mu_k - (\tfrac{-\frac{1}{2}\partial^2_{yy}u(x,y)}{\partial_x u})\sigma^2_k\} \text{ for all } (x,y)\,,$$

which simplifies in the obvious way under the constancy condition (17).

We discuss each part of the theorem in turn.

(i) Focus on (16). Intuition derives from interpretation given above of $-\partial^2_{yy}u/\partial_x u$ as a (local) measure of risk aversion.The relatively small degree of risk aversion indicated in (16) implies that the larger mean for arm 1 more than compensates for its larger variance. Moreover, this is true at each stage, regardless of history, because the inequality in (16) is satisfied globally.

(ii) This is an immediate consequence of (i) that we include in the statement because the consequence of the indicated constancy warrants emphasis. Two examples covered by this constancy are mean-variance and the special case of (u.2) where φ is an exponential.At first glance, the implication regarding the *unimportance of diversification* might seem surprising, especially given its central role in portfolio theory. Of course, diversification in portfolio theory refers to the simultaneous holding of several assets, which, interpreting each arm as an asset, is excluded here. But diversification over time is permitted and that is its meaning here. The result that specialization in one arm over time is always asymptotically optimal given (17) can be understood as follows. Considering the factors that might lead to different arms being chosen at two different stages, note first that the payoff distribution for each arm is unchanged by assumption. Second, though a finite-horizon induces a nonstationarity that can affect choices, our decision-maker is, roughly speaking, acting as if solving an infinite-horizon problem. That leaves only the variation of risk attitude with past outcomes, which is excluded if $-\partial^2_{yy}u/\partial_x u$ is constant.

(iii) Note first that it has often been argued, including by Markowitz (1959), that investors are more concerned with downside risk than with variance, and hence that *semivariance is a better measure of the relevant risk.* In our sequential choice context, the mean-semivariance model agrees partially with the mean-variance model in that for both (the inequality $\leq$ in) (18) implies the asymptotic optimality of choosing (the high mean, high variance) arm 1 throughout. However, their agreement ends there. In particular, there is *a role for time-diversification* for the semivariance model, in that, for $\underline{\alpha} < \alpha < \overline{\alpha}$, asymptotic optimality can be achieved only by a strategy that employs both arms. (In particular, if arm 2 is risk-free ($\sigma_2 = 0$), then time-diversification is necessary for asymptotic optimality if $\frac{3}{4}\alpha$ exceeds the risk-adjusted excess mean $(\mu_1 - \mu_2)/\sigma^2_1$.) Here is some intuition for the existence of a region with nonspecialization. Since only negative deviations are penalized, it is as though DM faces, or perceives, less risk than what is measured by σ^2. Alternatively, in our preferred interpretation, for any given risk measured by variance, DM is less averse to that risk in the present model as if her effective α is smaller than its nominal magnitude. Moreover, risk aversion varies across stages. For example, contingent on cumulative past deviations being positive (negative) at stage m, it is relatively unlikely

(likely) that future choices will lead later to negative cumulative deviations, and thus variance is less (more) of a concern. Such *endogenous changes in risk aversion* can lead to specialization in either single arm being dominated in large horizons.

In finance, it has been argued (Nantell and Price 1979; Klebaner et al 2017) that the change from variance to semivariance has limited consequences for received asset market theory. In contrast, a similar change in the bandit problem context leads to qualitative differences regarding the importance of time-diversification.

(iv) This utility specification, for which only the existence of a shortfall and not its size matters, implies that *it is never asymptotically optimal to specialize in the low mean, low variance arm.*[7] Indeed, by (9), specializing in the high mean, high variance arm is superior for large horizons without any regard to the numerical magnitudes of $\mu_1 - \mu_2$ and $\sigma_1^2 - \sigma_2^2$. However, specializing in the high mean, high variance arm is also ruled out for large enough α - those lying in $(\underline{\alpha}', \infty)$. Note that this set grows larger as σ_1 increases (keeping μ_1, μ_2 and σ_2 fixed) - a larger variance makes it more likely that repeated choice of arm 1 will produce a cumulative shortfall, which is tolerable only if the associated penalty parameter α is smaller. Therefore, as in the semivariance model (iii), for a range of parameter values *asymptotic optimality can be achieved only through diversification across time.*

(v) The utility specification $u(x, y) = \varphi(x) + \alpha y$ leads to an asymptotically optimal strategy that is diversified and that can be described explicitly. Condition (15) suggests that either nonmonotonicity (e.g. a change in the sign of $\partial_x u$), or variable risk aversion (e.g. a change in the sign or magnitude of $\partial_{yy}^2 u$) might lead to the asymptotic optimality of switching between arms. This utility specification, with φ not necessarily monotonic, illustrates the former case. The interpretation of the strategy θ^* defined in the theorem is that *DM targets x^*, a maximizer of φ* on $[\mu_2, \mu_1]$. (When $(\mu_2 + \mu_1)/2$ is a maximizer, then θ^* chooses arms according to the sequence 121212.... When φ is monotonic, θ^* specializes in arm 1 or in arm 2 according as φ is increasing or decreasing on $[\mu_2, \mu_1]$, respectively.) Irrespective of any nonlinearity of φ, and the implied non-neutrality to risk, *variances do not matter asymptotically as in the classic LLN.*

3 Concluding Comments

Our model has produced new results regarding sequential choice between repeated gambles, most notably in describing connections, expressed in simple formal terms, between the endogeneity of risk aversion and the value of time-diversification. Three features of the model that facilitate tractability are (i) the heuristic of approximate optimality for large horizons, which is the decision-maker's assumed response to a complex problem; (ii) the existence of a suitable measure of risk attitude (similar to, but distinct from, the Arrow-Pratt measure) that describes her risk/reward tradeoff; and (iii) the fact that without loss of generality gambles can be represented by their mean and variance alone, thus providing a new rationale for mean-variance analysis.

[7] Intuitively, relying exclusively on the more conservative arm increases the asymptotic likelihood of cumulative shortfalls. The problem is reminiscent of the classic introductory story in Dubins and Savage (1976, Ch. 1) of a gambler who must decide how to gamble in order to minimize the probability that cumulative winnings fall short of a fixed target. Their solution is that he should not gamble cautiously.

The results are general in the sense that payoff distributions are unrestricted except for the requirement that means and variances exist. However, results depend on our nonstandard specification of payoffs (via averages) and utility function. Specific assumptions are needed in order to derive analytical results, and our assumption compares favourably, in our view, with the assumption of risk neutrality adopted in much of the bandit literature. Given the complexity of dynamic decision problems under uncertainty, it is natural to wonder if behavior might be better described by "approximate optimality," and we view the paper as a modest first step in this modeling direction. Undoubtedly more needs to be done. Axiomatic analysis for such behavior poses an interesting challenge for decision theorists.

A Appendix: Proofs

We remind the reader of the following notation used in this section: $\overline{\mu}, \underline{\mu}$ and $\overline{\sigma}^2, \underline{\sigma}^2$ are the bounds of means and variances given in (2), $\mathcal{A}$ denotes the set of mean-variance pairs of all K arms, and $\mathcal{A}^{ext} \subset \mathcal{A}$ denotes the set of extreme points of $co\,(\mathcal{A})$. Pairs consisting of mean and standard deviation (rather than variance) will also be important, and thus it is convenient to define

$$\begin{aligned}[\mathcal{A}] &= \{(\mu,\sigma) : (\mu,\sigma^2) \in \mathcal{A}\}, \text{ and} \\ [\mathcal{A}]^{ext} &= \{(\mu,\sigma) : (\mu,\sigma^2) \in \mathcal{A}^{ext}\}\end{aligned}$$

Let $B = \{B_t = (B_t^{(1)}, B_t^{(2)})\}$ be a two-dimensional standard Brownian motion defined on $(\Omega, \mathcal{F}, P)$, and let $\{\mathcal{F}_t\}$ be the natural filtration generated by (B_t). For a fixed $T > 0$, and any $0 \leq t \leq s \leq T$, let $[\mathcal{A}](t,T)$ denote the set of all $\{\mathcal{F}_s\}$-progressively measurable processes, $a = \{a_s = (a_s^{(1)}, a_s^{(2)})\} : [t,T] \times \Omega \to [\mathcal{A}] \subset \mathbb{R}^2$. Finally, $[\mathcal{A}]^{ext}(t,T)$ is defined similarly by restricting the images of each process a to lie in $[\mathcal{A}]^{ext}$.

The following lemma gives properties of $\{Z_n^\theta\}$ that will be used repeatedly.

Lemma 4 *The rewards $\{Z_n^\theta : n \geq 1\}$ defined in (4) satisfy the following:*

(1) *For any $n \geq 1$,*

$$\begin{aligned}\overline{\mu} &= ess\sup_{\theta\in\Theta} E_P[Z_n^\theta | \mathcal{H}_{n-1}^\theta], \ \underline{\mu} = ess\inf_{\theta\in\Theta} E_P[Z_n^\theta | \mathcal{H}_{n-1}^\theta] \\ \overline{\sigma}^2 &= ess\sup_{\theta\in\Theta} E_P\left[\left(Z_n^\theta - E_P[Z_n^\theta|\mathcal{H}_{n-1}^\theta]\right)^2 |\mathcal{H}_{n-1}^\theta\right] \\ \underline{\sigma}^2 &= ess\inf_{\theta\in\Theta} E_P\left[\left(Z_n^\theta - E_P[Z_n^\theta|\mathcal{H}_{n-1}^\theta]\right)^2 |\mathcal{H}_{n-1}^\theta\right].\end{aligned}$$

(2) *For any $\theta \in \Theta$ and $n \geq 1$, let U_{n-1}^θ be any $\mathcal{H}_{n-1}^\theta$-measurable random variable. For any bounded measurable functions f_0, f_1 and f_2 on $\mathbb{R}$, let $\psi(x,y) = f_0(x) + f_1(x)y + f_2(x)y^2, (x,y) \in \mathbb{R}^2$. Then*

$$\sup_{\theta\in\Theta} E_P\left[\psi\left(U_{n-1}^\theta, Z_n^\theta\right)\right] = \sup_{\theta\in\Theta} E_P\left[\max_{1\leq k\leq K}\left\{\psi_k(U_{n-1}^\theta)\right\}\right]$$

where, for all $x \in\mathbb{R}$ and $1 \leq k \leq K$,

$$\psi_k(x) = E_P[\psi(x, X_{k,n})] = f_0(x) + \mu_k f_1(x) + (\mu_k^2 + \sigma_k^2)\, f_2(x). \tag{19}$$

Proof: (1) $\{Z_n^\theta\}$ satisfy, for any $\theta \in \Theta$ and $n \geq 1$,

$$\begin{aligned} E_P[Z_n^\theta|\mathcal{H}_{n-1}^\theta] &= \sum_{k=1}^K I_{\{\theta_n=k\}} E_P[X_{k,n}|\mathcal{H}_{n-1}^\theta] \\ &= \sum_{k=1}^K I_{\{\theta_n=k\}} E_P[X_{k,n}] = \sum_{k=1}^K I_{\{\theta_n=k\}} \mu_k. \end{aligned}$$

Combine with the definitions of $\overline{\mu}$ and $\underline{\mu}$ in (2) to derive

$$ess \sup_{\theta\in\Theta} E_P[Z_n^\theta|\mathcal{H}_{n-1}^\theta] = \overline{\mu}, \ ess \inf_{\theta\in\Theta} E_P[Z_n^\theta|\mathcal{H}_{n-1}^\theta] = \underline{\mu}.$$

The other two equalities can be proven similarly.

(2) For any $\theta \in \Theta$ and $n \geq 1$, let U_{n-1}^θ be a $\mathcal{H}_{n-1}^\theta$-measurable random variable. By direct calculation we obtain that

$$\begin{aligned} &\sup_{\theta\in\Theta} E_P\left[\psi\left(U_{n-1}^\theta, Z_n^\theta\right)\right] \\ =& \sup_{\theta\in\Theta} E_P\left[\sum_{k=1}^K I_{\{\theta_n=k\}} E_P[\psi\left(U_{n-1}^\theta, X_{k,n}\right)|\mathcal{H}_{n-1}^\theta]\right] \\ =& \sup_{\theta\in\Theta} E_P\left[\max_{1\leq k\leq K} \psi_k(U_{n-1}^\theta)\right], \end{aligned}$$

where ψ_k is given in (19). ■

Following Peng (2019), our arguments make use of nonlinear partial differential equations (PDEs) and viscosity solutions. The following is taken from Theorems 2.1.2, C.3.4 and C.4.5 in Peng's book.

Lemma 5 *For given $T > 0$, consider the following PDE:*

$$\begin{cases} \partial_t v(t,x,y) + G\left(\partial_x v(t,x,y), \partial_{yy}^2 v(t,x,y)\right) = 0, & (t,x,y) \in [0,T) \times \mathbb{R}^2 \\ v(T,x,y) = u(x,y), \end{cases} \tag{20}$$

where $u \in C(\mathbb{R}^2)$. Suppose that G is continuous on $\mathbb{R}^2$ and satisfies the following conditions, for all $(p,q),(p',q') \in \mathbb{R}^2$:

$$G(p,q) \leq G(p,q'), \quad \textit{whenever } q \leq q', \tag{21}$$

$$G(p,q) - G(p',q') \leq G(p-p', q-q'), \tag{22}$$

$$G(\lambda p, \lambda q) = \lambda G(p,q), \quad \textit{for } \lambda \geq 0. \tag{23}$$

Then, for any $u \in C(\mathbb{R}^2)$ satisfying a polynomial growth condition, there exists a unique $v \in C([0,T] \times \mathbb{R}^2)$ such that v is a viscosity solution of the PDE (20). Moreover, if $\exists\lambda > 0$ such that, for all $p \in \mathbb{R}$ and $q \geq q' \in \mathbb{R}$,

$$G(p,q) - G(p,q') \geq \lambda(q-q'),$$

and if the initial condition u is uniformly bounded, then for each $0 < \epsilon < T$, $\exists\beta \in (0,1)$ such that

$$\|v\|_{C^{1+\beta/2,2+\beta}([0,T-\epsilon]\times\mathbb{R}^2)} < \infty. \tag{24}$$

Here $\|\cdot\|_{C^{1+\beta/2,2+\beta}([0,T-\epsilon]\times\mathbb{R}^2)}$ is the Krylov (1987) norm on $C^{1+\beta/2,2+\beta}([0,T-\epsilon]\times\mathbb{R}^2)$, the set of (continuous and) suitably differentiable functions on $[0,T-\epsilon]\times\mathbb{R}^2$.[8]

[8] Some detail is provided in the Online Appendix. See also Peng (2019, Ch. 2.1).

A.1 Proof of Theorem 1

We first prove a nonlinear central limit theorem for the bandit problem. The values V_n and V are defined in (11) and (12) respectively.

Proposition 6 (CLT) *Let $u \in C_{b,Lip}(\mathbb{R}^2)$, the class of all bounded and Lipschitz continuous functions on $\mathbb{R}^2$, and adopt all other assumptions and the notation in Theorem 1. Then*

$$\lim_{n\to\infty} V_n = V = \sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right] \tag{25}$$

$$= \sup_{a\in[\mathcal{A}]^{ext}(0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right]. \tag{26}$$

The proof in this appendix assumes $\underline{\sigma}> 0$. The extension to $\underline{\sigma}= 0$ is proven in the Online Appendix. The boundedness assumption on utility indices excludes many interesting specifications. However, the Online Appendix shows that the Proposition is valid for all $u \in C\left(\mathbb{R}^2\right)$ satisfying a growth condition.

The following immediate corollary is used frequently in later proofs of Theorems 2 and 3 (the Online Appendix contains a proof).

Corollary 7 *For all $u \in C\left(\mathbb{R}^2\right)$ satisfying a polynomial growth condition, the limit in (25) can be described also by the solution of a PDE. Specifically,*

$$V = v(0,0,0), \tag{27}$$

where v is the solution of PDE (20), with function G given by

$$G(p,q) = \sup_{(\mu,\sigma^2)\in\mathcal{A}} \left[\mu p + \tfrac{1}{2}\sigma^2 q\right], \quad (p,q) \in \mathbb{R}^2. \tag{28}$$

Some related literature on CLTs was cited in the introduction. In addition, Chen and Epstein (2022) and Chen, Epstein and Zhang (2022) have nonlinear CLTs, which, when translated into the bandits context, restrict differences between arms either by assuming that they all have the identical variance (in the former paper), or the identical mean (in the latter paper). These restrictions preclude study of the risk/reward tradeoff. In addition, their objective is to obtain simple closed-form expressions for the limit (what we denote by V), and for that purpose they adopt very special functional forms for u.[9] In contrast, Proposition 6 and its corollary apply to a much more general class of utility indices. Moreover, as this paper shows, in spite of the complexity of the expression for V it is the basis for a range of results about the bandit problem even allowing unrestricted heterogeneity across arms.

Next we proceed with lemmas that will lead to a proof of the CLT. They assume $u \in C_b^3(\mathbb{R}^2)$ and relate to the functions $\{H_t\}_{t\in[0,1]}$ defined by, for all $(x,y) \in \mathbb{R}^2$,

$$H_t(x,y) = \sup_{a\in[\mathcal{A}](t,1+h)} E_P\left[u\left(x + \int_t^{1+h} a_s^{(1)}ds, y + \int_t^{1+h} a_s^{(2)}dB_s^{(2)}\right)\right], \tag{29}$$

where $h > 0$ is fixed and dependence on h is suppressed notationally. In addition, we often write $z = (z_1, z_2) = (x, y)$ and define $|z - z'|^\beta = |z_1 - z_1'|^\beta + |z_2 - z_2'|^\beta$.

[9] In particular, the second paper cited assumes $u\,(x,y) = \varphi\,(y)$, where $\varphi\,(y) = \varphi_1\,(y-c)$ if $y \geq c$, and $= -\lambda^{-1}\varphi_1\,(-\lambda(y-c))$ if $y < c$, for some function φ_1 and $c \in \mathbb{R}$. This functional form is motivated by loss aversion, but from the perspective of this paper is very special.

Lemma 8 *The functions $\{H_t\}_{t\in[0,1]}$ satisfy the following properties:*

(1) $H_t \in C_b^2(\mathbb{R}^2)$ *and the first and second derivatives of H_t are uniformly bounded for all $t \in [0,1]$.*

(2) *There exist constants $L > 0$ and $\beta \in (0,1)$, independent of t, such that for any $(z_1, z_2), (z_1', z_2') \in \mathbb{R}^2$,*

$$|\partial^2_{z_i z_j} H_t(z_1, z_2) - \partial^2_{z_i z_j} H_t(z_1', z_2')| \le L(|z_1 - z_1'|^\beta + |z_2 - z_2'|^\beta), \quad i, j = 1, 2.$$

(3) *Dynamic programming principle: For any $\delta \in [0, 1+h-t]$,*

$$H_t(x,y) = \sup_{a\in[\mathcal{A}](t,t+\delta)} E_P\left[H_{t+\delta}\left(x + \int_t^{t+\delta} a_s^{(1)} ds, y + \int_t^{t+\delta} a_s^{(2)} dB_s^{(2)}\right)\right], \ (x,y) \in \mathbb{R}^2.$$

(4) *For the function G given in (28), we have*

$$\lim_{n\to\infty} \sum_{m=1}^n \sup_{(x,y)\in\mathbb{R}^2} \left| H_{\frac{m-1}{n}}(x,y) - H_{\frac{m}{n}}(x,y) - \frac{1}{n} G\left(\partial_x H_{\frac{m}{n}}(x,y), \partial^2_{yy} H_{\frac{m}{n}}(x,y)\right)\right| = 0.$$

(5) *There exists a constant $C_0 > 0$ such that*

$$\sup_{(x,y)\in\mathbb{R}^2} |H_1(x,y) - u(x,y)| \le C_0 h$$

$$\sup_{(x,y)\in\mathbb{R}^2} |H_0(x,y) - \psi(x,y)| \le C_0 h,$$

where $\psi(x,y) = \sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(x + \int_0^1 a_s^{(1)} ds, y + \int_0^1 a_s^{(2)} dB_s^{(2)}\right)\right]$.

Proof: For any $t \in [0, 1+h]$ and $(x,y) \in \mathbb{R}^2$, we define the function $v(t,x,y) = H_t(x,y)$. Then v is the solution of the HJB-equation (20) with function G given in (28) (Yong and Zhou (1999, Theorem 5.2, Ch. 4)). By Lemma 5, $\exists \beta \in (0,1)$ such that

$$\|v\|_{C^{1+\beta/2, 2+\beta}([0,1]\times\mathbb{R}^2)} < \infty.$$

This proves both (1) and (2).

(3) follows directly from the classical dynamic programming principle (Yong and Zhou (1999, Theorem 3.3, Ch. 4)).

Prove (4): By Ito's formula,

$$\begin{aligned}
&\sum_{m=1}^n \sup_{(x,y)\in\mathbb{R}^2} \left| H_{\frac{m-1}{n}}(x,y) - H_{\frac{m}{n}}(x,y) - \frac{1}{n} G\left(\partial_x H_{\frac{m}{n}}(x,y), \partial^2_{yy} H_{\frac{m}{n}}(x,y)\right)\right| \\
=& \sum_{m=1}^n \sup_{(x,y)\in\mathbb{R}^2} \left| \sup_{\alpha\in[\mathcal{A}](\frac{m-1}{n}, \frac{m}{n})} E_P\left[H_{\frac{m}{n}}\left(x + \int_{\frac{m-1}{n}}^{\frac{m}{n}} a_s^{(1)} ds, y + \int_{\frac{m-1}{n}}^{\frac{m}{n}} a_s^{(2)} dB_s^{(2)}\right)\right] \right. \\
&\qquad \left. - H_{\frac{m}{n}}(x,y) - \frac{1}{n} G\left(\partial_x H_{\frac{m}{n}}(x,y), \partial^2_{yy} H_{\frac{m}{n}}(x,y)\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{n} \sup_{(x,y)\in\mathbb{R}^2} \left| \sup_{\alpha\in[\mathcal{A}](\frac{m-1}{n},\frac{m}{n})} E_P \left[\int_{\frac{m-1}{n}}^{\frac{m}{n}} \partial_x H_{\frac{m}{n}} \left(x + \int_{\frac{m-1}{n}}^{s} a_s^{(1)} ds, y + \int_{\frac{m-1}{n}}^{s} a_s^{(2)} dB_s^{(2)} \right) a_s^{(1)} ds \right.\right. \\
&\qquad\qquad \left. + \frac{1}{2} \int_{\frac{m-1}{n}}^{\frac{m}{n}} \partial_{yy}^2 H_{\frac{m}{n}} \left(x + \int_{\frac{m-1}{n}}^{s} a_s^{(1)} ds, y + \int_{\frac{m-1}{n}}^{s} a_s^{(2)} dB_s^{(2)} \right) (a_s^{(2)})^2 ds \right] \\
&\qquad\qquad\qquad \left. - \frac{1}{n} G\left(\partial_x H_{\frac{m}{n}}(x,y), \partial_{yy}^2 H_{\frac{m}{n}}(x,y) \right) \right| \\
&\leq \frac{C}{n} \sum_{m=1}^{n} \sup_{z\in\mathbb{R}^2} \left| \sup_{\alpha\in[\mathcal{A}](\frac{m-1}{n},\frac{m}{n})} E_P \left[\sup_{s\in[\frac{m-1}{n},\frac{m}{n}]} \left(\left| \int_{\frac{m-1}{n}}^{s} a_s^{(1)} ds \right| + \left| \int_{\frac{m-1}{n}}^{s} a_s^{(2)} dB_s^{(2)} \right| \right) \right.\right. \\
&\qquad\qquad \left.\left. + \sup_{s\in[\frac{m-1}{n},\frac{m}{n}]} \left(\left| \int_{\frac{m-1}{n}}^{s} a_s^{(1)} ds \right|^{\beta} + \left| \int_{\frac{m-1}{n}}^{s} a_s^{(2)} dB_s^{(2)} \right|^{\beta} \right) \right] \right| \\
&\quad \to 0, \quad \text{as } n \to \infty,
\end{aligned}
$$

where C is a constant that depends only on $\overline{\mu}, \underline{\mu}, \overline{\sigma}^2$, the uniform bound of $\partial_{xx}^2 H_t, \partial_{xy}^2 H_t$, and constant L in (2).

Prove (5): Use Ito's formula to check that

$$
\begin{aligned}
&\sup_{(x,y)\in\mathbb{R}^2} |H_1(x,y) - u(x,y)| \\
&= \sup_{(x,y)\in\mathbb{R}^2} \left| \sup_{a\in[\mathcal{A}](1,1+h)} E_P \left[\int_1^{1+h} \partial_x u \left(x + \int_1^s a_s^{(1)} ds, y + \int_1^s a_s^{(2)} dB_s^{(2)} \right) a_s^{(1)} ds \right.\right. \\
&\qquad\qquad \left.\left. + \frac{1}{2} \int_1^{1+h} \partial_{yy}^2 u \left(x + \int_1^s a_s^{(1)} ds, y + \int_1^s a_s^{(2)} dB_s^{(2)} \right) (a_s^{(2)})^2 ds \right] \right| \\
&\leq C_0 h,
\end{aligned}
$$

where the constant C_0 depends only on $\overline{\mu}, \underline{\mu}, \overline{\sigma}^2$ and the uniform bound of $\partial_x u, \partial_{yy}^2 u$.

Similarly, we can prove that $\sup_{(x,y)\in\mathbb{R}^2} |H_0(x,y) - \psi(x,y)| \leq C_0 h$. ■

Lemma 9 *Take G to be the function defined in (28), let $\{H_t\}_{t\in[0,1]}$ be the functions defined in (29), and define $\{L_{m,n}\}_{m=1}^n$ by*[10]

$$
L_{m,n}(z) = H_{\frac{m}{n}}(z) + \frac{1}{n} G\left(\partial_{z_1} H_{\frac{m}{n}}(z), \partial_{z_2 z_2}^2 H_{\frac{m}{n}}(z) \right), \ z \in \mathbb{R}^2. \tag{30}
$$

For any $\theta \in \Theta$ and $n \geq 1$, define

$$
S_n^{\theta} = \sum_{i=1}^{n} Z_i^{\theta}, \quad \overline{S}_n^{\theta} = \sum_{i=1}^{n} \overline{Z}_i^{\theta}, \quad \overline{Z}_n^{\theta} = Z_n^{\theta} - E_P[Z_n^{\theta} | \mathcal{H}_{n-1}^{\theta}].
$$

Then

$$
\lim_{n\to\infty} \sum_{m=1}^{n} \left| \sup_{\theta\in\Theta} E_P \left[H_{\frac{m}{n}} \left(\frac{S_m^{\theta}}{n}, \frac{\overline{S}_m^{\theta}}{\sqrt{n}} \right) \right] - \sup_{\theta\in\Theta} E_P \left[L_{m,n} \left(\frac{S_{m-1}^{\theta}}{n}, \frac{\overline{S}_{m-1}^{\theta}}{\sqrt{n}} \right) \right] \right| = 0. \tag{31}
$$

[10] Again, $z = (z_1, z_2) = (x, y)$.

Proof: We need only prove

$$\lim_{n\to\infty}\sum_{m=1}^{n}\left|\sup_{\theta\in\Theta}E_P\left[H_{\frac{m}{n}}\left(\frac{S_m^\theta}{n},\frac{\overline{S}_m^\theta}{\sqrt{n}}\right)\right]-e(m,n)\right|=0\ \text{ and } \tag{32}$$

$$\lim_{n\to\infty}\sum_{m=1}^{n}\left|e(m,n)-\sup_{\theta\in\Theta}E_P\left[L_{m,n}\left(\frac{S_{m-1}^\theta}{n},\frac{\overline{S}_{m-1}^\theta}{\sqrt{n}}\right)\right]\right|=0, \tag{33}$$

where $e(m,n)$ is given by

$$\begin{aligned}e(m,n)=&\sup_{\theta\in\Theta}E_P\left[H_{\frac{m}{n}}\left(\frac{S_{m-1}^\theta}{n},\frac{\overline{S}_{m-1}^\theta}{\sqrt{n}}\right)+\partial_{z_1}H_{\frac{m}{n}}\left(\frac{S_{m-1}^\theta}{n},\frac{\overline{S}_{m-1}^\theta}{\sqrt{n}}\right)\frac{Z_m^\theta}{n}\right.\\&\left.+\partial_{z_2}H_{\frac{m}{n}}\left(\frac{S_{m-1}^\theta}{n},\frac{\overline{S}_{m-1}^\theta}{\sqrt{n}}\right)\frac{\overline{Z}_m^\theta}{\sqrt{n}}+\partial^2_{z_2z_2}H_{\frac{m}{n}}\left(\frac{S_{m-1}^\theta}{n},\frac{\overline{S}_{m-1}^\theta}{\sqrt{n}}\right)\frac{(\overline{Z}_m^\theta)^2}{2n}\right].\end{aligned}$$

By Lemma 8, parts (1) and (2), $\exists C>0$, $\beta\in(0,1)$ such that

$$\sup_{t\in[0,1]}\sup_{z\in\mathbb{R}^2}|\partial^2_{z_iz_j}H_t(z)|\le C,$$

$$\sup_{t\in[0,1]}\sup_{z,z'\in\mathbb{R}^2,z\neq z'}\frac{|\partial^2_{z_iz_j}H_t(z)-\partial^2_{z_iz_j}H_t(z')|}{|z-z'|^\beta}\le C,\ \ i,j=1,2.$$

It follows from Taylor's expansion that $\forall\epsilon>0\ \exists\delta>0$ (depending only on C and ϵ), such that $\forall z,z'\in\mathbb{R}^2$, and $\forall t\in[0,1]$,[11]

$$\begin{aligned}&\left|H_t(z+z')-H_t(z)-D_zH_t(z)z'-\tfrac{1}{2}tr\left(z'^\top D_z^2H_t(z)z'\right)\right|\\\le&\epsilon|z'|^2I_{\{|z'|<\delta\}}+2C|z'|^2I_{\{|z'|\ge\delta\}}.\end{aligned}\tag{34}$$

Set $z=\left(\frac{S_{m-1}^\theta}{n},\frac{\overline{S}_{m-1}^\theta}{\sqrt{n}}\right)$ and $z'=\left(\frac{Z_m^\theta}{n},\frac{\overline{Z}_m^\theta}{\sqrt{n}}\right)$. Use (34) to obtain

$$\begin{aligned}&\sum_{m=1}^{n}\left|\sup_{\theta\in\Theta}E_P\left[H_{\frac{m}{n}}\left(\frac{S_m^\theta}{n},\frac{\overline{S}_m^\theta}{\sqrt{n}}\right)\right]-e(m,n)\right|\\\le&\frac{C}{2}\sum_{m=1}^{n}\sup_{\theta\in\Theta}E_P\left[\left|\frac{Z_m^\theta}{n}\right|^2+\left|\frac{Z_m^\theta}{n}\right|\left|\frac{\overline{Z}_m^\theta}{\sqrt{n}}\right|\right]\\&+\epsilon\sum_{m=1}^{n}\sup_{\theta\in\Theta}E_P\left[\left(\left|\frac{Z_m^\theta}{n}\right|^2+\left|\frac{\overline{Z}_m^\theta}{\sqrt{n}}\right|^2\right)I_{\left\{\sqrt{\left|\frac{Z_m^\theta}{n}\right|^2+\left|\frac{\overline{Z}_m^\theta}{\sqrt{n}}\right|^2}<\delta\right\}}\right]\\&+2C\sum_{m=1}^{n}\sup_{\theta\in\Theta}E_P\left[\left(\left|\frac{Z_m^\theta}{n}\right|^2+\left|\frac{\overline{Z}_m^\theta}{\sqrt{n}}\right|^2\right)I_{\left\{\sqrt{\left|\frac{Z_m^\theta}{n}\right|^2+\left|\frac{\overline{Z}_m^\theta}{\sqrt{n}}\right|^2}\ge\delta\right\}}\right]\\\to&0,\quad\text{ as } n\to\infty \text{ and } \epsilon\to0.\end{aligned}$$

[11] Here $D_z:=(\partial_{z_i})_{i=1}^2$ and $D_z^2:=(\partial^2_{z_iz_j})_{i,j=1}^2$.

The convergence is due to the finiteness of $\underline{\mu}, \overline{\mu}$ and $\overline{\sigma}$. This proves (32).

Combine with Lemma 4 and show that

$$
\begin{aligned}
& e(m,n) \\
= & \sup_{\theta\in\Theta} E_P \left[H_{\frac{m}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) + \partial_{z_1} H_{\frac{m}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right)\frac{Z^\theta_m}{n} \right. \\
& \qquad\qquad\qquad\qquad \left. + \partial^2_{z_2 z_2} H_{\frac{m}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right)\frac{(\overline{Z}^\theta_m)^2}{2n} \right] \\
= & \sup_{\theta\in\Theta} E_P \left[H_{\frac{m}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) + \max_{1\le k\le K} E_P \left[\partial_{z_1} H_{\frac{m}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right)\frac{\mu_k}{n} \right. \right. \\
& \qquad\qquad\qquad\qquad \left. \left. + \partial^2_{z_2 z_2} H_{\frac{m}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right)\frac{\sigma^2_k}{2n} \right] \right] \\
= & \sup_{\theta\in\Theta} E_P \left[L_{m,n}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) \right].
\end{aligned}
$$

This proves (33), and completes the proof of (31). ■

Proof of Proposition 6: We prove it for $u \in C^\infty_b(\mathbb{R}^2)$. This suffices because any $u \in C_{b,Lip}(\mathbb{R}^2)$ can be approximated uniformly by a sequence of functions in $C^\infty_b(\mathbb{R}^2)$ (see Approximation Lemma in Feller (1971, Ch. VIII)). The proof also assumes $\underline{\sigma}> 0$.

For small enough $h > 0$, we continue to use $\{H_t(x,y)\}_{t\in[0,1+h]}$ as defined in (29). Let $\{L_{m,n}(x,y)\}^n_{m=1}$ be the functions defined in (30). By direct calculation we obtain

$$
\begin{aligned}
& \sup_{\theta\in\Theta} E_P \left[H_1\left(\frac{S^\theta_n}{n}, \frac{\overline{S}^\theta_n}{\sqrt{n}}\right) \right] - H_0(0,0) \\
= & \sum_{m=1}^n \left\{ \sup_{\theta\in\Theta} E_P \left[H_{\frac{m}{n}}\left(\frac{S^\theta_m}{n}, \frac{\overline{S}^\theta_m}{\sqrt{n}}\right) \right] - \sup_{\theta\in\Theta} E_P \left[H_{\frac{m-1}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) \right] \right\} \\
= & \sum_{m=1}^n \left\{ \sup_{\theta\in\Theta} E_P \left[H_{\frac{m}{n}}\left(\frac{S^\theta_m}{n}, \frac{\overline{S}^\theta_m}{\sqrt{n}}\right) \right] - \sup_{\theta\in\Theta} E_P \left[L_{m,n}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) \right] \right\} \\
& + \sum_{m=1}^n \left\{ \sup_{\theta\in\Theta} E_P \left[L_{m,n}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) \right] - \sup_{\theta\in\Theta} E_P \left[H_{\frac{m-1}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) \right] \right\} \\
=: & I_{1n} + I_{2n}.
\end{aligned}
$$

Application of Lemma 9 implies that $|I_{1n}| \to 0$ as $n \to \infty$. Lemma 8 implies

$$
\begin{aligned}
|I_{2n}| \le & \sum_{m=1}^n \sup_{\theta\in\Theta} E_P \left[\left| L_{m,n}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) - H_{\frac{m-1}{n}}\left(\frac{S^\theta_{m-1}}{n}, \frac{\overline{S}^\theta_{m-1}}{\sqrt{n}}\right) \right| \right] \\
\le & \sum_{m=1}^n \sup_{(x,y)\in\mathbb{R}^2} \left| L_{m,n}(x,y) - H_{\frac{m-1}{n}}(x,y) \right| \\
\to & \ 0, \quad \text{as } n \to \infty,
\end{aligned}
$$

which implies that

$$\lim_{n\to\infty}\left|\sup_{\theta\in\Theta} E_P\left[H_1\left(\frac{S_n^\theta}{n},\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right]-H_0(0,0)\right|=0.$$

Combine the latter with Lemma 8, part (5), to obtain

$$\begin{aligned}
&\left|V-\sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds,\int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right]\right|\\
&=\lim_{n\to\infty}\left|\sup_{\theta\in\Theta} E_P\left[u\left(\frac{S_n^\theta}{n},\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right]-\sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds,\int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right]\right|\\
&\le\lim_{n\to\infty}\left|\sup_{\theta\in\Theta} E_P\left[\varphi\left(\frac{S_n^\theta}{n},\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right]-\sup_{\theta\in\Theta} E_P\left[H_1\left(\frac{S_n^\theta}{n},\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right]\right|\\
&\quad+\lim_{n\to\infty}\left|\sup_{\theta\in\Theta} E_P\left[H_1\left(\frac{S_n^\theta}{n},\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right]-H_0(0,0)\right|\\
&\quad+\left|H_0(0,0)-\sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds,\int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right]\right|\le C_0h,
\end{aligned}$$

where the constant C_0 depends only on $\underline{\mu},\overline{\mu},\overline{\sigma}$ and the uniform bound of $\partial_x u$ and $\partial_{yy}^2 u$. By the arbitrariness of h, the proof of (25) is completed.

Finally, prove (26). Let G be defined by (28), and define, for all $(p,q)\in\mathbb{R}^2$,

$$G^{ext}(p,q)=\sup_{(\mu,\sigma^2)\in\mathcal{A}^{ext}}\left[\mu p+\frac{1}{2}\sigma^2 q\right].$$

Then

$$G(p,q)=G^{ext}(p,q)\quad\forall(p,q)\in\mathbb{R}^2. \tag{35}$$

The proof is completed by applying a Comparison Theorem (Peng (2019, Theorem C.2.5)). ■

Proof of Theorem 1: All the results can be obtained from Proposition 6. That u need only satisfy continuity and the stated growth condition is implied by Lemma 2.4.12 and Exercise 2.5.7 in Peng (2019) (or by Rosenthal's inequality in Zhang (2016)). For the convenience of readers, we provide a proof in the Online Appendix. ■

A.2 Proof of Theorem 2

We are given that $u(x,y)$ is increasing in x and concave in y, and $(\overline{\mu},\underline{\sigma}^2)\in\mathcal{A}$.

For any $t\in[0,1]$ and $(x,y)\in\mathbb{R}^2$, define the function

$$v(t,x,y)=E_P[u(x+(1-t)\overline{\mu},y+\underline{\sigma}(B_1^{(2)}-B_t^{(2)}))].$$

Then

$$v(0,0,0)=E_P[u(\overline{\mu},\underline{\sigma}B_1^{(2)}]=\int u(\overline{\mu},\cdot)d\mathbb{N}(0,\underline{\sigma}^2).$$

By the (classic) Feynman-Kac formula (Mao (2008, Theorem 2.8.3)), v is the solution of the (linear parabolic) PDE

$$\left\{\begin{array}{l}\partial_t v(t,x,y)+\overline{\mu}\partial_x v(t,x,y)+\frac{1}{2}\underline{\sigma}^2\partial^2_{yy}v(t,x,y)=0,\ (t,x,y)\in[0,1)\times\mathbb{R}^2\\ v(1,x,y)=u(x,y).\end{array}\right. \tag{36}$$

Since $u(x,y)$ is increasing in x and concave in y, it follows that $v(t,x,y)$ is increasing in x and concave in y for any $t\in[0,1]$, that is,

$$\partial_x v(t,x,y)\geq 0 \text{ and } \partial^2_{yy}v(t,x,y)\leq 0,\quad \forall(t,x,y)\in[0,1)\times\mathbb{R}^2.$$

Given also $(\overline{\mu},\underline{\sigma}^2)\in\mathcal{A}$, it follows that

$$\sup_{(\mu,\sigma^2)\in\mathcal{A}}\left\{\mu\partial_x v+\tfrac{1}{2}\sigma^2\partial^2_{yy}v\right\}=\overline{\mu}\partial_x v+\tfrac{1}{2}\underline{\sigma}^2\partial^2_{yy}v,$$

and hence that v solves the PDE (20). By uniqueness of the solution (Lemma 5), and (27), conclude that

$$V=v(0,0,0)=\int u(\overline{\mu},\cdot)d\mathbb{N}(0,\underline{\sigma}^2).$$

■

A.3 Proof of Theorem 3

Throughout we assume that $\mathcal{A}=\{(\mu_1,\sigma_1^2),(\mu_2,\sigma_2^2)\}$.

Proof of (i): The proof consists of three steps.

Step 1: From Theorem 1(i) and (27), it follows that

$$\lim_{n\to\infty}V_n=\lim_{n\to\infty}\sup_{\theta\in\Theta}E_P\left[u\left(\frac{S_n^\theta}{n},\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right]=v(0,0,0)$$

where $v(t,x,y)$ solves the PDE (20).

Step 2: Prove that the following function v solves the above PDE:

$$\begin{aligned}\hat{v}(t,x,y)=&E_P[u(x+(1-t)\mu_1,y+\sigma_1(B_1^{(2)}-B_t^{(2)}))]\\ =&\int_{\mathbb{R}}u(x+(1-t)\mu_1,y+\sqrt{1-t}\sigma_1 r)\frac{1}{\sqrt{2\pi}}e^{-\frac{r^2}{2}}dr\end{aligned} \tag{37}$$

By the Feynman-Kac formula, $\hat{v}$ solves

$$\left\{\begin{array}{l}\partial_t\hat{v}(t,x,y)+\mu_1\partial_x\hat{v}(t,x,y)+\frac{1}{2}\sigma_1^2\partial^2_{yy}\hat{v}(t,x,y)=0,\ (t,x,y)\in[0,1)\times\mathbb{R}^2\\ \hat{v}(1,x,y)=u(x,y).\end{array}\right. \tag{38}$$

From (37) and assumption (15), it follows that, for all $(t,x,y)\in[0,1)\times\mathbb{R}^2$,

$$\tfrac{1}{2}\sigma_1^2\partial^2_{yy}\hat{v}(t,x,y)+\mu_1\partial_x\hat{v}(t,x,y)\geq\tfrac{1}{2}\sigma_2^2\partial^2_{yy}\hat{v}(t,x,y)+\mu_2\partial_x\hat{v}(t,x,y),$$

that is,

$$\sup_{(\mu,\sigma^2)\in\mathcal{A}}\left\{\mu\partial_x\hat{v}+\frac{1}{2}\sigma^2\partial^2_{yy}\hat{v}\right\}=\mu_1\partial_x\hat{v}+\tfrac{1}{2}\sigma_1^2\partial^2_{yy}\hat{v}. \tag{39}$$

Thus $\hat{v}$ solves the PDE (20). By uniqueness of the solution (Lemma 5), conclude that

$$\lim_{n\to\infty} V_n = v(0,0,0) = \hat{v}(0,0,0) = \int u(\mu_1,\cdot)d\mathbb{N}(0,\sigma_1^2).$$

Step 3: If θ^* denotes the strategy of choosing arm 1 always, then, using Step 1,

$$\lim_{n\to\infty} E_P\left[u\left(\frac{S_n^{\theta^*}}{n},\frac{\overline{S}_n^{\theta^*}}{\sqrt{n}}\right)\right] = E_P[u(\mu_1,\sigma_1 B_1^{(2)})] = v\left(0,0,0\right) = V.$$

Hence θ^* is asymptotically optimal.

Proof of (iii): Case 1 ($\alpha \leq \frac{\mu_1-\mu_2}{\sigma_1^2-\sigma_2^2}$): Define v by (37). Although u is not twice differentiable, we can calculate $\partial_x v$ and $\partial_{yy}^2 v$ directly to obtain $\partial_x v = 1$ and $\partial_{yy}^2 v = -2\alpha\Phi(\frac{-y}{\sigma_1\sqrt{1-t}})$. Therefore,

$$\begin{aligned} \alpha &< \frac{\mu_1-\mu_2}{\sigma_1^2-\sigma_2^2} \Longrightarrow \\ \mu_1 - \alpha\Phi(\frac{-y}{\sqrt{1-t}\sigma_1})\sigma_1^2 &> \mu_2 - \alpha\Phi(\frac{-y}{\sqrt{1-t}\sigma_1})\sigma_2^2 \Longrightarrow \\ \mu_1\partial_x v + \tfrac{1}{2}\sigma_1^2\partial_{yy}^2 v &> \mu_2\partial_x v + \tfrac{1}{2}\sigma_2^2\partial_{yy}^2 v. \end{aligned}$$

Proceed as in the proof of (i).[12]

Case 2 ($\underline{\alpha} < \alpha < \overline{\alpha}$): To prove that single-arm strategies are not asymptotically optimal, it is enough to show that

$$V = \sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right] > \max_{i=1,2} E_P\left[u\left(\mu_i,\sigma_i B_1^{(2)}\right)\right]. \quad (40)$$

Consider the bandit problem with set of arms given by

$$\mathcal{A}' = \{(\mu_1,\sigma_1^2),(\mu_2,\sigma_2^2),(\mu_3,\sigma_3^2)\},$$

where $(\mu_3,\sigma_3^2) = (1-\lambda)(\mu_1,\sigma_1^2)+\lambda(\mu_2,\sigma_2^2)$ for some $0<\lambda<1$ to be selected below. Because $\mathcal{A}'$ and $\mathcal{A}$ have the identical extreme points, Proposition 6 implies that

$$\begin{aligned} V &= \sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right] \\ &= \sup_{a\in[\mathcal{A}'](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right]. \end{aligned}$$

Take

$$(\hat{a}_s^{(1)},\hat{a}_s^{(2)}) = (\mu_1,\sigma_1)I_{\{W_s^{\sigma_1,\sigma_3}\geq 0\}} + (\mu_3,\sigma_3)I_{\{W_s^{\sigma_1,\sigma_3}<0\}}, \quad (41)$$

where

$$W_t^{\sigma_1,\sigma_3} = \int_0^t \left(\sigma_1 I_{\{W_s^{\sigma_1,\sigma_3}\geq 0\}} + \sigma_3 I_{\{W_s^{\sigma_1,\sigma_3}<0\}}\right)dB_s^{(2)};$$

[12] If we assume the reverse inequality in (18), then corresponding implications fail. For example, if $y>0$ is sufficiently large which would make $\Phi(\frac{-y}{\sqrt{1-t\sigma}})$ close to zero for $\sigma=\sigma_1,\sigma_2$. $t\geq 0$, then the last two inequalities above could remain valid even though $\alpha > \left(\mu_1-\mu_2\right)/\left(\sigma_1^2-\sigma_2^2\right)$.

$W_s^{\sigma_1,\sigma_3}$ is an oscillating Brownian motion, that is, the solution of the stochastic differential equation (SDE)

$$W_t^{\sigma_1,\sigma_3} = \int_0^t \left(\sigma_1 I_{\{W_s^{\sigma_1,\sigma_3} \geq 0\}} + \sigma_3 I_{\{W_s^{\sigma_1,\sigma_3} < 0\}}\right) dB_s^{(2)}.$$

By Keilson and Wellner (1978, Theorem 1), the probability density of $W_t^{\sigma_1,\sigma_3}$ is $q\,(t,\cdot)$, where

$$q\,(t,y) = \begin{cases} q^*\left(y;\sigma_1^2 t\right)\left[\frac{2\sigma_3}{\sigma_1+\sigma_3}\right] & y \geq 0 \\ \\ q^*\left(y;\sigma_3^2 t\right)\left[\frac{2\sigma_1}{\sigma_1+\sigma_3}\right] & y < 0 \end{cases} \tag{42}$$

and $q^*(y;\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-(y/\sigma)^2/2\right)$ is the pdf for $\mathbb{N}\left(0,\sigma^2\right)$. Using this pdf, we can calculate

$$\begin{aligned}
&E_P\left[u\left(\int_0^1 \hat{a}_s^{(1)}ds, \int_0^1 \hat{a}_s^{(2)}dB_s^{(2)}\right)\right] \\
=&E_P\left[\int_0^1 \left(\mu_1 I_{\{W_s^{\sigma_1,\sigma_3}\geq 0\}} + \mu_2 I_{\{W_s^{\sigma_1,\sigma_3}<0\}}\right)ds\right] - \alpha E_P\left[\left(W_1^{\sigma_1,\sigma_3}\right)^2 I_{\{W_1^{\sigma_1,\sigma_3}\leq 0\}}\right] \\
=&\mu_1\int_0^1 P(W_s^{\sigma_1,\sigma_3}\geq 0)ds + \mu_3\int_0^1 P(W_s^{\sigma_1,\sigma_3}<0)ds - \alpha\int_{-\infty}^0 y^2 q(1,y)dy \\
=&\mu_1\int_0^1\int_0^\infty q(s,y)dyds + \mu_3\int_0^1\int_{-\infty}^0 q(s,y)dyds - \alpha\int_{-\infty}^0 y^2 q(1,y)dy \\
=&\mu_1\frac{\sigma_3}{\sigma_1+\sigma_3} + \mu_3\frac{\sigma_1}{\sigma_1+\sigma_3} - \alpha\frac{\sigma_1\sigma_3^2}{\sigma_1+\sigma_3}.
\end{aligned}$$

Verify the inequality

$$\mu_1\frac{\sigma_3}{\sigma_1+\sigma_3} + \mu_3\frac{\sigma_1}{\sigma_1+\sigma_3} - \alpha\frac{\sigma_1\sigma_3^2}{\sigma_1+\sigma_3} > \mu_1 - \alpha\frac{\sigma_1^2}{2} = E_P\left[u\left(\mu_1,\sigma_1 B_1^{(2)}\right)\right],$$

and deduce that

$$\alpha > \frac{2(\mu_1-\mu_3)}{(\sigma_1+2\sigma_3)(\sigma_1-\sigma_3)} = \frac{\mu_1-\mu_2}{\sigma_1^2-\sigma_2^2+f(\lambda)}$$

where $f(\lambda) = \left(\sigma_1\sqrt{(1-\lambda)\sigma_1^2+\lambda\sigma_2^2}-\sigma_1^2\right)/2\lambda$. It can be verified that $f'(\lambda)<0$ for $\lambda\in(0,1)$ and $\lim_{\lambda\to 0} f(\lambda) = \left(\sigma_2^2-\sigma_1^2\right)/4$.

Therefore, for any $\alpha > \underline{\alpha} = \frac{4(\mu_1-\mu_2)}{3(\sigma_1^2-\sigma_2^2)}$, there exists $\lambda_0\in(0,1)$ such that

$$\alpha > \frac{\mu_1-\mu_2}{\sigma_1^2-\sigma_2^2+f(\lambda_0)} > \frac{4(\mu_1-\mu_2)}{3(\sigma_1^2-\sigma_2^2)}.$$

Choose $\lambda=\lambda_0$ in the definition (41) of $\hat{a} = (\hat{a}_s^{(1)},\hat{a}_s^{(2)})$ and deduce that

$$\begin{aligned}
V = &\sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right] \\
\geq &E_P\left[u\left(\int_0^1 \hat{a}_s^{(1)}ds, \int_0^1 \hat{a}_s^{(2)}dB_s^{(2)}\right)\right] \\
> &E_P\left[u\left(\mu_1,\sigma_1 B_1^{(2)}\right)\right].
\end{aligned}$$

That is, when $\alpha > \underline{\alpha}$, then specializing in arm 1 is NOT asymptotically optimal.

An analogous argument proves that specializing in arm 2 is not asymptotically optimal if $\alpha < \overline{\alpha}$. Details are provided in the Online Appendix.

Proof of (iv): It remains to prove only the claim for the case $\underline{\alpha}' < \alpha$. The proof is similar to that for (iii). Specifically, prove that (40) is satisfied for the process $(\hat{a}_s^{(1)}, \hat{a}_s^{(2)})$ where

$$(\hat{a}_s^{(1)}, \hat{a}_s^{(2)}) = (\mu_1, \sigma_1) I_{\{W_s^{\sigma_2,\sigma_1} < 0\}} + (\mu_2, \sigma_2) I_{\{W_s^{\sigma_2,\sigma_1} \geq 0\}},$$

and $W_s^{\sigma_2,\sigma_1}$ is the oscillating Brownian motion given by

$$W_t^{\sigma_2,\sigma_1} = \int_0^t \left(\sigma_1 I_{\{W_s^{\sigma_2,\sigma_1} < 0\}} + \sigma_2 I_{\{W_s^{\sigma_2,\sigma_1} \geq 0\}} \right) dB_s^{(2)}.$$

The process $W_t^{\sigma_2,\sigma_1}$ admits a probability density analogous to (42).

Proof of (v): For $i \geq 1$, we have $Z_i^{\theta^*} = X_{k,i}$ where $\theta_i^* = k$, and $\{X_{k,i} : i \geq 1\}$ are i.i.d. Then

$$E_P \left[\varphi \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{\theta^*} \right) \right] = E_P \left[\varphi \left(\frac{\psi_n}{n} \frac{\sum_{i=1}^{\psi_n} X_{1,i}}{\psi_n} + \frac{n - \psi_n}{n} \frac{\sum_{i=1}^{n-\psi_n} X_{2,i}}{n - \psi_n} \right) \right]$$

Since $\psi_n / n \to \lambda$ as $n \to \infty$, combine with the classical LLN for $\{X_{1,i} : i \geq 1\}$ and $\{X_{2,i} : i \geq 1\}$ to obtain

$$\lim_{n \to \infty} E_P \left[\varphi \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{\theta^*} \right) \right] = \varphi \left(\lambda \mu_1 + (1 - \lambda) \mu_2 \right) = \varphi(x^*).$$

Therefore, θ^* is asymptotically optimal because, by Proposition 6,

$$\begin{aligned} V &= \sup_{a \in [\mathcal{A}](0,1)} E_P \left[u \left(\int_0^1 a_s^{(1)} ds, \int_0^1 a_s^{(2)} dB_s^{(2)} \right) \right] \\ &= \sup_{a \in [\mathcal{A}](0,1)} E_P \left[\varphi \left(\int_0^1 a_s^{(1)} ds \right) \right] \leq \varphi(x^*). \end{aligned}$$

The remaining assertion is implied by the fact that $\lim_{n \longrightarrow \infty} U_n \left(\theta^{\mu,\sigma} \right) = \varphi(\mu)$ for each $\left(\mu, \sigma^2 \right)$. ■

References

[1] Aivaliotis, G. and J. Palczewski (2010). Tutorial for viscosity solutions in optimal control of diffusions. Available at SSRN 1582548.

[2] Benartzi, S. and R.H. Thaler (1999). Risk aversion or myopia? choices in repeated gambles and retirement investments. *Management Sci.* 45(3), 364-381.

[3] Berry, D. and B. Fristedt (1985). *Bandit Problems.* Chapman Hall, London.

[4] Cassel, A., S. Mannor and A. Zeevi (2018). A general approach to multi-armed bandits under risk criteria. *Proc. Machine Learn. Res.* 75, 1–12.

[5] Chen, Z. and L.G. Epstein (2022). A central limit theorem for sets of measures. *Stoch. Process. Appl.* 152, 424-451.

[6] Chen Z., L.G. Epstein, and G. Zhang (2023). A central limit theorem, loss aversion and multi-armed bandits. *J. Econom. Theory* 209, 105645.

[7] Chew S.H. and L.G. Epstein (1988). The law of large numbers and the attractiveness of compound gambles. *J. Risk Uncert.* 1, 125-132.

[8] Fang, X., S. Peng, Q.M. Shao and Y. Song (2019). Limit theorems with rate of convergence under sublinear expectations. *Bernoulli* 25(4A), 2564-2596.

[9] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*,Vol.II. Second Edition. John Wiley and Sons, New York.

[10] Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique priors. *J. Math. Econom.* 18, 141-153.

[11] Gollier C. (1996). Repeated optional gambles and risk aversion. *Management Sci.* 42(11), 1524-1530.

[12] Gollier, C. and R.J. Zeckhauser (2002). Time-horizon length and risk aversion. *J. Risk Uncert.* 24(3), 195-212.

[13] Huberman, G. and S. Ross (1983). Portfolio turnpike theorems, risk aversion, and regularly varying utility functions. *Econometrica* 51(5), 1345-1361.

[14] Keeney, R.L. and H. Raiffa. 1993. *Decisions with Multiple Objectives.* Cambridge U. Press, NewYork. Original publication, Wiley, NewYork, 1976.

[15] Keilson, J. and J.A. Wellner (1978). Oscillating Brownian motion. *J. Appl. Probab.* 15(2), 300-310.

[16] Klebaner, F., Z. Landsman, U. Makov and J. Yao (2017). Optimal portfolios with downside risk. *Quant. Finan.* 17, 315-325.

[17] Krylov, N.V.: (1987) *Nonlinear Parabolic and Elliptic Equations of the Second Order.* Reidel. Original Russian version by Nauka, Moscow (1985).

[18] Lopes L.L. (1996). When time is of the essence: averaging, aspiration and the short-run. *Organizational Behavior and Human Decision Processes* 65(3), 179-189.

[19] Mao, X. (2008). *Stochastic Differential Equations and Applications.* Woodhead Publishing.

[20] Markowitz H. (1959). *Portfolio Selection.* Yale U. Press, New Haven.

[21] McCardle, K.F. and R.I. Winkler (1992). Repeated gambles, learning and risk aversion. *Management Sci.* 38, 807-818.

[22] Nantell, T.J. and B. Price (1979). An analytical comparison of variance and semivariance capital market theories. *J. Finan. Quant. Anal.* 14, 221-242.

[23] Peng, S. (2007). G-expectation, G-Brownian motion and related stochastic calculus of Itô type. In: Benth, F.E., Di Nunno, G., Lindstrøm, T., Øksendal, B., Zhang, T. (eds) *Stoch. Anal. Appl.* Abel Symposia, vol 2. Springer, Berlin, https://doi.org/10.1007/978-3-540-70847-6_25.

[24] Peng, S. (2019). *Nonlinear Expectations and Stochastic Calculus under Uncertainty: with Robust CLT and G-Brownian Motion.* Springer Nature.

[25] Pham, H. (2009). *Continuous-Time Stochastic Control and Optimization with Financial Applications* (vol. 61). Springer Science & Business Media.

[26] Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica* 32(1/2), 122-136.

[27] Samuelson, P.A. (1963). Risk and uncertainty: the fallacy of the law of large numbers. *Scientia* 98, 108-113.

[28] Samuelson, P.A. (1989). The judgement of economic science on rational portfolio management: Indexing, timing and long-horizon effects. *J. Portfolio Management*, Fall, 3-12.

[29] Samuelson, P.A. (1997). Proof by certainty equivalents that diversification-across-time does worse, risk corrected, than diversification-throughout-time. *J. Risk Uncert.* 14, 129-142.

[30] Sani, A., A. Lazaric and R. Munos (2013). Risk-aversion in multi-armed bandits. arXiv:1301.1936v1 [cs.LG].

[31] Slivkins, A. (2022). *Introduction to Multi-Armed Bandits.* arXiv:1904.07272v7 [cs.LG].

[32] Vakili, S. and Q. Zhao (2016). Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE J. Selected Topics in Signal Processing*, Digital object identifier 10.1109/JSTSP.2016.2592622.

[33] Yong, J. and X.Y. Zhou (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations* (vol. 43). Springer Science & Business Media.

[34] Zhang, L. (2016). Rosenthal's inequalities for independent and negatively dependent random variables under sub-linear expectations with applications. *Science China Math.* 59(4), 751-768.

[35] Zimin, A., R. Ibsen-Jensen and K. Chatterjee (2014). Generalized risk-aversion in stochastic multi-armed bandits. arXiv:1405.0833 [cs.LG].

ONLINE APPENDIX

Lemma: Our CLT, Proposition 6, is valid also if $\underline{\sigma} = 0$.
Proof: As in the proof of Proposition 6, it suffices to take $u \in C_b^\infty(\mathbb{R}^2)$.

Given $\underline{\sigma} = 0$, we add a perturbation to the random returns of the K arms. For any $1 \leq k \leq K$ and $n \geq 1$, let $X_{k,n}^\epsilon = X_{k,n} + \epsilon\zeta_n$, where $\epsilon > 0$ is a fixed small constant and $\{\zeta_n\}$ is a sequence of i.i.d. standard normal random variables, independent with $\{X_{k,n}\}$. Then, for any $\theta \in \Theta$ and $n \geq 1$, the corresponding reward is denoted by $Z_n^{\theta,\epsilon} = Z_n^\theta + \epsilon\zeta_n$, and the corresponding set of mean-variance pairs is denoted by

$$\mathcal{A}_\epsilon = \{(\mu_{k,\epsilon}, \sigma_{k,\epsilon}^2) : 1 \leq k \leq K\},$$

where $\mu_{k,\epsilon} = \mu_k$ and $\sigma_{k,\epsilon}^2 = \sigma_k^2 + \epsilon^2$. The corresponding bounds are $\overline{\mu}_\epsilon, \underline{\mu}_\epsilon, \overline{\sigma}_\epsilon^2$, and $\underline{\sigma}_\epsilon^2 > 0$.

Define

$$V_n^\epsilon = \sup_{\theta\in\Theta} E_P \left[u \left(\frac{\sum_{i=1}^n Z_i^{\theta,\epsilon}}{n}, \frac{\sum_{i=1}^n (Z_i^{\theta,\epsilon} - E_P[Z_i^{\theta,\epsilon}|\mathcal{H}_{i-1}^\theta])}{\sqrt{n}} \right) \right]$$

By Proposition 6 for $\{Z_n^{\theta,\epsilon}\}$,

$$\lim_{n\to\infty} V_n^\epsilon = \sup_{a\in[\mathcal{A}_\epsilon](0,1)} E_P \left[u \left(\int_0^1 a_s^{(1)} ds, \int_0^1 a_s^{(2)} dB_s^{(2)} \right) \right] = v_\epsilon(0,0,0), \tag{43}$$

where $v_\epsilon(t,x,y)$ is the solution of PDE (20) with function G_ϵ instead of G,

$$G_\epsilon(p,q) = \sup_{(\mu,\sigma^2)\in\mathcal{A}_\epsilon} \left[\mu p + \frac{1}{2}\sigma^2 q \right], \quad (p,q) \in \mathbb{R}^2. \tag{44}$$

By Yong and Zhou (1999, Propn. 5.10, Ch. 4), $\exists C' > 0$ such that

$$|v_\epsilon(t,x,y) - v(t,x,y)| \leq C'\sqrt{\epsilon}, \quad \forall (t,x,y) \in [0,1) \times \mathbb{R}^2.$$

We also have

$$|V_n - V_n^\epsilon|^2 \leq C\epsilon^2 E_P \left[\left| \frac{\sum_{i=1}^n \zeta_i}{n} \right|^2 + \left| \frac{\sum_{i=1}^n \zeta_i}{\sqrt{n}} \right|^2 \right] \leq 2C\epsilon^2,$$

where the constant C depends only on the bounds of $\partial_x u$ and $\partial_y u$.

Letting as $\epsilon \to 0$ in (43), the CLT (25) is proven for $\underline{\sigma} = 0$. Similar arguments show that (26) is also valid. ■

Lemma: Our CLT, Proposition 6, is valid also if u is continuous and, for some $g \geq 1$ and $c > 0$, $|u(x,y)| \leq c(1 + ||(x.y)||^{g-1})$ and $\sup_{1\leq k\leq K} E_P[|X_k|^g] < \infty$.
Proof: We prove that (25) remains valid. Refer to it as "the CLT."
Step 1: Prove the CLT for any $u \in C_b(\mathbb{R}^2)$ with compact support (constant outside a

compact subset of $\mathbb{R}^2$). In this case, $\forall \epsilon > 0\ \exists \hat{u} \in C_{b,Lip}(\mathbb{R}^2)$ such that $\sup_{z\in\mathbb{R}^2} |u(z) - \hat{u}(z)| \leq \frac{\epsilon}{2}$. Then

$$\begin{aligned}
&\left|\sup_{\theta\in\Theta} E_P\left[u\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[u(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right| \\
\leq &\epsilon + \left|\sup_{\theta\in\Theta} E_P\left[\hat{u}\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[\hat{u}(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right|
\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty}\left|\sup_{\theta\in\Theta} E_P\left[u\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[u(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right| \leq \epsilon,$$

which proves the CLT since ϵ is arbitrary.
Step 2: Let $u \in C(\mathbb{R}^2)$ satisfy the growth condition $|u(z)| \leq c(1 + |z|^{g-1})$ for $g \geq 1$. For any $N > 0$, $\exists u_1, u_2 \in C(\mathbb{R}^2)$ such that $u = u_1 + u_2$, where u_1 has a compact support and $u_2(z) = 0$ for $|z| \leq N$, and $|u_2(z)| \leq |u(z)|$ for all z. Then

$$|u_2(z)| \leq \frac{2c(1 + |z|^g)}{N}, \quad \forall z \in \mathbb{R}^2,$$

and

$$\begin{aligned}
&\left|\sup_{\theta\in\Theta} E_P\left[u\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[u(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right| \\
\leq &\left|\sup_{\theta\in\Theta} E_P\left[u_1\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[u_1(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right| \\
&+ \sup_{\theta\in\Theta} E_P\left[\left|u_2\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right|\right] + \sup_{a\in[\mathcal{A}](0,1)} E_P[|u_2(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})|] \\
\leq &\left|\sup_{\theta\in\Theta} E_P\left[u_1\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[u_1(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right| \\
&+ \frac{2c}{N}\left(2 + \sup_{\theta\in\Theta} E_P\left[\left|\frac{S_n^\theta}{n}\right|^g + \left|\frac{\overline{S}_n^\theta}{\sqrt{n}}\right|^g\right] + \sup_{a\in[\mathcal{A}](0,1)} E_P\left[\left|\int_0^1 a_s^{(1)}ds\right|^g + \left|\int_0^1 a_s^{(2)}dB_s^{(2)}\right|^g\right]\right)
\end{aligned}$$

By the Burkholder-Davis-Gundy inequality (Mao (2008, Theorem 1.7.3)),

$$\begin{aligned}
&\limsup_{n\to\infty}\left|\sup_{\theta\in\Theta} E_P\left[u\left(\frac{S_n^\theta}{n}, \frac{\overline{S}_n^\theta}{\sqrt{n}}\right)\right] - \sup_{a\in[\mathcal{A}](0,1)} E_P[u(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)})]\right| \\
\leq &\frac{2c}{N}\left(2 + \max\{|\overline{\mu}|^g, |\underline{\mu}|^g\} + \overline{\sigma}^g + \sup_n \sup_{\theta\in\Theta} E_P\left[\left|\frac{S_n^\theta}{n}\right|^g + \left|\frac{\overline{S}_n^\theta}{\sqrt{n}}\right|^g\right]\right).
\end{aligned}$$

Since N can be arbitrarily large, it suffices to prove

$$\sup_n \sup_{\theta\in\Theta} E_P\left[\left|\frac{S_n^\theta}{n}\right|^g + \left|\frac{\overline{S}_n^\theta}{\sqrt{n}}\right|^g\right] < \infty$$

Step 3: Prove the preceding inequality. For any n,

$$\sup_{\theta\in\Theta} E_P\left[\left|\frac{S_n^\theta}{n}\right|^g\right] \leq \sup_{\theta\in\Theta} E_P\left[\frac{n^{g-1}}{n^g}\sum_{i=1}^n |Z_i^\theta|^g\right] \leq K \sup_{1\leq k\leq K} E_P[|X_k|^g].$$

For $1 \leq g \leq 2$,

$$\begin{aligned}\left(\sup_{\theta\in\Theta} E_P\left[\left|\frac{\overline{S}_n^\theta}{\sqrt{n}}\right|^g\right]\right)^{\frac{2}{g}} \leq& \sup_{\theta\in\Theta} E_P\left[\left(\frac{\overline{S}_n^\theta}{\sqrt{n}}\right)^2\right] \\ =&\frac{1}{n}\sup_{\theta\in\Theta} E_P\left[\left(\overline{S}_{n-1}^\theta\right)^2 + 2\overline{S}_{n-1}^\theta\overline{Z}_n^\theta + (\overline{Z}_n^\theta)^2\right] \\ \leq&\frac{1}{n}\sup_{\theta\in\Theta} E_P\left[\left(\overline{S}_{n-1}^\theta\right)^2 + \overline{\sigma}^2\right] \leq \overline{\sigma}^2.\end{aligned}$$

For $g > 2$,

$$|x+y|^g \leq 2^g g^2|x|^g + |y|^g + gx|y|^{g-1}sgn(y) + 2^g g^2 x^2|y|^{g-2},\ \forall x, y \in \mathbb{R}.$$

Let $T_k^\theta = \max\{\overline{S}_k^\theta, \overline{S}_k^\theta - \overline{S}_1^\theta, \cdots, \overline{S}_k^\theta - \overline{S}_{k-1}^\theta\}$. Then $T_k^\theta = \overline{Z}_k^\theta + (T_{k-1}^\theta)^+$ and

$$\begin{aligned}&\sup_{\theta\in\Theta} E_P[|T_k^\theta|^g] \\ \leq&2^g g^2 \sup_{\theta\in\Theta} E_P[|\overline{Z}_k^\theta|^g] + \sup_{\theta\in\Theta} E_P[|(T_{k-1}^\theta)^+|^g] \\ &+ g\sup_{\theta\in\Theta} E_P[\overline{Z}_k^\theta|(T_{k-1}^\theta)^+|^{g-1}] + 2^g g^2 \sup_{\theta\in\Theta} E_P[(\overline{Z}_k^\theta)^2|(T_{k-1}^\theta)^+|^{g-2}] \\ \leq&2^g g^2\sum_{i=1}^k \sup_{\theta\in\Theta} E_P[|\overline{Z}_i^\theta|^g] + 2^g g^2\sum_{i=2}^k \sup_{\theta\in\Theta} E_P[(\overline{Z}_i^\theta)^2|(T_{i-1}^\theta)^+|^{g-2}] \\ \leq&2^g g^2\sum_{i=1}^n \sup_{\theta\in\Theta} E_P[|\overline{Z}_i^\theta|^g] + 2^g g^2\overline{\sigma}^2\sum_{i=1}^n \left(\sup_{\theta\in\Theta} E_P[|(T_i^\theta)^+|^g]\right)^{\frac{g-2}{g}}.\end{aligned}$$

Let $A_n = \sup_{k\leq n}\sup_{\theta\in\Theta} E_P[|T_k^\theta|^g]$. Then, by Young's inequality (Peng (2019, Lemma 1.4.1)),[13]

$$\begin{aligned}A_n \leq&2^g g^2\sum_{i=1}^n \sup_{\theta\in\Theta} E_P[|\overline{Z}_i^\theta|^g] + 2^g g^2\overline{\sigma}^2 n A_n^{\frac{g-2}{g}} \\ \leq&2^g g^2\sum_{i=1}^n \sup_{\theta\in\Theta} E_P[|\overline{Z}_i^\theta|^g] + \frac{2}{g}(2^g g^2\overline{\sigma}^2 n)^{\frac{g}{2}} + \frac{g-2}{g}A_n.\end{aligned}$$

Therefore,

$$\begin{aligned}A_n \leq&C_{g,1}\sum_{i=1}^n \sup_{\theta\in\Theta} E_P[|\overline{Z}_i^\theta|^g] + C_{g,2}n^{\frac{g}{2}} \\ \leq&C_{g,1}\sum_{i=1}^n \sup_{\theta\in\Theta} E_P[|Z_i^\theta|^g + \max\{|\overline{\mu}|^g, |\underline{\mu}|^g\}] + C_{g,2}n^{\frac{g}{2}} \\ \leq&C_{g,1}nK\sup_{1\leq k\leq K} E_P[|X_k|^g] + C_{g,1}n\max\{|\overline{\mu}|^g, |\underline{\mu}|^g\} + C_{g,2}n^{\frac{g}{2}}.\end{aligned}$$

[13] $| ab |\leq p^{-1} \mid a \mid^p + q^{-1} \mid a \mid^q$ if $1 < p, q < \infty$ and $p^{-1} + q^{-1} = 1$.

Finally,

$$\begin{aligned}
&\sup_{\theta\in\Theta} E_P\left[\left|\frac{\overline{S}_n^\theta}{\sqrt{n}}\right|^g\right] \leq n^{-\frac{g}{2}} A_n \\
\leq & C_{g,1} n^{1-\frac{g}{2}} K \sup_{1\leq k\leq K} E_P[|X_k|^g] + C_{g,1} n^{1-\frac{g}{2}} \max\{|\overline{\mu}|^g, |\underline{\mu}|^g\} + C_{g,2}.
\end{aligned}$$

Since $\sup_{1\leq k\leq K} E_P[|X_k|^g] < \infty$, Step 3 is complete and the Lemma is proven. ■

Proof of Corollary 7: The preceding Lemma proves the extension for Proposition 6.

To prove (27), define

$$v(t,x,y) = \sup_{a\in[\mathcal{A}](t,1)} E_P\left[u\left(x + \int_t^{1+h} a_s^{(1)} ds, y + \int_t^{1+h} a_s^{(2)} dB_s^{(2)}\right)\right], \quad (x,y)\in\mathbb{R}^2.$$

As in the proof of Lemma 8(1), for $u \in C_{b,Lip}(\mathbb{R}^2)$, it can be checked that (Yong and Zhou (1999, Theorem 5.2 in Chapter 4)) v is the unique viscosity solution of the HJB-equation (20) with function G given in (28). Then we have

$$V = \sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(x + \int_t^{1+h} a_s^{(1)} ds, y + \int_t^{1+h} a_s^{(2)} dB_s^{(2)}\right)\right] = v(0,0,0).$$

For $u \in C(\mathbb{R}^2)$ with growth condition, the value function is still the unique viscosity solution of the PDE (20) with function G given in (28). Supporting details can be found in Pham (2009, p.66) or Aivaliotis and Palczewski (2010, Corollary 4.7). ■

The Krylov norm: W use the notation in Krylov (1987, Section 1.1); see also Peng (2019, Chapter 2.1). For $\Gamma \subset [0,\infty)\times\mathbb{R}^2$, $C(\Gamma)$ denotes the set of all real-valued functions v defined on Γ, continuous in the relative topology on Γ and having a finite norm,

$$\|v\|_{C(\Gamma)} = \sup_{(t,z)\in\Gamma} |v(t,z)|.$$

Similarly, given $\alpha,\beta\in(0,1)$,

$$\|v\|_{C^{\alpha,\beta}(\Gamma)} = \|v\|_{C(\Gamma)} + \sup_{(t,z),(t',z')\in\Gamma,(t,z)\neq(t',z')} \frac{|v(t,z)-v(t',z')|}{|t-t'|^\alpha + |z-z'|^\beta}$$

$$\|v\|_{C^{1+\alpha,1+\beta}(\Gamma)} = \|v\|_{C^{\alpha,\beta}(\Gamma)} + \|\partial_t v\|_{C^{\alpha,\beta}(\Gamma)} + \sum_{i=1}^2 \|\partial_{z_i} v\|_{C^{\alpha,\beta}(\Gamma)}.$$

$$\|v\|_{C^{1+\alpha,2+\beta}(\Gamma)} = \|v\|_{C^{1+\alpha,1+\beta}(\Gamma)} + \sum_{i,j=1}^2 \|\partial^2_{z_i z_j} v\|_{C^{\alpha,\beta}(\Gamma)}.$$

The corresponding subspaces of $C(\Gamma)$ in which the correspondent derivatives exist and the above norms are finite are denoted respectively by

$$C^{1+\alpha,1+\beta}(\Gamma) \text{ and } C^{1+\alpha,2+\beta}(\Gamma).$$

Therefore, the first and second derivatives $v(t,z)$ with respect to z exist and the related norms are finite. In particular, $\exists L > 0$ such that

$$\sup_{(t,z),(t,z')\in\Gamma, z\neq z'} \frac{|v(t,z) - v(t,z')|}{|z-z'|^{\beta}} < L.$$

In the proof of Lemma 8, we applied the preceding to $v(t,z) = H_t(z)$.

Completion of the proof of Theorem 3(iii): Show that specializing in arm 2 is not asymptotically optimal if $\alpha < \overline{\alpha}$.

Verify the inequality

$$\mu_1 \frac{\sigma_3}{\sigma_1+\sigma_3} + \mu_3 \frac{\sigma_1}{\sigma_1+\sigma_3} - \alpha \frac{\sigma_1\sigma_3^2}{\sigma_1+\sigma_3} > \mu_2 - \alpha\frac{\sigma_2^2}{2} = E_P\left[u\left(\mu_2, \sigma_2 B_1^{(2)}\right)\right],$$

and deduce that

$$\alpha < \frac{2(\mu_1-\mu_2)\left[(1-\lambda)\sigma_1 + \sqrt{(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}\right]}{2(1-\lambda)\sigma_1^3 + (2\lambda-1)\sigma_1\sigma_2^2 - \sigma_2^2\sqrt{(1-\lambda)\sigma_1^2+\lambda\sigma_2^2}} \equiv 2(\mu_1-\mu_2)g(\lambda).$$

It can be verified that, $g'(\lambda) > 0$ for $\lambda \in (0,1)$ and $\lim_{\lambda\to 1} g(\lambda) = \frac{1}{\sigma_2(\sigma_1-\sigma_2)}$.

Therefore, for any $\alpha < \overline{\alpha} = \frac{2(\mu_1-\mu_2)}{\sigma_2(\sigma_1-\sigma_2)}$, there exists $\lambda_1 \in (0,1)$ such that

$$\alpha < 2(\mu_1-\mu_2)g(\lambda_1) < \frac{2(\mu_1-\mu_2)}{\sigma_2(\sigma_1-\sigma_2)}.$$

Choose $\lambda = \lambda_1$ in the definition (41) of $\hat{a} = (\hat{a}_s^{(1)}, \hat{a}_s^{(2)})$ and deduce that

$$\begin{aligned} V = &\sup_{a\in[\mathcal{A}](0,1)} E_P\left[u\left(\int_0^1 a_s^{(1)}ds, \int_0^1 a_s^{(2)}dB_s^{(2)}\right)\right] \\ \geq &E_P\left[u\left(\int_0^1 \hat{a}_s^{(1)}ds, \int_0^1 \hat{a}_s^{(2)}dB_s^{(2)}\right)\right] \\ > &E_P\left[u\left(\mu_2, \sigma_2 B_1^{(2)}\right)\right]. \end{aligned}$$

Therfore, specializing in arm 2 is NOT asymptotically optimal.

When $\sigma_2 = 0$, we can set $\overline{\alpha} = \infty$, and the above proof still holds. ■