

# Approximate Optimality and the Risk/Reward Tradeoff in a Class of Bandit Problems\*

Zengjing Chen      Larry G. Epstein      Guodong Zhang

October 26, 2022

## Abstract

This paper studies a multi-armed bandit problem where payoff distributions are known but where the riskiness of payoffs matters. The decision-maker is assumed to pursue strategies that are approximately optimal for large horizons. By exploiting the tractability afforded by asymptotics, analytical results regarding the risk/reward tradeoff are derived. The key technical tool is a new central limit theorem.

Keywords: multi-armed bandit, risk/reward tradeoff, large-horizon approximations, central limit theorem, semivariance, asymptotics

## 1 Introduction

We study the following sequential choice problem, a version of the bandit problem. There are  $K$  arms (or actions), each yielding a random payoff. Payoffs

---

\*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).

are independent across arms and for a given arm across distinct trials. At each stage  $i = 1, 2, \dots, n$ , the decision-maker (DM) chooses one arm, knowing both the realized payoffs from previously chosen arms, and the distribution of the payoff for each arm. She chooses a strategy ex ante to maximize expected utility. Because we are interested in varying horizons, we define a strategy for an infinite horizon, and then to 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  $n \rightarrow \infty$  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 bandit problem and corresponding asymptotically optimal strategies.

The bandit framework has spawned numerous applications, many of which are covered, for example, in Berry and Fristedt (1985), in the more recent textbook-like treatment of the literature by Slivkins (2022), and, for economic applications, in Bergemann and Välimäki (2008). Consider three concrete settings that fit our model well.<sup>1</sup> *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.

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, given nonindifference to risk which is our focus here. Most of the

---

<sup>1</sup>The second and third are adapted from Slivkins.

literature assumes (a finite horizon and) that choices are driven by expected total rewards. 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 the exploration/exploitation tradeoff, while we assume known distributions and focus instead on the risk/reward tradeoff.<sup>2</sup> Theorem 3 gives analytical results on the latter tradeoff by exploiting the advantages of large-horizon approximations.

A second reason for studying asymptotics is that tractability may be a concern also for the decision-maker within the model who cannot fully comprehend her extremely complicated large (but finite) horizon optimization problem. Thus, she seeks a strategy that is approximately optimal if her horizon is sufficiently long. (Accordingly, our analysis should be viewed as more descriptive than prescriptive.) The presumption that a large-horizon heuristic can alleviate cognitive limitations is supported by two 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. However, in the bandit problem, we are interested

---

<sup>2</sup>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. Note also that we will show that only the means and variances of distributions need be known.

in evaluating strategies which, in general, permit switching arms, and hence also payoff distributions, at any stage. Accordingly, in our CLT means and variances of  $(X_i)$  can vary with  $i$  subject only to the restriction that they lie in a fixed set. The CLT (Proposition 6) is the key technical result underlying our results about bandits.

The central role played by limit theorems is reflected also in our specification of the von Neumann-Morgenstern (vNM) utility index  $u$ . Two attributes of random payoff streams are assumed to be important. Accordingly,  $u : \mathbb{R}^2 \rightarrow \mathbb{R}$  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 (for example, a form of mean-variance, and another specification where variance is replaced by semivariance).

The bandit model and main results follow in the next section. Most proofs are provided in section 3. Proofs of some details are collected in the Supplementary Appendix.

## 2 The bandit 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

$$\mu_k := E_P[X_k], \quad \sigma_k^2 := Var_P[X_k], \quad 1 \leq k \leq K. \quad (1)$$

The largest and smallest means and variances are given by

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

The set of mean-variance pairs is

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

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

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

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

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

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

The dependence of  $\mathcal{H}_{n-1}^\theta$  on the strategy captures the fact that the relevant history at any stage consists not only of past payoffs but also of which arms were chosen. As an example, the strategy of alternating between arms 1 and 2, as in Theorem 3(iv), is thus rendered admissible.

The set of all admissible strategies is  $\Theta$ . (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  $\theta$  and the payoff streams that they generate. Let  $u : \mathbb{R}^2 \rightarrow \mathbb{R}$  be the corresponding von-Neumann Morgenstern (vNM) utility index and define  $U_n$  by

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

The two arguments of  $u$  correspond to the two attributes or characteristics of a random payoff stream that are taken into account. The first argument of  $u$  is the sample average outcome under strategy  $\theta$ , and the second, the  $\sqrt{n}$ -weighted average of deviations from conditional means, represents sample volatility. The presence of conditional rather than unconditional means reflects the sequential nature of the setting. As for the  $\sqrt{n}$ -weighting, as is familiar from discussions of the classical LLN and CLT, the scaling by  $\frac{1}{n}$  implies "too little" weight for finite samples, particularly when considering 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  $u$ .

**Remark** *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 constant in its second argument can be axiomatized, but imposes a priori that only means matter when choosing between arms and hence is too special (Theorem 3(iv)). Take the further special case where  $u$  is linear 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  $(\mu, \sigma^2)$  repeatedly regardless of previous outcomes. In this case payoffs are i.i.d. with mean  $\mu$  and variance  $\sigma^2$ . Thus the conditional expectation appearing in (5) equals  $\mu$ , and the classical LLN and CLT imply that in the large horizon limit risk is described by the normal distribution  $\mathbb{N}(0, \sigma^2)$  and

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

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

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

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)**  $u(x, y) = \varphi(x) + \alpha y$ . Then

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

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

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

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

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

which is a form of the classic mean-variance specification for our setting.<sup>3</sup> For any arm  $(\mu, \sigma^2)$  that is played repeatedly,

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

---

<sup>3</sup>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.4) (Mean-semivariance)  $u(x, y) = x - \alpha y^2 I_{(-\infty, 0)}(y)$ . Only negative cumulative deviations from (conditional) means are penalized. Then, given  $\theta$  and letting  $Y = \sum_{i=1}^n (Z_i^\theta - E_P[Z_i^\theta | \mathcal{H}_{i-1}^\theta])$ ,  $Var_P[Y]$  in (7) is replaced by the *semi-variance*  $E_P[Y^2 I_{Y < 0}]$ .<sup>4</sup> If  $\theta = \theta^{\mu, \sigma}$ , then

$$U_n(\theta^{\mu, \sigma}) \xrightarrow{n \rightarrow \infty} \mu - \alpha \int_{-\infty}^0 y^2 d\mathbb{N}(0, \sigma^2) = \mu - \alpha \sigma^2 / 2.$$

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

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

### 2.3 Optimization and the value of a set of arms

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

$$V_n \equiv \sup_{\theta \in \Theta} E_P U_n(\theta). \quad (10)$$

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  $\varphi$  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 (see next subsection). Then we can accommodate a much more general class of utility indices.

The first step in developing asymptotics is to define

$$V \equiv \lim_{n \rightarrow \infty} V_n. \quad (11)$$

Our first theorem proves that  $V$  is well-defined, that is, values have a limit, and more.<sup>5</sup>

---

<sup>4</sup>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.

<sup>5</sup>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  $\underline{\sigma} > 0$ . 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:**  $\lim_{n \rightarrow \infty} V_n$  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  $(\mu'_k, \sigma_k'^2)$  be defined by the obvious counterpart of (1). Then

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

Thus we can write

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

(iii) **Extreme arms are enough:**

$$V(\mathcal{A}) = V(\mathcal{A}^{ext}). \tag{12}$$

**Remark** *The assumption that  $u$  is continuous rules out example (u.5). 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 our paper (2022, section A.3), for example, where we extend from continuous functions to indicators.) Similarly for results below. Because the details are standard, we will ignore the discontinuity of (u.5).*

Section 3 provides a proof of (i), based largely on our CLT (Proposition 6), 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  $(\mu, \sigma^2)$  as an arm. (iii) describes a further possible simplification for DM – she need only consider "extreme arms", that is, the extreme points of  $co(\mathcal{A})$ , the polygon generated by  $\mathcal{A}$ . All other arms are redundant. For example, given two arms  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ , 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

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

Moreover, 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. 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  $O(n^{-1})$ , is linear in  $(\mu_k, \sigma_k^2)$ . Therefore, the value when maximizing over the  $K$  arms (asymptotically) equals that when maximizing over the convex hull  $co(\mathcal{A})$ , or over its set of extreme points  $\mathcal{A}^{ext}$ , as asserted in (12). 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 by any one of the  $K$  arm(s).<sup>6</sup>

---

<sup>6</sup>Theorem 3 (iii)-(v) and their proofs give conditions under which there are gains from

## 2.4 Strategies and the risk/reward tradeoff

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

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

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

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

Say that  $(\mu, \sigma^2)$  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  $(\bar{\mu}, \underline{\sigma}^2)$  is feasible, then: the strategy of always choosing an arm exhibiting  $(\bar{\mu}, \underline{\sigma}^2)$  is asymptotically optimal, and the corresponding limiting value, defined in (11), is given by*

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

Here  $(B_t)$  denotes a standard Brownian motion under the probability space  $(\Omega, \mathcal{F}, P)$ .

Intuition argues for the choice of  $(\bar{\mu}, \underline{\sigma}^2)$  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

---

switching. Part (v) deals with the special case where variances can be ignored, (because DM is indifferent to differences in variances), and hence the extremes are defined by the means alone.

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  $(\underline{\mu}, \underline{\sigma}^2)$  ( $(\underline{\mu}, \bar{\sigma}^2)$ ) 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 ( $K = 2$ ), where only  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$  are feasible,<sup>7</sup> and where

$$\mu_1 > \mu_2, \quad \sigma_1 > \sigma_2 > 0. \quad (13)$$

Parts (i) and (ii) 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 (11).*

---

<sup>7</sup>By Theorem 1, results would be unaffected if there were other arms lying on the straight line joining  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ . Extensions to  $K > 2$  arms are outlined briefly in the remark near the end of this section.

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

$$\partial_x u(x, y) (\mu_1 - \mu_2) + \frac{1}{2} \partial_{yy}^2 u(x, y) (\sigma_1^2 - \sigma_2^2) \geq 0 \text{ for all } (x, y) \in \mathbb{R}^2. \quad (14)$$

Then specializing in arm 1 always is asymptotically optimal and, (by (6)),  $V = \int u(\mu_1, \cdot) d\mathbb{N}(0, \sigma_1^2)$ . If  $\partial_x u$  is everywhere positive, then (14) is equivalent to

$$\frac{-\frac{1}{2} \partial_{yy}^2 u(x, y)}{\partial_x u(x, y)} \leq \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2} \text{ for all } (x, y) \in \mathbb{R}^2. \quad (15)$$

When the inequality in (14) 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

$$\frac{-\frac{1}{2} \partial_{yy}^2 u}{\partial_x u} = \alpha > 0 \text{ for all } (x, y) \in \mathbb{R}^2. \quad (16)$$

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

$$\alpha \leq (\geq) \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2}. \quad (17)$$

Both strategies are asymptotically optimal when there is equality in (17).

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

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

where the critical values  $\underline{\alpha}$  and  $\bar{\alpha}$  are given by

$$\underline{\alpha} \equiv \frac{2(\mu_1 - \mu_2)}{(\sigma_1 + 2\sigma_2)(\sigma_1 - \sigma_2)}, \quad \bar{\alpha} \equiv \frac{2(\mu_1 - \mu_2)}{\sigma_2(\sigma_1 - \sigma_2)}.$$

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  $\alpha < \bar{\alpha}$ ), then specializing in arm 1 (arm 2) is *NOT* asymptotically optimal.

(iv) Let  $u(x, y) = x - \alpha I_{(-\infty, 0)}(y)$ ,  $\alpha > 0$ . Specializing in arm 2 is not asymptotically optimal for any  $\alpha$ , 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  $\psi_i$  the number times that arm 1 is chosen in first  $i$  stages. Let the strategy  $\theta^*$  choose arm 1 at stage 1, and also at stage  $i + 1$ , ( $i \geq 1$ ), if and only if  $\frac{\psi_i}{i} \leq \lambda$ .<sup>8</sup> Then  $\theta^*$  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  $\max\{\varphi(\mu_1), \varphi(\mu_2)\} = \max_{\mu_2 \leq x \leq \mu_1} \varphi(x)$ .

We discuss each part in turn.

(i) Focus on (15). Intuition derives from interpretation of  $-\partial_{yy}^2 u / \partial_x u$  as a (local) measure of risk aversion that is a (slight) variant of the Arrow-Pratt measure (Pratt, 1964). The relatively small degree of risk aversion indicated in (15) implies that the larger mean for arm 1 more than compensates for its larger variance. Moreover, this is true contingent at each stage, regardless of history, because the inequality in (15) is satisfied globally.

Though the Arrow-Pratt argument is well-known and applies also here (with the minor extension to risks with two attributes), it might be worthwhile to couch it in our context. To do so, fix  $(x, y)$ , and let DM use the utility index  $u(x + \cdot, y + \cdot)$ . Consider the arm  $(\epsilon^2\mu, \epsilon^2\sigma^2)$ , where  $\epsilon > 0$  has the effect, when small, of scaling down both the mean and variance of payoffs by  $\epsilon^2$ . By (6), the limiting expected utility of using this arm repeatedly equals<sup>9</sup>

$$v(\epsilon, x, y) = E_P [u(x + \epsilon^2\mu, y + \epsilon\sigma B_1)].$$

Set

$$\mu = \frac{-\frac{1}{2}\partial_{yy}^2 u(x, y)}{\partial_x u(x, y)} \sigma^2. \quad (18)$$

---

<sup>8</sup>Asymptotically optimal strategies are not unique. For example, if  $\lambda = 1/2$ , then alternating between arms (deterministically), that is, choosing arms according to the sequence 121212..., is also asymptotically optimal.

<sup>9</sup> $B_1$  is the time 1 value of a standard Brownian motion, and hence is distributed as  $\mathbb{N}(0, 1)$ .

Then  $v(\epsilon, x, y) = u(x, y)$  up to the second-order in a Taylor series expansion about  $\epsilon = 0$  (hence up to the first-order in  $\epsilon^2$  or in the corresponding variance). In that sense,  $-\partial_{yy}^2 u(x, y) / \partial_x u(x, y)$  gives twice the mean-variance ratio needed to render a small risk about  $(x, y)$  asymptotically neutral.

(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 of functions  $u$  covered by (ii) are the mean-variance model (u.3) and Example (u.2) when  $\varphi$  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 asymptotically optimal given (16) 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  $-\partial_{yy}^2 u / \partial_x u$  is constant.

(iii) The mean-semivariance model agrees partially with the mean-variance model in that for both (17) implies the asymptotic optimality of choosing (the high mean, high variance) arm 1 throughout. However, their agreement ends there. In particular, for  $\underline{\alpha} < \alpha < \bar{\alpha}$ , specializing in one arm is not asymptotically optimal. Here is some intuition. Since only negative deviations are penalized, it is as though DM faces, or perceives, less risk than what is measured by  $\sigma^2$ . 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  $\alpha$  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 a single arm being dominated. Thus, for example, such specialization is not even approximately optimal in large horizons if  $\underline{\alpha} < \alpha < \bar{\alpha}$ .

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 diversification.

(iv) For this utility specification, it is never asymptotically optimal to specialize in the low mean, low variance arm. Indeed, by (9), specializing in the high mean, high variance arm is superior for large horizons, and this is true given only the ordinal assumption (13) about their means and variances. However, the latter strategy is also not asymptotically optimal for large enough  $\alpha$ , and the set of parameter values  $(\underline{\alpha}', \infty)$  where asymptotic optimality of arm 1 fails depends on the numerical values of means and variances. For example, the set grows as  $\sigma_1$  increases (keeping  $\mu_1, \mu_2$  and  $\sigma_2$  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  $\alpha$  is even smaller.

(v) Condition (14) 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 case illustrates the former factor, with the interpretation that DM is targeting  $x^*$ , a maximizer of  $\varphi$ , while being indifferent to risk. Because of the linearity of  $u(x, \cdot)$ , variances do not matter. For example, when  $\varphi$  is increasing, arm 1 is chosen always because of its larger mean, regardless of how risky it is. Nonlinearity of  $\varphi$  does not matter asymptotically as in the classic LLN.

**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

$$j \in \arg \max_{k=1, \dots, K} \left\{ \mu_k - \left( \frac{-\frac{1}{2} \partial_{yy}^2 u(x, y)}{\partial_x u} \right) \sigma_k^2 \right\} \quad \text{for all } (x, y),$$

which simplifies in the obvious way under the constancy condition (16).

In conclusion, we emphasize that *payoff distributions are unrestricted* in our model - they are *not assumed* to be adequately summarized by means and variances. That is a result (Theorem 1). Accordingly, *it is only because of our asymptotic analysis that the conditions in the above theorem giving information about the risk/reward tradeoff take such a simple form.*

### 3 Proofs

We remind the reader of the following notation used in this section:  $\bar{\mu}, \underline{\mu}$  and  $\bar{\sigma}^2, \underline{\sigma}^2$  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,  $a = \{a_s = (a_s^{(1)}, a_s^{(2)})\} : [t, T] \times \Omega \rightarrow [\mathcal{A}] \subset \mathbb{R}^2$ . 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  $n \geq 1$ ,

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

(2) For any  $\theta \in \Theta$  and  $n \geq 1$ , let  $U_{n-1}^\theta$  be any  $\theta$ -dependent (dependent only on  $(\theta_1, \dots, \theta_{n-1})$ ) and  $\mathcal{H}_{n-1}^\theta$ -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

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

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

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

**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  $\bar{\mu}$  and  $\underline{\mu}$  in (2) to derive

$$\text{ess sup}_{\theta \in \Theta} E_P[Z_n^\theta | \mathcal{H}_{n-1}^\theta] = \bar{\mu}, \quad \text{ess inf}_{\theta \in \Theta} E_P[Z_n^\theta | \mathcal{H}_{n-1}^\theta] = \underline{\mu}.$$

The other two equalities can be proven similarly.

(2) For any  $\theta \in \Theta$  and  $n \geq 1$ , let  $U_{n-1}^\theta$  be a  $\mathcal{H}_{n-1}^\theta$ -measurable random variable, which thus depends on  $(\theta_1, \dots, \theta_{n-1})$ . By direct calculation we obtain that

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

where  $\psi_k$  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:*

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

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

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

$$G(p, q) - G(p', q') \leq G(p - p', q - q'), \quad (22)$$

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

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  $p, q, q' \in \mathbb{R}$ ,

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

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

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

Here  $\|\cdot\|_{C^{1+\beta/2, 2+\beta}([0, T-\epsilon] \times \mathbb{R}^2)}$  is a norm on  $C^{1+\beta/2, 2+\beta}([0, T-\epsilon] \times \mathbb{R}^2)$ , the set of (continuous and) suitably differentiable functions on  $[0, T-\epsilon] \times \mathbb{R}^2$ . (The condition (24) is due to Krylov (1987); see also Peng (2019, Ch. 2.1). Some detail is provided in the Appendix.)

### 3.1 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 (10) and (11) 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.<sup>10</sup> Then*

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

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

The boundedness assumption on utility indices excludes many interesting specifications. Therefore, Lemma 10 in the Appendix shows that the Proposition is valid for all  $u \in C(\mathbb{R}^2)$  satisfying a growth condition. The following immediate corollary is used frequently in later proofs of Theorems 2 and 3 (the Appendix contains a proof).

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

$$V = v(0, 0, 0), \quad (27)$$

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

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

**Remark** *There is related literature on CLTs. 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 specific functional forms for  $u$ , special cases of Example (u.2). 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*

---

<sup>10</sup>The extension to  $\underline{\sigma} = 0$  is valid. A proof is available on request.

$V$ , it is the basis for a range of results about the bandit problem even allowing unrestricted heterogeneity across arms. (To our knowledge, our earlier paper (2022) is the only other paper to apply a nonlinear CLT to study bandit problems, though subject to the restrictions noted above.<sup>11</sup>) Peng (2007, 2019) and Fang et al (2019) prove nonlinear CLTs that are motivated by robustness to ambiguity (see Theorem 2.4.8 in Peng (2019), for example). The connection to sequential decision-making is not addressed, for example, strategies do not appear in their formulation. Another difference is their adoption of a "sublinear expectation space" framework, while we work within the standard and more familiar probability space framework.

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  $(x, y) \in \mathbb{R}^2$ ,

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

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  $|z - z'|^\beta = |z_1 - z'_1|^\beta + |z_2 - z'_2|^\beta$ .

**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  $(z_1, z_2), (z'_1, z'_2) \in \mathbb{R}^2$ ,

$$|\partial_{z_i z_j}^2 H_t(z_1, z_2) - \partial_{z_i z_j}^2 H_t(z'_1, z'_2)| \leq L(|z_1 - z'_1|^\beta + |z_2 - z'_2|^\beta), \quad i, j = 1, 2.$$

---

<sup>11</sup>In particular, they adopt (u.2), with  $\alpha = 1$  and  $\varphi$  having the form  $\varphi(y) = \varphi_1(y - c)$  if  $y \geq c$ , and  $= -\lambda^{-1} \varphi_1(-\lambda(y - c))$  if  $y < c$ , for some function  $\varphi_1$  and  $c \in \mathbb{R}$ . This functional form is motivated by loss aversion, but from the perspective of this paper is very special.

(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

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

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

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

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

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

**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

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

(For the reader's convenience, we include the definition of the norm in the Appendix.) 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,

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

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

where  $C$  is a constant that depends only on  $\bar{\mu}, \underline{\mu}, \bar{\sigma}^2$ , 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

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

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

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

**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<sup>12</sup>

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

<sup>12</sup>Again,  $z = (z_1, z_2) = (x, y)$ .

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

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

Then

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

**Proof:** We need only prove

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

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

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

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

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

$$\begin{aligned} & \sup_{t \in [0, 1]} \sup_{z \in \mathbb{R}^2} |\partial_{z_i z_j}^2 H_t(z)| \leq C, \\ & \sup_{t \in [0, 1]} \sup_{z, z' \in \mathbb{R}^2, z \neq z'} \frac{|\partial_{z_i z_j}^2 H_t(z) - \partial_{z_i z_j}^2 H_t(z')|}{|z - z'|^\beta} \leq C, \quad i, j = 1, 2. \end{aligned}$$

It follows from Taylor's expansion that  $\forall \epsilon > 0 \exists \delta > 0$  (depending only on  $C$  and  $\epsilon$ ), such that  $\forall z, z' \in \mathbb{R}^2$ , and  $\forall t \in [0, 1]$ ,<sup>13</sup>

$$\begin{aligned} & |H_t(z + z') - H_t(z) - D_z H_t(z) z' - \frac{1}{2} \text{tr} (z'^\top D_z^2 H_t(z) z')| \\ & \leq \epsilon |z'|^2 I_{\{|z'| < \delta\}} + 2C |z'|^2 I_{\{|z'| \geq \delta\}}. \end{aligned} \quad (34)$$

<sup>13</sup>Here  $D_z := (\partial_{z_i})_{i=1}^2$  and  $D_z^2 := (\partial_{z_i z_j}^2)_{i, j=1}^2$ .



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

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

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

Combine with Lemma 4 and show that

$$\begin{aligned}
& e(m, n) \\
& = \sup_{\theta \in \Theta} E_P \left[ H_m^m \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) + \partial_{z_1} H_m^m \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \frac{Z_m^\theta}{n} \right. \\
& \quad \left. + \partial_{z_2 z_2}^2 H_m^m \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \frac{(\bar{Z}_m^\theta)^2}{2n} \right] \\
& = \sup_{\theta \in \Theta} E_P \left[ H_m^m \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) + \max_{1 \leq k \leq K} E_P \left[ \partial_{z_1} H_m^m \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \frac{\mu_k}{n} \right. \right. \\
& \quad \left. \left. + \partial_{z_2 z_2}^2 H_m^m \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \frac{\sigma_k^2}{2n} \right] \right] \\
& = \sup_{\theta \in \Theta} E_P \left[ L_{m,n} \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \right].
\end{aligned}$$

This proves (33), and completes the proof of (31).  $\blacksquare$

**Proof of Proposition 6:** We prove it for  $u \in C_b^\infty(\mathbb{R}^2)$ . This suffices because any  $u \in C_{b,Lip}(\mathbb{R}^2)$  can be approximated uniformly by a sequence of functions in  $C_b^\infty(\mathbb{R}^2)$  (see Approximation Lemma in Feller (1971, Ch. VIII)).

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

$$\begin{aligned}
& \sup_{\theta \in \Theta} E_P \left[ H_1 \left( \frac{S_n^\theta}{n}, \frac{\bar{S}_n^\theta}{\sqrt{n}} \right) \right] - H_0(0, 0) \\
&= \sum_{m=1}^n \left\{ \sup_{\theta \in \Theta} E_P \left[ H_{\frac{m}{n}} \left( \frac{S_m^\theta}{n}, \frac{\bar{S}_m^\theta}{\sqrt{n}} \right) \right] - \sup_{\theta \in \Theta} E_P \left[ H_{\frac{m-1}{n}} \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \right] \right\} \\
&= \sum_{m=1}^n \left\{ \sup_{\theta \in \Theta} E_P \left[ H_{\frac{m}{n}} \left( \frac{S_m^\theta}{n}, \frac{\bar{S}_m^\theta}{\sqrt{n}} \right) \right] - \sup_{\theta \in \Theta} E_P \left[ L_{m,n} \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \right] \right\} \\
&\quad + \sum_{m=1}^n \left\{ \sup_{\theta \in \Theta} E_P \left[ L_{m,n} \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \right] - \sup_{\theta \in \Theta} E_P \left[ H_{\frac{m-1}{n}} \left( \frac{S_{m-1}^\theta}{n}, \frac{\bar{S}_{m-1}^\theta}{\sqrt{n}} \right) \right] \right\} \\
&=: I_{1n} + I_{2n}.
\end{aligned}$$

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

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

which implies that

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

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

$$\begin{aligned}
& \left| V - \sup_{a \in [\mathcal{A}](0,1)} E_P \left[ u \left( \int_0^1 a_s^{(1)} ds, \int_0^1 a_s^{(2)} dB_s^{(2)} \right) \right] \right| \\
&= \lim_{n \rightarrow \infty} \left| \sup_{\theta \in \Theta} E_P \left[ u \left( \frac{S_n^\theta}{n}, \frac{\bar{S}_n^\theta}{\sqrt{n}} \right) \right] - \sup_{a \in [\mathcal{A}](0,1)} E_P \left[ u \left( \int_0^1 a_s^{(1)} ds, \int_0^1 a_s^{(2)} dB_s^{(2)} \right) \right] \right| \\
&\leq \lim_{n \rightarrow \infty} \left| \sup_{\theta \in \Theta} E_P \left[ \varphi \left( \frac{S_n^\theta}{n}, \frac{\bar{S}_n^\theta}{\sqrt{n}} \right) \right] - \sup_{\theta \in \Theta} E_P \left[ H_1 \left( \frac{S_n^\theta}{n}, \frac{\bar{S}_n^\theta}{\sqrt{n}} \right) \right] \right| \\
&\quad + \lim_{n \rightarrow \infty} \left| \sup_{\theta \in \Theta} E_P \left[ H_1 \left( \frac{S_n^\theta}{n}, \frac{\bar{S}_n^\theta}{\sqrt{n}} \right) \right] - H_0(0, 0) \right| \\
&\quad + \left| H_0(0, 0) - \sup_{a \in [\mathcal{A}](0,1)} E_P \left[ u \left( \int_0^1 a_s^{(1)} ds, \int_0^1 a_s^{(2)} dB_s^{(2)} \right) \right] \right| \leq C_0 h,
\end{aligned}$$

where the constant  $C_0$  depends only on  $\underline{\mu}, \bar{\mu}, \bar{\sigma}$  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

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

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

**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 Appendix (Lemma 10).  $\blacksquare$

## 3.2 Proof of Theorem 2

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

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

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

Then

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

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

$$\begin{cases} \partial_t v(t, x, y) + \bar{\mu} \partial_x v(t, x, y) + \frac{1}{2} \underline{\sigma}^2 \partial_{yy}^2 v(t, x, y) = 0, & (t, x, y) \in [0, 1) \times \mathbb{R}^2 \\ v(1, x, y) = u(x, y). \end{cases} \quad (36)$$

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,

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

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

$$\sup_{(\mu, \sigma^2) \in \mathcal{A}} \left\{ \mu \partial_x v + \frac{1}{2} \sigma^2 \partial_{yy}^2 v \right\} = \bar{\mu} \partial_x v + \frac{1}{2} \underline{\sigma}^2 \partial_{yy}^2 v,$$

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

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

■

### 3.3 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

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

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

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

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

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

$$\begin{cases} \partial_t \hat{v}(t, x, y) + \mu_1 \partial_x \hat{v}(t, x, y) + \frac{1}{2} \sigma_1^2 \partial_{yy}^2 \hat{v}(t, x, y) = 0, & (t, x, y) \in [0, 1) \times \mathbb{R}^2 \\ \hat{v}(1, x, y) = u(x, y). \end{cases} \quad (38)$$

From (37) and assumption (14), it follows that, for all  $(t, x, y) \in [0, 1) \times \mathbb{R}^2$ ,

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

that is,

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

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

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

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

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

Hence  $\theta^*$  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,

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

Proceed as in the proof of (i).<sup>14</sup>

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

$$E_P \left[ u \left( \int_0^1 \hat{a}_s^{(1)} ds, \int_0^1 \hat{a}_s^{(2)} dB_s^{(2)} \right) \right] > \max_{i=1,2} E_P \left[ u \left( \mu_i, \sigma_i B_1^{(2)} \right) \right], \quad (40)$$

for some  $\hat{a} = (\hat{a}_s^{(1)}, \hat{a}_s^{(2)}) \in [\mathcal{A}](0, 1)$ . Then Proposition 6 implies that

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

---

<sup>14</sup>If we assume the reverse inequality in (17), 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  $\alpha > (\mu_1 - \mu_2) / (\sigma_1^2 - \sigma_2^2)$ .

Take

$$(\hat{a}_s^{(1)}, \hat{a}_s^{(2)}) = (\mu_1, \sigma_1)I_{\{W_s^{\sigma_1, \sigma_2} \geq 0\}} + (\mu_2, \sigma_2)I_{\{W_s^{\sigma_1, \sigma_2} < 0\}}, \quad (41)$$

where  $W_s^{\sigma_1, \sigma_2}$  is an oscillating Brownian motion, that is, the solution of the stochastic differential equation (SDE)

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

By Keilson and Wellner (1978, Theorem 1), the probability density of  $W_t^{\sigma_1, \sigma_2}$  is  $q(t, \cdot)$ , where

$$q(t, y) = \begin{cases} q^*(y; \sigma_1^2 t) \left[ \frac{2\sigma_2}{\sigma_1 + \sigma_2} \right] & y \geq 0 \\ q^*(y; \sigma_2^2 t) \left[ \frac{2\sigma_1}{\sigma_1 + \sigma_2} \right] & y < 0 \end{cases} \quad (42)$$

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

$$\begin{aligned} & E_P \left[ u \left( \int_0^1 \hat{a}_s^{(1)} ds, \int_0^1 \hat{a}_s^{(2)} dB_s^{(2)} \right) \right] \\ &= E_P \left[ \int_0^1 \left( \mu_1 I_{\{W_s^{\sigma_1, \sigma_2} \geq 0\}} + \mu_2 I_{\{W_s^{\sigma_1, \sigma_2} < 0\}} \right) ds \right] - \alpha E_P \left[ (W_1^{\sigma_1, \sigma_2})^2 I_{\{W_1^{\sigma_1, \sigma_2} \leq 0\}} \right] \\ &= \mu_1 \int_0^1 P(W_s^{\sigma_1, \sigma_2} \geq 0) ds + \mu_2 \int_0^1 P(W_s^{\sigma_1, \sigma_2} < 0) ds - \alpha \int_{-\infty}^0 y^2 q(1, y) dy \\ &= \mu_1 \int_0^1 \int_0^\infty q(s, y) dy ds + \mu_2 \int_0^1 \int_{-\infty}^0 q(s, y) dy ds - \alpha \int_{-\infty}^0 y^2 q(1, y) dy \\ &= \mu_1 \frac{\sigma_2}{\sigma_1 + \sigma_2} + \mu_2 \frac{\sigma_1}{\sigma_1 + \sigma_2} - \alpha \frac{\sigma_1 \sigma_2^2}{\sigma_1 + \sigma_2}. \end{aligned}$$

Therefore, (40) is satisfied if and only if

$$\underline{\alpha} = \frac{2(\mu_1 - \mu_2)}{(\sigma_1 + 2\sigma_2)(\sigma_1 - \sigma_2)} < \alpha < \frac{2(\mu_1 - \mu_2)}{\sigma_2(\sigma_1 - \sigma_2)} = \bar{\alpha}. \quad (43)$$

Proofs for other assertions regarding cases  $\underline{\alpha} < \alpha$  and  $\alpha < \bar{\alpha}$  are apparent from the above.

**Proof of (iv):** The proof is similar to that for (iii). Specifically, prove that (40) is satisfied for the process  $(\hat{a}_s^{(1)}, \hat{a}_s^{(2)})$  if  $\alpha$  satisfies the asserted inequality  $\underline{\alpha}' < \alpha$ , 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

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

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

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

Therefore,  $\theta^*$  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  $\lim_{n \rightarrow \infty} U_n(\theta^{\mu, \sigma}) = \varphi(\mu)$  for each  $(\mu, \sigma^2)$ . ■

## References

- [1] Aivaliotis, G. and J. Palczewski (2010). Tutorial for viscosity solutions in optimal control of diffusions. Available at SSRN 1582548.
- [2] Bergemann, D. and J. Välimäki (2008). Bandit problems. In Palgrave Macmillan (eds.) *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, London.
- [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 (2022). A central limit theorem, loss aversion and multi-armed bandits. arXiv:2106.05472v2 [math.PR].
- [7] 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.
- [8] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol.II. Second Edition. John Wiley and Sons, New York.
- [9] Keilson, J. and J.A. Wellner (1978). Oscillating Brownian motion. *J. Appl. Probab.* 15(2), 300-310.
- [10] Klebaner, F., Z. Landsman, U. Makov and J. Yao (2017). Optimal portfolios with downside risk. *Quant. Finan.* 17, 315-325.
- [11] Krylov, N.V.: (1987) *Nonlinear Parabolic and Elliptic Equations of the Second Order*. Reidel. Original Russian version by Nauka, Moscow (1985).
- [12] Mao, X. (2008). *Stochastic Differential Equations and Applications*. Woodhead Publishing.
- [13] Markowitz H. (1959). *Portfolio Selection*. Yale U. Press, New Haven.
- [14] Nantell, T.J. and B. Price (1979). An analytical comparison of variance and semivariance capital market theories. *J. Finan. Quant. Anal.* 14, 221-242.
- [15] 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](https://doi.org/10.1007/978-3-540-70847-6_25).



- [16] Peng, S. (2019). *Nonlinear Expectations and Stochastic Calculus under Uncertainty: with Robust CLT and G-Brownian Motion*. Springer Nature.
- [17] Pham, H. (2009). *Continuous-Time Stochastic Control and Optimization with Financial Applications* (vol. 61). Springer Science & Business Media.
- [18] Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica* 32(1/2), 122-136.
- [19] Sani, A., A. Lazaric and R. Munos (2013). Risk-aversion in multi-armed bandits. arXiv:1301.1936v1 [cs.LG].
- [20] Slivkins, A. (2022). *Introduction to Multi-Armed Bandits*. arXiv:1904.07272v7 [cs.LG].
- [21] 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.
- [22] Yong, J. and X.Y. Zhou (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations* (vol. 43). Springer Science & Business Media.
- [23] 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.
- [24] Zimin, A., R. Ibsen-Jensen and K. Chatterjee (2014). Generalized risk-aversion in stochastic multi-armed bandits. arXiv:1405.0833 [cs.LG].

## A Supplementary Appendix

**Lemma 10** *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

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

Therefore,

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

which proves the CLT since  $\epsilon$  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

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

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

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

Since  $N$  can be arbitrarily large, it suffices to prove

$$\sup_n \sup_{\theta \in \Theta} E_P \left[ \left| \frac{S_n^\theta}{n} \right|^g + \left| \frac{\bar{S}_n^\theta}{\sqrt{n}} \right|^g \right] < \infty \quad (44)$$

Step 3: Prove (44). 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  $1 \leq g \leq 2$ ,

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

For  $g > 2$ ,

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

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

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

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)),<sup>15</sup>

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

Therefore,

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

Finally,

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

Since  $\sup_{1 \leq k \leq K} E_P[|X_k|^g] < \infty$ , (44), and subsequently also the Lemma, are proven.  $\blacksquare$

**Proof of Corollary 7:** Lemma 10 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

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

<sup>15</sup>  $|ab| \leq p^{-1} |a|^p + q^{-1} |a|^q$  if  $1 < p, q < \infty$  and  $p^{-1} + q^{-1} = 1$ .

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:** We use the notation in Krylov (1987, Section 1.1); see also in Peng (2019, Chapter 2.1). For  $\Gamma$  be a subset of  $[0, \infty) \times \mathbb{R}^2$ ,  $C(\Gamma)$  denotes all continuous functions  $v$  defined on  $\Gamma$ , in the relative topology on  $\Gamma$ , with a finite norm,

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

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

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

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

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

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  $v(t, z) = H_t(z)$ .