Choquet regularization for continuous-time reinforcement learning

Xia Han* Ruodu Wang† Xun Yu Zhou‡

7th May 2023

Abstract

We propose *Choquet regularizers* to measure and manage the level of exploration for reinforcement learning (RL), and reformulate the continuous-time entropy-regularized RL problem of Wang et al. (2020a) in which we replace the differential entropy used for regularization with a Choquet regularizer. We derive the Hamilton–Jacobi–Bellman equation of the problem, and solve it explicitly in the linear–quadratic (LQ) case via maximizing statically a mean–variance constrained Choquet regularizer. Under the LQ setting, we derive explicit optimal distributions for several specific Choquet regularizers, and conversely identify the Choquet regularizers that generate a number of broadly used exploratory samplers such as ε-greedy, exponential, uniform and Gaussian.

KEYWORDS: Reinforcement learning, Choquet integrals, continuous time, exploration, regularizers, quantile, HJB equations, linear–quadratic control.

1 Introduction

Reinforcement learning (RL) is one of the most active and fast developing subareas in machine learning. The foundation of RL is "trial and error" – to *strategically* explore different action plans in order to find the best plan as efficiently and economically as possible. A key question to this inherent exploratory approach for RL is to seek a proper tradeoff between exploration and exploitation, for which one needs to first quantify the level of exploration. Because exploration is typically captured by randomization in the RL study, entropy has been employed to measure the magnitude of the randomness and hence that of the exploration – a uniform distribution representing a completely blind search has the maximum entropy while a Dirac mass signifying no exploration at all has the minimum entropy. Discrete-time entropy-regularized (or "softmax") RL formulation has been proposed which introduces a weighted entropy value of the exploration as a regularization term

*School of Mathematical Sciences and LPMC, Nankai University, China. Email: xiahan@nankai.edu.cn
†Department of Statistics and Actuarial Science, University of Waterloo, Canada. Email: wang@uwaterloo.ca
‡Department of IEOR, Columbia University, USA. Email: xz2574@columbia.edu.

into the objective function (Haarnoja et al. (2018), Nachum et al. (2017) and Ziebart et al. (2008)). For continuous-time RL, Wang et al. (2020a) formulate an entropy-regularized, distribution-valued stochastic control problem for diffusion processes, and derive theoretically the Gibbs (or Boltzmann) measure as the optimal distribution for exploration which specializes to Gaussian when the problem is linear–quadratic (LQ). Gao et al. (2022) and Wang and Zhou (2020) apply the results of Wang et al. (2020a) to a Langevin diffusion for simulated annealing and a continuous-time entropy-regularized Markowitz's mean–variance portfolio selection problem, respectively. Guo et al. (2020) analyze both quantitatively and qualitatively the impact of entropy regularization for mean-field games with learning in a finite time horizon. There have been recently many other developments along this direction of RL in continuous time; see Jia and Zhou (2022a,b,c), Mou et al. (2021) and Tang et al. (2022) and the references therein.

While the entropy is a reasonable metric to quantify the information gain of exploring the environment and entropy regularization can indeed explain some broadly used exploration distributions such as Gaussian, there are two closely related open questions:

1. Distributions other than Gaussian, such as exponential or uniform, are also widely used for exploration in RL. What regularizer(s) can theoretically justify the use of a given class of exploratory distributions?

2. What are the optimal exploratory distributions for regularizers other than the entropy?

In this paper, we study these two questions in the setting of continuous-time diffusion processes, by introducing a new class of regularizers borrowing from the literature of risk metrics. Risk metrics, roughly speaking, include risk measures and variability measures, which are two separate and active research streams in the general area of risk management. Value-at-risk (VaR), expected shortfall (ES) and various coherent or convex risk measures, introduced by Artzner et al. (1999), Delbaen (2002) and Föllmer and Schied (2002), are popular examples of risk measures. Variance, the Gini deviation, interquantile range and deviation measures of Rockafellar et al. (2006) are instances of variability measures. There has been a rich body of study on risk metrics in the past two decades; see Föllmer and Schied (2016) and the references therein.

We introduce what we call *Choquet regularizers*, which belong to the class of the signed Choquet integrals recently studied by Wang et al. (2020c) in the context of risk management. A signed Choquet integral in general gives rise to a nonlinear and non-monotone expectation in which the state of nature is weighted by a probability weighting or distortion function in calculating the expectation. It includes as special cases Yaari's dual utility (Yaari (1987)) and distortion risk measures (Acerbi (2002) and Kusuoka (2001)), which are commonly used monotone functionals, and appears in rank-dependent utility (RDU) theory; see De Waegenaere and Wakker (2001), Gilboa and Schmeidler (1989), Quiggin (1982) and Tversky and Kahneman (1992) in the related literature of behavioral economic theory.

There are several reasons to use Choquet regularizers for RL due to a number of theoretical and practical advantages. First, they satisfy several "good" properties such as quantile additivity, normalization, concavity, and consistency with convex order (mean-preserving spreads) that facilitate analysis as regularizers. Second, Choquet regularizers are non-monotone. In order to measure exploration, monotonicity is irrelevant, in contrast to assessing risk or wealth. For instance, a degenerate distribution should be associated with no-exploration regardless of its position, in which case non-monotone mappings should be used. Moreover, the use of Choquet regularizers is closely connected to distributionally robust optimization (DRO) where a Wasserstein distance naturally induces a special class of Choquet regularizers, whereas DRO itself is an important approach for learning and for correcting the inherent flaws suffered by classical model-based estimation and optimization. Finally, as we will see later in the paper, for any given location–scale class of distributions, there exists a common Choquet regularizer such that the corresponding regularized continuous-time LQ control for RL has optimal distributions in that class.

We take the same continuous-time exploratory stochastic control problem as in Wang et al. (2020a), except that the entropy regularizer is replaced with a Choquet regularizer. In the general case we derive the Hamilton–Jacobi–Bellman (HJB) equation. However, in sharp contrast to Wang et al. (2020a) in which the optimal control distributions are proved to be Gibbs measures, obtaining the class of optimal distributional policies via verification theorem remains a significant open question. To obtain explicit solutions, we focus on the LQ case. The form of the LQ-specialized HJB equation suggests that the problem boils down to a static optimization in which the given Choquet regularizer is to be maximized over distributions with given mean and variance. It turns out this last problem has been solved explicitly by Liu et al. (2020). The optimal distributions form a location–scale family, whose shape depends on the choices of the Choquet regularizer. The solutions to the static problem are then employed to solve the original LQ-based exploratory HJB equation explicitly and to derive the optimal samplers for exploration under the given Choquet regularizer. As expected, optimal distributions are no longer necessarily Gaussian as in Wang et al. (2020a), and are now dictated by the choice of Choquet regularizers. However, the following feature of the entropy-regularized solutions revealed in Wang et al. (2020a) remains intact: the means of the optimal distributions are linear in the current state and independent of the exploration, whereas the variances are determined by the exploration but irrespective of the current state. Along an opposite line of inquiry, we are able to identify a proper Choquet regularizer in order to interpret a given exploratory distribution. Specifically, we derive the regularizers that generate some common exploration measures including ε-greedy, three-point, exponential, uniform and Gaussian.

The rest of the paper is organized as follows. We introduce Choquet regularizers in Section 2, and present their basic properties as well as an axiomatic characterization based on existing results of Wang et al. (2020b,c). In Section 3, we formulate the continuous-time Choquet-regularized RL control problem and derive the HJB equation. We then motivate a mean–variance constrained

Choquet regularizer maximization problem for LQ control. This problem is studied in details in Section 4, including discussions on some special regularizers arising from problems in finance, optimization, and risk management. In Section 5, we return to the exploratory LQ control problem and solve it completely. We also present examples linking specific exploratory distributions with the corresponding Choquet regularizers. In Section 6, we discuss the connections between the exploratory LQ problem and the classical LQ problem. Finally, Section 7 concludes the paper.

2 Choquet regularizers

Throughout the paper, we assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is an atomless probability space. With a slight abuse of notation, let $\mathcal{M}$ denote both the set of (probability) distribution functions of real random variables and the set of Borel probability measures on $\mathbb{R}$, with the obvious identity $\Pi(x) \equiv \Pi((-\infty, x])$ for $x \in \mathbb{R}$ and $\Pi \in \mathcal{M}$. We denote by $\mathcal{M}^p \subset \mathcal{M}$, $p \in [1, \infty)$, the set of distribution functions or probability measures with finite p-th moment. For a random variable X and a distribution Π, we write $X \sim \Pi$ if the distribution of X is Π under $\mathbb{P}$, and $X \overset{\mathrm{d}}{=} Y$ if two random variables X and Y have the same distribution. We denote by μ and σ^2 the mean and variance functionals on $\mathcal{M}^2$, respectively; that is, $\mu(\Pi)$ is the mean of Π and $\sigma^2(\Pi)$ the variance of Π for $\Pi \in \mathcal{M}^2$.

Given a function $h : [0, 1] \to \mathbb{R}$ of bounded variation with $h(0) = 0$ and $\Pi \in \mathcal{M}$, the functional I_h on $\mathcal{M}$ is defined as

$$I_h(\Pi) \equiv \int h \circ \Pi([x, \infty)) \mathrm{d}x := \int_{-\infty}^{0} [h \circ \Pi([x, \infty)) - h(1)] \, \mathrm{d}x + \int_0^{\infty} h \circ \Pi([x, \infty)) \mathrm{d}x, \tag{2.1}$$

assuming that (2.1) is well defined (which could take the value ∞). The function h is called a *distortion function*, and the functional I_h is called a *signed Choquet integral* by Wang et al. (2020c). If $h(x) \equiv x$ then I_h reduces to the mean functional; thus, I_h is a nonlinear generalization of the mean/expecation. If h is increasing and satisfies $h(0) = 1 - h(1) = 0$, then I_h is called an *increasing Choquet integral*, which include as special cases the two most important risk measures used in current banking and insurance regulation, VaR and ES.[1]

Next, we define the *Choquet regularizer*, a main object of this paper. We are particularly interested in a subclass of signed Choquet integrals, where h satisfies the following properties:

(i) h is concave;

(ii) $h(1) = h(0) = 0$.

[1]This functional I_h is termed differently in different fields. For example, it is known as Yaari's dual utility (Yaari (1987)) in decision theory, distorted premium principles (Denneberg (1994) and Wang et al. (1997)) in insurance and distortion risk measures (Acerbi (2002) and Kusuoka (2001)) in finance.

Let us briefly explain the interpretations and implications of the above two conditions. Condition (i) is equivalent to several other properties, and in particular, to that I_h is a concave mapping and to that I_h is consistent with *convex order*;[2] see Theorem 3 of Wang et al. (2020c) for this claim and several other equivalent properties. Here, concavity of I_h means

$$I_h(\lambda\Pi_1 + (1-\lambda)\Pi_2) \geqslant \lambda I_h(\Pi_1) + (1-\lambda) I_h(\Pi_2), \quad \text{for all } \Pi_1, \Pi_2 \in \mathcal{M} \text{ and } \lambda \in [0,1],$$

and consistency with convex order means

$$I_h(\Pi_1) \leqslant I_h(\Pi_2), \quad \text{for all } \Pi_1, \Pi_2 \in \mathcal{M} \text{ with } \Pi_1 \preceq_{\mathrm{cx}} \Pi_2.$$

If $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$, then Π_2 is also called a *mean-preserving spread* of Π_1 (Rothschild and Stiglitz (1970)), which intuitively means that Π_2 is more spread-out (and hence "more random") than Π_1. The above two properties do indeed suggest that $I_h(\Pi)$ serves as a measure of randomness for Π, since both a mixture and a mean-preserving spread introduce extra randomness; see e.g., Acciaio and Svindland (2013) for related discussions. Condition (ii), on the other hand, is equivalent to $I_h(\delta_c) = 0\ \forall c \in \mathbb{R}$, where δ_c is the Dirac mass at c. That is, degenerate distributions do not have any randomness measured by I_h.

Definition 2.1. *Let $\mathcal{H}$ be the set of $h : [0,1] \to \mathbb{R}$ satisfying (i)-(ii). A functional $\Phi : \mathcal{M} \to (-\infty, \infty]$ is a* Choquet regularizer *if there exists $h \in \mathcal{H}$ such that $\Phi = I_h$, that is,*

$$\Phi(\Pi) = \int_{\mathbb{R}} h \circ \Pi([x, \infty)) \mathrm{d}x, \tag{2.2}$$

and this Φ will henceforth be denoted by Φ_h.

To clarify on notation, we require $h \in \mathcal{H}$ for Φ_h, while there is no such requirement for I_h. Moreover, we call Φ_h to be location invariant and scale homogeneous if $\Phi_h(\Pi') = \lambda \Phi_h(\Pi)$ where Π' is the distribution of $\lambda X + c$ for $\lambda > 0$, $c \in \mathbb{R}$ and $X \sim \Pi$.

We summarize some useful properties of Φ_h in the following lemma.

Lemma 2.2. *For $h \in \mathcal{H}$, Φ_h is well defined, non-negative, and location invariant and scale homogeneous.*

Proof. First, a concave h with $h(0) = h(1)$ has to be first increasing and then decreasing on $[0,1]$. Hence h has bounded variation, and the two integrals in (2.1) are well defined. Moreover, (i) and (ii) imply $h \geqslant 0$, which further yields that both terms in (2.1) are non-negative. So Φ_h is well defined and non-negative. Location invariance and scale homogeneity follow from Proposition 2 (iii) and (iv) of Wang et al. (2020b). □

[2] Let Π_1 and Π_2 be two distribution functions with finite means. Then, Π_1 is smaller than Π_2 in *convex order*, denoted by $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$, if $\mathbb{E}[f(\Pi_1)] \leqslant \mathbb{E}[f(\Pi_2)]$ for all convex functions f, provided that the two expectations exist. It is immediate that $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$ implies $\mathbb{E}[\Pi_1] \leqslant \mathbb{E}[\Pi_2]$.

Each property in Lemma 2.2 has a simple interpretation for a regularizer that measures the level of randomness, or the level of exploration in the RL context of this paper.

(a) Well-posedness: Any distribution for exploration can be measured.[3]

(b) Non-negativity: Randomness is measured in non-negative values.

(c) Location invariance: The measurement of randomness does not depend on the location.

(d) Scale homogeneity: The measurement of randomness is linear in its scale.

For a distribution $\Pi \in \mathcal{M}$, let its left-quantile for $p \in (0,1]$ be defined as, recalling that $\Pi(x) = \Pi((-\infty, x])$ for $x \in \mathbb{R}$,

$$Q_\Pi(p) = \inf\{x \in \mathbb{R} : \Pi(x) \geqslant p\},$$

whereas its right-quantile function for $p \in [0,1)$ be defined as

$$Q_\Pi^+(p) = \inf\{x \in \mathbb{R} : \Pi(x) > p\}.$$

It is useful to note that Φ_h admits a quantile representation as follows; see Lemma 1 of Wang et al. (2020b).

Lemma 2.3. *For $h \in \mathcal{H}$ and $\Pi \in \mathcal{M}$,*

(i) if h is right-continuous, then $\Phi_h(\Pi) = \int_0^1 Q_\Pi^+(1-p)\mathrm{d}h(p)$;

(ii) if h is left-continuous, then $\Phi_h(\Pi) = \int_0^1 Q_\Pi(1-p)\mathrm{d}h(p)$;

(iii) if Q_Π is continuous, then $\Phi_h(\Pi) = \int_0^1 Q_\Pi(1-p)\mathrm{d}h(p)$.

Choquet regularizers include, for instance, range, mean-median deviation, the Gini deviation, and inter-ES differences. Moreover, standard deviation can be written as the supremum of Choquet regularizers; see Examples 1, 3 and 4 of Wang et al. (2020c). Variance also has a related representation (Example 2.2 of Liu et al. (2020)):

$$\sigma^2(\Pi) = \sup_{h \in \mathcal{H}} \left\{ \Phi_h(\Pi) - \frac{1}{4}||h'||_2^2 \right\}, \quad \Pi \in \mathcal{M},$$

where $||h'||_2^2 = \int_0^1 (h'(p))^2 \mathrm{d}p$ if h is continuous with a.e. right-derivative h', and $||h'||_2^2 := \infty$ if h is not continuous.

[3]This property is technically important since functional properties of I_h could be very difficult to analyze if one faces a quantity such as $\infty - \infty$. As an example, consider $h(x) = x$ leading to I_h being the mean functional. In this case, I_h is only well defined on some subsets of $\mathcal{M}$.

Concave signed Choquet integrals are characterized by, e.g., Wang et al. (2020c), which is essentially a consequence of the seminal works of Schmeidler (1989) and Yaari (1987); see also Theorem 2.4 below. In what follows, we say that $\Phi = \Phi_h$ is *quantile additive* if for all $\Pi_1, \Pi_2 \in \mathcal{M}$, $\Phi(\Pi_1 \oplus \Pi_2) = \Phi(\Pi_1) + \Phi(\Pi_2)$ where the quantile function of $\Pi_1 \oplus \Pi_2$ is the sum of those of Π_1 and Π_2. In other words, $Q_{\Pi_1 \oplus \Pi_2} = Q_{\Pi_1} + Q_{\Pi_2}$. Moreover, we say that Φ is *continuous at infinity* if $\lim_{M \to 1} \Phi((\Pi \wedge M) \vee (1 - M)) = \Phi(\Pi)$, and Φ is *uniform sup-continuity* if for any $\varepsilon > 0$, there exists $\delta > 0$, such that $|\Phi(\Pi_1) - \Phi(\Pi_2)| < \varepsilon$ whenever ess-sup$|\Pi_1 - \Pi_2| < \delta$, where ess-sup is the essential supremum defined by $\Pi^{-1}(1)$.

We give the following simple characterization for our Choquet regularizers based on Theorems 1 and 3 of Wang et al. (2020b).

Theorem 2.4. *A functional Φ_h is a Choquet regularizer in (2.2) if and only if it satisfies all of the following properties*

(i) Φ_h *is quantile additive;*

(ii) Φ_h *is concave or* $\preceq_{\mathrm{cx}}$*-consistent;*

(iii) $\Phi_h \geqslant 0$ *and* $\Phi_h(\delta_c) = 0$ *for all* $c \in \mathbb{R}$*;*

(iv) Φ_h *is continuous at infinity and uniformly sup-continuous.*

Note that Theorems 1 and 3 of Wang et al. (2020b) are stated in terms of a risk measure defined on the space of real random variables, say $\mathcal{X}$, while here Φ_h is defined on $\mathcal{M}$. To use these results, we can define $\rho : \mathcal{X} \to \mathbb{R}$ by $\rho(X) = \Phi_h(\Pi)$ where $X \sim \Pi$, which is automatically law-invariant.[4] On the other hand, Theorem 1 in Wang et al. (2020b) requires an extra continuity condition to imply that h has bounded variation on $[0, 1]$, which is guaranteed here by condition (iii). In fact, condition (i) is equivalent to comonotonic additivity of ρ.[5] Continuity at infinity and uniform sup-continuity of ρ can be defined in paralell to those of Φ_h. Finally, $h(1) = h(0) = 0$ is equivalent to $\Phi_h(\delta_c) = 0$ for all $c \in \mathbb{R}$. Theorem 2.4 hence follows directly from Theorems 1 and 3 of Wang et al. (2020b).

Remark 2.5. *If h is not constantly* 0*, Choquet regularizers belong to the class of* generalized deviation measures *in Grechuk et al. (2009) and Rockafellar et al. (2006). Moreover, Choquet regularizers can be used to construct law-invariant generalized deviation measures. Indeed, combining characterization of generalized deviation measures in Proposition 2.2 of Grechuk et al. (2009) and the quantile representation of signed Choquet integrals in Lemma 2.3, all law-invariant generalized deviation measures can be represented as a supremum of some Choquet regularizers of the type (2.2). This includes standard deviation and mean absolute deviation as special cases.*

[4] Law-invariance means that $\rho(X) = \rho(Y)$ for $X \overset{\mathrm{d}}{=} Y$.

[5] A random vector $(X_1, \dots, X_n)$ is called *comonotonic* if there exists a random variable $Z \in \mathcal{X}$ and increasing functions $f_1, \dots, f_n$ on $\mathbb{R}$ such that $X_i = f_i(Z)$ almost surely for all $i = 1, \dots, n$. Comonotoic-additivity means that $\rho(X + Y) = \rho(X) + \rho(Y)$ if X and Y are comonotonic.

We conclude this section by comparing the Choquet regularization with the differential entropy regularization, the latter having been used for exploration–exploitation balance in RL; see Guo et al. (2020), Wang et al. (2020a) and Wang and Zhou (2020). For an absolutely continuous Π, we define DE, Shannon's differential entropy, as

$$\mathrm{DE}(\Pi) := -\int_{\mathbb{R}} \Pi'(x)\log(\Pi'(x))\mathrm{d}x. \tag{2.3}$$

Sunoj and Sankaran (2012) show that (2.3) admits a different quantile representation

$$\mathrm{DE}(\Pi) = \int_0^1 \log(Q'_\Pi(p))\mathrm{d}p. \tag{2.4}$$

It is clear that DE is location invariant, but not scale homogeneous. It is not quantile additive either. Therefore, DE is *not* a Choquet regularizer.

3 Exploratory control with Choquet regularizers

In this section, we first introduce an exploratory stochastic control problem for RL in continuous time and spaces which was originally proposed in Wang et al. (2020a), and then reformulate it with Choquet regularizers.

Let $\mathbb{F} = \{\mathcal{F}_t\}_{t\geqslant 0}$ be a filtration defined on $(\Omega, \mathcal{F}, \mathbb{P})$ along with an $\{\mathcal{F}_t\}_{t\geqslant 0}$-adapted Brownian motion $W = \{W_t\}_{t\geqslant 0}$, the filtered probability space satisfying the usual assumptions of completeness and right continuity. All stochastic processes introduced below are supposed to be adapted processes in this space.

The classical stochastic control problem is to control the state dynamic described by a stochastic differential equation (SDE)

$$\mathrm{d}X_t^u = b\left(X_t^u, u_t\right)\mathrm{d}t + \xi\left(X_t^u, u_t\right)\mathrm{d}W_t,\ t>0; \quad X_0^u = x \in \mathbb{R}, \tag{3.1}$$

where $u = \{u_t\}_{t\geqslant 0}$ is the control process taking value in a given action space U. The aim of the problem is to achieve the maximum expected total discounted reward represented by the value function

$$V^{\mathrm{cl}}(x) := \sup_{u\in\mathcal{A}^{\mathrm{cl}}(x)} \mathbb{E}_x\left[\int_0^\infty e^{-\rho t} r\left(X_t^u, u_t\right)dt\right], \tag{3.2}$$

where r is the reward function, $\rho > 0$ is the discount rate, and $\mathcal{A}^{\mathrm{cl}}(x)$ denotes the set of all admissible controls which in general may depend on x. Throughout this paper, for ease of notation we assume that the state and Brownian motion are scalar-valued processes. Moreover, we suppose that the control is also one-dimensional, which is however an essential assumption because the Choquet

regularizer to be involved is defined only for distributions on $\mathbb{R}$.[6]

With the complete knowledge of the model parameters, the theory for solving the classical, model-based problem (3.1)–(3.2) has been developed and established thoroughly. In the RL setting, where those parameters are partly or completely unknown and therefore dynamic learning is needed, the agent employs exploration to interact with and learn the unknown environment through trial and error. The key idea is to model exploration by a distribution of controls $\Pi = \{\Pi_t\}_{t\geqslant 0}$ over the control space U from which each "trial" is sampled. Thus, the notion of controls is extended to distributions. The agent executes controls for N rounds over the same time horizon, while at each round, a classical control is sampled from the distribution Π. The reward of such a policy becomes accurate enough when N is large.

Thus, similarly to Wang et al. (2020a), we give the "exploratory" version of the state dynamic (3.1) motivated by repetitive learning in RL. The control process is now randomized, leading to a distributional or exploratory control process $\Pi = \{\Pi_t\}_{t\geqslant 0}$, where $\Pi_t \in \mathcal{M}(U)$ is the probability distribution function for control at time t, with $\mathcal{M}(U)$ being the set of distribution functions on U. For a given such distributional control Π, the exploratory version of the state dynamics is

$$\mathrm{d}X_t^{\Pi} = \tilde{b}\left(X_t^{\Pi}, \Pi_t\right)\mathrm{d}t + \tilde{\xi}\left(X_t^{\Pi}, \Pi_t\right)\mathrm{d}W_t, \ t > 0; \quad X_0^{\Pi} = x \in \mathbb{R}, \tag{3.3}$$

where the coefficients $\tilde{b}(\cdot,\cdot)$ and $\tilde{\xi}(\cdot,\cdot)$ are defined as

$$\tilde{b}(y,\Pi) = \int_U b(y,u)\mathrm{d}\Pi(u), \quad y \in \mathbb{R}, \ \Pi \in \mathcal{M}(U), \tag{3.4}$$

and

$$\tilde{\xi}(y,\Pi) = \sqrt{\int_U \xi^2(y,u)\mathrm{d}\Pi(u)}, \quad y \in \mathbb{R}, \ \Pi \in \mathcal{M}(U). \tag{3.5}$$

The "exploratory state process" $\left\{X_t^{\Pi}\right\}_{t\geqslant 0}$ describes the average of the state processes under (infinitely) many different classical control processes sampled from the exploratory control $\Pi = \{\Pi_t\}_{t\geqslant 0}$. Further, the reward function r in (3.2) needs also to be modified to the exploratory reward

$$\tilde{r}(y,\Pi) = \int_U r(y,u)\mathrm{d}\Pi(u), \quad y \in \mathbb{R}, \ \Pi \in \mathcal{M}(U). \tag{3.6}$$

A detailed explanation of where this exploratory formulation comes from is provided in Wang et al. (2020a, pp. 6–8). We reiterate that the exploratory state process $\left\{X_t^{\Pi}\right\}_{t\geqslant 0}$ is the *average* of the sample state trajectories under infinitely many actions generated from the same distribution Π and is in itself *not* a sample state trajectory nor observable. The exploratory formulation above just provides a framework for *theoretical* analysis. See Jia and Zhou (2022b, p. 9) for more discussion on this point.

[6] See Section 7 for a discussion about how we may extend the notion of Choquet regularizer to multi-dimensions.

Next, we use a Choquet regularizer Φ_h to measure the level of exploration, and the aim of the exploratory control is to achieve the maximum expected total discounted and regularized exploratory reward represented by the optimal value function

$$V(x) = \sup_{\Pi \in \mathcal{A}(x)} \mathbb{E}_x \left[\int_0^\infty e^{-\rho t} \left(\tilde{r}(X_t^\Pi, \Pi) + \lambda \Phi_h(\Pi) \right) \mathrm{d}t \right], \tag{3.7}$$

where $\lambda > 0$ is the *temperature* parameter representing the weight on exploration, $\mathcal{A}(x)$ is the set of admissible distributional controls (which may in general depend on x), and $\mathbb{E}_x$ represents the conditional expectation given $X_0^\Pi = x$.

The precise definition of $\mathcal{A}(x)$ depends on the specific dynamic model under consideration and the specific problems one wants to solve, which may vary from case to case. We will define $\mathcal{A}(x)$ precisely later for the linear–quadratic (LQ) control case, which will be the main focus of the paper. Note that (3.7) is mathematically a so-called relaxed stochastic control problem; see Wang et al. (2020a, Footnote 7) for a detailed discussion about the connection between the exploratory formulation and relaxed control.

Controls in $\mathcal{A}(x)$ are measure (distribution function)-valued stochastic adapted processes, which are open-loop controls in the control terminology. A more important notion in RL is the feedback (control) *policy*. Specifically, a deterministic mapping $\Pi(\cdot;\cdot)$ is called a feedback policy if i) $\Pi(\cdot; x)$ is a distribution function for each $x \in \mathbb{R}$; ii) the following SDE (which is the system dynamic after the feedback law $\Pi(\cdot;\cdot)$ is applied)

$$\mathrm{d}X_t = \tilde{b}\left(X_t, \Pi\left(\cdot; X_t\right)\right) \mathrm{d}t + \tilde{\xi}\left(X_t^\Pi, \Pi\left(\cdot; X_t\right)\right) \mathrm{d}W_t, \quad t > 0; \quad X_0 = x \in \mathbb{R}$$

has a unique strong solution $\{X_t\}_{t \geqslant 0}$; and iii) the open-loop control $\Pi = \{\Pi_t\}_{t \geqslant 0} \in \mathcal{A}(x)$ where $\Pi_t := \Pi\left(\cdot; X_t\right)$. In this case, the resulting open-loop control Π is said to be *generated* from the feedback policy $\Pi(\cdot;\cdot)$ with respect to the initial state x.

On the other hand, for a continuous $h \in \mathcal{H}$, we have

$$\Phi_h(\Pi) = \int_0^1 Q_\Pi(1-p)\mathrm{d}h(p) \quad = \int_U u h'(1-\Pi(u))\mathrm{d}\Pi(u).$$

We present the general procedure for solving the problem (3.7), following Wang et al. (2020a). Applying the classical Bellman principle of optimality, we deduce that the optimal value function V satisfies the Hamilton-Jacobi-Bellman (HJB) equation

$$\rho v(x) = \max_{\Pi \in \mathcal{M}(U)} \left(\tilde{r}(x, \Pi) + \lambda \int_U u h'(1-\Pi(u))\mathrm{d}\Pi(u) + \frac{1}{2}\tilde{\xi}^2(x,\Pi)v''(x) + \tilde{b}(x,\Pi)v'(x) \right), \tag{3.8}$$

or equivalently,

$$\rho v(x) = \max_{\Pi \in \mathcal{M}(U)} \int_U \left(r(x,u) + \lambda u h'(1-\Pi(u)) + \frac{1}{2}\xi^2(x,u)v''(x) + b(x,u)v'(x) \right) \mathrm{d}\Pi(u),$$

where v denotes the generic unknown solution of the equation. The verification theorem then yields that the feedback policy Π^* defined as

$$\Pi^*(x) := \arg\max_{\Pi \in \mathcal{M}(U)} \int_U \left(r(x,u) + \lambda u h'(1-\Pi(u)) + \frac{1}{2}\xi^2(x,u)v''(x) + b(x,u)v'(x) \right) \mathrm{d}\Pi(u) \tag{3.9}$$

is an optimal policy if it generates an admissible open-loop control for any x.

When the regularizer is the entropy, Wang et al. (2020a) applied the corresponding verification theorem to conclude that the Gibbs (or Boltzmann) measures are generally optimal samplers for exploration, which specialize to Gaussian in the LQ case. However, no general study on the entropy-regularized exploratory HJB equation was available until Tang et al. (2022) established the well-posedness and regularity of its viscosity solution. With the current Choquet regularizers, studying (3.8) and solving the maximization problem in (3.9) generally remain (significant) open questions because (3.8) is very different from its entropy counterpart and it is unclear whether the analyses in Tang et al. (2022) and Wang et al. (2020a) carry over.

In this paper, we focus on the LQ setting, in which the exploratory HJB equation (3.8) can be explicitly solved, to study how different Choquet regularizers may generate the optimal policy distributions. Specifically, we consider

$$b(x,u) = Ax + Bu \quad \text{and} \quad \xi(x,u) = Cx + Du, \quad x, u \in \mathbb{R}, \tag{3.10}$$

where $A, B, C, D \in \mathbb{R}$, and

$$r(x,u) = -\left(\frac{M}{2}x^2 + Rxu + \frac{N}{2}u^2 + Px + Lu \right), \quad x, u \in \mathbb{R}, \tag{3.11}$$

where $M \geqslant 0$, $N > 0$, and $R, P, L \in \mathbb{R}$. Moreover, as in standard LQ theory we assume henceforth that $U = \mathbb{R}$ and thus write $\mathcal{M} = \mathcal{M}(U)$ and $\mathcal{M}^2 = \mathcal{M}^2(U)$.

Remark 3.1. *LQ control plays a vitally important role in the classical control literature, not only because it usually admits elegant and simple solutions, but also because more complex, nonlinear problems can be approximated by LQ problems. Indeed, one can simply apply a second-order Taylor approximation to the reward function and a first-order Taylor approximation to the dynamics coefficient functions to define an approximate LQ problem; see Benigno and Woodford (2003, 2012), Judd (1998), Li and Todorov (2007), Todorov and Li (2005) and the reference therein for more details.*

Fix an initial state $x \in \mathbb{R}$. For each open-loop control $\Pi \in \mathcal{A}(x)$, denote its mean and variance

processes $\{\mu_t\}_{t\geqslant 0}$ and $\{\sigma_t^2\}_{t\geqslant 0}$ by

$$\mu_t \equiv \mu(\Pi_t) = \int_U u \mathrm{d}\Pi_t(u) \qquad \text{and} \qquad \sigma_t^2 \equiv \sigma^2(\Pi_t) = \int_U u^2 \mathrm{d}\Pi_t(u) - \mu_t^2.$$

By (3.4) and (3.5), we have

$$\tilde{b}(x,\Pi) = Ax + B\mu(\Pi), \quad \tilde{\xi}(x,\Pi) = \sqrt{C^2x^2 + 2CDx\mu(\Pi) + D^2[\mu^2(\Pi) + \sigma^2(\Pi)]}. \tag{3.12}$$

Thus, the state dynamic X^Π in (3.3) is given by

$$\mathrm{d}X_t^\Pi = (AX_t^\Pi + B\mu_t)\mathrm{d}t + \sqrt{(CX_t^\Pi + D\mu_t)^2 + D^2\sigma_t^2}\,\mathrm{d}W_t, \quad X_0^\Pi = x \in \mathbb{R}, \tag{3.13}$$

which implies that the state process only depends on the mean process $\{\mu_t\}_{t\geqslant 0}$ and the variance process $\{\sigma_t^2\}_{t\geqslant 0}$ of the given distributional control $\{\Pi_t\}_{t\geqslant 0}$. Let $\mathcal{B}$ be the Borel algebra on $\mathbb{R}$. A control process Π is said to be admissible, denoted by $\Pi \in \mathcal{A}(x)$, if (i) for each $t \geqslant 0$, $\Pi_t \in \mathcal{M}$ a.s.; (ii) for each $A \in \mathcal{B}$, $\{\Pi_t(A), t \geqslant 0\}$ is $\mathcal{F}_t$-progressively measurable; (iii) for each $t \geqslant 0$, $\mathbb{E}[\int_0^t (\mu_s^2 + \sigma_s^2)\mathrm{d}s] < \infty$; (iv) with $\{X_t^\Pi\}_{t\geqslant 0}$ solving (3.3), $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}[(X_T^\Pi)^2] = 0$; (v) with $\{X_t^\Pi\}_{t\geqslant 0}$ solving (3.3), $\mathbb{E}[\int_0^\infty e^{-\rho t}|\tilde{r}(X_t^\Pi, \Pi_t) + \lambda\Phi_h(\Pi_t)|\mathrm{d}t] < \infty$.

In the above, condition (iii) is to ensure that for any $\Pi \in \mathcal{A}(x)$, both the drift and volatility terms of (3.3) satisfy a global Lipschitz condition and a linear growth condition in the state variable and, hence, the SDE (3.3) admits a unique strong solution X^Π. Condition (iv) is used to ensure that dynamic programming and verification theorem are applicable, as will be evident in the sequel. Finally, the reward is finite under condition (v).

By (3.6) and (3.11), we have

$$\tilde{r}(x,\Pi) = -\frac{M}{2}x^2 - Rx\mu(\Pi) - \frac{N}{2}[\mu^2(\Pi) + \sigma^2(\Pi)] - Px - L\mu(\Pi). \tag{3.14}$$

Thus, plugging (3.12) and (3.14) back into (3.8), we can derive the HJB equation for LQ control as

$$\begin{aligned}\rho v(x) = \max_{\Pi\in\mathcal{M}^2} \Big\{ &- Rx\mu(\Pi) - \frac{N}{2}\left[\mu^2(\Pi) + \sigma^2(\Pi)\right] - L\mu(\Pi) + \lambda\Phi_h(\Pi) \\ &+ CDx\mu(\Pi)v''(x) + \frac{1}{2}D^2\left[\mu^2(\Pi) + \sigma^2(\Pi)\right]v''(x) + B\mu(\Pi)v'(x)\Big\} \\ &+ Axv'(x) - \frac{M}{2}x^2 - Px + \frac{1}{2}C^2x^2v''(x).\end{aligned} \tag{3.15}$$

To analyze and solve this equation, we need to study the maximization problem therein. Denote by $\varphi(x,\Pi)$ the term inside the max operator above. Observe that $\varphi(x,\Pi)$ depends on Π via only its mean $\mu(\Pi)$ and variance $\sigma^2(\Pi)$, except for the term $\Phi_h(\Pi)$, which motivates us to write

$$\max_{\Pi\in\mathcal{M}^2} \varphi(x,\Pi) = \max_{m\in\mathbb{R}, s>0} \max_{\Pi\in\mathcal{M}^2, \mu(\Pi)=m, \sigma^2(\Pi)=s^2} \varphi(x,\Pi). \tag{3.16}$$

The inner maximization problem is in turn equivalent to

$$\max_{\Pi\in\mathcal{M}^2} \Phi_h(\Pi) \qquad \text{subject to } \mu(\Pi) = m \text{ and } \sigma^2(\Pi) = s^2. \tag{3.17}$$

This is a *static* optimization problem, which holds the key to solve the HJB equation (3.15) and thus to our exploratory problem with Choquet regularizers. It is interesting to note that when the regularizer is the entropy, the optimal solution to the above problem is Gaussian, which is indeed the essential reason behind the Gaussian exploration derived in Wang et al. (2020a). More specifically, for LQ control any regularized payoff function depends only on the mean and variance processes of the distributional control, and the Gaussian distribution maximizes the entropy when the mean and variance are fixed. The natural question in our setting is what distribution with given mean and variance maximizes a Choque regularizer, which is exactly the problem (3.17). The next section is devoted to solving explicitly this maximization problem (3.17) of "mean–variance constrained Choquet regularizers" with a variety of specific Choquet regularizers.

4 Maximizing mean–variance constrained Choquet regularizers

4.1 General results

For given $h \in \mathcal{H}$, $m \in \mathbb{R}$ and $s > 0$, we consider the problem (3.17), which has been motivated by the exploratory control for RL as discussed in the previous section. Note that since Φ_h is location-invariant and scalable, (3.17) is equivalent to the following problem

$$s \max_{\Pi\in\mathcal{M}^2} \Phi_h(\Pi) \qquad \text{subject to } \mu(\Pi) = 0 \text{ and } \sigma^2(\Pi) = 1.$$

In what follows, h' represents the right-derivative of h, which exists on $[0, 1)$ since h is concave on $[0, 1]$.

It turns out that a general solution to (3.17) has been given by Theorem 3.1 of Liu et al. (2020).

Lemma 4.1. *If h is continuous and not constantly zero, then a maximizer Π^* to (3.17) has the following quantile function*

$$Q_{\Pi^*}(p) = m + s\frac{h'(1-p)}{||h'||_2}, \quad a.e.\ p \in (0, 1), \tag{4.1}$$

and the maximum value of (3.17) is $\Phi_h(\Pi^*) = s||h'||_2$.

In the context of RL, an interesting question arises: Given a distribution used for exploration, what is the regularizer that leads to that distribution? This is a practically important question that

can provide interpretability to some widely used samplers for exploration in practice. Theoretically, answering this question is in some sense a converse of Lemma 4.1 at least in the LQ setting.

In what follows, we denote by $\mathcal{M}^2(m, s^2)$ the set of $\Pi \in \mathcal{M}^2$ satisfying $\mu(\Pi) = m \in \mathbb{R}$ and $\sigma^2(\Pi) = s^2 > 0$. Also, recall that given a distribution Π the *location-scale family* of Π is the set of all distributions $\Pi_{a,b}$ parameterized by $a \in \mathbb{R}$ and $b > 0$ such that $\Pi_{a,b}(x) = \Pi((x-a)/b)$ for all $x \in \mathbb{R}$.

Proposition 4.2. *Let $\Pi \in \mathcal{M}^2(m, s^2)$ be given, where $m \in \mathbb{R}$ and $s > 0$. Then Π maximizes Φ_h as well as $\Phi_{\lambda h}$ for any $\lambda > 0$ over $\mathcal{M}^2(m, s^2)$ for a continuous $h \in \mathcal{H}$ specified by*

$$h'(p) = Q_\Pi(1-p) - m, \quad \text{a.e. } p \in (0,1). \tag{4.2}$$

Moreover, for any $\hat{\Pi}$ in the location-scale family of Π, $\hat{\Pi}$ also maximizes Φ_h over $\mathcal{M}^2(\mu(\hat{\Pi}), \sigma^2(\hat{\Pi}))$.

Proof. By Lemma 4.1, given a continuous $h \in \mathcal{H}$, we have

$$h'(p) = \frac{||h'||_2}{s}(Q_\Pi(1-p) - m), \quad \text{a.e. } p \in (0,1),$$

where Π maximizes Φ_h over $\mathcal{M}^2(m, s^2)$. Since $\Phi_{\lambda h}(\Pi) = \lambda \Phi_h(\Pi)$ for any $\lambda > 0$, Π that maximizes Φ_h also maximizes $\Phi_{\lambda h}$, which means that a positive constant multiplier in Φ_h does not affect problem (3.17). Hence, Π maximizes Φ_h over $\mathcal{M}^2(m, s^2)$ with $h'(p) = Q_\Pi(1-p) - m$ for $p \in (0,1)$ a.e. Moreover, if $\hat{\Pi}$ is in the location-scale family of Π, then we have $\hat{\Pi}(x) = \Pi((x-a)/b)$ for some $a \in \mathbb{R}$ and $b > 0$ for all $x \in \mathbb{R}$, which implies that

$$h'(p) = Q_\Pi(1-p) - m = (Q_{\hat{\Pi}}(1-p) - a)/b - m \quad \text{for } p \in (0,1) \text{ a.e.}$$

Since $\mu(\hat{\Pi}) = a + bm$, it follows that $\hat{\Pi}$ maximizes Φ_h over $\mathcal{M}^2(\mu(\hat{\Pi}), \sigma^2(\hat{\Pi}))$. □

A simple but important implication from Proposition 4.2 is that *every non-degenerate distribution with finite first and second moments is the optimizer of some Φ_h in (3.17) over $\mathcal{M}^2(m, s^2)$ for some $m \in \mathbb{R}$ and $s > 0$*. Therefore, any distribution used for static exploration can be interpreted by certain suitable Choquet regularizer Φ_h. Moreover, there is a common distortion function h, which is explicitly specified by Proposition 4.2, for any given location-scale family, in the sense that any distribution function Π belonging to this location-scale family maximizes Φ_h over $\mathcal{M}^2(\mu(\Pi), \sigma^2(\Pi))$. In other words, a single Φ_h can serve as the same regularizer for a whole location-scale family of distributions. We remark that optimization of a general functional I_h may also be feasible where h is not necessarily concave (see Pesenti et al. (2020) for inverse S-shaped distortion functions); however, this is not desirable for an exploration regularizer.

In the following subsections, we present specific examples applying the above general results, involving several samplers commonly used in RL for exploration, as well as measures commonly

used in finance and operations research for evaluating distribution variability.

4.2 Some common exploratory distributions

We first present some simple distributions which have been widely used for exploration in the RL literature.

Example 4.3 (Bang–bang exploration)**.** *Let* Π *be a Bernoulli distribution with* $\Pi(\{0\}) = 1-\varepsilon \in (0,1)$ *and* $\Pi(\{1\}) = \varepsilon$*. In this case, the RL agent explores only two states 0 and 1, which is called a bang–bang exploration. In particular, in the classical two-armed bandit problem, 0 is the currently more promising arm and 1 is the other arm. Proposition 4.2 gives*

$$h'(p) = \mathbb{1}_{\{p<\varepsilon\}} - \varepsilon, \quad \text{a.e. } p \in (0,1),$$

and thus $h(p) = p \wedge \varepsilon - \varepsilon p$*. The corresponding regularizer* Φ_h *is given by, using the quantile representation in Lemma 2.3,*

$$\Phi_h(\Pi) = \int_0^\varepsilon Q_\Pi(1-p)\mathrm{d}p - \varepsilon \int_0^1 Q_\Pi(1-p)\mathrm{d}p = \varepsilon(\mu_\varepsilon(\Pi) - \mu(\Pi)),$$

where $\mu_\varepsilon(\Pi)$ *is the* ε*-tail mean defined by*

$$\mu_\varepsilon(\Pi) := \frac{1}{\varepsilon}\int_0^\varepsilon Q_\Pi(1-p)\mathrm{d}p.$$

Since a constant multiplier in Φ_h *does not affect problem (3.17), a Bernoulli distribution with parameter* ε *maximizes* $\Phi_h = \mu_\varepsilon - \mu$*. Note that the tail mean corresponds to ES in risk management with an axiomatic foundation laid out in Wang and Zitikis (2021). The difference between an ES and the mean,* $\mu_\varepsilon - \mu$*, is an example of generalized deviation measures in Example 3 of Rockafellar et al. (2006), which has an axiomatic characterization similar to ES.*

Example 4.4 (ε-greedy exploration)**.** *Let* Π *be a discrete distribution satisfying* $\Pi(\{0\}) = 1-\varepsilon \in (0,1)$ *and* $\Pi(\{j\}) = \varepsilon/(2n)$ *for* $j \in \{-n,\dots,-1,1,\dots,n\}$*. In this case, the RL agent explores* $2n+1$ *states where* 0 *is the currently most "exploitative" state and* $\{-n,\dots,-1,1,\dots,n\}$ *represent the other states surrounding* 0*. From Proposition 4.2, we have*

$$h'(p) = \sum_{i=1}^{n}(n-i+1)\mathbb{1}_{\left\{\frac{(i-1)\varepsilon}{2n} \leqslant p < \frac{i\varepsilon}{2n}\right\}} - \sum_{i=n+1}^{2n}(i-n)\mathbb{1}_{\left\{\frac{(i-1)\varepsilon}{2n}+1-\varepsilon \leqslant p < \frac{i\varepsilon}{2n}+1-\varepsilon\right\}} \tag{4.3}$$

for $p \in (0,1)$ *a.e.; and thus* h *is a piece-wise linear function. An example of* h *in (4.3) is plotted in FIG. 1. Using the quantile representation in Lemma 2.3, the corresponding regularizer* Φ_h *is given*

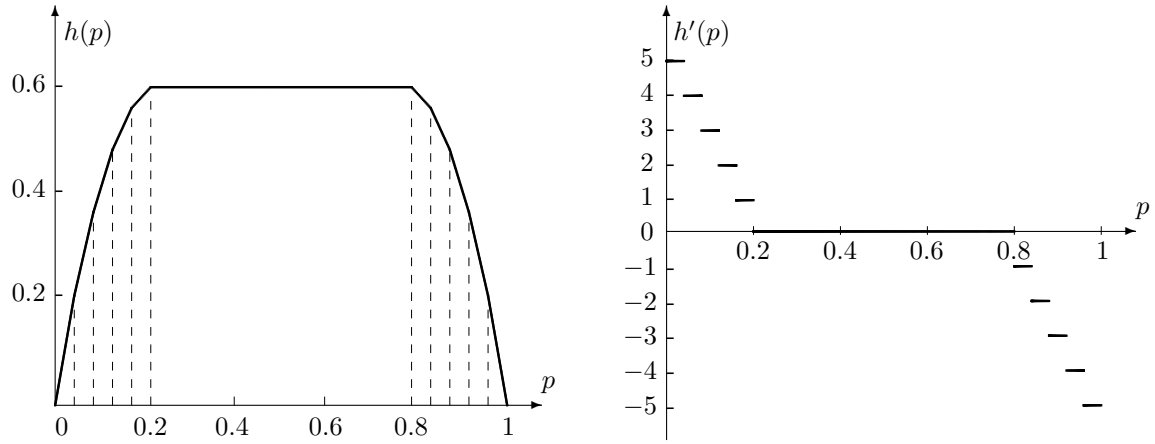

Figure 1: The plots of h (left panel) and h' (right panel) in Example 4.4 corresponding to a discrete distribution Π where $n = 5$ and $\varepsilon = 0.4$.

by

$$\Phi_h(\Pi) = \varepsilon \left(\sum_{i=1}^{n} \mu_\varepsilon^+(i, \Pi) - \sum_{i=n+1}^{2n} \mu_\varepsilon^-(i, \Pi) \right),$$

where $\mu_\varepsilon^+(i, \Pi)$ *and* $\mu_\varepsilon^-(i, \Pi)$ *are defined by*

$$\mu_\varepsilon^+(i, \Pi) := \frac{n - i + 1}{\varepsilon} \int_{\frac{(i-1)\varepsilon}{2n}}^{\frac{i\varepsilon}{2n}} Q_\Pi(1 - p)\mathrm{d}p \quad \textit{for} \quad i = 1, \ldots, n, \tag{4.4}$$

and

$$\mu_\varepsilon^-(i, \Pi) := \frac{i - n}{\varepsilon} \int_{\frac{(i-1)\varepsilon}{2n} + (1-\varepsilon)}^{\frac{i\varepsilon}{2n} + (1-\varepsilon)} Q_\Pi(1 - p)\mathrm{d}p \quad \textit{for} \quad i = n + 1, \ldots, 2n. \tag{4.5}$$

This example is related to the ε-greedy strategy in multi-armed bandit problem, where ε signifies the probability of exploring. To be specific, the ε-greedy exploration is to select the current best arm with probability $1 - \varepsilon$, and the other $2n$ arms uniformly with probability $\varepsilon/(2n)$. It is worth noting that ES is also used as a criterion in the multi-armed bandit problem with exploration; see Benac and Godin (2021) and Chang et al. (2020).

Example 4.5 (Exponential exploration)**.** *Let Π be an exponential distribution with mean 1. It follows from Proposition 4.2 that*

$$h'(p) = -\log(p) - 1, \quad \textit{a.e. } p \in (0, 1),$$

and thus $h(p) = -p\log(p)$. The corresponding Choquet regularizer Φ_h is given by

$$\Phi_h(\Pi) = -\int_0^1 Q_\Pi(1 - p)(\log(p) + 1)\mathrm{d}p =: \mathrm{CRE}(\Pi), \quad \Pi \in \mathcal{M},$$

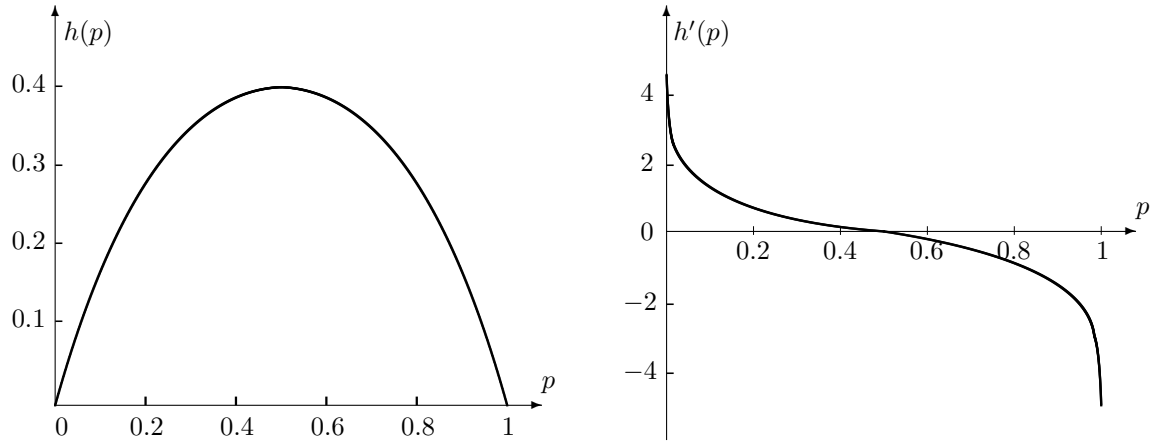

Figure 2: The plots of h (left panel) and h' (right panel) in Example 4.6 corresponding to a Gaussian distribution.

where

$$\mathrm{CRE}(\Pi) := -\int_0^\infty \Pi([x,\infty))\log(\Pi([x,\infty)))\mathrm{d}x,$$

which is called the cumulative residual entropy (CRE) and studied by Hu and Chen (2020) and Rao et al. (2004). Toomaj et al. (2017) argue that CRE can be viewed as a measure of dispersion or variability. Thus, the exponential exploration can be interpreted by the CRE regularizer.

Example 4.6 (Gaussian exploration)**.** *If Π is a Gaussian distribution, then Proposition 4.2 gives*

$$h'(p) = z(1-p), \quad \text{a.e. } p \in (0,1),$$

where z is the quantile function of a standard normal distribution.[7] *This gives $h(p) = \int_0^p z(1-s)\mathrm{d}s$, which is plotted in FIG. 2. The corresponding regularizer Φ_h is given by*

$$\Phi_h(\Pi) = \int_0^1 Q_\Pi(1-p)z(1-p)\mathrm{d}p = \int_0^1 Q_\Pi(p)z(p)\mathrm{d}p, \quad \Pi \in \mathcal{M}. \tag{4.6}$$

Thus, any Gaussian distribution maximizes the regularize Φ_h given by $\Phi_h(\Pi) = \int_0^1 Q_\Pi(p)z(p)\mathrm{d}p$. This example also indicates that there are multiple regularizers (including the above regularizer and differential entropy) that induce Gaussian exploration.

4.3 The inter-ES difference as a Choquet regularizer

We look at a regularizer based on ES. For $\Pi \in \mathcal{M}$, ES at level p is defined as

$$\mathrm{ES}_p(\Pi) := \frac{1}{1-p}\int_p^1 Q_\Pi(r)\mathrm{d}r, \quad p \in (0,1),$$

[7] In statistics, the quantile of a standard normal distribution corresponding to a test statistic is often referred to as a z-score – hence the notation z.

and the left-ES is defined as

$$\mathrm{ES}_p^-(\Pi) := \frac{1}{p}\int_0^p Q_\Pi(r)\mathrm{d}r, \quad p \in (0,1).$$

For $\alpha \in (0,1)$, let

$$h_\alpha(p) := p/(1-\alpha) \wedge 1 + (\alpha - p)/(1-\alpha) \wedge 0, \quad p \in [0,1]. \tag{4.7}$$

Define $\Phi_{h_\alpha} = \mathrm{IER}_\alpha$ by

$$\mathrm{IER}_\alpha(\Pi) := \mathrm{ES}_\alpha(\Pi) - \mathrm{ES}_{1-\alpha}^-(\Pi),$$

which is known as the inter-ES difference. Here, we assume $\alpha \in [1/2, 1)$. The inter-ES difference is a relatively new notion: it appears in Example 4 of Wang et al. (2020c) as a signed Choquet integral. In a recent work by Bellini et al. (2022), various properties are studied to underline the special role the inter-ES difference plays among other variability measures.

Proposition 4.7. *Suppose that $\alpha \in [1/2, 1)$. For $m \in \mathbb{R}$ and $s^2 > 0$, the optimization problem*

$$\max_{\Pi \in \mathcal{M}^2} \mathrm{IER}_\alpha(\Pi) \qquad \textit{subject to } \mu(\Pi) = m \textit{ and } \sigma^2(\Pi) = s^2$$

is solved by a three-point distribution Π^* with its quantile function uniquely specified as

$$Q_{\Pi^*}(p) = m + \frac{s}{\sqrt{2(1-\alpha)}}\left[\mathbb{1}_{\{p>\alpha\}} - \mathbb{1}_{\{p\leqslant 1-\alpha\}}\right], \quad \textit{a.e. } p \in (0,1). \tag{4.8}$$

Proof. Note that for $\Phi_h = \mathrm{IER}_\alpha$, we have

$$h'(p) = \frac{1}{1-\alpha}\mathbb{1}_{\{p<1-\alpha\}} - \frac{1}{1-\alpha}\mathbb{1}_{\{p\geqslant\alpha\}}$$

for $\alpha \in [1/2, 1)$, By (4.1), we can show that a maximizer Π^* satisfies (4.8), which is a three-point distribution. □

So the inter-ES difference regularizer encourages exploration at three points. One of them is the mean m corresponding to the best single-point exploitation without exploration, while the other two spots are symmetric to m capturing the exploration part.

Remark 4.8. *For $\alpha \in [1/2, 1)$, if we take the function $h_\alpha(p) = \mathbb{1}_{[1-\alpha,\alpha]}(p)$, $p \in [0,1]$, the inter-quantile difference $\Phi_{h_\alpha} := \mathrm{IQR}_\alpha$ is given by*

$$\mathrm{IQR}_\alpha(\Pi) := Q_\Pi^+(\alpha) - Q_\Pi(1-\alpha),$$

which is a classical measure of statistical dispersion widely used in e.g., box plots. Unlike the inter-ES difference, the distortion function h_α for IQR_α is not concave. However, the concave envelopes

of h is give by $h^*(p) = p/(1-\alpha) \wedge 1 + (\alpha - p)/(1-\alpha) \wedge 0$, $p \in [0,1]$, which is exactly (4.7). According to Theorem 1 in Pesenti et al. (2020), we have $\sup_{\Pi \in \mathcal{M}^2} \mathrm{IQR}_\alpha(\Pi) = \sup_{\Pi \in \mathcal{M}^2} \mathrm{IER}_\alpha(\Pi)$ and the maximizer is obtained by Π^* which satisfies (4.8). Thus, the optimization problem is still solvable even if h is not concave.

4.4 The L^1-Wasserstein distance to Dirac measures as a Choquet regularizer

Let $W : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ be a statistical distance between two distributions, such as a Wasserstein distance. Since an exploration is essentially to move away from degenerate (Dirac) distributions, a natural way to encourage exploration is to use $W(\Pi, \delta_x)$, where δ_x is the Dirac measure at $x \in \mathbb{R}$, as a regularizer. Moreover, to remove the location dependence, we modify the regularizer to be $\min_{x \in \mathbb{R}} W(\Pi, \delta_x)$. For any statistical distance satisfying $W(\Pi, \hat{\Pi}) = 0$ if and only if $\Pi = \hat{\Pi}$, it is clear that $\min_{x \in \mathbb{R}} W(\Pi, \delta_x) = 0$ if and only if Π itself is a Dirac measure (hence deterministic).

The use of Wasserstein distance to model distributional uncertainty in other settings naturally gives rise to a regularization term, yielding a theoretical justification for its use in practice; see for example Blanchet et al. (2021), Esfahani and Kuhn (2017) and Pflug and Wozabal (2007) that formulate different models with distributional robustness based on Wasserstein distances.

We focus on the case where W is the Wasserstein L^1 distance, defined as

$$W_1(\Pi, \hat{\Pi}) := \int_0^1 |Q_\Pi(p) - Q_{\hat{\Pi}}(p)| \mathrm{d}p.$$

In this case, $W_1(\Pi, \delta_x)$ is the L^1 distance between x and $X \sim \Pi$, and it is well known via L^1 loss minimization that the minimizers of $\min_{x \in \mathbb{R}} W_1(\Pi, \delta_x)$ are the medians of Π (unique if Q_Π is continuous):

$$\operatorname*{arg\,min}_{x \in \mathbb{R}} W_1(\Pi, \delta_x) = [Q_\Pi(1/2), Q_\Pi^+(1/2)].$$

Moreover, for a median of Π, $x^* \in [Q_\Pi(1/2), Q_\Pi^+(1/2)]$, we have that $W_1(\Pi, \delta_{x^*})$ is the mean-median deviation; namely

$$\begin{aligned}
\min_{x \in \mathbb{R}} W_1(\Pi, \delta_x) &= W_1(\Pi, \delta_{x^*}) \\
&= \int_0^{1/2} (x^* - Q_\Pi(p)) \mathrm{d}p + \int_{1/2}^1 (Q_\Pi(p) - x^*) \mathrm{d}p \\
&= \int_{1/2}^1 Q_\Pi(p) \mathrm{d}p - \int_0^{1/2} Q_\Pi(p) \mathrm{d}p.
\end{aligned}$$

This in turn shows that $\arg\min_{x \in \mathbb{R}} W_1(\Pi, \delta_x)$ belongs to the class of Choquet regularizers.

Proposition 4.9. *For $m \in \mathbb{R}$ and $s^2 > 0$, the optimization problem*

$$\max_{\Pi \in \mathcal{M}^2} \min_{x \in \mathbb{R}} W_1(\Pi, \delta_x) \qquad \textit{subject to } \mu(\Pi) = m \textit{ and } \sigma^2(\Pi) = s^2,$$

is solved by a unique Π^* *with the quantile function specified as*

$$Q_{\Pi^*}(p) = m + s\mathbb{1}_{\{p>1/2\}} - s\mathbb{1}_{\{p\leqslant 1/2\}}, \quad a.e.\ p \in (0,1). \tag{4.9}$$

Proof. Applying Lemma 2.3 to get $\min_{x\in\mathbb{R}} W_1(\Pi, \delta_x) = \Phi_h(\Pi)$ with $h'(p) = 1$ for $p < 1/2$ and $h'(p) = -1$ for $p \geqslant 1/2$. Using (4.1) in Lemma 4.1 yields (4.9), which implies a symmetric two-point distribution. □

As $\Phi_h(\Pi) = \min_{x\in\mathbb{R}} W_1(\Pi, \delta_x)$ induces a symmetric exploration around the mean, we call it a symmetric Wasserstein regularizer with $h(p) = p\mathbb{1}_{\{p<1/2\}} + (1-p)\mathbb{1}_{\{p\geqslant 1/2\}}$. Next, let us discuss two-point asymmetric exploration. Suppose that two directions are not symmetric, and we would like to regularize in a way to encourage more exploration in a certain direction. Take a constant $\alpha \in (0,1)$, and choose W as an asymmetric Wasserstein distance

$$W_1^\alpha(\Pi, \hat{\Pi}) = \int_0^1 \big(\alpha(Q_\Pi(p) - Q_{\hat{\Pi}}(p))_+ + (1-\alpha)(Q_\Pi(p) - Q_{\hat{\Pi}}(p))_-\big)\mathrm{d}p.$$

The corresponding minimizers are the α-quantiles

$$\arg\min_{x\in\mathbb{R}} W_1^\alpha(\Pi, \delta_x) = [Q_\Pi(\alpha), Q_\Pi^+(\alpha)],$$

and for $x^* \in [Q_\Pi(\alpha), Q_\Pi(\alpha)]$, we have

$$\begin{aligned}\min_{x\in\mathbb{R}} W_1^\alpha(\Pi, \delta_x) &= W_1^\alpha(\Pi, \delta_{x^*}) \\ &= \int_0^\alpha (1-\alpha)(x^* - Q_\Pi(p))\mathrm{d}p + \int_\alpha^1 \alpha(Q_\Pi(p) - x^*)\mathrm{d}p \\ &= \alpha\int_\alpha^1 Q_\Pi(p)\mathrm{d}p - (1-\alpha)\int_0^\alpha Q_\Pi(p)\mathrm{d}p.\end{aligned}$$

We call $\Phi_h(\Pi) = \min_{x\in\mathbb{R}} W_1^\alpha(\Pi, \delta_x)$ an asymmetric Wasserstein regularizer with $h(p) = \alpha p\mathbb{1}_{\{p<1-\alpha\}} + (1-\alpha)(1-p)\mathbb{1}_{\{p\geqslant 1-\alpha\}}$.

Proposition 4.10. *For* $m \in \mathbb{R}$ *and* $s^2 > 0$*, the optimization problem*

$$\max_{\Pi\in\mathcal{M}^2} \min_{x\in\mathbb{R}} W_1^\alpha(\Pi, \delta_x) \qquad \textit{subject to } \mu(\Pi) = m \textit{ and } \sigma^2(\Pi) = s^2$$

has a unique maximizer Π^* *with the quantile function uniquely specified as*

$$Q_{\Pi^*}(p) = m + s\left(\frac{\alpha}{1-\alpha}\right)^{1/2}\mathbb{1}_{\{p>\alpha\}} - s\left(\frac{1-\alpha}{\alpha}\right)^{1/2}\mathbb{1}_{\{p\leqslant\alpha\}}, \quad a.e.\ p \in (0,1). \tag{4.10}$$

Proof. For $\Phi_h(\Pi) = \min_{x\in\mathbb{R}} W_1^\alpha(\Pi, \delta_x)$, we have

$$h'(p) = \alpha \quad \text{for} \quad p < 1-\alpha \quad \text{and} \quad h'(p) = -1+\alpha \quad \text{for} \quad p \geqslant 1-\alpha.$$

Using (4.1), the optimization problem has a solution Π^* satisfying (4.10), which is an asymmetric two-point distribution. □

To recap, the Wasserstein L^1 regularization encourages possibly asymmetric (with respect to the mean) two-point exploration, which is an instance of the bang-bang exploration in Example 4.3.

4.5 The Gini mean difference or maxiance as a Choquet regularizer

By letting $h(p) = p - p^2$, $p \in [0,1]$, we consider the regularizer $\Phi_\sigma := \Phi_h$ given by

$$\Phi_\sigma(\Pi) = \int_{\mathbb{R}} \Big(\Pi([x,\infty)) - \Pi^2([x,\infty))\Big)\mathrm{d}x.$$

There are two ways to represent $\Phi_\sigma(\Pi)$ in terms of two iid copies X_1 and X_2 from the distribution Π. First, Φ_σ can be rewritten as

$$\Phi_\sigma(\Pi) = \frac{1}{2}\mathbb{E}[|X_1 - X_2|],$$

which is the *Gini mean difference* (e.g., Furman et al. (2017); sometimes without the factor 1/2). Alternatively, Φ_σ can be represnted as

$$\Phi_\sigma(\Pi) = \mathbb{E}[\max\{X_1, X_2\}] - \mu(\Pi),$$

which is called the *maxiance* by Eeckhoudt and Laeven (2022). The two representations are identical as seen from the following equality

$$\begin{aligned}
\mathbb{E}[\max\{X_1, X_2\}] - \mu(\Pi) &= \mathbb{E}\left[\max\{X_1, X_2\} - \frac{1}{2}(X_1 + X_2)\right] \\
&= \mathbb{E}\left[\max\{X_1, X_2\} - \frac{1}{2}\left(\max\{X_1, X_2\} + \min\{X_1, X_2\}\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\max\{X_1, X_2\} - \min\{X_1, X_2\}\right] = \frac{1}{2}\mathbb{E}[|X_1 - X_2|].
\end{aligned}$$

As argued by Eeckhoudt and Laeven (2022), the maxiance can be seen as the dual version of the variance, due to the following identities

$$\sigma^2(\Pi) = \int_{\mathbb{R}} (x - \mu(\Pi))^2 \mathrm{d}\Pi, \qquad \Phi_\sigma(\Pi) = \int_{\mathbb{R}} (x - \mu(\Pi))\mathrm{d}\Pi^2.$$

Moreover, the maxiance can be used to approximate a local index of absolute risk aversion in Yaari (1987)'s dual theory of choice under risk, which is similar to the role of variance in the classic

expected utility theory.

We now show that the maxiance regularizer Φ_σ leads to a uniform distribution for exploration.

Proposition 4.11. *For $m \in \mathbb{R}$ and $s^2 > 0$, the optimization problem*

$$\max_{\Pi \in \mathcal{M}^2} \Phi_\sigma(\Pi) \qquad \text{subject to } \mu(\Pi) = m \text{ and } \sigma^2(\Pi) = s^2 \tag{4.11}$$

has a unique maximizer $\Pi^* = \mathrm{U}[m - \sqrt{3}s, m + \sqrt{3}s]$.

Proof. Note that for $\Phi_h = \Phi_\sigma$, we have $h'(p) = 1 - 2p$. It follows from (4.1) that a maximizer Π^* is a uniform distribution. By matching the moments in (4.11), we obtain $\Pi^* = \mathrm{U}[m - \sqrt{3}s, m + \sqrt{3}s]$. The uniqueness statement is guaranteed by e.g. Theorem 2 of Pesenti et al. (2020). □

Proposition 4.11 provides a foundation for a uniformly distributed exploration strategy on $\mathbb{R}$. Note that this is different from the result of uniform distributions maximizing entropy on a fixed, given bounded region: here in our setting the region is *not* fixed, since we allow Π to be chosen from arbitrary distributions on $\mathbb{R}$, and thus the bounded region $[m - \sqrt{3}s, m + \sqrt{3}s]$ is endogenously derived rather than exogenously given.

Remark 4.12. *The inequality*

$$\sigma(\Pi) \geqslant \sqrt{3}\Phi_\sigma(\Pi) \text{ for all } \Pi \in \mathcal{M}^2$$

is known as Glasser's inequality (Glasser (1962)). For the uniform distribution Π^* in Proposition 4.11 with $\sigma(\Pi^*) = s$, we have $\Phi_\sigma(\Pi^*) = \sqrt{3}s/3$ by Lemma 4.1. Thus, Π^* attains the sharp bound of Glasser's inequality, which holds naturally since Π^* maximizes Φ_σ for a fixed σ^2.

5 Solving the exploratory stochastic LQ control problem

We are now ready to solve the exploratory stochastic LQ control problem presented in Section 3. Let

$$W(x, \Pi) = \mathbb{E}_x \left[\int_0^\infty e^{-\rho t} \left(\tilde{r}(X_t^\Pi, \Pi_t) + \lambda \Phi_h(\Pi_t) \right) \mathrm{d}t \right], \quad x \in \mathbb{R}, \ \Pi \in \mathcal{A}(x). \tag{5.1}$$

We have the following result based on Lemma 4.1.

Proposition 5.1. *Let a continuous $h \in \mathcal{H}$ be given. For any $\Pi = \{\Pi_t\}_{t \geqslant 0} \in \mathcal{A}(x)$ with mean process $\{\mu_t\}_{t \geqslant 0}$ and variance process $\{\sigma_t^2\}_{t \geqslant 0}$, there exists $\Pi^* = \{\Pi_t^*\}_{t \geqslant 0} \in \mathcal{A}(x)$ given by*

$$Q_{\Pi_t^*}(p) = \mu_t + \sigma_t \frac{h'(1-p)}{||h'||_2}, \qquad \text{a.e. } p \in (0,1), \ \ t \geqslant 0, \tag{5.2}$$

which has the same mean and variance processes satisfying $W(x, \Pi^*) \geqslant W(x, \Pi)$.

Proof. It follows from (3.13) and (3.14) that the term $\mathbb{E}_x\left[\int_0^\infty e^{-\rho t}\tilde{r}(X_t^\Pi, \Pi_t)\mathrm{d}t\right]$ in (5.1) only depends on the mean process $\{\mu_t\}_{t\geqslant 0}$ and variance process $\{\sigma_t^2\}_{t\geqslant 0}$ of $\{\Pi_t\}_{t\geqslant 0}$. Thus, for any fixed $t \geqslant 0$, choose Π_t^* with mean μ_t and variance σ_t^2 that maximizes $\Phi_h(\Pi)$. From Lemma 4.1, it follows that Π_t^* satisfies (5.2) and the maximum value is $\Phi_h(\Pi_t) = \sigma_t||h'||_2$. Clearly, the strategy $\Pi^* = \{\Pi_t^*\}_{t\geqslant 0} \in \mathcal{A}(x)$ is the desired one. □

Proposition 5.1 indicates that the control problem (3.7) in the LQ setting is maximized within a location–scale family of distributions, which is determined only by h.

We go back to the HJB equation (3.15). It follows from (3.16)–(3.17) along with Lemma 4.1 that (3.15) is equivalent to

$$\begin{aligned}\rho v(x) = \max_{\mu\in\mathbb{R},\sigma>0}\Big[&- Rx\mu - \frac{N}{2}\left(\mu^2+\sigma^2\right) - L\mu + \lambda\sigma\left\|h'\right\|_2 + CDx\mu v''(x) \\ &+ \frac{1}{2}D^2\left(\mu^2+\sigma^2\right)v''(x) + B\mu v'(x)\Big] + Axv'(x) - \frac{M}{2}x^2 - Px + \frac{1}{2}C^2x^2v''(x).\end{aligned} \tag{5.3}$$

Applying the first-order conditions, we get the maximizers

$$\mu^*(x) = \frac{CDxv''(x) + Bv'(x) - Rx - L}{N - D^2v''(x)} \quad \text{and} \quad (\sigma^*(x))^2 = \frac{\lambda^2\left\|h'\right\|_2^2}{(N - D^2v''(x))^2}$$

of the max operator in (5.3), which in turn leads to the optimal distributional policy $\Pi^*(\cdot; x)$ prescribed by Lemma 4.1.

Bringing the above expressions of $\mu^*(x)$ and $\sigma^*(x)$ back into (5.3), we can further write the HJB equation as

$$\begin{aligned}\rho v(x) = &\frac{[CDxv''(x) + Bv'(x) - Rx - L]^2 + \lambda^2\left\|h'\right\|_2^2}{2[N - D^2v''(x)]} \\ &+ \frac{1}{2}\left[C^2v''(x) - M\right]x^2 + [Av'(x) - P]x.\end{aligned} \tag{5.4}$$

We now solve this equation explicitly. Denote

$$\Delta = [\rho - (2A + C^2)]N + 2(B + CD)R - D^2M.$$

Under the assumptions that $\rho > 2A + C^2$ and $MN > R^2$, a smooth solution to (5.4) is given by

$$v(x) = \frac{1}{2}k_2x^2 + k_1x + k_0,$$

where[8]

$$k_2 = \frac{\Delta - \sqrt{\Delta^2 - 4[(B + CD)^2 + (\rho - (2A + C^2))D^2](R^2 - MN)}}{2[(B + CD)^2 + D^2(\rho - (2A + C^2))]}, \tag{5.5}$$

[8]Values of k_2, k_1 and k_0 are obtained by solving the system of equations $\rho k_2 = \frac{(k_2(B+CD)-R)^2}{N-k_2D^2} + k_2\left(2A + C^2\right) - M$, $\rho k_1 = \frac{(k_1B-L)(k_2(B+CD)-R)}{N-k_2D^2} + k_1A - P$, and $\rho k_0 = \frac{(k_1B-L)^2+\lambda^2\|h'\|_2^2}{2(N-k_2D^2)}$.

$$k_1 = \frac{P\left(N - k_2 D^2\right) - LR}{k_2 B(B + CD) + (A - \rho)\left(N - k_2 D^2\right) - BR}, \tag{5.6}$$

and

$$k_0 = \frac{(k_1 B - L)^2 + \lambda^2 \|h'\|_2^2}{2\rho(N - D^2 k_2)}. \tag{5.7}$$

We can verify easily that $k_2 < 0$. Hence, v is concave, a property that is essential for v to be actually the value function. Next, we state the main result of this section, whose proof follows essentially the same lines of that of Theorem 4 in Wang et al. (2020a), thanks to the analysis above and the results obtained. We omit the details here.

Theorem 5.2. *Consider the LQ control specified by (3.10)–(3.11), where we assume $M \geqslant 0$, $N > 0$, $MN > R^2$ and* [9]

$$\rho > 2A + C^2 + \max\left(\frac{D^2 R^2 - 2NR(B + CD)}{N}, 0\right).$$

Then the value function in (3.7) is given by

$$V(x) = \frac{1}{2} k_2 x^2 + k_1 x + k_0, \quad x \in \mathbb{R},$$

where k_2, k_1 and k_0 are as in (5.5)-(5.7), respectively. The optimal feedback policy has the distribution function $\Pi^*(\cdot; x)$ whose quantile function is

$$Q_{\Pi^*(\cdot;x)}(p) = \frac{(k_2(B + CD) - R)x + k_1 B - L}{N - k_2 D^2} + \frac{\lambda h'(1 - p)}{N - k_2 D^2}, \quad a.e.\ p \in (0, 1), \quad x \in \mathbb{R}, \tag{5.8}$$

with the mean and variance given by

$$\mu^*(x) = \frac{(k_2(B + CD) - R)x + k_1 B - L}{N - k_2 D^2} \quad \text{and} \quad (\sigma^*(x))^2 = \frac{\lambda^2 \|h'\|_2^2}{(N - k_2 D^2)^2}, \quad x \in \mathbb{R}. \tag{5.9}$$

Finally, the associated optimal state process $\{X_t^*\}_{t \geqslant 0}$ with $X_0^* = x$ under $\Pi^*(\cdot;\cdot)$ is the unique solution of the SDE

$$\begin{aligned} \mathrm{d}X_t^* = &\left[\left(A + \frac{B\left(k_2(B + CD) - R\right)}{N - k_2 D^2}\right) X_t^* + \frac{B\left(k_1 B - L\right)}{N - k_2 D^2}\right] \mathrm{d}t \\ &+ \sqrt{\left[\left(C + \frac{D\left(k_2(B + CD) - R\right)}{N - k_2 D^2}\right) X_t^* + \frac{D\left(k_1 B - L\right)}{N - k_2 D^2}\right]^2 + \frac{D^2 \lambda^2 \|h'\|_2^2}{(N - k_2 D^2)^2}} \mathrm{d}W_t. \end{aligned}$$

Some remarks are in order. First of all, (5.8) implies that for any Choquet regularizer, the optimal exploratory distribution in the regularized LQ problem is uniquely determined by h'. Note that $h'(x)$ is the "probability weight" put on x when calculating the (nonlinear) Choquet expectation; see e.g. Gilboa and Schmeidler (1989) and Quiggin (1982). Second, we can see from (5.9)

[9] The constraint on ρ is used not only to ensure $k_2 < 0$ but also to show $\liminf_{T \to \infty} e^{-\rho T} \mathbb{E}[(X_T^\Pi)^2] = 0$; see the proof of Theorem 4 in Wang et al. (2020a) for more details.

that the mean of the optimal distribution does not depend on the exploration represented by h and λ, and only the variance does. In particular, the mean is exactly the same as the one in Wang et al. (2020a) when the differential entropy is used as a regularizer, which is also identical to the optimal control of the classical, non-exploratory LQ problem. Third, the mean of the exploration distributions is a linear function of the state, while its variance is independent of the state.

These observations are intuitive in the context of RL. Different h's correspond to different Choquet regularizers; hence they will certainly affect the way and the level of exploration. Also, the more weight put on the level of exploration, the more spreaded out the exploration becomes around the current position. Furthermore, the second and third observations above show a perfect separation between exploitation and exploration, as the former is captured by the mean and the latter by the variance of the optimal exploration distributions. This property is also consistent with the LQ case studied in Wang et al. (2020a) and Wang and Zhou (2020) even though a different type of regularizer is applied therein.

Next, we investigate optimal exploration samplers under the LQ framework for some concrete choices of h studied in Section 4. For convenience, we denote

$$\tilde{\sigma}^*(x) := \frac{\sigma^*(x)}{\|h'\|_2} \equiv \frac{\lambda}{N - k_2 D^2}.$$

Theorem 5.2 yields that the mean of the optimal distribution is independent of h; so we will specify only its quantile function and variance for each h discussed below. Recall that the expressions of $\mu^*(x)$ and $(\sigma^*(x))^2$ for a general h are given by (5.9).

(i) Let $h(p) = (p \wedge \varepsilon - \varepsilon p)$, leading to $\Phi_h(\Pi) = \varepsilon(\mu_\varepsilon(\Pi) - \mu(\Pi))$; see Example 4.3. The optimal policy is ε-greedy, given as

$$\Pi^*\left(\{\mu^*(x) + (1-\varepsilon)\tilde{\sigma}^*(x)\}\right) \equiv \Pi^*\left(\left\{\frac{(k_2(B+CD) - R)x + k_1 B - L + (1-\varepsilon)\lambda}{N - k_2 D^2}\right\}\right) = \varepsilon,$$

and

$$\Pi^*\left(\{\mu^*(x) - \varepsilon\tilde{\sigma}^*(x)\}\right) \equiv \Pi^*\left(\left\{\frac{(k_2(B+CD) - R)x + k_1 B - L - \varepsilon\lambda}{N - k_2 D^2}\right\}\right) = 1 - \varepsilon.$$

At each state x, the control policy takes a more "promising" action at $\mu^*(x) - \varepsilon\tilde{\sigma}^*(x)$ with a large probability $1-\varepsilon$, and tries an alternative action $\mu^*(x) + (1-\varepsilon)\tilde{\sigma}^*(x)$ with probability ε.[10] Since $\|h'\|_2^2 = \varepsilon(1-\varepsilon)$, the variance of Π^* is

$$(\sigma^*(x))^2 = \frac{\varepsilon(1-\varepsilon)\lambda^2}{(N - k_2 D^2)^2}.$$

[10] Precisely speaking, the policy presented here is not exactly the ε-greedy strategy in the classical two-arm bandit problem because the two "arms" in our setting depend on the current state x and hence are dynamically changing over time. However, at any point of time one needs to explore only two action points.

(ii) Let $h(p)$ be specified by the discrete exploration in (4.3), leading to

$$\Phi_h(\Pi) = \varepsilon \left(\sum_{i=1}^{n} \mu_\varepsilon^+(i, \Pi) - \sum_{i=n+1}^{2n} \mu_\varepsilon^-(i, \Pi) \right),$$

where $\mu_\varepsilon^+(i, \Pi)$ and $\mu_\varepsilon^-(i, \Pi)$ are defined by (4.4) and (4.5); see Example 4.4. The optimal policy is a $(2n+1)$-point distribution given as

$$\Pi^* \left(\{\mu^*(x) + j\tilde{\sigma}^*(x)\} \right) \equiv \Pi^* \left(\left\{ \frac{(k_2(B+CD) - R)x + k_1 B - L + j\lambda}{N - k_2 D^2} \right\} \right) = \frac{\varepsilon}{2n},$$

for $j \in \{-n, \ldots, -1, 1, \ldots, n\}$, and

$$\Pi^* \left(\{\mu^*(x)\} \right) \equiv \Pi^* \left(\left\{ \frac{(k_2(B+CD) - R)x + k_1 B - L}{N - k_2 D^2} \right\} \right) = 1 - \varepsilon.$$

Similarly, at each state x, the control policy takes a more "exploitative" action at $\mu^*(x)$ with a large probability $1 - \varepsilon$, and tries $2n$ alternative actions $\mu^*(x) + j\tilde{\sigma}^*(x)$ for $j \in \{-n, \ldots, -1, 1, \ldots, n\}$, each with probability $\varepsilon/(2n)$. Since $\|h'\|_2^2 = \varepsilon(n+1)(2n+1)/6$, the variance of Π^* is given by

$$(\sigma^*(x))^2 = \frac{\varepsilon(n+1)(2n+1)\lambda^2}{6(N - k_2 D^2)^2}.$$

(iii) Let $h(p) = -p \log(p)$, corresponding to

$$\Phi_h(\Pi) = \int_0^\infty \Pi([x, \infty)) \log(\Pi([x, \infty))) \mathrm{d}x;$$

see Example 4.5. The optimal policy is a shifted-exponential distribution given as

$$\Pi^*(u; x) = 1 - \exp\left\{ \frac{[(k_2(B+CD) - R)x + k_1 B - L]}{\lambda} - 1 \right\} \exp\left\{ -\frac{(N - D^2 k_2)u}{\lambda} \right\}.$$

Since $\|h'\|_2^2 = 1$, the variance of Π^* is given by

$$(\sigma^*(x))^2 = \frac{\lambda^2}{(N - k_2 D^2)^2}.$$

(iv) Let $h(p) = \int_0^p z(1-s)\mathrm{d}s$ where z is the standard normal quantile function. We have $\Phi_h(\Pi) = \int_0^1 Q_\Pi(p) z(p) \mathrm{d}p$; see Example 4.6. The optimal policy is a normal distribution given by

$$\Pi^*(\cdot; x) = \mathrm{N}\left(\frac{(k_2(B+CD) - R)x + k_1 B - L}{N - k_2 D^2}, \frac{\lambda^2}{(N - k_2 D^2)^2} \right),$$

owing to the fact that $\|h'\|_2^2 = 1$. Recall that the optimal distribution is also Gaussian in Wang et al. (2020a) using the entropy regularizer. This is an example of different regularizers leading to the

same class of exploration samplers. On the other hand, examining more closely the Gaussian policy derived above and the one in Wang et al. (2020a, eq. (40)), we observe that the means of the two are identical but the variance of the former is the square of that of the latter. The reason of the discrepency in variance is because the maximized mean–variance constrained Choquet regularizer $\Phi_h(\Pi)$ is always linear in the given standard deviation σ whereas the corresponding maximized entropy regularizer $\mathrm{DE}(\Pi)$ is logorithmic in σ.

(v) Let $h(p) = p/(1-\alpha) \wedge 1 + (\alpha - p)/(1-\alpha) \wedge 0$ with $\alpha \in [1/2, 1)$. Then $\Phi_h(\Pi) = \mathrm{ES}_\alpha(\Pi) - \mathrm{ES}^-_{1-\alpha}(\Pi)$; see Section 4.3. The optimal policy is a three-point distribution given as

$$\Pi^*\left(\left\{\frac{(1-\alpha)[(k_2(B+CD)-R)x + k_1 B - L] + \lambda}{(1-\alpha)(N-k_2D^2)}\right\}\right) = 1-\alpha,$$

$$\Pi^*\left(\left\{\frac{(k_2(B+CD)-R)x + k_1 B - L}{N-k_2D^2}\right\}\right) = 2\alpha - 1,$$

and

$$\Pi^*\left(\left\{\frac{(1-\alpha)[(k_2(B+CD)-R)x + k_1 B - L] - \lambda}{(1-\alpha)(N-k_2D^2)}\right\}\right) = 1-\alpha.$$

Since $\|h'\|_2^2 = 2a/(1-\alpha)^2$, the variance of Π^* is given by

$$(\sigma^*(x))^2 = \frac{2\alpha\lambda^2}{(1-\alpha)^2(N-k_2D^2)^2}.$$

(vi) Let $h(p) = \alpha p \mathbb{1}_{\{p<1-\alpha\}} + (1-\alpha)(1-p)\mathbb{1}_{\{p\geqslant 1-\alpha\}}$ with $\alpha \in (0,1)$. Then $\Phi_h(\Pi) = \min_{x\in\mathbb{R}} W_1(\Pi, \delta_x)$; see Section 4.4. The optimal feedback policy is an asymmetric two-point distribution given as

$$\Pi^*\left(\left\{\frac{(k_2(B+CD)-R)x + k_1 B - L + \alpha\lambda}{N-k_2D^2}\right\}\right) = 1-\alpha,$$

and

$$\Pi^*\left(\left\{\frac{(k_2(B+CD)-R)x + k_1 B - L - (1-\alpha)\lambda}{N-k_2D^2}\right\}\right) = \alpha.$$

Since $\|h'\|_2^2 = \alpha(1-\alpha)$, the variance of Π^* is given by

$$(\sigma^*(x))^2 = \frac{\alpha(1-\alpha)\lambda^2}{(N-k_2D^2)^2}.$$

(vii) Let $h(p) = p - p^2$. Then $\Phi_h(\Pi) = \mathbb{E}[|X_1 - X_2|]/2$; see Section 4.5. The optimal policy $\Pi^*(\cdot; x)$ is a uniform distribution given as

$$\mathrm{U}\left[\frac{(k_2(B+CD)-R)x + k_1 B - L - \lambda}{N-k_2D^2}, \frac{(k_2(B+CD)-R)x + k_1 B - L + \lambda}{N-k_2D^2}\right].$$

Since $\|h'\|_2^2 = 1/3$, the variance of Π^* is given by

$$(\sigma^*(x))^2 = \frac{\lambda^2}{3(N - k_2 D^2)^2}.$$

Note here the uniform distribution is on a state-dependent bounded region centering around the mean $\mu^*(x)$, rather than on a pre-specified bounded region.

6 Relationship between classical and exploratory problems

In this section, similarly to the discussions in Wang et al. (2020a) and Wang and Zhou (2020), we study the relationship between the classical (unregularized and non-exploratory) and exploratory stochastic LQ problems. Since most results are parallel , we will make the exposition brief.

Recall the classical LQ problem (3.2) where the reward function is given by (3.11). The explicit forms of optimal control and value function, denoted respectively by u^* and V^{cl}, were given by Theorem 9-(b) of Wang et al. (2020a). We now provide the solvability equivalence between the problems (3.2) and (3.7).

Theorem 6.1. *The following two statements (a) and (b) are equivalent.*

(a) The function $V(x) = \frac{1}{2}\alpha_2 x^2 + \alpha_1 x + \alpha_0 + \frac{\lambda^2 \|h'\|_2^2}{2\rho(N - \alpha_2 D^2)}, x \in \mathbb{R}$, *with* $\alpha_0, \alpha_1 \in \mathbb{R}$ *and* $\alpha_2 < 0$, *is the value function of the exploratory problem (3.7) and the corresponding optimal feedback policy has the distribution function* $\Pi^*(\cdot; x)$ *whose quantile function is*

$$Q_{\Pi^*(\cdot;x)}(p) = \frac{(\alpha_2(B + CD) - R)x + \alpha_1 B - L}{N - \alpha_2 D^2} + \frac{\lambda h'(1-p)}{N - \alpha_2 D^2},$$

with the mean and variance given by

$$\mu^*(x) = \frac{(\alpha_2(B + CD) - R)x + \alpha_1 B - L}{\alpha_2 D^2} \quad \text{and} \quad (\sigma^*(x))^2 = \frac{\lambda^2 \|h'\|_2^2}{(N - \alpha_2 D^2)^2}.$$

(b) The function $w(x) = \frac{1}{2}\alpha_2 x^2 + \alpha_1 x + \alpha_0, x \in \mathbb{R}$, *with* $\alpha_0, \alpha_1 \in \mathbb{R}$ *and* $\alpha_2 < 0$, *is the value function of the classical problem (3.2) and the corresponding optimal feedback control is*

$$u^*(x) = \frac{(\alpha_2(B + CD) - R)\, x + \alpha_1 B - L}{N - \alpha_2 D^2}.$$

Proof. We rewrite the exploratory dynamics of X^* under Π^* as

$$\begin{aligned}\mathrm{d}X_t^* &= \left(AX_t^* + B\frac{(\alpha_2(B+CD)-R)\,X_t^* + \alpha_1 B - L}{N-\alpha_2 D^2}\right)\mathrm{d}t \\ &\quad + \sqrt{\left(CX_t^* + D\frac{(\alpha_2(B+CD)-R)\,X_t^* + \alpha_1 B - L}{N-\alpha_2 D^2}\right)^2 + \frac{D^2\lambda^2\left\|h'\right\|_2^2}{(N-\alpha_2 D^2)^2}}\mathrm{d}W_t \\ &\equiv (A_1 X_t^* + A_2)\,\mathrm{d}t + \sqrt{(B_1 X_t^* + B_2)^2 + C_1}\mathrm{d}W_t,\end{aligned} \tag{6.1}$$

where $A_1 := A + \frac{B(\alpha_2(B+CD)-R)}{N-\alpha_2 D^2}$, $A_2 := \frac{B(\alpha_1 B-L)}{N-\alpha_2 D^2}$, $B_1 := C + \frac{D(\alpha_2(B+CD)-R)}{N-\alpha_2 D^2}$, $B_2 := \frac{D(\alpha_1 B-L)}{N-\alpha_2 D^2}$ and $C_1 := \frac{D^2\lambda^2\|h'\|_2^2}{(N-\alpha_2 D^2)^2}$. This has exactly the same form as that appearing in the proof of Theorem 9 in Appendix C of Wang et al. (2020a), except that the values of C_1 are different.[11] Thus, the rest of the proof is the same as in Wang et al. (2020a). □

Note that, although the value function V of the exploratory problem (3.7) has been explicitly given by Theorem 5.2, the above theorem focuses on the equivalence of *solvability* of the two problems without having to know the explicit expression of the value function of either problem. Hence we use generic letters $(\alpha_0, \alpha_1, \alpha_2)$ instead of (k_0, k_1, k_2) to express the value functions.

The following result shows that the Choquet-regularized LQ problem converges to its classical counterpart if the exploration weight λ goes to zero.

Proposition 6.2. *Assume that statement (a) (or equivalently, (b)) of Theorem 6.1 holds. Then, for each $x \in \mathbb{R}$,*

$$\lim_{\lambda\to 0}\Pi^*(\cdot;x) = \delta_{u^*(x)}(\cdot) \quad \textit{weakly.}$$

Moreover, for each $x \in \mathbb{R}$, $\lim_{\lambda\to 0}\left|V(x) - V^{cl}(x)\right| = 0$.

Proof. The proof is the same as that of Theorem 11 in Wang et al. (2020a), noting that

$$\lim_{\lambda\to 0}\frac{\lambda^2\|h'\|_2^2}{2\rho(N-\alpha_2 D^2)} = 0.$$

□

Finally, we examine the "cost of exploration" – the loss in the original (i.e., non-regularized) objective due to exploration, which was originally defined and derived in Wang et al. (2020a) for problems with entropy regularization. The notion can be extended readily to the current Choquet setting, namely, it is the difference between the two optimal value functions, adjusting for the additional contribution coming from the Choquet regularizer of the optimal exploratory strategy:

$$\mathcal{C}^{u^*,\Pi^*}(x) := V^{\mathrm{cl}}(x) - \left(V(x) - \lambda\mathbb{E}_x\left[\int_0^\infty e^{-\rho t}\left(\int_U uh'(1-\Pi_t^*(u))\mathrm{d}\Pi_t^*(u)\right)\mathrm{d}t\right]\right), \tag{6.2}$$

for $x \in \mathbb{R}$.

[11] There is a typo in the title of Appendix C of Wang et al. (2020a): it should be the proof of Theorem 9, instead of Theorem 7.

Theorem 6.3. *Assume that statement (a) (or equivalently, (b)) of Theorem 6.1 holds. Then, the exploration cost for the stochastic LQ problem is*

$$\mathcal{C}^{u^*,\Pi^*}(x) = \frac{\lambda^2 \|h'\|_2^2}{2\rho(N - \alpha_2 D^2)}, \text{ for } x \in \mathbb{R}. \tag{6.3}$$

Proof. Let $\{\Pi_t^*\}_{\{t\geqslant 0\}}$ be the open-loop control generated by the feedback control $\Pi^*(\cdot; x)$ given in statement (a) with respect to the initial state x whose quantile function is

$$Q_{\Pi^*(\cdot;x)}(p) = \frac{(\alpha_2(B + CD) - R)x + \alpha_1 B - L}{N - \alpha_2 D^2} + \frac{\lambda h'(1-p)}{N - \alpha_2 D^2},$$

with the mean and variance given by

$$\mu^*(x) = \frac{(\alpha_2(B + CD) - R)x + \alpha_1 B - L}{N - \alpha_2 D^2} \quad \text{and} \quad (\sigma^*(x))^2 = \frac{\lambda^2 \|h'\|_2^2}{(N - \alpha_2 D^2)^2}.$$

By Lemma 4.1, it is straightforward to calculate

$$\int_U uh'(1 - \Pi_t^*(u))\mathrm{d}\Pi_t^*(u) = \frac{\lambda ||h'||_2^2}{N - \alpha_2 D^2}.$$

The desired result now follows immediately from the definition (6.2) and the expressions of $V(\cdot)$ in (a) and $V^{\mathrm{cl}}(\cdot)$ in (b). □

The costs of exploration derived in Wang et al. (2020a) and Wang and Zhou (2020) for the entropy setting depend on only the temperature parameter and the discounting rate or time horizon which are chosen by the agents, but not on the state dynamics or the reward coefficients which the agents generally do not know about. In contrast, the derived exploration cost (6.3) for the Choquet setting does depend on the unknown model parameters in a complicated way (mainly through α_2), which seems to be a disadvantage from the learning perspective. However, a bit of reflection reveals that it is more important to know *what* impact the cost than to know the *precise* value of the cost. For example, (6.3) suggests a way to strategically select the regularizers: other things being equal, to reduce the exploration cost one should choose regularizers with smaller values of $\|h'\|_2$. Moreover, $\mathcal{C}^{u^*,\Pi^*}(x) \leqslant \frac{\lambda^2 \|h'\|_2^2}{2\rho N}$ noting $\alpha_2 < 0$; so the cost is bounded above by a constant that is inversely proportional to the unknown parameter N, the control weight in the reward function. As a result, when executing controls becomes increasingly costly, the exploration cost diminishes because the agent is less motivated to do exploration. Furthermore, $\mathcal{C}^{u^*,\Pi^*}(x) = \frac{\lambda \|h'\|_2}{2\rho}\sigma^*(x)$, meaning that the cost is proportional to the standardized deviation of the exploratory control, a feature that is not presented in the entropy setting Wang et al. (2020a). Finally, the exploration cost (6.3) depends on λ and ρ in a rather intuitive way: it increases as λ increases, due to more emphasis placed on exploration, or as ρ decreases, indicating an effectively longer horizon for exploration.

7 Conclusion

This paper develops a framework for continuous-time RL that can generate or indeed interpret/explain many broadly practiced distributions for exploration. The main contributions are conceptual/theoretical rather than algorithmic: Theorem 5.2 does not lead directly to an algorithm to compute optimal policies, because the expression (5.8) involves the model parameters which are unknown in the RL context. That said, our results do provide important guidance for devising RL algorithms. First, Theorem 5.2 may imply a provable policy improvement theorem and hence result in a q-learning theory analogous to that in the entropy-regularized setting recently established by Jia and Zhou (2022c). Second, the explicit form (5.8) can suggest special structure of function approximators for learning optimal distributions, thereby greatly reduce the number of parameters needed for function approximation and improve the efficiency of the resulting learning algorithms. Finally, the availability of a large class of Choquet regularizers makes it possible to compare and choose specific regularizers to achieve certain objectives specific to each learning problem.

Another conceptual contribution of the paper is that it establishes a link between risk metrics and RL. This paper is the first to do so, and the attempt is by no means comprehensive. The rich literature on decision theory and risk metrics is expected to further bring in new insights and directions into the RL study, not only related to regularization, but also in terms of motivating new objective functions and axiomatic approaches for learning.

The theory developed in this paper opens up several research directions. Here we comment on some. One is to develop the corresponding q-learning theory mentioned earlier. Another is to find the "best Choquet regularizer" in terms of efficiency of the resulting RL algorithms. Yet another problem is in financial application: to formulate a continuous-time mean–variance portfolio selection problem with a Choquet regularizer and compare the performance with its entropy counterpart solved in Wang and Zhou (2020).

Last but not least, the Choquet regularizers proposed in this paper are defined for distributions on $\mathbb{R}$, while many RL applications involve multi-dimensional action spaces. Because Choquet regularizers are characterized by quantile additivity as in Theorem 2.4 while quantile functions are not well defined for distributions on $\mathbb{R}^d$ with $d > 1$, it is very challenging to study Choquet regularizers in high dimensions. To overcome the difficulty, the first possible attempt is to minic (2.2) by defining, for distributions Π on $\mathbb{R}^d$, the functional $\Phi_h^{\text{joint}}(\Pi) = \int_{\mathbb{R}^d} h \circ \Pi([\mathbf{x}, \infty))\mathrm{d}\mathbf{x}$. This formulation requires some further conditions on $h \in \mathcal{H}$ to guarantee desirable properties, and it is unclear whether we can derive the corresponding optimizers in a form similar to Proposition 4.2. Another possible idea is to use

$$\Phi_h^{\text{sum}}(\Pi) = \sum_{i=1}^{d} \int_{\mathbb{R}} h \circ \Pi_i([x, \infty))\mathrm{d}x \quad \text{or} \quad \Phi_h^{\text{prod}}(\Pi) = \prod_{i=1}^{d} \int_{\mathbb{R}} h \circ \Pi_i([x, \infty))\mathrm{d}x,$$

where Π_i is the i-th marginal distribution of Π. This formulation relies only on the marginal distributions of Π, allowing us to utilize the existing results for Choquet regularizers on $\mathbb{R}$. Either formulation mentioned above requires a thorough analysis in a future study.

Acknowledgements. Han is supported by the Fundamental Research Funds for the Central Universities, Nankai University (Grant No. 63231138). Wang is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03823, RGPAS-2018-522590). Zhou is supported by a start-up grant and the Nie Center for Intelligent Asset Management at Columbia University. His work is also part of a Columbia-CityU/HK collaborative project that is supported by the InnotHK Initiative, The Government of the HKSAR, and the AIFT Lab. The authors are grateful to the two anonymous referees for their constructive comments that have led to an improved version of the paper.

References

Acciaio, B. and Svindland, G. (2013). Are law-invariant risk functions concave on distributions? *Dependence Modeling*, **1**, 54–64.

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance*, **26**(7), 1505–1518.

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

Bellini, F., Fadina, T., Wang, R. and Wei, Y. (2022). Parametric measures of variability induced by risk measures. *Insurance: Mathematics and Economics*, **106**, 270–284.

Benac, L. and Godin, F. (2021). Risk averse non-stationary multi-armed bandits. *arXiv:* 2109.13977.

Benigno, P. and Woodford, M. (2003). Optimal monetary and fiscal policy: A linear-quadratic approach. *NBER macroeconomics annual*, **18**, 271–333.

Benigno, P. and Woodford, M. (2012). Linear-quadratic approximation of optimal policy problems. *Journal of Economic Theory*, **147**, 1–42.

Blanchet, J., Chen, L. and Zhou, X. (2022). Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Science*, forthcoming.

Chang, J. Q., Zhu, Q. and Tan, V. Y. (2020). Risk-constrained thompson sampling for CVaR bandits. *arXiv:* 2011.08046.

De Waegenaere, A. and Wakker P. P. (2001). Nonmonotonic Choquet integrals. *Journal of Mathematical Economics*, **36**(1), 45–60.

Delbaen, F. (2002). Coherent risk measures on general probability spaces. In *Advances in Finance and Stochastics* (pp. 1–37). Springer, Berlin, Heidelberg.

Denneberg, D. (1994). *Non-additive Measure and Integral*. Springer Science & Business Media.

Eeckhoudt, L. and Laeven, R. (2022). Dual moments and risk attitudes. *Operations Research*, **70**(3),

1330–1341.

Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein meric: Performance guarantees and tractable reformulations. *Mathematical Programming*, **171**(1), 115–166.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, **6**(4), 429–447.

Föllmer, H. and Schied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time*. Fourth Edition. Walter de Gruyter, Berlin.

Furman, E., Wang, R. and Zitikis, R. (2017). Gini-type measures of risk and variability: Gini shortfall, capital allocation and heavy-tailed risks. *Journal of Banking and Finance*, **83**, 70–84.

Gao, X., Xu, Z. and Zhou, X (2022). State-dependent temperature control for Langevin diffusions. *SIAM Journal on Control and Optimization*, **60**, 1250–1268.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, **18**(2), 141–153.

Glasser, G. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, **57**(299), 648–654.

Grechuk, B., Molyboha, A. and Zabarankin, M. (2009). Maximum entropy principle with general deviation measures. *Mathematics of Operations Research*, **34**(2), 445–467.

Guo, X., Xu, R. and Zariphopoulou, T. (2020). Entropy regularization for mean field games with learning. *arXiv*: 2010.00145.

Haarnoja, T., Zhou, T., Abbeel, P. and Levine, S. (2018). Soft actor-critic: Offpolicy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning* (pp. 1856—1865).

Hu, T. and Chen, O. (2020). On a family of coherent measures of variability. *Insurance: Mathematics and Economics*, **95**, 173–182.

Jia, J. and Zhou, X. (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, **23**(154), 1–55.

Jia, J. and Zhou, X. (2022b). Policy gradient and actor–critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, **23**(275),1–50.

Jia, J. and Zhou, X. (2022c). q-Learning in continuous time. *arXiv*: 2207.00713.

Judd, K. (1998). Numerical Methods in Economomics. MIT Press, Cambridge, MA.

Kusuoka, S. (2001). On law invariant coherent risk measures. *Advances in Mathematical Economics*, **3**, 83–95.

Li, W. and Todorov, E. (2007). Iterative linearization methods for approximately optimal control and estimation of non-linear stochastic system. *International Journal of Control*, **80**, 1439–1453.

Liu, F., Cai, J., Lemieux, C. and Wang, R. (2020). Convex risk functionals: Representation and applications. *Insurance: Mathematics and Economics*, **90**, 66–79.

Mou, C., Zhang, W. and Zhou, C. (2021). Robust exploratory mean-variance problem with drift uncertainty. *arXiv*: 2108.04100.

Nachum, O., Norouzi, M., Xu, K. and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems* (pp. 2775—2785).

Pesenti, S., Wang, Q. and Wang R. (2020). Optimizing distortion risk metrics with distributional uncertainty. *arXiv*: 2011.04889.

Pflug, G. and Wozabal, D. (2007). Ambiguity in portfolio selection. *Quantitative Finance*, **7**, 435–442.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, **3**(4), 323–343.

Rao, M., Chen, Y., Vemuri, B. C. and Wang, F. (2004). Cumulative residual entropy: A new measure of information. *IEEE Transactions on Information Theory*, **50**, 1220–1228.

Rockafellar, R. T., Uryasev, S. and Zabarankin, M. (2006). Generalized deviation in risk analysis. *Finance and Stochastics*, **10**, 51–74.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, **2**(3), 225–243.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, **57**(3), 571–587.

Sunoj, S. M. and Sankaran, P. G. (2012). Quantile based entropy function. *Statistics and Probability Letters*, **82**(6), 1049–1053.

Tang, W., Zhang, P. and Zhou, X. (2022). Exploratory HJB equations and their convergence. *SIAM Journal on Control and Optimization*, **60**(6), 3191–3216.

Todorov, E. and Li, W. (2005). A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In: *Proceedings of the 2005, American Control Conference, 2005*. IEEE, 300–306.

Toomaj, A., Sunoj, S. M. and Navarro, J. (2017). Some properties of the cumulative residual entropy of coherent and mixed systems. *Journal of Applied Probability*, **54**(2), 379–393.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of Uncertainty. *Journal of Risk and Uncertainty*, **5**(4), 297–323.

Wang, H., Zariphopoulou, T. and Zhou, X. (2020a). Exploration versus exploitation in reinforcement learning: A stochastic control approach. *Journal of Machine Learning Research*, **21**(198), 1–34.

Wang, H. and Zhou, X. (2020). Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, **30**(4), 1273–1308.

Wang, Q., Wang, R. and Wei, Y. (2020b). Distortion risk metrics on general spaces. *ASTIN Bulletin*, **50**(4), 827–851.

Wang, R., Wei, Y. and Willmot, G. E. (2020c). Characterization, robustness and aggregation of

signed Choquet integrals. *Mathematics of Operations Research*, **45**(3), 993–1015.

Wang, R. and Zitikis, R. (2021). An axiomatic foundation for the Expected Shortfall. *Management Science*, **67**(3), 1413–1429.

Wang, S., Young, V. R. and Panjer, H. H. (1997). Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics*, **21**(2), 173–183.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica*, **55**(1), 95–115.

Ziebart, B. D., Maas, A. L., Bagnell, J. A. and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence* (volume 8, pp. 1433–1438). Chicago, IL, USA.