Exploration versus exploitation in reinforcement learning: a stochastic control approach*

Haoran Wang† Thaleia Zariphopoulou‡ Xun Yu Zhou§

First draft: March 2018
This draft: February 2019

Abstract

We consider reinforcement learning (RL) in continuous time and study the problem of achieving the best trade-off between exploration and exploitation. We propose an entropy-regularized reward function involving the differential entropy of the distributions of actions, and motivate and devise an exploratory formulation for the feature dynamics that captures learning under exploration. The resulting optimization problem is a revitalization of the classical relaxed stochastic control. We carry out a complete analysis of the problem in the linear–quadratic (LQ) setting and deduce that the optimal feedback control distribution for balancing exploitation and exploration is Gaussian. This in turn interprets the widely adopted Gaussian exploration in RL, beyond its simplicity for sampling. Moreover, the exploitation and exploration are captured respectively by the mean and variance of the Gaussian distribution. We also find that a more random environment contains more learning opportunities in the sense that less exploration is needed. We characterize the cost of exploration, which, for the LQ case, is shown to be proportional to the entropy regularization weight and inversely proportional to the discount rate. Finally, as the weight of exploration decays to zero, we prove the convergence of the solution of the entropy-regularized LQ problem to the one of the classical LQ problem.

Key words. Reinforcement learning, entropy regularization, stochastic control, linear–quadratic, Gaussian distribution.

*We are grateful for comments from the seminar participants at UC Berkeley and Stanford, and from the participants at the Columbia Engineering for Humanity Research Forum "Business Analytics; Financial Services and Technologies" in New York and The Quantitative Methods in Finance 2018 Conference in Sydney. We thank Jose Blanchet, Wendell Fleming, Kay Giesecke, Xin Guo, Josef Teichmann and Renyuan Xu for helpful discussions and comments on the paper.

†Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA. Email: hw2718@columbia.edu.

‡Department of Mathematics and IROM, The University of Texas at Austin, Austin, USA and the Oxford-Man Institute, University of Oxford, Oxford, UK. Email: zariphop@math.utexas.edu.

§Department of Industrial Engineering and Operations Research, and The Data Science Institute, Columbia University, New York, NY 10027, USA. Email: xz2574@columbia.edu.

1 Introduction

Reinforcement learning (RL) is currently one of the most active and fast developing subareas in machine learning. In recent years, it has been successfully applied to solve large scale real world, complex decision making problems, including playing perfect-information board games such as Go (AlphaGo/AlphaGo Zero, Silver et al. (2016), Silver et al. (2017)), achieving human-level performance in video games (Mnih et al. (2015)), and driving autonomously (Levine et al. (2016), Mirowski et al. (2016)). An RL agent does not pre-specify a structural model or a family of models but, instead, gradually learns the best (or near-best) strategies based on trial and error, through interactions with the random (black box) environment and incorporation of the responses of these interactions, in order to improve the overall performance. This is a case of "kill two birds with one stone": *the agent's actions (controls) serve both as a means to explore (learn) and a way to exploit (optimize).*

Since exploration is inherently costly in terms of resource, time and opportunity, a natural and crucial question in RL is to address the dichotomy between exploration of uncharted territory and exploitation of existing knowledge. Such question exists in both the stateless RL settings (e.g. the multi-armed bandit problem) and the more general multi-state RL settings (e.g. Sutton and Barto (2018), Kaelbling et al. (1996)). Specifically, the agent must balance between greedily exploiting what has been learned so far to choose actions that yield near-term higher rewards, and continuously exploring the environment to acquire more information to potentially achieve long-term benefits. Extensive studies have been conducted to find strategies for the best trade-off betweeen exploitation and exploration.[1]

However, most of the contributions to balancing exploitation and exploration do not include exploration explicitly as a part of the optimization objective; the attention has mainly focused on solving the classical optimization problem maximizing the accumulated rewards, while exploration is typically treated separately as an *ad-hoc* chosen exogenous component, rather than being endogenously *derived* as a part of the solution to the over-

[1]For the multi-armed bandit problem, well known strategies include Gittins index approach (Gittins (1974)), Thompson sampling (Thompson (1933)), and upper confidence bound algorithm (Auer et al. (2002)), whereas theoretical optimality is established, for example, in Russo and Van Roy (2013, 2014). For general RL problems, various efficient exploration methods have been proposed that have been proved to induce low sample complexity, among other advantages (see, for example, Brafman and Tennenholtz (2002), Strehl and Littman (2008), Strehl et al. (2009)).

all RL problem. The recently proposed discrete time entropy-regularized (also termed as "entropy-augmented" or "softmax") RL formulation, on the other hand, explicitly incorporates exploration into the optimization objective as a regularization term, with a trade-off weight imposed on the entropy of the exploration strategy (Ziebart et al. (2008), Nachum et al. (2017), Fox et al. (2016)). An exploratory distribution with a greater entropy signifies a higher level of exploration, reflecting a bigger weight on the exploration front. On the other hand, having the minimal entropy, the extreme case of Dirac measure implies no exploration, reducing to the case of classical optimization with a complete knowledge about the underlying model. Recent works have been devoted to the designing of various algorithms to solve the entropy regularized RL problem, where numerical experiments have demonstrated remarkable robustness and multi-modal policy learning (Haarnoja et al. (2017), Haarnoja et al. (2018)).

In this paper, we study the trade-off between exploration and exploitation for RL in a continuous-time setting with both continuous control (action) and state (feature) spaces.[2] Such a continuous-time formulation is especially appealing if the agent can interact with the environment at ultra-high frequency, examples including high frequency stock trading, autonomous driving and snowboard riding. More importantly, once cast in continuous time, it is possible, thanks in no small measure to the tools of stochastic calculus and differential equations, to derive elegant and insightful results which, in turn, lead to theoretical understanding of some of the fundamental issues in RL, give guidance to algorithm design and provide *interpretability* to the underlying learning technologies.

Our first main contribution is to propose an *entropy-regularized reward function* involving the differential entropy for exploratory probability distributions over the continuous action space, and motivate and devise an "exploratory formulation" for the state dynamics that captures repetitive learning under exploration in the continuous time limit. Existing theoretical works on exploration mainly concentrate on the analysis at the algorithmic level, including proving convergence of the proposed exploration algorithms to the solutions of the classical optimization problems (see, for example, Singh et al. (2000), Jaakkola et al. (1994)). However, they rarely look into the impact of the exploration on changing significantly the underlying dynamics (e.g. the transition probabilities in the discrete time

[2]The terms "feature" and "action" are typically used in the RL literature, whose counterparts in the control literature are "state" and "control", respectively. Since this paper uses the control approach to study RL problems, we will interchangeably use these terms whenever there is no confusion.

context). Indeed, exploration not only substantially enriches the space of control strategies (from that of Dirac measures to that of all possible probability distributions) but also, as a result, enormously expands the reachable space of states. This, in turn, sets out to change both the underlying state transitions and the system dynamics.

We show that our exploratory formulation can account for the effects of learning in both the rewards received and the state transitions observed from the interactions with the environment. It, thus, unearths the important characteristics of learning at a more refined and in-depth level, beyond merely devising and analyzing learning algorithms. Intriguingly, the proposed formulation of the state dynamics coincides with that in the *relaxed control* framework in classical control theory (see, for example, Fleming and Nisio (1984); El Karoui et al. (1987); Zhou (1992); Kurtz and Stockbridge (1998, 2001)), which was motivated by entirely different reasons. Specifically, relaxed controls were introduced to mainly deal with the *theoretical* question of whether an optimal control exists. The approach essentially entails randomization to convexify the universe of control strategies. To the best of our knowledge, the present paper is the first to bring back the formulation of relaxed control, guided by a practical motivation: exploration and learning.

We then carry out a complete analysis on the continuous-time entropy-regularized RL problem, assuming that the original system dynamics is linear in both the control and the state, and that the original reward function is quadratic in the two. This type of linear–quadratic (LQ) problems has occupied the center stage for research in classical control theory for its elegant solutions and its ability to approximate more general nonlinear problems. One of the most important, conceptual contributions of this paper is to show that the optimal *feedback* control distribution for balancing exploitation and exploration is *Gaussian*. Precisely speaking, if, at any given state, the agent sets out to engage in exploration, then she needs look no further than Gaussian distributions. As is well known, a pure exploitation optimal distribution is Dirac, and a pure exploration optimal distribution is uniform. Our results reveal that Gaussian is the right choice if one seeks a balance between those two extremes. Moreover, we find that the mean of this optimal exploratory distribution is a function of the current state *independent* of the intended exploration level, whereas the variance is a linear function of the entropy regularizing weight (also called the "temperature parameter" or "exploration weight") *irrespective* of the current state. This result highlights a *separation* between exploitation and exploration: the former is reflected in the mean and the latter in the variance of the optimal Gaussian distribution.

There is yet another intriguing result. The higher impact actions have on the volatility of the underlying dynamic system, the smaller the variance of the optimal Gaussian distribution needs to be. Conceptually, this implies that a more random environment in fact contains more learning opportunities and, hence, is less costly for learning. This theoretical finding provides an interpretation of the recent RL heuristics where injecting noises leads to better effect of exploration; see, for example, Lillicrap et al. (2016); Plappert et al. (2018).

Another contribution of the paper is that we establish a direct connection between the solvability of the exploratory LQ problem and that of the classical LQ problem. We prove that as the exploration weight in the former decays to zero, the optimal Gaussian control distribution and its value function converge respectively to the optimal Dirac measure and the value function of the classical LQ problem, a desirable result for practical learning purposes.

Finally, we observe that, beyond the LQ problems and under proper conditions, the Gaussian distribution remains optimal for a much larger class of control problems, namely, problems with drift and volatility linear in control and reward functions linear or quadratic in control even if the dependence on state is nonlinear. Such a family of problems can be seen as the local-linear-quadratic approximation to more general stochastic control problems whose state dynamics are linearized in the control variables and the reward functions are locally approximated by quadratic control functions (Todorov and Li (2005), Li and Todorov (2007)). Note also that although such iterative LQ approximation generally has different parameters at different local state-action pairs, our result on the optimality of Gaussian distribution under the exploratory LQ framework still holds at any local point, and therefore justifies, from a stochastic control perspective, why Gaussian distribution is commonly used in the RL practice for exploration (see, among others, Haarnoja et al. (2017), Haarnoja et al. (2018), Nachum et al. (2018)), beyond its simplicity for sampling.

The rest of the paper is organized as follows. In section 2, we motivate and propose the relaxed stochastic control formulation involving an exploratory state dynamics and an entropy-regularized reward function for our RL problem. We then present the associated Hamilton-Jacobi-Bellman (HJB) equation and the optimal control distribution for general entropy-regularized stochastic control problems in section 3. In section 4, we study the special LQ problem in both the state-independent and state-dependent reward cases, corresponding respectively to the multi-armed bandit problem and the general RL problem in discrete time, and derive the optimality of

Gaussian exploration. We discuss the connections between the exploratory LQ problem and the classical LQ problem in section 5, establish the solvability equivalence of the two and the convergence result for vanishing exploration, and finally characterize the cost of exploration. We conclude in section 6. Some technical contents and proofs are relegated to Appendices.

2 An Entropy-Regularized Relaxed Stochastic Control Problem

We introduce an entropy-regularized relaxed stochastic control problem and provide its motivation in the context of RL.

Consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{F}_t\}_{t\geq 0})$ in which we define an $\{\mathcal{F}_t\}_{t\geq 0}$-adapted Brownian motion $W = \{W_t,\, t \geq 0\}$. An "action space" U is given, representing the constraints on an agent's decisions ("controls" or "actions"). An admissible (*open-loop*) control $u = \{u_t,\, t \geq 0\}$ is an $\{\mathcal{F}_t\}_{t\geq 0}$-adapted measurable process taking values in U.

The classical stochastic control problem is to control the state (or "feature") dynamics[3]

$$dx_t^u = b(x_t^u, u_t)dt + \sigma(x_t^u, u_t)dW_t,\ t > 0; \quad x_0^u = x \in \mathbb{R}, \tag{1}$$

where (and throughout this paper) x is a generic variable representing a current state of the system dynamics. The aim of the control is to achieve the maximum expected total discounted reward represented by the value function

$$V^{\mathrm{cl}}(x) := \sup_{u\in\mathcal{A}^{\mathrm{cl}}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} r\left(x_t^u, u_t\right) dt \middle| x_0^u = x\right], \tag{2}$$

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.

In the classical setting, where the model is fully known (namely, when the functions b, σ and r are fully specified) and dynamic programming is applicable, the optimal control can be derived and represented as a *deterministic* mapping from the current state to the action space U, $u_t^* = \boldsymbol{u}^*(x_t^*)$. The mapping $\boldsymbol{u}^*$ is called an optimal *feedback* control (or "policy" or "law"); this feedback control is derived at $t = 0$ and *will* be carried out through $[0, \infty)$.[4]

[3] We assume that both the state and the control are scalar-valued, only for notational simplicity. There is no essential difficulty to carry out our analysis with these being vector-valued.

[4] In general, feedback controls are easier to implement as they respond directly to the *current* state of the controlled dynamics.

In contrast, in the RL setting, where the underlying model is not known 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 $\pi = \{\pi_t(u), t \geq 0\}$ over the control space U from which each "trial" is sampled.[5] We can therefore extend the notion of controls to distributions.[6] The agent executes a control 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. This procedure, known as *policy evaluation*, is considered as a fundamental element of most RL algorithms in practice (Sutton and Barto (2018)). Hence, for evaluating such a policy distribution in our continuous time setting, it is necessary to consider the limiting situation as $N \to \infty$.

In order to capture the essential idea for doing this, let us first examine the special case when the reward depends only on the control, namely, $r(x_t^u, u_t) = r(u_t)$. One then considers N identical independent copies of the control problem in the following way: at round i, $i = 1, 2, \ldots, N$, a control u^i is sampled under the (possibly random) control distribution π, and executed for its corresponding copy of the control problem (1)–(2). Then, at each fixed time t, it follows, from the law of large numbers (and under certain mild technical conditions), that the average reward over $[t, t + \Delta t]$, with Δt small enough, should satisfy, as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N} e^{-\rho t} r(u_t^i)\Delta t \xrightarrow{\text{a.s.}} \mathbb{E}\left[e^{-\rho t}\int_U r(u)\pi_t(u)du\Delta t\right].$$

For a general reward $r(x_t^u, u_t)$ which also depends on the state, we first need to describe how exploration might alter the state dynamics (1) by defining appropriately its "exploratory" version. For this, we look at the effect of repetitive learning under a given control distribution, say π, for N rounds. Let W_t^i, $i = 1, 2, \ldots, N$, be N independent sample paths of the Brownian motion W_t, and x_t^i, $i = 1, 2, \ldots, N$, be the copies of the state

[5] As will be evident in the sequel, rigorously speaking, $\pi_t(\cdot)$ is a probability *density* function for each $t \geq 0$. With a slight abuse of terminology, *we will not distinguish a density function from its corresponding probability distribution or probability measure and thus will use these terms interchangeably in this paper*. Such nomenclature is common in the RL literature.

[6] A classical control $u = \{u_t,\ t \geq 0\}$ can be regarded as a Dirac distribution (or "measure") $\pi = \{\pi_t(u), t \geq 0\}$ where $\pi_t(\cdot) = \delta_{u_t}(\cdot)$. In a similar fashion, a feedback policy $u_t = \boldsymbol{u}(x_t^u)$ can be embedded as a Dirac measure $\pi_t(\cdot) = \delta_{\boldsymbol{u}(x_t^u)}(\cdot)$, parameterized by the current state x_t^u.

process respectively under the controls u^i, $i = 1, 2, \ldots, N$, each sampled from π. Then, the increments of these state process copies are, for $i = 1, 2, \ldots, N$,

$$\Delta x_t^i \equiv x_{t+\Delta t}^i - x_t^i \;\approx\; b(x_t^i, u_t^i)\Delta t + \sigma(x_t^i, u_t^i)\left(W_{t+\Delta t}^i - W_t^i\right), \quad t \geq 0. \quad (3)$$

Each such process x^i, $i = 1, 2, \ldots, N$, can be viewed as an independent sample from the exploratory state dynamics X^π. The superscript π of X^π indicates that each x^i is generated according to the classical dynamics (3), with the corresponding u^i sampled independently under this policy π.

It then follows from (3) and the law of large numbers that, as $N \to \infty$,

$$\begin{aligned} & \tfrac{1}{N}\textstyle\sum_{i=1}^N \Delta x_t^i \;\approx\; \tfrac{1}{N}\sum_{i=1}^N b(x_t^i, u_t^i)\Delta t + \tfrac{1}{N}\sum_{i=1}^N \sigma(x_t^i, u_t^i)\left(W_{t+\Delta t}^i - W_t^i\right) \\ \xrightarrow{\text{a.s.}} \;& \mathbb{E}\left[\textstyle\int_U b(X_t^\pi, u)\pi_t(u)du\Delta t\right] + \mathbb{E}\left[\textstyle\int_U \sigma(X_t^\pi, u)\pi_t(u)du\right]\mathbb{E}\left[W_{t+\Delta t} - W_t\right] \\ = \;& \mathbb{E}\left[\textstyle\int_U b(X_t^\pi, u)\pi_t(u)du\Delta t\right]. \end{aligned} \quad (4)$$

In the above, we have implicitly applied the (reasonable) assumption that both π_t and X_t^π are independent of the increments of the Brownian motion sample paths, which are identically distributed over $[t, t + \Delta t]$.

Similarly, as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^N \left(\Delta x_t^i\right)^2 \;\approx\; \frac{1}{N}\sum_{i=1}^N \sigma^2(x_t^i, u_t^i)\Delta t \xrightarrow{\text{a.s.}} \mathbb{E}\left[\int_U \sigma^2(X_t^\pi, u)\pi_t(u)du\Delta t\right]. \quad (5)$$

As we see, not only Δx_t^i but also $(\Delta x_t^i)^2$ are affected by repetitive learning under the given policy π.

Finally, as the individual state x_t^i is an independent sample from X_t^π, we have that Δx_t^i and $(\Delta x_t^i)^2$, $i = 1, 2, \ldots, N$, are the independent samples from ΔX_t^π and $(\Delta X_t^\pi)^2$, respectively. As a result, the law of large numbers gives that as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^N \Delta x_t^i \xrightarrow{\text{a.s.}} \mathbb{E}\left[\Delta X_t^\pi\right] \qquad \text{and} \qquad \frac{1}{N}\sum_{i=1}^N (\Delta x_t^i)^2 \xrightarrow{\text{a.s.}} \mathbb{E}\left[(\Delta X_t^\pi)^2\right].$$

This interpretation, together with (4) and (5), *motivates* us to propose the *exploratory version* of the state dynamics, namely,

$$dX_t^\pi = \tilde{b}(X_t^\pi, \pi_t)dt + \tilde{\sigma}(X_t^\pi, \pi_t)dW_t, \; t > 0; \quad X_0^\pi = x \in \mathbb{R}, \quad (6)$$

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

$$\tilde{b}(y,\pi) := \int_U b\left(y,u\right)\pi(u)du, \quad y\in\mathbb{R},\ \pi\in\mathcal{P}\left(U\right), \tag{7}$$

and

$$\tilde{\sigma}(y,\pi) := \sqrt{\int_U \sigma^2\left(y,u\right)\pi(u)du}, \quad y\in\mathbb{R},\ \pi\in\mathcal{P}\left(U\right), \tag{8}$$

with $\mathcal{P}\left(U\right)$ being the set of density functions of probability measures on U that are absolutely continuous with respect to the Lebesgue measure.

We will call (6) the *exploratory formulation* of the controlled state dynamics, and $\tilde{b}(\cdot,\cdot)$ and $\tilde{\sigma}(\cdot,\cdot)$ in (7) and (8), respectively, the *exploratory drift* and the *exploratory volatility.*[7]

In a similar fashion, as $N\to\infty$,

$$\frac{1}{N}\sum_{i=1}^{N} e^{-\rho t} r(x_t^i,u_t^i)\Delta t \xrightarrow{\text{a.s.}} \mathbb{E}\left[e^{-\rho t}\int_U r(X_t^\pi,u)\pi_t(u)du\Delta t\right]. \tag{10}$$

Hence, the reward function r in (2) needs to be modified to the *exploratory reward*

$$\tilde{r}\left(y,\pi\right) := \int_U r\left(y,u\right)\pi(u)du, \quad y\in\mathbb{R},\ \pi\in\mathcal{P}\left(U\right). \tag{11}$$

If, on the other hand, the model is fully known, exploration would not be needed at all and the control distributions would all degenerate to the

[7]The exploratory formulation (6), inspired by repetitive learning, is consistent with the notion of relaxed control in the control literature (see, for example, Fleming and Nisio (1984); El Karoui et al. (1987); Zhou (1992); Kurtz and Stockbridge (1998, 2001)). Indeed, let $f:\mathbb{R}\mapsto\mathbb{R}$ be a bounded and twice continuously differentiable function, and consider the infinitesimal generator associated to the classical controlled process (1),

$$\mathbb{L}[f](x,u) := \frac{1}{2}\sigma^2(x,u)f''(x) + b(x,u)f'(x), \quad x\in\mathbb{R},\ u\in U.$$

In the classical relaxed control framework, the controlled dynamics is replaced by the six-tuple $(\Omega,\mathcal{F},\mathbb{F}=\{\mathcal{F}_t\}_{t\geq 0},\mathbb{P},X^\pi,\pi)$, such that $X_0^\pi = x$ and

$$f(X_t^\pi) - f(x) - \int_0^t\int_U \mathbb{L}[f](X_t^\pi,u)\pi_t(u)duds, \quad t\geq 0, \quad \text{is a } \mathbb{P}-\text{martingale.} \tag{9}$$

It is easy to verify that our proposed exploratory formulation (6) agrees with the above martingale formulation. However, even though the mathematical formulations are equivalent, the motivations of the two are entirely different. Relaxed control was introduced to mainly deal with the existence of optimal controls, whereas the exploratory formulation here is motivated by learning and exploration in RL.

Dirac measures, and we would then be in the realm of the classical stochastic control. Thus, in the RL context, we need to add a "regularization term" to account for model uncertainty and to encourage exploration. We use Shanon's *differential entropy* to measure the level of exploration:

$$\mathcal{H}(\pi) := -\int_U \pi(u)\ln\pi(u)du, \quad \pi \in \mathcal{P}(U).$$

We therefore introduce the following entropy-regularized relaxed stochastic control problem

$$V(x) := \sup_{\pi\in\mathcal{A}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\int_U r(X_t^\pi, u)\,\pi_t(u)\,du - \lambda\int_U \pi_t(u)\ln\pi_t(u)du\right)dt \middle| X_0^\pi = x\right] \tag{12}$$

where $\lambda > 0$ is an exogenous exploration weight parameter capturing the trade-off between exploitation (the original reward function) and exploration (the entropy), $\mathcal{A}(x)$ is the set of the admissible control distributions (which may in general depend on x), and V is the value function.[8]

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. Here, we first provide some of the "minimal" requirements for $\mathcal{A}(x)$. Denote by $\mathcal{B}(U)$ the Borel algebra on U. An admissible control distribution is a measure-valued (or precisely a density-function-valued) process $\pi = \{\pi_t,\, t \geq 0\}$ satisfying at least the following properties:

(i) for each $t \geq 0$, $\pi_t \in \mathcal{P}(U)$ a.s.;

(ii) for each $A \in \mathcal{B}(U)$, $\{\int_A \pi_t(u)du, t \geq 0\}$ is $\mathcal{F}_t$-progressively measurable;

(iii) the stochastic differential equation (SDE) (6) has a unique strong solution $X^\pi = \{X_t^\pi, t \geq 0\}$ if π is applied;

(iv) the expectation on the right hand side of (12) is finite.

Naturally, there could be additional requirements depending on specific problems. For the linear–quadratic control case, which will be the main focus of the paper, we define $\mathcal{A}(x)$ precisely in section 4.

Finally, analogous to the classical control formulation, $\mathcal{A}(x)$ contains *open-loop* control distributions that are measure-valued *stochastic processes*. We will also consider *feedback* control distributions. Specifically, a *deterministic* mapping $\boldsymbol{\pi}(\cdot;\cdot)$ is called a feedback control (distribution) if i) $\boldsymbol{\pi}(\cdot; x)$ is

[8] In the RL community, λ is also known as the temperature parameter, which we will be using occasionally.

a density function for each $x \in \mathbb{R}$; ii) the following SDE (which is the system dynamics after the feedback law $\boldsymbol{\pi}(\cdot;\cdot)$ is applied)

$$dX_t = \tilde{b}(X_t, \boldsymbol{\pi}(\cdot; X_t))dt + \tilde{\sigma}(X_t^{\pi},\ \boldsymbol{\pi}(\cdot; X_t))dW_t,\ t > 0; \quad X_0 = x \in \mathbb{R} \quad (13)$$

has a unique strong solution $\{X_t; t \geq 0\}$; and iii) the open-loop control $\pi = \{\pi_t,\, t \geq 0\} \in \mathcal{A}(x)$ where $\pi_t := \boldsymbol{\pi}(\cdot; X_t)$. In this case, the open-loop control π is said to be *generated* from the feedback control law $\boldsymbol{\pi}(\cdot;\cdot)$ with respect to x.

3 HJB Equation and Optimal Control Distributions

We present the general procedure for solving the optimization problem (12). The arguments are informal and a rigorous analysis will be carried out in the next section.

To this end, applying the classical Bellman's principle of optimality, we have

$$V(x) = \sup_{\pi \in \mathcal{A}(x)} \mathbb{E}\left[\int_0^s e^{-\rho t}\left(\tilde{r}\left(X_t^{\pi}, \pi_t\right) + \lambda \mathcal{H}\left(\pi_t\right)\right) dt + e^{-\rho s} V\left(X_s^{\pi}\right) \middle| X_0^{\pi} = x\right],\ s > 0.$$

Proceeding with standard arguments, we deduce that V satisfies the Hamilton-Jacobi-Bellman (HJB) equation

$$\rho v(x) = \max_{\pi \in \mathcal{P}(U)} \left(\tilde{r}(x,\pi) - \lambda \int_U \pi(u) \ln \pi(u) du + \frac{1}{2}\tilde{\sigma}^2(x,\pi) v''(x) \right.$$

$$\left. + \tilde{b}(x,\pi) v'(x) \right), \quad x \in \mathbb{R}, \quad (14)$$

or

$$\rho v(x) = \max_{\pi \in \mathcal{P}(U)} \int_U \left(r(x,u) - \lambda \ln \pi(u) + \frac{1}{2}\sigma^2(x,u) v''\left(x\right) + b\left(x,u\right) v'(x) \right) \pi(u) du, \quad (15)$$

where v denotes the generic unknown solution of the equation.

Recalling that $\pi \in \mathcal{P}\left(U\right)$ if and only if

$$\int_U \pi(u) du = 1 \quad \text{and} \quad \pi(u) \geq 0 \text{ a.e. on } U, \quad (16)$$

we can solve the (constrained) maximization problem on the right hand side of (15) to get a feedback control:

$$\boldsymbol{\pi}^*(u;x) = \frac{\exp\left(\frac{1}{\lambda}\left(r(x,u) + \frac{1}{2}\sigma^2(x,u)\,v''(x) + b(x,u)\,v'(x)\right)\right)}{\int_U \exp\left(\frac{1}{\lambda}\left(r(x,u) + \frac{1}{2}\sigma^2(x,u)\,v''(x) + b(x,u)\,v'(x)\right)\right)du}. \quad (17)$$

For each given initial state $x \in \mathbb{R}$, this feedback control in turn generates an optimal open-loop control

$$\pi_t^* := \boldsymbol{\pi}^*(u;X_t^*) = \frac{\exp\left(\frac{1}{\lambda}\left(r(X_t^*,u) + \frac{1}{2}\sigma^2(X_t^*,u)v''(X_t^*) + b(X_t^*,u)v'(X_t^*)\right)\right)}{\int_U \exp\left(\frac{1}{\lambda}\left(r(X_t^*,u) + \frac{1}{2}\sigma^2(X_t^*,u)v''(X_t^*) + b(X_t^*,u)v'(X_t^*)\right)\right)du}, \quad (18)$$

where $\{X_t^*, t \geq 0\}$ solves (6) when the feedback control law $\boldsymbol{\pi}^*(\cdot;\cdot)$ is applied and assuming that $\{\pi_t^*, t \geq 0\} \in \mathcal{A}(x)$.[9]

Formula (17) above elicits qualitative understanding about optimal explorations. We further investigate this in the next section.

4 The Linear–Quadratic Case

We now focus on the family of entropy-regularized (relaxed) stochastic control problems with linear state dynamics and quadratic rewards, in which

$$b(x,u) = Ax + Bu \quad \text{and} \quad \sigma(x,u) = Cx + Du, \;\; x,u \in \mathbb{R}, \quad (19)$$

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

$$r(x,u) = -\left(\frac{M}{2}x^2 + Rxu + \frac{N}{2}u^2 + Px + Qu\right), \;\; x,u \in \mathbb{R} \quad (20)$$

where $M \geq 0$, $N > 0$, $R, P, Q \in \mathbb{R}$.

In the classical control literature, this type of linear–quadratic (LQ) control problems is one of the most important, not only because it admits elegant and simple solutions but also because more complex, nonlinear problems can be approximated by LQ problems. As is standard with LQ control, we assume that the control set is unconstrained, namely, $U = \mathbb{R}$.

Fix an initial state $x \in \mathbb{R}$. For each open-loop control $\pi \in \mathcal{A}(x)$, denote its mean and variance processes μ_t, $\sigma_t^2, t \geq 0$, by

$$\mu_t := \int_{\mathbb{R}} u\pi_t(u)du \quad \text{and} \quad \sigma_t^2 := \int_{\mathbb{R}} u^2\pi_t(u)du - \mu_t^2\,. \quad (21)$$

[9] We stress that the procedure described in this section, while standard, is informal. A rigorous treatment requires a precise definition of $\mathcal{A}(x)$ and a verification that indeed $\{\pi_t^*, t \geq 0\} \in \mathcal{A}(x)$. This will be carried out in the study of the linear–quadratic case in the following sections.

Then, the state SDE (6) becomes

$$\begin{aligned} dX_t^\pi = &\ (AX_t^\pi + B\mu_t)\,dt + \sqrt{C^2(X_t^\pi)^2 + 2CDX_t^\pi\mu_t + D^2(\mu_t^2+\sigma_t^2)}\,dW_t \\ = &\ (AX_t^\pi + B\mu_t)\,dt + \sqrt{(CX_t^\pi + D\mu_t)^2 + D^2\sigma_t^2}\,dW_t,\ t>0; \quad X_0^\pi = x. \end{aligned} \tag{22}$$

Further, denote

$$L(X_t^\pi, \pi_t) := \int_{\mathbb{R}} r(X_t^\pi, u)\pi_t(u)du - \lambda\int_{\mathbb{R}} \pi_t(u)\ln\pi_t(u)du.$$

Next, we specify the associated set of admissible controls $\mathcal{A}(x)$: $\pi \in \mathcal{A}(x)$, if

(i) for each $t \geq 0$, $\pi_t \in \mathcal{P}(\mathbb{R})$ a.s.;

(ii) for each $A \in \mathcal{B}(\mathbb{R})$, $\{\int_A \pi_t(u)du, t \geq 0\}$ is $\mathcal{F}_t$-progressively measurable;

(iii) for each $t \geq 0$, $\mathbb{E}\left[\int_0^t \left(\mu_s^2 + \sigma_s^2\right) ds\right] < \infty$;

(iv) with $\{X_t^\pi, t \geq 0\}$ solving (22), $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\big[\,(X_T^\pi)^2\,\big] = 0$;

(v) with $\{X_t^\pi, t \geq 0\}$ solving (22), $\mathbb{E}\left[\int_0^\infty e^{-\rho t}\,|L(X_t^\pi, \pi_t)|\,dt\,\right] < \infty$.

In the above, condition (iii) is to ensure that for any $\pi \in \mathcal{A}(x)$, both the drift and volatility terms of (22) satisfy a global Lipschitz condition and a type of linear growth condition in the state variable and, hence, the SDE (22) admits a unique strong solution X^π. Condition (iv) will be used to ensure that dynamic programming and verification are applicable for this model, as will be evident in the sequel. Finally, the reward is finite under condition (v).

We are now ready to introduce the entropy-regularized relaxed stochastic LQ problem

$$V(x) = \sup_{\pi\in\mathcal{A}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\int_{\mathbb{R}} r(X_t^\pi, u)\pi_t(u)du - \lambda\int_{\mathbb{R}} \pi_t(u)\ln\pi_t(u)du\right)dt\Big|X_0^\pi = x\right] \tag{23}$$

with r as in (20) and X^π as in (22).

In the following two subsections, we derive explicit solutions for both cases of state-independent and state-dependent rewards.

4.1 The case of state-independent reward

We start with the technically less challenging case $r(x,u) = -\left(\frac{N}{2}u^2 + Qu\right)$, namely, the reward is state (feature) independent. In this case, the system

dynamics becomes irrelevant. However, the problem is still interesting in its own right as it corresponds to the state-independent RL problem, which is known as the continuous-armed bandit problem in the continuous time setting (Mandelbaum (1987); Kaspi and Mandelbaum (1998)).

Following the derivation in the previous section, the optimal feedback control in (17) reduces to

$$\begin{aligned}\boldsymbol{\pi}^*(u;x) &= \frac{\exp\left(\frac{1}{\lambda}\left(\left(-\frac{N}{2}u^2+Qu\right)+\frac{1}{2}(Cx+Du)^2v''(x)+(Ax+Bu)v'(x)\right)\right)}{\int_{\mathbb{R}}\exp\left(\frac{1}{\lambda}\left(\left(-\frac{N}{2}u^2+Qu\right)+\frac{1}{2}(Cx+Du)^2v''(x)+(Ax+Bu)v'(x)\right)\right)du} \\ &= \frac{\exp\left(-\left(u-\frac{CDxv''(x)+Bv'(x)-Q}{N-D^2v''(x)}\right)^2/\frac{2\lambda}{N-D^2v''(x)}\right)}{\int_{\mathbb{R}}\exp\left(-\left(u-\frac{CDxv''(x)+Bv'(x)-Q}{N-D^2v''(x)}\right)^2/\frac{2\lambda}{N-D^2v''(x)}\right)du}. \end{aligned} \tag{24}$$

Therefore, the optimal feedback control distribution appears to be *Gaussian*. More specifically, at any present state x, the agent should embark on exploration according to the Gaussian distribution with mean and variance given, respectively, by $\frac{CDxv''(x)+Bv'(x)-Q}{N-D^2v''(x)}$ and $\frac{\lambda}{N-D^2v''(x)}$. Note that in deriving the above, we have used that $N-D^2v''(x)>0$, $x\in\mathbb{R}$, a condition that will be justified and discussed later on.

Remark 1 *If we examine the derivation of (24) more closely, we easily see that the optimality of the Gaussian distribution still holds as long as the state dynamics is linear in control and the reward is quadratic in control, whereas the dependence of both on the state can be generally nonlinear.*

Substituting (24) back to (14), the HJB equation becomes, after straightforward calculations,

$$\begin{aligned}\rho v(x) = \ & \frac{\left(CDxv''(x)+Bv'(x)-Q\right)^2}{2(N-D^2v''(x))}+\frac{\lambda}{2}\left(\ln\left(\frac{2\pi e\lambda}{N-D^2v''(x)}\right)-1\right) \\ & +\frac{1}{2}C^2x^2v''(x)+Axv'(x).\end{aligned} \tag{25}$$

In general, this nonlinear equation has *multiple* smooth solutions, even among quadratic polynomials that satisfy $N-D^2v''(x)>0$. One such solution is a constant, given by

$$v(x)=v:=\frac{Q^2}{2\rho N}+\frac{\lambda}{2\rho}\left(\ln\frac{2\pi e\lambda}{N}-1\right), \tag{26}$$

with the corresponding optimal feedback control distribution (24) being

$$\boldsymbol{\pi}^*(u;x) = \frac{e^{-\left(u+\frac{Q}{N}\right)^2/\frac{2\lambda}{N}}}{\int_{\mathbb{R}} e^{-\left(u+\frac{Q}{N}\right)^2/\frac{2\lambda}{N}}\,du}. \tag{27}$$

It turns out that the right hand side of the above is independent of the current state x. So the optimal feedback control distribution is the same across different states. Note that the classical LQ problem with the state-independent reward function $r(x,u) = -\left(\frac{N}{2}u^2 + Qu\right)$ clearly has the optimal control $u^* = -\frac{Q}{N}$, which is also state-independent and is nothing else than the mean of the optimal Gaussian feedback control $\boldsymbol{\pi}^*$.

The following result establishes that the constant v is indeed the value function V and that the feedback control $\boldsymbol{\pi}^*$ defined by (27) is optimal. Henceforth, we denote, for notational convenience, by $\mathcal{N}(\cdot|\mu,\sigma^2)$ the density function of a Gaussian random variable with mean μ and variance σ^2.

Theorem 2 *If $r(x,u) = -\left(\frac{N}{2}u^2 + Qu\right)$, then the value function in (23) is given by*

$$V(x) = \frac{Q^2}{2\rho N} + \frac{\lambda}{2\rho}\left(\ln\frac{2\pi e\lambda}{N} - 1\right), \quad x \in \mathbb{R},$$

and the optimal feedback control distribution is Gaussian, with

$$\boldsymbol{\pi}^*(u;x) = \mathcal{N}\left(u\,\middle|\,-\frac{Q}{N}, \frac{\lambda}{N}\right).$$

Moreover, the associated optimal state process, $\{X_t^*, t \geq 0\}$, under $\boldsymbol{\pi}^*(\cdot;\cdot)$ is the unique solution of the SDE

$$dX_t^* = \left(AX_t^* - \frac{BQ}{N}\right)dt + \sqrt{\left(CX_t^* - \frac{DQ}{N}\right)^2 + \frac{\lambda D^2}{N}}\,dW_t, \; X_0^* = x. \tag{28}$$

Proof. Let $v(x) \equiv v$ be the constant solution to the HJB equation (25) defined by (26). Then, the corresponding feedback optimizer $\boldsymbol{\pi}^*(u;x) = \mathcal{N}\left(u\,\middle|\,-\frac{Q}{N}, \frac{\lambda}{N}\right)$ follows immediately from (24). Let $\pi^* = \{\pi_t^*, t \geq 0\}$ be the open-loop control generated from $\boldsymbol{\pi}^*(\cdot;\cdot)$. It is straightforward to verify that $\pi^* \in \mathcal{A}(x)$.[10]

[10]Since the state process is irrelevant in the current case, it is not necessary to verify the admissibility condition (iv).

Now, for any $\pi \in \mathcal{A}(x)$ and $T \geq 0$, it follows from the HJB equation (14) that

$$e^{-\rho T} v = v - \int_0^T e^{-\rho t} \rho v dt$$

$$\leq v + \mathbb{E}\left[\int_0^T e^{-\rho t} \left(\int_{\mathbb{R}} \left(\frac{N}{2}u^2 + Qu\right) \pi_t(u) du + \lambda \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du\right) dt\right].$$

Since $\pi \in \mathcal{A}(x)$, the dominated convergence theorem yields that, as $T \to \infty$,

$$v \geq \mathbb{E}\left[\int_0^\infty e^{-\rho t} \left(\int_{\mathbb{R}} - \left(\frac{N}{2}u^2 + Qu\right) \pi_t(u) du - \lambda \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du\right) dt\right]$$

and, thus, $v \geq V(x)$, for $\forall x \in \mathbb{R}$. On the other hand, $\boldsymbol{\pi}^*$ has been derived as the maximizer for the right hand side of (14); hence

$$\rho v = \int_{\mathbb{R}} - \left(\frac{N}{2}u^2 + Qu\right) \pi_t^*(u) du - \lambda \int_{\mathbb{R}} \pi_t^*(u) \ln \pi_t^*(u) du.$$

Replacing the inequalities by equalities in the above argument and sending T to infinity, we conclude that

$$V(x) = v = \frac{Q^2}{2\rho N} + \frac{\lambda}{2\rho}\left(\ln \frac{2\pi e \lambda}{N} - 1\right),$$

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

Finally, the exploratory dynamics equation (28) follows readily from substituting $\mu_t^* = -\frac{Q}{N}$ and $(\sigma_t^*)^2 = \frac{\lambda}{N}$, $t \geq 0$, into (22). ■

It is possible to obtain *explicit solutions* to the SDE (28) for most cases, which may be useful in designing exploration algorithms based on the theoretical results derived in this paper. We relegate this discussion about solving (28) explicitly to Appendix A.

The above solution suggests that when the reward is independent of the state, so is the optimal feedback control distribution with density $\mathcal{N}(\cdot \mid -\frac{Q}{N}, \frac{\lambda}{N})$. This is intuitive since objective (12) in this case does not explicitly distinguish between states.[11]

[11] Similar observation can be made for the (state-independent) pure entropy maximization formulation, where the goal is to solve

$$\sup_{\pi \in \mathcal{A}(x)} \mathbb{E}\left[-\int_0^\infty e^{-\rho t} \left(\int_U \pi_t(u) \ln \pi_t(u) du\right) dt \Big| X_0^\pi = x\right]. \tag{29}$$

This problem becomes relevant when $\lambda \to \infty$ in the entropy-regularized objective (23),

A remarkable feature of the derived optimal distribution $\mathcal{N}(\cdot \mid -\frac{Q}{N}, \frac{\lambda}{N})$ is that its mean coincides with the optimal control of the original, non-exploratory LQ problem, whereas the variance is determined by the temperature parameter λ. In the context of continuous-armed bandit problem, this result stipulates that the mean is concentrated on the current incumbent of the best arm and the variance is determined by the temperature parameter. The more weight put on the level of exploration, the more spread out the exploration becomes around the current best arm. This type of exploration/exploitation strategies is clearly intuitive and, in turn, gives a guidance on how to actually choose the temperature parameter in practice: it is nothing else than the variance of the exploration the agent wishes to engage in (up to a scaling factor being the quadratic coefficient of the control in the reward function).

However, we shall see in the next section that when the reward depends on the local state, the optimal feedback control distribution genuinely depends on the state.

4.2 The case of state-dependent reward

We now consider the general case with the reward depending on both the control and the state, namely,

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

We will be working with the following assumption.

Assumption 3 *The discount rate satisfies* $\rho > 2A + C^2 + \max\left(\frac{D^2R^2 - 2NR(B+CD)}{N}, 0\right)$.

This assumption requires a sufficiently large discount rate, or (implicitly) a sufficiently short planning horizon. Such an assumption is standard in infinite horizon problems with running rewards.

---

corresponding to the extreme case of pure exploration without considering exploitation (i.e., without maximizing any reward). To solve problem (29), we can pointwisely maximize its integrand, leading to the state-independent optimization problem

$$\sup_{\pi \in \mathcal{P}(U)} \left(-\int_U \pi(u) \ln \pi(u) du\right). \tag{30}$$

It is then straightforward that the optimal control distribution π^* is, for all $t \geq 0$, the uniform distribution. This is in accordance with the traditional static setting where uniform distribution achieves maximum entropy (Shannon (2001)).

Following an analogous argument as for (24), we deduce that a candidate optimal feedback control is given by

$$\boldsymbol{\pi}^*(u;x) = \mathcal{N}\left(u \,\middle|\, \frac{CDxv''(x) + Bv'(x) - Rx - Q}{N - D^2v''(x)}\,,\ \frac{\lambda}{N - D^2v''(x)}\right). \tag{31}$$

In turn, denoting by $\mu^*(x)$ and $(\sigma^*(x))^2$ the mean and variance of $\boldsymbol{\pi}^*(\cdot;x)$ given above, the HJB equation (14) becomes

$$\begin{aligned}
\rho v(x) = \ & \int_{\mathbb{R}} -\left(\tfrac{M}{2}x^2 + Rxu + \tfrac{N}{2}u^2 + Px + Qu\right)\mathcal{N}\left(u \ \middle|\mu^*(x), (\sigma^*(x))^2\right)du \\
& + \lambda \ln\left(\sqrt{2\pi e}\sigma^*(x)\right) + v'(x)\int_{\mathbb{R}}(Ax+Bu)\mathcal{N}\left(u \ \middle|\mu^*(x), (\sigma^*(x))^2\right)du \\
& + \tfrac{1}{2}v''(x)\int_{\mathbb{R}}(Cx+Du)^2\mathcal{N}\left(u \ \middle|\mu^*(x), (\sigma^*(x))^2\right)du \\
= \ & -\tfrac{M}{2}x^2 - \tfrac{N}{2}\left(\left(\tfrac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)}\right)^2 + \tfrac{\lambda}{N-D^2v''(x)}\right) \\
& -(Rx+Q)\tfrac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)} - Px + \lambda\ln\sqrt{\tfrac{2\pi e\lambda}{N-D^2v''(x)}} \\
& + Axv'(x) + Bv'(x)\tfrac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)} + \tfrac{1}{2}C^2x^2v''(x) \\
& + \tfrac{1}{2}D^2\left(\left(\tfrac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)}\right)^2 + \tfrac{\lambda}{N-D^2v''(x)}\right)v''(x) \\
& + CDxv''(x)\tfrac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)}.
\end{aligned}$$

Reorganizing, the above reduces to

$$\begin{aligned}
\rho v(x) = \ & \tfrac{(CDxv''(x)+Bv'(x)-Rx-Q)^2}{2(N-D^2v''(x))} + \tfrac{\lambda}{2}\left(\ln\left(\tfrac{2\pi e\lambda}{N-D^2v''(x)}\right) - 1\right) \\
& + \tfrac{1}{2}(C^2v''(x) - M)x^2 + (Av'(x) - P)x.
\end{aligned} \tag{32}$$

Under Assumption 3 and the additional condition $R^2 < MN$ (which holds automatically if $R = 0$, $M > 0$ and $N > 0$, a standard case in the classical LQ problems), one smooth solution to the HJB equation (32) is given by

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

where[12]

$$\begin{aligned}
k_2 := \ & \tfrac{1}{2}\tfrac{(\rho-(2A+C^2))N+2(B+CD)R-D^2M}{(B+CD)^2+(\rho-(2A+C^2))D^2} \\
& -\tfrac{1}{2}\tfrac{\sqrt{((\rho-(2A+C^2))N+2(B+CD)R-D^2M)^2-4((B+CD)^2+(\rho-(2A+C^2))D^2)(R^2-MN)}}{(B+CD)^2+(\rho-(2A+C^2))D^2},
\end{aligned} \tag{36}$$

[12]In general, there are multiple solutions to (32). Indeed, applying, for example, a generic quadratic function ansatz $v(x) = \frac{1}{2}a_2x^2 + a_1x + a_0$, $x \in \mathbb{R}$, in (32) yields the

$$k_1 := \frac{P(N - k_2 D^2) - QR}{k_2 B(B + CD) + (A - \rho)(N - k_2 D^2) - BR}, \tag{37}$$

and

$$k_0 := \frac{(k_1 B - Q)^2}{2\rho(N - k_2 D^2)} + \frac{\lambda}{2\rho}\left(\ln\left(\frac{2\pi e\lambda}{N - k_2 D^2}\right) - 1\right). \tag{38}$$

For this particular solution, given by $v(x)$ above, we can verify that $k_2 < 0$, due to Assumption 3 and $R^2 < MN$. Hence, v is concave, a property that is essential in proving that it is actually the value function.[13] On the other hand, $N - D^2 v''(x) = N - k_2 D^2 > 0$, ensuring that k_0 is well defined.

Next, we state one of the main results of this paper.

Theorem 4 *Suppose the reward function is given by*

$$r(x, u) = -\left(\frac{M}{2}x^2 + Rxu + \frac{N}{2}u^2 + Px + Qu\right),$$

with $M \geq 0$, $N > 0$, $R, Q, P \in \mathbb{R}$ and $R^2 < MN$. Furthermore, suppose that Assumption 3 holds. Then, the value function in (23) is given by

$$V(x) = \frac{1}{2}k_2 x^2 + k_1 x + k_0, \quad x \in \mathbb{R}, \tag{39}$$

where k_2, k_1 and k_0 are as in (36), (37) and (38), respectively. Moreover, the optimal feedback control is Gaussian, with its density function given by

$$\boldsymbol{\pi}^*(u; x) = \mathcal{N}\left(u \;\middle|\; \frac{(k_2(B + CD) - R)x + k_1 B - Q}{N - k_2 D^2}\,,\; \frac{\lambda}{N - k_2 D^2}\right). \tag{40}$$

---

system of algebraic equations

$$\rho a_2 = \frac{(a_2(B + CD) - R)^2}{N - a_2 D^2} + a_2(2A + C^2) - M, \tag{33}$$

$$\rho a_1 = \frac{(a_1 B - Q)(a_2(B + CD) - R)}{N - a_2 D^2} + a_1 A - P, \tag{34}$$

$$\rho a_0 = \frac{(a_1 B - Q)^2}{2(N - a_2 D^2)} + \frac{\lambda}{2}\left(\ln\left(\frac{2\pi e\lambda}{N - a_2 D^2}\right) - 1\right). \tag{35}$$

This system has two sets of solutions (as the quadratic equation (33) has, in general, two roots), leading to two quadratic solutions to the HJB equation (32). The one given through (36)–(38) is one of the two solutions.

[13] Under Assumption 3 and $R^2 < MN$, the HJB equation has an additional quadratic solution, which however is *convex*.

Finally, the associated optimal state process $\{X_t^*, t \geq 0\}$ *under* $\boldsymbol{\pi}^*(\cdot;\cdot)$ *is the unique solution of the SDE*

$$dX_t^* = \left(\left(A + \frac{B(k_2(B+CD)-R)}{N-k_2D^2}\right)X_t^* + \frac{B(k_1B-Q)}{N-k_2D^2}\right)dt$$
$$+\sqrt{\left(\left(C + \frac{D(k_2(B+CD)-R)}{N-k_2D^2}\right)X_t^* + \frac{D(k_1B-Q)}{N-k_2D^2}\right)^2 + \frac{\lambda D^2}{N-k_2D^2}}\,dW_t,\ X_0^* = x. \tag{41}$$

A proof of this theorem follows essentially the same idea as that of Theorem 2, but it is more technically involved, mainly for verifying the admissibility of the candidate optimal control. To ease the presentation, we defer it to Appendix B.

Remark 5 *As in the state-independent case (see Appendix A), the solution to the SDE (41) can be expressed through the Doss-Saussman transformation if* $D \neq 0$.

Specifically, if $C + \frac{D(k_2(B+CD)-R)}{N-k_2D^2} \neq 0$, *then*

$$X_t^* = F(W_t, Y_t), \quad t \geq 0,$$

where the function F *is given by*

$$F(z,y) = \frac{\sqrt{\tilde{D}}}{|\tilde{C}_1|}\sinh\left(|\tilde{C}_1|z + \sinh^{(-1)}\left(\frac{|\tilde{C}_1|}{\sqrt{\tilde{D}}}\left(y + \frac{\tilde{C}_2}{\tilde{C}_1}\right)\right)\right) - \frac{\tilde{C}_2}{\tilde{C}_1},$$

and the process Y_t, $t \geq 0$, *is the unique pathwise solution to the random ODE*

$$\frac{dY_t}{dt} = \frac{\tilde{A}F(W_t,Y_t) + \tilde{B} - \frac{\tilde{C}_1}{2}\left(\tilde{C}_1F(W_t,Y_t) + \tilde{C}_2\right)}{\frac{\partial}{\partial y}F(z,y)\big|_{z=W_t, y=Y_t}},\ Y_0 = x,$$

with $\tilde{A} := A + \frac{B(k_2(B+CD)-R)}{N-k_2D^2}$, $\tilde{B} := \frac{B(k_1B-Q)}{N-k_2D^2}$, $\tilde{C}_1 := C + \frac{D(k_2(B+CD)-R)}{N-k_2D^2}$,

$\tilde{C}_2 := \frac{D(k_1B-Q)}{N-k_2D^2}$ *and* $\tilde{D} := \frac{\lambda D^2}{N-k_2D^2}$.

If $C + \frac{D(k_2(B+CD)-R)}{N-k_2D^2} = 0$ *and* $\tilde{A} \neq 0$, *then it follows from direct computation that*

$$X_t^* = xe^{\tilde{A}t} - \frac{\tilde{B}}{\tilde{A}}(1-e^{\tilde{A}t}) + \sqrt{\tilde{C}_1^{\,2} + \tilde{D}}\int_0^t e^{\tilde{A}(t-s)}dW_s, \quad t \geq 0.$$

We leave the detailed derivations to the interested readers.

The above results demonstrate that, for the general state and control dependent reward case, the optimal actions over $\mathbb{R}$ also depend on the current state x, which are selected according to a state-dependent Gaussian distribution (40) with a state-independent variance $\frac{\lambda}{N-k_2D^2}$. Note that if $D \neq 0$, then $\frac{\lambda}{N-k_2D^2} < \frac{\lambda}{N}$ (since $k_2 < 0$). Therefore, the exploration variance in the general state-dependent case is *strictly smaller* than $\frac{\lambda}{N}$, the one in the state-independent case. Recall that D is the coefficient of the control in the diffusion term of the state dynamics, generally representing the level of randomness of the environment.[14] Therefore, volatility impacting actions reduce the need for exploration. Moreover, the greater D is, the smaller the exploration variance becomes, indicating that even less exploration is required. As a result, the need for exploration is further reduced if an action has a greater impact on the volatility of the system dynamics. This hints that a more volatile environment renders more learning opportunities.

On the other hand, the mean of the Gaussian distribution does not explicitly depend on λ. The implication is that the agent should concentrate on the most promising region in the action space while randomly selecting actions to interact with the unknown environment. It is intriguing that the entropy-regularized RL formulation separates the exploitation from exploration, respectively through the mean and variance of the resulting optimal Gaussian distribution.

Remark 6 *It should be noted that it is the optimal* **feedback** *control distribution, not the open-loop control generated from the feedback control, that has the Gaussian distribution. More precisely, $\boldsymbol{\pi}^*(\cdot;x)$ defined by (40) is Gaussian for each and every x, but the measure-valued process with the density function*

$$\pi_t^*(u) := \mathcal{N}\left(u \;\middle|\; \frac{(k_2(B+CD)-R)X_t^* + k_1B - Q}{N-k_2D^2}\;,\; \frac{\lambda}{N-k_2D^2}\right),\quad t \geq 0, \tag{42}$$

where $\{X_t^*, t \geq 0\}$ is the solution of the exploratory dynamics under the feedback control $\boldsymbol{\pi}^*(\cdot;\cdot)$ with any fixed initial state, say, $X_0^* = x_0$, is in general not Gaussian for any $t > 0$. The reason is that, for each $t > 0$, the right hand side of (42) is a composition of the Gaussian density function and a random variable X_t^* whose distribution is, in general, unknown. We stress that the Gaussian property of the optimal feedback control is more important and relevant in the RL context, as it stipulates that, at any given

[14] For example, in the Black–Scholes market, D is the volatility parameter of the underlying stock.

state, if one undertakes exploration then she should follow Gaussian. The open-loop control $\{\pi_t^*, t \geq 0\}$*, generated from the Gaussian feedback control, is just what the agent would end up with if she follows Gaussian exploration at every state.*

Finally, as noted earlier (see Remark 1), the optimality of the Gaussian distribution is still valid for problems with dynamics

$$dx_t = (A(x_t) + B(x_t)u_t)\ dt + (C(x_t) + D(x_t)u_t)\ dW_t,$$

and reward function in the form $r(x,u) = r_2(x)u^2 + r_1(x)u + r_0(x)$, where the functions A, B, C, D, r_2, r_1 and r_0 are possibly nonlinear (pending some additional assumptions for the verification arguments to hold).

5 The Cost and Effect of Exploration

Motivated by the necessity of exploration facing the typically unknown environment in an RL setting, we have formulated and analyzed a new class of stochastic control problems that combine entropy-regularized criteria and relaxed controls. We have also derived closed-form solutions and presented verification results for the important class of LQ problems. A natural question arises, namely, how to quantify the cost and effect of the exploration. This can be done by comparing our results to the ones for the classical stochastic LQ problems, which have neither entropy regularization nor control relaxation.

We carry out this comparison analysis next.

5.1 The classical LQ problem

We first briefly recall the classical stochastic LQ control problem in an infinite horizon with discounted reward. Let $\{W_t,\ t \geq 0\}$ be a standard Brownian motion defined on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ that satisfies the usual conditions. The controlled state process $\{x_t^u, t \geq 0\}$ solves

$$dx_t^u = (Ax_t^u + Bu_t)\ dt + (Cx_t^u + Du_t)\, dW_t\,, \quad t \geq 0,\ \ x_0^u = x, \tag{43}$$

with given constants A, B, C and D, and the process $\{u_t,\ t \geq 0\}$ being a (classical, non-relaxed) control.

The value function is defined as in (2),

$$V^{\mathrm{cl}}(x) := \sup_{u \in \mathcal{A}^{\mathrm{cl}}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} r(x_t^u, u_t)\ dt \middle|\ x_0^u = x\right], \tag{44}$$

for $x \in \mathbb{R}$, where the reward function $r(\cdot,\cdot)$ is given by (20). Here, the admissible set $\mathcal{A}^{\rm cl}(x)$ is defined as follows: $u \in \mathcal{A}^{\rm cl}(x)$ if

(i) $\{u_t, t \geq 0\}$ is $\mathcal{F}_t$-progressively measurable;
(ii) for each $t \geq 0$, $\mathbb{E}\left[\int_0^t (u_s)^2\, ds\right] < \infty$;
(iii) with $\{x_t^u, t \geq 0\}$ solving (43), $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\big[(x_T^u)^2\big] = 0$;
(iv) with $\{x_t^u, t \geq 0\}$ solving (43), $\mathbb{E}\big[\int_0^\infty e^{-\rho t}|r(x_t^u, u_t)|\, dt\big] < \infty$.

The associated HJB equation is

$$\rho w(x) = \max_{u\in\mathbb{R}} \left(r(x,u) + \frac{1}{2}(Cx+Du)^2 w''(x) + (Ax+Bu)w'(x) \right)$$

$$= \max_{u\in\mathbb{R}} \left(-\frac{1}{2}\left(N - D^2 w''(x)\right)u^2 + \left(CDxw''(x) + Bw'(x) - Rx - Q\right)u \right)$$

$$+\frac{1}{2}(C^2 w''(x) - M)x^2 + (Aw'(x) - P)x$$

$$= \frac{\left(CDxw''(x) + Bw'(x) - Rx - Q\right)^2}{2(N - D^2 w''(x))} + \frac{1}{2}(C^2 w''(x) - M)x^2 + (Aw'(x) - P)x, \tag{45}$$

with the maximizer being, provided that $N - D^2 w''(x) > 0$,

$$\boldsymbol{u}^*(x) = \frac{CDxw''(x) + Bw'(x) - Rx - Q}{N - D^2 w''(x)}, \quad x \in \mathbb{R}. \tag{46}$$

Standard verification arguments then yield that $\boldsymbol{u}$ is the optimal feedback control.

In the next section, we will establish a solvability equivalence between the entropy-regularized relaxed LQ problem and the classical one.

5.2 Solvability equivalence of classical and exploratory problems

Given a reward function $r(\cdot,\cdot)$ and a classical controlled process (1), the relaxed formulation (6) under the entropy-regularized objective is, naturally, a technically more challenging problem, compared to its classical counterpart.

In this section, we show that there is actually a solvability equivalence between the exploratory and the classical stochastic LQ problems, in the sense that the value function and optimal control of one problem lead directly to those of the other. Such equivalence enables us to readily establish the convergence result as the exploration weight λ decays to zero. Furthermore, it makes it possible to quantify the exploration cost, which we introduce in the sequel.

Theorem 7 *The following two statements (a) and (b) are equivalent.*

(a) *The function* $v(x) = \frac{1}{2}\alpha_2 x^2 + \alpha_1 x + \alpha_0 + \frac{\lambda}{2\rho}\left(\ln\left(\frac{2\pi e\lambda}{N-\alpha_2 D^2}\right) - 1\right)$, $x \in \mathbb{R}$, *with* $\alpha_0, \alpha_1 \in \mathbb{R}$ *and* $\alpha_2 < 0$, *is the value function of the exploratory problem (23) and the corresponding optimal feedback control is*

$$\boldsymbol{\pi}^*(u;x) = \mathcal{N}\left(u \;\middle|\; \frac{(\alpha_2(B+CD)-R)x + \alpha_1 B - Q}{N-\alpha_2 D^2}\,,\; \frac{\lambda}{N-\alpha_2 D^2}\right).$$

(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 (44) and the corresponding optimal feedback control is*

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

Proof. See Appendix C. ■

The above equivalence between statements (a) and (b) yields that if one problem is solvable, so is the other; and conversely, if one is not solvable, neither is the other.

5.3 Cost of exploration

We define the exploration cost for a general RL problem to be the difference between the discounted accumulated rewards following the corresponding optimal *open-loop* controls under the classical objective (2) and the exploratory objective (12), net of the value of the entropy. Note that the solvability equivalence established in the previous subsection is important for this definition, not least because the cost is well defined only if both the classical and the exploratory problems are solvable.

Specifically, let the classical maximization problem (2) with the state dynamics (1) have the value function $V^{\mathrm{cl}}(\cdot)$ and optimal strategy $\{u_t^*, t \geq 0\}$, and the corresponding exploratory problem have the value function $V(\cdot)$ and optimal control distribution $\{\pi_t^*, t \geq 0\}$. Then, we define the *exploration cost* as

$$\mathcal{C}^{u^*,\pi^*}(x) := V^{\mathrm{cl}}(x) - \left(V(x) + \lambda\mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\int_U \pi_t^*(u)\ln\pi_t^*(u)du\right)dt \,\middle|\, X_0^{\pi^*} = x\right]\right), \tag{47}$$

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

The term in the parenthesis represents the total discounted rewards incurred by π^* after taking out the contribution of the entropy term to the value function $V(\cdot)$ of the exploratory problem. The exploration cost hence measures the best outcome due to the explicit inclusion of exploratory strategies in the entropy-regularized objective, relative to the benchmark $V^{\mathrm{cl}}(\cdot)$ which is the best possible objective value should the model be *a priori* fully known.

We next compute the exploration cost for the LQ case. As we show, this cost is surprisingly simple: it depends only on two "agent-specific" parameters: the temperature parameter λ and the discounting parameter ρ.

Theorem 8 *Assume that statement (a) (or equivalently, (b)) of Theorem 7 holds. Then, the exploration cost for the stochastic LQ problem is*

$$\mathcal{C}^{u^*,\pi^*}(x) = \frac{\lambda}{2\rho}, \quad \textit{for } x \in \mathbb{R}. \tag{48}$$

Proof. Let $\{\pi_t^*, t \geq 0\}$ be the open-loop control generated by the feedback control $\boldsymbol{\pi}^*$ given in statement (a) with respect to the initial state x, namely,

$$\pi_t^*(u) = \mathcal{N}\left(u \,\middle|\, \frac{(\alpha_2(B+CD)-R)X_t^* + \alpha_1 B - Q}{N - \alpha_2 D^2}, \frac{\lambda}{N-\alpha_2 D^2} \right)$$

where $\{X_t^*, t \geq 0\}$ is the associated state process of the exploratory problem, starting from the state x, when $\boldsymbol{\pi}^*$ is applied. Then, we easily deduce that

$$\int_{\mathbb{R}} \pi_t^*(u) \ln \pi_t^*(u) du = -\frac{1}{2} \ln\left(\frac{2\pi e \lambda}{N - \alpha_2 D^2} \right).$$

The desired result now follows immediately from the general definition in (47) and the expressions of $V(\cdot)$ in (a) and $V^{\mathrm{cl}}(\cdot)$ in (b). ■

In other words, the exploration cost for stochastic LQ problems can be completely pre-determined by the learning agent through choosing her individual parameters λ and ρ, since the cost relies neither on the specific (unknown) linear state dynamics, nor on the quadratic reward structure.

Moreover, the exploration cost (48) 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.[15]

[15]The connection between a discounting parameter and an effective length of time horizon is well known in the discrete time discounted reward formulation $\mathbb{E}[\sum_{t\geq 0} \gamma^t R_t]$

5.4 Vanishing exploration

Herein, the exploration weight λ has been taken as an exogenous parameter reflecting the level of exploration desired by the learning agent. The smaller this parameter is, the more emphasis is placed on exploitation. When this parameter is sufficiently close to zero, the exploratory formulation is sufficiently close to the problem without exploration. Naturally, a desirable result is that if the exploration weight λ goes to zero, then the entropy-regularized LQ problem would converge to its classical counterpart. The following result makes this precise.

Theorem 9 *Assume that statement (a) (or equivalently, (b)) of Theorem 7 holds. Then, for each $x \in \mathbb{R}$,*

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

Moreover, for each $x \in \mathbb{R}$,

$$\lim_{\lambda \to 0} |V(x) - V^{cl}(x)| = 0.$$

Proof. The weak convergence of the feedback controls follows from the explicit forms of $\boldsymbol{\pi}^*$ and $\boldsymbol{u}^*$ in statements (a) and (b), and the fact that α_1, α_2 are independent of λ. The pointwise convergence of the value functions follows easily from the forms of $V(\cdot)$ and $V^{\mathrm{cl}}(\cdot)$, together with the fact that

$$\lim_{\lambda \to 0} \frac{\lambda}{2\rho} \left(\ln \left(\frac{2\pi e \lambda}{N - \alpha_2 D^2} \right) - 1 \right) = 0.$$

■

6 Conclusions

This paper approaches RL from a stochastic control perspective. Indeed, control and RL both deal with the problem of managing dynamic and stochastic systems by making the best use of available information. However,

for classical Markov Decision Processes (MDP) (see, among others, Derman (1970)). This infinite horizon discounted problem can be viewed as an undiscounted, finite horizon problem with a random termination time T that is geometrically distributed with parameter $1 - \gamma$. Hence, an effectively longer horizon with mean $\frac{1}{1-\gamma}$ is applied to the optimization problem as γ increases. Since a smaller ρ in the continuous time objective (2) or (12) corresponds to a larger γ in the discrete time objective, we can see the similar effect of a decreasing ρ on the effective horizon of continuous time problems.

as a recent survey paper Recht (2018) points out, "*...That the RL and control communities remain practically disjoint has led to the co-development of vastly different approaches to the same problems....*" It is our view that communication and exchange of ideas between the two fields are of paramount importance to the progress of both fields, for an old idea from one field may well be a fresh one to the other. The continuous-time relaxed stochastic control formulation employed in this paper exemplifies such a vision.

The main contributions of this paper are *conceptual* rather than *algorithmic*: casting the RL problem in a continuous-time setting and with the aid of stochastic control and stochastic calculus, we interpret and explain why the Gaussian distribution is best for exploration in RL. This finding is independent of the specific parameters of the underlying dynamics and reward function structure, as long as the dependence on actions is linear in the former and quadratic in the latter. The same can be said about other main results of the paper, such as the separation between exploration and exploitation in the mean and variance of the resulting Gaussian distribution, and the cost of exploration. The explicit forms of the derived optimal Gaussian distributions do indeed depend on the model specifications which are unknown in the RL context. With regards to implementing RL algorithms based on our results for LQ problems, we can either do it in continuous time and space directly following, for example, Doya (2000), or modify the problem into an MDP one by discretizing the time, and then learn the parameters of the optimal Gaussian distribution following standard RL procedures (e.g. the so-called Q-learning). For that, our results may again be useful: they suggest that we only need to learn among the class of simpler Gaussian policies, i.e., $\pi = \mathcal{N}(\cdot\,|\theta_1 x + \theta_2, \phi)$ (cf. (40)), rather than generic (nonlinear) parametrized Gaussian policy $\pi_{\theta,\phi} = \mathcal{N}(\cdot\,|\theta(x), \phi(x))$. We expect that this simpler functional form can considerably increase the learning speed.

Appendix A: Explicit Solutions to (28)

For a range of parameters, we derive explicit solutions to SDE (28) satisfied by the optimal state process $\{X_t^*, t \geq 0\}$.

If $D = 0$, the SDE (28) reduces to

$$dX_t^* = \left(AX_t^* - \frac{BQ}{N}\right) dt + |C|\,|X_t^*|\, dW_t, \quad X_0^* = x.$$

If $x \geq 0$ and $BQ \leq 0$, the above equation has a nonnegative solution given

by

$$X_t^* = xe^{\left(A-\frac{C^2}{2}\right)t+|C|W_t} - \frac{BQ}{N}\int_0^t e^{\left(A-\frac{C^2}{2}\right)(t-s)+|C|(W_t-W_s)}ds.$$

If $x \leq 0$ and $BQ \geq 0$, it has a nonpositive solution

$$X_t^* = xe^{\left(A-\frac{C^2}{2}\right)t-|C|W_t} - \frac{BQ}{N}\int_0^t e^{\left(A-\frac{C^2}{2}\right)(t-s)-|C|(W_t-W_s)}ds.$$

These two cases cover the special case when $Q = 0$ which is standard in the LQ control formulation. We are unsure if there is an explicit solution when neither of these assumptions is satisfied (e.g. when $x \geq 0$ and $BQ > 0$).

If $C = 0$, the SDE (28) becomes

$$dX_t^* = \left(AX_t^* - \frac{BQ}{N}\right)dt + \frac{|D|}{N}\sqrt{Q^2+\lambda N}\,dW_t,$$

and its unique solution is given by

$$X_t^* = xe^{At} - \frac{BQ}{AN}(1-e^{At}) + \frac{|D|}{N}\sqrt{Q^2+\lambda N}\int_0^t e^{A(t-s)}dW_s, \quad t \geq 0,$$

if $A \neq 0$, and by

$$X_t^* = x - \frac{BQ}{N}t + \frac{|D|}{N}\sqrt{Q^2+\lambda N}W_t, \quad t \geq 0,$$

if $A = 0$.

If $C \neq 0$ and $D \neq 0$, then the diffusion coefficient of SDE (28) is C^2 in the unknown, with the first and second order derivatives being bounded. Hence, (28) can be solved explicitly using the Doss-Saussman transformation (see, for example, Karatzas and Shreve (1991), pp 295-297). This transformation uses the ansatz

$$X_t^*(\omega) = F(W_t(\omega), Y_t(\omega)), \quad t \geq 0,\ \omega \in \Omega \tag{49}$$

for some deterministic function F and an adapted process Y_t, $t \geq 0$, solving a random ODE. Applying Itô's formula to (49) and using the dynamics in (28), we deduce that F solves, for each fixed y, the ODE

$$\frac{\partial F}{\partial z} = \sqrt{\left(CF(z,y) - \frac{DQ}{N}\right)^2 + \frac{\lambda D^2}{N}}, \quad F(0,y) = y. \tag{50}$$

Moreover, Y_t, $t \geq 0$, is the unique pathwise solution to the random ODE

$$\frac{d}{dt}Y_t(\omega) = G(W_t(\omega), Y_t(\omega)), \quad Y_0(\omega) = x, \tag{51}$$

where

$$G(z,y) = \frac{AF(z,y) - \frac{BQ}{N} - \frac{C}{2}\left(CF(z,y) - \frac{DQ}{N}\right)}{\frac{\partial}{\partial y}F(z,y)}.$$

It is easy to verify that both equations (50) and (51) have a unique solution. Solving (50), we obtain

$$F(z,y) = \sqrt{\frac{\lambda}{N}}\left|\frac{D}{C}\right| \sinh\left(|C|z + \sinh^{(-1)}\left(\sqrt{\frac{N}{\lambda}}\left|\frac{C}{D}\right|\left(y - \frac{DQ}{CN}\right)\right)\right) + \frac{DQ}{CN}.$$

This, in turn, leads to the explicit expression of the function $G(z,y)$.

Appendix B: Proof of Theorem 4

Recall that the function v, where $v(x) = \frac{1}{2}k_2x^2 + k_1x + k_0, \quad x \in \mathbb{R}$, where k_2, k_1 and k_0 are defined by (36), (37) and (38), respectively, satisfies the HJB equation (14).

Throughout this proof we fix the initial state $x \in \mathbb{R}$. Let $\pi \in \mathcal{A}(x)$ and X^π be the associated state process solving (22) with π being used. Let $T > 0$ be arbitrary. Define the stopping times $\tau_n^\pi := \{t \geq 0 : \int_0^t \left(e^{-\rho t}v'(X_t^\pi)\tilde{\sigma}(X_t^\pi, \pi_t)\right)^2 dt \geq n\}$, for $n \geq 1$. Then, Itô's formula yields

$$e^{-\rho(T\wedge\tau_n^\pi)}v(X_{T\wedge\tau_n^\pi}^\pi) = v(x) + \int_0^{T\wedge\tau_n^\pi} e^{-\rho t}\Big(- \rho v(X_t^\pi) + \frac{1}{2}v''(X_t^\pi)\tilde{\sigma}^2(X_t^\pi, \pi_t)$$

$$+v'(X_t^\pi)\tilde{b}(X_t^\pi, \pi_t)\Big)\, dt + \int_0^{T\wedge\tau_n^\pi} e^{-\rho t}v'(X_t^\pi)\tilde{\sigma}(X_t^\pi, \pi_t)\, dW_t.$$

Taking expectations, using that v solves the HJB equation (14) and that π is in general suboptimal yield

$$\mathbb{E}\left[e^{-\rho(T\wedge\tau_n^\pi)}v(X_{T\wedge\tau_n^\pi}^\pi)\right]$$

$$= v(x) + \mathbb{E}\left[\int_0^{T\wedge\tau_n^\pi} e^{-\rho t}\left(-\rho v(X_t^\pi) + \frac{1}{2}v''(X_t^\pi)\tilde{\sigma}^2(X_t^\pi, \pi_t) + v'(X_t^\pi)\tilde{b}(X_t^\pi, \pi_t)\right)\, dt\right]$$

$$\leq v(x) - \mathbb{E}\left[\int_0^{T\wedge\tau_n^\pi} e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) - \lambda\int_{\mathbb{R}} \pi_t(u)\ln\pi_t(u)du\right) dt\right].$$

Classical results yield that $\mathbb{E}\left[\sup_{0\leq t\leq T}|X_t^\pi|^2\right] \leq K(1+x^2)e^{KT}$, for some constant $K > 0$ independent of n (but dependent on T and the model

coefficients). Sending $n \to \infty$, we deduce that

$$\mathbb{E}\left[e^{-\rho T} v(X_T^\pi)\right] \leq v(x) - \mathbb{E}\left[\int_0^T e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) - \lambda \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du\right) dt\right],$$

where we have used the dominated convergence theorem and that $\pi \in \mathcal{A}(x)$.

Next, we recall the admissibility condition $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\left[(X_T^\pi)^2\right] = 0$. This, together with the fact that $k_2 < 0$, lead to $\limsup_{T\to\infty} \mathbb{E}\left[e^{-\rho T} v(X_T^\pi)\right] = 0$. Applying the dominated convergence theorem once more yields

$$v(x) \geq \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) - \lambda \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du\right) dt\right],$$

for each $x \in \mathbb{R}$ and $\pi \in \mathcal{A}(x)$. Hence, $v(x) \geq V(x)$, for all $x \in \mathbb{R}$.

On the other hand, we deduce that the right hand side of (14) is maximized at

$$\boldsymbol{\pi}^*(u; x) = \mathcal{N}\left(u \,\middle|\, \frac{CDxv''(x) + Bv'(x) - Rx - Q}{N - D^2 v''(x)}, \frac{\lambda}{N - D^2 v''(x)}\right).$$

Let $\pi^* = \{\pi_t^*, t \geq 0\}$ be the open-loop control distribution generated from the above feedback law along with the corresponding state process $\{X_t^*, t \geq 0\}$ with $X_0^* = x$, and assume for now that $\pi^* \in \mathcal{A}(x)$. Then

$$\mathbb{E}\left[e^{-\rho T} v(X_T^*)\right] = v(x) - \mathbb{E}\left[\int_0^T e^{-\rho t}\left(\tilde{r}(X_t^*, \pi_t^*) - \lambda \int_{\mathbb{R}} \pi_t^*(u) \ln \pi_t^*(u) du\right) dt\right].$$

Noting that $\liminf_{T\to\infty} \mathbb{E}\left[e^{-\rho T} v(X_T^*)\right] \leq \limsup_{T\to\infty} \mathbb{E}\left[e^{-\rho T} v(X_T^*)\right] = 0$, and applying the dominated convergence theorem yield

$$v(x) \leq \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^*, \pi_t^*) - \lambda \int_{\mathbb{R}} \pi_t^*(u) \ln \pi_t^*(u) du\right) dt\right],$$

for any $x \in \mathbb{R}$. This proves that v is indeed the value function, namely $v \equiv V$.

It remains to show that $\pi^* \in \mathcal{A}(x)$. First, we verify that

$$\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\left[(X_T^*)^2\right] = 0, \tag{52}$$

where $\{X_t^*, t \geq 0\}$ solves the SDE (41). To this end, Itô's formula yields, for any $T \geq 0$,

$$(X_T^*)^2 = x^2 + \int_0^T \left(2\left(\tilde{A}X_t^* + \tilde{B}\right) X_t^* + (\tilde{C}_1 X_t^* + \tilde{C}_2)^2 + D^2\right) dt$$

$$+\int_0^T 2X_t^*\sqrt{\left(\tilde{C}_1X_t^*+\tilde{C}_2\right)^2+\tilde{D}^2}\,dW_t. \tag{53}$$

Following similar arguments as in the proof of Lemma 10 in Appendix C, we can show that $\mathbb{E}[(X_T^*)^2]$ contains the terms $e^{(2\tilde{A}+\tilde{C_1}^2)T}$ and $e^{\tilde{A}T}$.

If $2\tilde{A}+\tilde{C_1}^2\leq\tilde{A}$, then $\tilde{A}\leq 0$, in which case (52) easily follows. Therefore, to show (52), it remains to consider the case in which the term $e^{(2\tilde{A}+\tilde{C_1}^2)T}$ dominates $e^{\tilde{A}T}$, as $T\to\infty$. In turn, using that k_2 solves the equation (33), we obtain

$$2\tilde{A}+\tilde{C_1}^2-\rho=2A+\frac{2B(k_2(B+CD)-R)}{N-k_2D^2}+\left(C+\frac{D(k_2(B+CD)-R)}{N-k_2D^2}\right)^2-\rho$$

$$=2A+C^2-\rho+\frac{2(B+CD)(k_2(B+CD)-R)}{N-k_2D^2}+\frac{D^2(k_2(B+CD)-R)^2}{(N-k_2D^2)^2}$$

$$=2A+C^2-\rho+\frac{k_2(2N-k_2D^2)(B+CD)^2}{N-k_2D^2}-\frac{2NR(B+CD)-D^2R^2}{N-k_2D^2}. \tag{54}$$

Notice that the first fraction is nonpositive due to $k_2<0$, while the second fraction is bounded for any $k_2<0$. Using Assumption 3 on the range of ρ, we then easily deduce (52).

Next, we establish the admissibility constraint

$$\mathbb{E}\left[\int_0^\infty e^{-\rho t}\left|L(X_t^*,\pi_t^*)\right|dt\right]<\infty.$$

The definition of L and the form of $r(x,u)$ yield

$$\mathbb{E}\left[\int_0^\infty e^{-\rho t}\left|L(X_t^*,\pi_t^*)\right|\,dt\right]$$

$$=\mathbb{E}\left[\int_0^\infty e^{-\rho t}\left|\int_{\mathbb{R}}r(X_t^*,u)\pi_t^*(u)du-\lambda\int_{\mathbb{R}}\pi_t^*(u)\ln\pi_t^*(u)du\right|\,dt\right]$$

$$=\mathbb{E}\Big[\int_0^\infty e^{-\rho t}\Big|\int_{\mathbb{R}}-\left(\frac{M}{2}\left(X_t^*\right)^2+RX_t^*u+\frac{N}{2}u^2+PX_t^*+Qu\right)\pi_t^*(u)du$$

$$+\frac{\lambda}{2}\ln\left(\frac{2\pi e\lambda}{N-k_2D^2}\right)\Big|\,dt\Big],$$

where we have applied similar computations as in the proof of Theorem 8.

Recall that

$$\pi_t^*(u)=\mathcal{N}\left(u\;\middle|\;\frac{(k_2(B+CD)-R)X_t^*+k_1B-Q}{N-k_2D^2}\;,\;\frac{\lambda}{N-k_2D^2}\right),\quad t\geq 0.$$

It is then clear that it suffices to prove $\mathbb{E}\left[\int_0^\infty e^{-\rho t}(X_t^*)^2 dt\right] < \infty$, which follows easily since, as shown in (54), we obtained that $\rho > 2\tilde{A} + \tilde{C_1}^2$ under Assumption 3. The remaining admissibility conditions for π^* can be easily verified.

Appendix C: Proof of Theorem 7

We first note that when (a) holds, the function v solves the HJB equation (32) of the exploratory LQ problem. Similarly for w of the classical LQ problem when (b) holds.

Next, we prove the equivalence between (a) and (b). First, a comparison between the two HJB equations (32) and (45) yields that if v in (a) solves the former, then w in (b) solves the latter, and vice versa.

Throughout this proof, we let x be fixed, being the initial state of both the exploratory problem in statement (a) and the classical problem in statement (b). Let $\pi^* = \{\pi_t^*, t \geq 0\}$ and $u^* = \{u_t^*, t \geq 0\}$ be respectively the open-loop controls generated by the feedback controls $\boldsymbol{\pi}^*$ and $\boldsymbol{u}^*$ of the two problems, and $X^* = \{X_t^*, t \geq 0\}$ and $x^* = \{x_t^*, t \geq 0\}$ be respectively the corresponding state processes, both starting from x. It remains to establish the equivalence between the admissibility of π^* for the exploratory problem and that of u^* for the classical problem. To this end, we first compute $\mathbb{E}[(X_T^*)^2]$ and $\mathbb{E}[(x_T^*)^2]$.

To ease the presentation, we rewrite the exploratory dynamics of X^* under π^* as

$$\begin{aligned} dX_t^* &= \left(AX_t^* + B\frac{(\alpha_2(B+CD)-R)X_t^* + \alpha_1 B - Q}{N-\alpha_2 D^2}\right)\, dt \\ &+\sqrt{\left(CX_t^* + D\frac{(\alpha_2(B+CD)-R)X_t^* + \alpha_1 B - Q}{N-\alpha_2 D^2}\right)^2 + \frac{\lambda D^2}{N-\alpha_2 D^2}}\, dW_t \\ &= (A_1 X_t^* + A_2)\, dt + \sqrt{(B_1 X_t^* + B_2)^2 + C_1}\, dW_t, \end{aligned}$$

where $A_1 := A + \frac{B(\alpha_2(B+CD)-R)}{N-\alpha_2 D^2}$, $A_2 := \frac{B(\alpha_1 B - Q)}{N-\alpha_2 D^2}$, $B_1 := C + \frac{D(\alpha_2(B+CD)-R)}{N-\alpha_2 D^2}$, $B_2 := \frac{D(\alpha_1 B - Q)}{N-\alpha_2 D^2}$ and $C_1 := \frac{\lambda D^2}{N-\alpha_2 D^2}$.

Similarly, the classical dynamics of x^* under u^* solves

$$dx_t^* = (A_1 x_t^* + A_2)\, dt + (B_1 x_t^* + B_2)\, dW_t.$$

The desired equivalence of the admissibility then follows from the following lemma.

Lemma 10 *We have that (i)* $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\big[\big(X_T^*\big)^2\big] = 0$ *if and only if* $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\big[\big(x_T^*\big)^2\big] = 0$ *and (ii)* $\mathbb{E}\left[\int_0^\infty e^{-\rho t}\big(X_t^*\big)^2 dt\right] < \infty$ *if and only if* $\mathbb{E}\left[\int_0^\infty e^{-\rho t}\big(x_t^*\big)^2 dt\right] < \infty$.

Proof. Denote $n(t) := \mathbb{E}[X_t^*]$, for $t \geq 0$. Then, a standard argument involving a series of stopping times and the dominated convergence theorem yields the ODE

$$\frac{dn(t)}{dt} = A_1 n(t) + A_2, \quad n(0) = x,$$

whose solution is $n(t) = \left(x + \frac{A_2}{A_1}\right) e^{A_1 t} - \frac{A_2}{A_1}$, if $A_1 \neq 0$, and $n(t) = x + A_2 t$, if $A_1 = 0$. Similarly, the function $m(t) := \mathbb{E}\big[(X_t^*)^2\big]$, $t \geq 0$, solves the ODE

$$\frac{dm(t)}{dt} = (2A_1 + B_1^2)m(t) + 2(A_2 + B_1 B_2)n(t) + B_2^2 + C_1, \quad m(0) = x^2.$$

We can also show that $n(t) = \mathbb{E}\big[x_t^*\big]$, and deduce that $\hat{m}(t) := \mathbb{E}\big[(x_t^*)^2\big]$, $t \geq 0$, satisfies

$$\frac{d\hat{m}(t)}{dt} = (2A_1 + B_1^2)\hat{m}(t) + 2(A_2 + B_1 B_2)n(t) + B_2^2, \quad \hat{m}(0) = x^2.$$

Next, we find explicit solutions to the above ODEs corresponding to various conditions on the parameters.

(a) If $A_1 = B_1^2 = 0$, then direct computation gives $n(t) = x + A_2 t$, and

$$m(t) = x^2 + A_2(x + A_2 t)t + (B_2^2 + C_1)t,$$

$$\hat{m}(t) = x^2 + A_2(x + A_2 t)t + B_2^2 t.$$

(b) If $A_1 = 0$ and $B_1^2 \neq 0$, we have $n(t) = x + A_2 t$, and

$$m(t) = \left(x^2 + \frac{2(A_2 + B_1 B_2)\left(A_2 + B_1^2(x + B_2^2 + C_1)\right)}{B_1^4}\right) e^{B_1^2 t}$$
$$- \frac{2(A_2 + B_1 B_2)\left(A_2 + B_1^2(x + B_2^2 + C_1)\right)}{B_1^4},$$

$$\hat{m}(t) = \left(x^2 + \frac{2(A_2 + B_1 B_2)\left(A_2 + B_1^2(x + B_2^2)\right)}{B_1^4}\right) e^{B_1^2 t}$$
$$- \frac{2(A_2 + B_1 B_2)\left(A_2 + B_1^2(x + B_2^2)\right)}{B_1^4}.$$

(c) If $A_1 \neq 0$ and $A_1 + B_1^2 = 0$, then $n(t) = \left(x + \frac{A_2}{A_1}\right) e^{A_1 t} - \frac{A_2}{A_1}$. Further calculations yield

$$m(t) = \left(x^2 + \frac{A_1(B_2^2 + C_1) - 2A_2(A_2 + B_1 B_2)}{A_1^2}\right) e^{A_1 t}$$
$$+ \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1} t e^{A_1 t} - \frac{A_1(B_2^2 + C_1) - 2A_2(A_2 + B_1 B_2)}{A_1^2},$$
$$\hat{m}(t) = \left(x^2 + \frac{A_1 B_2^2 - 2A_2(A_2 + B_1 B_2)}{A_1^2}\right) e^{A_1 t}$$
$$+ \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1} t e^{A_1 t} - \frac{A_1 B_2^2 - 2A_2(A_2 + B_1 B_2)}{A_1^2}.$$

(d) If $A_1 \neq 0$ and $2A_1 + B_1^2 = 0$, we have $n(t) = \left(x + \frac{A_2}{A_1}\right) e^{A_1 t} - \frac{A_2}{A_1}$, and

$$m(t) = \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1^2} e^{A_1 t}$$
$$+ \frac{A_1(B_2^2 + C_1) - 2A_2(A_2 + B_1 B_2)}{A_1^2} t + x^2 - \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1^2},$$
$$\hat{m}(t) = \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1^2} e^{A_1 t}$$
$$+ \frac{A_1 B_2^2 - 2A_2(A_2 + B_1 B_2)}{A_1^2} t + x^2 - \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1^2}.$$

(e) If $A_1 \neq 0$, $A_1 + B_1^2 \neq 0$ and $2A_1 + B_1^2 \neq 0$, then we arrive at $n(t) = \left(x + \frac{A_2}{A_1}\right) e^{A_1 t} - \frac{A_2}{A_1}$, and

$$m(t) =$$
$$\left(x^2 + \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1(A_1 + B_1^2)} + \frac{A_1(B_2^2 + C_1) - 2A_2(A_2 + B_1 B_2)}{A_1(2A_1 + B_1^2)}\right) e^{(2A_1 + B_1^2)t}$$
$$- \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1(A_1 + B_1^2)} e^{A_1 t} - \frac{A_1(B_2^2 + C_1) - 2A_2(A_2 + B_1 B_2)}{A_1(2A_1 + B_1^2)},$$
$$\hat{m}(t) = \left(x^2 + \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1(A_1 + B_1^2)} + \frac{A_1 B_2^2 - 2A_2(A_2 + B_1 B_2)}{A_1(2A_1 + B_1^2)}\right) e^{(2A_1 + B_1^2)t}$$
$$- \frac{2(A_2 + B_1 B_2)(A_1 x + A_2)}{A_1(A_1 + B_1^2)} e^{A_1 t} - \frac{A_1 B_2^2 - 2A_2(A_2 + B_1 B_2)}{A_1(2A_1 + B_1^2)}.$$

It is easy to see that for all cases (a)–(e), the assertions in the Lemma follow and we conclude. ■

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Ronen I Brafman and Moshe Tennenholtz. R-max—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Cyrus Derman. *Finite state Markovian decision processes*. Acedemic Press, New York, 1970.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245, 2000.

Nicole El Karoui, Nguyen Du Huu, and Monique Jeanblanc-Picqué. Compactification methods in the control of degenerate diffusions: existence of an optimal control. *Stochastics*, 20(3):169–219, 1987.

Wendell H Fleming and Makiko Nisio. On stochastic relaxed control for partially observed diffusions. *Nagoya Mathematical Journal*, 93:71–108, 1984.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pages 202–211, 2016.

John Gittins. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266, 1974.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1352–1361, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, pages 703–710, 1994.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Ioannis Karatzas and Steven E Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, 2nd edition, 1991.

Haya Kaspi and Avishai Mandelbaum. Multi-armed bandits in discrete and continuous time. *Annals of Applied Probability*, pages 1270–1290, 1998.

Thomas Kurtz and Richard Stockbridge. Existence of Markov controls and characterization of optimal Markov controls. *SIAM Journal on Control and Optimization*, 36(2):609–653, 1998.

Thomas Kurtz and Richard Stockbridge. Stationary solutions and forward equations for controlled and singular martingale problems. *Electronic Journal of Probability*, 6, 2001.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Weiwei Li and Emanuel Todorov. Iterative linearization methods for approximately optimal control and estimation of non-linear stochastic system. *International Journal of Control*, 80(9):1439–1453, 2007.

Timothy Lillicrap, Jonathan Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Avi Mandelbaum. Continuous multi-armed bandits and multi-parameter processes. *The Annals of Probability*, pages 1527–1556, 1987.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, and Koray Kavukcuoglu. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2775–2785, 2017.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Trust-pcl: An off-policy trust region method for continuous control. In *International Conference on Learning Representations*, 2018.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2018.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *arXiv preprint arXiv:1806.09460v2*, 2018.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

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

Xun Yu Zhou. On the existence of optimal relaxed controls of stochastic partial differential equations. *SIAM Journal on Control and Optimization*, 30(2):247–261, 1992.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.