# Inverse stochastic optimal controls

Yumiharu Nakano

Department of Mathematical and Computing Science, School of Computing
Tokyo Institute of Technology
W8-28, 2-12-1, Ookayama, Meguro-ku, Tokyo 152-8550, Japan
e-mail: nakano@c.titech.ac.jp

May 26, 2020

**Abstract**

We study an inverse problem of the stochastic control of general diffusions with performance index having the quadratic penalty term of the control process. Under mild conditions on the drift, the volatility, the cost functions of the state, and under the assumption that the optimal control belongs to the interior of the control set, we show that our inverse problem is well-posed. Then, with the well-posedness, we reduce the inverse problem to some root finding problem of the expectation of a random variable involved with the value function, which has a unique solution. Based on this result, we propose a numerical method for our inverse problem by replacing the expectation above with arithmetic mean of observed optimal control processes and the corresponding state processes. Several numerical experiments show that the numerical method recover the unknown weight parameter with high accuracy.

**Key words**: Inverse problems, stochastic control, stochastic maximum principle, Hamilton-Jacobi-Bellman equations, Kernel-based collocation methods.

## 1 Introduction

The *inverse optimal controls* refers to the problem of determining the performance index that makes a given control law optimal. Bellman and Kalaba [2] derives some equations for cost functions using the feedback function of the optimal control for finite horizon problems. Kalman [12] analyses the case of linear quadratic models in infinite horizon in details. We also refer to Thau [21] and Casti [6] for another early contributions. Dvijotham and Todorov [8] works under a linearly solvable class of stochastic optimal control in continuous-time, where explicit forms of the optimal feedback laws are available. In literature of reinforcement learning, the inverse problems are studied by, e.g., Ng and Russell [18], Abbeel and Ng [1], and Ziebart et.al [23].

In this paper, we are concerned with the inverse problem of optimal stochastic controls of general diffusion processes. Consider the $d$-dimensional controlled stochastic differential equation

$$dX(t) = b(t, X(t), u(t))dt + \sigma(t, X(t), u(t))dW(t) \tag{1.1}$$

with initial condition $X_0$, where $\{W(t)\}_{t\geqslant 0}$ is an $m$-dimensional standard Brownian motion on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the random variable $X_0$ is independent of $\{W(t)\}_{t\geqslant 0}$. The functions $b$ and $\sigma$ are assumed to be measurable and satisfy suitable conditions for ensuring the existence and uniqueness of (1.1), which will be precisely described in Section 2 below. The control processes $\{u(t)\}_{0\leqslant t\leqslant T}$ are assumed to take values in a closed set $U$ of $\mathbb{R}^k$. Denote by $\{\mathcal{F}(t)\}_{t\geqslant 0}$ the augmented filtration generated by $\{W(t)\}_{t\geqslant 0}$ and $X_0$. The class $\mathcal{U}$ of controls is then defined by the set of all $U$-valued $\{\mathcal{F}(t)\}$-adapted processes $\{u(t)\}_{0\leqslant t\leqslant T}$ satisfying

$$\mathbb{E}\int_0^T |u(t)|^2 dt < \infty.$$

We are concerned with the following optimal control problem:

**Problem** $(\mathcal{C}_\theta)$**.** Minimize

$$J[u] := \mathbb{E}\left[g(X(T)) + \int_0^T \left\{f(t, X(t)) + \theta|u(t)|^2\right\} dt\right]$$

over $u \in \mathcal{U}$.

Here we have assumed that the running cost is decomposed into that for the state and that for the penalty of the control input with weight parameter $\theta > 0$. Then, we consider the inverse problem with respect to $\theta$ given an optimal control and the corresponding state trajectory, i.e., our problem is to recover $\theta$ from a solution $\{u^*(t)\} \in \mathcal{U}$ for $\mathcal{C}(\theta)$ and the corresponding state process $\{X^*(t)\}$ under the assumption that $b$, $\sigma$, $g$ and $f$ are known.

Our first aim is to study Hadamard's well-posedness for the inverse problem above. Namely, we investigate the *existence*, the *uniqueness*, and the *stability* of $\theta$ given optimal $\{u^*(t)\} \in \mathcal{U}$ for $\mathcal{C}(\theta)$ and the corresponding state process $\{X^*(t)\}$. The existence is trivial. That is, for any observed control policy there exists $\theta > 0$ that makes the control process is optimal for $(\mathcal{C}_\theta)$, since, in our framework, observed controls are assumed to be optimal for $(\mathcal{C}_{\theta'})$ for some $\theta' > 0$. As one might predicted, the uniqueness and the stability do not hold in general. We will give such an example. On the other hand, under additional mild conditions on $b$, $\sigma$, $g$, $f$, and the optimal control process, we will show that the uniqueness and the stability do hold using the stochastic maximum principle.

Our second aim is to propose a numerical method for the inverse problem above. The problem is, given $N$ independent samples $\{(u^{(j)}(t_i), X^{(j)}(t_i))\}_{i=0}^n$, $j = 1, \ldots, N$, of $\{(u^*(t_i), X^*(t_i))\}_{i=0}^n$ for $\mathcal{C}(\theta)$, to determine $\theta$ computationally, where $0 = t_0 < t_1 < \ldots < t_n = T$. Our approach is based on the simple optimally condition $J(u^*) = \inf_{u\in\mathcal{U}} J(u)$ and is a reduction to some root-finding problem. This of course involves the value function computations, and so we rely on the recent progress of numerical analyses of Hamilton-Jacobi-Bellman equations. Thanks to the uniqueness result of the inverse problem, in our numerical experiments performed in Section 3, the positive root $\theta$'s are recovered with high accuracy.

The present paper is organized as follows: Section 2 is devoted to the issue of the well-posedness. In Section 3, we propose a numerical method for our inverse problems and validate it. Section 4 concludes.

## 2 Well-posedness

First we shall discuss Hadamard's well-posedness of the inverse problem of the stochastic optimal control. As stated in Section 1, the existence is trivial. Then, as in many other inverse problems, the uniqueness and the stability do not hold in general for continuous-time optimal controls.

**Example 2.1.** Consider the state equation

$$\frac{dX(t)}{dt} = u(t)X(t)dt,$$

with nonrandom initial condition $X_0$. The control processes $\{u(t)\}$ are taken from the class of $[0,1]$-valued Borel function on $[0,T]$. The objective function $J[u]$ is assumed to be given by

$$J[u] = X(T) + \theta \int_0^T u(t)dt,$$

where $\theta > 0$. Since $X(T) = X_0 e^{\int_0^T u(t)dt}$, we obtain $\inf_u J[u] = \min_{0\leqslant\gamma\leqslant T}\{X_0 e^{\gamma} + \theta\gamma\}$. Thus, if $X_0 \geqslant 0$ then $\gamma \equiv 0$, i.e., $u \equiv 0$, is optimal for any problem $(\mathcal{C}_\theta)$ with $\theta > 0$.

Denote by $\mathrm{int}(U)$ the interior of $U$. Denote also by $D_x\varphi$ and $D_x^2\varphi$ the gradient vector and the Hessian matrix of $\varphi$ with respect to $x$, respectively. To discuss the uniqueness, we impose the following:

**Assumption 2.2.** (i) The random variable $X_0$ is a constant.

(ii) The set $\mathrm{int}(U)$ is nonempty.

(iii) The functions $b$, $\sigma$, $f$, and $g$ are of $C^2$-class in $x$.

(iv) There exists a constant $C_0 > 0$ and a modulus of continuity $\rho : [0,\infty) \to [0,\infty)$ such that for $\varphi(t,x,u) = b(t,x,u)$, $\sigma(t,x,u)$, $f(t,x)$, $g(x)$, we have

$$\begin{aligned}
|\varphi(t,x,u) - \varphi(t,x',u')| &\leqslant C_0|x-x'| + \rho(|u-u'|),\\
|\varphi(t,0,u)| &\leqslant C_0,\\
|D_x\varphi(t,x,u) - D_x\varphi(t,x',u')| &\leqslant C_0|x-x'| + \rho(|u-u'|),\\
|D_x^2\varphi(t,x,u) - D_x^2\varphi(t,x',u')| &\leqslant \rho(|x-x'| + |u-u'|),
\end{aligned}$$

for $t \in [0,T]$, $x, x' \in \mathbb{R}^d$, and $u, u' \in U$.

It should be noted that Assumption 2.2 (i) means that the filtration $\{\mathcal{F}(t)\}_{t\geqslant 0}$ is generated by the Brownian motion only. Assumption 2.2 (ii) excludes the case where $U$ is countable. The first two conditions in Assumption 2.2 (iv) are standard ones for stochastic optimal control problems. Under these two requirements, there exists a unique strong solution $\{X(t)\}_{0\leqslant t\leqslant T}$ of (1.1) for each $u \in \mathcal{U}$. See Fleming and Soner [11].

Denote by $D_u\varphi$ the gradient of $\varphi$ with respect to $u$. We write $\mathbb{S}^d$ for the totality of symmetric $d \times d$ real matrices, and $a^{\mathsf{T}}$ for the transpose of a vector or matrix $a$. Under Assumption 2.2, we have the following:

**Theorem 2.3.** *Let Assumption* 2.2 *hold. Suppose that* $\{u^*(t)\}_{0\leqslant t\leqslant T} \in \mathcal{U}$ *be optimal both for the problem* $(\mathcal{C}_{\theta_1})$ *and* $(\mathcal{C}_{\theta_2})$ *for some* $\theta_1, \theta_2 > 0$. *Moreover, suppose that* $\mathbb{P}(u^*(t_0) \in \mathrm{int}(U)\backslash\{0\}) > 0$ *for some* $t_0 \in [0,T]$. *Then* $\theta_1 = \theta_2$.

*Proof.* Our basic tool is a stochastic maximum principle as described in Yong and Zhou [22]. Since $(X^*(t), u^*(t))$ is optimal both for $(\mathcal{C}_{\theta_1})$ and $(\mathcal{C}_{\theta_2})$, for each $\theta = \theta_1, \theta_2$, there exists a unique solution $(p(t), q(t))$, $0 \leqslant t \leqslant T$, of the backward stochastic differential equation (BSDE)

$$\begin{aligned} dp(t) = -\Big\{ & D_x b(t, X^*(t), u^*(t))^{\mathsf{T}} p(t) + \sum_{j=1}^m D_x \sigma_j(t, X^*(t), u^*(t))^{\mathsf{T}} q_j(t) \\ & - D_x f(t, X^*(t)) \Big\} dt + q(t) dW(t), \\ p(T) = -D_x g(X^*(T)), & \end{aligned} \tag{2.1}$$

as well as there exists a unique solution $(P(t), Q(t))$, $0 \leqslant t \leqslant T$, of the BSDE

$$\begin{aligned} dP(t) = -\Big\{ & D_x b(t, X^*(t), u^*(t))^{\mathsf{T}} P(t) + P(t) D_x b(t, X^*(t), u^*(t))^{\mathsf{T}} \\ & + \sum_{j=1}^m D_x \sigma_j(t, X^*(t), u^*(t))^{\mathsf{T}} P(t) D_x \sigma_j(t, X^*(t), u^*(t)) \\ & + \sum_{j=1}^m \Big\{ D_x \sigma_j(t, X^*(t), u^*(t))^{\mathsf{T}} Q_j(t) + Q_j(t) D_x \sigma_j(t, X^*(t), u^*(t)) \Big\} \\ & + D_x^2 H_0(t, X^*(t), u^*(t), p(t), q(t)) \Big\} dt + \sum_{j=1}^m Q_j(t) dW(t), \\ P(T) = -D_x^2 g(X^*(T)), & \end{aligned} \tag{2.2}$$

where $q(t) = (q_1(t), \dots, q_m(t))$, $Q(t) = (Q_1(t), \dots, Q_m(t))$, $\sigma_j(t,x,u) \in \mathbb{R}^d$ for each $j = 1, \dots, m$ such that $\sigma(t,x,u) = (\sigma_1(t,x,u), \dots, \sigma_m(t,x,u))$, and

$$H_0(t,x,u,p,q) := p^{\mathsf{T}} b(t,x,u) + \mathrm{tr}(q^{\mathsf{T}} \sigma(t,x,u)) - f(t,x) - \theta |u|^2.$$

In particular, $p(t)$, $q_1(t), \dots, q_m(t)$ are $\mathbb{R}^d$-valued and $Q_1(t), \dots, Q_m(t)$ are $\mathbb{S}^d$-valued, all of which are adapted processes satisfying

$$\mathbb{E} \int_0^T |\varphi(t)|^2 dt < \infty$$

for $\varphi = p, q_1, \dots, q_m, Q_1, \dots, Q_m$. Moreover, with the generalized Hamiltonian

$$\begin{aligned} \mathcal{H}(t,x,u) := H_0(t,x,u,p(t),q(t)) & - \frac{1}{2} \mathrm{tr} \Big[ \sigma(t, X^*(t), u^*(t))^{\mathsf{T}} P(t) \sigma(t, X^*(t), u^*(t)) \Big] \\ & + \frac{1}{2} \mathrm{tr} \Big\{ [\sigma(t,x,u) - \sigma(t, X^*(t), u^*(t))]^{\mathsf{T}} P(t) [\sigma(t,x,u) - \sigma(t, X^*(t), u^*(t))] \Big\} \end{aligned}$$

we have

$$\mathcal{H}(t, X^*(t), u^*(t)) = \max_{u \in U} \mathcal{H}(t, X^*(t), u). \tag{2.3}$$

See Chapter 3 in [22]. Now, we write $(p_i(t), q_i(t), P_i(t), Q_i(t))$ for the corresponding $(p(t), q(t), P(t), Q(t))$ for $\theta = \theta_1, \theta_2$. Similarly, we write $\mathcal{H}_i$ for the corresponding $\mathcal{H}$ for $\theta = \theta_1, \theta_2$. By the optimality condition (2.3) and our assumptions, we obtain

$$D_u \mathcal{H}_i(t_0, X^*(t_0), u^*(t_0)) = 0, \quad i = 1, 2,$$

with positive probability. So,

$$\begin{aligned} 0 &= K(t_0, X^*(t_0), u^*(t_0), p_1(t), P_1(t)) - 2\theta_1 u^*(t_0) \\ &= K(t_0, X^*(t_0), u^*(t_0), p_2(t), P_2(t)) - 2\theta_2 u^*(t_0) \end{aligned}$$

where for $t \in [0, T]$, $x, p \in \mathbb{R}^d$, and $P \in \mathbb{R}^{d \times d}$,

$$K(t, x, u, p, P) = D_u \left[ \frac{1}{2} \mathrm{tr} \left( \sigma(t, x, u)^\mathsf{T} P \sigma(t, x, u) \right) + p^\mathsf{T} b(t, x, u) \right].$$

Since the uniqueness of the BSDEs immediately leads to $p_1 = p_2$ and $P_1 = P_2$, the equalities above yield $\theta_1 = \theta_2$, as claimed. □

To guarantee the stability, we need to impose additional conditions.

**Assumption 2.4.** The functions $b$, $\sigma$, $f$, and $g$ are of $C^1$-class in $u$. Further, for $\varphi(t, x, u) = b(t, x, u)$, $\sigma(t, x, u)$, $f(t, x)$, $g(x)$, we have

$$\begin{aligned} &|\sigma(t, x, u)| + |D_u \sigma(t, x, u)| \leqslant C_0, \\ &|D_u \varphi(t, x, u) - D_u \varphi(t, x', u')| \leqslant \rho(|x - x'| + |u - u'|), \quad x, x' \in \mathbb{R}^d, \ \ u, u' \in U, \end{aligned}$$

where $C_0$ and $\rho$ are as in Assumption 2.2.

Then we have the following:

**Theorem 2.5.** *Let Assumptions* 2.2 *and* 2.4 *hold. Suppose that* $\{u^*(t)\}_{0 \leqslant t \leqslant T} \in \mathcal{U}$ *and* $\{u_n(t)\}_{0 \leqslant t \leqslant T} \in \mathcal{U}$ *are optimal for the problems* $(\mathcal{C}_{\theta^*})$ *and* $(\mathcal{C}_{\theta_n})$ *for some* $\theta^*, \theta_n > 0$, *respectively,* $n \in \mathbb{N}$, *such that*

$$\mathbb{E} \int_0^T |u_n(t) - u^*(t)|^2 dt \to 0, \quad n \to \infty.$$

*Moreover, suppose that there exists a measurable set* $E \subset [0, T] \times \Omega$ *with positive measure such that*

$$\lim_{n \to \infty} u_n(t, \omega) = u^*(t, \omega) \in \mathrm{int}(U) \backslash \{0\}, \quad (t, \omega) \in E.$$

*Then we have* $\lim_{n \to \infty} \theta_n = \theta^*$.

*Proof.* By $C$ we denote positive constants that may vary from line to line. Since $u_n$ converges to $u^* \in \mathrm{int}(U)$ pointwise on $E$, it follows that $u_n \in \mathrm{int}(U)$ for any sufficiently large $n$. Thus by (2.3),

$$2\theta_n u_n(t,\omega) - 2\theta^* u^*(t,\omega) = L_n(t,\omega) - L^*(t,\omega), \quad (t,\omega) \in E, \tag{2.4}$$

where $L_n(t,\omega) = K(t, X_n(t,\omega), u_n(t,\omega), p_n(t,\omega), P_n(t,\omega))$ with $X_n$ being the state process corresponding to $u_n$, and $(p_n, q_n)$ and $(P_n, Q_n)$ the solutions of the BSDE (2.1) and (2.2) with $u = u_n$ and $X = X_n$, respectively. $L^*$ is similarly defined. The usual arguments with Gronwall's lemma yield

$$\sup_{0\leqslant t\leqslant T} \mathbb{E}|X_n(t) - X^*(t)|^2 \leqslant C\mathbb{E}\int_0^T \rho(|u_n(t) - u^*(t)|)^2 dt. \tag{2.5}$$

Since $\rho$ is a modulus of continuity for $D_x^2 b(t,\cdot,u)$ uniformly in $t$ and $u$, we may assume that $\rho$ is bounded. Hence, for any $\varepsilon > 0$ there exists $\delta > 0$ such that $\rho(r) \leqslant \varepsilon + \delta r$, $r \geqslant 0$. From this and the $L^2$-convergence of $u_n$ we obtain

$$\sup_{0\leqslant t\leqslant T} \mathbb{E}|X_n(t) - X^*(t)|^2 \to 0, \quad n \to \infty. \tag{2.6}$$

Furthermore, by the a priori estimates for the BSDE (see, e.g., [22, Theorem 3.3, Chapter 7]),

$$\begin{aligned}
&\mathbb{E}\int_0^T |p_n(t) - p^*(t)|^2 dt \\
&\leqslant C\mathbb{E}|D_x g(X_n(T)) - D_x g(X^*(T))|^2 \\
&\quad + C\mathbb{E}\int_0^T |D_x b(t, X_n(t), u_n(t)) - D_x b(t, X^*(t), u^*(t))|^2 |p^*(t)|^2 dt \\
&\quad + C\mathbb{E}\int_0^T |D_x \sigma(t, X_n(t), u_n(t)) - D_x \sigma(t, X^*(t), u^*(t))|^2 |q^*(t)|^2 dt \\
&\quad + C\mathbb{E}\int_0^T |D_x f(t, X_n(t), u_n(t)) - D_x f(t, X^*(t), u^*(t))|^2 dt.
\end{aligned}$$

By Assumption 2.2 (iv) and (2.5),

$$\begin{aligned}
&\mathbb{E}|D_x g(X_n(T)) - D_x g(X^*(T))|^2 + \mathbb{E}\int_0^T |D_x f(t, X_n(t), u_n(t)) - D_x f(t, X^*(t), u^*(t))|^2 dt \\
&\leqslant C\mathbb{E}\int_0^T \rho(|u_n(t) - u^*(t)|)^2 dt \to 0, \quad n \to \infty.
\end{aligned}$$

Using Assumption 2.2 (iv) again, for $\varepsilon > 0$, we observe

$$\begin{aligned}
&\mathbb{E}\int_0^T |D_x b(t, X_n(t), u_n(t)) - D_x b(t, X^*(t), u^*(t))|^2 |p^*(t)|^2 dt \\
&\leqslant \mathbb{E}\int_0^T |D_x b(t, X_n(t), u_n(t)) - D_x b(t, X^*(t), u^*(t))|^2 |p^*(t)|^2 \\
&\qquad \times 1_{\{|X_n(t)-X^*(t)|+\rho(|u_n(t)-u^*(t)|)>\varepsilon\}} dt \\
&\quad + \mathbb{E}\int_0^T |D_x b(t, X_n(t), u_n(t)) - D_x b(t, X^*(t), u^*(t))|^2 |p^*(t)|^2 \\
&\qquad \times 1_{\{|X_n(t)-X^*(t)|+\rho(|u_n(t)-u^*(t)|)\leqslant\varepsilon\}} dt \\
&\leqslant C\mathbb{E}\int_0^T |p^*(t)|^2 1_{\{|X_n(t)-X^*(t)|+\rho(|u_n(t)-u^*(t)|)>\varepsilon\}} dt + C\varepsilon\mathbb{E}\int_0^T |p^*(t)|^2 dt.
\end{aligned}$$

Thus, letting $n \to \infty$ and then $\varepsilon \to 0$, we have

$$\lim_{n\to\infty} \mathbb{E}\int_0^T |D_x b(t, X_n(t), u_n(t)) - D_x b(t, X^*(t), u^*(t))|^2 |p^*(t)|^2 dt = 0.$$

Similarly,

$$\lim_{n\to\infty} \mathbb{E}\int_0^T D_x \sigma(t, X_n(t), u_n(t)) - D_x \sigma(t, X^*(t), u^*(t))|^2 |q^*(t)|^2 dt = 0.$$

Therefore,

$$\lim_{n\to\infty} \mathbb{E}\int_0^T |p_n(t) - p^*(t)|^2 = 0. \tag{2.7}$$

By the same way, we obtain

$$\lim_{n\to\infty} \mathbb{E}\int_0^T |P_n(t) - P^*(t)|^2 = 0. \tag{2.8}$$

For notational simplicity we denote $D_u b_n(t) = D_u b(t, X_n(t), u_n(t))$. Analogously we use the notation $D_u b^*(t)$, $\sigma_n(t)$, and $D_u \sigma_n(t)$. With this notation, by Assumption 2.4 we see

$$\begin{aligned}
&|L_n(t) - L^*(t)| \\
&\leqslant |D_u b_n(t)||p_n(t) - p^*(t)| + |p^*(t)||D_u b_n(t) - D_u b^*(t)| \\
&\quad + C|\sigma_n(t)||P_n(t)||D_u \sigma_n(t) - D_u \sigma^*(t)| + C|D_u \sigma^*(t)||P_n(t)||\sigma_n(t) - \sigma^*(t)| \\
&\quad + |D_u \sigma^*(t)||\sigma^*(t)|)|P_n(t) - P^*(t)| \\
&\leqslant C(1 + |X_n(t)| + |u_n(t)|)(|p_n(t) - p^*(t)| + |P_n(t) - P^*(t)|) \\
&\quad + C_\varepsilon(1 + |p^*(t)| + |p_n(t)|)(\varepsilon + |X_n(t) - X^*(t)| + |u_n(t) - u^*(t)|)
\end{aligned}$$

for any $\varepsilon > 0$ with constant $C_\varepsilon$ depending on $\varepsilon$. Thus, by Cauchy-Schwartz inequality and (2.6)–(2.8),

$$\lim_{n\to\infty} \mathbb{E}\int_0^T |L_n(t) - L^*(t)| dt = 0.$$

From this and (2.4) it follows that

$$2|\theta_n-\theta^*|\int_E |u_n|dt\times d\mathbb{P} \leqslant 2|\theta^*|\mathbb{E}\int_0^T |u_n(t)-u^*(t)|dt+\mathbb{E}\int_0^T |L_n(t)-L^*(t)|dt \to 0, \quad n\to\infty.$$

On the other hand,

$$\lim_{n\to\infty}\int_E |u_n|dt\times d\mathbb{P} = \int_E |u^*|dt\times d\mathbb{P} > 0.$$

Therefore, $\theta_n \to \theta^*$, as claimed. □

Next we consider the case of the linear quadratic regulator problems. We need a special treatment since Assumptions 2.2 and 2.4 (iii) exclude the case where the state processes are affine in controls taking values in unbounded sets.

**Assumption 2.6.** (i) The random variable $X_0$ satisfies $\mathbb{E}|X_0|^2 < \infty$.

(ii) The set $U$ is given by $U = \mathbb{R}^k$.

(iii) The functions $b$, $\sigma$, $f$, and $g$ are given respectively by

$$b(t,x,u) = b_0(t)x + b_1(t)u, \quad \sigma(t,x,u) = \sigma_0(t),$$
$$f(t,x,u) = x^{\mathsf{T}}P(t)x, \quad g(x) = x^{\mathsf{T}}Rx,$$

for $t\in[0,T]$, $x\in\mathbb{R}^d$, and $u\in\mathbb{R}^k$, where $b_0$ is $\mathbb{R}^{d\times d}$-valued, $b_1$ is $\mathbb{R}^{d\times k}$-valued, $\sigma_0$ is $\mathbb{R}^{d\times m}$, and $P$ is $\mathbb{R}^{d\times d}$-valued, all of which are continuous on $[0,T]$, and $R\in\mathbb{R}^{d\times d}$. Further $R$ and $P(t)$ are positive semidefinite for any $t\in[0,T]$.

**Theorem 2.7.** *Let Assumption* 2.6 *hold. Suppose that* $\{u^*(t)\}_{0\leqslant t\leqslant T}\in\mathcal{U}$ *is optimal both for the problem* $(\mathcal{C}_{\theta_1})$ *and* $(\mathcal{C}_{\theta_2})$ *for some* $\theta_1,\theta_2>0$. *Moreover, suppose that* $\mathbb{P}(u^*(t_0)\in \mathrm{int}(U)\backslash\{0\})>0$ *for some* $t_0\in[0,T]$. *Then* $\theta_1=\theta_2$.

*Proof.* For each $\theta=\theta_1$ and $\theta_2$, an optimal control for $(\mathcal{C}_\theta)$ uniquely exists and we necessarily have

$$u^*(t) = -\frac{1}{\theta}b_1(t)^{\mathsf{T}}F(t)X^*(t), \quad 0\leqslant t\leqslant T, \tag{2.9}$$

where $\{F(t)\}_{0\leqslant t\leqslant T}$ is a unique solution of the matrix Riccati equation

$$\frac{d}{dt}F(t) + b_0^{\mathsf{T}}(t)F(t) + F(t)b_0(t) - \frac{1}{\theta}F(t)b_1(t)b_1(t)^{\mathsf{T}}F(t) + P(t) = 0, \quad F(T)=R. \tag{2.10}$$

We refer to, e.g., Bensoussan [3] for this result. A simple application of Itô formula then yields

$$dF(t)X^*(t) = -(b_0(t)^{\mathsf{T}}F(t)+P(t))X^*(t)dt + F(t)\sigma_0(t)dW(t).$$

Let $F_1$ and $F_2$ be the solution of the Riccati equation (2.10) corresponding to $\theta_1$ and $\theta_2$, respectively. Then we have

$$d(F_1(t)-F_2(t))X^*(t) = -b_0(t)^{\mathsf{T}}X^*(t)dt + (F_1(t)-F_2(t))\sigma_0(t)dW(t).$$

From $F_1(T) = F_2(T)$ it follows that

$$(F_1(t) - F_2(t))X^*(t) = \mathbb{E}\left[\int_t^T b_0(s)^{\mathsf{T}}(F_1(s) - F_2(s))X^*(s)ds\middle|\mathcal{F}(t)\right],$$

whence

$$\mathbb{E}|(F_1(t) - F_2(t))X^*(t)|^2 \leqslant C\int_t^T \mathbb{E}|(F_1(s) - F_2(s))X^*(s)|^2 ds$$

for some positive constant $C > 0$. Therefore we have $F_1 = F_2$ and so $\theta_1 = \theta_2$. Thus the theorem follows. □

**Theorem 2.8.** *Let Assumption* 2.6 *hold. Suppose that* $\{u^*(t)\}_{0\leqslant t\leqslant T} \in \mathcal{U}$ *and* $\{u_n(t)\}_{0\leqslant t\leqslant T} \in \mathcal{U}$ *are optimal for the problems* $(\mathcal{C}_{\theta^*})$ *and* $(\mathcal{C}_{\theta_n})$ *for some* $\theta^*, \theta_n > 0$*, respectively,* $n \in \mathbb{N}$*, such that*

$$\lim_{n\to\infty} \mathbb{E}\int_0^T |u_n(t) - u^*(t)|^2 dt = 0.$$

*Moreover, suppose that there exists a measurable set* $E \subset [0,T]\times\Omega$ *with positive measure such that*

$$\lim_{n\to\infty} u_n(t,\omega) = u^*(t,\omega) \in \mathrm{int}(U)\backslash\{0\}, \quad (t,\omega) \in E.$$

*Then,* $\lim_{n\to\infty}\theta_n = \theta^*$.

*Proof.* Let $X^*(t)$ and $X^*(t)$ be the state processes corresponding to $u^*$ and $u_n$, respectively. Further, let $F^*$ and $F_n$ be the solution of the Riccati equation (2.10) corresponding to $\theta^*$ and $\theta_n$, respectively, $n \in \mathbb{N}$. Then, as in the proof of Theorem 2.7,

$$\begin{aligned}&F^*(t)X^*(t) - F_n(t)X_n(t)\\ &= \mathbb{E}\left[R(X^*(T) - X_n(T)) + \int_t^T b_0(s)^{\mathsf{T}}(F^*(s)X^*(s) - F_n(s)X_n(s))ds\middle|\mathcal{F}(t)\right],\end{aligned}$$

whence by Gronwall inequality,

$$\sup_{0\leqslant t\leqslant T}\mathbb{E}|F^*(t)X^*(t) - F_n(t)X_n(t)|^2 \leqslant C\mathbb{E}|X^*(T) - X_n(T)|^2 \to 0, \quad n\to\infty.$$

Thus,

$$\begin{aligned}|\theta_n - \theta^*|\int_E u_n dt\times d\mathbb{P} \leqslant |\theta^*|\mathbb{E}\int_0^T |u_n(t) - u^*(t)|dt\\ + \mathbb{E}\int_0^T |F^*(t)X^*(t) - F_n(t)X_n(t)|dt \to 0, \quad n\to\infty.\end{aligned}$$

Consequently, $\theta_n \to \theta^*$, as required. □

## 3 Numerical method

Here we propose a method for determining the weight parameter $\theta$ given observed data of optimal controls. To this end, recall that the value function $V$ for the problem $(\mathcal{C}_\theta)$ is given by

$$V(t,x;\theta) = \inf_{u\in\mathcal{U}_t} \mathbb{E}\left[ g(X(T)) + \int_t^T (f(s,X(s)) + \theta|u(s)|^2)ds \middle| X(t) = x \right], \tag{3.1}$$

for $(t,x) \in [0,T]\times\mathbb{R}^d$, where $\mathcal{U}_t$ is the set of all $U$-valued $\{\mathcal{F}(s)\}$-adapted processes $\{u(s)\}_{t\leqslant s\leqslant T}$ satisfying

$$\mathbb{E}\int_t^T |u(s)|^2 ds < \infty.$$

It is well-known that under Assumption 2.2 the value fuction $V(t,x) \equiv V(t,x;\theta)$ is a unique continuous viscosity solution of the Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{cases} \partial_t V(t,x) + H(t,x,D_xV(t,x),D_x^2V(t,x)) = 0, \quad (t,x)\in[0,T)\times\mathbb{R}^d, \\ V(T,x) = g(x), \quad x\in\mathbb{R}^d, \end{cases} \tag{3.2}$$

where

$$H(t,x,p,M) = \inf_{u\in U}\left[ b(t,x,u)^\mathsf{T} p + \frac{1}{2}\mathrm{tr}(\sigma(t,x,u)\sigma(t,x,u)^\mathsf{T} M) + f(t,x,u)\right]$$

for $(t,x,p,M)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{S}^d$ (see, e.g., [11]). Under Assumption 2.6, i.e., the case of the linear quadratic regulator problems, the value function $V$ is given by

$$V(t,x) = x^\mathsf{T} F(t)x + G(t), \quad (t,x)\in[0,T]\times\mathbb{R}^d, \tag{3.3}$$

where $F$ is as in (2.10) and $G$ is the unique solution of

$$\frac{d}{dt}G(t) + \mathrm{tr}(\sigma_0(t)^\mathsf{T} F(t)\sigma_0(t)) = 0, \quad G(T) = 0.$$

Then we have the following basic result:

**Theorem 3.1.** *Let $\{u^*(t)\}_{0\leqslant t\leqslant T}\in\mathcal{U}$ solves $(\mathcal{C}_{\theta^*})$ for some $\theta^*>0$, and $\{X^*(t)\}_{0\leqslant t\leqslant T}$ the corresponding state trajectory. Suppose that Assumption 2.4 holds and $\mathbb{P}(u^*(t_0)\in \mathrm{int}(U)\backslash\{0\}) > 0$ for some $t_0\in[0,T]$. Then, $\theta^*$ is a unique positive root of the equation*

$$\Phi(\theta) := \mathbb{E}\left[ g(X^*(T)) + \int_0^T (f(t,X^*(t)) + \theta|u^*(t)|^2)dt - V(0,X_0;\theta)\right] = 0. \tag{3.4}$$

*Proof.* Since $X_0$ is a constant and $\{u^*(t)\}$ is optimal, clearly we have $V(0,X_0;\theta^*) = \inf_{u\in\mathcal{U}} J[u] = J[u^*]$, whence $\Phi(\theta^*) = 0$. If $\theta' > 0$ satisfies $\Phi(\theta') = 0$, then $\{u^*(t)\}$ is also optimal for $(\mathcal{C}_{\theta'})$. By Theorem 2.3, we obtain $\theta' = \theta^*$. □

We have the same result in the case of the linear quadratic regulator problems.

**Theorem 3.2.** *Let $\{u^*(t)\}_{0\leqslant t\leqslant T} \in \mathcal{U}$ solves $(\mathcal{C}_{\theta^*})$ for some $\theta^* > 0$, and $\{X^*(t)\}_{0\leqslant t\leqslant T}$ the corresponding state trajectory. Suppose that Assumption 2.6 holds and $\mathbb{P}(u^*(t_0) \in \mathrm{int}(U)\backslash\{0\}) > 0$ for some $t_0 \in [0,T]$. Then, $\theta^*$ is a unique positive root of the equation $\Phi(\theta) = 0$, defined as in (3.2).*

*Proof.* Since $\{u^*(t)\}$ is necessarily given by (2.9), it is clear that the process $\{u^*(t)|_{X_0=x}\}$ is optimal for any $x \in \mathbb{R}^d$. Moreover, the process $\{X^*(t)|_{X_0=x}\}$ is the state process corresponding to $\{u^*(t)|_{X_0=x}\}$. Thus,

$$\begin{aligned} J[u^*] &= \mathbb{E}\left[X^*(T)^\mathsf{T} R X^*(T) + \int_0^T \left(X^*(t)^\mathsf{T} P(t) X^*(t) + \theta^* |u^*(t)|^2\right) dt\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X^*(T)^\mathsf{T} R X^*(T) + \int_0^T \left(X^*(t)^\mathsf{T} P(t) X^*(t) + \theta^* |u^*(t)|^2\right) dt \middle| X_0\right]\right] \\ &= \mathbb{E}[V(0, X_0; \theta^*)]. \end{aligned}$$

The rest of the proof can be done by the same way as in the proof of Theorem 3.1. □

Now, suppose that the $N$ independent samples $\{u^{(j)}(t_i), X^{(j)}(t_i)\}$, $i = 0, \dots, n$, $j = 1, \dots, N$, of optimal control process for $\mathcal{P}(\theta^*)$ at time $t_i$ and the corresponding state at time $t_i$ are available, where $0 = t_0 < \cdots < t_n = T$. An inverse control problem is then to determine the unknown $\theta^*$ from the observations. Our approach is to focus on the following problem:

**Problem** $(\mathcal{I})$**.** Find a positive root of the equation

$$\frac{1}{N}\sum_{j=1}^{N}\left\{g(X^{(j)}(T)) + \sum_{i=0}^{n-1}(f(t_i, X^{(j)}(t_i)) + \theta|u^{(j)}(t_i)|^2)(t_{i+1} - t_i) - V(0, X^{(j)}(0); \theta)\right\} = 0.$$

In views of the strong law of large number and Theorems 3.1 and 3.2, we can obtain an approximate solution of the inverse problem by solving the problem $(\mathcal{I})$ for sufficiently large $N$ and $n$.

**Example 3.3.** Consider the case where the state is described by the one-dimensional equation

$$dX(t) = u(t)dt + \frac{1}{10}dW(t)$$

with initial condition $X_0$ having the standard normal distribution, and the control objective $J[u]$ is given by

$$J[u] = \mathbb{E}\int_0^1 \left(10|X(t)|^2 + \theta|u(t)|^2\right) dt,$$

where $U = \mathbb{R}$. This problem is of course a particular case of the linear quadratic regulator ones. The value function $V$ is explicitly given by $V(t, x; \theta) = F(t; \theta)x^2 + G(t; \theta)$, where

$F(t;\theta) = \sqrt{10\theta}\tanh((1-t)\sqrt{10/\theta})$ and $G(t;\theta) = (\theta/100)\log(\cosh((1-t)\sqrt{10/\theta}))$. The unique optimal control $u^*(t)$ is given by

$$u^*(t) = -\sqrt{\frac{10}{\theta}}\tanh\left(\sqrt{\frac{10}{\theta}}(1-t)\right)X^*(t).$$

To test our approach, we independently generate the samples $(X^{(j)}(t_i), u^{(j)}(t_i))$, $j = 1,\dots,10000$, $i = 0,\dots,1000$, of the optimal state and control for $\theta = 1$, where $t_i = i/1000$, and consider these samples as observed data. We solve the root-finding problem by minimizing

$$\frac{1}{10000}\left|\sum_{j=1}^{10000}\sum_{i=0}^{999}\left(10|X^{(j)}(t_i)|^2 + \theta|u^{(j)}(t_i)|^2\right)\Delta t - V(0, X_0^{(j)};\theta)\right|$$

over $\theta \in [0.0001, 10]$. The estimated $\theta$ over 100 trials has the mean 0.9971 and the standard deviation $4.4181 \times 10^{-4}$.

In most of nonlinear problems, analytical solutions of HJB equations are rarely available. So we need to numerically solve (3.4) to approximate the value functions. Existing numerical methods applicable to (3.4) include the finite difference methods (see, e.g., Kushner and Dupuis [15] and Bonnans and Zidani [4]), the finite-element like methods (see, e.g., Camilli and Falcone [5] and Debrabant and Jakobsen [7]), the kernel-based collocation methods (see, e.g., Kansa [13, 14] and Nakano [17]), and the regression-based methods (see, e.g., Fahim et al. [10], E et al. [9], Sirignano and Spiliopoulos [20], and the references therein). It is well-known that the use of the finite difference methods is limited to low dimensional problems since the number of the spatial grids points has an exponential growth in dimension. In the kernel-based collocation methods, we seek an approximate solution of the form of a linear combination of a radial basis function (e.g., multiquadrics in the Kansa's original work) by solving finite dimensional linear equations. In general, this procedure allows for a simpler numerical implementation compared to the finite-element like methods and the regression-based methods. The regression-based methods, in particular the ones with neural networks, are prominent in high dimensional problems although they are computationally expensive.

In Example 3.4 below, we will deal with some population control problem having a 3-dimensional state space. We adopt the kernel-based collocation methods to compute the value function for that problem since they are useful in multi but relatively low dimensional problems.

The kernel-based collocation methods are obtained by the interpolations with positive definite kernels applied backward recursively in time. Given a points set $\Gamma = \{x^{(1)},\dots,x^{(N)}\} \subset \mathbb{R}^d$ such that $x^{(j)}$'s are pairwise distinct, and a positive definite function $\Phi : \mathbb{R}^d \to \mathbb{R}$, the function

$$\sum_{j=1}^{N}(A^{-1}\psi|_\Gamma)_j\Phi(x - x^{(j)}), \quad x \in \mathbb{R}^d,$$

interpolates $\psi$ on $\Gamma$. Here, $A = \{\Phi(x^{(j)} - x^{(\ell)})\}_{j,\ell=1,\dots,N}$, $\psi|_\Gamma$ is the column vector composed of $\psi(x_j)$, $j = 1, \dots, N$, and $(z)_j$ denotes the $j$-th component of $z \in \mathbb{R}^N$. Thus, with time grid $\{t_0, \dots, t_n\}$ such that $0 = t_0 < \cdots < t_n = T$, the function $\tilde{V}(t_n, \cdot)$ defined by

$$\tilde{V}(t_n, x) = \sum_{j=1}^{N} (A^{-1}\tilde{V}_n)_j \Phi(x - x^{(j)}), \quad x \in \mathbb{R}^d$$

approximates $g$, where $\tilde{V}_n = g|_\Gamma$. Then, for any $k = 0, 1, \dots, n-1$, with a vector $\tilde{V}_{k+1} \in \mathbb{R}^N$ of an approximate solution at $\{t_{k+1}\} \times \Gamma$, we set

$$\tilde{V}(t_{k+1}, x) = \sum_{j=1}^{N} (A^{-1}\tilde{V}_{k+1})_j \Phi(x - x^{(j)}), \quad x \in \mathbb{R}^d,$$
$$\tilde{V}_k = \tilde{V}_{k+1} + (t_{k+1} - t_k) H_{k+1}(v^h_{k+1}),$$

where $H_{k+1}(\tilde{V}_{k+1}) = (H_{k+1,1}(\tilde{V}_{k+1}), \dots, H_{k+1,N}(\tilde{V}_{k+1})) \in \mathbb{R}^N$ with

$$H_{k+1,j} = H(t_{k+1}, x^{(j)}, D\tilde{V}(t_{k+1}, x^{(j)}), D^2\tilde{V}(t_{k+1}, x^{(j)})).$$

The method above can be described in a matrix form. See Nakano [16] for details.

**Example 3.4.** Consider the following simple SIR (Susceptible-Infectious-Recovery) epidemic model with vaccination studied in Rachah and Torres [19]:

$$\frac{dS(t)}{dt} = -\beta S(t)I(t) - u(t)S(t),$$
$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \mu I(t),$$
$$\frac{dR(t)}{dt} = \mu I(t) + u(t)S(t),$$

with initial conditions $S(0) = 0.95$, $I(0) = 0.05$, $R(0) = 0$, where $\beta = 0.2$ and $\mu = 0.1$. The control objective is given by

$$J[u] = \int_0^{10} (10I(t) + \theta|u(t)|^2)dt.$$

We assume that $\theta = 1$ is a true parameter and solve the corresponding control problem by the kernel-based collocation method to obtain an approximate value function $\tilde{V}$, nearly optimal trajectories $S^*(t_i)$, $I^*(t_i)$, $R^*(t_i)$, and a nearly optimal control $u^*(t_i)$, where $t_i = i/1000$, $i = 0, 1, \dots, 10000$. Specifically, the kernel-based method is performed with the following ingredients: the Gaussian kernel with $\alpha = 1$ and the collocation points $\Gamma$ consisting of $8^3$ uniform grids points in $[S(0) - 0.5, S(0) + 0.5] \times [I(0) - 0.5, I(0) + 0.5] \times [R(0) - 0.5, R(0) + 0.5]$ including the boundary. With this method, we generate 100 independent copies $X^{(j)}(t_i)$, $j = 1, \dots, 100$, of $(S^*(t_i), I^*(t_i), R^*(t_i)) + 0.01 \times Z_i$, $i = 0, 1, \dots, n$, where $Z = (Z_0, \dots, Z_n)$ is an $(n+1) \times 3$-dimensional Gaussian vector with all components being independent of each other and having zero mean and unit

variance. Then we consider these samples with noises as observed data. As in Example 3.3, we solve the root-finding problem $(\mathcal{I})$ by minimizing

$$\frac{1}{100}\left|\sum_{j=1}^{100}\sum_{i=0}^{9999}\left(10|X^{(j)}(t_i)|^2+\theta|u^{(j)}(t_i)|^2\right)\Delta t-\tilde{V}(0,X_0^{(j)};\theta)\right|$$

subject to $\theta\in[0.0001,10]$. The estimated $\theta$ over 100 trials has the mean 0.9496 and the standard deviation 0.0558.

# 4 Conclusion

We study the inverse problem of the stochastic control of general diffusions with performance index

$$g(X(T))+\int_0^T\left(f(t,X(t))+\theta|u(t)|^2\right)dt.$$

Precisely, we formulate the inverse problem as the one of determining the weight parameter $\theta>0$ given the dynamics, $g$, $f$, and an optimal control process. Under mild conditions on the drift, the volatility, $g$, and $f$, and under the assumption that the optimal control belongs to the interior of the control set, we show that our inverse problem is well-posed. It should be noted that whether the latter assumption holds true or not is easily confirmed in practice. Then, with the well-posedness, we reduce the inverse problem to some root finding problem of the expectation of a random variable involved with the value function, which has a unique solution. Based on this result, we propose a numerical method for our inverse problem by replacing the expectation above with arithmetic mean of observed optimal control processes and the corresponding state processes. Several numerical experiments show that the numerical method recover the unknown $\theta$ with high accuracy. In particular, with the help of the kernel-based collocation method for Hamilton-Jacobi-Bellman equations, our method for the inverse problems still works well even when an explicit form of the value function is unavailable.

Possible future studies include the well-posedness when the function $g$ or $f$ is also unknown, and numerical methods under non-uniqueness of the inverse problems. In the latter issue, we may need to use an additional criterion to determine unknowns from multiple candidates such as the maximum entropy principle adopted in [23].

### Acknowledgements

This study is supported by JSPS KAKENHI Grant Number JP17K05359.

# References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, 2004.

[2] R. Bellman and R. Kalaba. An inverse problem in dynamic programming and automatic control. *J. Math. Anal. Appl.*, 7:322–325, 1963.

[3] A. Bensoussan. *Stochastic control of partially observable systems*. Cambridge University Press, Cambridge, 1992.

[4] F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic HJB equation. *SIAM J. Numer. Anal.*, 41:1008–1021, 2003.

[5] F. Camilli and M. Falcone. An approximation scheme for the optimal control of diffusion processes. *Math. Model. Numer. Anal.*, 29:97–122, 1995.

[6] J. Casti. On the general inverse problem of optimal control theory. *J. Optim. Theory Appl.*, 32:491–497, 1980.

[7] K. Debrabant and E. R. Jakobsen. Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. *Math. Comp.*, 82:1433–1462, 2013.

[8] K. Dvijotham and E. Todorov. Inverse optimal control with linearly-solvable MDPs. In *Proc. ICML*, 2010.

[9] W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat.*, 5:349–380, 2017.

[10] A. Fahim, N. Touzi, and X. Warin. A probabilistic numerical method for fully nonlinear parabolic PDEs. *Ann. Appl. Probab.*, 21:1322–1364, 2011.

[11] W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*. Springer-Verlag, New York, 2nd edition, 2006.

[12] R. E. Kalman. When is a linear control system optimal? *Trans. ASME Ser. D: J. Basic Eng.*, 86:51–60, 1964.

[13] E. J. Kansa. Multiquadricsa scattered data approximation scheme with application to computational fluid-dynamicsI. *Computers Math. Applic.*, 19:127–145, 1990.

[14] E. J. Kansa. Multiquadricsa scattered data approximation scheme with application to computational fluid-dynamicsII. *Computers Math. Applic.*, 19:147–161, 1990.

[15] H. J. Kushner and P. Dupuis. *Numerical methods for stochastic control problems in continuous time*. Springer-Verlag, New York, 2001.

[16] Y. Nakano. Convergent kernel-based methods for parabolic equations. arXiv:1803.09446[Math.NA].

[17] Y. Nakano. Convergence of meshfree collocation methods for fully nonlinear parabolic equations. *Numer. Math.*, 136:703–723, 2017.

[18] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proc. ICML*, 2000.

[19] A. Rachah and D. F. Torres. Mathematical modelling, simulation, and optimal control of the 2014 ebola outbreak in west africa. *Discrete Dyn. Nat. Soc.*, 2015, 2015.

[20] J. Sirignano and K. Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *J. Comput. Phys.*, 375:1339–1364, 2018.

[21] F. Thau. On the inverse optimum control problem for a class of nonlinear autonomous systems. *IEEE Trans. Automat. Control*, 12:674–681, 1967.

[22] J. Yong and X. Y. Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*. Springer-Verlag, New York, 1999.

[23] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. D. Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438, 2008.