Stochastic Game Theoretic Trajectory Optimization In Continuous Time

Wei Sun Evangelos A. Theodorou Panagiotis Tsiotras

Abstract— **A Stochastic Game Theoretic Differential Dynamic Programming (SGT-DDP) algorithm is derived to solve a differential game under stochastic dynamics. We present the update law for the minimizing and maximizing controls for both players and provide a set of backward differential equations for the second order value function approximation. We compute the extra terms in the backward propagation equations that arise from the stochastic assumption compared with the original GT-DDP. We present the SGT-DDP algorithm and analyze how the design of the cost function affects the feed-forward and feedback parts of the control policies under the game theoretic formulation. The performance of SGT-DDP is then investigated through simulations on two examples, namely, a first order nonlinear system, the inverted pendulum and the cart pole problems with conflicting controls. We conclude with some possible future extensions.**

I. Introduction

Over the recent years, autonomy has become one of the most active areas of research, with many applications in the areas of robotics, automotive and aerospace systems. From the different computational frameworks used to achieve autonomy in engineered systems, stochastic trajectory optimization plays a key role since it provides a framework for computing the best possible action in the presence of exogenous stochastic disturbances. While there has been an extensive amount of work on stochastic and deterministic trajectory optimization, most of the prior work in this area has been on discrete time representations. In cases where the initial problem formulation is in continuous time, the standard approach is to discretize the problem at hand and then perform optimization in discrete time.

In this work we derive a method for stochastic trajectory optimization using the framework of Differential Dynamic Programming (DDP) [1]. DDP is one of the most well-known trajectory optimization methods that iteratively finds a local optimal control policy starting from a nominal control and state trajectory. There has been a plethora of variations and applications of DDP within the controls and robotics communities. Starting with a differential game theoretic formulation and its application on bipedal locomotion [2] to receding horizon [3], and stochastic control formulations [4], [5], DDP has become one of the standard methods for trajectory optimization with a broad range of applications [4], [6]–[13].

In this paper we approach the problem of stochastic trajectory optimization in continuous time from a game-theoretic point of view, and present an algorithm that relies on first order expansion of the dynamics and second order expansion of the value function. In particular, we derive the equations for the backward propagation of the value function for the case of stochastic differential games. The resulting algorithm has the attractive characteristics of DDP in terms of scalability and numerical efficiency, while it also features robustness to deterministic and stochastic disturbances due to stochastic min-max formulation. The contribution of this paper is twofold:

i) For the first time in the literature, we derive a stochastic game theoretic Differential Dynamic Programming algorithm in continuous time and provide a set of backward ordinary differential equations for the zeroth, first and second order terms in the approximation of the value function along a nominal trajectory.

ii) The resulting algorithm is a generalization of our previous work on Game Theoretic DDP (GT-DDP) [14] to a stochastic settings. Applications of the proposed algorithm include robust stochastic trajectory optimization and stochastic control under non-zero mean stochastic disturbances.

The motivation of this work comes from the fact that there is a fundamental connection between min-max extensions of optimal control and risk-sensitive stochastic control formulations [15]. This relationship was first investigated by Jacobson in [16]. The work in [12] also investigated risk-sensitive stochastic control in an LQG setting while the work in [13] addressed risk-sensitive control for nonlinear stochastic systems and infinite horizon control tasks. In a risk-sensitive setting, the control objective is to minimize a performance index, which is expressed as a function of the mean and variance of a given state- and control-dependent cost. Therefore, the element of risk sensitivity arises from the minimization of the variance of that cost. Thus, risk-sensitive optimal control problems are directly related to stochastic differential games [17] considered in this paper.

II. Problem Formulation

We consider the problem of a differential game between two players

$$V(\mathbf{x}(t_0), t_0) = \min_{\mathbf{u}} \max_{\mathbf{v}} J(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

W. Sun is a Ph.D. candidate at the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta. GA 30332-0150, USA. Email: wsun42@gatech.edu

E. Theodorou is an Assistant Professor at the School of Aerospace Engineering and the Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta. GA 30332-0150, USA. Email: evangelos.theodorou@ae.gatech.edu

P. Tsiotras is a Professor at the School of Aerospace Engineering and the Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta. GA 30332-0150, USA. Email: tsiotras@gatech.edu

$$= \min_{\mathbf{u}} \max_{\mathbf{v}} \mathbb{E}\left[\phi(\mathbf{x}(t_f)) + \int_{t_0}^{t_f} \mathcal{L}(\mathbf{x},\mathbf{u},\mathbf{v})\mathrm{d}t\right], \quad (1)$$

subject to the stochastic dynamics

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x},\mathbf{u},\mathbf{v},t)\mathrm{d}t + \mathbf{G}(\mathbf{x})\mathrm{d}w, \\ \mathbf{x}(t_0) &= \mathbf{x}_0, \end{aligned} \quad (2)$$

where V stands for the value function (expected cost-to-go), the term J represents the performance index, and $\mathbf{x} \in \mathbb{R}^n$ represents the state of the dynamical system. The term $\mathbf{u} \in \mathbb{R}^p$ stands for the input of the minimizing player, whose objective is to minimize the performance index. Similarly, $\mathbf{v} \in \mathbb{R}^q$ represents the input of the maximizing player, which tries to maximize the performance index. The function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \mapsto \mathbb{R}$ is the *running cost* and $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ is the *terminal cost*, where the terminal time t_f is a prescribed constant. The term $\mathrm{d}w$ represents an increment of a m-dimensional Wiener process (standard Brownian motion), and $\mathbf{G} : \mathbb{R}^n \mapsto \mathbb{R}^{n\times m}$ is introduced to scale $\mathrm{d}w$ and match the dimension of $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \mapsto \mathbb{R}^n$. It is assumed that $\mathrm{d}w \sim \mathcal{N}(0, I_{m\times m}\mathrm{d}t)$.

Denote by $\mathcal{U}$ the admissible feedback control set of the minimizing player, that is, $\mathcal{U} = \{\mathbf{u} : [t_0, t_f] \times \mathbb{R}^n \mapsto \mathbb{R}^p, \mathbf{u}(\tau, \cdot)$ is $\mathcal{F}_\tau$-measurable, $\forall \tau \in [t, t_f]$, and $\mathbf{u}(\cdot, \mathbf{x})$ is Lebesgue measurable, $\forall \mathbf{x} \in \mathbb{R}^n\}$. Similarly, the admissible feedback control set of the maximizing player is given by $\mathcal{V} = \{\mathbf{v} : [t_0, t_f] \times \mathbb{R}^n \mapsto \mathbb{R}^q, \mathbf{v}(\tau, \cdot)$ is $\mathcal{F}_\tau$-measurable, $\forall \tau \in [t, t_f]$, and $\mathbf{v}(\cdot, \mathbf{x})$ is Lebesgue measurable, $\forall \mathbf{x} \in \mathbb{R}^n\}$. Here $\mathcal{F}_t$ denotes the corresponding filtration with respect to the Brownian motion, which can be interpreted as representing all historical information available up to time t about the stochastic process.

We assume in this paper that the value of the game exists, that is,

$$V = \min_{\mathbf{u}} \max_{\mathbf{v}} J(\mathbf{x},\mathbf{u},\mathbf{v}) = \max_{\mathbf{v}} \min_{\mathbf{u}} J(\mathbf{x},\mathbf{u},\mathbf{v}). \quad (3)$$

Next, we derive the stochastic min-max DDP framework.

III. Optimal Control Variations

Given a nominal mean trajectory of the state and initial controls $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$, and letting $\delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$, $\delta\mathbf{u} = \mathbf{u} - \bar{\mathbf{u}}$, $\delta\mathbf{v} = \mathbf{v} - \bar{\mathbf{v}}$, from

$$\begin{aligned} \mathrm{d}(\bar{\mathbf{x}} + \delta\mathbf{x}) &= \mathbf{f}(\bar{\mathbf{x}} + \delta\mathbf{x}, \bar{\mathbf{u}} + \delta\mathbf{u}, \bar{\mathbf{v}} + \delta\mathbf{v})\mathrm{d}t + \mathbf{G}(\bar{\mathbf{x}} + \delta\mathbf{x})\mathrm{d}w \\ &\approx (\mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}}) + \nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x} + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u} + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v})\mathrm{d}t \\ &\quad + (\mathbf{G}(\bar{\mathbf{x}}) + \mathbf{G}_{\mathbf{x}}\delta\mathbf{x})\mathrm{d}w, \quad (4) \\ \mathrm{d}\bar{\mathbf{x}} &= \mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})\mathrm{d}t, \quad (5) \end{aligned}$$

we obtain

$$\begin{aligned} \mathrm{d}\delta\mathbf{x} &= (\nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x} + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u} + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v})\mathrm{d}t \\ &\quad + (\mathbf{G}(\bar{\mathbf{x}}) + \mathbf{G}_{\mathbf{x}}(\delta\mathbf{x}))\mathrm{d}w, \end{aligned} \quad (6)$$

where $\mathbf{G}_{\mathbf{x}}(\delta\mathbf{x}) = [\nabla_{\mathbf{x}}\mathbf{G}^{(1)}\delta\mathbf{x}, \ldots, \nabla_{\mathbf{x}}\mathbf{G}^{(m)}\delta\mathbf{x}]$ and $\mathbf{G}^{(j)}$ denotes the j-th column vector of $\mathbf{G}$, $j = 1, \ldots, m$. The arguments of the functions in the previous derivation are omitted when they are evaluated along the nominal trajectory $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$.

In order to derive the update law for the minimizing and maximizing controls, we start our analysis with Bellman/Isaac's principle, which states

$$V(\mathbf{x}_t, t) = \\ \min_{\mathbf{u}} \max_{\mathbf{v}} \mathbb{E}\left[\int_t^{t+\mathrm{d}t} \mathcal{L}(\mathbf{x},\mathbf{u},\mathbf{v})\mathrm{d}t + V(\mathbf{x}_{t+\mathrm{d}t}, t+\mathrm{d}t)\middle|\mathbf{x}_t\right], \quad (7)$$

where the subscript t and $t + \mathrm{d}t$ are introduced to denote the evaluation of the variables at time t and $t + \mathrm{d}t$, respectively.

The main idea is to take expansions of the terms in both sides of equation (7) around the nominal state and control trajectories $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ to find the update equations for the minimizing control, maximizing control and backward differential equations for the zeroth, first and second order approximation terms of the value function. Starting with the left-hand side of (7), the second order expansion of the cost-to-go function around a nominal trajectory $\bar{\mathbf{x}}$ is obtained as follows

$$\begin{aligned} V(\mathbf{x}_t, t) &= V(\mathbf{x}_t + \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_t, t) = V(\bar{\mathbf{x}}_t + \delta\mathbf{x}_t, t) \\ &\approx V_t + \nabla_{\mathbf{x}}V_t\delta\mathbf{x}_t + \frac{1}{2}\delta\mathbf{x}_t^{\mathsf{T}}\nabla_{\mathbf{x}\mathbf{x}}V_t\delta\mathbf{x}_t, \end{aligned}$$

where $V_t = V(\bar{\mathbf{x}}_t, t)$. As for the right-hand side of (7), the first term is approximated as follows

$$\begin{aligned} \mathbb{E}\left[\int_t^{t+\mathrm{d}t} \mathcal{L}(\mathbf{x},\mathbf{u},\mathbf{v})\mathrm{d}t\middle|\mathbf{x}_t\right] &\approx \mathcal{L}(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t)\mathrm{d}t \\ &= \mathcal{L}(\bar{\mathbf{x}}_t + \delta\mathbf{x}_t, \bar{\mathbf{u}}_t + \delta\mathbf{u}_t, \bar{\mathbf{v}}_t + \delta\mathbf{v}_t)\mathrm{d}t. \end{aligned} \quad (8)$$

This expression can be approximated as

$$\mathcal{L}\mathrm{d}t + (\nabla_{\mathbf{x}}\mathcal{L}\delta\mathbf{x} + \nabla_{\mathbf{u}}\mathcal{L}\delta\mathbf{u} + \nabla_{\mathbf{v}}\mathcal{L}\delta\mathbf{v})\mathrm{d}t \\ + \frac{1}{2}\begin{bmatrix} \delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathbf{v}_t \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \nabla_{\mathbf{x}\mathbf{x}}\mathcal{L} & \nabla_{\mathbf{x}\mathbf{u}}\mathcal{L} & \nabla_{\mathbf{x}\mathbf{v}}\mathcal{L} \\ \nabla_{\mathbf{u}\mathbf{x}}\mathcal{L} & \nabla_{\mathbf{u}\mathbf{u}}\mathcal{L} & \nabla_{\mathbf{u}\mathbf{v}}\mathcal{L} \\ \nabla_{\mathbf{v}\mathbf{x}}\mathcal{L} & \nabla_{\mathbf{v}\mathbf{u}}\mathcal{L} & \nabla_{\mathbf{v}\mathbf{v}}\mathcal{L} \end{bmatrix} \begin{bmatrix} \delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathbf{v}_t \end{bmatrix} \mathrm{d}t, \quad (9)$$

where the function $\mathcal{L}$ and its derivatives in the last equation are all evaluated at $(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t, \bar{\mathbf{v}}_t)$ and thus omitted for simplicity of notation. Henceforth, all the terms are evaluated at $(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t, \bar{\mathbf{v}}_t)$, unless specified otherwise.

Before we expand the term $\mathbb{E}\left[V(\mathbf{x}_{t+\mathrm{d}t}, t+\mathrm{d}t)\right]$ around $\bar{x}_{t+\mathrm{d}t}$ and make it compatible with the left-hand side of (7), we need to find an expression for $\delta\mathbf{x}_{t+\mathrm{d}t}$ in terms of $\delta\mathbf{x}_t$. Indeed, from (6), we get

$$\begin{aligned} \delta\mathbf{x}_{t+\mathrm{d}t} &= \delta\mathbf{x}_t + (\nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x}_t + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u}_t + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v}_t)\mathrm{d}t \\ &\quad + (\mathbf{G} + \mathbf{G}_{\mathbf{x}}(\delta\mathbf{x}_t))\mathrm{d}w, \end{aligned}$$

where $\mathbf{G}_{\mathbf{x}}(\delta\mathbf{x}_t) = [\nabla_{\mathbf{x}}\mathbf{G}^{(1)}\delta\mathbf{x}_t, \ldots, \nabla_{\mathbf{x}}\mathbf{G}^{(m)}\delta\mathbf{x}_t]$. Returning to the expansion of $\mathbb{E}[V(\mathbf{x}_{t+\mathrm{d}t}, t+\mathrm{d}t)|\mathbf{x}_t]$, and letting $V_{t+\mathrm{d}t} = V(\bar{\mathbf{x}}_{t+\mathrm{d}t}, t+\mathrm{d}t)$, we have

$$\mathbb{E}\left[V(\mathbf{x}_{t+\mathrm{d}t}, t+\mathrm{d}t)\middle|\mathbf{x}_t\right] = \mathbb{E}\left[V(\bar{\mathbf{x}}_{t+\mathrm{d}t} + \delta\mathbf{x}_{t+\mathrm{d}t}, t+\mathrm{d}t)\middle|\mathbf{x}_t\right]$$

By expanding the last term, we obtain

$$
\begin{aligned}
&\mathbb{E}\big[V_{t+\mathrm{d}t} + \nabla_{\mathbf{x}} V_{t+\mathrm{d}t}\delta\mathbf{x}_{t+\mathrm{d}t} + \frac{1}{2}\delta\mathbf{x}_{t+\mathrm{d}t}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\delta\mathbf{x}_{t+\mathrm{d}t}\big|\mathbf{x}_t\big] \\
&= V_{t+\mathrm{d}t} + \nabla_{\mathbf{x}}V_{t+\mathrm{d}t}\big[\delta\mathbf{x}_t + (\nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x}_t + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u}_t + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v}_t)\mathrm{d}t\big] \\
&+ \frac{1}{2}\big[\delta\mathbf{x}_t + (\nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x}_t + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u}_t + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v}_t)\mathrm{d}t\big]^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t} \\
&\cdot\big[\delta\mathbf{x}_t + (\nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x}_t + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u}_t + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v}_t)\mathrm{d}t\big] \\
&+ \frac{1}{2}\mathrm{tr}\big(\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}(\mathbf{G}+\mathbf{G}_{\mathbf{x}}(\delta\mathbf{x}_t))(\mathbf{G}+\mathbf{G}_{\mathbf{x}}(\delta\mathbf{x}_t))^{\mathsf{T}}\big)\mathrm{d}t, \qquad (10)
\end{aligned}
$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. In the previous derivation, we make use of the fact that $\mathrm{d}w \sim \mathcal{N}(0, I_{m\times m}\mathrm{d}t)$.

We proceed by taking expansions of all the terms. After combining (9) with (10) and grouping the terms with respect to $\delta\mathbf{x}_t$, $\delta\mathbf{u}_t$ and $\delta\mathbf{v}_t$, we can represent the right-hand side of (7) in a compact form, that is,

$$
\begin{aligned}
&\mathbb{E}\left[\int_t^{t+\mathrm{d}t}\mathcal{L}(\mathbf{x},\mathbf{u},\mathbf{v})\mathrm{d}t + V(\mathbf{x}_{t+\mathrm{d}t}, t+\mathrm{d}t)\middle|\mathbf{x}_t\right] \\
&= V_{t+\mathrm{d}t} + Q_0\mathrm{d}t + \nabla_{\mathbf{x}}V_{t+\mathrm{d}t}\delta\mathbf{x}_t \\
&+ \big(Q_{\mathbf{x}}\delta\mathbf{x}_t + Q_{\mathbf{u}}\delta\mathbf{u}_t + Q_{\mathbf{v}}\delta\mathbf{v}_t\big)\mathrm{d}t + \frac{1}{2}\delta\mathbf{x}_t^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\delta\mathbf{x}_t \\
&+ \frac{1}{2}\begin{bmatrix}\delta\mathbf{x}_t\\ \delta\mathbf{u}_t\\ \delta\mathbf{v}_t\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}Q_{\mathbf{xx}} & Q_{\mathbf{xu}} & Q_{\mathbf{xv}}\\ Q_{\mathbf{ux}} & Q_{\mathbf{uu}} & Q_{\mathbf{uv}}\\ Q_{\mathbf{vx}} & Q_{\mathbf{vu}} & Q_{\mathbf{vv}}\end{bmatrix}\begin{bmatrix}\delta\mathbf{x}_t\\ \delta\mathbf{u}_t\\ \delta\mathbf{v}_t\end{bmatrix}\mathrm{d}t, \quad (11)
\end{aligned}
$$

where

$$
\begin{aligned}
Q_0 &= \mathcal{L} + \frac{1}{2}\mathrm{tr}(\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\mathbf{G}\mathbf{G}^{\mathsf{T}}),\\
Q_{\mathbf{x}} &= \nabla_{\mathbf{x}}\mathcal{L} + \nabla_{\mathbf{x}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{f} + \sum_{j=1}^{m}\mathbf{G}^{(j)\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{G}^{(j)},\\
Q_{\mathbf{u}} &= \nabla_{\mathbf{u}}\mathcal{L} + \nabla_{\mathbf{x}}V_{t+\mathrm{d}t}\nabla_{\mathbf{u}}\mathbf{f},\\
Q_{\mathbf{v}} &= \nabla_{\mathbf{v}}\mathcal{L} + \nabla_{\mathbf{x}}V_{t+\mathrm{d}t}\nabla_{\mathbf{v}}\mathbf{f}
\end{aligned}
$$

and the second partials,

$$
\begin{aligned}
Q_{\mathbf{xx}} &= \nabla_{\mathbf{xx}}\mathcal{L} + \nabla_{\mathbf{x}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{f}\mathrm{d}t + 2\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{f}\\
&\quad + \sum_{j=1}^{m}\nabla_{\mathbf{x}}\mathbf{G}^{(j)\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{G}^{(j)},\\
Q_{\mathbf{uu}} &= \nabla_{\mathbf{uu}}\mathcal{L} + \nabla_{\mathbf{u}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{u}}\mathbf{f}\mathrm{d}t,\\
Q_{\mathbf{vv}} &= \nabla_{\mathbf{vv}}\mathcal{L} + \nabla_{\mathbf{v}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{v}}\mathbf{f}
\end{aligned}
$$

and the mixed partials,

$$
\begin{aligned}
Q_{\mathbf{ux}} &= \nabla_{\mathbf{ux}}\mathcal{L} + \nabla_{\mathbf{u}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t} + \nabla_{\mathbf{u}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{f}\mathrm{d}t,\\
Q_{\mathbf{vx}} &= \nabla_{\mathbf{vx}}\mathcal{L} + \nabla_{\mathbf{v}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t} + \nabla_{\mathbf{v}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{x}}\mathbf{f}\mathrm{d}t,\\
Q_{\mathbf{uv}} &= \nabla_{\mathbf{uv}}\mathcal{L} + \nabla_{\mathbf{u}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\nabla_{\mathbf{v}}\mathbf{f}\mathrm{d}t,\\
Q_{\mathbf{xu}} &= Q_{\mathbf{ux}}^{\mathsf{T}}, \qquad Q_{\mathbf{xv}} = Q_{\mathbf{vx}}^{\mathsf{T}}, \qquad Q_{\mathbf{vu}} = Q_{\mathbf{uv}}^{\mathsf{T}}.
\end{aligned}
$$

All the parameters in the previous expressions are henceforth denoted as the Q-functions. The reason we single out $V_{t+\mathrm{d}t}$, $\nabla_{\mathbf{x}}V_{t+\mathrm{d}t}\delta\mathbf{x}_t$ and $\frac{1}{2}\delta\mathbf{x}_t^{\mathsf{T}}\nabla_{\mathbf{xx}}V_{t+\mathrm{d}t}\delta\mathbf{x}_t$ instead of appending them in the Q-functions will become clear later on, as we derive the backward differential equations with respect to the value function and its derivatives.

In order to find the optimal control updates $\delta\mathbf{u}_t^*$ and $\delta\mathbf{v}_t^*$, we take the derivative of (11) with respect to $\delta\mathbf{u}_t$ and $\delta\mathbf{v}_t$, respectively, and set them equal to zero to obtain

$$
\begin{aligned}
\delta\mathbf{u}_t^* &= -Q_{\mathbf{uu}}^{-1}\big(Q_{\mathbf{ux}}\delta\mathbf{x} + Q_{\mathbf{uv}}\delta\mathbf{v}_t + Q_{\mathbf{u}}\big), \qquad (12)\\
\delta\mathbf{v}_t^* &= -Q_{\mathbf{vv}}^{-1}\big(Q_{\mathbf{vx}}\delta\mathbf{x} + Q_{\mathbf{vu}}\delta\mathbf{u}_t + Q_{\mathbf{v}}\big). \qquad (13)
\end{aligned}
$$

By replacing the $\delta\mathbf{v}_t$ term in (12) with (13) and solving for $\delta\mathbf{u}_t^*$, we can eliminate $\delta\mathbf{v}_t$ in the expression of $\delta\mathbf{u}_t^*$. We can solve for $\delta\mathbf{v}_t^*$ in a similar manner and obtain

$$
\delta\mathbf{u}_t^* = \mathbf{l}_{\mathbf{u}} + \mathbf{L}_{\mathbf{u}}\delta\mathbf{x} \quad \text{and} \quad \delta\mathbf{v}_t^* = \mathbf{l}_{\mathbf{v}} + \mathbf{L}_{\mathbf{v}}\delta\mathbf{x}, \qquad (14)
$$

with the feed-forward gains $\mathbf{l}_{\mathbf{v}}, \mathbf{l}_{\mathbf{u}}$ and feedback gains $\mathbf{L}_{\mathbf{v}}, \mathbf{L}_{\mathbf{u}}$ defined as:

$$
\begin{aligned}
\mathbf{l}_{\mathbf{u}} &= -\big(Q_{\mathbf{uu}} - Q_{\mathbf{uv}}Q_{\mathbf{vv}}^{-1}Q_{\mathbf{vu}}\big)^{-1}\big(Q_{\mathbf{u}} - Q_{\mathbf{uv}}Q_{\mathbf{vv}}^{-1}Q_{\mathbf{v}}\big),\\
\mathbf{L}_{\mathbf{u}} &= -\big(Q_{\mathbf{uu}} - Q_{\mathbf{uv}}Q_{\mathbf{vv}}^{-1}Q_{\mathbf{vu}}\big)^{-1}\big(Q_{\mathbf{ux}} - Q_{\mathbf{uv}}Q_{\mathbf{vv}}^{-1}Q_{\mathbf{vx}}\big),\\
\mathbf{l}_{\mathbf{v}} &= -\big(Q_{\mathbf{vv}} - Q_{\mathbf{vu}}Q_{\mathbf{uu}}^{-1}Q_{\mathbf{uv}}\big)^{-1}\big(Q_{\mathbf{v}} - Q_{\mathbf{vu}}Q_{\mathbf{uu}}^{-1}Q_{\mathbf{u}}\big),\\
\mathbf{L}_{\mathbf{v}} &= -\big(Q_{\mathbf{vv}} - Q_{\mathbf{vu}}Q_{\mathbf{uu}}^{-1}Q_{\mathbf{uv}}\big)^{-1}\big(Q_{\mathbf{vx}} - Q_{\mathbf{vu}}Q_{\mathbf{uu}}^{-1}Q_{\mathbf{ux}}\big).
\end{aligned}
$$

IV. BACKWARD PROPAGATION OF THE VALUE FUNCTION

Notice that the feed-forward and feedback gains are functions of the value function and its first and second order partial derivatives with respect to $\mathbf{x}$. Therefore, we need to find a way to obtain these values, and this is presented in the following proposition.

Proposition 4.1: The value function and its first and second order partial derivatives with respect to $\mathbf{x}$ can be determined by the following backward ordinary differential equations

$$
\begin{aligned}
-\frac{\mathrm{d}V}{\mathrm{d}t} &= Q_0 + \mathbf{l}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{u}} + \mathbf{l}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{v}} + \frac{1}{2}\mathbf{l}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{uu}}\mathbf{l}_{\mathbf{u}}\\
&\quad + \mathbf{l}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{uv}}\mathbf{l}_{\mathbf{v}} + \frac{1}{2}\mathbf{l}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{vv}}\mathbf{l}_{\mathbf{v}},\\
-\frac{\mathrm{d}(\nabla_{\mathbf{x}}V)}{\mathrm{d}t} &= Q_{\mathbf{x}} + \mathbf{L}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{u}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{v}}\\
&\quad + Q_{\mathbf{ux}}^{\mathsf{T}}\mathbf{l}_{\mathbf{u}} + Q_{\mathbf{vx}}^{\mathsf{T}}\mathbf{l}_{\mathbf{v}} + \mathbf{L}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{uu}}\mathbf{l}_{\mathbf{u}} + \mathbf{L}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{uv}}\mathbf{l}_{\mathbf{v}}\\
&\quad + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{vu}}\mathbf{l}_{\mathbf{u}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{vv}}\mathbf{l}_{\mathbf{v}},\\
-\frac{\mathrm{d}(\nabla_{\mathbf{xx}}V)}{\mathrm{d}t} &= Q_{\mathbf{xx}} + 2\mathbf{L}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{ux}} + 2\mathbf{L}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{vx}}\\
&\quad + 2\mathbf{L}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{vu}}\mathbf{L}_{\mathbf{u}} + \mathbf{L}_{\mathbf{u}}^{\mathsf{T}}Q_{\mathbf{uu}}\mathbf{L}_{\mathbf{u}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}}Q_{\mathbf{vv}}\mathbf{L}_{\mathbf{v}},
\end{aligned}
\qquad (15)
$$

where the Q-functions are in the form

$$
\begin{aligned}
Q_0 \;&= \mathcal{L} + \frac{1}{2}\mathrm{tr}(\nabla_{\mathbf{xx}}V_t\mathbf{G}\mathbf{G}^{\mathsf{T}}),\\
Q_{\mathbf{x}} \;&= \nabla_{\mathbf{x}}\mathcal{L} + \nabla_{\mathbf{x}}V_t\nabla_{\mathbf{x}}\mathbf{f} + \sum_{j=1}^{m}\mathbf{G}^{(j)\mathsf{T}}\nabla_{\mathbf{xx}}V_t\nabla_{\mathbf{x}}\mathbf{G}^{(j)},\\
Q_{\mathbf{u}} \;&= \nabla_{\mathbf{u}}\mathcal{L} + \nabla_{\mathbf{x}}V_t\nabla_{\mathbf{u}}\mathbf{f}, \qquad Q_{\mathbf{v}} = \nabla_{\mathbf{v}}\mathcal{L} + \nabla_{\mathbf{x}}V_t\nabla_{\mathbf{v}}\mathbf{f},\\
Q_{\mathbf{xx}} &= \nabla_{\mathbf{xx}}\mathcal{L} + 2\nabla_{\mathbf{xx}}V_t\nabla_{\mathbf{x}}\mathbf{f} + \sum_{j=1}^{m}\nabla_{\mathbf{x}}\mathbf{G}^{(j)\mathsf{T}}\nabla_{\mathbf{xx}}V_t\nabla_{\mathbf{x}}\mathbf{G}^{(j)},\\
Q_{\mathbf{uu}} &= \nabla_{\mathbf{uu}}\mathcal{L}, \qquad Q_{\mathbf{vv}} = \nabla_{\mathbf{vv}}\mathcal{L}, \qquad Q_{\mathbf{uv}} = \nabla_{\mathbf{uv}}\mathcal{L},\\
Q_{\mathbf{ux}} &= \nabla_{\mathbf{ux}}\mathcal{L} + \nabla_{\mathbf{u}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_t, \; Q_{\mathbf{vx}} = \nabla_{\mathbf{vx}}\mathcal{L} + \nabla_{\mathbf{v}}\mathbf{f}^{\mathsf{T}}\nabla_{\mathbf{xx}}V_t,\\
Q_{\mathbf{xu}} &= Q_{\mathbf{ux}}^{\mathsf{T}}, \qquad Q_{\mathbf{xv}} = Q_{\mathbf{vx}}^{\mathsf{T}}, \qquad Q_{\mathbf{vu}} = Q_{\mathbf{uv}}^{\mathsf{T}}, \qquad (16)
\end{aligned}
$$

subject to the terminal conditions

$$
\begin{aligned}
V(t_f) = \phi(\bar{\mathbf{x}}(t_f), t_f), \quad \nabla_{\mathbf{x}} V(t_f) = \nabla_{\mathbf{x}} \phi(\bar{\mathbf{x}}(t_f), t_f),\\
\nabla_{\mathbf{x}\mathbf{x}} V(t_f) = \nabla_{\mathbf{x}\mathbf{x}} \phi(\bar{\mathbf{x}}(t_f), t_f). \qquad (17)
\end{aligned}
$$

Proof: In order to find the update law of the value function and its first and second order partial derivatives, we substitute the optimal minimizing control (12) and maximizing control (13) in the expansion of (7) to obtain

$$
\begin{aligned}
& V_t + \nabla_{\mathbf{x}} V_t \delta\mathbf{x}_t + \frac{1}{2}\delta\mathbf{x}_t^{\mathsf{T}} \nabla_{\mathbf{x}\mathbf{x}} V_t \delta\mathbf{x}_t \\
&= V_{t+\mathrm{d}t} + Q_0 \mathrm{d}t + \nabla_{\mathbf{x}} V_{t+\mathrm{d}t} \delta\mathbf{x}_t \\
&+ Q_{\mathbf{x}} \mathrm{d}t \delta\mathbf{x}_t + Q_{\mathbf{u}} \mathrm{d}t \delta\mathbf{u}_t + Q_{\mathbf{v}} \mathrm{d}t \delta\mathbf{v}_t + \frac{1}{2}\delta\mathbf{x}_t^{\mathsf{T}} \nabla_{\mathbf{x}\mathbf{x}} V_{t+\mathrm{d}t} \delta\mathbf{x}_t \\
&+ \frac{1}{2} \begin{bmatrix} \delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathbf{v}_t \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{\mathbf{x}\mathbf{x}}\mathrm{d}t & Q_{\mathbf{x}\mathbf{u}}\mathrm{d}t & Q_{\mathbf{x}\mathbf{v}}\mathrm{d}t \\ Q_{\mathbf{u}\mathbf{x}}\mathrm{d}t & Q_{\mathbf{u}\mathbf{u}}\mathrm{d}t & Q_{\mathbf{u}\mathbf{v}}\mathrm{d}t \\ Q_{\mathbf{v}\mathbf{x}}\mathrm{d}t & Q_{\mathbf{v}\mathbf{u}}\mathrm{d}t & Q_{\mathbf{v}\mathbf{v}}\mathrm{d}t \end{bmatrix} \begin{bmatrix} \delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathbf{v}_t \end{bmatrix}.
\end{aligned} \qquad (18)
$$

After grouping terms on the right-hand side of (18) as zeroth order, first order and second order expressions of $\delta\mathbf{x}_t$, we can equate the coefficients on the left-hand side and right-hand side of (18) and after some mathematical manipulations, we arrive at

$$
\begin{aligned}
-\frac{\mathrm{d}V_t}{\mathrm{d}t} &= Q_0 + \mathbf{l}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}} + \mathbf{l}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}} + \frac{1}{2}\mathbf{l}_{\mathbf{u}} Q_{\mathbf{u}\mathbf{u}} \mathbf{l}_{\mathbf{u}} \\
&+ \mathbf{l}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}\mathbf{v}} \mathbf{l}_{\mathbf{v}} + \frac{1}{2}\mathbf{l}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}\mathbf{v}} \mathbf{l}_{\mathbf{v}}, \qquad (19)\\
-\frac{\mathrm{d}\nabla_{\mathbf{x}} V_t}{\mathrm{d}t} &= Q_{\mathbf{x}} + \mathbf{L}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}} + Q_{\mathbf{u}\mathbf{x}}^{\mathsf{T}} \mathbf{l}_{\mathbf{u}} + Q_{\mathbf{v}\mathbf{x}}^{\mathsf{T}} \mathbf{l}_{\mathbf{v}} \\
+\mathbf{L}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}\mathbf{u}} \mathbf{l}_{\mathbf{u}} &+ \mathbf{L}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}\mathbf{v}} \mathbf{l}_{\mathbf{v}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}\mathbf{u}} \mathbf{l}_{\mathbf{u}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}\mathbf{v}} \mathbf{l}_{\mathbf{v}}, \qquad (20)\\
-\frac{\mathrm{d}\nabla_{\mathbf{x}\mathbf{x}} V_t}{\mathrm{d}t} &= Q_{\mathbf{x}\mathbf{x}} + 2\mathbf{L}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}\mathbf{x}} + 2\mathbf{L}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}\mathbf{x}} \\
&+ 2\mathbf{L}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}\mathbf{u}} \mathbf{L}_{\mathbf{u}} + \mathbf{L}_{\mathbf{u}}^{\mathsf{T}} Q_{\mathbf{u}\mathbf{u}} \mathbf{L}_{\mathbf{u}} + \mathbf{L}_{\mathbf{v}}^{\mathsf{T}} Q_{\mathbf{v}\mathbf{v}} \mathbf{L}_{\mathbf{v}}. \qquad (21)
\end{aligned}
$$

Letting $\mathrm{d}t \to 0$ in (19) through (21), we readily obtain (15). Similarly, the expressions of the Q-functions are turned into (16).

At the final time, we have $V(\mathbf{x}(t_f), t_f) = \phi(\mathbf{x}(t_f))$. By taking the expansions around $\bar{\mathbf{x}}(t_f)$ we get

$$
\begin{aligned}
\phi(\mathbf{x}(t_f)) &= \phi(\bar{\mathbf{x}}(t_f) + \delta\mathbf{x}(t_f)) \\
&\approx \phi(\bar{\mathbf{x}}(t_f)) + \nabla_{\mathbf{x}} \phi(\bar{\mathbf{x}}(t_f)) \delta\mathbf{x}(t_f) \\
&+ \delta\mathbf{x}(t_f)_{\mathsf{T}} \nabla_{\mathbf{x}\mathbf{x}} \phi(\bar{\mathbf{x}}(t_f)) \delta\mathbf{x}(t_f). \qquad (22)
\end{aligned}
$$

Therefore, the boundary conditions at $t = t_f$ for the backward differential equations are represented by (17), and this completes the proof. ■

Now that we have found a method to obtain the value function and its first and second order partial derivatives with respect to the state through backward propagation, we put all the pieces together and provide the Stochastic Game Theoretic Differential Dynamic Programming (SGT-DDP) algorithm in a pseudocode form shown in Algorithm 1.

The cost function is chosen depending on the application. The roles of minimizing and maximizing controls in the control design are determined by the choices of the Hessian of $\mathcal{L}$ with respect to the controls. In order to see this feature, recall that $Q_{\mathbf{u}\mathbf{u}} = \nabla_{\mathbf{u}\mathbf{u}}\mathcal{L}$ and $Q_{\mathbf{v}\mathbf{v}} = \nabla_{\mathbf{v}\mathbf{v}}\mathcal{L}$. Furthermore, since $\nabla_{\mathbf{u}\mathbf{u}}\mathcal{L}$ and $\nabla_{\mathbf{v}\mathbf{v}}\mathcal{L}$ are design parameters, they can be chosen such that $\nabla_{\mathbf{u}\mathbf{u}}\mathcal{L}$ is positive definite and $\nabla_{\mathbf{v}\mathbf{v}}\mathcal{L}$ is negative definite. Such design makes sure that the role of the controller $\mathbf{u}$ is to minimize the cost whereas the controller $\mathbf{v}$ aims to maximize it. Since $Q_{\mathbf{u}\mathbf{u}} > 0$ and $Q_{\mathbf{v}\mathbf{v}} < 0$, we can deduce that $\left(Q_{\mathbf{u}\mathbf{u}} - Q_{\mathbf{u}\mathbf{v}} Q_{\mathbf{v}\mathbf{v}}^{-1} Q_{\mathbf{v}\mathbf{u}}\right)^{-1} > 0$, and $\left(Q_{\mathbf{v}\mathbf{v}} - Q_{\mathbf{v}\mathbf{u}} Q_{\mathbf{u}\mathbf{u}}^{-1} Q_{\mathbf{u}\mathbf{v}}\right)^{-1} < 0$. Combining these two matrix inequalities and the form of the feed-forward and feedback gains of the control policies in the expressions for $\mathbf{l}_{\mathbf{u}}, \mathbf{L}_{\mathbf{u}}, \mathbf{l}_{\mathbf{v}}$ and $\mathbf{L}_{\mathbf{v}}$, it can be seen that the controls are updated such that the control $\mathbf{u}$ tends to reduce the cost while the control $\mathbf{v}$ tends to increase it.

Algorithm 1 Pseudocode of the SGT-DDP Algorithm

Given:

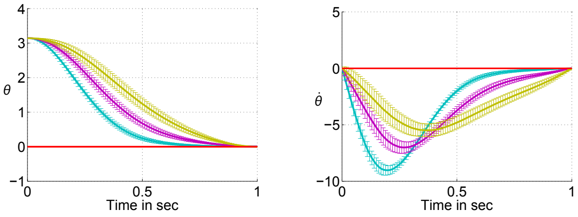
- Stochastic dynamics $\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{v}, t)\mathrm{d}t + \mathbf{G}(\mathbf{x})\mathrm{d}w$
- Initial condition of the dynamics $\mathbf{x}_0$
- Initial minimizing control $\bar{\mathbf{u}}$ and maximizing control $\bar{\mathbf{v}}$
- Terminal time t_f
- Multiplier γ
- A constant N

```
procedure UPDATE_CONTROL($\mathbf{x}_0$, $\bar{\mathbf{u}}$, $\bar{\mathbf{v}}$, $t_f$, $\gamma$, $N$)
    for $i$ from 1 to $N$ do
        Find the initial mean trajectory $\bar{\mathbf{x}}$ by integrating the deterministic part of the controlled dynamics forward with $\mathbf{x}_0$, $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$;
        Find the value of $V, V_{\mathbf{x}}, V_{\mathbf{x}\mathbf{x}}$ at $t_f$ according to (17);
        Compute the quadratic approximation of the value function $V, V_{\mathbf{x}}, V_{\mathbf{x}\mathbf{x}}$ in $[0, t_f]$ by integrating backward the equations (15);
        Compute $\mathbf{l}_{\mathbf{u}}, \mathbf{L}_{\mathbf{u}}, \mathbf{l}_{\mathbf{v}}, \mathbf{L}_{\mathbf{v}}$ with the $Q$-functions from (16);
        Compute $\delta\mathbf{x}(t)$ through $\delta\mathbf{x}_{t+\mathrm{d}t} = \delta\mathbf{x}_t + (\nabla_{\mathbf{x}}\mathbf{f}\delta\mathbf{x}_t + \nabla_{\mathbf{u}}\mathbf{f}\delta\mathbf{u}_t + \nabla_{\mathbf{v}}\mathbf{f}\delta\mathbf{v}_t)\mathrm{d}t$ while replacing $\delta\mathbf{u}$ and $\delta\mathbf{v}$ with $(\mathbf{l}_{\mathbf{u}} + \mathbf{L}_{\mathbf{u}}\delta\mathbf{x})$ and $(\mathbf{l}_{\mathbf{v}} + \mathbf{L}_{\mathbf{v}}\delta\mathbf{x})$, respectively;
        Compute $\delta\mathbf{u} = \mathbf{l}_{\mathbf{u}} + \mathbf{L}_{\mathbf{u}}\delta\mathbf{x}$ and $\delta\mathbf{v} = \mathbf{l}_{\mathbf{v}} + \mathbf{L}_{\mathbf{v}}\delta\mathbf{x}$;
        Update control $\mathbf{u}^* = \mathbf{u}^* + \gamma\delta\mathbf{u}$, where $\gamma \in (0, 1]$ is chosen as the learning rate;
        Set $\bar{\mathbf{u}} = \mathbf{u}^*$ and $\bar{\mathbf{v}} = \bar{\mathbf{v}}^*$;
    end for
    return $\mathbf{x}^*$, $\mathbf{u}^*$, $\mathbf{v}^*$, $\mathbf{l}_{\mathbf{u}}, \mathbf{L}_{\mathbf{u}}, \mathbf{l}_{\mathbf{v}}, \mathbf{L}_{\mathbf{v}}$.
end procedure
```

V. Simulation Results

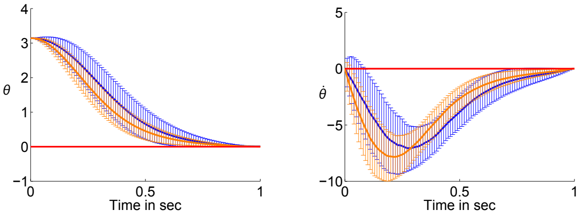
In this section, we apply the proposed SGT-DDP algorithm to two systems. The first system is the inverted pendulum and the second one is the cart pole problem. Specifically, the first system is governed by the equations

$$
\mathrm{d}\mathbf{x} = \begin{bmatrix} \mathbf{x}(2) \\ (mg\ell/I)\sin\mathbf{x}(1) - (b/I)\mathbf{x}(2) + (1/I)(\mathbf{u} + \mathbf{v}) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} 0 \\ \alpha\mathbf{x}(1) \end{bmatrix} \mathrm{d}w, \qquad (23)
$$



(a) (b)

Fig. 1: (a) Plots of mean and standard deviation of 1000 trajectories of θ. Cyan, magenta and yellow plots correspond to the case of $R_\mathbf{v} = 0.13, 0.2$ and 10, respectively. The red line at the bottom depicts the desired state $\theta = 0$. (b) Plots of mean and standard deviation of 1000 trajectories of $\dot{\theta}$.



(a) (b)

Fig. 2: (a) Comparison of plots of mean and standard deviation of 1000 trajectories of θ with respect to the SGT-DDP and the GT-DDP control in orange and blue, respectively. (b) Comparison of plots of $\dot{\theta}$ with respect to the SGT-DDP and the GT-DDP control.

where $\mathbf{x} = [\theta, \dot{\theta}]^\intercal$ and the parameters are chosen as $m = 1$ Kg, $\ell = 0.5$ m, $b = 0.1$, $I = ml^2$, $g = 9.81\text{Kg} \cdot \text{m}/\sec^2$ and $\alpha = 1$. Our goal is to bring the pendulum from the initial state $[\theta, \dot{\theta}] = [\pi, 0]$ to the target position $[\theta, \dot{\theta}] = [0, 0]$. The cost function is given by

$$J = \mathbf{x}(t_f)^\intercal Q_f \mathbf{x}(t_f) + \int_0^{t_f} (\mathbf{u}^\intercal R_\mathbf{u} \mathbf{u} - \mathbf{v}^\intercal R_\mathbf{v} \mathbf{v})\, \mathrm{d}t, \tag{24}$$

where

$$Q_f = \begin{bmatrix} 100, & 0 \\ 0, & 5 \end{bmatrix}. \tag{25}$$

For the simulation, we set $R_\mathbf{u} = 0.1$ and $R_\mathbf{v} = 0.13, 0.2, 10$ to observe how the change of control authority of the maximizing control affects the outcome of the simulation.

We set the initial control to be $\bar{\mathbf{u}} \equiv 0$, $\bar{\mathbf{v}} \equiv 0$, the terminal time to be $t_f = 1$ and the multiplier $\gamma = 0.8$. For each value of $R_\mathbf{v}$, we run the inverted pendulum system with feedback minimizing control for 1000 times. In Fig. 1a, we have three colored plots, where cyan, magenta and dark yellow plots correspond to the case of $R_\mathbf{v} = 0.13, 0.2$ and 10, respectively. The plot of each color contains the mean of the trajectories of θ with respect to time and an error bar with a distance of the standard deviation above and below the curve is drawn at every time step. Similarly, the mean and standard deviation of the trajectories of $\dot{\theta}$ for these values of $R_\mathbf{v}$ are shown in Fig. 1b.

It can be observed from Fig. 1 that the feedforward and feedback parts of the control policy change with $R_\mathbf{v}$. In particular, as $R_\mathbf{v}$ decreases the feedforward control steers the mean trajectory towards the desired state earlier. Moreover, the optimal feedback gains reduce the variability of the trajectories when $R_\mathbf{v}$ gets small. This behavior indicates that the game-theoretic formulation can give rise to robust policies that shape both the mean and the variance of the optimal trajectories.

Furthermore, we compare the performance of the feedback control emerging from our algorithm with the control that results from the deterministic game theoretic DDP in [14]. This time, we fix $R_\mathbf{v} = 1$ and $\alpha = 4$. The other parameters remain unchanged. The result is shown in Fig. 2 where the orange plots depict the mean and error bars of the state trajectories subject to the feedback control originated from the algorithm proposed in this paper and the blue plots are associated with the deterministic game theoretic DDP from [14]. It can be seen that the SGT-DDP algorithm returns a control that drives the mean trajectory towards the desired state earlier. One explanation of this behavior is as follows. Recall from (15) that

$$\begin{aligned} -\frac{\mathrm{d}(\nabla_{\mathbf{xx}} V)}{\mathrm{d}t} &= Q_{\mathbf{xx}} + 2\mathbf{L}_\mathbf{u}^\intercal Q_{\mathbf{ux}} + 2\mathbf{L}_\mathbf{v}^\intercal Q_{\mathbf{vx}} \\ &\quad + 2\mathbf{L}_\mathbf{v}^\intercal Q_{\mathbf{vu}} \mathbf{L}_\mathbf{u} + \mathbf{L}_\mathbf{u}^\intercal Q_{\mathbf{uu}} \mathbf{L}_\mathbf{u} + \mathbf{L}_\mathbf{v}^\intercal Q_{\mathbf{vv}} \mathbf{L}_\mathbf{v}. \end{aligned} \tag{26}$$

Let $Q_{\mathbf{vv}} = \nabla_{\mathbf{vv}} \mathcal{L}$ be negative definite. Then the more authority the maximizing control has (the smaller $Q_{\mathbf{vv}}$ is), the larger the right-hand side of (26) becomes. Similarly, by the expression of $Q_{\mathbf{xx}}$, as the state-dependent noise gets larger, the right-hand side of (26) also increases. Therefore, the noise and the maximizer affect the update of $\nabla_{\mathbf{xx}} V$ in a similar fashion. Hence, it is expected that the enhancement of the control authority of the maximizing control and the inclusion of noise in SGT-DDP result in similar behavior, as shown in Figs. 1 and 2.

In the next example, we consider the cart pole problem with conflicting controls under stochastic disturbances. This is an underactuated mechanical system and the corresponding dynamics is given by

$$\dot{\mathbf{x}} = f(\mathbf{x}) + G(\mathbf{x})(\mathbf{u} + \mathbf{v} + \mathrm{d}w), \tag{27}$$

where

$$f(\mathbf{x}) = \begin{bmatrix} \mathbf{x}(2) \\ \dfrac{m \sin \mathbf{x}(3)(-\ell \mathbf{x}\dot{(4)}^2 + g \cos \mathbf{x}(3))}{M + m \sin^2 \mathbf{x}(3)} \\ \mathbf{x}(4) \\ \dfrac{-m\ell \mathbf{x}(4)^2 \cos \mathbf{x}(3) \sin \mathbf{x}(3) + (M + m) g \sin \mathbf{x}(3)}{\ell(M + m \sin^2 \mathbf{x}(3))} \end{bmatrix}, \tag{28}$$

and

$$G(\mathbf{x}) = \begin{bmatrix} 0 \\ \dfrac{1}{M + m \sin(\mathbf{x}(3))^2} \\ 0 \\ \dfrac{\cos(\mathbf{x}(3))}{\ell(M + m \sin(\mathbf{x}(3))^2)} \end{bmatrix}, \tag{29}$$

The state $\mathbf{x} = [x, \dot{x}, \theta, \dot{\theta}]^\intercal$ where x represents the displacement of the cart and θ stands for the angle of the pole.

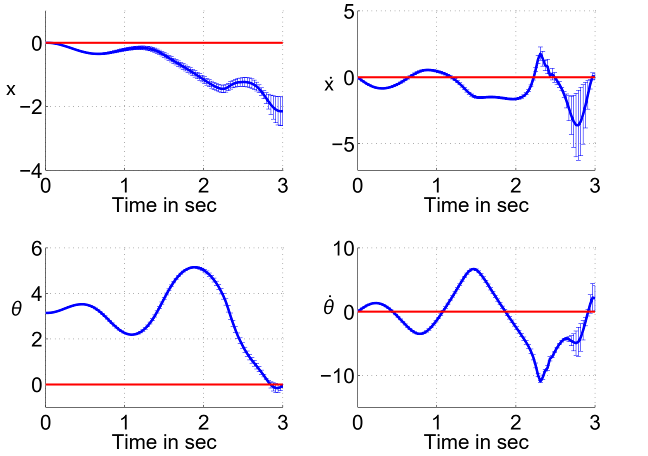


Fig. 3: Mean and standard variance of 100 trajectories of the four states with respect to time under conflicting controls in blue. The red lines represents the goal states $\theta = 0$.

$\ell = 0.5$ is the length of the pole, $M = 10$ is the mass of the cart and $m = 1$ is the mass of the pole, and $g = 9.8$ is the gravitational constant. The cost function is in the form

$$J = (\mathbf{x}(t_f) - x_f)^{\mathsf{T}} Q_f (\mathbf{x}(t_f) - x_f) \tag{30}$$

$$+ \int_0^{t_f} (\mathbf{u}^{\mathsf{T}} R_{\mathbf{u}} \mathbf{u} - \mathbf{v}^{\mathsf{T}} R_{\mathbf{v}} \mathbf{v}) \mathrm{d}t, \tag{31}$$

where $Q_f = \mathrm{diag}([0, 500, 5000, 50])$. The other parameters in the cost function are given by $R_{\mathbf{u}} = 0.01$, $R_{\mathbf{v}} = 0.1$. The minimizing control $\mathbf{u}$ aims to bring the system from the initial state $\mathbf{x}_0 = [0, 0, \pi, 0]^{\mathsf{T}}$ to the desired state $x_f = [0, 0, 0, 0]^{\mathsf{T}}$, whereas the maximizing control $\mathbf{v}$ attempts to stop this from happening. Note that the terminal displacement is actually not restricted to reach zero since $Q_f(1, 1) = 0$ in the cost function. The initial controls are set to $\bar{\mathbf{u}} \equiv 0$, $\bar{\mathbf{v}} \equiv 0$, the terminal time $t_f = 3$ and the multiplier is set to $\gamma = 0.3$. The mean of 100 trajectories of the states under conflicting feedback controls and stochastic disturbances are depicted in Fig. 3 in blue. Error bars of the standard deviation are drawn around the mean trajectories.

VI. Conclusion

We consider a differential game involving two conflicting controls under stochastic dynamics. Starting from the Bellman-Isaacs equation, we take expansions of the value function and its derivatives around a nominal trajectory and find the update law of the minimizing and maximizing controls of both players, as well as the backward differential equations of the approximation of the value function up to the second order. We present the SGT-DDP algorithm and analyze the effect of the game theoretic formulation in the feed-forward and feedback parts of the control policies.

The SGT-DDP algorithm is tested on three distinct systems: one is a first-order nonlinear system and the other two are the inverted pendulum and cart pole problems with conflicting controls. We investigate how the intensity of the stochastic noise affects the behavior of the controls and the corresponding trajectories.

Possible extensions of this research include applications of this method to more realistic systems with higher order dynamics, including many applications starting from neuromuscular and bio-mechanical systems to stochastic pursuit-evasion problems. The extension of SGT-DDP to systems with control and state constraints is another direction of the research.

Acknowledgement: Funding for this research was provided by the NSF (award no. CMMI-1160780) and AFOSR (award no. FA9550-13-1-0029).

References

[1] D. H. Jacobson and D. Q. Mayne, *Differential Dynamic Programming*. New York,: American Elsevier Pub. Co., 1970.

[2] J. Morimoto and C. Atkeson, "Minimax differential dynamic programming: An application to robust biped walking," in *In Advances in Neural Information Processing Systems 15*, Cambridge, MA: MIT Press, 2002.

[3] Y. Tassa, T. Erez, and W. D. Smart, "Receding horizon differential dynamic programming," in *Advances in Neural Information Processing Systems*, pp. 1465–1472, 2008.

[4] E. Todorov and W. Li, "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems," in *American Control Conference*, (Portland, OR), pp. 300–306, June 8-10 2005.

[5] E. Theodorou, Y. Tassa, and E. Todorov, "Stochastic differential dynamic programming," in *American Control Conference*, (Baltimore, MD), pp. 1125–1132, June 30 - July 2 2010.

[6] Y. Tassa, T. Erez, and E. Todorov, "Synthesis and stabilization of complex behaviors through online trajectory optimization," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Vilamoura, Algarve, Portugal), pp. 4906–4913, Oct. 7-12 2012.

[7] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng, "An application of reinforcement learning to aerobatic helicopter flight," in *In Advances in Neural Information Processing Systems 19*, p. 2007, MIT Press, 2007.

[8] T. Erez, Y. Tassa, and E. Todorov, "Infinite-horizon model predictive control for periodic tasks with contacts," in *Proceedings of Robotics: Science and Systems*, (Los Angeles, CA), June 2011.

[9] Y. Tassa, N. Mansard, and E. Todorov, "Control-limited differential dynamic programming," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, (Hong Kong, China), pp. 1168–1175, May 31 - June 07 2014.

[10] C. G. Atkeson and J. Morimoto, "Nonparametric representation of policies and value functions: A trajectory-based approach," in *Neural Information Processing Systems*, pp. 1611–1618, MIT Press, 2003.

[11] C. Atkeson and B. Stephens, "Random sampling of states in dynamic programming," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, pp. 924–929, Aug 2008.

[12] P. Whittle, "Risk-sensitive Linear/Quadratic/Gaussian control," *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981.

[13] W. H. Fleming and W. M. McEneaney, "Risk-sensitive control on an infinite time horizon," *SIAM Journal on Control and Optimization*, vol. 33, no. 6, pp. 1881–1915, 1995.

[14] W. Sun, E. Theodorou, and P. Tsiotras, "Game theoretic continuous time differential dynamic programming," in *American Control Conference*, (Chicago, IL), pp. 5593–5598, July 1–3 2015.

[15] T. Basar and P. Berhard, *H_∞ Optimal Control and Related Minimax Design*. Boston: Birkhauser, 1995.

[16] D. H. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions of Automatic Control*, vol. AC - 18, pp. 124–131, 1973.

[17] W. H. Fleming and P. E. Souganidis, "On the existence of value-functions of two-player, zero-sum stochastic differential-games," *Indiana University Mathematics Journal*, vol. 38, no. 2, pp. 293–314, 1989.