

Minimax Iterative Dynamic Game: Application to Nonlinear Robot Control Tasks.

Olalekan Ogunmolu, Nicholas Gans, and Tyler Summers.

Abstract—Multistage decision policies provide useful control strategies in high-dimensional state spaces, particularly in complex control tasks. However, they exhibit weak performance guarantees in the presence of disturbance, model mismatch, or model uncertainties. This brittleness limits their use in high-risk scenarios. We present how to quantify the sensitivity of such policies in order to inform of their robustness capacity. We also propose a minimax iterative dynamic game framework for designing robust policies in the presence of disturbance/uncertainties. We test the quantification hypothesis on a carefully designed deep neural network policy; we then pose a minimax iterative dynamic game (iDG) framework for improving policy robustness in the presence of adversarial disturbances. We evaluate our iDG framework on a mecanum-wheeled robot, whose goal is to find a locally robust optimal multistage policy that achieve a given goal-reaching task. The algorithm is simple and adaptable for designing meta-learning/deep policies that are robust against disturbances, model mismatch, or model uncertainties, up to a disturbance bound. Videos of the results are on the author’s [website](#), while the codes for reproducing our experiments are on [github](#). A self-contained environment for reproducing our results is on [docker](#).

I. INTRODUCTION

Multistage decision policies are often brittle to deploy on real-world systems owing to their lack of robustness [1], [2] and the data inefficiency of the learning process.

Methods of designing scalable high-dimensional policies often rely on heuristics e.g. [3], which do not always produce repeatable results. Quite often, policies are learned under partial observability, but sampling with partial observations can be unstable [4]. In the presence of model uncertainties or model mismatch between the source and target environments [5], we must therefore devise policies that are robust to perturbations. Our goal in this paper is to provide an underpinning for designing robust policies, leveraging on methods from H_∞ control theory [6], dynamic programming (DP) [7], differential dynamic programming (DDP) [8], and iterative LQG [9]. Essentially, we consider the performance of policies in the presence of various adversarial agents [10]–[12] using a minimax method. While recent DRL techniques produce performance efficiency for agent tasks in the real world [13]–[19], there are sensitivity concerns that need to be addressed, e.g. the trade-off between a system’s nominal performance and its performance in the face of uncertainty or model mismatch.

Olalekan Ogunmolu and Nicholas Gans are with the Department of Electrical Engineering, Tyler Summers is with the Department of Mechanical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA. {olalekan.ogunmolu, ngans, tyler.summers}@utdallas.edu

Contributions

- We provide a framework for *testing the brittleness of a policy*: Given a nominal policy, we pit a disturbing input against it, whose magnitude is controlled by a disturbance parameter, γ . If an adversarial agent causes significant performance degradation, then γ indicates the upper bound on the efficiency of such policy.
- In an *iterative, dynamic two player zero-sum Markov game* (iDG), we let each agent execute an opposite reaction to its pair: a concave-convex problem ensues, and we seek to drive the policy toward a saddle point equilibrium. This iDG framework generates local control laws – providing an alternating best response update of global control and adversarial policies, leading to a saddle-point equilibrium convergence. This is essentially a meta-algorithm that can be extended to quantify and design the robustness of model-free, model-based RL, as well as DP/iterative LQG family of policies.

We evaluate our proposal for control of robot motor tasks using policy search [15], [20] and the ILQG algorithm [9]. The rest of this paper is thus organized: we provide a formal treatment policy sensitivity quantification and the iDG algorithm within a linearly solvable MDP [21] in [Sec. II](#). The dynamics and model of the robot we that evaluates our iDG hypothesis is presented in [Sec. III](#). Results for the two proposals of this work are discussed in [Sec. IV](#), and we conclude this work in [Sec. V](#). A more extensive discussion of the modeling methods in this paper is detailed in [22].

II. TWO-PLAYER TRAJECTORY OPTIMIZATION

Consider two agents interacting in an environment, \mathcal{E} , over a finite horizon, T ; the states evolve according to a discrete-time stochastic dynamics,

$$\mathbf{x}_{t+1} = f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t), \quad t = 0, \dots, T-1, \quad \mathbf{x}_0 = \bar{\mathbf{x}}_0,$$

where $\mathbf{x}_t \in \mathcal{X}_t$ is the n -dimensional state vector, $\mathbf{u}_t \in \mathcal{U}_t$ is the m -dimensional nominal agent’s action (or control law), and $\mathbf{v}_t \in \mathcal{V}_t$ denote the disturbing agent’s p -dimensional action. The nominal agent chooses its action under a (stochastic) policy $\{\pi = \pi_0, \pi_1, \dots, \pi_T\} \subseteq \Pi$, while the uncertainty’s actions are governed by a policy $\{\psi = \psi_0, \psi_1, \dots, \psi_T\} \subseteq \Psi$. For the policy pair (π, ψ) , we define the *cost-to-go*, $\mathcal{J}(\mathbf{x}, \pi, \psi)$, of a trajectory $\{\mathbf{x}_t\}_{t=0, \dots, T}$ with initial condition \mathbf{x}_0 , as a partial sum of costs from t to T ,

$$\mathcal{J}_0(\mathbf{x}_0, \pi, \psi) = \mathbf{E}_{\mathbf{u}_t, \mathbf{v}_t} \sum_{t=0}^{T-1} \ell_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) + L_T(\mathbf{x}_T),$$

where ℓ_t is a nonnegative function of $(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t)$, denoting the stage cost, and L_T is a nonnegative function of \mathbf{x}_T , denoting the final cost. We seek a pair of *saddle point equilibrium* policies, (π^*, ψ^*) that satisfy,

$$\mathcal{J}_0(\mathbf{x}_0, \pi^*, \psi) \leq \mathcal{J}_0(\mathbf{x}_0, \pi^*, \psi^*) \leq \mathcal{J}_0(\mathbf{x}_0, \pi, \psi^*),$$

$\forall \pi \in \Pi, \psi \in \Psi$ and \mathbf{x}_0 . For the general case where we start from an initial condition \mathbf{x}_t , one may write the dynamic programming (DP) equation above as

$$\mathcal{J}_t^*(\mathbf{x}_t) = \min_{\pi \in \Pi} \max_{\psi \in \Psi} \mathcal{J}_t(\mathbf{x}_t, \pi, \psi),$$

where π and ψ contain the control sequences $\{\mathbf{u}_t\}$ and $\{\mathbf{v}_t\}$, and $\mathcal{J}(\cdot)$ denote the Hamilton-Jacobi Bellman cost function. The saddle point equilibrium for an optimal control sequence pair $\{\mathbf{u}_t^*, \mathbf{v}_t^*\}$ can be obtained with

$$\begin{aligned} \mathcal{J}_t^*(\mathbf{x}_t) &= \min_{\pi \in \Pi} \max_{\psi \in \Psi} \mathcal{J}_t(\mathbf{x}_t, \pi, \psi) \\ &= \min_{\pi \in \Pi} \max_{\psi \in \Psi} [\ell_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) + \mathcal{J}_{t+1}^*(f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t))] \end{aligned} \quad (1)$$

A. Quantifying a policy's robustness

Suppose that the nominal policy, π , of an agent has been found, and consider an interacting disturbing agent so that the closed-loop dynamics is describable by the discretized Euler equation,

$$\begin{aligned} \mathbf{x}_{t+1} &= f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t), \quad \mathbf{u}_t \sim \pi_t \\ &= \tilde{f}_t(\mathbf{x}_t, \mathbf{v}_t), \quad t = 0, \dots, T-1. \end{aligned} \quad (2)$$

For stage costs of the form, $\ell_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) = c_t(\mathbf{x}_t, \mathbf{u}_t) - \gamma g_t(\mathbf{v}_t)$, where $c_t(\mathbf{x}_t, \mathbf{u}_t)$ represents the nominal stage cost, $g_t(\cdot)$ is a norm on the adversarial input¹, penalizing the adversary's actions, and $\gamma > 0$ is a disturbance term that controls the strength of the adversary; the adversary faces a maximization problem of the form

$$\max_{\psi \in \Psi} \mathbf{E}_{\mathbf{u}_t \sim \pi_t} \sum_{t=0}^T c(\mathbf{x}_t, \mathbf{u}_t) - \gamma g(\mathbf{v}_t) = \max_{\psi \in \Psi} \mathbf{E} \sum_{t=0}^T \tilde{\ell}_t^{\gamma}(\mathbf{x}_t, \mathbf{v}_t).$$

Varying γ increases/decreases the penalty incurred by the adversarial agent's actions. As $\gamma \rightarrow \infty$, the adversary's optimal policy is to do nothing, since any action will incur an infinite penalty; as γ decreases, the adversary incurs lower penalties, causing large system disturbance. The (inverse of the) smallest γ -value for which the adversary causes unacceptable performance degradation (e.g., instability) provides a measure of robustness of the nominal agent's policy π . The parameter γ is a distinguishing feature of our work; γ quantifies the \mathcal{H}_{∞} norm of the closed-loop system, a measure of its robustness to an adversarial input. The adversary need not represent a malicious input, but can be interpreted as a worst possible disturbance of a given magnitude.

In systems with nonlinear dynamics, one can use differential dynamic programming (DDP) [8] or iterative LQG [23] to optimize the *adversary* against the closed-loop system under the nominal agent's policy π . Indeed,

¹This formulation takes \mathcal{V}_t as a vector space but one can as well define a nonnegative adversarial penalty term when \mathcal{V}_t is a finite set.

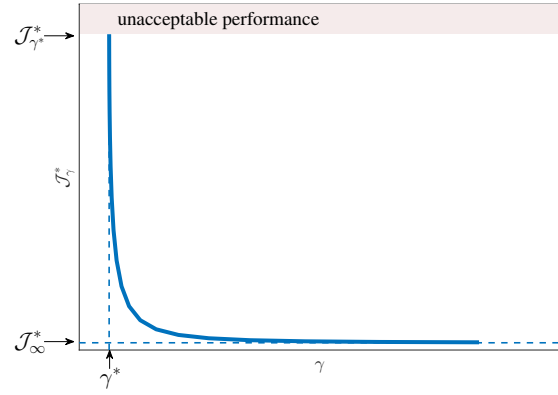


Fig. 1: Prototypical Performance-Robustness Tradeoff Curve.

Algorithm 1 Robustness Curve via iLQG

- 1: **for** $\gamma \in \text{range}(\gamma_{\min}, \gamma_{\max}, N)$ **do**
- 2: **for** $i \in \{1, \dots, I\}$ **do**
- 3: Initialize $x_0 = x(0)$
- 4: $p_i \leftarrow \arg \min_{p_i} \mathbf{E}_{p_i} \left[\sum_{t=0}^T \tilde{\ell}_t^{\gamma}(\mathbf{x}_t, \mathbf{v}_t) \right]$
- 5: **end for**
- 6: **end for**

Morimoto (in [24]) applied minimax DDP to a walking biped robot, but we noticed errors in the value function recursions². We are motivated by the improved convergence guarantees of iterative LQG (iLQG) over DDP, and its suitability for solving constrained nonlinear control problems by iteratively linearizing a nonlinear system about a local neighboring trajectory and computing the optimal local control laws. Our minimax iLQG framework facilitates learning control decision strategies that are robust in the presence of disturbances and modeling errors – improving upon nominal iLQG policies.

A performance-robustness tradeoff curve can be computed by optimizing the adversary for various values of γ . A prototypical curve is illustrated in Fig. 1. This curve quantifies the robustness of a fixed policy π , defining an upper bound \mathcal{J}_{γ}^* on the cost function that does not yield acceptable performance: this is the critical value of γ to be identified. This γ^* -value determines how much adversarial effort we need to suffer performance degradation; large γ^* values will give poor robustness of π – even with little adversarial effort. The algorithm for optimizing adversaries and generating the trade-off curve is summarized in Algorithm 1. The adversary's policy may not be globally optimal due to the approximation of policies in continuous spaces and because the algorithm may only converge to a locally optimal solution, we obtain an upper bound on the robustness of the policy π .

B. Achieving Robustness via IDG

We propose to arrive at a saddle point equilibrium with (1) by continually solving an online finite-horizon trajectory optimization problem. This is essentially a minimax framework and is applicable in principle to any off/on-policy RL algorithm. In this work, we generalize the online trajectory

²These corrections are given in (7).

optimization algorithm of [23], to a two-player, zero-sum dynamic game as follows:

- we approximate the nonlinear system dynamics (c.f. (2)), starting with a schedule of the nominal agent's local controls, $\{\bar{\mathbf{u}}_t\}$, and nominal adversarial agent's local controls $\{\bar{\mathbf{v}}_t\}$ which are assumed to be available (when these are not available, we can initialize them to 0)
- we then run the system's passive dynamics with $\{\bar{\mathbf{u}}_t\}, \{\bar{\mathbf{v}}_t\}$ to generate a nominal state trajectory $\{\bar{\mathbf{x}}_t\}$, with neighboring trajectories $\{\mathbf{x}_t\}$
- we choose a small neighborhood, $\{\delta\mathbf{x}_t\}$ of $\{\mathbf{x}_t\}$, which provides an optimal reduction in cost as the dynamics no longer represent those of $\{\mathbf{x}_t\}$
- discretizing time, the new state and control sequence pairs become $\delta\mathbf{x}_t = \mathbf{x}_t - \bar{\mathbf{x}}_t$, $\delta\mathbf{u}_t = \mathbf{u}_t - \bar{\mathbf{u}}_t$, $\delta\mathbf{v}_t = \mathbf{v}_t - \bar{\mathbf{v}}_t$.

Setting $V(\mathbf{x}, T) = \ell_T(\mathbf{x}_T)$, the min-max over the entire control sequence reduces to a stepwise optimization over a single control, going backward in time with

$$V(\mathbf{x}_t) = \min_{\mathbf{u}_t \sim \pi} \max_{\mathbf{v}_t \sim \psi} [\ell(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) + V(f(\mathbf{x}_{t+1}, \mathbf{u}_{t+1}, \mathbf{v}_{t+1}))].$$

If we consider the Hamiltonian, $\ell(\cdot) + V(\cdot)$, as a perturbation around the tuple $\{\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t\}$, the cost over the local neighborhood via an approximate Taylor series expansion becomes

$$Q_t = \ell(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t, t) + V(\mathbf{x}_{t+1}, t+1).$$

A second-order approximation of the perturbed Q -coefficients of the LQR problem around the neighborhood $\{\delta\mathbf{x}_t\}$ of the trajectory $\{\mathbf{x}_t\}$ is defined as,

$$Q(\cdot) \approx \frac{1}{2} \begin{bmatrix} 1 \\ \delta\mathbf{x}_t^T \\ \delta\mathbf{u}_t^T \\ \delta\mathbf{v}_t^T \end{bmatrix}^T \begin{bmatrix} 1 & Q_{\mathbf{x}t}^T & Q_{\mathbf{u}t}^T & Q_{\mathbf{v}t}^T \\ Q_{\mathbf{x}t} & Q_{\mathbf{xx}t} & Q_{\mathbf{xu}t} & Q_{\mathbf{xvt}} \\ Q_{\mathbf{u}t} & Q_{\mathbf{ux}t} & Q_{\mathbf{uu}t} & Q_{\mathbf{uv}t} \\ Q_{\mathbf{v}t} & Q_{\mathbf{vx}t} & Q_{\mathbf{vu}t} & Q_{\mathbf{vv}t} \end{bmatrix} \begin{bmatrix} 1 \\ \delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathbf{v}_t \end{bmatrix}, \quad (3)$$

where,

$$\begin{aligned} Q_{\mathbf{x}t} &= \ell_{\mathbf{x}t} + f_{\mathbf{x}t}^T V_{\mathbf{x}t+1}, & Q_{\mathbf{u}t} &= \ell_{\mathbf{u}t} + f_{\mathbf{u}t}^T V_{\mathbf{x}t+1} \\ Q_{\mathbf{v}t} &= \ell_{\mathbf{v}t} + f_{\mathbf{v}t}^T V_{\mathbf{x}t+1}, & Q_{\mathbf{xx}t} &= \ell_{\mathbf{xx}t} + f_{\mathbf{x}t}^T V_{\mathbf{xx}t+1} f_{\mathbf{x}t} \\ Q_{\mathbf{ux}t} &= \ell_{\mathbf{ux}t} + f_{\mathbf{u}t}^T V_{\mathbf{xx}t+1} f_{\mathbf{x}t}, & Q_{\mathbf{vx}t} &= \ell_{\mathbf{vx}t} + f_{\mathbf{v}t}^T V_{\mathbf{xx}t+1} f_{\mathbf{x}t} \\ Q_{\mathbf{uu}t} &= \ell_{\mathbf{uu}t} + f_{\mathbf{u}t}^T V_{\mathbf{xx}t+1} f_{\mathbf{u}t}, & Q_{\mathbf{vv}t} &= \ell_{\mathbf{vv}t} + f_{\mathbf{v}t}^T V_{\mathbf{xx}t+1} f_{\mathbf{v}t} \\ Q_{\mathbf{uv}t} &= \ell_{\mathbf{uv}t} + f_{\mathbf{u}t}^T V_{\mathbf{xx}t+1} f_{\mathbf{v}t}. \end{aligned}$$

This is consistent with linearized methods, where linearized second moment terms will dominate the higher order terms. The LQR approximation to the state and the optimal control performance index become,

$$\begin{aligned} \delta\mathbf{x}_{t+1} &\approx f_{\mathbf{x}t} \delta\mathbf{x}_t + f_{\mathbf{u}t} \delta\mathbf{u}_t + f_{\mathbf{v}t} \delta\mathbf{v}_t \\ \ell(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) &\approx \frac{1}{2} \begin{bmatrix} 1 \\ \delta\mathbf{x}_t^T \\ \delta\mathbf{u}_t^T \\ \delta\mathbf{v}_t^T \end{bmatrix}^T \begin{bmatrix} \ell_{0t} & \ell_{\mathbf{x}t}^T & \ell_{\mathbf{u}t}^T & \ell_{\mathbf{v}t}^T \\ \ell_{\mathbf{x}t} & \ell_{\mathbf{xx}t} & \ell_{\mathbf{ux}t}^T & \ell_{\mathbf{vx}t}^T \\ \ell_{\mathbf{u}t} & \ell_{\mathbf{ux}t} & \ell_{\mathbf{uu}t} & \ell_{\mathbf{uv}t} \\ \ell_{\mathbf{v}t} & \ell_{\mathbf{vx}t} & \ell_{\mathbf{vu}t} & \ell_{\mathbf{vv}t} \end{bmatrix} \begin{bmatrix} 1 \\ \delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathbf{v}_t \end{bmatrix} \end{aligned} \quad (4)$$

where single and double subscripts denote first and second-order derivatives³. The best possible nominal agent's action

³ Note that $\delta\mathbf{x}_k, \delta\mathbf{u}_k$, and $\delta\mathbf{v}_k$ are measured w.r.t the nominal vectors $\bar{\mathbf{x}}_k, \bar{\mathbf{u}}_k, \bar{\mathbf{v}}_k$ and are not necessarily small.

and the worst possible adversarial action can be found by performing the respective arg min and arg max operations on the Q -function in (3) so that

$$\begin{aligned} \delta\mathbf{u}_t^* &= -Q_{\mathbf{uu}t}^{-1} [Q_{\mathbf{u}t}^T + Q_{\mathbf{ux}t} \delta\mathbf{x}_t + Q_{\mathbf{uv}t} \delta\mathbf{v}_t], \\ \delta\mathbf{v}_t^* &= -Q_{\mathbf{vv}t}^{-1} [Q_{\mathbf{v}t}^T + Q_{\mathbf{vx}t} \delta\mathbf{x}_t + Q_{\mathbf{vu}t} \delta\mathbf{u}_t]. \end{aligned} \quad (5)$$

Note that the control strategies in (5) depend on the action of the other player. Say the nominal agent first implements its strategy, then transmits its information to the adversary, which subsequently chooses its strategy; it follows that the adversary can choose a more favorable outcome since it knows what the nominal agent's strategy is. It becomes obvious that the *best* action for the nominal agent is to choose a control strategy that is an optimal response to the action choice of the adversary. Similarly, if the roles of the players are changed, the nominal agent's response to the adversary's *worst* choice will be more favorable since it knows what the adversarial agent's strategy is. Therefore, it does not matter that the order of play is predetermined. We end up with an *iterative dynamic game*, where each agent's strategy depends on its opponent's actions. This ensures that we have a *cooperative game* in which the nominal and adversarial agent alternate between taking best possible and worst possible actions during the trajectory optimization phase. This helps maintain equilibrium around the system's desired trajectory, while ensuring robustness in local policies. Suppose we set,

$$\begin{aligned} \mathbf{K}_{\mathbf{u}t} &= [(I - Q_{\mathbf{uu}t}^{-1} Q_{\mathbf{uv}t} Q_{\mathbf{vv}t}^{-1} Q_{\mathbf{uv}t}^T) Q_{\mathbf{uu}t}^{-1}]^{-1}, \\ \mathbf{K}_{\mathbf{v}t} &= [(I - Q_{\mathbf{vv}t}^{-1} Q_{\mathbf{uv}t}^T Q_{\mathbf{uu}t}^{-1} Q_{\mathbf{uv}t}) Q_{\mathbf{vv}t}^{-1}]^{-1}, \\ \mathbf{g}_{\mathbf{u}t} &= \mathbf{K}_{\mathbf{u}t} (Q_{\mathbf{uv}t} Q_{\mathbf{vv}t}^{-1} Q_{\mathbf{v}t} - Q_{\mathbf{u}t}), \\ \mathbf{g}_{\mathbf{v}t} &= \mathbf{K}_{\mathbf{v}t} (Q_{\mathbf{uv}t}^T Q_{\mathbf{uu}t}^{-1} Q_{\mathbf{u}t} - Q_{\mathbf{v}t}), \\ \mathbf{G}_{\mathbf{u}t} &= \mathbf{K}_{\mathbf{u}t} (Q_{\mathbf{uv}t} Q_{\mathbf{vv}t}^{-1} Q_{\mathbf{vx}t} - Q_{\mathbf{ux}t}), \\ \mathbf{G}_{\mathbf{v}t} &= \mathbf{K}_{\mathbf{v}t} (Q_{\mathbf{uv}t}^T Q_{\mathbf{uu}t}^{-1} Q_{\mathbf{ux}t} - Q_{\mathbf{vx}t}), \end{aligned}$$

then it follows that we can rewrite (5) as

$$\delta\mathbf{u}_t^* = \mathbf{g}_{\mathbf{u}t} + \mathbf{G}_{\mathbf{u}t} \delta\mathbf{x}_t, \quad \delta\mathbf{v}_t^* = \mathbf{g}_{\mathbf{v}t} + \mathbf{G}_{\mathbf{v}t} \delta\mathbf{x}_t. \quad (6)$$

Equation 6 gives the open and closed-loop components of the control equations for both agents. Comparing coefficients in (4), we find that the value function coefficients can be thus written

$$\begin{aligned} \Delta V_t &= \mathbf{g}_{\mathbf{u}t} Q_{\mathbf{u}t} + \mathbf{g}_{\mathbf{v}t} Q_{\mathbf{v}t} + \mathbf{g}_{\mathbf{u}t} Q_{\mathbf{uv}t} \mathbf{g}_{\mathbf{v}t} \\ &\quad + \frac{1}{2} (\mathbf{g}_{\mathbf{u}t} Q_{\mathbf{uu}t} \mathbf{g}_{\mathbf{u}t} + \mathbf{g}_{\mathbf{v}t} Q_{\mathbf{vv}t} \mathbf{g}_{\mathbf{v}t}) \\ V_{\mathbf{x}t} &= Q_{\mathbf{x}t} + \mathbf{G}_{\mathbf{u}t}^T Q_{\mathbf{u}t} + \mathbf{G}_{\mathbf{v}t}^T Q_{\mathbf{v}t} + \mathbf{G}_{\mathbf{u}t}^T Q_{\mathbf{uu}t} \mathbf{g}_{\mathbf{u}t} + \mathbf{g}_{\mathbf{u}t} Q_{\mathbf{ux}t} \\ &\quad + \mathbf{g}_{\mathbf{v}t} Q_{\mathbf{vx}t} + \mathbf{G}_{\mathbf{v}t}^T Q_{\mathbf{vv}t} \mathbf{g}_{\mathbf{v}t} + \mathbf{G}_{\mathbf{v}t}^T Q_{\mathbf{uv}t}^T \mathbf{g}_{\mathbf{u}t} + \mathbf{G}_{\mathbf{u}t}^T Q_{\mathbf{uv}t} \mathbf{g}_{\mathbf{v}t} \\ V_{\mathbf{xx}t} &= \frac{1}{2} (Q_{\mathbf{xx}t} + \mathbf{G}_{\mathbf{u}t}^T Q_{\mathbf{uu}t} \mathbf{G}_{\mathbf{u}t} + \mathbf{G}_{\mathbf{v}t}^T Q_{\mathbf{vv}t} \mathbf{G}_{\mathbf{v}t}) + \mathbf{G}_{\mathbf{u}t}^T Q_{\mathbf{ux}t} \\ &\quad + \mathbf{G}_{\mathbf{v}t}^T Q_{\mathbf{vx}t} + \mathbf{G}_{\mathbf{u}t}^T Q_{\mathbf{uv}t} \mathbf{G}_{\mathbf{v}t}. \end{aligned} \quad (7)$$

These recursive value functions, essentially differentiate our value coefficient recursion equations from Morimoto's DDP recursions.

C. Improved Regularization

For nonlinear systems, the inverse of the Hessian must be strictly positive definite. When the inverse of the Hessian is non-positive-definite, we can add a suitably large positive quantity to it [26], [27], or replace the elements of the diagonal matrix in its eigen decomposition, $[V, \mathcal{D}] = \text{eig}(Q)$, that are smaller than an adequately small ρ , and then set $Q = V\mathcal{D}V^T$ [9]. In this work, ρ is added to $\tilde{\mathbf{Q}}$ when the Hessian is not well-posed. Our update rule is $\tilde{\mathbf{Q}}_{\text{uu}} = \mathbf{Q}_{\text{uu}} + \rho \mathbf{I}_m$. However, the regularization can have adverse effects on the system arising from the control/disturbance transition matrices f_{ut} and f_{vt} ; therefore, we introduce a similar penalty term used in [23] to deviations from states so that the regularization yields a quadratic state cost about the previous policy:

$$\begin{aligned}\tilde{\mathbf{Q}}_{\text{uu}t} &= l_{\text{uu}t} + f_{\text{ut}}^T (V\mathbf{xx}_{t+1} + \rho \mathbf{I}_n) f_{\text{xt}} + V\mathbf{x}_{t+1} f_{\text{uu}t} \\ \tilde{\mathbf{Q}}_{\text{vv}t} &= l_{\text{vv}t} + f_{\text{vt}}^T (V\mathbf{xx}_{t+1} + \rho \mathbf{I}_n) f_{\text{xt}} + V\mathbf{x}_{t+1} f_{\text{vv}t} \\ \tilde{\mathbf{Q}}_{\text{ux}t} &= l_{\text{ux}t} + f_{\text{ut}}^T (V\mathbf{xx}_{t+1} + \rho \mathbf{I}_n) f_{\text{xt}} + V\mathbf{x}_{t+1} f_{\text{ux}t} \\ \tilde{\mathbf{Q}}_{\text{vx}t} &= l_{\text{vx}t} + f_{\text{vt}}^T (V\mathbf{xx}_{t+1} + \rho \mathbf{I}_n) f_{\text{xt}} + V\mathbf{x}_{t+1} f_{\text{vx}t}.\end{aligned}\quad (8)$$

The adjusted gains therefore become

$$\begin{aligned}\tilde{\mathbf{K}}_{\text{ut}} &= \left[\left(I - \tilde{\mathbf{Q}}_{\text{uu}t}^{-1} \tilde{\mathbf{Q}}_{\text{uv}t} \tilde{\mathbf{Q}}_{\text{vv}t}^{-1} \tilde{\mathbf{Q}}_{\text{vu}t} \right) \tilde{\mathbf{Q}}_{\text{uu}t}^{-1} \right]^{-1}, \\ \tilde{\mathbf{K}}_{\text{vt}} &= \left[\left(I - \tilde{\mathbf{Q}}_{\text{vv}t}^{-1} \tilde{\mathbf{Q}}_{\text{vu}t} \tilde{\mathbf{Q}}_{\text{uu}t}^{-1} \tilde{\mathbf{Q}}_{\text{uv}t} \right) \tilde{\mathbf{Q}}_{\text{vv}t}^{-1} \right]^{-1} \\ \mathbf{g}_{\text{ut}} &= \tilde{\mathbf{K}}_{\text{ut}} (\tilde{\mathbf{Q}}_{\text{uv}t} \tilde{\mathbf{Q}}_{\text{vv}t}^{-1} \tilde{\mathbf{Q}}_{\text{vt}} - \tilde{\mathbf{Q}}_{\text{ut}}), \\ \mathbf{G}_{\text{ut}} &= \tilde{\mathbf{K}}_{\text{ut}} \left(\tilde{\mathbf{Q}}_{\text{uv}t} \tilde{\mathbf{Q}}_{\text{vv}t}^{-1} \tilde{\mathbf{Q}}_{\text{vxt}} - \tilde{\mathbf{Q}}_{\text{uxt}} \right) \\ \mathbf{g}_{\text{vt}} &= \tilde{\mathbf{K}}_{\text{vt}} (\tilde{\mathbf{Q}}_{\text{vu}t} \tilde{\mathbf{Q}}_{\text{uu}t}^{-1} \tilde{\mathbf{Q}}_{\text{ut}} - \tilde{\mathbf{Q}}_{\text{vt}}), \\ \mathbf{G}_{\text{vt}} &= \tilde{\mathbf{K}}_{\text{vt}} \left(\tilde{\mathbf{Q}}_{\text{vu}t} \tilde{\mathbf{Q}}_{\text{uu}t}^{-1} \tilde{\mathbf{Q}}_{\text{uxt}} - \tilde{\mathbf{Q}}_{\text{vxt}} \right).\end{aligned}\quad (9)$$

The improved value functions are updated in (7) accordingly.

D. Regularization Schedule

To accurately tweak the regularization term, ρ , we adopt a regularization schedule that penalizes $\mathbf{Q}_{\text{uu}t}$ or $\mathbf{Q}_{\text{vv}t}$ when the backward pass fails. When the backward pass is successful, we would desire rapid decrease in ρ in order to assure fast convergence; otherwise, we would want to quickly increase ρ , albeit in a bumpless manner since the minimum value of ρ that prevents divergence is of linear order. We let ρ_0 denote some minimal modification value for ρ (set to 1.0), and we adjust ρ as follows:

increase ρ:	reduce ρ:
$\rho \leftarrow 1.1 \rho_0$	$\rho \leftarrow 0.09 \rho_0$
$\rho_0 = \rho$	$\rho_0 = \rho$

III. DYNAMICS MODELING AND SIMULATION

We consider the KUKA youbot⁴ platform with four mecanum wheels, capable of spatial $\{x, y\}$ motion, *i.e.* sideways, and forward, and an in-place θ -rotation about the z -axis (see Fig. 2a). It is equipped with a 5-DoF arm, mounted on its base. We use the complete kinematic and dynamic model of the youbot platform, accounting for the wheels' friction and mass, while neglecting the links' masses and

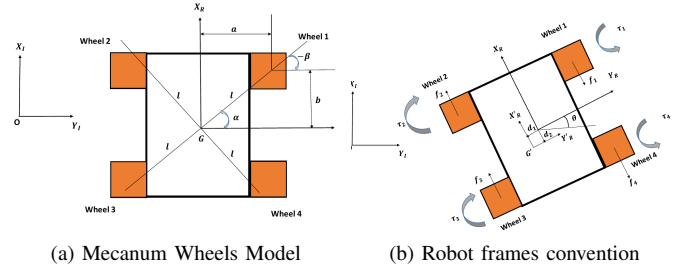


Fig. 2: Robot Geometry

their associated inertia forces. The coordinates of the robot in the world frame are denoted $\mathbf{x}_R = [x_R, y_R, \theta_R]^T$, where given as the x_R, y_R are coordinates of the origin of the robot frame and θ_R is the relative angle between the world and robot x axes (see Fig. 2b).

The torques that govern the robot's motion are obtained from [28]. We run our experiments in the Gazebo physics engine, which has its reference frame defined as x pointing forward, y pointing sideways, and z pointing up. Therefore, our reference frame and robot geometry are as illustrated in Figs 2a and 2b. Formally, we define the generalized Lagrangian equation of the robot as,

$$\mathbf{M}(\mathbf{x})\ddot{\mathbf{x}} + \mathbf{C}(\mathbf{x}, \dot{\mathbf{x}})\dot{\mathbf{x}} + \mathbf{B}^T \mathbf{S} \mathbf{f} = \frac{1}{r} \mathbf{B}^T \boldsymbol{\tau} \quad (10)$$

where $\boldsymbol{\tau} = [\tau_1, \tau_2, \tau_3, \tau_4]$ is the wheel torque vector, r is the wheel radius, $\mathbf{f} = [f_1, f_2, f_3, f_4]^T$ is the friction vector, and \mathbf{S} and \mathbf{B} map the inverse kinematics, gravity, external forces and robot's angle, θ , to each wheel torque; matrices \mathbf{M} and \mathbf{C} denote the inertia and coriolis properties of the robot. \mathbf{B} and \mathbf{S} are given by,

$$\mathbf{B} = \begin{bmatrix} -(\cos \theta + \sin \theta) & -(\cos \theta + \sin \theta) & -\sqrt{2}l \sin(\zeta) \\ -(\cos \theta + \sin \theta) & (\cos \theta - \sin \theta) & -\sqrt{2}l \sin(\zeta) \\ (\cos \theta - \sin \theta) & (\cos \theta + \sin \theta) & -\sqrt{2}l \sin(\zeta) \\ (\cos \theta + \sin \theta) & -(\cos \theta - \sin \theta) & -\sqrt{2}l \sin(\zeta) \end{bmatrix}$$

$$\mathbf{S} = \text{diag} [sgn(\dot{\phi}_1), sgn(\dot{\phi}_2), sgn(\dot{\phi}_3), sgn(\dot{\phi}_4)];$$

$\zeta = \pi/4 - \alpha$, l is the mounting distance of the wheels as shown in Fig. 2a, and $\dot{\phi}_i$, is the rotation speed of each wheel about its axis of rotation. We apply the generalized force/torque vector, F_i , to the base frame of the robot, defined as,

$$F_i = \sum_{j=1}^4 \left(\tau_j - r sgn(\dot{\phi}_j) f_j \right) \frac{\partial \dot{\phi}_j}{\partial \dot{\mathbf{x}}_i}, \quad i = \{1, 2, 3\} \quad (11)$$

A. Trajectory Optimization: ILQR

This section describes the navigation of the robot using the nominal ILQR algorithm, the objective function design, and specific initializations for the robot. The goal is for the robot to move optimally from the center of the environment in Fig. 3a to within a 0.1m radius of the orange box attached to the right-hand corner of Fig. 3b. We separate the state and control cost terms for easier manipulation of cost components. For the state's instantaneous cost, we define a pseudo "smooth-

⁴<https://goo.gl/CYTjvD>

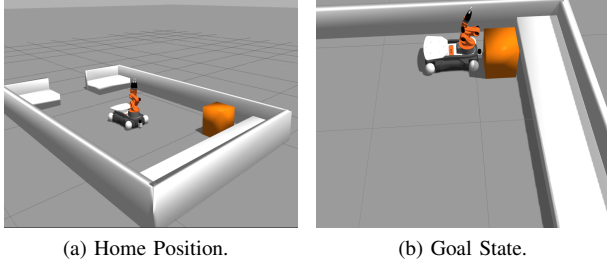


Fig. 3: Goal Navigation Illustration

abs" function,

$$\ell(\mathbf{x}_t) = \sqrt{\alpha + (\mathbf{x}_t - \mathbf{x}^*)^T \text{diag}(\mathbf{w}_x) (\mathbf{x}_t - \mathbf{x}^*)}, \quad (12)$$

where \mathbf{x}^* denotes the desired state, \mathbf{w}_x is a state penalty vector, and α is a constant that controls the curvature of the cost function: the lower the α value, the smoother is the robot's trajectory near the goal state. The $(\mathbf{x} - \mathbf{x}^*)$ term encourages the robot to follow the nominal trajectory while driving toward the goal state. This l_{12} function enforces reaching a desired target exactly in 3D space. Equation 12 encourages the relative weighting of state-cost terms along different axes of motions. Inspired by [23], we choose a hyperbolic cosine control cost function (instead of a quadratic cost which gives disproportional controls in different portions of the state space) defined as:

$$\ell(\mathbf{u}_t) = \alpha^2 (\cosh(\mathbf{w}_u^T \mathbf{u}_t) - 1), \quad (13)$$

where \mathbf{w}_u is a control penalty vector. This cost function limits control outputs to an α -sized neighborhood in \mathbf{u} -space. We set α to 10^{-4} , \mathbf{w}_u to $[1, 1, 1, 1]$ and \mathbf{w}_x to $[1, 1, 0.8]$. We found a time horizon, $T = 150$ to be appropriate for this task. We proceed as follows:

- starting with an open-loop control schedule $\{\bar{\mathbf{u}}_t\}$, (initialized to $[1.3, 0.8, 0.1]$), we generate the nominal states $\{\bar{\mathbf{x}}_t\}$
- replacing τ with \mathbf{u} in (10), we compute the forward dynamics:

$$\ddot{\mathbf{x}} = -\mathbf{M}^{-1} \mathbf{C} \dot{\mathbf{x}} - \mathbf{M}^{-1} \mathbf{B}^T \left(\mathbf{S} \mathbf{f} - \frac{1}{r} \mathbf{u} \right) \quad (14)$$

- we then obtain derivatives of \mathbf{f} and those of ℓ from (4)
- starting at time $t = T - 1$, we compute the associated nominal Q coefficients in (4), obtain the open and closed-loop gains, \mathbf{g}_u , \mathbf{G}_u , and obtain the value function derivatives
- in the forward pass, we proceed from $t = \{0, \dots, T - 1\}$, and update the trajectories with a backtracking line search parameter, $0 < \varsigma \leq 1$, as follows

$$\hat{\mathbf{x}}(1) = \mathbf{x}(1), \hat{\mathbf{u}}(t) = \mathbf{u}(t) + \varsigma \mathbf{g}_u(t) + \mathbf{G}_u(t)(\hat{\mathbf{x}}(t) - \mathbf{x}(t))$$

$$\hat{\mathbf{x}}(t+1) = \mathbf{f}(\hat{\mathbf{x}}(t), \hat{\mathbf{u}}(t)) \quad (15)$$

- the backward/forward pass informs of the change in cost ΔJ^* , which is compared to the estimated reduction

$\eta = (J(\mathbf{u}_{1,\dots,T-1}) - J(\hat{\mathbf{u}}_{1,\dots,T-1}) / \Delta(J(\rho)))$, where

$$\Delta(J(\rho)) = \rho \sum_{t=1}^{T-1} \mathbf{g}_{ut}^T \mathbf{Q}_{ut} + \frac{\rho^2}{2} \sum_{i=1}^{T-1} \mathbf{g}_{ut}^T \mathbf{Q}_{uu} \mathbf{g}_{ut}.$$

- we accept a trajectory only if $0 < c < \eta$ (where we have chosen $c = 0.5$) following [8]'s recommendation.
- set $\bar{\mathbf{x}}_t = \mathbf{x}_t (t = 1, \dots, T)$, $\mathbf{u}_t = \bar{\mathbf{u}}_t$, reset ρ and repeat the forward pass
- stop when $\Delta(J(\rho)) < \chi$ (where χ is a problem dependent term).

B. Trajectory Optimization: IDG

We initialized the adversarial inputs \mathbf{v} from a Gaussian distribution $\sim \mathcal{N}(\mathbf{0}, \mathbf{2I})$ and augment the stage control cost of (13) as:

$$\ell(\mathbf{u}_t, \mathbf{v}_t) = \alpha^2 (\cosh(\mathbf{w}_u^T \mathbf{u}_t) - \gamma \cosh(\mathbf{w}_v^T \mathbf{v}_t)). \quad (16)$$

$\gamma \cosh(\mathbf{w}_v^T \mathbf{v}_t)$ introduces a weighting term in the disturbing input, with γ as the robustness parameter. We set \mathbf{w}_v to $[1, 1, 1, 1]$ and run the two player, zero-sum game erstwhile described. Again, we compute the derivatives \mathbf{f}_{ut} and \mathbf{f}_{vt} , calculate the associated derivatives in (4) and give the optimized iDG torque to the robot. We initialized the nominal disturbance vectors $\bar{\mathbf{v}}$, and \mathbf{v}_t as a multivariate Gaussian-filtered, random noise vectors $\sim \mathcal{N}(\mathbf{0}, \mathbf{2I})$. Our iDG process goes thus:

- starting with an open-loop control and disturbance schedules $\{\bar{\mathbf{u}}_t\}$, $\{\bar{\mathbf{v}}_t\}$, we generate the nominal states $\{\bar{\mathbf{x}}_t\}$
- we then obtain the derivatives \mathbf{f}_{ut} , \mathbf{f}_{vt} , \mathbf{f}_{uv} , \mathbf{f}_{uu} , \mathbf{f}_{vv} and those of ℓ from (4)
- in a backward pass, we obtain the associated nominal Q coefficients, the open and closed-loop gains, and obtain the improved value function coefficients with (7)
- in the forward pass, we proceed from $t = \{0, \dots, T-1\}$, and update the trajectories with a backtracking linesearch parameter, $0 < \varsigma \leq 1$, as follows

$$\hat{\mathbf{x}}(1) = \mathbf{x}(1), \hat{\mathbf{u}}(t) = \mathbf{u}(t) + \varsigma \mathbf{g}_u(t) + \mathbf{G}_u(t)(\hat{\mathbf{x}}(t) - \mathbf{x}(t))$$

$$\hat{\mathbf{v}}(t) = \mathbf{v}(t) + \varsigma \mathbf{g}_v(t) + \mathbf{G}_v(t)(\hat{\mathbf{x}}(t) - \mathbf{x}(t))$$

$$\hat{\mathbf{x}}(t+1) = \mathbf{f}(\hat{\mathbf{x}}(t), \hat{\mathbf{u}}(t), \hat{\mathbf{v}}(t)) \quad (17)$$

- the backward/forward pass informs of the change in cost ΔJ^* , which is compared to the estimated reduction $\eta = \frac{J(\mathbf{u}_{1,\dots,T-1}, \mathbf{v}_{1,\dots,T-1}) - J(\hat{\mathbf{u}}_{1,\dots,T-1}, \hat{\mathbf{v}}_{1,\dots,T-1})}{\Delta(J(\rho))}$, where

$$\Delta(J(\rho)) = \rho \sum_{t=1}^{T-1} [\mathbf{g}_u(t)^T \mathbf{Q}_u(t) + \mathbf{g}_v(t)^T \mathbf{Q}_v(t)] + \dots$$

$$\dots \frac{\rho^2}{2} \sum_{i=1}^{T-1} [\mathbf{g}_u(t)^T \mathbf{Q}_{uu}(t) \mathbf{g}_u(t) + \mathbf{g}_v(t)^T \mathbf{Q}_{vv}(t) \mathbf{g}_v(t)]$$

- we accept a trajectory only if $0 < c < \eta$ (choosing $c = 0.5$)
- set $\bar{\mathbf{x}}_t = \mathbf{x}_t, \{t=1, \dots, T\}$, $\mathbf{u}_t = \bar{\mathbf{u}}_t$, $\mathbf{v}_t = \bar{\mathbf{v}}_t$, reset ρ and repeat the forward pass
- stop when $\Delta(J(\rho)) < \chi$ (where χ is a problem dependent term)

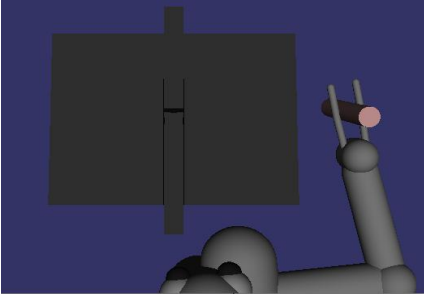


Fig. 4: Policy Robustness Quantification Experiment

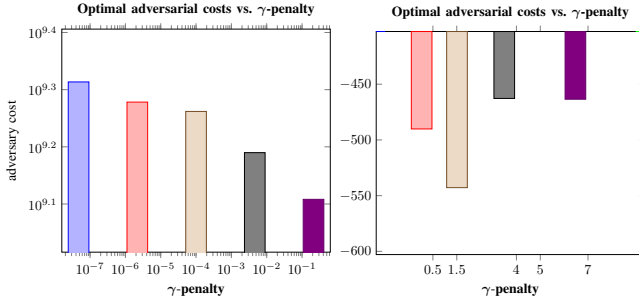


Fig. 5: Robustness Analysis for Peg Insertion Task dependent term).

IV. RESULTS

We implement the policy sensitivity analysis using the guided policy search algorithm of [16]. We then run the iDG optimization algorithm on the youbot and analyze the effectiveness of using our method against standard ILQG. The codes for reproducing the experiments can be found on github at <https://github.com/lakehanne/youbot/tree/rilqg>.

A. Robustness Margins

We first generate data used to train a global neural network policy by running our ILQG optimization scheme for the local controllers. We then run an adversarial local controller in closed-loop with the nominal agent’s controller for various disturbance values, γ , in a supervised learning of global neural network policies. We task the PR2 robot (shown in Fig. 4 and simulated in the MuJoCo physics simulator [29]), to insert a peg into the hole at the bottom of the slot. The model was learned with a mixture of $N = 40$ Gaussians and we parameterized the local control laws with the deep neural network described in [16]. The difficulty of this task stems from the discontinuity in dynamics from the contact between the peg and the slot’s surface.

We choose quadratic stage costs for both the state and the two controllers

$$\ell(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) = \mathbf{w}_u \mathbf{u}_t^T \mathbf{u}_t + \mathbf{w}_x \|\mathbf{d}(\mathbf{x}_t) - \mathbf{d}^*\|^2 - \gamma \mathbf{v}_t^T \mathbf{v}_t \quad (18)$$

where $\mathbf{d}(\mathbf{x}_t)$ denotes the end effector position at state \mathbf{x}_t and \mathbf{d}^* denotes the desired end effector position at the base of the slot. The cost function quantifies the squared distance of the end effector to its desired position and energy of the the controller and adversary torque inputs. We set \mathbf{w}_u and \mathbf{w}_x to 10^{-6} and 1, respectively. For various γ -values, we evaluate the robustness of the trained policy by training the adversary

with the GPS algorithm, and observe its effect on the task performance. We run each experiment for 11 GPS iterations. Fig. 5 shows that the adversary causes a sharp degradation in the controller performance for values of $\gamma < 1.5$. This corresponds to when the optimized policy is destabilized and the arm fails to reach the desired target. As values of γ increase above 1.5, however, we observe that the adversary has a reduced effect on the given task. Video of this result is on the robustness margins link on our website.

B. ILQG-based Trajectory Optimization

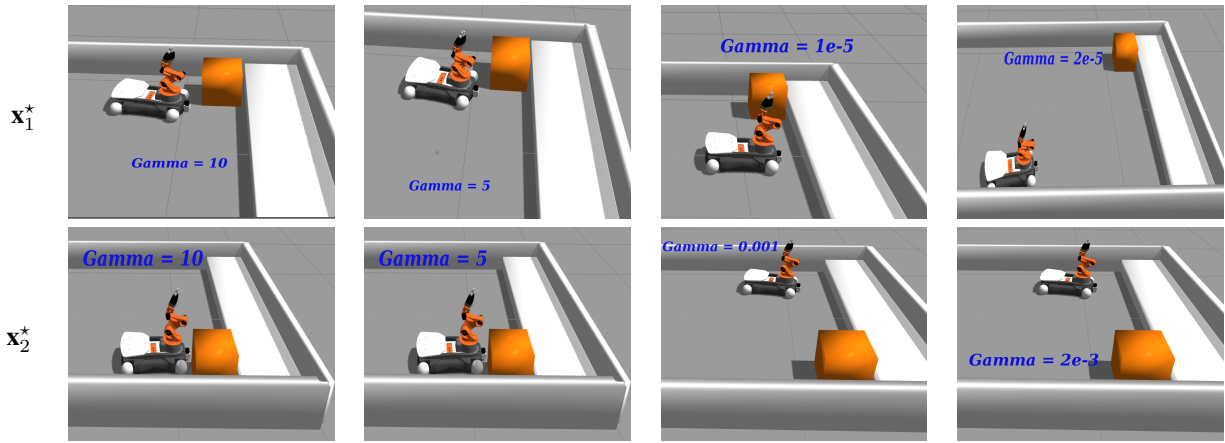
We evaluate the trajectory optimization algorithm on the youbot KUKA robotic platform. The iLQG result is best understood by watching the ILQG video result available here: <https://goo.gl/JhshTB>. In the video, observe that the arm on the robot vibrates with abrupt motions while navigating toward the desired goal. This is expected as the dynamics of the arm links were not included in the model of the platform. Regardless, the robot reaches the goal state after 76 seconds.

C. Trajectory Optimization: iDG

We found $T = 150$ to be a desirable stage horizon, initiate the disturbance from a multivariate Gaussian filtered noise $\mathcal{N}(0, 2)$, set the gains for the torques to $\{k_{F_x} = 10, k_{F_y} = 15, k_{F_\theta} = 1\}$ respectively, and run the algorithm described in III-B for various γ -disturbance values. We run two experiments: one with the goal at far left corner of the environment and the second with the goal state at the far right corner of the environment. We pose various adversarial inputs against the controller with values of γ in the range $\{10, 5, 3, 1.5, 0.65, 0.1, 0.1, 10^{-5}, 2 \times 10^{-3}, 2 \times 10^{-5}\}$ for both experiments. The result showing the trajectory evolution and the position of the youbot after each iDG run for various γ values is better appreciated by watching the videos available on our website: <https://goo.gl/JhshTB>.

In both experiments I and II, observe that for values of γ in the range $1.5 \leq \gamma \leq 10$, the robot drives smoothly and reaches the goal without the arm vibrating on the platform as in the ILQG video. This is despite that did not take the mass and inertia matrix components of the arm links into consideration. When $\gamma \leq 1.5$, we notice that the adversarial disturbance’s effect on the overall system start getting pronounced. While the robot still reaches its desired goal state for $\gamma = 0.5$ and 1.5 , the motion of the arm is no longer stable. Indeed for critical values of $\gamma \leq 10^{-3}$, the trajectory of the robot becomes perturbed as well as the balance of the robot on the arm. As γ becomes very small (below 10^{-5}), the robot’s trajectory is undefined, and it converges to a spurious minimum as both experiments show. This γ indices would correspond to the \mathcal{J}_γ^* that depicts unacceptable performance in Fig. 1.

There is a time-robustness tradeoff in solving a policy optimization scheme with our iDG algorithm/other nonlinear methods. While iLQG does achieve the goal in generally 3-4 iterations, iDG solves the same task in about 5-7 iterations. When speed is not crucial, and the safety of a real-world agent/its environment is important, iDG provides a viable alternative for designing safe policies. Without loss



State convergence of the KUKA robot with our iDG formulation given different goal states and varying γ -values

of generality, we envisage that our minimax iDG scenario is extensible to similar systems that compute gradients of a cost function or reward in achieving an optimal control task. Compared to Morimoto’s work [24], our minimax iDG does not have to learn the unmodeled disturbance with reinforcement learning after implementing the minimax player. We conjecture that [24]’s model lacked robustness due to the incorrect second order derivative of the value function recursions. We thus identify the critical value of γ to be between $0.1 - 10^{-3}$ as being the value that causes maximal disturbance in the robot’s torque. To design a robust policy, one can consider γ values in this range in order to design a smooth trajectory for this particular using the minimax iDG algorithm.

V. CONCLUSIONS

Despite exhibiting near-optimal performance, high-dimensional policies often exhibit brittleness in the presence of adversarial agents, model mismatch or policy transfer [2], [3]. In this work, we have presented a way of identifying the greatest upper bound on a policy’s sensitivity via a minimax framework. We have also presented an iDG framework that informs of how to make a policy robust to unmodeled disturbances, and uncertainties, up to a disturbance bound by converging to a saddle equilibrium. We ran four experiments in total to validate our hypothesis and the videos of our results are available on our website.

REFERENCES

- [1] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, “Tactics of Adversarial Attack on Deep Reinforcement Learning Agents,” *arXiv preprint arXiv:1703.06748*, 2017. 1
- [2] J. Kos and D. Song, “Delving into Adversarial Attacks on Deep Policies,” *arXiv preprint arXiv:1705.06452*, 2017. 1, 7
- [3] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, “Robust Adversarial Reinforcement Learning,” *arXiv preprint arXiv:1703.02702*, 2017. 1, 7
- [4] J. Garcia and F. Fernández, “A Comprehensive Survey on Safe Reinforcement Learning,” *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015. 1
- [5] A. Mandlekar, Y. Zhu, A. Garg, L. Fei-Fei, and S. Savarese, “Adversarially Robust Policy Learning: Active Construction of Physically-Plausible Perturbations,” 2017. 1
- [6] K. Zhou and J. C. Doyle, *Essentials of Robust Control*. Prentice hall Upper Saddle River, NJ, 1998, vol. 104. 1
- [7] R. Bellman, *Dynamic programming*. Princeton University Press, 1957. 1
- [8] D. H. Jacobson and D. Q. Mayne, *Differential Dynamic Programming*. American Elsevier Publishing Company, Inc., New York, NY, 1970. 1, 2, 5
- [9] E. Todorov and W. Li, “A Generalized Iterative LQG Method For Locally-optimal Feedback Control Of Constrained Nonlinear Stochastic Systems,” *IEEE Conference on Decision and Control*, 2004. 1, 4
- [10] Basar, Tamer and Olsder, Geert Jan, *Dynamic Noncooperative Game Theory*. Academic Press, New York, 1999. 1
- [11] M. L. Littman, “Markov Games as a Framework for Multi-agent Reinforcement Learning,” in *Proceedings of the Eleventh International Conference on Machine Learning*, vol. 157, 1994, pp. 157–163. 1
- [12] J. Morimoto and K. Doya, “Robust Reinforcement Learning,” *Neural computation*, vol. 17, no. 2, pp. 335–359, 2005. 1
- [13] I. Mordatch, K. Lowrey, G. Andrew, Z. Popovic, and E. V. Todorov, “Interactive control of diverse complex characters with neural networks,” in *Advances in Neural Information Processing Systems*, 2015, pp. 3132–3140. 1
- [14] T. Zhang, G. Kahn, S. Levine, and P. Abbeel, “Learning Deep Control Policies for Autonomous Aerial Vehicles with MPC-Guided Policy Search,” in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 528–535. 1
- [15] S. Levine and V. Koltun, “Learning complex neural network policies with trajectory optimization,” in *International Conference on Machine Learning*, 2014, pp. 829–837. 1
- [16] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-End Training of Deep Visuomotor Policies,” *Journal of Machine Learning Research*, vol. 17, pp. 1–40, 2016. 1, 6
- [17] W. Montgomery and S. Levine, “Guided Policy Search as Approximate Mirror Descent,” *arXiv preprint arXiv:1607.04614*, 2016. 1
- [18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015. 1
- [19] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous Methods for Deep Reinforcement Learning,” in *International Conference on Machine Learning*, 2016, pp. 1928–1937. 1
- [20] M. P. Deisenroth, G. Neumann, and J. Peters, “A Survey on Policy Search for Robotics,” *Foundations and Trends in Robotics*, vol. 2, no. 1, pp. 1–142, 2011. 1
- [21] K. Dvijotham and E. Todorov, “Inverse Optimal Control with Linearly-Solvable MDPs,” in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 335–342. 1
- [22] O. Ogunmolu, N. Gans, and T. Summers, “Robust two player, zero-sum deep reinforcement learning,” 2017. 1
- [23] Y. Tassa, T. Erez, and E. Todorov, “Synthesis and Stabilization of Complex Behaviors through Online Trajectory Optimization,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2012. 2, 3, 4, 5
- [24] J. Morimoto, G. Zeglin, and C. Atkeson, “Minimax Differential Dynamic Programming: Application to A Biped Walking Robot,”

- IEEE/RSJ International Conference on Intelligent Robots and Systems.*, vol. 2, no. October, pp. 1927–1932, 2003. [2](#), [7](#)
- [25] M. M. Polycarpou and P. A. Ioannou, “Modelling, Identification And Stable Adaptive Control Of Continuous-time Nonlinear Dynamical Systems Using Neural Networks,” in *American Control Conference, 1992.* IEEE, 1992, pp. 36–40.
- [26] H. Kelley, *AIAA Astrodynamics Specialists Conference, Yale University*, August 1963. [4](#)
- [27] T. Bullock and G. Franklin, *IEEE Transactions on Automatic Control*, pp. AC-12, 666, 1967. [4](#)
- [28] L.-C. Lin and H.-Y. Shih, “Modeling and adaptive control of an omnimecanum-wheeled robot,” *Intelligent Control and Automation*, vol. 4, no. 02, p. 166, 2013. [4](#)
- [29] E. Todorov, T. Erez, and Y. Tassa, “MuJoCo: A Physics Engine for Model-based Control,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE, 2012, pp. 5026–5033. [6](#)