The driver and the engineer: Reinforcement learning and robust control

Natalie Bernat, Jiexin Chen, Nikolai Matni and John Doyle

Abstract— **Reinforcement learning (RL) and other AI methods are exciting approaches to data-driven control design, but RL's emphasis on maximizing expected performance contrasts with robust control theory (RCT), which puts central emphasis on the impact of model uncertainty and worst case scenarios. This paper argues that these approaches are potentially complementary, roughly analogous to that of a driver and an engineer in, say, formula one racing. Each is indispensable but with radically different roles. If RL takes the driver seat in safety critical applications, RCT may still play a role in plant design, and also in diagnosing and mitigating the effects of performance degradation due to changes or failures in component or environments. While much RCT research emphasizes synthesis of controllers, as does RL, in practice RCT's impact has perhaps already been greater in using hard limits and tradeoffs on robust performance to provide insight into plant design, interpreted broadly as including sensor, actuator, communications, and computer selection and placement in addition to core plant dynamics. More automation may ultimately require more rigor and theory, not less, if our systems are going to be both more efficient and robust. Here we use the simplest possible toy model to illustrate how RCT can potentially augment RL in finding mechanistic explanations when control is not merely hard, but impossible, and issues in making them more compatibly data-driven. Despite the simplicity, questions abound. We also discuss the relevance of these ideas to more realistic challenges.**

I. Introduction

From vision-based control to agile robotics, learning based methods have been applied to continuous control problems with tremendous and dramatic success. Perhaps even more impressive is that in many such cases, the most successful learning based control methods have been *model-free*, in that no explicit representation of the system dynamics (e.g., the function mapping current state and action to next state) is learned.[1] Indeed, such methods have many favorable properties, such as being broadly applicable and simple to implement, and not suffering from bias due to improper model class selection.

Although undeniably impressive, many (if not most) of these demonstrations were performed in highly-controlled and stylized settings, wherein safety and robustness were not primary concerns. However, recent catastrophic failures of learning based control systems (e.g., autonomous vehicle involvement in fatal collisions) underscore the need to ensure that such methods be both interpretable, i.e., root causes of failures can be identified and remedied, and provably safe. This need for robustness and safety has motivated the development of model-based methods, wherein an approximate system model is learned, its uncertainty is quantified, and then an optimal or robust controller is computed with respect to this nominal model and uncertainty set. Such methods are reminiscent of classical approaches to robust control, with the main difference being that contemporary results focus on providing finite-data sample guarantees.

The vast range and scope of model-free and model-based methods has made it difficult to quantitatively compare them. In light of this, the optimal linear quadratic (LQ) control of an unknown linear-time-invariant system has proved to be a useful and perhaps surprisingly challenging benchmark for learning based control methods [15]. Initiated in [9], and revisited in [3], [7], the contemporary study of this problem has focused on providing bounds on the amount of data needed for near optimal performance. These two references, as well as [2], [4]–[6], [13], [14], [16], take a model based approach wherein estimates of the linear dynamics are learned and used to synthesize a controller, and prove sub-linear *regret bounds* on the performance of adaptive control strategies under a variety of assumptions. In a parallel line of work, model-free methods have also been shown to converge to optimal control policies for the LQ control problem [1], [8], [10], [18], again with regret bound guarantees provided under a variety of assumptions. We refer the interested reader to [11] for an exhaustive and historical perspective on the interplay between learning and control.

We emphasize here that all of the aforementioned results focus on characterizing *upper bounds* on the performance of learning based control strategies on the LQ control problem, once again making it difficult to rigorously and quantitatively compare methods. In fact, while such upper bounds are common in the literature, *lower bounds* are few and far between. To the best of our knowledge, the only such lower bounds can be found in [17], the authors derive asymptotic lower bounds on the number of samples needed by both the classical least-squares-temporal-differencing (LSTD) estimator for policy evaluation, and policy gradient methods for policy improvement, and in doing so, demonstrate a provable gap between model-free and model-based methods for the LQ problem.

Although lower bounds are uncommon in the learning

Natalie Bernat, Jiexin Chen and John C. Doyle are with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA. Nikolai Matni is with the Department of Electrical and Systems Engineering, University of Pennsylvania nbernat@caltech.edu, jch3n@caltech.edu, doyle@caltech.edu, nmatni@seas.upenn.edu

[1]Although there is no agreed upon definition of model-free or model-based methods, for the purposes of this paper, we restrict ourselves to the following simple dichotomy: a method is model-based if it learns a function mapping current state and action to next state; otherwise, it is model free.

based control literature, a rich set of results on the fundamental limits of control can be found in the robust control literature. For example, it is well known that if a system has an open-loop unstable pole p and unstable zero q, and that these are close in value (i.e., there is a near unstable pole/zero cancellation), then this lower bounds the $\mathcal{H}_\infty$ norm of the complementary sensitivity function as $\Omega(\frac{1}{|p-q|})$. To the best of our knowledge, no investigation of the effects of such fundamental limits on learning based control methods exist in the literature – this absence may be due to the fact that many demonstrations of learning based control methods are on over-engineered systems for which sensing and actuation are not an issue. However, as learning systems move from the lab into the real world, systems with full-sensing that are overly-actuated may become prohibitive or impossible to build.

This paper provides a first step towards connecting fundamental limits from robust control to the performance, robustness, and sample-efficiency of both model-based and model-free algorithms. We perform an empirical study comparing and contrasting a simple model-based baseline with a simple model-free approach. In particular, for our model-based baseline, we use the Certainty Equivalent (CE) LQ optimal controller, obtained by first learning a nominal estimate of the system dynamics, and then computing an optimal controller assuming these estimates describe the true behavior of the system. For the model-free method, we use the direct policy gradient (DPG) method suggested in [8]. We apply these methods on a simple toy model that is open-loop unstable, and that is parameterized by a scalar that can be used to vary the controllability of the system (with loss of controllability mimicking the unstable pole/zero cancellation mentioned above) and instability of the system. In this way, our toy model captures the essence of the aforementioned fundamental limits, while still being intuitive, simple, and amenable to exhaustive numerical search.

Given that for the systems considered, even an optimal controller performs poorly, our focus is less on sample efficiency and performance (although we do comment on these when appropriate), but rather on fault diagnosis. In particular, given the behavior of either a CE or DPG derived controller, can diagnostic insights be drawn from the data as to the cause of poor performance, i.e., can data-driven methods be used to identify if the system is uncontrollable? Such insights are valuable from an engineering perspective, as they suggest that poor performance is due not a bad control policy, but rather to a fundamental limit of the system being controlled and its control architecture (i.e., sensing and actuation).

The rest of this paper is structured as follows. In section II, we describe the experimental system and control synthesis methods in detail. In section III-A, we show that the fundamental performance limits set by model-based control theory are indeed observed in our toy system. In section III-B, we present an unexpected phenomenon observed in the model-based experiments, which exposes an inherent vulnerability to control errors and which arises in the process of identifying system dynamics. Finally, we conclude with a short discussion of the many research questions evoked by this simple empirical study.

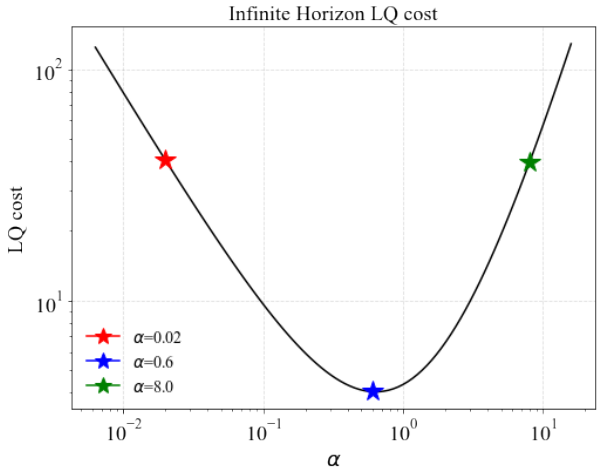

Fig. 1: Optimal infinite time LQ state feedback cost as a function of α. Choices of 3 α values for detailed study are highlighted. These values differ by 400x in α and by 10x in cost.

II. Methods

A. System Model

We ran experiments on the simple dynamical feedback system (1). This system has a tunable parameter, α, which sets performance limits on any controller.

$$x_{t+1} = \begin{bmatrix} 1+\alpha & 0 \\ 0 & 1 \end{bmatrix} x_t + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u_t + w_t \tag{1}$$

where $w_t \sim \mathcal{N}(0,\, \Sigma)$, $\Sigma = \sigma^2 I$, $\sigma = 0.4$, and

$$u_t = K(x_{0:t}, u_{0:t}) \tag{2}$$

We generate the control function K by a variety of methods, with varying degrees of prior information about the true system dynamics (see section II-B for more details). To evaluate the performance of any controller, we use the standard LQ cost of the state and control input:

$$J_{\text{inf}} = \mathbb{E}_w \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (||x_t||^2 + u_t^2) \tag{3}$$

where the expectation is taken over Gaussian-distributed disturbances to the dynamics.

With decreasing α the system becomes increasingly harder to control until at $\alpha = 0$ it becomes uncontrollable[2] (and in fact, unstabilizable). On the other end, as α increases, the system becomes more and more unstable. In both extremes, the optimal LQG control cost degrades (see Figure 1).

To further simplify our experimental setting, we chose three values of α, which represent the three qualitative

[2]as measured by $\sigma_{min}(\Lambda_C)$, the minimum singular value of the controllability Gramian

regions of control challenges: an uncontrollable regime, a benign regime, and an unstable regime. In both the uncontrollable and unstable regimes, we chose a value of α with a roughly 10x performance degradation compared to the benign regime (see Figure 1).

This system is intended to be almost trivially simple to highlight surprising subtle questions, if not complete answers. The impulse responses of each system can be visualized on two plots (see Appendix, Figure 9). We can analytically derive and visualize the space of stabilizing controllers on a single plot (see Appendix, Figure 6). We can easily solve for the optimal controller and the optimal infinite horizon cost (see Figure 1). We can easily do system ID from data using simple least squares, and with certainty equivalence (IDCE), control works surprisingly well. And we can easily explore a variety of RL methods. Nevertheless, the seemingly simple problem of finding data-driven mechanistic explanations when control is unavoidably bad is rich and has surprising twists. This is emphatically not the last word on this subject but a tentative early exploration.

B. Control Methods

1) Direct Policy Gradient (DPG): We explored the performance of the simple model-free policy gradient method from [8]. This algorithm collects simulated trajectories on the system (1) (aka rollouts), with Gaussian noise (with variance σ_η^2) injected into the control signal. These trajectories, and their associated LQ costs, are used to construct a gradient of the constant K, which is then used to implement stochastic gradient descent. Rather than directly optimizing the LQ cost function, this process biases the controller towards policies with lower cost. No estimates of A or B are explicitly constructed.

This method requires an initial stabilizing controller K_0. Using the stabilizing regions found in Figure 6, an initial K_0 was manually chosen for each α such that the infinite horizon LQ cost that it achieved was one order of magnitude above the optimal cost for that α. Choosing such a family of K_0's allows us to explore the optimal performance of this method with relatively little sensitivity to hyperparameters.

To tune the hyperparameters, we searched over a grid of the controller noise σ_η = [1e-4, 1e-3, 1e-2, 0.1, 0.2, 0.3], the step size μ=[1e-7, 1e-6, 1e-5, 1e-4, 1e-3], as well as the time horizon of the rollouts T = [5, 10, 15, 25, 50, 100] (while adjusting the number of iterations as to keep the total number of data points fixed). From this procedure, we chose hyperparameters that produced a stabilizing controller for 100 percent of trials, and then optimized for final cost. This procedure resulted in a time horizon of T=5 for all α's, and (σ_η, μ)=$(0.3, 1e-6)$ for α=0.02, $(0.3, 1e-5)$ for α=0.6, and $(0.2, 1e-7)$ for α=8.0.

2) Least Squares ID and Certainty Equivalent LQ (IDCE), with or without Priors: Next, we consider a naive Ordinary Least-Squares system identification (ID) process coupled with a standard certainty equivalent LQ control synthesis procedure. In this method, data is collected by simulating trajectories of the dynamical system (1) when Gaussian-distributed noise is injected into the control input. Least-Squares estimates of the system parameters $\hat{A}$ and $\hat{B}$ are made from the trajectories according to:

$$(\hat{A}, \hat{B}) = \text{argmin}_{A,B} \sum_{i=1}^{N} \sum_{t=0}^{T-1} ||x_{t+1}^i - Ax_t^i - Bu_t^i||_2 \quad (4)$$

These estimates are used for the standard LQ controller synthesis procedure:

$$K = -(R + B^T P B)^{-1} B^T P A \quad (5)$$

where P is the solution to the Discrete Algebraic Riccati Equation. This entire procedure can be carried out with a varying degree of prior assumptions about the system parameters A and B. In the "no priors" setting, the full ($\hat{A}$,$\hat{B}$) are estimated from data. In the "priors" setting, only the value of α is estimated, while the rest of the structure of (A, B) is fixed. To test the performance of the controllers, infinite horizon LQ cost are evaluated on the true system (A, B) and the ID'd system $(\hat{A}, \hat{B})$.

The following hyperparameters were used: For the ID process, time horizon of $T = 60$ was used for $\alpha = 0.02$, and $T = 8$ was used for $\alpha = 0.6, 8$; maximum $N = 60$ trajectories were used for $\alpha = 0.02$, while maximum $N = 130$ was used for $\alpha = 0.6, 8$. To observe the amount of data necessary for convergence in performance, the IDCE procedure was repeated on the subset of the whole data set. 100 trials were performed for the procedure above for each α.

III. Results

A. Lower Bounds on Data-Driven Control Performance: You can be stumped

In both DPG and IDCE methods, we find that with a sufficient amount of data, performance converges to the optimal LQ control performance (See Figures 2 and 3). Though perhaps unsurprising, this highlights an important problem: without theoretical bounds to compare performance with, someone using a data-driven method on a plant with intrinsically poor performance (e.g. in the α=0.02 uncontrollable or α=8.0 unstable cases) would not be aware that they were performing optimally, and might blame poor performance on their algorithm, when the plant itself is actually at fault.

In the DPG setting (Figure 2), the convergence rates depend heavily on hyperparameter choices. Our choice was conservative: we required that 100% of controllers produced by the algorithm were stabilizing throughout. When this constraint is relaxed, the controllers which do converge can do so much more quickly. We do see convergence towards the optimum, but due to this hyperparameter sensitivity, we will not comment on its convergence rates.

On the other hand, in the IDCE (with no priors[3]) setting (Figure 3) we see clear trends in convergence that we can

[3]IDCE with prior information performed qualitatively similarly to the no priors case, but with much less data required, as illustrated in the Appendix, Figure 7.

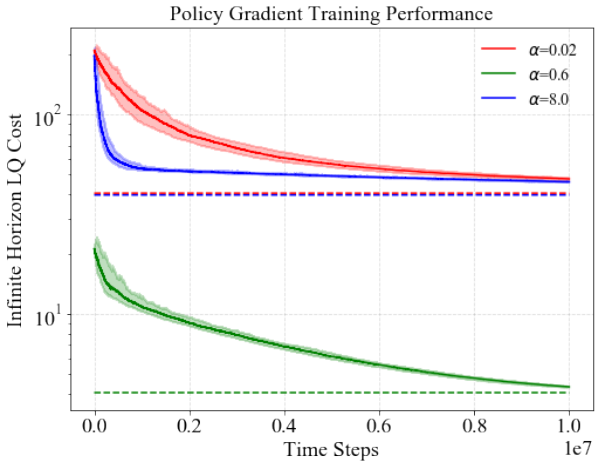

Fig. 2: For each α, the LQ control performance achieved by DPG approaches near optimal (shown in dashed lines) over the course of training. The solid line is the median cost for 100 randomly seeded trials, and the shading indicates 10% to 90% quantiles.

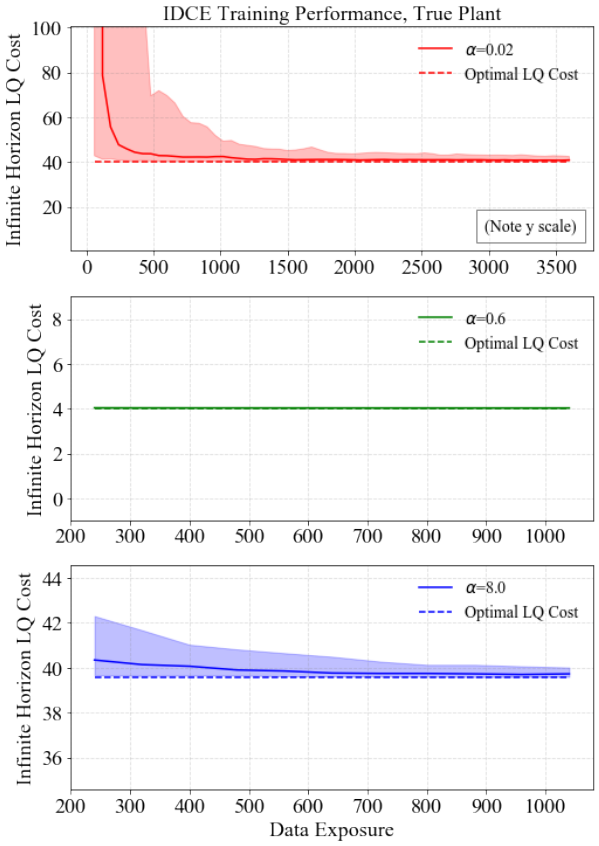

Fig. 3: The true LQ cost achieved by IDCE also approaches optimal when exposed to a sufficient amount of data. Note the very different axes, indicating the relative difficulty of IDing the .02 case. The solid line is the median cost for 100 randomly seeded trials, and the shading indicates 10% to 90% quantiles.

reasonably understand from the system properties. In the benign regime (green), the IDCE method converges quickly to the optimum. In both the unstable (blue) and uncontrollable (red) regimes, we see a slower rate of convergence towards their optimal LQ costs, but much more dramatically so in the uncontrollable case (note the scale of the LQ cost axes). It is important to emphasize however that using rich priors can vastly improve every aspect of this problem (as long as the priors are correct), as we see in the Appendix, Figure 7.

In the in the α=0.02 uncontrollable or α=8.0 unstable cases, IDCE provides a plant estimate $(\hat{A}, \hat{B})$ which can be used to infer the controllability of the true model (by measuring $\sigma_{min}(\Lambda_C)$, the minimum singular value of the Controllability Gramian), to help explain the high LQ cost. Because the model-free DPG setting provides no such estimate, no diagnostic would be available to explain the poor control performance.

B. True cost vs. ID'd cost: Fooled by a smile?

There are two approaches to control performance evaluation: the controller can be evaluated by its performance on the true system's dynamics (in the IDCE setting, we referred to this as the "true LQ performance"), but this is only possible in sufficiently representative model systems, in simulations with sufficient data & computational resources, or in physical experimental systems which can tolerate a sufficient amount of failure for repeated experimentation. Since these requirements may be too restrictive, another approach to performance evaluation is to leverage estimated models of system dynamics to (hopefully) approximate the true cost.

In the IDCE setting, we can measure the performance of controllers using the ID'd system dynamics $(\hat{A}, \hat{B})$ in place of the true (A,B) within the CE control synthesis procedure (5) (referred to as "ID'd LQ performance"). However, an potential problem occurs when using the ID'd performance metric on the IDCE method: we observe cases in which the ID'd cost metric reports costs well below the optimal (see Figure 4).

To explore this phenomenon further, the true LQ cost was plotted against the ID'd LQ cost for the system $\alpha = 0.02$, where the phenomenon was most dramatic. We see that the true vs. ID'd cost forms a "smile" shape, which is approximately quadratic on a loglog scale (see Figure 5, plot 1). The left half of the "smile" contains controllers whose ID'd costs are smaller than their true costs (i.e. plant estimates that are overly optimistic about their control performance), while the right half of the the "smile" contains the reverse (i.e. plant estimates that are overly pessimistic). (See Appendix, section D for a discussion on the origins of the "smile" shape.)

We plotted these "smiles" as a function the amount of IDCE training data (see Figure 5, plots 1-4). As the IDCE is exposed to more data, it improves its estimate of the plant dynamics, and we see that the "smile" curve narrows towards a more accurate estimate of the LQ cost. This, in conjunction with the convergence rates of the benign and

unstable cases in Figure 4, highlights a property of the (nearly) uncontrollable plant: a relatively large amount of data is needed to confidently verify that the controller has poor performance[4].

Even worse, any single plant estimate can be actively misleading, with extreme consequences. Early in the training process, without a sufficient amount of data to ID the dynamics, the IDCE yields many unstable closed-loop controllers (see the black X's in Figure 5, plot 1). Most of the unstable controllers lie far in the left half of the "smile", indicating that the estimated cost is lower than the true optimal cost. This result is particularly alarming: the controllers which appear to be performing best early on are actually performing the worst. a

IV. Conclusion

In this paper, we carried out an empirical study of a simple state feedback system to explore the effects of uncontrollability and instability on the LQ performance achieved by two data-driven methods. We demonstrated that the model-free DPG method was able to produce near-optimal controllers, but argued that it provides no diagnostics for poor performance (in the unstable or uncontrollable settings), and that model-based methods would be required to provide any kind of lower bound for performance. In the model-based IDCE setting, we also explored a practically relevant performance metric, the ID'd LQ cost. We demonstrated that, in an uncontrollable setting, this metric is slow to gain confidence in its system estimate (in comparison to the controllable settings). Using our knowledge of the true LQ performance, we showed a weakness when there is limited data in the ID'd LQ performance metric, which can lead to deceivingly optimistic assessments of bad control performance.

This simple empirical setting has set the stage to think about broader research questions, towards the ultimate goal of providing model-free methods with plant diagnostics (akin to those provided by robust control). Here, we highlight a few specific topics for further exploration:

- In this setting, IDCE works well at efficiently finding good controllers, but could be vastly improved in explaining bad ones, if that objective was considered at the ID stage. Could data-driven performance metrics be developed to give early warning that a plant is bad? For example, given that "smiles" exist, could we leverage that knowledge to better estimate our true cost, particularly when it is bad?
- We used RL here in a purely MF mode, but there are MB versions that might blend well with more Robust-Control-like analysis, see [6] for example.
- We focused on minimal RL and simplified RCT ideas here to maximize accessibility to audiences in both Control and Learning, but related control fields like system ID, adaptive control, fault and failure diagnosis, and model based systems engineering (MBSE) all have vast literatures relevant to diagnosing bad cost control and sophisticated use of data and models. Could we leverage any of this work to provide more general solutions to the problem of identifying bad plants from data?
- The example here had almost trivial mechanisms for high costs and everything depended crucially on state feedback and linearity. Much more interesting hard limits arise in output feedback and also localized, delayed, and distributed settings, and with nonlinearities, most importantly actuator saturation. How could the methods here be upgraded to deal with these generalities?
- One specific toy system that may be fruitful in extending this work to more relevant settings (and which inspired this paper's choice of plant), is the standard stick on

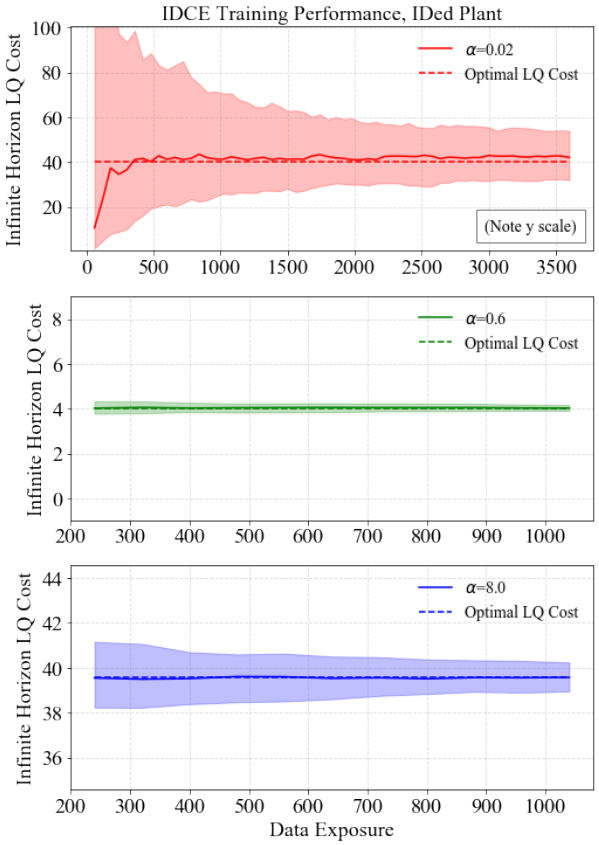

Fig. 4: The ID'd LQ cost achieved by IDCE also approaches optimal when exposed to a sufficient amount of data, but this metric requires much more data to converge than when using the true LQ cost. Additionally, this cost metric, which is meant to approximate the true LQ cost, can be overly optimistic, since it frequently reports costs below optimal (which we know to be impossible). Again, note the extremely different axes in these plots, indicating the extreme relative difficulty of the uncontrollable (low α) case. The solid line is the median cost for 100 randomly seeded trials, and the shading indicates 10% to 90% quantiles.

[4]Intuitively, an uncontrollable state will grow according to the disturbances it experiences in its uncontrollable mode. For systems with Gaussian noise into one uncontrollable mode, the expected state deviation grows like $\sqrt{(t)}$. This relatively slow-growing error signal takes a large amount of time/data to appear.

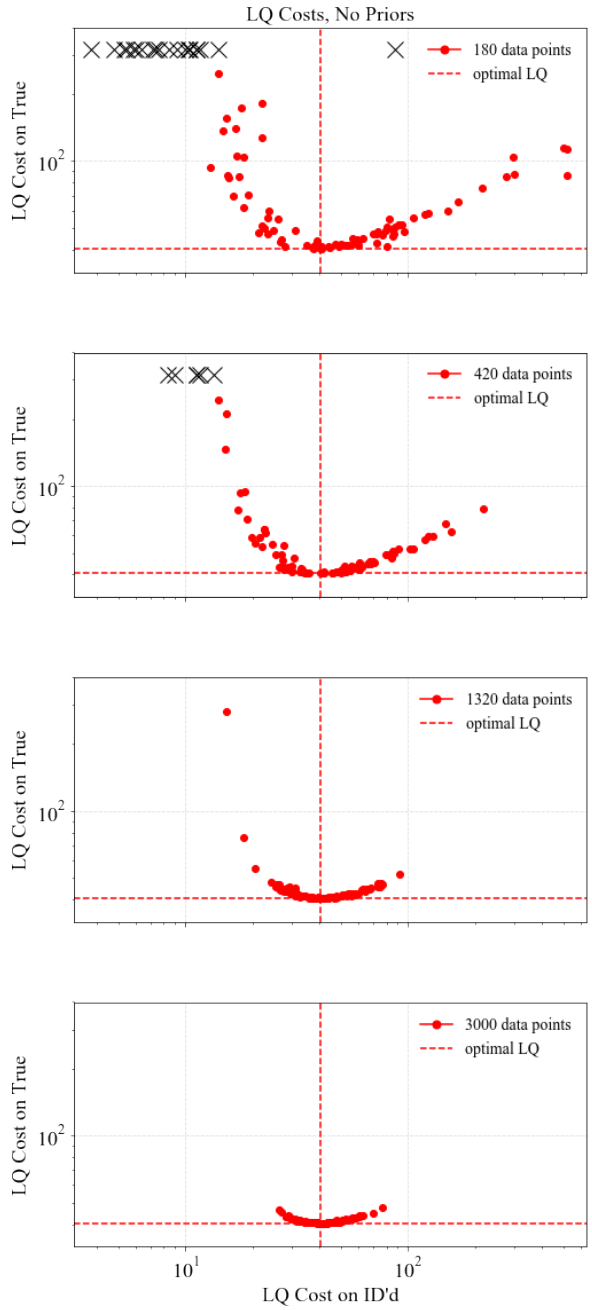

Fig. 5: Comparison of True LQ costs and ID'd LQ costs, for 100 trials of IDCE control synthesis on a near-uncontrollable plant (α=0.02). Each plot shows the performances of each controller K as the amount of training data increases, with the optimal LQ controller's cost shown in a dashed line. Black X's denote infinite true costs for K's that do not stabilize the plant. With enough data, both true and ID'd LQ costs converge towards the true optimal cost (as seen in Figures 3 and 4), but the ID'd cost does not converge as quickly as evident by the spread in the x-direction (the "smile" shape) which persists through all amounts of training data. The presence of any points to the left of the vertical optimal line demonstrates that the ID'd LQ cost metric can be deceivingly optimistic, since the true cost is above the horizontal optimal line.

an actuated cart with enough sensors to have full state feedback. This plant can be made more and more unstable by shortening the stick length, and it can be made uncontrollable by adding a second stick and varying the relative stick lengths. Since this system has analogous configurations to this paper's toy system, and because is is physically amenable to adding properties like delayed or quantized actuation, this provides a logical next step for generalizing this type of problem's results.

References

[1] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*, 2018.

[2] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117, 2019.

[3] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[4] Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

[5] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

[6] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

[7] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sampling complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 2019.

[8] Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for linearized control problems. 2018.

[9] Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 72–80. Citeseer, 1997.

[10] Dhruv Malik, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, , and Martin J. Wainwright. Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems. In *AISTATS*, 2019.

[11] Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again. *arXiv preprint arXiv:1906.11392*, 2019.

[12] Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again, 2019.

[13] Y. Ouyang, M. Gagrani, and R. Jain. Control of unknown linear systems with thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205, Oct 2017.

[14] Anders Rantzer. Concentration bounds for single parameter adaptive control. In *2018 Annual American Control Conference (ACC)*, pages 1862–1866. IEEE, 2018.

[15] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2018.

[16] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends on Machine Learning*, 11(1):1–96, July 2018.

[17] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*, 2018.

[18] Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5012–5021, 2018.

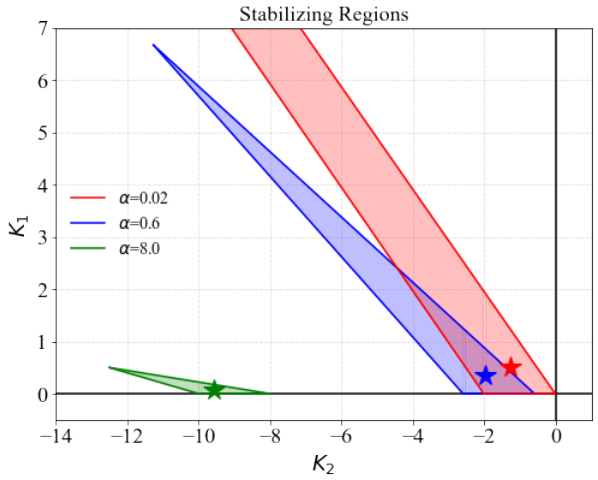

Fig. 6: Stabilizing controllers for the plant under investigation (1), indicated by stars, presented in the space of $K = [K1, K2]$. For reference, the highly uncontrollable (red, α=0.02) and highly unstable (green, α=8.0) plant both have LQ control costs that are 40x worse than the "mild" (blue, α=0.6) plant.

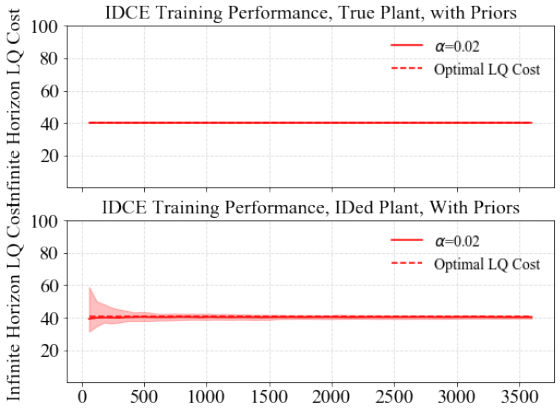

Fig. 7: In the IDCE setting which includes priors on the structure of the plant, both the true LQ cost and the ID'd LQ cost converge towards the optimal, but faster than the IDCE setting without priors. The α=0.6,8.0 figures were omitted, as they both converged so quickly that they appear as lines (much like the true LQ performance plot here). The solid line is the median cost for 100 randomly seeded trials, and the shading indicates 10% to 90% quantiles.

APPENDIX

A. Stable Regions for Controllers K

Using Jury's criterion, the space of stabilizing controllers $K = [K_1, K_2]$ can be derived analytically. The resultant inequalities are

$$\begin{array}{lcl} \alpha K_2 & > & 0 \\ 2K_1 + (\alpha+2)K_2 + 2\alpha + 4 & > & 0 \\ K_1 + K_2 + \alpha K_2 + \alpha & < & 0 \end{array} \tag{6}$$

and are plotted for the three chosen values of α in Figure 6.

B. Impulse Responses for System (1)

To better visualize the dynamics of system (1) with feedback from its optimal controller K^*, we show impulse responses, generated by SciPy's signal processing toolkit. See Figure 9.

C. Performance of IDCE with Priors

Though the performance of the IDCE method in the "with priors" setting is qualitatively similar to the "no priors" setting, we show one example of its performance, for the system with α=0.02 to demonstrate that it is qualitatively similar, but unsurprisingly, that it performs better than the IDCE setting without priors. See Figure 7.

D. "Smile" experiments

When looking at plots of true LQ cost vs ID'd LQ cost, one might initially expect that the true cost should cluster around a vertical line, or perhaps spread in a more uniform distribution above the optimal performance line. The fact that the errors in ID'd cost form such a strong trend suggests that we might be able to better understand the ID'd LQ cost as it relates to some underlying property of IDCE's controller synthesis method.

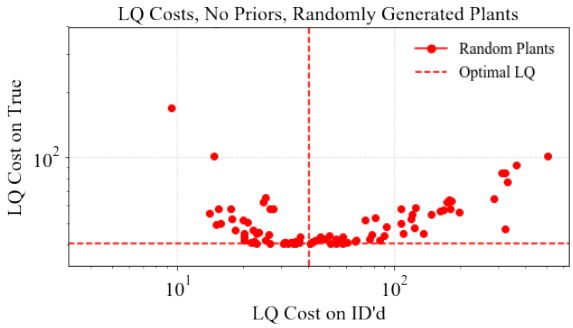

Fig. 8: Comparison of True LQ costs and ID'd LQ costs, for 100 controllers $\tilde{K}$ on a near-uncontrollable plant (α=0.02). $\tilde{K}$'s were generated by adding small Gaussian noise to the true plant matrices, and using these random plants $(\tilde{A}, \tilde{B})$ in the standard control synthesis procedure (5). Since the "smile" trend from Figure 5 persists, we can conclude that the IDCE plant estimation step does not contribute to the form of the "smile", and that this shape must result from the properties of the LQ control procedure.

To explore whether the underlying ID process has any effect on the "smile", we generated plants $(\tilde{A}, \tilde{B})$ according to a normal distribution centered around the true plants, with a variance of 0.005^2, and generated controllers $\tilde{K}$ from each of those systems using the control synthesis procedure (5). We found the same "smile" trend (see Figure 8), indicating that this trend is not intrinsic to the IDCE's ID process, but merely a consequence of the trends in LQ control costs for controllers that are near optimal. See [12], particularly section V.C., for a theoretical discussion of regret (upper) bounds for uncertain plants.

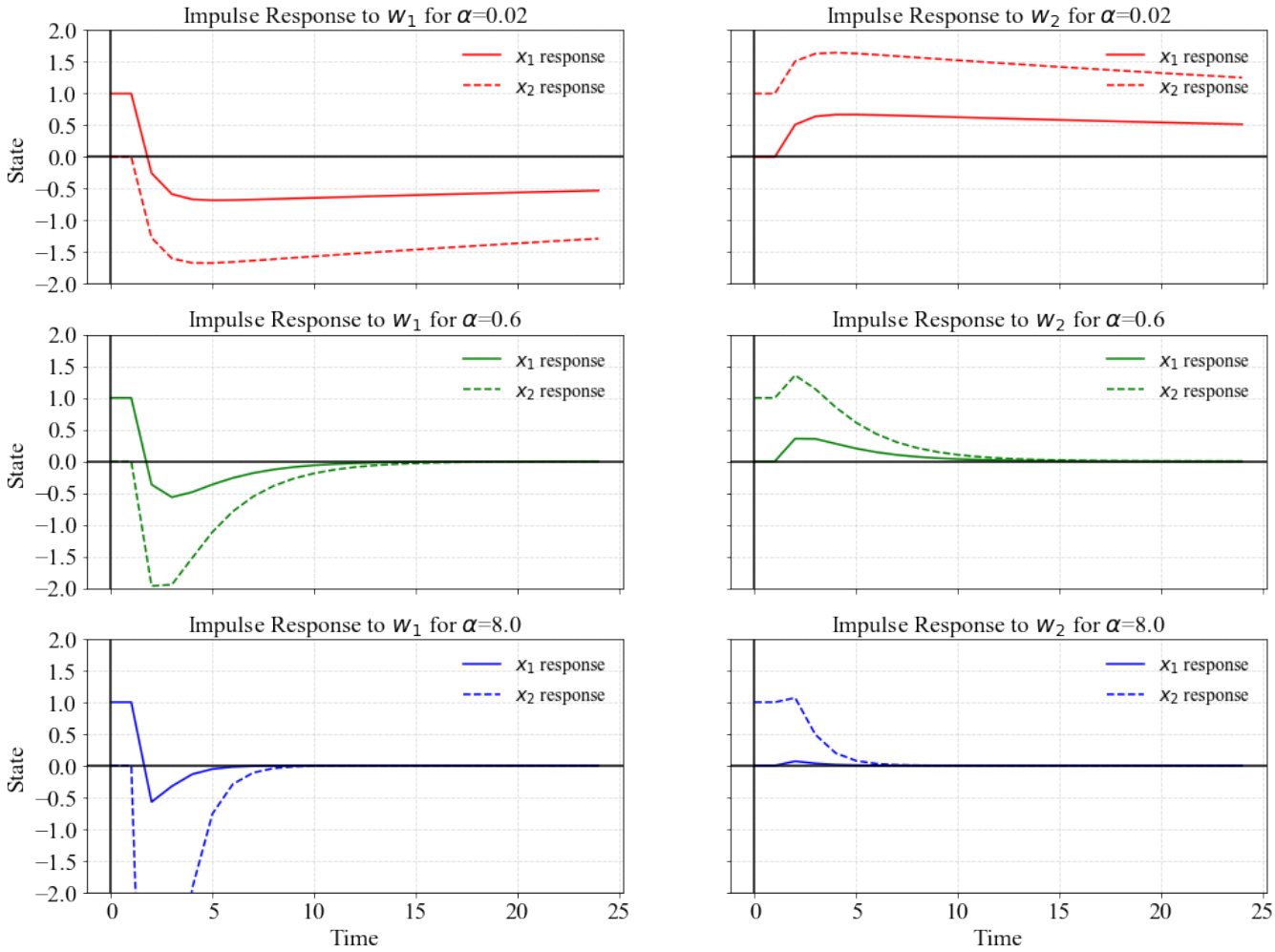

Fig. 9: Closed-loop impulse responses of the system (1) with the optimal controllers for each value of α. w_1 is the impulse into the first state of the system, while w_2 is into the second. The α=0.02 and α=8.0 cases have roughly equal cost but very different transients, and much larger cost than α=0.6. Note that α=0.02 has the same initial transient as α=0.6 but slower settling, which hints that making diagnosis of the large cost might require more data for the α=0.02 plant.