Type of the article (Article, Review, Communication, etc.)

Safety-Critical Policy Iteration Algorithm for Control under Model Uncertainty

Navid Moshtaghi Yazdani1,\*,

Reihaneh Kardehi Moghaddam2 and

Mohammad Hasan Olyaei3

1 PhD, Department of Electrical Engineering, Mashhad branch, Islamic Azad University, Mashhad, Iran

 2 PhD, Department of Electrical Engineering, Mashhad branch, Islamic Azad University, Mashhad, Iran

3 M. Sc., Department of Electrical Engineering, Sadjad University of Technology, Mashhad, Iran.

\*Corresponding Author: Navid Moshtaghi Yazdani, PhD, Department of Electrical Engineering, Mashhad branch, Islamic Azad University, Mashhad, Iran; Email: navid.moshtaghi@ut.ac.ir

**Abstract:** Safety is an important aim in designing safe-critical systems. To design such systems, many policy iterative algorithms are introduced to find safe optimal controllers. Due to the fact that in most practical systems, finding accurate information from the system is rather impossible, a new online training method is presented in this paper to perform an iterative reinforcement learning based algorithm using real data instead of identifying system dynamics. In addition, in this paper the effect of model uncertainty is investigated on control Lyapunov functions and control barrier functions dynamic constraints. The Sum of Square program is used to iteratively find an optimal safe control solution. The simulation results which are applied on a quarter car model show the efficiency of the proposed method in the fields of optimality and robustness.

**Keywords:** safe-critical, optimal controller, reinforcement learning, Lyapunov, Sum-of-Square.

# Introduction

Safety is an integral part and a central requirement for any safe-critical system such as power systems, automatic devices, industrial robots, and chemical reactors. Considering the increasing demand for safe systems in the future generation of industrial systems, and also the importance of an interaction with systems` surroundings and uncertainties, there is a real need for the development of safe controllers, which can meet the already-mentioned demand. In the absence or violation of these safety conditions, the system is likely to suffer from some faults, including the system stabilization problem and its simultaneous survival in the given safety system; which lead to the rise of multiple serious challenges to designing controllers. The optimal control design, as well as the safe control design for the feedback state, are discussed separately in the literature review. Developing such safe controllers to optimize the performance of dynamic systems with uncertainties, primarily resulted from lack of safe optimal controllers with uncertainty conditions.

*1.1. Related works*

The official design for the stabilization of non-linear dynamic systems is often obtained by employing the Control Lyapunov Functions (CLFs). , The optimal feedback controllers necessary for general non-linear systems can be designed by solving Hamilton-Jacobi-Bellman equations (HJB), which have been done approximately by through the use of Galerkin method [1] and neural networks method [2-8]. However, due to the lack of robustness and computational infeasibility for online performance, the open-loop form of calculating these solutions seems problematic. Consequently, in this paper the optimal control of constrained systems equipped with penalty functions in the performance function. [9]. However, the application of these methods is only limited to linear state constraints.

Real-time safety in dynamic systems has gained large attention in [10-14], followed by the introduction of the barrier functions in [15]; through which the risk of the system states entering the given non-safety zones can be removed. Also, control methods using Control Lyapunov function (CLF) and control barrier function (CBF) have been considered as successful methods to achieve safety-stability control. References [16-17] have shown that for the performance of movement tasks (manipulation and locomotion), CLF-based quadratic programs (CLF-QP) with constraints can be solved online. They have also combined CBFs with CLF-QP in [10] to effectively for the effective management of safety constraints in real time. Nevertheless, a detailed knowledge of the system model is required for all these CLF-based and CBF-based methods.

Taylor et.al [18-19] addressed how a minimization method for experimental risk can lead to the uncertainties in CLF and CBF constraints, respectively. Westernbroek et al. [20] have also proposed an RL-based method to learn model uncertainty compensation for the input-output linearization control. In references [21, 22], learning-based control is obtained in dynamical systems with high uncertainty regardless of the safety constraint. Moreover, in references [23-24], probabilistic models such as Gaussian process can be used to learn about model uncertainties. Through the use of these methods, the detailed analysis of the learned model or policy is allowed; however, they can scale poorly with state dimension and using them in high-ordered systems will not be easy.

*1.2. Contributions and Outline*

In [25], a policy iteration algorithm is introduced as a way to build the safe optimal controller for a class of certain nonlinear systems. However, due to the difficulty of practically obtaining accurate system information, an online training method is presented in this study to replace identifying system dynamics with an iterative algorithm featured with real data. In this paper, the effect of model uncertainty is, also, investigated on CLF and CBF dynamic constraints. For each of them, the purpose of the RL agent and the policy to be learned will be defined. The Sum-of-Square program is utilized to iteratively discover an optimal safe control solution. Finally, in order for the efficiency of the proposed method to be validated, a simulation example is employed.

The remaining part of the present paper is organized as follows: Section 2 formulates the problem and presents a new safe optimal control framework. Section 3 presents reinforcement learning for optimal safe control under uncertain dynamics, and Section 4 provides the numerical examples to validate the efficiency of the proposed method.

*1.3. Notations*

The term  denotes the set of all continuous differential functions. Then,  denotes the set of all existing functions in  that are positive, definite and proper. The polynomial  is Sum-of-Squares (SOS) (i.e.,  in whichis a set of SOS polynomials,  where) . Function is an extended class function and. Alludes to the gradient of the V function: .The Li derivative of function  with respect to f is defined as  . For any positive integer and where  ,  is the vector of all distinct monic monomial sets in with minimum degree of  and maximum degree of . Moreover,  represents a set of all polynomials in  with degrees less than and greater than .

# Problem Formulation and details

In this part, we talk about safety, stability and optimization of the control systems. The initial results of each are also mentioned. Then the formulas of the optimal safe control design will be performed.

*2.1. Optimal control of dynamical systems*

Consider the following nonlinear system:

 (1)

In which is the system state vector,  is the control input vector, and  are both locally Lipschitz continuous with  . We expect the system as a stabilizable one.

The main goal of standard optimal control design is to find a control policy to minimize the predefined performance index over the system trajectories (1) defined as follows:

 (2)

In relation (2), ,  and  can be considered as reward function, positive definite function and positive definite matrix, respectively. The reward function is defined such that optimizing (2) guarantees the achievement of control objectives (e.g., minimizing the control effort to achieve the desired transient response) as well as system stability.

The existence of an optimal stabilizing solution is guaranteed under mild assumptions about the system dynamics and reward function [24].

Assumption 1. Considering system (1), there exists a Lyapunov function and a feedback control policy u which satisfies the following inequality:

 (3)

The system stability conditions are guaranteed by this assumption, implying that the cost  is ﬁnite

Theorem 1. [26, Theorem 10.1.2] Consider system (1) with performance function (2), there must be a positive semi-definite function  satisfying the Hamilton-Jacobi-Belman (HJB) equation as follows:



In which

 (4)

Therefore, the following feedback control

 (5)

Optimizes the performance index (2) and results in the achievement of asymptotic stability of the equilibrium. Also, the optimal value function is given as follows:

 (6)

Assumption 1 appears that it is vital to solve the HJB equation (4) to find an optimal control solution.

Assumption 2: There are proper mappings  and, such that and are SOS.

*2.2. About Control barrier functions and its relation with safe control of dynamical systems*

In a safety-critical system, it is important to prevent its state starting from any initial conditions in set to enter some special unsafe regions like. To design a safe controller, control barrier functions (CBF), inspired by Control Lyapunov Function (CLF), can be employed. Now Equation (1) and the function can be considered as follows:

 (7)

The following function is also defined as:

 (8)

 having ZCBF  the admissible control space is defined as follows:

 (9)

The following theorem demonstrates the way a controller is designed using the ZCBF concept to ensure that the forward invariance of the safe set and system stability.

Theorem 2. For  given in (8) and a ZCBF defined by h in (9), each controller for the system (1) presents a safe set L forward invariant.

Proof: see [27]

The barrier functions for exponential controls are introduced in [12]. They are improved in a work by Ams et.al [28].

This translates to the  time-derivative of 



The authors expanded the CBFs having an arbitrary relative degree  to  functions. To do so, we define . As well, we assume that u can be selected so that  for  which is a slack input. We have:



Where,  are,



If a set  is defined as the super level set for an r-times functions which are continuously differentiable , then h is considered as an exponential control barrier function (ECBF) for the control system (1). Therefore, the acceptable space SE(x) (if exists) is defined as follows,



Where, 

As Assumption 3, the admissible control space S (x) can be considered not empty.

# Reinforcement learning for safe optimal control under uncertain dynamics

In this section, the potential mismatch between the model and the plant dynamics is discussed, while there is paucity of accurate knowledge of the true plant vector fields  ,. Moreover, its effects on the dynamics of CLF and CBF will be examined.

Let the nominal model used in the controller be defined as follows:

  (10)

Assume that the vectors  are Lipschitz continuous.and Where, 

Problem 1. (Safe Optimal Control under uncertainty dynamics): Find a controller that solves the following equation:

 (11)

 In relation (11),  is an area in which the system performance is expected to be improved, is the design parameter that acts as a trade-off between the system aggressiveness toward performance and safety, and  is the Stability relaxation factor. Note that δ can be defined as the Aspiration level for a performance that shows the level of performance sacrificed as a result of failure in satisfying safety and performance. However, this parameter is minimized to achieve the highest possible performance.

First, the relaxed optimal control problem for system (1) with performance (2) is examined as follows:

 (12)

In Which is defined by Equation (4) and is an arbitrary compact set containing the origin. Problem 1 actually solves a relaxed version of HJB (4) in which the HJB equation is relaxed with the HJB inequality. In Reference 29, researchers have shown that the solution of problem 1 is unique and if  is a solution for (9), then, through

 (13)

the stability of the system is guaranteed and  plays the role of an upper bound or an overestimate for the actual cost. The superscript  is used here to indicate that  is a performance-oriented controller. However, with a safe control policy,  ،  and  are derived to solve the following optimization problem.

 This control policy does not verify system safety.

 (14)

In SOS framework, this optimization problem is defined as follows:



Based on Assumption 1, there is a safe control policy u. Now we can write the control policy as in which is a part of the controller that is applied to optimize performance regardless of safety and  has been added to  in order to guarantee safety.

*3.1. Deriving under uncertainty situation*

Lemma 1: Consider system (10). Suppose that  is a global safe control policy and  is also existed. Then the system (11) is feedforward.

Proof: According to the assumptions 1 and 2,  . Then by sum of squares, we conclude that

 (15)

According to [24, Result 2.11], system (11) is feed forward:

There is a fixed matrix  in which such that . It is also assumed that there is a fixed vector in which so that . Then, the following terms can be defined along with the solutions of the system (11):



Note that two terms and , depend on and. Since there is uncertainty in these terms, we should solve them without identifying and.

For a similar abovementioned pair , we can find a fixed vector  , in which  and is a fixed matrix, such that

 (16)

 (17)

Therefore, and  are calculated to find  and  By substituting equations (16) and (17) in equation (15), we have:

 (18)

By integrating (18) into the time interval [t, t + δt]:

 (19)

Now, and  can be calculated without having accurate information about and  by using real online data.

1) Initial value:

Find the pair  that satisfies Assumption 1. Consider a fixed vector such that , and.

2) Online data collection:

First, apply  to the system and then find an optimal solution for the following SOS program.

 (20)

So, we have  .Then, we can derive the value of and proceed to step 2) where

*3.2. Reinforcement learning for CBFs*

The control rule for the computed input-output linearization has the following form based on the nominal model and :

 (21)

In which μ is also an auxiliary input.

Under the uncertainty situation, it can be written:

 (22)



Where  are



Terms obtained from the mismatch existing between model and plant. It should also be noted that if α, β are zero, we have the same equation as (22).

Using an estimator made of  that in the form 

RL's goal is to learn  policies so that is close to .as much as possible. Thereby, using RL, the uncertainty terms for CBF can be estimated. Therefore, there is a need for designing the reward function to minimize policy estimation errors. Therefore, it can be defined as follows:



The RL factor embraces a policy that considers the uncertainty terms in CBF, which are summed with the SOS constraints as they are extracted from the nominal model, resulting in accurate estimates. One can consider the focal RL problem with the considered reward for a given state  as the summation of the negative objective functions plus an arbitrary penalty (s) selected by the user

 (23)

Where b is the number of CBFs. One can solve RL, using common algorithms

# Applications

The reason of this part is to demonstrate that our proposed system can make possible the critical safe control, even in the presence of uncertain conditions. Two simulation examples are presented in this section in order to approve the efficiency of the proposed model.

 Example1:

Consider the car quarter suspension model shown in figure 1. Its non-linear dynamic is defined as follows [30]. However, it is worth mentioning that while the training experiences or the simulations are operating, the car quarter suspension model is assumed to be under the proper dynamics (given its uncertainties)





 (24)



**Where, , , and  are the car position, velocity, and its mass, respectively. , , and  are also the wheel position, velocity, and their total mass. , , , and  shows the tire hardness, the linear suspension system, the non-linear suspension hardness, and the damping rate of the suspension system, respectively.

*Figure 1- Quarter car model*

The uncertainty for the significant model in this experiment is introduced by weighing all the components with a weighing coefficient of 2. During the training (process) of the RL agent, we only know the nominal model.

Let, , , ,, , .Then, it can be easily observed that the system establishment has been done in a global level asymptotically, with an absence of input control. The purpose of the proposed method is to design an active suspension control system which reduces the performance index, while retains the global asymptote stability, simultaneously. As well, reducing the disorder effects in the set can improve the system performance.

The reinforcement learning factor is taught using a Deep Deterministic Policy Gradient algorithm (DDPG, Silver et al. [31]). The 4 observed state variables, and the CBF component of the simulation constitute the inputs for the actor neural network. The output dimension is equal to which corresponds to, and.

There exist hidden layers as wide as 200 and 300 in both the actor and the critic neural networks in example 1. This agent is trained by simulation in the interval between  , and  .

A time step of  is employed (in this regard). The simulations have been carried out on a 6-core laptop with Intel Core™ i7-9400 (2.7 GHz) processor and 4 GB RAM.

Use SOSTOOLS [32] to obtain an initial cost function, V0 for the simulated system having non-determined parameters.

Then, we apply the proposed method in which u1=0. The primary condition has been selected randomly. To do the training, we apply the noise from to  till the convergence is obtained after 8 repetitions.

The obtained control policy is as follows,

 (25)

To test the trained controller, we choose the road disorder as a single-impact as follows,

  (26)

In addition, as an indication of a car carrying a load, an overweight of 260 kg is applied to the vehicle assembly.

So that,the departure of position is relative to the origin. The proposed control policy performance is compared to the primary system performance without any control, as shown in figure 2. In figure 3, these two performances of the costs are compared by the constraint wheel position, wheel velocity When they are zero. As can be seen,  has been reduced significantly compared to  .

*Figure 2- Comparison of performance car position and car velocity*

*Figure 3- Comparison of performance wheel position and wheel velocity*

**

*Figure 4- Comparison of learned value functions*

Example2:

Now consider the following system equations:

 (27)

In which  are uncertain parameters, and  and  are mode and system control, respectively. The unsafe space was coded with a polynomial inequality 

With the following details:



Using SOS techniques, it has been shown in [33] that the following robust control policy can stabilize the system (27) at the source level globally and asymptotically.

 (28)

However, the optimality of the closed-loop system has not been fully addressed.

The primary goal of the control is to find more improved safeguard policies under uncertainty using the iterative safeguard policy algorithm. By solving the following feasibility study using SOS-TOOLS we have:

 (29)

The  function is obtained as follows:



If we put  and , the initial condition is arbitrarily set to  and .

  (30)

The  function is as follows:



The indefinite cost function and the initial cost function are compared in Figure 5.

*Figure 5- Comparison of learned value functions*

Both operator and critical neural networks in example 2 have hidden layers with a width of 100 and 200. The training environment using the learning environment was the same as the previous example, the proposed method was learning took 2 seconds per episode. Control policy Obtained after 5 episodes.

In addition, the safe set is equal to:



In which:

 (31)

Note that it is necessary for the safe set to be a member of the complementary set of the unsafe set, as well as being invariable in a way that it never leaves the set in the future. The safe set is obtained using CBF  . Be attention that barrier certificate is bounded to a second-order polynomial. In Figure 6, the estimated safe sets for both the initial control policy and the optimal control policy are shown.

*Figure 6- Estimated safe area using the proposed optimal safe controller in the presence of uncertainty*

# Conclusions

A safe optimization is proposed for the control of dynamics systems under model uncertainty. In order for the performance and safety to be guaranteed, a Hamilton-Jacobi-Bellman (HJB) inequality replaces the HJB equality; besides, a safe policy iteration algorithm is presented certifying the safety of the improved policy and finding a value function corresponding to it. Also, the RL factor was also presented in the proposed method to reduce model uncertainty. The effectiveness of the proposed method is illustrated through two simulation examples.

# References

1. Beard, R. W., Saridis, G. N., & Wen, J. T. (1997). Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation. Automatica, 33(12), 2159-2177. [https://doi.org/10.1016/S0005-1098(97)00128-3](https://doi.org/10.1016/S0005-1098%2897%2900128-3).
2. K. G. Vamvoudakis and F. L. Lewis, “Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem, ”Automatica, vol. 46, no. 5, pp. 878 – 888, 2010.
3. F. L. Lewis and K. G. Vamvoudakis, “Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, pp. 14–25, Feb 2011.
4. B. Kiumarsi and F. L. Lewis, “Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems,” IEEE Transactions on Neural Networks and Learning Systems, vol. 26, pp. 140–151, Jan 2015.
5. H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, “Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems,” Automatica, vol. 50, no. 1, pp. 193 – 202, 2014.
6. D. Wang, D. Liu, Y. Zhang, and H. Li, “Neural network robust tracking control with adaptive critic framework for uncertain nonlinear systems,” Neural Networks, vol. 97, pp. 11 – 18, 2018.
7. S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. Lewis, and W. Dixon, “A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems,” Automatica, vol. 49, no. 1, pp. 82 – 92, 2013.
8. W. Gao and Z. Jiang, “Learning-based adaptive optimal trackingcontrol of strict-feedback nonlinear systems,” IEEE Transactions on Neural Networks and Learning Systems, vol. 29, pp. 2614–2624, June 2018.
9. M. Abu-Khalaf and F. L. Lewis, “Nearly optimal state feedback control of constrained nonlinear systems using a neural networks hjb approach,” Annual Reviews in Control, vol. 28, no. 2, pp. 239 – 251, 2004.
10. A. D. Ames, J. W. Grizzle, and P. Tabuada, “Control barrier function based quadratic programs with application to adaptive cruise control,” in 53rd IEEE Conference on Decision and Control, 12 2014, pp. 6271–6278.
11. A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, “Control barrier function based quadratic programs for safety critical systems,” IEEE Transactions on Automatic Control, vol. 62, no. 8, pp. 3861–3876, 8 2017.
12. Q. Nguyen and K. Sreenath, “Exponential control barrier functions for enforcing high relative-degree safety-critical constraints,” in 2016 American Control Conference (ACC), 7 2016, pp. 322–328.
13. M. Z. Romdlony and B. Jayawardhana, “Uniting control Lyapunov and control barrier functions,” in 53rd IEEE Conference on Decision and Control, 12 2014, pp. 2293–2298.
14. X. Xu, P. Tabuada, J. W. Grizzle, and A. D. Ames, “Robustness of control barrier functions for safety critical control,” IFACPapersOnLine, vol. 48, no. 27, pp. 54 – 61, 2015, analysis and Design of Hybrid Systems ADHS.
15. S. Prajna and A. Rantzer, “On the necessity of barrier certiﬁcates,” IFAC Proceedings Volumes, vol. 38, no. 1, pp. 526 – 531, 2005, 16thIFAC World Congress
16. A. D. Ames and M. Powell. “Towards the uniﬁcation of locomotion and manipulation through control lyapunov nctions and quadratic programs”. In Control of CyberPhysical Systems, pages 219–240. Springer, 2013.
17. Galloway, K., Sreenath, K., Ames, A. D., & Grizzle, J. W. Torque saturation in bipedal robotic walking through control Lyapunov function-based quadratic programs. 2015. 323-332. https://doi.org/10.1109/ACCESS.2015.2419630
18. A. J. Taylor, V. D. Dorobantu, H. M. Le, Y. Yue, and A. D. Ames. Episodic learning with control lyapunov functions for uncertain robotic systems. arXiv preprint arXiv:1903.01577, 2019. URL <https://arxiv.org/abs/1903>. 01577.
19. A. J. Taylor, A. Singletary, Y. Yue, and A. D. Ames. Learning for safety-critical control with control barrier functions. arXiv preprint arXiv:1912.10099, 2019. URL <https://arxiv.org/abs/1912.10099>.
20. T. Westenbroek, D. Fridovich-Keil, E. Mazumdar, S. Arora, V. Prabhu, S. S. Sastry, and C. J. Tomlin. Feedback linearization for unknown systems via reinforcement learning. arXiv preprint arXiv:1910.13272, 2019. URL <https://arxiv.org/abs/1910.13272>.
21. J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter. Learning agile and dynamic motor skills for legged robots. Science Robotics, 4(26):eaau5872, 2019.
22. S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-toend training of deep visuomotor policies. J. Mach. Learn. Res., 17(1):13341373, January 2016. ISSN 1532-4435.
23. S. Bansal, R. Calandra, T. Xiao, S. Levine, and C. J. Tomlin. Goal-driven dynamics learning via Bayesian optimization. In 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pages 5168–5173, Dec 2017. doi: 10.1109/CDC.2017.8264425.
24. J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. IEEE Transactions on Automatic Control, 64 (7):2737–2752, July 2019. ISSN 2334-3303. doi: 10. 1109/TAC.2018.2876389. URL <https://ieeexplore.ieee>. org/abstract/document/8493361.
25. Prajna, S., & Jadbabaie, A. (2004, March). Safety verification of hybrid systems using barrier certificates. In International Workshop on Hybrid Systems: Computation and Control. Springer, 477-492.
26. [Navid Moshtaghi Yazdani](https://ieeexplore.ieee.org/author/37088406652); [Reihaneh Kardehi Moghaddam](https://ieeexplore.ieee.org/author/37073960500),et al,” A Safety-Certified Policy Iteration Algorithm for Control of Constrained Nonlinear Systems”,  [IEEE Control Systems Letters](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=7782633) ( Volume: 4 , [Issue: 3](https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8995821) , July 2020 )
27. F. L. Lewis, D. L. Vrabie, and V. L. Syrmos, Optimal Control. John Wiley, 2012
28. L. Wang, A. Ames, and M. Egerstedt, “Safety barrier certiﬁcates for heterogeneous multi-robot systems,” in 2016 American Contro Conference (ACC), pp. 5213–5218, IEEE, 2016.
29. Ames, A.D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. (2019). “Control barrier functions: Theory and applications. ” In Proc. 2019 European Control Conference. Naples, Italy.
30. Y. Jiang and Z. Jiang, “Global adaptive dynamic programming for continuous-time nonlinear systems,” IEEE Transactions on Automatic Control, vol. 60, pp. 2917–2929, Nov 2015.
31. P. Gaspar, I. Szaszi, and J. Bokor, “Active suspension design using linear parameter varying control,” International Journal of Vehicle Autonomous Systems, vol. 1, no. 2, pp. 206–221, 2003.
32. Silver, David, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. , “Deterministic policy gradient algorithms” 2014.
33. A. Papachristodoulou, J. Anderson, G. Valmorbida, S. Prajna, P. Seiler, and P. A. Parrilo, “SOSTOOLS: Sum of squares optimization toolbox for MATLAB,” 2013. [Online]. Available: <http://arxiv.org/abs/1310.4716>
34. J. Xu, L. Xie, and Y. Wang, “Simultaneous stabilization and robust control of polynomial nonlinear systems using SOS techniques,” IEEE Transactions on Automatic Control, vol. 54, no. 8, pp. 1892–1897, 2009.