Responses to Reviewer Comments for *Optimization Landscape of Policy Gradient Methods for Discrete-time Static Output Feedback*

Jingliang Duan, Jie Li, Xuyang Chen, Kai Zhao, Shengbo Eben Li, Lin Zhao

We would like to thank the editor and reviewers for their constructive and valuable comments to improve the quality of this manuscript (previous paper ID CYB-E-2022-01-0106). Revisions have been made and the main changes are summarized below.

Associate Editor

This paper investigates the static output feedback problem of discrete-time LTI systems with different optimization landscape of policy gradient methods. The paper may contain publishable material, however, the motivation to tackle the static output feedback problem with policy gradient is confused, the contribution is not quite clear, and the demonstration with simple model is not well convinced.

Reply: Thank you for handling our manuscript and for your valuable comments! We have revised the full text accordingly to further clarify our motivation and contribution. Besides, a new numerical experiment based on a four-dimensional circuit system has been carried out to make our theoretical convergence results more convincing. Please see the details in our responses to the reviewers' comments.

Reviewer 1

In this paper, the authors study the optimization landscape of policy gradient methods for SOF control of discrete-time LTI systems. In response to the SOF problems, some interesting properties are proposed. Based on these properties, the convergence of three policy gradient methods is given. In addition, an example is given to verify the theoretical results. The contribution is novel and the theory is sufficient. But there are still some unprofessional and low-level problems. I have the following minor comments before accepting the paper. Some detailed comments include:

Q 1.1 Where is the title of the introduction?

Reply: Thanks for the reviewer's comment! We are really sorry for this low-level mistake. We have revised this problem accordingly.

Q 1.2 The manuscript should be formatted better and some spelling and grammar should be checked carefully, such as in the introduction part "co nsiders" should be "considers".

Reply: Thank you for your helpful suggestion! We are sorry for our careless mistakes in the manuscript. We have carefully checked the full text and made corrections to spelling and grammatical errors found.

Q 1.3 Almost the same statement appears in the abstract and introduction.

Reply: Thanks for the reviewer's comment! To avoid similar statement, we have revised the Abstract and the contribution part of Introduction accordingly.

Q 1.4 The manuscript is greatly related to the field of RL and control theory with a lack of some necessary background materials. For example: The intelligent critic framework for advanced optimal control (Springer); System stability of learning-based linear optimal control with general discounted value iteration.

Reply: Thanks for your valuable suggestion! We have added some additional RL-related studies in Introduction, para. 2, to address this comment:

Optimal control problems have been commonly utilized to help reveal various properties of RL, such as stability [11], [12].

[11] D. Wang, M. Ha, and M. Zhao, "The intelligent critic framework for advanced optimal control," *Artificial Intelligence Review*, pp. 1–22, 2022.

[12] D. Wang, J. Ren, M. Ha, and J. Qiao, "System stability of learning-based linear optimal control with general discounted value iteration," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Q 1.5 Too many meaningless italics are inappropriate to reading. The authors are suggested to revise and polish the paper greatly and to make it a technical one.

Reply: Thanks for your valuable suggestion! We have changed all italics in Lemmas, Theorems, and Propositions to normal for better presentation. Besides, we have also polished the full text carefully to improve readability and technicality.

Q 1.6 There are layout problems with formula (8a) (current equation (13a)).

Reply: Thanks for your valuable suggestion! We have revised the layout of this equation.

Reviewer 2

This paper carries out the analysis of several properties of optimal control for linear time-invariant systems using static output feedback control, and the proofs of several policy gradient methods are presented. The topic is interesting and this paper is well organized. However, the following issues still need to be well addressed.

Q 2.1 One can find that analyzing properties and proving convergences are main tasks in this paper. How to design policy gradient methods is out of their concern. But the authors pointed out "finding the globally optimal SOF controller is a challenging task." when compared with existing convergence results of gradient descent using state feedback control. This is a different thing from the discussion of convergence. What challenges would be met when proving convergence of output feedback control-based policy gradient methods compared with state feedback control should be explored for showing the research value of this paper. Specifically, the innovations or contributions of this paper should be clearly justified by making comparisons with existing convergence results of policy gradient built on state feedback.

Reply: Thanks for the reviewer's comment! We apologize for being unclear. The reason we pointed out that "finding the globally optimal SOF controller is a challenging task" is to show that we cannot expect something better than converging to stationary points for SOF settings. Therefore, in our manuscript, we focus on the analysis of convergence (and the corresponding convergence rate) to stationary points. Note that the convergence rate of SOF is for the stationary point, while the convergence rate of state-feedback LQR is for the globally optimal point.

For the state-feedback LQR, the global optimality and convergence are not unrelated. The global optimality of the state-feedback policy is guaranteed by the gradient dominance condition, which indicates that there are no stationary points except the global minimum [32, 34]. This condition can also facilitate convergence rate identification. Therefore, both the optimality and convergence analysis of state-feedback LQR relies on the gradient dominance condition.

However, due to the disconnectivity of the stabilizing SOF domain and the possible existence of multiple

stationary points, the gradient dominance does not hold for SOF problems, which means the analysis in the SOF setting is more complicated and quite different from the state-feedback LQR. The key component of our analysis is based on the compactness and local L-smoothness properties of the SOF cost. This allows us to use the L-smoothness inequality (27) during the whole learning process and obtain the novel results on convergence (and rate of convergence) to stationary points. Besides, with the help of Lipschitz continuity of Hessian, we further demonstrate that the vanilla policy gradient method will converge linearly to a local minimum if the starting point is sufficiently close to one. To the best of our knowledge, Lipschitz continuity of the Hessian of the SOF cost has not been characterized in related studies.

To further clarify the contribution of this manuscript, we have revised the contribution part on page 2 as

1) Despite the non-convexity of the SOF problem, we are still able to establish several important properties of the SOF cost, including the compact sublevel set, L-smoothness, and M-Lipschitz continuous Hessian. These results will play a key role in the subsequent convergence analysis. To the best of our knowledge, Lipschitz continuity of the Hessian has not been characterized in related studies on both SOF and state-feedback LQR. However, this property is important since it can be used to obtain a stronger result of convergence to a local minimum for SOF problems.

2) The convergence analysis of state-feedback LQR heavily relies on the gradient dominance condition [13, 14, 15, 16, 18, 19]. However, due to the disconnectivity of the stabilizing SOF domain and the possible existence of multiple stationary points, the gradient dominance does not hold for SOF problems, which means the convergence analysis in the SOF setting is more complicated. Upon the compactness and L-smoothness properties of the SOF cost, we finally show that three types of policy gradient methods, including the vanilla policy gradient method, the natural policy gradient method, and the Gauss-Newton method, converge to stationary points at a dimension-free rate if an initial stabilizing policy is given.

3) Besides, with the help of Lipschitz continuity of Hessian, we further demonstrate that the vanilla policy gradient method will converge linearly to a local minimum if the starting point is sufficiently close to one.

added the following text in para. 3, Section I (page 1)

Unlike the state feedback LQR, the domain of the output control gain of SOF can be disconnected, while stationary points in each component can be local minima, saddle points, or even local maxima. Therefore, finding the globally optimal SOF controller using gradient descent is generally intractable. However, it is still of great significance to analyze the optimization landscape of SOF, such as the convergence to stationary points, which will bring new insights for understanding the performance of policy gradient methods on partially observed control problems.

and added Remark 1 on page 5

Remark 1 The coercivity, compactness on the sublevel set, and L-smoothness of the SOF cost, can be deemed as partially observed counterparts to the properties of the state-feedback LQR cost. The associated proofs follow similar steps as the state-feedback LQR case [14, 20]. Different from these properties, to the best of our knowledge, we are the first to establish the M-Lipschitz Hessian property for both SOF and state-feedback LQR problems. These established properties are essential for convergence analysis of policy gradient methods. □

Q 2.2 Three types of policy gradient methods are taken into account, and the explicit form of convergence rate can be derived, then one wonders if the explicit forms of convergence rate for natural policy gradient and Gauss-Newton method can be derived. More discussions or further derivations should be added.

Reply: Thanks for the reviewer's comment! We apologize for the misunderstanding caused. Actually, the explicit forms of the convergence rate of all three policy gradient methods are derived in our manuscript. In particular, Theorem 1 shows the explicit convergence rate of the vanilla policy gradient methods, Theorem 3 shows the explicit convergence rate of the natural policy method, and Theorem 4 shows the explicit

convergence rate of the Gauss-Newton method. In addition to the convergence rate for SOF settings, we also provide the explicit forms of the state-feedback LQR for completeness, which are consistent with the results given in [13, Theorem 7]. We have added more discussions after Theorems 1, 3, and 4 to address this comment:

1. (page 6)

From Theorem 1, given an initial stabilizing SOF controller, one can ensure that the vanilla policy gradient method will stay in the stabilizing set and lead to monotonic policy improvement. In particular, it converges to a stationary point with a dimension-free rate. To provide a unified view of the SOF and state-feedback LQR, results also show that the vanilla policy gradient method globally converges to the single minimum point at a linear rate when the state is completely observed. In this case, the derived convergence rate in (49) is consistent with the result given in [13, Theorem 7]. Besides, different from the result of [13, Theorem 7], we provide an explicit upper bound of the stepsize η such that (52) is satisfied.

2. (page 7)

Theorem 3 shows that the natural policy gradient method also converges to a stationary point for SOF with a dimension-free rate. Besides, the explicit form of the convergence rate (60) for the fully observed case ($C \in \mathbb{C}$) is also provided for completeness, which is consistent with the result given in [13, Theorem 7].

3. (page 7)

Theorem 4 establishes the result of dimension-free convergence to stationary points of the Gauss-Newton method. The explicit form of the convergence rate (63) for the fully observed case ($C \in \mathbb{C}$) is consistent with the result given in [13, Theorem 7].

Q 2.3 This paper demonstrates that vanilla policy gradient method will converge linearly to a local minimum if the starting point is sufficiently close to one. This conclusion seems very strictly, what if the starting point is not close to one, what would happen?

Reply: Thanks for this comment! We apologize for being unclear. In Theorem 1, we prove that the vanilla policy gradient method converges to a stationary point with a dimension-free rate given an initial stabilizing SOF controller. This claim holds for any stabilizing starting point. Therefore, if the starting point is not close to a local minimum, it will at least converge to a stationary point. To address this comment, we have added the following text before Theorem 2 on page 6:

Although the convergence to stationary points of the vanilla policy gradient method for SOF has been established, the converged stationary points can be local minima, saddle points, or even local maxima. Next, we will further show that the vanilla method can converge to a local minimum under mild assumptions.

Q 2.4 The meaning of E in (2) should be given, and whether it means expectation of x0? "The objective of infinite-horizon LQR performance is given by…". This sentence expresses unclear meaning. Actually (2) presents a performance index, one cannot exactly know the objective without minimizing or maximizing.

Reply: Thanks for the reviewer's comment! We apologize for being unclear. $\mathbb{E}_{x_0}$ in (2) represents taking expectation over x_0. And the LQR problem aims to find a control policy to minimize the performance given in (2). To address this comment, we have added the following sentence in the Notation part (page 2)

$\mathbb{E}_x$ means taking expectation over x.

and revised the text before (2) to

The linear quadratic regulator (LQR) problem aims to find a control policy to minimize the infinite-horizon accumulated linear quadratic cost

$$\mathbb{E}_{x_0\sim\mathcal{D}}\Big[\sum_{t=0}^{\infty}(x_t^\top Q x_t + u_t^\top R u_t)\Big].$$

Q 2.5 The equations appeared in this paper should be numbered, or one cannot identify and point out the equations that are put after (6) (current (8)). The second form after (6) (current (8)) cannot be straightforwardly understood, more derivations or explanations should be added.

Reply: Thanks for this valuable suggestion! We have numbered all equations that appear in the main text. To address the unclear expression, the equations after (8) have been revised as

If the initial state correlation matrix is positive definite, i.e.,

$$X_0 := \mathbb{E}_{x_0 \sim \mathcal{D}} x_0 x_0^\top \succ 0,$$

one has that the minimal singular value of X_0

$$\mu := \sigma_{\min}(X_0) \ > 0.$$

Since $\Sigma_K \succeq X_0$, it is straightforward that

$$\sigma_{\min}(\Sigma_K) \geq \mu.$$

Q 2.6 Are there any differences of all lemmas proved in this paper from those in existing literature with state feedback policy gradient methods? More explanations of difficulties or differences need to be presented.

Reply: Thanks for the reviewer's comment! In Lemmas 1-2, we derive the analytical expressions of the policy gradient and Hessian of the SOF cost, which are the basis for the optimization landscape analysis. The proofs of these two lemmas follow similar lines as the state-feedback LQR case [13, 20]. Lemma 3-7 establish several key properties of the SOF cost, which play an important role in the final convergence analysis. Specifically, the coercivity, compactness on the sublevel set, and L-smoothness of the SOF cost, can be deemed as partially observed counterparts to the properties of state-feedback LQR cost. The associated proofs follow similar steps as the state-feedback LQR case [14, 20]. Different from these properties, to the best of our knowledge, we are the first to establish the M-Lipschitz Hessian property for both SOF and state-feedback LQR problems. Compared with the local L-smoothness that can be proved by analyzing the upper bound of the spectral norm of the Hession matrix $\nabla^2 J(\text{vec}(K))$, the M-Lipschitz Hessian need to be established by directly proving $\|\nabla^2 J(\text{vec}(K)) - \nabla^2 J(\text{vec}(K'))\|_F \leq M\|K - K'\|_F$ since the analysis of tensor $\nabla^3 J(\text{vec}(K))$ is very complicated. To address this comment, we have added the the following sentence in para. 1, Section III (page 3)

The derivations follow similar lines as the state-feedback LQR case [13, 20].

and added Remark 1 on page 5

Remark 1 The coercivity, compactness on the sublevel set, and L-smoothness of the SOF cost, can be deemed as partially observed counterparts to the properties of the state-feedback LQR cost. The associated proofs follow similar steps as the state-feedback LQR case [14, 20]. Different from these properties, to the best of our knowledge, we are the first to establish the M-Lipschitz Hessian property for both SOF and state-feedback LQR problems. These established properties are essential for convergence analysis of policy gradient methods. □

Q 2.7 This paper claimed "Lipschitz continuity of the Hessian for policy optimization of LTI systems has not been studied before.", but it also said that this property has also been widely used in [35, 36, 37] to achieve efficient strict saddle point escape. It seems there is contradiction in there. More explanations need to be provided to make it clear.

Reply: Thanks for the reviewer's comment! We are sorry for the misunderstanding caused. Lipschitz continuity of the Hessian is not a new concept. Existing studies [35, 36, 37] have shown that for general non-convex optimization problems, the Hessian Lipschitz property can be used to achieve efficient strict saddle point escape. These studies consider general non-convex optimization problems whose Hessian matrices are known to be Lipschitz continuous. Therefore, these studies are not contradictory to the claim that the

Lipschitz continuity of the Hessian of the SOF cost of LTI systems has not been studied before. To avoid misunderstanding, we have revised the corresponding text before Remark 1 on page 5 as:

To the best of our knowledge, Lipschitz continuity of the Hessian of the SOF cost has not been studied before. However, this result is remarkable since it can be used to obtain a stronger result of convergence to a local minimum for non-convex optimization under mild assumptions [33]. Besides, recent studies [35, 36, 37] have shown that for general gradient-based non-convex optimization problems, the Hessian Lipschitz property can be used to achieve efficient strict saddle point escape.

Q 2.8 Some parameters in three policy gradient methods are not given in Simulation part.

Reply: Thanks for the reviewer's comment! We have added hyperparameters for the employed methods in Section VI.B (page 8) and provided a link to our code in the footnote.

The stepsize of all methods is set as $\eta = 0.2$. Besides, other hyperparamters of Algorithm 1 are set as: $r = 0.001$, $z = 2^{14}$, and $l = 50$[1].

[1]Our code is available at `https://github.com/jieli18/sof`.

Q 2.9 "correlation metrix" should be "correlation matrix".

Reply: Thank you for your helpful suggestion! We are sorry for this careless mistake. We have corrected this mistake and carefully checked the full text.

Reviewer 3

This paper analyzes the optimization landscape of policy gradient methods for solving the problem of static output feedback (SOF) LQR control of discrete-time LTI systems. Several properties of the SOF cost (J(K)), convergence (and rate of convergence) to stationary points are analyzed. Some main issues I am concerned about are as follows:

Q 3.1 What is the benefit of using a policy gradient method to solve the LQR control problem? Especially, the problem of state feedback LQR control for an LTI system is a convex optimization problem, and the optimal controller can be simply obtained by solving an ARE. I don't quite understand why so many researchers are interested in using policy gradient methods to solve the state feedback LQR control problem. It seems that all those papers [13, 14, 15, 16, 17, 18, 19] addressing the LQR control problem with policy gradient methods only gave theoretical and numerical simulation results. I wonder if there exist some practical applications of this kind of method to design LQR controllers in the literature.

Reply: Thanks for the reviewer's comment! We are sorry for the misunderstanding caused. The key point of related papers [13, 14, 15, 16, 17, 18, 19] is not to propose a new method to solve the state-feedback LQR controller. While many policy gradient-based RL algorithms are easy to describe and run in practice, certain theoretical aspects of their behavior, such as the convergence, complexity, and optimality, remain mysterious, even when they are applied in relatively simple settings. Linear quadratic regulator (LQR) is one of the most basic settings. It is well known that the optimal state-feedback gain of LQR can be obtained by solving the algebraic Riccati equation (ARE). **However, from a modern optimization perspective, directly optimizing the state-feedback policy using the LQR cost shown in** (R1) **is generally a non-convex problem:**

$$\begin{aligned} \min_{K\in\mathbb{K}} \quad & J(K) = \mathbb{E}_{x_0\sim\mathcal{D}}\Big[\sum_{t=0}^{\infty}(x_t^\top Q x_t + u_t^\top R u_t)\Big] \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bu_t \\ & u_t = -Kx_t. \end{aligned} \tag{R1}$$

This is because the stabilizing set $\mathbb{K}$ is non-convex for both discrete-time and continuous-time settings (see Example 1).

Example 1 *Let A and B be 3×3 identity matrices and*

$$K_1 = \begin{bmatrix} 1 & 0 & -10 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad K_2 = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

Then the spectra of $A - BK_1$ and $A - BK_2$ are both concentrated at the origin, yet two of the eigenvalues of $A - B\hat{K}$ with $\hat{K} = (K_1 + K_2)/2$ are outside of the unit circle in the complex plane, i.e., $J(\hat{K})$ is infinite.

Due to the lack of convexity of Problem R1, the optimization landscape analysis of the policy gradient method when applied to LQR is non-trivial. Therefore, existing studies [13, 14, 15, 16, 17, 18, 19] choose LQR as a basic testbed to analyze the optimization landscape of policy gradient, which will shed light on the efficiency of policy gradient-based RL in general policy optimization problems. As for this manuscript, we focus on the case of the SOF problem. Compared with state-feedback LQR, the analysis of SOF is more complicated since the stabilizing set of the SOF gain can be even disconnected. To avoid misunderstanding, we add the following text in para. 2, Section I (page 1):

It is well known that the optimal state-feedback gain of LQR can be obtained by solving the algebraic Riccati equation (ARE). However, to understand the performance of policy gradient in solving LQR, we need to directly optimize the linear policy using the LQR cost without solving the associated ARE. This generally leads to a non-convex optimization problem since the domain of stabilizing state-feedback gains can be non-convex [13].

and also added the following text on page 3 (after Assumption 1):

Note that $\mathbb{K}$ can be disconnected and the stationary points of Problem 1 in each component can be local minima, saddle points, or even local maxima [25, 26, 27]; therefore, Problem 1 is generally a non-convex optimization problem and the corresponding convergence analysis is non-trivial.

Q 3.2 For the model-free SOF LQR control case, the system state x is used to calculate J(K). When the system state x is measured, why not use a state feedback controller?

Reply: Thanks for the reviewer's comment! The objective function given in (R2) is widely used in static output-feedback and dynamic output-feedback studies [23, 25]:

$$\min_{u_{0,1,\cdots}} \ \mathbb{E}_{x_0 \sim \mathcal{D}} \Big[\sum_{t=0}^{\infty} (x_t^\top Q x_t + u_t^\top R u_t) \Big]. \tag{R2}$$

Although the system state cannot be completely observed, these studies still aim to learn an output-feedback controller that minimizes the accumulated state-based cost function. For the model-free SOF LQR case, we assume the model parameters, A, B, C, Q, R, are totally unknown. This means that even if we have complete state information (which we do not actually have), we still cannot calculate the running cost by $c_t = x_t^\top Q x_t + u_t^\top R u_t$. A widely-used assumption for the model-free LQR is that we treat the running cost c_t as the output information of the unknown simulation platform [13]. In other words, although x_t, Q, and R are not observed by the controller, the unknown environment will directly output the running cost c_t as a feedback of the control inputs. We have added the following text and Algorithm 1 in Section VI.A on page 8 to address this comment:

In the model-free setting, the model parameters, A, B, C, Q, R, are unknown. In keeping with other work in the literature [13], we assume the algorithm has access to the observation y_t and running cost c_t at each time step, where $c_t := x_t^\top Q x_t + u_t^\top R u_t$. Using the zeroth-order optimization approach [38, 39, 13], Algorithm 1 provides a data-driven procedure to estimate the gradients of both vanilla and natural policy gradient methods.

Q 3.3 How to ensure the stability of the closed-loop system when using the controllers obtained by policy gradient methods?

Algorithm 1 Model-Free Vanilla Policy Gradient and Natural Policy Gradient

Input: stabilizing policy gain K_0, number of trajectories z, roll-out length l, perturbation amplitude r, stepsize η

repeat

Gradient Estimation:

for $i = 1, \cdots, z$ **do**

Sample x_0 from $\mathcal{D}$

Simulate K_j for l steps starting from x_0 and observe $y_0, \cdots, y_{l-1}$ and $c_0, \cdots, c_{l-1}$.

Draw U_i uniformly at random over matrices such that $\|U_i\|_F = 1$, and generate a policy $K_{j,U_i} = K_j + rU_i$.

Simulate K_{j,U_i} for l steps starting from x_0 and observe $c'_0, \cdots, c'_{l-1}$.

Calculate empirical estimates:

$$\widehat{J^i_{K_j}} = \sum_{t=0}^{l-1} c_t,\ \widehat{\mathcal{L}^i_{K_j}} = \sum_{t=0}^{l-1} y_t y_t^\top,\ \widehat{J_{K_{j,U_i}}} = \sum_{t=0}^{l-1} c'_t.$$

end for

Return estimates:

$$\widehat{\nabla J(K_j)} = \frac{1}{z}\sum_{i=1}^{z} \frac{\widehat{J_{K_{j,U_i}}} - \widehat{J^i_{K_j}}}{r} U_i,\ \widehat{\mathcal{L}_{K_j}} = \frac{1}{z}\sum_{i=1}^{z} \widehat{\mathcal{L}^i_{K_j}}.$$

Policy Update:

Vanilla policy gradient $K_{j+1} = K_j - \eta\widehat{\nabla J(K_j)}$.

Natural policy gradient $K_{j+1} = K_j - \eta\widehat{\nabla J(K_j)}\widehat{\mathcal{L}_{K_j}}^{-1}$.

$j = j + 1$.

until $\|\widehat{\nabla J(K_{j-1})}\|_F \leq \epsilon$

Reply: Thanks for the reviewer's comment! In Theorems 1, 3, and 4, we show that the SOF cost $J(K_i)$ is monotonically decreasing during the learning process for all three policy gradient methods. This indicates that if $K_0 \in \mathbb{K}_\alpha$, then $K_i \in \mathbb{K}_\alpha$ for $\forall i \in \mathbb{N}^+$. Since the sublevel set $\mathbb{K}_\alpha \subseteq \mathbb{K}$, it is clear that $K_i \in \mathbb{K}$, i.e., K_i is stabilizing during the whole learning process. To be more clear, we have added the following sentence in Theorems 1, 3, and 4:

$J(K_i)$ is monotonically decreasing (which indicates $K_i \in \mathbb{K}_\alpha \subseteq \mathbb{K}$, i.e., K_i is stabilizing)

Q 3.4 Design algorithms should be given.

Reply: Thanks for the reviewer's comment! As we pointed out in the response to Q 3.1, the focus of this manuscript is to investigate the optimization landscape of existing policy gradient methods when applied to SOF problems, rather than to propose a new algorithm for solving the SOF controller. For completeness, we have added the pseudocode of Model-Free Vanilla Policy Gradient and Natural Policy Gradient in Section VI, page 8 (see the response to Q 3.2 for details). The link to our code for all methods mentioned in Section VI is also provided on the footnote of page 8:

The stepsize of all methods is set as $\eta = 0.2$. Besides, other hyperparamters of Algorithm 1 are set as: $r = 0.001$, $z = 2^{14}$, and $l = 50$[1].

[1]Our code is available at `https://github.com/jieli18/sof`.

Q 3.5 The numerical example is too simple, a model of practical control problem should be used to show the effectiveness of the methods. It would be much more interesting and convincing if the authors could give experimental or practical application results.

Reply: Thanks for this helpful suggestion! We have conducted a new numerical experiment based on a four-dimensional circuit system to address this comment. Due to the page length limit required by the journal, this experiment is shown in the supplementary material:

Consider a circuit system given in [S1] with

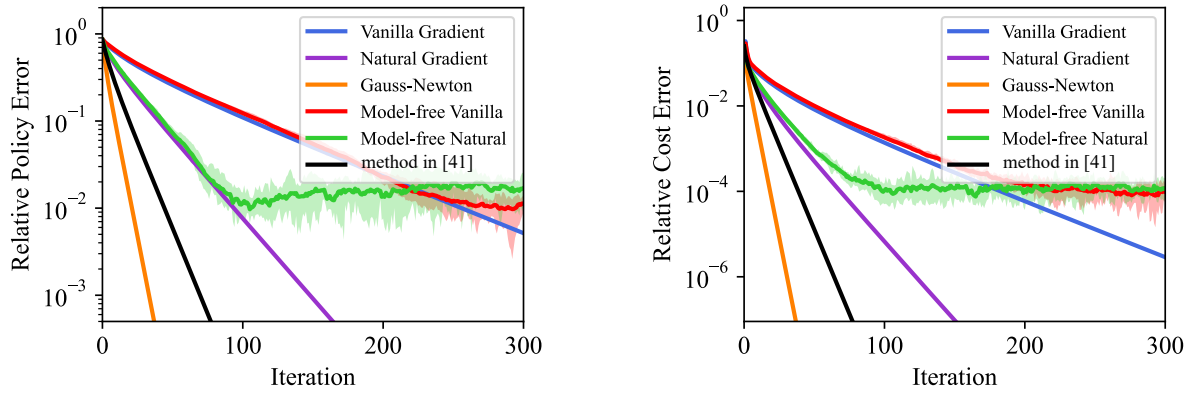
$$A = \begin{bmatrix} 0.90031 & -0.00015 & 0.09048 & -0.00452 \\ -0.00015 & 0.90031 & 0.00452 & -0.09048 \\ -0.09048 & -0.00452 & 0.90483 & -0.09033 \\ 0.00452 & 0.09048 & -0.09033 & 0.90483 \end{bmatrix},$$

$$B = \begin{bmatrix} 0.00468 & -0.00015 \\ 0.00015 & -0.00468 \\ 0.09516 & -0.00467 \\ -0.00467 & 0.09516 \end{bmatrix}, C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

where $Q = \mathrm{diag}([0.1, 0.2, 0, 0])$, $R = \mathrm{diag}([10^{-6}, 10^{-4}])$, and $X_0 = 10I_4$. According to [41, Theorem 1], the optimal gain is

$$K^* = \begin{bmatrix} 2.9738 & -7.2907 \\ 2.1067 & -12.5384 \end{bmatrix}.$$

We set $K_0 = \begin{bmatrix} 0 & -1 \\ 0 & -2 \end{bmatrix}$ for all methods and adopt the same hyperparameters as in Section VI-B. The relative errors of the control gain and the cost function of different methods are shown in Fig. S1. The observed trend of this example is quite similar to the example given in Section VI-B. Overall, the numerical results are consistent with our convergence analysis.



(a) policy error

(b) cost error

Figure S1: Learning curves of different methods. The solid lines correspond to the mean and the shaded regions correspond to interval between maximum and minimum values over 10 runs.

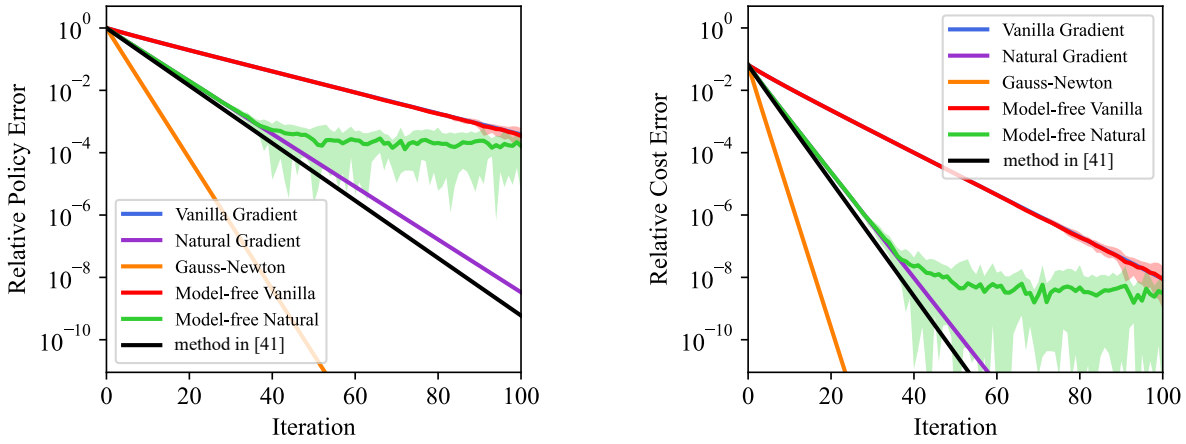
Q 3.6 The convergence rate of the algorithm of [41] should also be shown in the figure for comparison.

Reply: Thanks for the reviewer's valuable suggestion! Actually, the method proposed in [41] is not a policy gradient-based method. For completeness, the learning results of this method are also provided in Section VI.B:

The optimal gain $K^* = 0.3746$ can be found by solving several Lyapunov equations given in [41, Theorem 1]; therefore, the associated model-based method given in [41] for iteratively solving these equations is also employed for comparison. The stepsize of all methods is set as $\eta = 0.2$.

The relative errors of the control gain and the cost function are presented in Fig. 1, which are computed

as $\|K - K^*\|_F / \|K^*\|_F$ and $|J(K) - J(K^*)|/|J(K^*)|$, respectively. We can easily observe that all model-based policy gradient methods converge to the optimal solution within 100 iterations, which demonstrates the convergence results obtained in Section V. As expected, the two model-free methods, especially the model-free nature policy gradient method, converge more slowly and unsteadily than their model-based counterparts due to gradient estimation errors. These results provide numerical evidence for our theoretical convergence analysis.



(a) policy error

(b) cost error

Figure 1. Learning curves of different methods. The solid lines correspond to the mean and the shaded regions correspond to interval between maximum and minimum values over 10 runs.