RL theory project: Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator

Chi-Lun Lin

June 2020

1 Introduction

1.1 Paper overview

Sampling based reinforcement learning algorithms such as policy gradient methods, is little theoretical guarantees to their efficiency. In contrast, control theory has a rich body of tools, with provable guarantees, for related sequential decision making problems, particularly those that involve continuous control. They often estimate an explicit dynamical model first (via system identification) and then design optimal controllers.
This paper leverages the optimal control theory and mathematical optimization, to derive the theoretical guarantee of the policy gradient method.

1.2 Main contributions

This paper proves that while using local search method to deal with a non-convex problem, it may find the globally optimal policy. It is divided into three cases:

- Exact gradient evaluation: This paper shows that descent gradient method indeed converges to the optimal policy.
- Model free cases: Instead of the model, using simulated trajectories in a stochastic policy gradient method is proofed to converge to a globally optimal policy, with polynomially computational and sample complexities.
- The natural policy gradient: This paper shows it improves convergence rate considerably compared too its naive gradient counterpart.

2 Problem Formulation

2.1 Optimal control problem and LQR

The system of the optimal control problem can be described as

$$x_{t+1} = f_t(x_t, u_t, w_t), \tag{1}$$

where x is the state of system, and u, w are referred as the action and the disturbance. Function f maps the state, the action, and the disturbance to the next state x_{t+1}. The objective is to find u_t which minimizes the cost c_t,

$$\text{minimize } \Sigma_{t=0}^{T} c_t(x_t, u_t) \tag{2}$$

$$\text{such that } x_{t+1} = f_t(x_t, u_t, w_t), t = 0...T \tag{3}$$

Considering the linearized control problem, the system is approximated by $x_{t+1} = A_t x_t + B_t x_t + w_t$, and the cost function can be approximated by a quadratic function in state and action. The linear quadratic regulator (LQR) problem is the linearized control problem, combined with homogeneous time, infinite horizon condition, and can be written as

$$\text{minimize } \mathbb{E}[\Sigma_{t=0}^{\infty}(x_t^T Q x_t + u_t^T R u_t)] \tag{4}$$

$$\text{such that } x_{t+1} = Ax_t + Bu_t, x_0 \sim D \tag{5}$$

The initial state is distributed as distribution D. The mactrices A and B are referred as transition mactrices. Q and R are positive definite matrices which parameterize the costs.
Optimal control theory shows that the action can be written as a linear function in the state,

$$u_t = -K^* x_t \tag{6}$$

$$K^* = -(B^T P B + R)^{-}1 B^T P A \tag{7}$$

For the LQR problem, planning P can be achieved by solving the Algebraic Riccati Equation (ARE),

$$P = A^T P A + Q - A^T P B (B^T P B + R)^{-}1 B^T P A \tag{8}$$

2.2 Technical assumption

The paper assumes that cost of state-action map function K ($C(K_0)$) is finite, and does not consider the initial state x_0 with the noise disturbance.

3 Theoretical Analysis

.

3.1 Model-based optimization

Theorem 3.1 (Global Convergence of Gradient Methods) *Suppose* $C(K_0)$ *is finite, and* $\mu > 0$. *Given step size, N, and the update rule, the gradient methods have the performance bound:*

$$C(K_N) - C(K^*) \leq \epsilon \tag{9}$$

- *Gauss-Newton method:*
  The step size is

$$\eta = 1 \tag{10}$$

  , with the update rule:

$$K_{n+1} = K_n - \eta(R + B^T P_{K_n} B)^{-1} \nabla C(K_n) \Sigma_{K_n}^{-1} \tag{11}$$

  , and

$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu} log \frac{C(K_0) - C(K^*)}{\epsilon} \tag{12}$$

- *Natural policy gradient:*
  The step size is

$$\eta = \frac{1}{\|R\| + \frac{\|B\|^2 C(K_0)}{\mu}} \tag{13}$$

  , with the update rule:

$$K_{n+1} = K_n - \eta \nabla C(K_n) \Sigma_{K_n}^{-1} \tag{14}$$

, and

$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu}(\frac{\|R\|}{\sigma_{min}(R)} + \frac{\|B\|^2 C(K_0)}{\mu\sigma_{min}(R)})log\frac{C(K_0) - C(K^*)}{\epsilon} \tag{15}$$

- *Gradient descent:*
  The step size is

$$\eta = poly(\frac{\mu\sigma_{min}(Q)}{C(K_0)}, \frac{1}{\|A\|}, \frac{1}{\|B\|}, \frac{1}{\|R\|}, \sigma_{min}(R)) \tag{16}$$

, with the update rule:

$$K_{n+1} = K_n - \eta\nabla C(K_n) \tag{17}$$

, and

$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu} log\frac{C(K_0) - C(K^*)}{\epsilon} poly(\frac{C(K_0)}{\mu\sigma_{min}(Q)}, \|A\|, \|B\|, \|R\|, \frac{1}{\sigma_{min}(R)}) \tag{18}$$

3.2 Model free optimization

Theorem 3.2 (Global Convergence in the model free setting) *For model free enviroment, given step size, N, the update rules, and the algorithm mentioned in the paper (Algorithm 1: Model-Free Policy Gradient Estimation), it also has the performance bound:*

$$C(K_N) - C(K^*) \leq \epsilon \tag{19}$$

- *Natural policy gradient: The step size is*

$$\eta = \frac{1}{\|R\| + \frac{\|B\|^2 C(K_0)}{\mu}} \tag{20}$$

, with the update rule:

$$K_{n+1} = K_n - \eta\nabla C\hat{(}K_n)\hat{\Sigma^{-1}_{K_n}} \tag{21}$$

, and

$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu}(\frac{\|R\|}{\sigma_{min}(R)} + \frac{\|B\|^2 C(K_0)}{\mu\sigma_{min}(R)})log\frac{C(K_0) - C(K^*)}{\epsilon} \tag{22}$$

- *Gradient descent:*
  The step size is

$$\eta = poly(\frac{\mu\sigma_{min}(Q)}{C(K_0)}, \frac{1}{\|A\|}, \frac{1}{\|B\|}, \frac{1}{\|R\|}, \sigma_{min}(R)) \tag{23}$$

, with the update rule:

$$K_{n+1} = K_n - \eta\nabla\hat{C(K_n)} \tag{24}$$

, and

$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu} log\frac{C(K_0) - C(K^*)}{\epsilon} poly(\frac{C(K_0)}{\mu\sigma_{min}(Q)}, \|A\|, \|B\|, \|R\|, \frac{1}{\sigma_{min}(R)}) \tag{25}$$

4 Conclusion

4.1 Future work

- Variance reduction: This paper proved efficiency from a polynomial sample size perspective. How to further decrease sample size may be an interesting future direction.
- Robust control: In model based approaches, optimal control theory provides efficient procedures to deal with model mis-specification. An important issue is how to provably understand robustness in a model free setting.