ECE 586 BH: Interplay between Control and Machine Learning **Fall 2023**

Lecture 6

Policy-Based Reinforcement Learning for Control, Part II

Lecturer: Bin Hu, Date:09/07/2023

Although policy-based RL has shown great promise for many complex control tasks, its theoretical properties are still underexplored. There are many theoretical questions. For example, we can ask the following questions.

Does the policy gradient method converge provably? If so, to what points? What is the sample complexity to achieve an ε-approximate solution in some certain sense?

Since policy optimization is non-convex in general, the answers to the above questions are non-trivial. Specifically, consider $\min_{K\in\mathcal{K}} \mathcal{C}(K)$ with non-convex $\mathcal{C}$ and $\mathcal{K}$. Then it is non-trivial to show the convergence properties of $K_{l+1} = K_l - \alpha_l \nabla\mathcal{C}(K_l)$. As a matter of fact, it is even non-trivial to show that the gradient method will always maintain the feasibility and stay in the non-convex feasible set $\mathcal{K}$. Very recently, there is a research trend studying the theoretical properties of policy optimization methods on relatively simple linear control benchmarks. In this lecture, we will review a few representative results along this research direction, and list several open questions. More discussions can be found in [3].

6.1 Background: Non-Convex Optimization Theory

Finding stationary points in unconstrained non-convex optimization. Let's first look at the unconstrained optimization setting without worrying about the feasible set $\mathcal{K}$. We will review several basic facts from non-convex optimization theory. In the unconstrained setting, a function $\mathcal{C}(K)$ is L-smooth if the following inequality holds for all (K, K'):

$$\mathcal{C}(K') \leq \mathcal{C}(K) + \langle \nabla\mathcal{C}(K), (K' - K)\rangle + \frac{L}{2}\|K' - K\|_F^2.$$

For L-smooth functions which are bounded below by a finite number, the gradient descent method $K_{l+1} = K_l - \alpha_l \nabla\mathcal{C}(K_l)$ with a constant stepsize $\alpha_l = \alpha < \frac{2}{L}$ can be guaranteed to find an ε-approximate stationary point within $O\left(\frac{1}{\varepsilon^2}\right)$ steps. Specifically, we have

$$\begin{aligned}\mathcal{C}(K_{l+1}) &\leq \mathcal{C}(K_l) + \langle \nabla\mathcal{C}(K_l), K_{l+1} - K_l\rangle + \frac{L}{2}\|K_{l+1} - K_l\|_F^2\\ &= \mathcal{C}(K_l) + \left(-\alpha + \frac{L\alpha^2}{2}\right)\|\nabla\mathcal{C}(K_l)\|_F^2.\end{aligned}$$

Summing the above inequality from $l = 0$ to N

$$\left(\alpha - \frac{L\alpha^2}{2}\right)\sum_{l=0}^{N}\|\nabla\mathcal{C}(K_l)\|_F^2 \leq \mathcal{C}(K_0) - \mathcal{C}(K_{l+1})$$

If $\alpha < \frac{2}{L}$, then $D = \alpha - \frac{L\alpha^2}{2} > 0$. We know $\mathcal{C}(K_{l+1}) \geq \mathcal{C}^*$ for some $\mathcal{C}^*$. Then we must have

$$\sum_{l=0}^{N} \|\nabla \mathcal{C}(K_l)\|_F^2 \leq \frac{\mathcal{C}(K_0) - \mathcal{C}^*}{D}$$

$$\Longrightarrow \min_{0 \leq l \leq N} \|\nabla \mathcal{C}(K_l)\|_F^2 \leq \frac{1}{N+1} \sum_{l=0}^{N} \|\nabla \mathcal{C}(K_l)\|_F^2 \leq \frac{\mathcal{C}(K_0) - \mathcal{C}^*}{D(N+1)}.$$

To find a point whose gradient norm is not bigger than ε, we need to run N steps with

$$N = \frac{\mathcal{C}(K_0) - \mathcal{C}^*}{D\varepsilon^2} - 1 = O\left(\frac{1}{\varepsilon^2}\right).$$

which is the worst-case iteration complexity for finding ε-approximate stationary point in the non-convex smooth setting. As a matter of fact, one can further show that the gradient descent method with the diminishing stepsize $\alpha_l = \frac{1}{(l+1)L}$ is guaranteed to convergence to a stationary point in this setting. If $\alpha_l \leq \frac{1}{L}$, we have $\frac{\alpha_l}{2} \leq \alpha_l - \frac{L\alpha_l^2}{2}$, and the following inequality holds

$$\sum_{l=0}^{\infty} \alpha_l \|\nabla \mathcal{C}(K_l)\|_F^2 \leq 2(\mathcal{C}(K_0) - \mathcal{C}^*).$$

If $\|\nabla \mathcal{C}(K_l)\|_F$ does not converge to 0, then we can use the fact that $\sum_{l=0}^{\infty} \alpha_l = \infty$ to prove that the above inequality cannot hold, leading to a contradiction. Combining the above argument with the fact that the iterates generated by the gradient descent method form a Cauchy sequence (this can be proved using the special form of the gradient descent method), we can show that the gradient descent method with the diminishing stepsize must converge to a stationary point of $\mathcal{C}$.

What if we can show stationary is global minimum? There are many non-convex optimization problems where all the stationary points are global minimum. Then the gradient descent method with the diminishing stepsize is guaranteed to converge to a global minimum! Notice that $\{\mathcal{C}(K_l)\}_{l=0}^{\infty}$ is monotonically decreasing, and hence has a unique limit point. By the properties of the gradient descent method and the diminishing stepsize rule, this limit point has to be equal to the cost value of some stationary point. If every stationary point is global, then this limit point has to be the global optimal value $\mathcal{C}^*$. And hence we must have $\mathcal{C}(K_l) \to \mathcal{C}^*$ monotonically. Many global convergence results in non-convex optimization boil down to proving that every stationary point is global.

From the unconstrained setting to the constrained setting. For policy optimization, it is quite often that we need to pose constraints on the controllers. The constrained set is typically non-convex, and hence we cannot just use projection. Can we just apply the gradient descent method (without projection) to achieve the same convergence result as in the unconstrained setting? If the cost function is a barrier function on the feasible set by itself, then the answer is yes! In this case, the gradient descent method (without projection) will maintain feasibility via decreasing the cost value, and all the above arguments for unconstrained optimization still work! Now we formalize this fact. Notice that a function $\mathcal{C}$ on a constrained feasible set $\mathcal{K}$ is coercive if for any sequence $\{K_l\}_{l=1}^{\infty} \subset \mathcal{K}$, we have $\mathcal{C}(K_l) \to +\infty$ when $\|K_l\|_F \to +\infty$, or K^l converges to an element on the boundary $\partial\mathcal{K}$. The following formal result (adopted from [3]) is useful.

Proposition 1. *Suppose $\mathcal{C}(K)$ is coercive over $\mathcal{K}$. Assume further that $\mathcal{C}$ is twice continuously differentiable over $\mathcal{K}$. Then the following statements hold:*

1. *The sublevel set $\mathcal{K}_\gamma := \{K \in \mathcal{K} : \mathcal{C}(K) \le \gamma\}$ is compact for any $\gamma \ge \mathcal{C}^*$.*

2. *The function $\mathcal{C}(K)$ is L-smooth on $\mathcal{K}_\gamma$, and the constant L depends on γ and the problem parameters. Specifically, for any (K, K') satisfying $tK + (1-t)K' \in \mathcal{K}_\gamma$ $\forall t \in [0,1]$, the following inequality holds*

$$\mathcal{C}(K') \le \mathcal{C}(K) + \langle \nabla\mathcal{C}(K), (K' - K)\rangle + \frac{L}{2}\|K' - K\|_F^2.$$

3. *Consider the gradient descent method $K_{l+1} = K_l - \alpha_l \nabla\mathcal{C}(K_l)$. Suppose $K_0 \in \mathcal{K}$. Let $\gamma_0 = \mathcal{C}(K_0)$. Suppose L is the smoothness constant of $\mathcal{C}(K)$ on $\mathcal{K}_{\gamma_0}$. Then, for any constant stepsize $0 < \alpha < \frac{2}{L}$, we have $K_l \in \mathcal{K}_{\gamma_0}$ for all l, and the gradient descent method is guaranteed to find an ε-approximate stationary point in $O\left(\frac{1}{\varepsilon^2}\right)$ steps.*

4. *If every stationary point of $\mathcal{C}$ is actually a global minimum, then the gradient descent method with the diminishing stepsize $\alpha_l = \frac{1}{L(l+1)}$ is guaranteed to stay in the feasible set and eventually find the global minimum of $\mathcal{C}$. In addition, we have $\mathcal{C}(K_l) \to \mathcal{C}^*$ monotonically.*

We briefly discuss some key proof ideas here. Statement 1 is just a consequence of the coercive property and makes perfect sense geometrically. Statement 2 is an important result. Since $\mathcal{C}$ is twice continuously differentiable, we know that the function $\|\nabla^2\mathcal{C}(K)\|$ (with $\|\cdot\|$ being the operator norm) is continuous. By Weierstrass theorem, we know that $\|\nabla^2\mathcal{C}(K)\|$ has to be bounded on the compact set $\mathcal{K}_\gamma$. We denote this uniform upper bound as L, and hence J is L-smooth on $\mathcal{K}_\gamma$ (the associated inequality can be proved using mean value theorem). Statements 3 and 4 can be viewed as the constrained versions of the unconstrained convergence results that we have covered before. The only new thing here is that we need to argue that the gradient descent update directions guarantee the iterations to stay in the compact sublevel set $\mathcal{K}_{\gamma_0}$. This is quite intuitive since $\mathcal{C}$ is a barrier function by itself. See [3] for more details. For the purpose of applying Proposition 1, we need to show $\mathcal{C}$ is coercive, twice continuously differentiable, and satisfies the condition that every stationary point is global. As a matter of fact, we can prove all these properties for the LQR policy optimization problem in a relatively self-contained manner.

6.2 Global Convergence of Policy Gradient on LQR

Now we apply Proposition 1 to show that the gradient descent method is guaranteed to achieve global convergence on the LQR problem. Recall that given a discrete-time LTI system $x_{t+1} = Ax_t + Bu_t$, we can formulate the LQR problem as a policy optimization problem $\min_{K \in \mathcal{K}} \mathcal{C}(K)$ with the following choice of $(K, \mathcal{C}, \mathcal{K})$.

- Decision variable K: K is simply the feedback gain matrix parameterizing the linear state-feedback policy, i.e. $u_t = -Kx_t$ for all t.
- Cost: We have $\mathcal{C}(K) = \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} x_0^\mathsf{T} ((A-BK)^\mathsf{T})^t (Q + K^\mathsf{T} R K)(A-BK)^t x_0 \right]$, which is a function of K. This cost can also be computed as $\mathcal{C}(K) = \operatorname{trace}(P_K \Sigma_0)$, where $\Sigma_0 = \mathbb{E} x_0 x_0^\mathsf{T}$ is the (full-rank) covariance matrix of x_0, and P_K is the solution of the following Lyapunov equation:
$$(A-BK)^\mathsf{T} P_K (A-BK) + Q + K^\mathsf{T} R K = P_K. \tag{6.1}$$
- Feasible set: The feasible set consists of all the stabilizing LTI state-feedback policies, i.e. $\mathcal{K} = \{K : \rho(A-BK) < 1\}$. For any $K \notin \mathcal{K}$, our LQR cost is not well-defined.

Before applying Proposition 1 to the above policy optimization problem, we need to check the following items:

1. **Is $\mathcal{C}$ twice continuously differentiable? The answer is yes.** To see this, notice that the analytical solution of the Lyapunov equation (6.1) can be calculated as $\operatorname{vec}(P_K) = \left(I - (A-BK)^\mathsf{T} \otimes (A-BK)^\mathsf{T}\right)^{-1} \operatorname{vec}(Q + K^\mathsf{T} R K)$. Hence P_K is a rational function of the elements of K. Then we know that $\mathcal{C}$ is a rational function of the elements of K. Therefore, J is real analytical and twice continuously differentiable.
2. **Is stationary global? Yes!** We can use the analytical form of the gradient $\nabla \mathcal{C}$ to verify this. Suppose $K^\dagger$ is a stationary point such that $\nabla \mathcal{C}(K^\dagger) = 0$. Hence we have
$$\nabla \mathcal{C}(K^\dagger) = 2((R + B^\mathsf{T} P_{K^\dagger} B) K^\dagger - B^\mathsf{T} P_{K^\dagger} A) \Sigma_{K^\dagger} = 0 \tag{6.2}$$
where $\Sigma_{K^\dagger} = \sum_{t=0}^{\infty} \mathbb{E}(x_t x_t^\mathsf{T}) \geq \Sigma_0$ is guaranteed to be full rank (we assume that Σ_0 is full rank). So we must have $(R + B^\mathsf{T} P_{K^\dagger} B) K^\dagger - B^\mathsf{T} P_{K^\dagger} A = 0$. This leads to $K^\dagger = (R + B^\mathsf{T} P_{K^\dagger} B)^{-1} B^\mathsf{T} P_{K^\dagger} A$, which can be substituted back to the Lyapunov equation (6.1) to yield the Riccati equation. This shows that $K^\dagger$ must be the global optimal policy for LQR.
3. **Is $\mathcal{C}$ coercive? If Q and R are both positive definite, then one can show the LQR cost is coercive.** See [3] for more details.

Therefore, we can directly apply Proposition 1 to establish the global convergence of the gradient descent method on the LQR problem.

Convergence rate and sample complexity. Arguably the above result establishes the global convergence of policy optimization on LQR via using least amount of LQR cost properties. However, the downside is that the above analysis does not give a satisfying convergence rate. Notice that the global convergence of policy gradient methods on LQR was first established in [1], from which the original results leverage a stronger property of the LQR cost to prove a much faster rate. Specifically, [1] proves that the LQR cost satisfies the following gradient dominance property:

$$\mathcal{C}(K) - \mathcal{C}(K^*) \leq \frac{1}{2\mu}\|\nabla \mathcal{C}(K)\|_F^2, \quad \forall K \in \mathcal{K}, \tag{6.3}$$

where K^* is the optimal policy, and μ is some positive constant. Notice that the gradient dominance property automatically guarantees that the stationary point is global (what happens in (6.3) if setting $\nabla \mathcal{C}(K) = 0$?). However, this property is much stronger, and can be combined with the L-smoothness property to establish a linear convergence rate $\mathcal{C}(K_l) - \mathcal{C}(K^*) \leq (1 - 2\mu\alpha + \mu L\alpha^2)^l(\mathcal{C}(K_0) - \mathcal{C}(K^*))$ (why? verify this by yourself!). In addition, when the model is unknown, [1] provides the first sample complexity result for model-free policy optimization on LQR. See [1] for more details.

6.3 Global Results for $\mathcal{H}_\infty$ State-Feedback Synthesis

Sometimes the cost function for policy optimization can be nonsmooth. This is especially true for robust control tasks that address the worst-case disturbances. The resultant policy optimization problem becomes non-convex and non-smooth. In this setting, we need to use more advanced optimization algorithms such as Goldstein's subgradient method. In this section, we briefly review one such example, namely the policy optimization for $\mathcal{H}_\infty$ state-feedback synthesis. The problem formulation is stated as follows. Consider the LTI system $x_{t+1} = Ax_t + Bu_t + w_t$ initialized at $x_0 = 0$. The design objective of $\mathcal{H}_\infty$ control is to choose $\{u_t\}$ to minimize the quadratic cost $\mathcal{C} := \sum_{t=0}^{\infty}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t)$ in the presence of the worst-case ℓ_2 disturbance satisfying $\sum_{t=0}^{\infty}\|w_t\|^2 \leq 1$ (the constant 1 can be changed to any other positive number). Suppose the state measurement is available. From robust control theory, it suffices to parameterize the controller as $u_t = -Kx_t$ (this fact is non-trivial). Then the $\mathcal{H}_\infty$ state-feedback synthesis can be reformulated as $\min_{K\in\mathcal{K}} \mathcal{C}(K)$ with the cost $\mathcal{C}(K)$ being defined as the following closed-loop $\mathcal{H}_\infty$ norm:

$$\mathcal{C}(K) = \sup_{\omega\in[0,2\pi]} \lambda_{\max}^{1/2}\big((e^{-j\omega}I - A + BK)^{-\mathsf{T}}(Q + K^\mathsf{T} R K)(e^{j\omega}I - A + BK)^{-1}\big). \tag{6.4}$$

The reason is that the above cost actually satisfies

$$\mathcal{C}^2(K) = \max_{\sum_{t=0}^{\infty}\|w_t\|^2\leq 1} \sum_{t=0}^{\infty} x_t^\mathsf{T}(Q + K^\mathsf{T} R K)x_t = \max_{\sum_{t=0}^{\infty}\|w_t\|^2\leq 1} \sum_{t=0}^{\infty}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t).$$

The above cost is well defined only for K satisfying $\rho(A - BK) < 1$. Therefore, this problem yields the same feasible set as the LQR problem, i.e. $\mathcal{K} = \{K : \rho(A - BK) < 1\}$.

Difficulty: Non-smoothness. A new technical challenge here is that this $\mathcal{H}_\infty$ cost (6.4) can be non-differentiable at some important feasible points, e.g., the optimal points, and hence Proposition 1 cannot be applied. From (6.4), we can see that this cost function is subject to two sources of nonsmoothness: the largest eigenvalue for a fixed frequency ω is nonsmooth, and the optimization step over $\omega \in [0, 2\pi]$ is also nonsmooth. We need to use advanced concepts and algorithms from non-convex non-smooth optimization.

Clarke stationary points. For differentiable functions, a stationary point $K^\dagger$ must satisfy $\nabla\mathcal{C}(K^\dagger) = 0$. We need to generalize the concept of stationary point for non-convex non-smooth optimization. Most functions used in engineering are at least locally Lipschitz.[1] By Rademacher's theorem, a locally Lipschitz function is differentiable almost everywhere, and the Clarke subdifferential is well defined for all feasible points. For any $K \in \mathcal{K}$, the Clarke subdifferential is defined as $\partial_C\mathcal{C}(K) := \text{conv}\{\lim_{i\to\infty} \nabla\mathcal{C}(K_i) : K_i \to K,\ K_i \in \text{dom}(\nabla\mathcal{C}) \subset \mathcal{K}\}$, where conv denotes the convex hull. A point $K^\dagger$ is Clarke stationary if $0 \in \partial_C\mathcal{C}(K^\dagger)$. The condition $0 \in \partial_C\mathcal{C}(K^\dagger)$ just generalizes our original condition $\nabla\mathcal{C}(K^\dagger) = 0$. Now the question becomes how to find Clarke stationary points efficiently.

Goldstein's subgradient method for policy optimization. In the setting of non-convex non-smooth optimization, Goldstein's subgradient method provides a nice generalization of the gradient descent method via generating good descent directions in a highly non-trivial manner (finding a good descent direction for nonsmooth optimization is not easy!). The concept of Goldstein subdifferential is particularly relevant. Given a point $K \in \mathcal{K}$ and a parameter $\delta > 0$, the Goldstein subdifferential of $\mathcal{C}$ at K is defined as $\partial_\delta\mathcal{C}(K) := \text{conv}\left\{\cup_{K'\in\mathbb{B}_\delta(K)}\partial_C\mathcal{C}(K')\right\}$ where $\mathbb{B}_\delta(K)$ denotes the δ-ball around K (we implicitly assume $\mathbb{B}_\delta(K) \subset \mathcal{K}$). The minimal norm element of the Goldstein subdifferential generates a good descent direction, i.e. we have $\mathcal{C}(K - \delta F/\|F\|_F) \leq \mathcal{C}(K) - \delta\|F\|_F$, where F is the minimal norm element in $\partial_\delta\mathcal{C}(K)$. This fact has inspired the developments of Goldstein's subgradient method $K_{l+1} = K_l - \alpha_l F_l$ with F_l being the minimal norm element in $\partial_{\alpha_l}\mathcal{C}(K_l)$, and many other implementable variants. Based on the descent property, we can show that if $\mathcal{C}$ is coercive over $\mathcal{K}$, Goldstein's subgradient method with diminishing stepsize can be guaranteed to find Clarke stationary points. If we can further show that any Clarke stationary point is global minimum, then Goldstein's subgradient method can be guaranteed to find global minimum provably. This turns out to be exactly the case for $\mathcal{H}_\infty$ state-feedback synthesis, i.e. the $\mathcal{H}_\infty$ cost is coercive for positive definite (Q, R) and satisfies the condition that every Clarke stationary point is global minimum. Therefore, Goldstein's subgradient method can be directly applied with provable global convergence guarantees. See [2] for more details.

[1] A function $\mathcal{C} : \mathcal{K} \to \mathbb{R}$ is locally Lipschitz if for any bounded $S \subset \mathcal{K}$, there exists a constant $L > 0$ such that $|\mathcal{C}(K) - \mathcal{C}(K')| \leq L\|K - K'\|_F$ for all $K, K' \in S$.

6.4 Open Issues in LQG and Other Problems

For linear state-feedback control problems, the understanding of the theoretical properties of non-convex policy optimization is more or less mature. Now we have a big picture of how policy optimization can achieve global optimality on this type of non-convex problems. See [3] for more discussions. However, for output feedback problems such as LQG and full-order $\mathcal{H}_\infty$ synthesis, there are many open questions. Two key issues are summarized as follows.

- **Is stationary global?** For output feedback problems such as LQG, the stationary points may not be global. Notice that the policy for LQG is a state-space system. When the stationary point is not controllable or observable, then the optimality can be lost. We need to understand why policy optimization methods can avoid this type of stationary points.

- **Is the cost coercive? The answer is also no.** The LQG cost is not a barrier function any more. How do the policy optimization methods maintain feasibility and stay in the feasible set? This is also unclear. Some new theoretical developments are needed to address this issue.

If you want to read more about these topics, see [3, 5, 4, 6].

Bibliography

[1] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018.

[2] X. Guo and B. Hu. Global convergence of direct policy search for state-feedback $\mathcal{H}_\infty$ robust control: A revisit of nonsmooth synthesis with Goldstein subdifferential. *Advances in Neural Information Processing Systems*, 2022.

[3] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123–158, 2023.

[4] B. Hu and Y. Zheng. Connectivity of the feasible and sublevel sets of dynamic output feedback control with robustness constraints. *IEEE Control Systems Letters*, 7:442–447, 2022.

[5] Y. Tang, Y. Zheng, and N. Li. Analysis of the optimization landscape of linear quadratic gaussian (lqg) control. *Mathematical Programming*, pages 1–46, 2023.

[6] Y. Zheng, Y. Sun, M. Fazel, and N. Li. Escaping high-order saddles in policy optimization for linear quadratic gaussian (lqg) control. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5329–5334, 2022.