Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies

Maryam Fazel, Bin Hu, Kaiqing Zhang

Joint with Tamer Başar, Na Li, Mehran Mesbahi, Yang Zheng

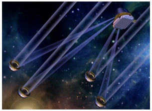
L4DC Tutorial, Philadelphia, PA

June 14, 2023

Motivation

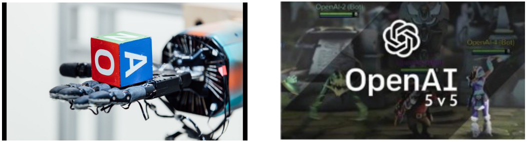
Data-guided decision-making for complex tasks in dynamical systems, e.g., game playing, robotics, networked systems,...

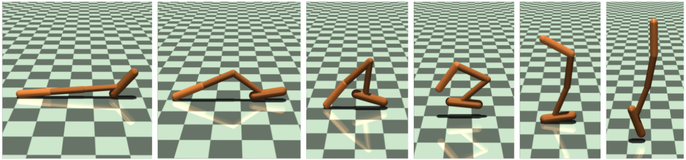
Many recent successes via Reinforcement Learning



Motivation : Policy Optimization

- A workhorse of (deep) RL : (direct) policy optimization methods
- Robotic manipulation, locomotion, video games, ChatGPT, etc.





Policy Optimization – One Workhorse of (Deep) RL

Why is policy optimization popular?

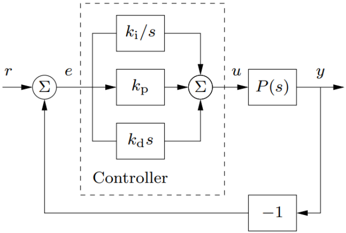
- Easy-to-implement & scalable to high-dimensional problems
- Enable model-free search for complex dynamics (e.g. with rich contact) or rich observations (e.g. images)
- This tutorial : Does policy optimization have guarantees on linear control benchmarks (e.g. LQR, LQG, $\mathcal{H}_\infty$ control, etc)?

What is Policy Optimization (PO) ?

- PO is an old idea from control : Fix the controller structure, and optimize a control metric over the parameters of the controllers

$$\min_K \quad J(K)$$

  - Parametrized controller/policy K
  - Cost function J (tracking errors, closed-loop $\mathcal{H}_2/\mathcal{H}_\infty$ norms, etc)
- Policy gradient method : $K' = K - \alpha \nabla J(K)$
- Example : Optimization-based PID Tuning $K = [K_p, K_i, K_d]^\top \in \mathbb{R}^3$



Credit : Astrom & Murray, 2020

History : Convex LMIs vs. PO

Key points :

- In 1980s, convex optimization methods become dominant due to strong global guarantees and efficient interior point methods
- PO problem formulation is generally not convex
- Reparameterize as convex optimization problems (one does not optimize the controller parameters directly) ; Lyapunov theory, stability/performance certificates, HJB, ...
- Examples of LMIs : state-feedback or full-order output-feedback $\mathcal{H}_2/\mathcal{H}_\infty$ control
- e.g., Boyd *et al.*, "Linear Matrix Inequalities in System and Control Theory", 1994, SIAM

History : Convex LMIs vs. PO

Historically, PO is used for control problems that can't be convexified ; often no theory

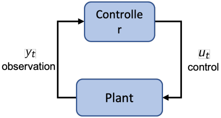
- Sometimes the plant order is unknown : PID tuning, feedback iterative tuning, etc
- Static output feedback LQ control
- Fixed-order structured $\mathcal{H}_\infty$ synthesis : HIFOO and Hinfstruct [Apkarian and Dominikus, '06]
- Distributed control design : Martensson/Rantzer ('09)

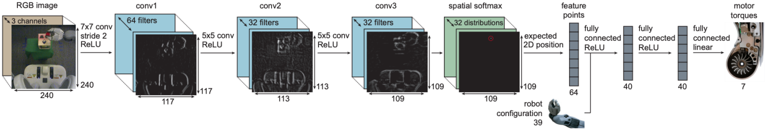
In recent years, new reason to revisit PO for classical control : help provide theory towards understanding model-free RL

A Modern Perspective from Deep RL

A common practice nowadays in deep RL for robotic control : visuomotor policy learning/image-to-torque [Levine et al., '16]

- A type of perception-based control : purely model-free
- Train perception and control systems jointly end-to-end





Advantages :

- Direct and relatively simple to implement
- Mitigate compounding error as in model-based RL (separately train perception and control)
- Make better use of deep NNs' abstraction and perception capabilities to handle high-dimensional visual signals

Policy Optimization : Old & New

Vanilla policy gradient :

- Policy Gradient Theorem [Sutton et al., '99]
$$\nabla J(K) = \mathbb{E}\big[Q_K(x,u) \cdot \nabla \log \pi_K(u \mid x)\big]$$
- REINFORCE estimator [Williams '92] : from N trajectories of length T – $(x_{t,i}, u_{t,i}, c_{t,i})_{i\in[N],t\in[T]}$
$$\nabla J(K) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\Bigg[\underbrace{\left(\sum_{\tau=t}^{T} c_{\tau,i}\right)}_{\text{accumulated cost}} \cdot \underbrace{\nabla \log \pi_K(u_{t,i} \mid x_{t,i})}_{\text{score function}}\Bigg]$$
- Others estimators : G(PO)MDP [Baxter & Bartlett, '01], actor-critic [Konda & Tsitsiklis, '99], natural policy gradient [Kakade '01] (will come back to it !)
- Essentially stochastic gradient descent (SGD) (heart of modern machine learning) !

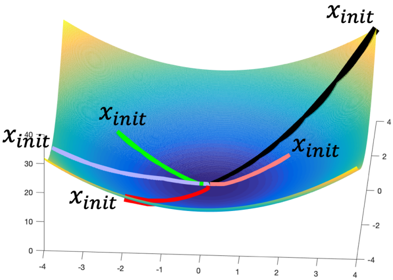
Modern variants (benefit from the advances of optimization theory) :

- Deep deterministic PG (DDPG) [Silver et al., '14], Trust-region PO (TRPO) [Schulman et al., '15], Proximal PO (PPO) [Schulman et al., '17], soft actor-critic (SAC) [Haarnoja et al., '18], variance-reduced PG [Papini et al., '18]...
- Default algorithm in OpenAI Gym, Dota 5v5, ChatGPT training – PPO

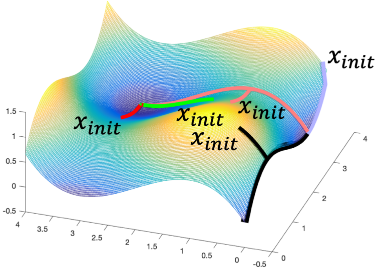
Missing Perspectives in Deep RL Literature

- Convergence guarantees : Nonconvex optimization in policy parameter spaces, e.g., weights of neural networks
- Sample efficiency guarantees : How many samples are needed ? Polynomial in problem parameters ?
- Constraints : Stability and robustness of the closed-loop systems

Missing Perspectives in Classic Control Literature



GD on convex landscape



GD for nonconvex landscape

- Landscape : Is convexity really needed for optimization ?
- Finite-iteration/sample complexity : If an algorithm converges, how fast and how many samples are needed ?

Tutorial Overview

Tutorial Overview : RL/Control $\rightarrow$ PO

This tutorial : Understanding policy optimization via examining guarantees on linear control benchmarks

- Start from simpler contexts and gain insights
  - Classical control benchmarks (c.f. [Recht et al., '17])
- Identify issues for establishing guarantees of PO for control
- Employ modern optimization perspective : iteration/sample complexity, first-order & zeroth-order oracle models, etc

Big picture :

- One perspective to bridge **control theory** and **RL**
- Understand and connect "model-free" & "model-based" views
- Towards a general framework for learning-based control

Schedule

- Now-2 :30pm : Preview and Some Optimization Background
- 2 :30-3 :00pm : PO Theory for LQR
- 3 :00-3 :30pm : PO Theory for Risk-sensitive & $\mathcal{H}_2/\mathcal{H}_\infty$ Robust Control
- 3 :30-4 :00pm : Coffee Break
- 4 :00-4 :30pm : PO Theory for State-feedback $\mathcal{H}_\infty$ Synthesis
- 4 :30-5 :00pm : PO Theory for LQG
- 5 :00-5 :15pm : Role of convex parameterization
- 5 :15-5 :30pm : Future work and Q&A/discussions

Preview : Big Picture

- Revisit linear control problems as benchmarks for PO

$$\min_K \quad J(K), \quad s.t. \;\; K \in \mathcal{K}$$

  - Parametrized policy K (e.g. linear mapping, neural networks)
  - Cost function J (tracking errors, closed-loop $\mathcal{H}_2/\mathcal{H}_\infty$ norms, etc)
  - Constraint set $\mathcal{K}$ (stability, robustness, safety, etc)

- Policy gradient : $K' = K - \alpha \nabla J(K)$
  - The gradient J can be estimated from data in a model-free manner (policy gradient theorem or stochastic finite difference)
  - For nonsmooth problems, replace the gradient with some subgradient

- Recent progress on PO theory (Nonconvexity is the key issue)
  - Landscape : Is stationary global minimum ?
  - Feasibility : Does the policy search stay in the feasible set $\mathcal{K}$?
  - Global convergence & sample complexity

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

Preview : Linear Quadratic Regulator as PO

- Linear quadratic regulator (LQR) as PO : Consider $x_{t+1} = Ax_t + Bu_t + w_t$

$$\min_K \quad J(K) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t)\right], \qquad s.t. \;\; K \;\; \text{is stabilizing}$$

  - $u_t = -Kx_t$ for gain matrix K
  - $\mathcal{K} = \{K : \rho(A - BK) < 1\}$; $\mathcal{K}$ a nonconvex constraint set
- PO theory for LQR
  - Landscape : Feasible set is connected, and stationary is global
  - Feasibility : The LQR cost is coercive and serves as a barrier on $\mathcal{K}$
  - Global convergence & sample complexity : Linear rate and finite sample complexity via the gradient dominance/smoothness property
- Main Ref :

  M. Fazel, R. Ge, S. Kakade, M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator, ICML 2018.

Preview : Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control as PO

- Mixed design : $\mathcal{H}_\infty$ constraints are crucial for robustness

$$\min_K \quad J(K), \qquad s.t. \;\; K \;\; \text{is stabilizing and robust in the } \mathcal{H}_\infty \text{ sense}$$

  - $J(K)$ is an upper bound on the $\mathcal{H}_2$ performance
  - $u_t = -Kx_t$ for gain matrix K
  - $\mathcal{K} = \{K : \rho(A - BK) < 1; \|\mathcal{T}(K)\|_\infty < \gamma\}$; add robustness constrains
  - $\gamma \to \infty$ reduces to LQR

- PO theory for mixed design
  - Key issue : The cost is not coercive ! How to maintain feasibility ?
  - Fix : Implicit regularization via Natural policy gradient (NPG) and Gauss-Newton
  - Global sublinear convergence for NPG and Gauss-Newton

- Main Ref :

  K. Zhang, B. Hu, T. Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee : Implicit regularization and global convergence, *SIAM Journal on Control and Optimization (SICON)*, 2021.

Preview : $\mathcal{H}_\infty$ State-Feedback Synthesis as PO

- $\mathcal{H}_\infty$ state-feedback synthesis : $x_{t+1} = Ax_t + Bu_t + w_t$ with $x_0 = 0$

$$\min_K \quad J(K), \qquad s.t. \ \ K \ \text{ is stabilizing}$$

  - $J = \sum_{t=0}^{\infty}(x_t^\top Q x_t + u_t^\top R u_t)$ subject to the worst-case disturbance satisfying $\sum_{t=0}^{\infty} \|w_t\|^2 \le 1$
  - $u_t = -Kx_t$ for gain matrix K
  - $\mathcal{K} = \{K : \rho(A - BK) < 1\}$
  - $J(K)$ is the closed-loop $\mathcal{H}_\infty$ norm (nonsmooth in K !)
- PO theory for $\mathcal{H}_\infty$ state-feedback synthesis (Nonconvex and nonsmooth)
  - Key issue : The cost may not be differentiable at important points
  - Fix : Show that Clarke stationary points are global, and apply Goldstein's subgradient method to guarantee sufficient descent
  - Global convergence : Goldstein's subgradient method achieves global convergence provably
- Main Ref :

  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, NeurIPS 2022.

Preview : Linear Quadratic Gaussian as PO

- Linear quadratic Gaussian (LQG) is the partially observable variant of LQR, and can be treated as PO (more details later)

- PO theory for LQG
  - Issue 1 : Feasible set may not be connected
  - Issue 2 : Stationary may not be global
  - Today's talk : Some positive results and many open questions

- Main Ref :
  Y. Zheng, Y. Tang, N. Li. Analysis of the optimization landscape of linear quadratic Gaussian (LQG) control, Mathematical Programming, 2023.

Background : Optimization Theory

Optimization of Smooth Nonconvex Functions

Definition : A function $J(K)$ is L-smooth if the following inequality holds for all (K, K') :

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

The above definition is equivalent to ∇J being L-Lipschitz.

Complexity : Gradient descent method $K^{n+1} = K^n - \alpha \nabla J(K^n)$ is guaranteed to find ϵ-stationary point of J within $O\left(\frac{1}{\epsilon^2}\right)$ steps

$$\begin{aligned} J(K^{n+1}) &\leq J(K^n) + \langle \nabla J(K^n), K^{n+1} - K^n \rangle + \frac{L}{2} \|K^{n+1} - K^n\|_F^2 \\ &= J(K^n) + \left(-\alpha + \frac{L\alpha^2}{2}\right) \|\nabla J(K^n)\|_F^2, \end{aligned}$$

Summing the above inequality from $n = 0$ to T

$$\left(\alpha - \frac{L\alpha^2}{2}\right) \sum_{n=0}^{T} \left\|\nabla J(K^n)\right\|_F^2 \leq J(K^0) - J(K^{n+1})$$

Optimization of Smooth Nonconvex Functions

Complexity : Gradient descent method $K^{t+1} = K^n - \alpha \nabla J(K^n)$ is guaranteed to find ϵ-stationary point of J within $O\left(\frac{1}{\epsilon^2}\right)$ steps

$$\left(\alpha - \frac{L\alpha^2}{2}\right) \sum_{n=0}^{T} \left\|\nabla J(K^n)\right\|_F^2 \leq J(K^0) - J(K^{n+1})$$

If $\alpha < \frac{2}{L}$, then $C = \alpha - \frac{L\alpha^2}{2} > 0$. We know $J(K^{n+1}) \geq J^*$ for some J^*.

$$\sum_{n=0}^{T} \left\|\nabla J(K^n)\right\|_F^2 \leq \frac{J(K^0) - J^*}{C}$$

$$\Longrightarrow \min_{0 \leq n \leq T} \left\|\nabla J(K^n)\right\|_F^2 \leq \frac{1}{T+1} \sum_{n=0}^{T} \left\|\nabla J(K^n)\right\|_F^2 \leq \frac{J(K^0) - J^*}{C(T+1)}.$$

To find a point whose gradient norm is smaller than or equal to ϵ, we need to run T steps with

$$T = \frac{J(K^0) - J^*}{C\epsilon^2} - 1 = O\left(\frac{1}{\epsilon^2}\right).$$

which is the complexity for finding ϵ-approximate stationary point

Optimization of Smooth Nonconvex Functions

Complexity : Gradient descent method $K^{t+1} = K^n - \alpha \nabla J(K^n)$ is guaranteed to find ϵ-stationary point of J within $O\left(\frac{1}{\epsilon^2}\right)$ steps

Convergence : Gradient descent method is guaranteed to convergence to a stationary point eventually

Question : What if we can show stationary is global ?

Answer : Then the gradient descent method converges to global minimum ! We have $J(K^n) \to J^*$!

Take-away : Nonconvex optimization may not be that terrifying if stationary is global !

Gradient Dominance and Linear Rate to Global Minimum

Definition : A function $J(K)$ is gradient dominant if it is continuously differentiable and satisfies

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

Landscape : Stationary is global!

Complexity : Gradient descent method $K^{n+1} = K^n - \alpha \nabla J(K^n)$ is guaranteed to find ϵ-optimal point of J within $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ steps

$$\begin{aligned} J(K^{n+1}) &\leq J(K^n) + \left(-\alpha + \frac{L\alpha^2}{2}\right) \|\nabla J(K^n)\|_F^2 \\ &\leq J(K^n) - 2\mu\left(\alpha - \frac{L\alpha^2}{2}\right)(J(K^n) - J^*) \end{aligned}$$

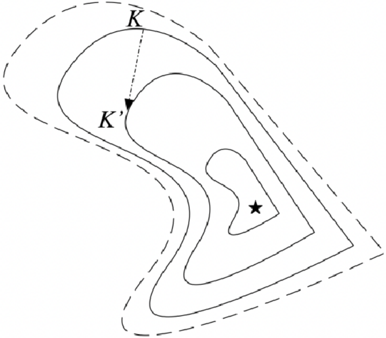
$$\Longrightarrow J(K^{n+1}) - J^* \leq (1 - 2\mu\alpha + \mu L\alpha^2)(J(K^n) - J^*)$$

$$\Longrightarrow J(K^T) - J^* \leq (1 - 2\mu\alpha + \mu L\alpha^2)^T (J(K^0) - J^*)$$

Running T steps with $T = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ guarantees $J(K^T) - J^* \leq \epsilon$

Coercive Functions and Compact Sublevel Sets

What if there are constraints ? If the cost is coercive, then it is a barrier function by itself !



Definition : A function $J(K)$ is coercive on $\mathcal{K}$ if for any sequence $\{K^l\}_{l=1}^{\infty} \subset \mathcal{K}$ we have $J(K^l) \to +\infty$ if either $\|K^l\|_2 \to +\infty$, or K^l converges to an element on the boundary $\partial\mathcal{K}$.

A Useful Result for Constrained Optimization

If J is coercive and twice continuously differentiable on $\mathcal{K}$, we have

- The sublevel set $\mathcal{K}_\gamma := \{K \in \mathcal{K} : J(K) \leq \gamma\}$ is compact.
- The function $J(K)$ is L-smooth on $\mathcal{K}_\gamma$, and the constant L depends on γ and the problem parameters.
- Suppose running GD method $K^{n+1} = K^n - \alpha \nabla J(K^n)$ initialized from $K^0 \in \mathcal{K}$. Let L be the smoothness parameter for $\mathcal{K}_{J(K^0)}$. Then GD finds an ϵ-approximate stationary point with $O\left(\frac{1}{\epsilon^2}\right)$ steps with $\alpha = 1/L$.
- If J is gradient dominant with parameter μ, then GD achieves linear convergence rate.

$$J(K^T) - J^* \leq (1 - 2\mu\alpha + \mu L \alpha^2)^T (J(K^0) - J^*)$$

PO Theory for LQR

Linear quadratic theory – background

Standard LQR problem (discrete-time, infinite horizon) : linear dynamics

$$x_{t+1} = Ax_t + Bu_t$$

with given initial state x_0, choose control sequence

$$u_0, u_1, \dots, u_t, \dots$$

in order to minimize the total cost

$$\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t$$

with given cost matrices $Q, R \succ 0$.

Linear quadratic theory

Classical solution via dynamic programming (when A,B known, stabilizable) : solve the *algebraic Riccati equation* (for P)

$$P = Q + A^T PA - (A^\top PB)(R + B^\top PB)^{-1}(B^\top PA)$$

then let

$$u_t = -K^* x_t = -(R + B^\top PB)^{-1}(B^T PA)\, x_t$$

- a "go-to" model-based control design (since Kalman in 60's)
- extensive theory, computational methods for solving Riccati equation (Laub ; Kleinman '68 ; Hewer '71)

Value and policy iterations

The solution of ARE determines the value matrix

$$\min_u J(x_0, u) = x_0^\top P^* x_0$$

one can develop an iteration on P s. t. $P \to P^*$, then recover the optimal control policy (this would be called value iteration)

PO for LQR, on the other hand, would directly update K, e.g.,

$$K^{n+1} = K^n - \eta \, d_K$$

when d_K is some sort of gradient update and η is (possibly time-varying) stepsize ; this is a first order method

this tutorial : can we develop direct PO methods with guarantees for some typical control synthesis problems ?

Direct policy optimization

towards writing LQR as "$J(K)$" ...

First, note that when A is Schur stable, the sequence

$$\sum_t (A^\top)^t Q A^t \to P \quad \text{converges, where} \quad P = A^\top P A + Q$$

so with stabilizing feedback in place, the LQR cost for the dynamics,

$$x_{t+1} = (A - BK)x_t$$

with an initial condition x_0, can be written as $x_0^T P_K x_0$, where

$$P_K = (A - BK)^\top P_K (A - BK) + K^\top R K + Q$$

so in this case, LQR is really optimizing

$$\min_{P,K} \quad \textbf{trace}\, P\Sigma_0 \qquad (\text{with} \quad \Sigma_0 = x_0 x_0{}^\top)$$

$$P = (A - BK)^\top P (A - BK) + K^\top R K + Q$$

However, as stated, this problem is a bilinear matrix optimization ...

When $K \in \mathcal{S}$ (set of stabilizing K), equation

$$P = (A - BK)^\top P(A - BK) + K^\top RK + Q$$

has a unique solution $P(K)$; hence, the LQR can be written (for a given Σ) as

$$\min_{K \in \mathcal{S}} J(K)$$

We take $x_0 \sim \mathcal{D}(0, \Sigma_0)$ where Σ_0 is a full-rank covariance (equivalently, can take Σ to correspond to a spanning set of initial conditions)
thus J is real analytic function over its domain.

Questions

Consider now PO algorithms :

- iterate on policy K,
- using gradient of cost, $\nabla J(K)$ (exact or approximate)

- does GD (with exact gradients) converge ? under what assumptions ? does it converge to the global opt K^* ?
- rate of convergence ?
- how about related algorithms, e.g., "natural gradient" descent ?
- "model-free" version : if gradients not available, would sampling $J(K)$ work ? finite-sample complexity ?

note : challenging as $J(K)$ is **not convex**

LQR and policy gradient methods

- Consider LQR without state noise (for simplicity), random initial condition $x_0 \sim \mathcal{D}$
- let $J(K)$ be the cost as function of policy K
- define covariance matrices :

$$\Sigma_K = \mathbb{E}\left[\sum_{t=0}^{\infty} x_t x_t^T\right], \qquad \Sigma_0 = \mathbb{E}\left[x_0 x_0^T\right]$$

- consider algorithms :

$$\text{Gradient descent :} \qquad K \leftarrow K - \eta \nabla J(K)$$

$$\text{Natural GD :} \qquad K \leftarrow K - \eta \nabla J(K) {\Sigma_K}^{-1}$$

[Kakade '01]

Oracles

Suppose algorithms have only **oracle access** to the model (A,B not known explicitly), e.g.,

- exact gradient oracle : $\nabla J(K)$
- "approximate gradient" oracle : use sample values of $J(K)$

i.e., first-order and zeroth order oracle in optimization

Optimization landscape I

$$\begin{aligned} J(K) &= \langle \Sigma_0, P_K \rangle \\ &= \mathrm{vec}(\Sigma_0)^T (I - (A - BK) \otimes (A - BK))^{-1} \mathrm{vec}(Q + K^T R K) \end{aligned}$$

observation : $J(K)$ is **not** convex in K (or quasiconvex, star convex) for $n \geq 3$ (convex for single input case when $n = 2$) ; we can computed the $\nabla J(K) = 2(RK - B^\top P_K A_K)\Sigma_K$, where,

$$\Sigma_K = \mathbb{E}\left[\sum_{t=0}^{\infty} x_t x_t^T\right]; \qquad \Sigma_0 = \mathbb{E}\left[x_0 x_0^T\right]$$

lemma : if $\nabla J(K) = 0$, then either

- K is optimal, or
- covariance Σ_K is rank deficient.

if Σ_0 is full rank, then Σ_K is full rank $\Longrightarrow$ stationary points globally optimal

can also show this leveraging the convex LMI reformulation—later research has shown this deeper connection and its uses.)

Optimization landscape I

$$\begin{aligned} J(K) &= \langle \Sigma_0, P_K \rangle \\ &= \mathrm{vec}(\Sigma_0)^T (I - (A - BK) \otimes (A - BK))^{-1} \mathrm{vec}(Q + K^T R K) \end{aligned}$$

observation : $J(K)$ is **not** convex in K (or quasiconvex, star convex) for $n \geq 3$.

$$\Sigma_K = \mathbb{E}\left[\sum_{t=0}^{\infty} x_t x_t^T\right], \qquad \Sigma_0 = \mathbb{E}\left[x_0 x_0^T\right]$$

lemma : if $\nabla J(K) = 0$, then either

- K is optimal, or
- covariance Σ_K is rank deficient.

if Σ_0 is full rank, then Σ_K is full rank $\Longrightarrow$ stationary points globally optimal

can also examine this via transformation to convex LMI (but proofs no simpler)

Optimization landscape II

lemma : Suppose Σ_0 is full rank, then

$$J(K) - J(K^*) \leq \frac{\|\Sigma_{K^*}\|}{\sigma_{\min}(\Sigma_0)\sigma_{\min}(R)} \|\nabla J(K)\|^2.$$

i.e., $J(K)$ gradient-dominated ([Polyak '63],...)

Main Theory : Summary

[Fazel, Ge, Kakade, Mesbahi 2018] for discrete-time LQR

- $J(K)$ is generally **not** convex/quasiconvex/star-convex
- convex combination of two stabilizing K's may not stabilize
- coerciveness : LQR cost is coercive on the set of stabilizing policies
- LQR cost is gradient dominant
- hence, gradient descent converges to K^* from any stabilizing initial K_0 ! (with a linear rate)
- similarly for related algorithms, e.g., **natural policy gradients** and policy iteration algorithm

Global convergence in the exact case

Theorem 1 : Suppose $J(K_0)$ is finite (i.e., K_0 is stabilizing), Σ_0 is full rank. With stepsize η chosen appropriately, and # of iterations N as

- for natural policy GD :

$$N \geq \frac{\|\Sigma_{K^*}\|}{\sigma_{\min}(\Sigma_0)} \left(\frac{\|R\|}{\sigma_{\min}(R)} + \frac{\|B\|^2 J(K_0)}{\sigma_{\min}(\Sigma_0)\sigma_{\min}(R)} \right) \log \frac{J(K_0) - J(K^*)}{\epsilon},$$

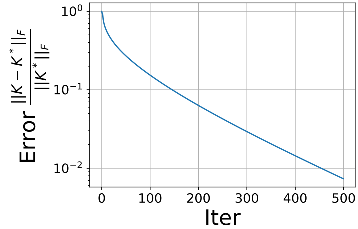
- for GD :

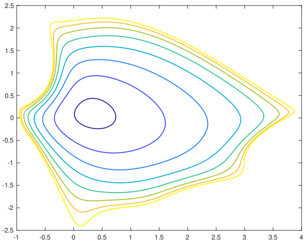
$$N \geq \frac{\|\Sigma_{K^*}\|}{\sigma_{\min}(\Sigma_0)} \log \frac{J(K_0) - J(K^*)}{\epsilon} \ \text{poly}(\text{everything else}),$$

then,

K_N has cost ϵ-close to optimum.

Numerical experiment





Left : Gradient descent for continuous time LQR
Right : 2-dim projection of the LQ cost contour

Unknown model case

$$\text{Gradient descent}: \quad K \leftarrow K - \eta \widehat{\nabla J(K)}$$
$$\text{Natural policy GD}: \quad K \leftarrow K - \eta \widehat{\nabla J(K)} \widehat{\Sigma}_K^{-1}$$

- we do not know (or directly learn) A, B
  - but have the ability to explore by perturbing K
- model-free estimation : add Gaussian noise to actions during rollouts
- similar to zeroth order (derivative-free) optimization
- **issues :** how much noise ? length of rollouts ? overall sample complexity ?

Algorithm

Input : K, # trajectories m, rollout length ℓ, parameter r, dimension d

for $i = 1, \cdots m$,

- draw U_i is uniformly at random from $\|U\|_F \leq r$
- sample policy $\widehat{K}_i = K + U_i$
- simulate $\widehat{K}_i$ for ℓ steps starting from $x_0 \sim \mathcal{D}$.
- get empirical estimates

$$\widehat{C}_i = \sum_{t=1}^{\ell} c_t\,, \quad \widehat{\Sigma}_i = \sum_{t=1}^{\ell} x_t x_t^\top$$

where c_t, x_t are costs and states on this trajectory

end for

use following estimates for PGD/NPGD :

$$\widehat{\nabla J(K)} = \frac{1}{m} \sum_{i=1}^{m} \frac{d}{r^2} \widehat{C}_i U_i\,, \quad \widehat{\Sigma_K} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\Sigma}_i$$

Idea for the Proof of Convergence

1. Prove when rollout length ℓ is large enough, cost function C and covariance Σ are close to infinite horizon quantities
2. Show with enough samples, alg can estimate gradient and covariance matrix within the desired accuracy
3. Show GD and NPGD converge with a similar rate, despite bounded perturbations in gradient/natural gradient estimates

Suppose $J(K_0)$ is finite, $\mu > 0$, x_0 has norm bounded by L almost surely ; and GD and the NPGD are called with parameters :

$$m, \ell, 1/r = \text{poly}\left(J(K_0), \frac{1}{\mu}, \frac{1}{\sigma_{\min}(Q)}, \|A\|, \|B\|, \|R\|, \ \frac{1}{\sigma_{\min}(R)}, d, 1/\epsilon, L^2/\mu\right).$$

- **NPGD :** for stepsize $\eta = \frac{1}{\|R\| + \frac{\|B\|^2 J(K_0)}{\mu}}$ and
$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu}\left(\frac{\|R\|}{\sigma_{\min}(R)} + \frac{\|B\|^2 J(K_0)}{\mu\sigma_{\min}(R)}\right) \log \frac{2(J(K_0) - J(K^*))}{\epsilon},$$
with high probability NPGD satisfies : $J(K_N) - J(K^*) \leq \epsilon$
- **GD :** for appropriate stepsize η,
$$\eta = \text{poly}\left(\frac{\mu\sigma_{\min}(Q)}{J(K_0)}, \frac{1}{\|A\|}, \frac{1}{\|B\|}, \frac{1}{\|R\|}, \sigma_{\min}(R)\right)$$
and
$$N \geq \frac{\|\Sigma_{K^*}\|}{\mu} \log \frac{J(K_0) - J(K^*)}{\epsilon} \text{poly}\left(\frac{J(K_0)}{\mu\sigma_{\min}(Q)}, \|A\|, \|B\|, \|R\|, \frac{1}{\sigma_{\min}(R)}\right),$$
with high probability, GD satisfies : $J(K_N) - J(K^*) \leq \epsilon$

Related Results

A burst of recent research interest :

LQR, continuous-time : [Mohammadi et al., 2019], [Bu et al., 2020]
LQR, discrete-time : [Fazel et al., 2018], [Bu et al., 2019]
Stabilization : [Perdomo et al., 2021]
Decentralized finite-horizon LQR under QI : [Furieri et al., 2020]
LQ games : [Zhang et al., 2019], [Bu et al., 2019], [Mazumdar et al., 2019], [Hambly et al., 2021]
Markov jump linear systems : [Jansch-Porto et al, 2020]
Output estimation with differentiable convex liftings (DCL) framework : [Umenberger et al, 2022]
$\mathcal{H}_\infty$ **control :** [Tang and Zheng, 2023]
Fundamental limits of policy gradient : [Ziemann et al, 2022]

And many other variants ! (See our survey article)

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

Coming up ...

- 3 :00-3 :30pm : PO Theory for Risk-sensitive & $\mathcal{H}_2/\mathcal{H}_\infty$ Robust Control
- 3 :30-4 :00pm : Coffee Break
- 4 :00-4 :30pm : PO Theory for State-feedback $\mathcal{H}_\infty$ Synthesis
- 4 :30-5 :00pm : PO Theory for LQG
- 5 :00-5 :15pm : Role of convex parameterization
- 5 :15-5 :30pm : Future work and Q&A/discussions

Policy Optimization for Risk-Sensitive Control, $\mathcal{H}_2/\mathcal{H}_\infty$ Robust Control, and Linear Quadratic Games

Kaiqing Zhang

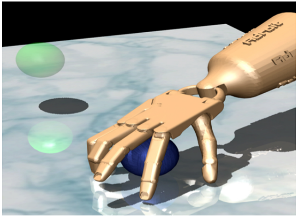
University of Maryland, College Park

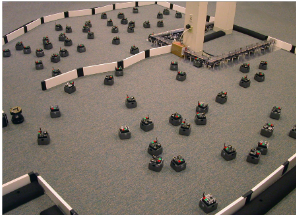
5^{th} Annual Learning for Dynamics & Control Conference (L4DC)

University of Pennsylvania, PA June 14, 2023

General Background: Recall

- Reinforcement learning (RL) has achieved tremendous empirical successes, including in some continuous-space control tasks
  - Game of Go, video games, robotics, etc[1].





- A resurgence of interest in the theoretical understandings of RL

[1] source: google images

Motivation

- Most scenarios involve more than one agent, e.g., game of Go, video games, robot team motion-planning, autonomous driving
- Most control systems are safety-critical, e.g., robotics, unmanned (aerial) vehicles, cyber-physical systems

Goal: Provable *RL* with *robustness* and *multi-agency* considerations

Background

Theoretical Understanding of Policy Optimization

- One workhorse in RL: Direct policy search/policy optimization
- Whether, where, how fast, PO methods converge?
  - Nonconvex in policy parameter space
- Let's start with benchmark RL/control tasks before deep RL?
- PO for linear quadratic regulator (LQR) (and variants) has been extensively studied recently [Fazel et al., '18][Tu & Recht, '19][Malik et al., '19][Bu et al., '19][Mohammadi et al., '19][Gravell et al. '19][Li et al., '19][Furieri et al., '19]...

All models are wrong, so is LQR $\Longrightarrow$ robustness concern is critical

- Question: whether & how PO methods can address benchmark control/RL with robustness/risk-sensitivity concerns?
- Side motivation: multi-agent RL (will come back later)

PO for RL/Control $\Longrightarrow$ Optimization

PO for RL/Control $\Longrightarrow$ (Constrained Nonconvex) Opt.

- PO for RL/Control can be generally written as

$$\min_{K} \ \mathcal{J}(K), \qquad s.t. \ \ K \in \mathcal{K}$$

  - Parametrized policy/controller K
  - Objective to optimize $\mathcal{J}$
  - Constraint set $\mathcal{K}$ (important but sometimes implicit!)

- Linear quadratic regulator (LQR) as an example

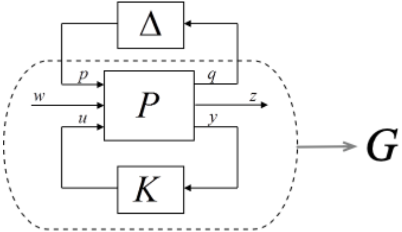
$$\min_{K} \ \mathcal{J}(K) := \sum_{t=0}^{\infty} \mathbb{E}[x_t^\top Q x_t + u_t^\top R u_t], \qquad s.t. \ \ K \ \text{ is stabilizing}$$

  - $u_t = -Kx_t$ for gain matrix K
  - $\mathcal{K} = \{K \,|\, \rho(A - BK) < 1\}$; $\mathcal{K}$ a nonconvex constraint set

- Other examples of $\mathcal{K}$: boundedness of K's norm, probability simplex, safety-constraint on states, etc.

Robustness Constraint on K

- Beyond stability, robustness is a core topic in control theory
- Need a controller robust to disturbance/model-uncertainty



- G–Δ model covers many robustness considerations
  - Parametric uncertainty $\Delta A, \Delta B$ in A, B (most popular in recent ML for control literature)
  - Time-varying parameters A_t, B_t
  - Time-varying delay $u_t = -Kx_{t-\tau}$
  - Even dynamical uncertainty (unknown model-order)
- Robustness constraint $\mathcal{K}$?

Questions

- How to enforce/maintain robustness for policy-optimization RL methods during learning?
- What are the global convergence guarantees, if any, of PO methods in *learning for robust control*?

Problem Statement

Starting Point: LEQG [Jacobson '73]

- LQR/LQG $\Longrightarrow$ linear exponential quadratic Gaussian (LEQG)
- Simple but benchmark risk-sensitive control setting
- Linear system dynamics:

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

system state $x_t \in \mathbb{R}^d$ with $x_0 \sim \mathcal{N}(\mathbf{0}, X_0)$, noise $w_t \sim \mathcal{N}(\mathbf{0}, W)$

- One-stage cost $c(x, u) = x^\top Qx + u^\top Ru$, with objective

$$\min \ \mathcal{J} := \lim_{T\to\infty} \ \frac{1}{T}\frac{2}{\beta} \log \mathbb{E} \exp \left[\frac{\beta}{2} \sum_{t=0}^{T-1} \big(\underbrace{x_t^\top Qx_t + u_t^\top Ru_t}_{c(x_t,u_t)}\big)\right]$$

- Intuition: by Taylor expansion around $\beta = 0$

$$\mathcal{J} \approx \lim_{T\to\infty} \ \frac{1}{T}\left\{\mathbb{E}\left[\sum_{t=0}^{T-1} c(x_t, u_t)\right] + \frac{\beta}{4}\mathrm{Var}\left[\sum_{t=0}^{T-1} c(x_t, u_t)\right]\right\} + O(\beta^2).$$

Starting Point: LEQG

- Optimal controller is LTI state-feedback [ZHB, ′21], conjectured in [Glover and Doyle, ′88]

$$\mu_t(x_{0:t}, u_{0:t-1}) = -K^* x_t$$

- See more results on LEQG specifically in [ZHB, ′21]
- Implicit robustness constraint in LEQG

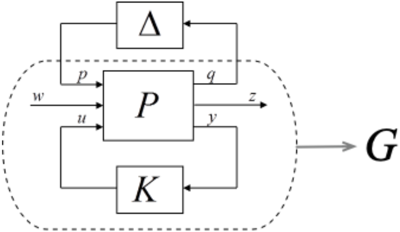
Lemma (Glover and Doyle ′88)

The feasible set of $\mathcal{J}(K)$ is the $1/\sqrt{\beta}$-sublevel set of the $\mathcal{H}_\infty$-norm of $\mathcal{T}(K)$, i.e.,

$$\{K \mid K \text{ stabilizing; } \|\mathcal{T}(K)\|_\infty < 1/\sqrt{\beta}\}.$$

Bigger Picture: Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control

- Linear dynamic systems:
$$x_{t+1} = Ax_t + Bu_t + Dw_t \qquad z_t = Cx_t + Eu_t$$
- $\mathcal{H}_\infty$-norm: $\ell_2 \to \ell_2$ operator norm from $\{w\}$ to $\{z\}$
$$\|\mathcal{T}(K)\|_\infty := \sup_{\theta\in[0,2\pi)} \lambda_{\max}^{1/2}\left[[\mathcal{T}(K)(e^{-j\theta})]^\top \mathcal{T}(K)(e^{j\theta})\right]$$



Bigger Picture: Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control

- Solve

$$\min_K \quad \mathcal{J}(K) \qquad s.t. \quad \underbrace{\rho(A-BK)<1 \;\;\&\;\; \|\mathcal{T}(K)\|_\infty<\gamma}_{\text{define} \quad \mathcal{K}}$$

- $\mathcal{J}(K)$ upper-bounds $\mathcal{H}_2$-norm:

  $$\mathcal{J}(K)=\mathrm{Tr}(P_K DD^\top), \quad \text{or} \quad \mathcal{J}(K)=-\gamma^2 \log\det(I-\gamma^{-2}P_K DD^\top),$$

  where P_K solves a Riccati equation given K

  $$P_K=\widetilde{A}_K^\top P_K \widetilde{A}_K+\widetilde{A}_K^\top P_K D(\gamma^2 I-D^\top P_K D)^{-1}D^\top P_K \widetilde{A}_K+C^\top C+K^\top RK$$

  with $\widetilde{A}_K=A-BK$
- If $\gamma\to\infty$, then $\mathcal{J}(K)\to\mathcal{H}_2$-norm, e.g., it reduces to LQR

Why an Important & Interesting Model?

- Intuition: Small gain theorem – if $\|\mathcal{T}(K)\|_\infty < \gamma$ and $\|\Delta\|_{\ell_2 \to \ell_2} < 1/\gamma$, then G–Δ is input-output stable
- Choosing

$$\mathcal{K} = \{K \mid \rho(A - BK) < 1; \|\mathcal{T}(K)\|_\infty < \gamma\}$$

$\Longrightarrow$ certain level of robust stability

  - $\gamma \to \infty$ reduces to stability region in LQR
- Second choice of $\mathcal{J}(K)$ with $D = W^{1/2}$ and $\gamma = 1/\sqrt{\beta}$ coincides with the risk-sensitive LEQG objective
- It also unifies maximum-entropy/$\mathcal{H}_\infty$ control, LQG/$\mathcal{H}_\infty$ control [Glover and Doyle, ′88; Mustafa, ′89]; γ-level disturbance attenuation [Başar, ′91]; and zero-sum LQ dynamic games (come to it later) [Jacobson, ′73; Başar and Bernhard, ′95]. Also used for solving $\mathcal{H}_\infty$-optimal control
- Also used in Economics to model sequential decision-making under model uncertainty [Hansen and Sargent, ′08]

Algorithms and Landscape

Policy Gradient Algorithms

- Following the naming convention in [Fazel et al., '18] for LQR

Policy Gradient:

$$K' = K - \eta \cdot \nabla \mathcal{J}(K)$$
$$= K - 2\eta \cdot \big[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A\big] \cdot \Delta_K,$$

Natural PG:

$$K' = K - \eta \cdot \nabla \mathcal{J}(K) \cdot \Delta_K^{-1}$$
$$= K - 2\eta \cdot \big[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A\big],$$

Gauss-Newton:

$$K' = K - \eta \cdot (R + B^\top \widetilde{P}_K B)^{-1} \cdot \nabla \mathcal{J}(K) \cdot \Delta_K^{-1}$$
$$= K - 2\eta \cdot \big[K - (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A\big]$$

- Recall P_K is the solution to some Riccati equation dictated by K, and Δ_K is the solution to another (dual) Riccati equation

Landscape

Lemma (Nonconvexity)

There is a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem that is nonconvex for policy optimization.

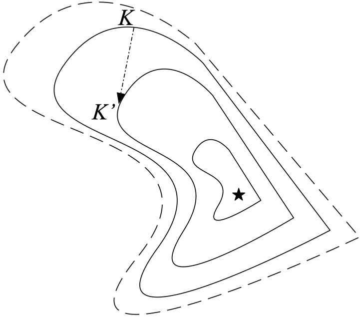
Lemma (Non-Coercivity)

There is a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem whose cost function $\mathcal{J}(K)$ is non-coercive. Particularly, as $K \to \partial\mathcal{K}$, $\mathcal{J}(K)$ does not necessarily approach infinity.

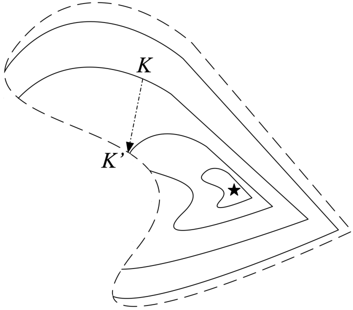
- Coercivity is key in LQR analysis [Fazel et al., '18][Bu et al., '19][Malik et al., '19][Mohammadi et al., '19]
  - Ensures any descent direction to be feasible
  - Can confine everything to the sub-level set (reduces to standard smooth optimization)

Convergence

Landscape Illustration



LQR



Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control

- Descent in $\mathcal{J}(K)$ does not necessarily ensure *feasibility/robust stability*
- How to enforce $K \in \mathcal{K}$ during iteration?

Implicit Regularization

- Regularization: *iterate K remains inside $\mathcal{K}$,* i.e., robustly stable
- Can be made explicit via projection onto $\mathcal{K}$. But $\mathcal{K}$ is nonconvex, and defined in frequency domain
- Implicit regularization:
  - The convergence of certain algorithms behaves as if certain regularization is used
  - Borrowed from machine learning literature: observed in learning overparametrized neural nets/nonlinear models [Kubo et al., '19][Azizan et al., '19], phase retrieval and matrix completion [Ma et al., '17], etc., with (stochastic) gradient (mirror) descent
  - Property of both the nonconvex problem and algorithm
- Gauss-Newton and natural PG enjoy implicit regularization!

Theory: Implicit Regularization

Theorem (Implicit Regularization)

Suppose the stepsizes η satisfy:

- *Gauss-Newton:* $\eta \le 1/2$,
- *Natural policy gradient:* $\eta \le 1/(2\|R + B^\top \widetilde{P}_{K_0} B\|)$.

Then, $K \in \mathcal{K} \Longrightarrow K' \in \mathcal{K}$.

- General descent directions of $\mathcal{J}(K)$ may not work, but certain directions do

Implicit Regularization: Proof Idea

- Linear matrix inequalities (LMIs)-based approach
- A new use of Bounded Real Lemma [Başar & Bernhard, ′95][Zhou et al., ′96][Dullerud & Paganini, ′00]:
  $K \in \mathcal{K} \Longleftrightarrow$ Riccati equation $\Longleftrightarrow$ strict Riccati inequality (RI)
- Observation: two consecutive iterates $K \to K'$ are related – previous P_K is a candidate for the non-strict RI under K' to hold
- Perturb P_K in a certain way $\Longrightarrow$ strict RI
  - Perturb $P = P_K + \alpha \bar{P}$ for small enough $\alpha > 0$, where $\bar{P} > 0$ solves the Lyapunov equation as before, with $C^\top C + K^\top RK$ replaced by $-I$

Theory: Global Convergence

Theorem (Global Convergence)

Suppose $K_0 \in \mathcal{K}$, then both the Gauss-Newton and natural PG updates converge to the global optimum $K^* = (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A$ with $\mathcal{O}(1/N)$ rate.

Theory: Local (Super-)Linear Rates

- Much faster rates around the global optimum

Theorem (Local Faster Rates)

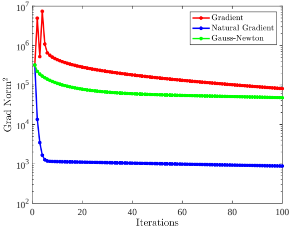
Suppose the conditions above hold, and additionally $DD^\top > 0$. Then, both the Gauss-Newton and natural PG updates converge to the optimal control gain K^* with locally linear rate. In addition, if $\eta = 1/2$, the Gauss-Newton update converges to K^* with locally Q-quadratic rate.

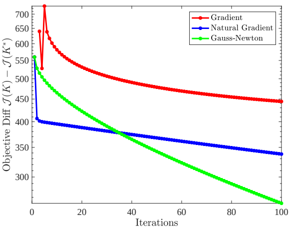
- Gradient domination (Polyak-Łojasiewicz) property holds locally
- Q-quadratic rate mirrors that of policy iteration for LQR [Lanchaster and Rodman '95]

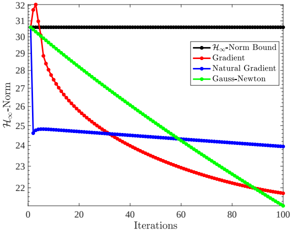
Simulations

Simulations

- Initialization near the boundary $\partial\mathcal{K}$; infinitesimal stepsize η for PG







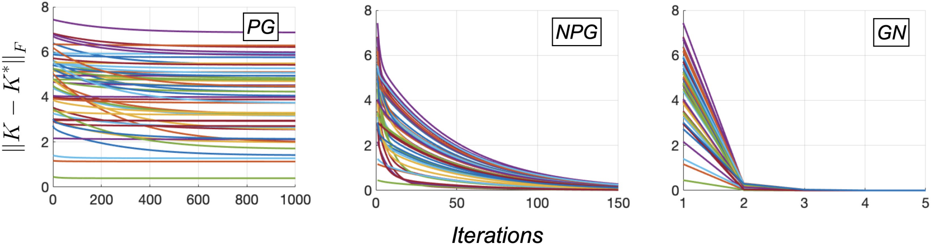
Ave. Grad. Norm Square

$\mathcal{J}(K) - \mathcal{J}(K^*)$

$\mathcal{H}_\infty$-norm $\|\mathcal{T}(K)\|_\infty$

Simulations: Global Convergence

- Escaping suboptimal stationary points



Simulations: Scalability

- Computationally more efficient than existing general robust control solvers – HIFOO [Arzelier et al., ′11] & Matlab `h2hinfsyn` function [Mahmoud and Pascal, ′96] and `systune` function [Apkarian et al., ′08]

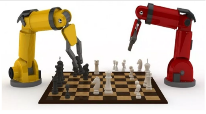
System Dim.	*HIFOO*	`h2hinfsyn`	`systune`	*NPG*	*GN*	*Speedup*
15×15	$0.3742s$	$95.2663s$	$0.4276s$	$0.0481s$	$0.0420s$	$\sim 8/2117/8.8\times$
60×60	$18.4380s$	fail, $> 7200s$	$171.7855s$	$0.3906s$	$0.3902s$	$\sim 47/ > 18461/440\times$
90×90	$241.4416s$	fail, $> 7200s$	$4126.9s$	$0.8167s$	$0.8103s$	$\sim 295/ > 36922/5093\times$

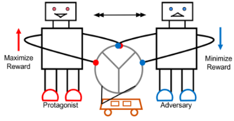
Table: Average runtime comparison

Connection to Multi-Agent RL (MARL)

Multi-Agent RL

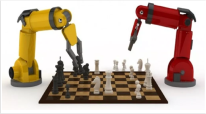
- Usually studied under framework of Markov games [Shapley '53]
- The most basic MARL model ever since [Littman, '94]: two-player zero-sum Markov games

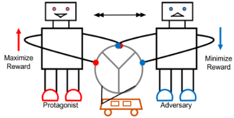




Multi-Agent RL

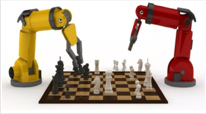
- Usually studied under framework of Markov games [Shapley '53]
- The most basic MARL model ever since [Littman, '94]: two-player zero-sum Markov games
- Benchmark in continuous control: linear quadratic zero-sum dynamic games (DG) [Başar and Bernhard, '95] (mirrors LQR for single-agent RL)

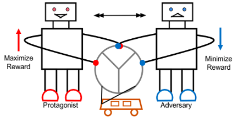




Multi-Agent RL

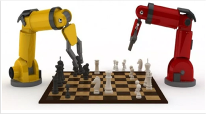
- Usually studied under framework of Markov games [Shapley '53]
- The most basic MARL model ever since [Littman, '94]: two-player zero-sum Markov games
- Benchmark in continuous control: linear quadratic zero-sum dynamic games (DG) [Başar and Bernhard, '95] (mirrors LQR for single-agent RL)
- PO methods widely used in modern empirical MARL, while its convergence guarantees remain largely open

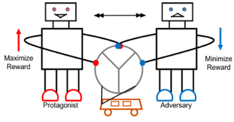




Multi-Agent RL

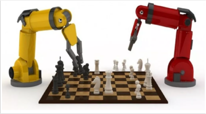
- Usually studied under framework of Markov games [Shapley '53]
- The most basic MARL model ever since [Littman, '94]: two-player zero-sum Markov games
- Benchmark in continuous control: linear quadratic zero-sum dynamic games (DG) [Başar and Bernhard, '95] (mirrors LQR for single-agent RL)
- PO methods widely used in modern empirical MARL, while its convergence guarantees remain largely open
  - (Projected) PO for LQ zero-sum DGs [ZYB, '19][Bu et al., '19]
  - Negative/Non-convergence results for (multi-player) LQ general-sum DGs [Mazumdar, Ratliff, Jordan, and Sastry, '19]
  - PG methods (and variants) for tabular zero-sum Markov games [Daskalakis et al., '20][Zhao et al., '21][Cen et al., '21,'22]...

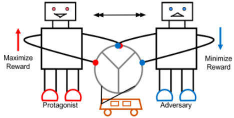




Multi-Agent RL

- Usually studied under framework of Markov games [Shapley '53]
- The most basic MARL model ever since [Littman, '94]: two-player zero-sum Markov games
- Benchmark in continuous control: linear quadratic zero-sum dynamic games (DG) [Başar and Bernhard, '95] (mirrors LQR for single-agent RL)
- PO methods widely used in modern empirical MARL, while its convergence guarantees remain largely open
  - (Projected) PO for LQ zero-sum DGs [ZYB, '19][Bu et al., '19]
  - Negative/Non-convergence results for (multi-player) LQ general-sum DGs [Mazumdar, Ratliff, Jordan, and Sastry, '19]
  - PG methods (and variants) for tabular zero-sum Markov games [Daskalakis et al., '20][Zhao et al., '21][Cen et al., '21,'22]...
- Nonconvex-nonconcave [ZYB, '19][Daskalakis et al., '20], PO can easily diverge if not designed carefully





LQ Zero-Sum Dynamic Games

- $\mathcal{H}_2/\mathcal{H}_\infty$ control is strongly tied to LQ zero-sum dynamic games
- Let $u_t = -Kx_t$ and $w_t = -Lx_t$ then solve:

$$\mathcal{J}(K, L) := \mathbb{E}_{x_0 \sim \mathcal{D}} \left\{ \sum_{t=0}^{\infty} \left[x_t^\top Q x_t + (Kx_t)^\top R^u (Kx_t) - (Lx_t)^\top R^v (Lx_t) \right] \right\}$$

$$\text{Solve: } \min_K \max_L \; \mathcal{J}(K, L) \quad \Longleftrightarrow \quad \min_K \; \mathcal{J}(K, L(K))$$

with $x_{t+1} = Ax_t + Bu_t + Cw_t$

- For fixed K (outer-loop), take max over L (inner-loop), the Riccati equation becomes
  the same Riccati equation as in $\mathcal{H}_2/\mathcal{H}_\infty$ control

Implication for MARL

- Previous results $\Longrightarrow$ double-loop update provably works
  - Double-loop/nested-grad.: fix K and improve L, then improve K
  - Aligned with the empirical tricks to stabilize nonconvex-nonconcave minimax opt. with timescale separation [Lin, Jin, & Jordan, '18], as in training GANs [Heusel et al., '18]
- Gives global convergence of PO in competitive MARL (zero-sum Markov/dynamic games)

Benefit from MARL: Model-Free $\mathcal{H}_2/\mathcal{H}_\infty$ Control

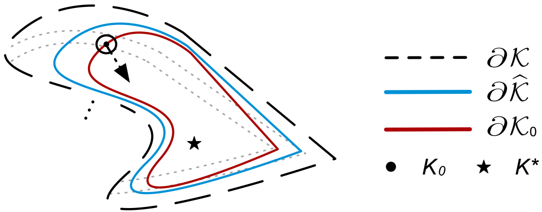
- Recall the policy gradient form

$$\nabla \mathcal{J}(K) = 2\big[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A\big]\Delta_K,$$

  while Δ_K cannot be estimated from sampled trajectories
- Instead solve the equivalent game using data
  - Build up a virtual adversary $w_t = -Lx_t$
  - Double-loop/nested-grad.: fix K and improve L, then improve K
- Derivative-free methods for LQR [Fazel et al., '18][Malik et al., '19] cannot work directly
  - Non-coercive & only certain direction works $\Longrightarrow$ no uniform margin
  - Caveat: quantities (cost, action space, control gain matrices) in the LQ setting are continuous, and can easily go unbounded!
  - Leads to no-global-smoothness + nonconvexity-nonconcavity
- Can be addressed under the unified LQ game formulation, for finite-horizon settings [ZZHB, '21]

Illustration for Derivative-Free PO Convergence

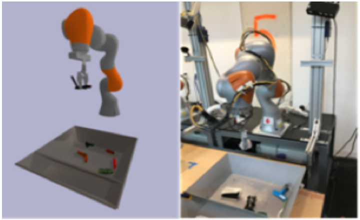
- Proof idea illustrated with figures
  - $\mathcal{K}_0$ is the "level-set" corresponding to initialization K_0;
  - $\widehat{\mathcal{K}}$ is a larger set that the iterates will not leave (with high probability); due to stochastic errors when using samples
  - both are compact $\Longrightarrow$ uniform smoothness constant over $\widehat{\mathcal{K}}$



Connection to Robust Adversarial RL (RARL)

Robust Adversarial RL [Pinto et al., ′17]

- RL hardly generalizes due to Sim2real and/or training-testing gap



[Google AI, ′16]

[Tobin et al., ′17]

- One remedy: RARL [Pinto et al., ′17]
  - Idea: introduce an adversary, playing against the agent
  - Dates back to [Morimoto and Doya, ′05], under the name robust RL, and "inspired by $\mathcal{H}_\infty$-theory"
  - Made popular by the empirical work [Pinto et al., ′17]
  - Question: Any robustness interpretation and convergence guarantee?

LQ RARL

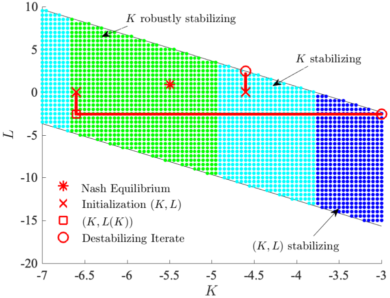
- RARL setting $\Longleftrightarrow$ zero-sum dynamic game
- LQ RARL: View w_t as model-uncertainty, or the model-misspecification when linearizing a nonlinear model
- Recall the RARL scheme in [Pinto et al., '17]

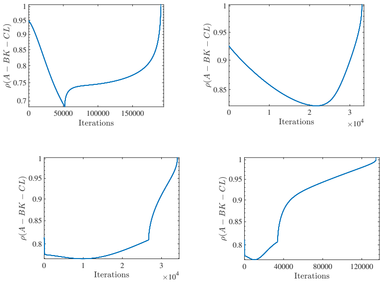
Algorithm 1 Policy-Based LQ RARL Scheme (Pinto et al., 2017)

Input: LQ RARL environment, initial policies (K_0, L_0)
for $n = 1, \ldots, N$ **do**
  Update $L_n \leftarrow L_{n-1}$
  for $j = 1, \ldots, N_L$ **do**
    Update $L_n \leftarrow$ PolicyOptimizer(K_{n-1}, L_n)
  end for
  Update $K_n \leftarrow K_{n-1}$
  for $i = 1, \ldots, N_K$ **do**
    Update $K_n \leftarrow$ PolicyOptimizer(K_n, L_n)
  end for
end for
Return: policy pair (K_N, L_N)

RARL in [Pinto et al., '17] Easily Fails [ZHB, '20]

- Stability issue due to bad initialization
- Stability issue due to bad choices of $(N_K,\ N_L)$ (K_0, L_0)





- What is a good combination of initialization & update rule?

Implication from Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control

- By implicit regularization, we find a provably convergent pair of (initialization, update rule): $\big((K_0 \in \mathcal{K},\ L_0 = 0),\ (N_K = 1,\ N_L = \infty)\big)$

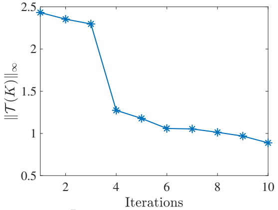
Algorithm Double-loop Update

Input: Initialize $K_0 \in \mathcal{K}$, L_0 stabilizing, e.g., $L_0 = 0$
for $n = 0, \cdots$ **do**
  for $i = 0, \cdots$ **do**
    Update $L_{i+1} \leftarrow \text{PolicyOptimizer}(K_n,\ L_i)$
  end for
  Update $K_{n+1} \leftarrow \text{PolicyOptimizer}(K_n,\ L_\infty)$
end for

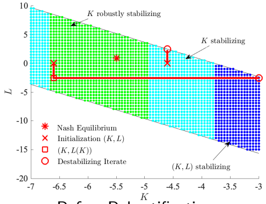
Implication from Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control

- By implicit regularization, we find a provably convergent pair of (initialization, update rule): $\big((K_0 \in \mathcal{K},\ L_0 = 0),\ (N_K = 1,\ N_L = \infty)\big)$
- How to find such a $K_0 \in \mathcal{K}$ (in a model-free way) – robustify K_0?
- For any stabilizing K, perform $K' = K - \alpha g$ with $g \in \mathbb{R}^{m\times n}$ the finite-difference estimate of the subgradient of $\|\mathcal{T}(K)\|_\infty$, where

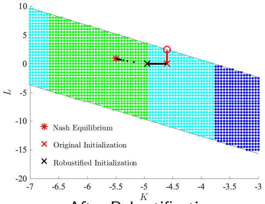
$$g_{ij} = \frac{\|\mathcal{T}(K + \epsilon d_{ij})\|_\infty - \|\mathcal{T}(K - \epsilon d_{ij})\|_\infty}{2\epsilon}$$



Decrease $\mathcal{H}_\infty$-norm



Before Robustification

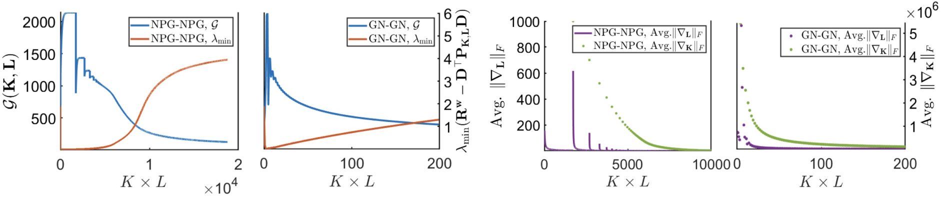


After Robustification

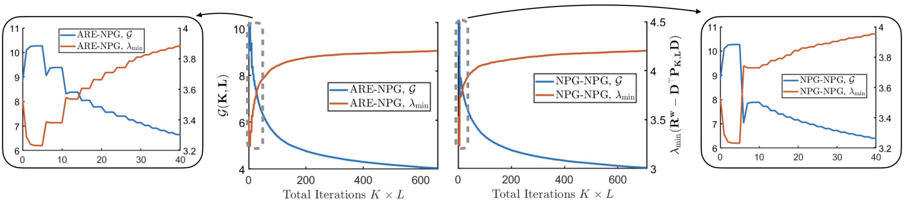
Additional Simulations

Convergent Cases

- Exact update:



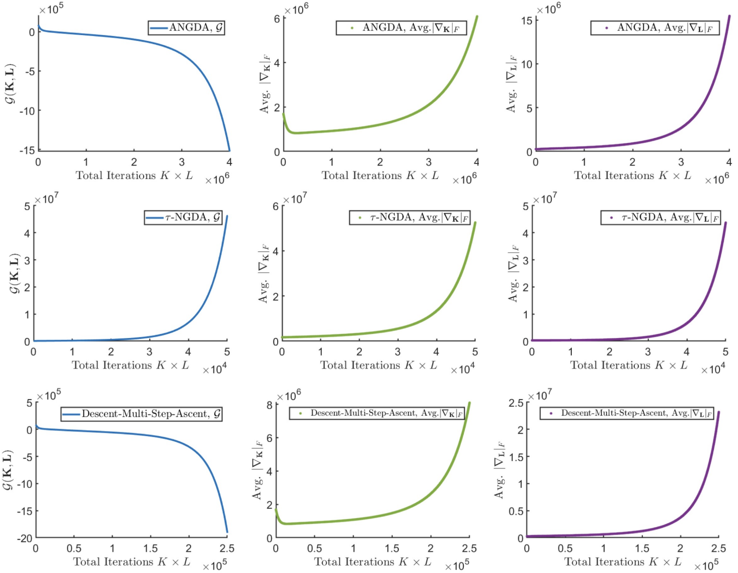
- Derivative-free update:



Some Divergent Cases

- Update-rules other than double-loop may easily diverge, even with infinitesimal stepsizes
  - `ANGDA`: alternative-update of natural PG descent & ascent
  - τ-`NGDA`: simultaneous-update with stepsizes ratio $\frac{\eta}{\alpha} = \tau > 1$
  - `Descent-Multi-Step-Ascent`: multiple ascent steps per descent step

Some Divergent Cases

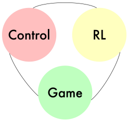


- Without global smoothness, controlling the iterates' boundedness is critical and challenging

Concluding Remarks

Concluding Remarks

- Studied policy optimization landscape for risk-sensitive/robust control, with fundamental challenges diff. from that of LQR – deepened our understanding of existing results on LQR
- Developed two PO methods, identified their implicit regularization property, and established global convergence + sample complexity
- Along the way
  - Global convergence and sample complexity of PO for competitive MARL, in the LQ zero-sum setting
  - Some theoretical understanding and critical thinking on RARL, from robust control perspective
  - Explicit regularization and convex-reformulation can also be useful —- a unified differentiable convex liftings (DCL) framework [USPZT, '22]
- A natural intersection of control, RL, and game theory



Thank You!

Direct Policy Search for Robust Control: A Nonsmooth Optimization Perspective

Bin Hu

ECE & CSL, University of Illinois Urbana-Champaign

L4DC Tutorial 2023

Joint work with Xingang Guo

Outline

- **Motivation and Problem Formulation**
- Main Results
- Conclusions and Future Directions

Motivation: Reinforcement Learning for Control

- Many robust control problems are solved via lifting into convex spaces. Recently, reinforcement learning (RL) has shown great promise for control!

- Main workhorse: direct policy search/policy optimization (PO)

$$\min_{K} \quad J(K), \qquad s.t. \quad K \in \mathcal{K}$$

  - Parametrized policy K (e.g. linear mapping, neural networks)
  - Cost function J (tracking errors, closed-loop $\mathcal{H}_2/\mathcal{H}_\infty$ norms, etc)
  - Constraint set $\mathcal{K}$ (stability, robustness, safety, etc)
  - PO algorithm: $K' = K - \alpha \nabla J(K)$ (nonconvex problem!)
  - Theory: Landscape, feasibility, convergence, complexity

- Question: How to tailor policy-based RL for robust control?
- **This talk: Guarantees of PO on $\mathcal{H}_\infty$ control benchmarks**

PO Theory for Robust Control

- PO theory for mixed design (maintaining robustness via improving average)
  - Landscape: Feasible set is connected, and stationary is global
  - Feasibility: The cost is nonconvex and non-coercive! Fortunately, double-loop natural policy gradient (NPG) can implicitly regularize
  - Global sublinear convergence for NPG
  - Ref: Zhang, Hu, Başar. Policy optimization for $\mathcal{H}_2$ linear Control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence, SICON 2021.
- **PO theory for $\mathcal{H}_\infty$ state-feedback synthesis (improving robustness)**
  - Feature: Nonconvex nonsmooth
  - Landscape: Any Clarke stationary points are global
  - Feasibility: The cost is coercive and serves as a barrier function on $\mathcal{K}$
  - Global convergence: Goldstein's subgradient method achieves global convergence provably
  - Ref: Guo and Hu. Global convergence of direct policy search for state-feedback $\mathcal{H}_\infty$ robust control: A revisit of nonsmooth synthesis with Goldstein subdifferential, NeurIPS 2022.

Review: Linear Quadratic Regulator

- LQR as PO: Consider $x_{t+1} = Ax_t + Bu_t + w_t$ with w_t being stochastic IID

$$\min_K \quad J(K) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t)\right], \quad s.t. \;\; K \;\; \text{is stabilizing}$$

  - $u_t = -Kx_t$ for gain matrix K
  - $\mathcal{K} = \{K \mid \rho(A - BK) < 1\}$; $\mathcal{K}$ a nonconvex constraint set

- PO theory for LQR
  - Landscape: Stationary is global
  - Feasibility: The LQR cost is coercive and serves as a barrier on $\mathcal{K}$
  - Global convergence & sample complexity: Linear rate via the gradient dominance/smoothness property

- Main Ref:

  M. Fazel, R. Ge, S. Kakade, M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator, ICML 2018.

Problem Formulation: State-feedback $\mathcal{H}_\infty$ Control

Consider the following linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad x_0 = 0.$$

- We assume that (A, B) is stabilizable
- $\mathbf{u} := \{u_0, u_1, \cdots\}$, $\mathbf{w} := \{w_0, w_1, \cdots\}$, and $\|\mathbf{w}\| = (\sum_{t=0}^{\infty} \|w_t\|^2)^{1/2}$.
- Our goal is to find a sequence $\mathbf{u}$ to minimize the quadratic cost function

$$\min_{\mathbf{u}} \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_{t=0}^{\infty} (x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t)$$

  in the presence of the worst case disturbance $\|\mathbf{w}\| \leq 1$.
- This is different than the LQR problem, where $\mathbf{w}$ is stochastic.
- $\|\mathbf{w}\| \leq 1$ is not restrictive, we can choose arbitrary ℓ_2 bound.
- We assume that Q and R are positive definite.
- It is well known that the optimal solution is using a linear state-feedback policy $u_t = -Kx_t$ (Başar and Bernhard 2008).

Problem Formulation: State-feedback $\mathcal{H}_\infty$ Control

Consider $u_t = -Kx_t$, the closed-loop system becomes $x_{t+1} = (A - BK)x_t + w_t$. We have the following PO problem:

$$\min_K \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_{t=0}^{\infty} x_t^\mathsf{T}(Q + K^\mathsf{T}RK)x_t.$$

The above optimization problem equivalent to the following PO problem

$$\min_K \; J(K) := \sup_{\omega\in[0,2\pi]} \sigma_{\max}\big((Q + K^\mathsf{T}RK)^{1/2}(e^{j\omega}I - A + BK)^{-1}\big)$$
$$\text{s.t. } K \in \mathcal{K} := \{K : \rho(A - BK) < 1\}.$$

- This is a constrained nonconvex nonsmooth optimization problem.
- $\mathcal{K}$ can be nonconvex.
- The nonsmoothness comes from two sources:
  1. The computation of the maximum singular value.
  2. The operator $\sup$ over $\omega \in [0, 2\pi]$.

Convex LMIs vs. Direct Policy Search

- In 1980s, convex optimization methods become popular for control study due to global guarantees and efficient interior point methods
- Reparameterize problems as convex optimization problems (one does not optimize the controller parameters directly)

$$\begin{aligned} &\{K \in \mathcal{K} : J(K) \leq \gamma\} \\ \Longleftrightarrow &\{K = LY^{-1} : \mathrm{LMI}(Y, L, \gamma) \preceq 0 \text{ is feasible, } Y \succ 0\}. \end{aligned}$$

- Minimizing γ over $\mathrm{LMI}(Y, L, \gamma) \preceq 0$ and $Y \succ 0$ is a SDP problem
- See for example, Boyd *et al.*, "Linear Matrix Inequalities in System and Control Theory", 1994, SIAM
- PO is not convex!
- In this past, PO has been a popular approach for problems that cannot be convexified, e.g. structured $\mathcal{H}_\infty$ synthesis! (HIFOO and Hinfstrcuct)
- **This talk:** View $\mathcal{H}_\infty$ synthesis as a benchmark for understanding PO

Some Background on Nonsmooth Optimization

Clarke subdifferential:

$$\partial_C J(K) := \mathrm{conv}\{\lim_{i\to\infty} \nabla J(K_i) : K_i \to K,\ K_i \in \mathrm{dom}(\nabla J) \subset \mathcal{K}\}.$$

- $\partial_C J(K)$ is well defined for any $K \in \mathcal{K}$.
- $J(K)$ is locally Lipschitz and hence almost everywhere differentiable.

Proposition

If K is a local minimum of J, then $0 \in \partial_C J(K)$ and K is a Clarke stationary point.

Some Background on Nonsmooth Optimization

Generalized Clarke directional derivative:

$$J^\circ(K, d) := \lim_{K' \to K} \sup_{t \searrow 0} \frac{J(K' + td) - J(K')}{t}.$$

Directional derivative:

$$J'(K, d) := \lim_{t \searrow 0} \frac{J(K + td) - J(K)}{t}.$$

- $J^\circ(K, d)$ and $J'(K, d)$ are different in general.
- $J'(K, d) = J^\circ(K, d)$ if $J(K)$ is subdifferentially regular.

Subdifferentially Regular Property

Let $K^\dagger$ be a Clarke stationary point for J. If J is subdifferentially regular, then $J'(K^\dagger, d) \geq 0$ for all d[a].

[a] This result is known, see Theorem 10.1 in Rockafellar and Wets 2009.

Some Background on Nonsmooth Optimization

Goldstein subdifferential:

$$\partial_\delta J(K) := \text{conv}\left\{\cup_{K' \in \mathbb{B}_\delta(K)} \partial_C J(K')\right\},$$

- $\mathbb{B}_\delta(K)$ is the δ-ball around K
- requires $\mathbb{B}_\delta(K) \subset \mathcal{K}$.

Generating a good descent direction (Goldstein1977):

Descent inequality

Let F be the minimal norm element in $\partial_\delta J(K)$. Suppose $K - \alpha F/\|F\|_2 \in \mathcal{K}$ for any $0 \le \alpha \le \delta$. Then we have:

$$J(K - \delta F/\|F\|_2) \le J(K) - \delta\|F\|_2.$$

Outline

- Motivation and Problem Formulation
- **Main Results**
- Conclusions and Future Directions

Summary of Known Facts

$$\min_{K} \; J(K) := \sup_{\omega\in[0,2\pi]} \sigma_{\max}\big((Q+K^{\mathsf{T}}RK)^{1/2}(e^{j\omega}I - A + BK)^{-1}\big)$$

$$\text{s.t. } K \in \mathcal{K} := \{K : \; \rho(A-BK) < 1\}.$$

- $\mathcal{K}$ is open, can be unbounded, and nonconvex.
- $J(K)$ is continuous, nonsmooth, and can be nonconvex in K.
- $J(K)$ is locally Lipschitz, subdifferentially regular over the feasible set $\mathcal{K}$.

High Level Ideas

Goldstein's subgradient method:

$$K^{n+1} = K^n - \delta^n F^n / \|F^n\|_2,$$

- F^n is the minimum norm element of $\partial_{\delta^n} J(K^n)$.
- $K^0 \in \mathcal{K}$ is known.

High level ideas:

- Goldstein's subgradient method finds Clarke stationary point
- Coerciveness ensures K^n stay within the nonconvex feasible set.
- Clarke stationary points are global, and hence global optimum is found.

Main Results

Theorem (Guo and Hu, NeurIPS2022)

Suppose (Q, R) are positive definite, and (A, B) is stabilizable. We have

1. *$J(K)$ is coercive over the set $\mathcal{K}$. (Proved via the properties of (Q, R))*
2. *For any $K \in \mathcal{K}$ satisfying $J(K) > J^*$, there exists $V \neq 0$ s.t. $J'(K, V) < 0$.*
3. *Any Clarke stationary points of the $\mathcal{H}_\infty$ cost are global minimum.*
4. *For any $\gamma > J^*$, the sublevel set $\mathcal{K}_\gamma = \{K \in \mathcal{K} : J(K) \leq \gamma\}$ is compact. There is a strict separation between $\mathcal{K}_\gamma$ and $\mathcal{K}^c$.*
5. *Suppose $K^0 \in \mathcal{K}$. Set $\Delta_0 := \mathrm{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c) > 0$, and $\delta^n = \frac{0.99\Delta_0}{n+1}$. Then Goldstein's subgradient method $K^{n+1} = K^n - \delta^n F^n / \|F^n\|_F$ with F^n being the minimum norm element of $\partial_{\delta^n} J(K^n)$ is guaranteed to stay in $\mathcal{K}$ for all n. In addition, we have $J(K^n) \to J^*$ as $n \to \infty$.*
6. *There is also a complexity result for finding (ε, δ)-stationary points.*

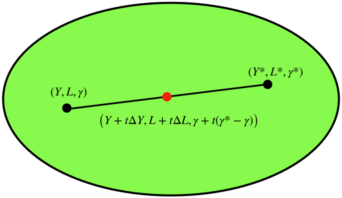
The most technical part of the proof is for Step 2. It requires the use of non-strict version of the KYP lemma.

Step 2 of Main Result

Lemma

For any $K \in \mathcal{K}$ that $J(K) > J^*$, there exists a direction $d \neq 0$ such that the directional derivative $J'(K, d) \leq J^* - J(K) < 0$.

Proof Sketch:



$$\mathrm{LMI}(Y, L, \gamma) \preceq 0$$
$$\mathrm{LMI}(Y^*, L^*, \gamma^*) \preceq 0$$

By convexity, we have

$$\mathrm{LMI}(Y + t\Delta Y, L + t\Delta L, \gamma + t(\gamma^* - \gamma)) \preceq 0$$

with $\Delta Y = Y^* - Y$, $\Delta L = L^* - L$, and $t \in [0, 1]$. In addition, we must have

$$J((L + t\Delta L)(Y + t\Delta Y)^{-1}) \leq \gamma + t(\gamma^* - \gamma).$$

Step 2 of Main Result

Lemma

For any $K \in \mathcal{K}$ that $J(K) > J^*$, there exists a direction $d \neq 0$ such that the directional derivative $J'(K,d) \leq J^* - J(K) < 0$.

Proof Sketch Con:
Based on the fact $J(K^*) < J(K)$, we can construct a direction d such that $J'(K,d) < 0$. Specifically, consider $d = \Delta L Y^{-1} - L Y^{-1} \Delta Y Y^{-1}$. Then we have

$$
\begin{aligned}
J'(K,d) &= \lim_{t \searrow 0} \frac{J(K + t(\Delta L Y^{-1} - L Y^{-1} \Delta Y Y^{-1})) - J(K)}{t} \\
&\leq \lim_{t \searrow 0} \left(\frac{J((L + t\Delta L)(Y + t\Delta Y)^{-1}) - J(K)}{t} + O(t) \right) \\
&\leq \lim_{t \searrow 0} \left(\frac{J(K) + t(J(K^*) - J(K)) - J(K)}{t} + O(t) \right) \\
&= J(K^*) - J(K) < 0,
\end{aligned}
$$

we use the fact that $(Y + t\Delta Y)^{-1} = Y^{-1} - tY^{-1}\Delta Y Y^{-1} + O(t^2)$. ■

Finite-time complexity for (δ, ε)-stationary points

Goldstein's subdifferential: $\partial_\delta J(K) :=$ conv $\left\{\cup_{K' \in \mathbb{B}_\delta(K)} \partial_C J(K')\right\}$.

Definition

A point K is said to be (δ, ε)-stationary if dist$(0, \partial_\delta J(K)) \leq \varepsilon$.

Theorem 3

If we choose $\delta^n = \delta < \Delta_0$, then we have:

- $K^n \in \mathcal{K}$ for all n
- $\min_{n:0 \leq n \leq N} \|F^n\|_2 \leq \frac{J(K^0) - J^*}{(N+1)\delta}$, i.e., the complexity of finding a (δ, ε)-stationary point is $\mathcal{O}\left(\frac{\Delta}{\delta\varepsilon}\right)$

- (δ, ε)-stationarity does not imply being δ-close to an ε-stationary point of J.
- Finite time bounds for $(J(K^n) - J^*)$ is possible via exploiting other advanced properties $J(K)$.

Implementable Algorithms

Finding minimum norm element of Goldstein's subdifferential may not be easy. Fortunately, there are many implementable variants:

- **Gradient Sampling (GS)** (The HIFOO toolbox): The main idea is to randomly generate differentiable samples over $\mathbb{B}_{\delta^n}(K^n)$ with probability 1. The convex hull of the gradients of these samples can be used as an approximation of $\partial_{\delta^n} J(K^n)$.
- **Nonderivative Sampling (NS)**(Kiwiel2010): The NS method can be viewed as the derivative-free version of the GS algorithm by only using the zeroth-order oracle.
- **Interpolated normalized gradient descent (INGD)** (Zhang, J., et al. 2020; Davis, D., et al. 2022): INGD uses an iterative sampling strategy to generate a descent direction. The INGD algorithm is guaranteed to find the (δ, ε)-stationary point with the high-probability finite-time complexity bound.

Numerical Example

To support our theory, we provide some numerical simulations. Consider the following example:

$$A = \begin{bmatrix} 1 & 0 & -5 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \; Q = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \; R = 1.$$

For this example, we have $J^* = 7.3475$. We initialize from

$$K^0 = \begin{bmatrix} 0.4931 & -0.1368 & -2.2654 \end{bmatrix},$$

which satisfies $\rho(A - BK^0) = 0.5756 < 1$.

Numerical Example

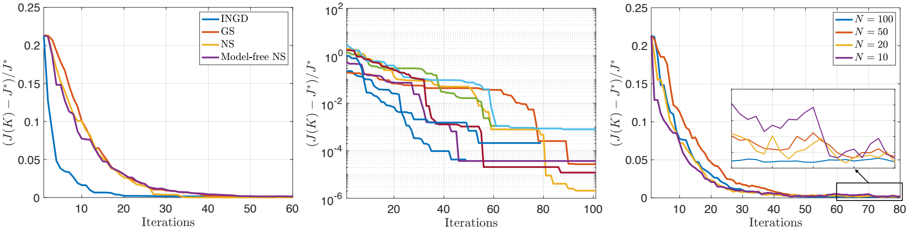


Figure: Simulation results. Left: The trajectory of relative error of GS, NS, INGD, and Model-free NS methods. Middle: The trajectory of the relative optimality gap of 8 randomly generated cases for the NS method. Right: The trajectory of the Model-free NS method with more noisy oracle.

Outline

- Motivation and Problem Formulation
- Main Results
- **Conclusions and Future Directions**

Take Aways

We studied the global convergence of direct policy search on state-feedback $\mathcal{H}_\infty$ robust control synthesis.

- State-feedback $\mathcal{H}_\infty$ synthesis is a constrained nonconvex nonsmooth policy optimization problem.
- Any Clarke stationary points for this problem are actually global minimum.
- Goldstein's subgradient methods are guaranteed to stay within the nonconvex feasible set and converge to the global optimal.
- (δ, ε)-stationary points can be found with finite-time guarantees.
- (δ, ε)-stationarity does not imply being δ-close to an ε-stationary point of J.

Future Work

- Finite-time bounds for the optimality gap (i.e. $J(K^n) - J^*$)
- The sample complexity of direct policy search on model-free $\mathcal{H}_\infty$ control
- Other $\mathcal{H}_\infty$ synthesis problems (static/dynamic output feedback, etc)

Thanks!

If you are interested, feel free to send an email to binhu7@illinois.edu

Funding & Support: NSF

Analysis of the Optimization Landscape of Linear Quadratic Gaussian (LQG) Control

Work by Yang Zheng, Yujie Tang, and Na Li

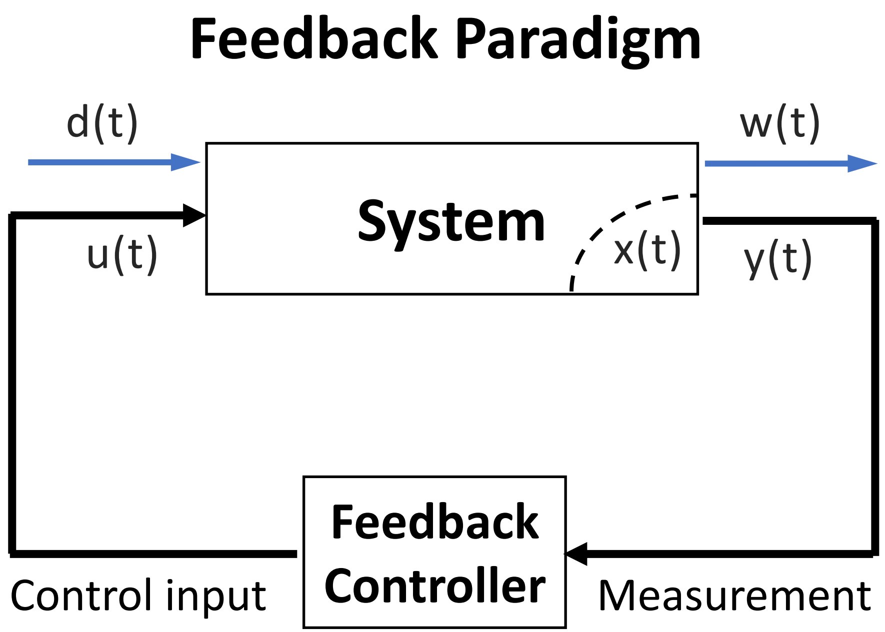
Presented by Bin Hu

5th Annual Learning for Dynamics & Control Conference

University of Pennsylvania. June 14-16, 2023

Today's talk

❑ Optimal Control



Linear Quadratic Optimal control

$$\min_{u_1,u_2,\ldots,} \quad \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(x_t^{\mathsf{T}} Q x_t + u_t^{\mathsf{T}} R u_t\right)\right]$$

$$\text{subject to} \quad x_{t+1} = Ax_t + Bu_t + w_t$$

$$y_t = Cx_t + v_t$$

- Many practical applications
- **Linear Quadratic Regulator (LQR)** when the state x_t is directly observable
- **Linear Quadratic Gaussian (LQG) control** when only partial output y_t is observed
- Extensive classical results (Dynamic programming, Separation principle, Riccati equations, etc.)

They are all model-based. Are there any guarantees for non-convex policy optimization?

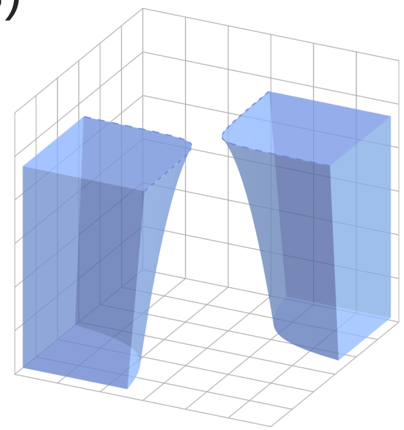
Challenges for partially observed LQG

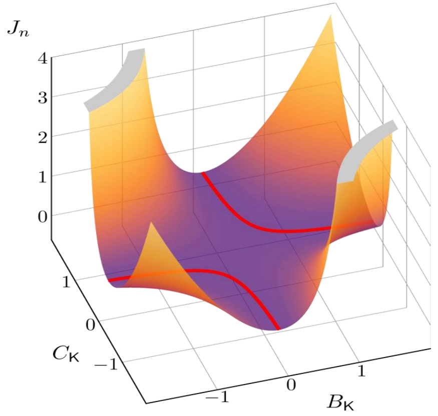
❑ Policy optimization for LQG control

- LQG is more complicated than LQR
- Requires dynamical controllers
- Its non-convex landscape properties are much richer and more complicated than LQR

Our focus: non-convex LQG landscape

- **Q1: Properties of the domain** (set of stabilizing controllers)
  - convexity, connectivity, open/closed?
- **Q2: Properties of the accumulated LQG cost**
  - convexity, differentiability, coercivity?
  - set of stationary points/local minima/global minima?

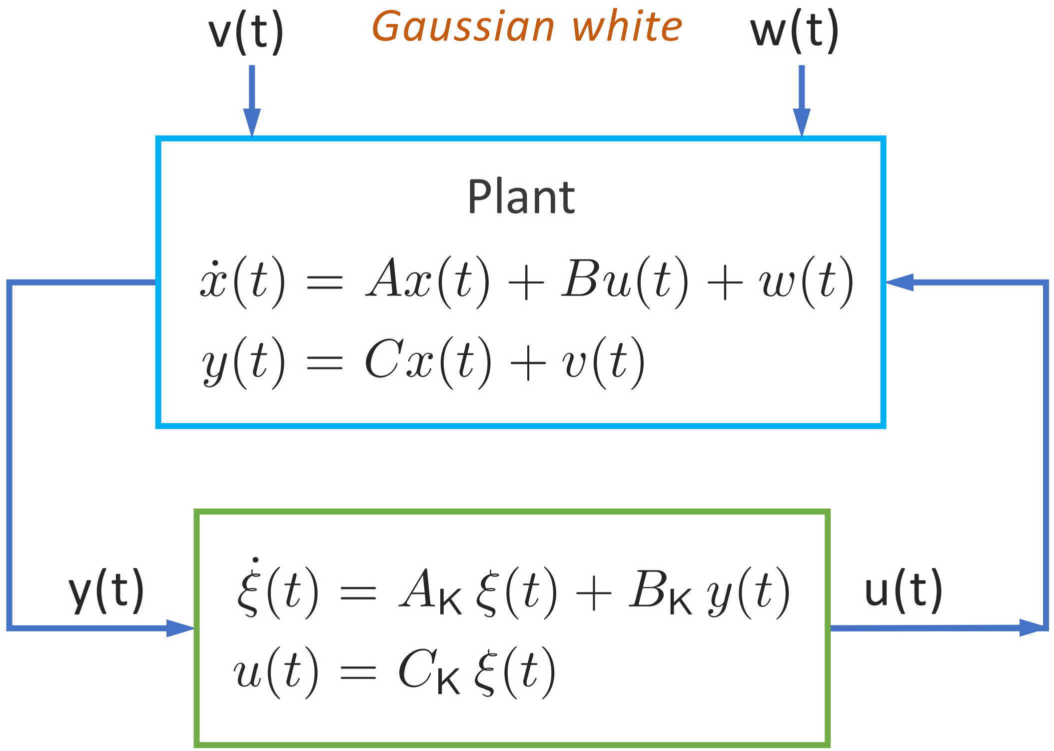




Outline

- LQG problem Setup
- Connectivity of the Set of Stabilizing Controllers
- Structure of Stationary Points of the LQG cost

LQG Problem Setup



dynamical controller

$\mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}})$

Standard Assumption	(A, B), $(A, W^{1/2})$	Controllable
	(C, A), $(Q^{1/2}, A)$	Observable

Objective: The LQG cost

$$\lim_{T\to+\infty} \frac{1}{T}\mathbb{E}\int_0^T (x^\top Q x + u^\top R u)\, dt$$

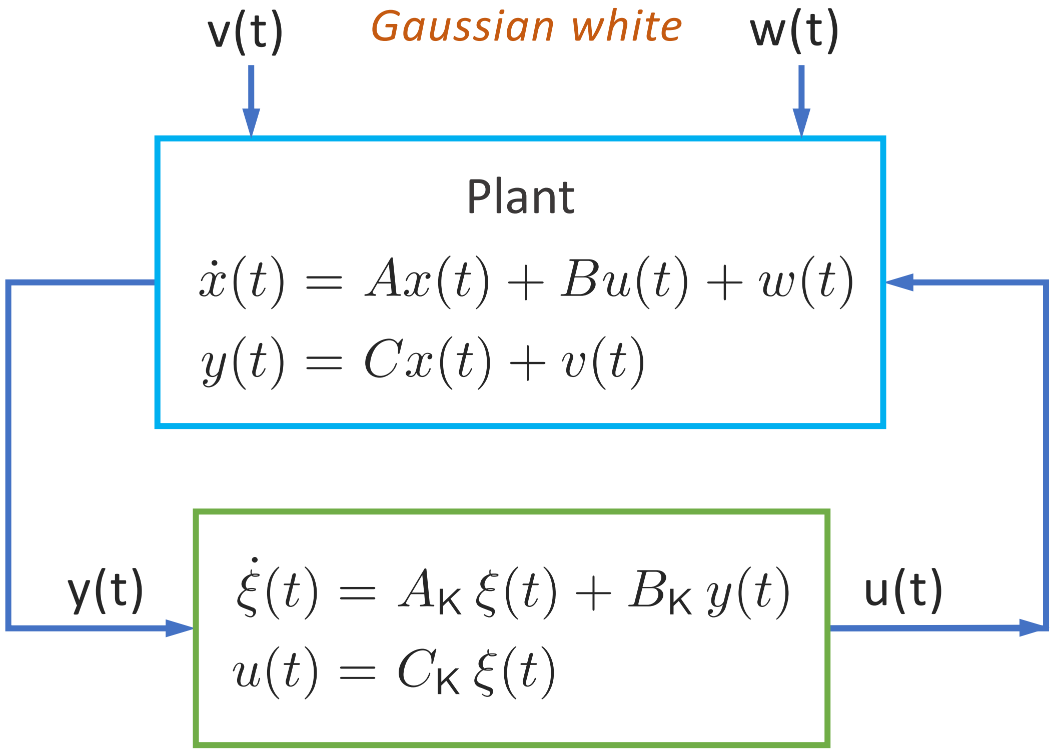
- $\xi(t)$ internal state of the controller
- $\dim \xi(t)$ order of the controller
- $\dim \xi(t) = \dim x(t)$ full-order
- $\dim \xi(t) < \dim x(t)$ reduced-order

Minimal controller

The input-output behavior cannot be replicated by a lower order controller.

* $(A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}})$ controllable and observable

Separation principle



dynamical controller

$$\mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}})$$

Explicit dependence on the dynamics

Objective: The LQG cost

$$\lim_{T \to +\infty} \frac{1}{T} \mathbb{E} \int_0^T (x^\top Q x + u^\top R u)\, dt$$

Solution: Kalman filter for state estimation
+ LQR based on the estimated state

$$\dot{\xi} = (A - BK)\xi + L(y - C\xi),$$
$$u = -K\xi.$$

Two Riccati equations

- **Kalman gain** $L = PC^{\mathsf{T}} V^{-1}$

$$AP + PA^{\mathsf{T}} - PC^{\mathsf{T}} V^{-1} CP + W = 0,$$

- **Feedback gain** $K = R^{-1} B^{\mathsf{T}} S$

$$A^{\mathsf{T}} S + SA - SBR^{-1} B^{\mathsf{T}} S + Q = 0$$

Policy Optimization formulation

Closed-loop dynamics

$$\frac{d}{dt}\begin{bmatrix} x \\ \xi \end{bmatrix} = \boxed{\begin{bmatrix} A & BC_{\mathsf{K}} \\ B_{\mathsf{K}}C & A_{\mathsf{K}} \end{bmatrix}} \begin{bmatrix} x \\ \xi \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & B_{\mathsf{K}} \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix},$$

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & C_{\mathsf{K}} \end{bmatrix} \begin{bmatrix} x \\ \xi \end{bmatrix} + \begin{bmatrix} v \\ 0 \end{bmatrix}.$$

Feasible region of the controller parameters

$$\mathcal{C}_{\text{full}} = \left\{ \mathsf{K} \,\middle|\, \mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \text{ is full order} \right.$$
$$\left. \begin{bmatrix} A & BC_{\mathsf{K}} \\ B_{\mathsf{K}}C & A_{\mathsf{K}} \end{bmatrix} \text{ is Hurwitz stable} \right\}$$

Cost function $\displaystyle \lim_{T\to+\infty} \frac{1}{T}\mathbb{E}\int_0^T (x^\top Q x + u^\top R u)\, dt$

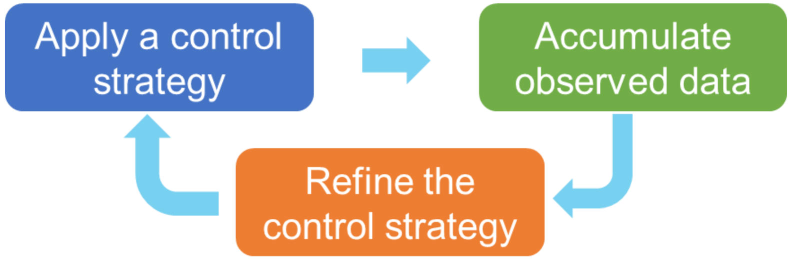
$$J(\mathsf{K}) = \operatorname{tr}\left(\begin{bmatrix} Q & 0 \\ 0 & C_{\mathsf{K}}^{\mathsf{T}} R C_{\mathsf{K}} \end{bmatrix} X_{\mathsf{K}} \right) = \operatorname{tr}\left(\begin{bmatrix} W & 0 \\ 0 & B_{\mathsf{K}} V B_{\mathsf{K}}^{\mathsf{T}} \end{bmatrix} Y_{\mathsf{K}} \right)$$

$X_{\mathsf{K}}, Y_{\mathsf{K}}$ Solution to Lyapunov equations

Policy optimization for LQG

$$\min_{\mathsf{K}} \quad J(\mathsf{K})$$
$$\text{s.t.} \quad \mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \in \mathcal{C}_{\text{full}}$$

Direct policy search $\quad \mathsf{K}_{i+1} = \mathsf{K}_i - \alpha_i \nabla J(\mathsf{K}_i)$

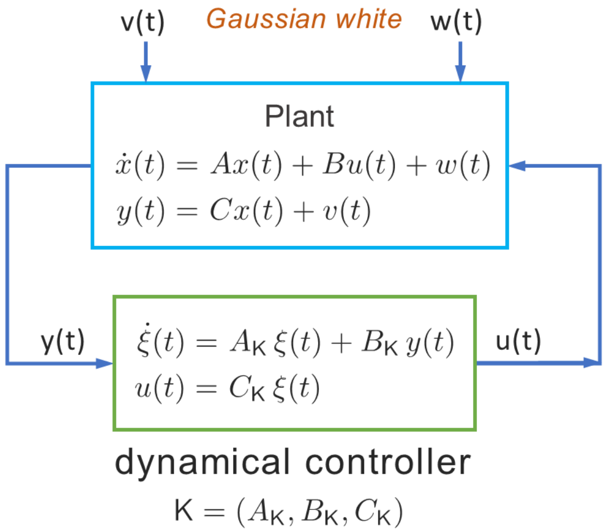


- ✓ **Does it converge at all?**
- ✓ **Converge to which point?**
- ✓ **Convergence speed?**

Optimization Landscape Analysis

Hyland, David, and Dennis Bernstein. "The optimal projection equations for fixed-order dynamic compensation." *IEEE Transactions on Automatic Control* 29.11 (1984): 1034-1037.

Main questions



Non-convex Landscape Analysis

Policy optimization for LQG

$$\min_{\mathsf{K}} \quad J(\mathsf{K})$$
$$\text{s.t.} \quad \mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \in \mathcal{C}_{\text{full}}$$

- **Q1: Connectivity of the feasible region $\mathcal{C}_{\text{full}}$**
  - Is it connected?
  - If not, how many connected components can it have?
- **Q2: Structure of stationary points of $J(\mathsf{K})$**
  - Are there spurious (strictly suboptimal, saddle) stationary points?
  - How to check if a stationary point is globally optimal?

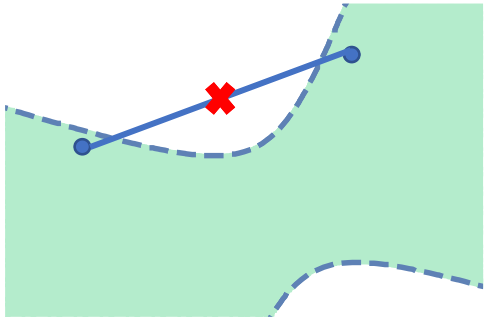
Outline

- ❑ LQG problem Setup
- ❑ **Connectivity of the Set of Stabilizing Controllers**
- ❑ Structure of Stationary Points of the LQG cost

Connectivity of the feasible region

❑ Simple observation: non-convex and unbounded

Lemma 1: the set $\mathcal{C}_{\text{full}}$ is non-empty, unbounded, and can be non-convex.



Example

$$\dot{x}(t) = x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$

$$\mathcal{C}_{\text{full}} = \left\{ \mathsf{K} = \begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \in \mathbb{R}^{2\times 2} \,\middle|\, \begin{bmatrix} 1 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \text{ is stable} \right\}.$$

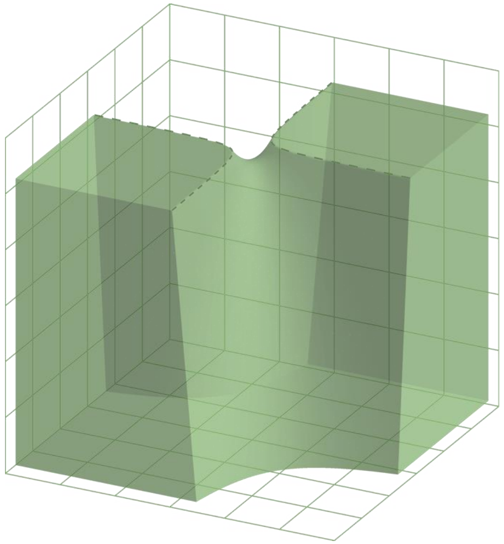
$$\mathsf{K}^{(1)} = \begin{bmatrix} 0 & 2 \\ -2 & -2 \end{bmatrix}, \qquad \mathsf{K}^{(2)} = \begin{bmatrix} 0 & -2 \\ 2 & -2 \end{bmatrix}$$ Stabilize the plant, and thus belong to $\mathcal{C}_{\text{full}}$

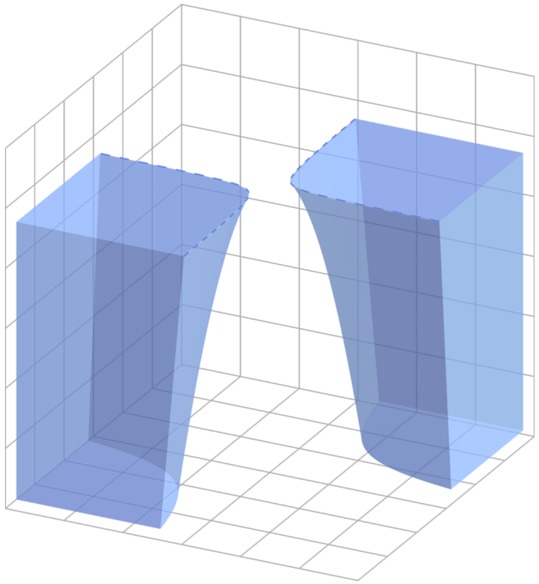
$$\hat{\mathsf{K}} = \frac{1}{2}\left(\mathsf{K}^{(1)} + \mathsf{K}^{(2)}\right) = \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}$$ Fails to stabilize the plant, and thus outside $\mathcal{C}_{\text{full}}$

Connectivity of the feasible region

❑ Main Result 1: dis-connectivity

Theorem 1: The set $\mathcal{C}_{\text{full}}$ can be disconnected but has at most 2 connected components.



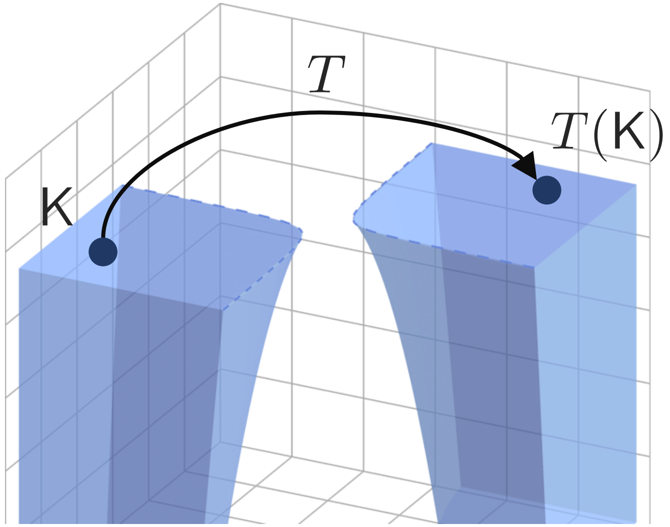


- ✓ Different from the connectivity of static stabilizing state-feedback controllers, which is always connected!
- ✓ **Is this a negative result for gradient-based algorithms?** → **No**

Connectivity of the feasible region

❑ Main Result 2: dis-connectivity

Theorem 2: If $\mathcal{C}_{\text{full}}$ has 2 connected components, then there is a smooth bijection T between the 2 connected components that has the same cost function value.



$J(\mathsf{K}) = J(T(\mathsf{K}))$

✓ In fact, the bijection T is defined by a similarity transformation (change of controller state coordinates)

$$\mathscr{T}_T(\mathsf{K}) := \begin{bmatrix} D_{\mathsf{K}} & C_{\mathsf{K}}T^{-1} \\ TB_{\mathsf{K}} & TA_{\mathsf{K}}T^{-1} \end{bmatrix}.$$

Positive news: For gradient-based local search methods, it makes no difference to search over either connected component.

Connectivity of the feasible region

❑ Main Result 3: conditions for connectivity

Theorem 3: 1) $\mathcal{C}_{\text{full}}$ is connected if there exists a reduced-order stabilizing controller.

2) The sufficient condition above becomes necessary if the plant is single-input or single-output.

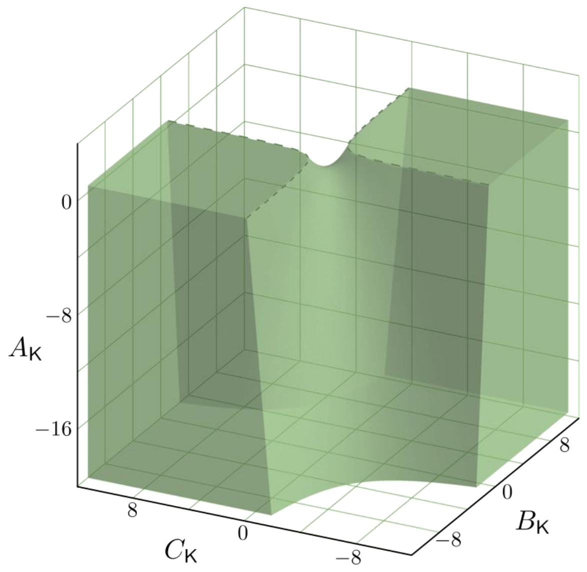
Corollary 1: Given any open-loop stable plant, the set of stabilizing controllers $\mathcal{C}_{\text{full}}$ is connected.

Example: Open-loop stable system

$$\dot{x}(t) = -x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$

Routh--Hurwitz stability criterion

$$\mathcal{C}_{\text{full}} = \left\{ \mathsf{K} = \begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \in \mathbb{R}^{2\times 2} \,\middle|\, A_{\mathsf{K}} < 1, B_{\mathsf{K}} C_{\mathsf{K}} < -A_{\mathsf{K}} \right\}.$$



Connectivity of the feasible region

❑ Main Result 3: conditions for connectivity

<u>**Example: Open-loop unstable system (SISO)**</u>

$$\dot{x}(t) = x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$

- **Routh--Hurwitz stability criterion**

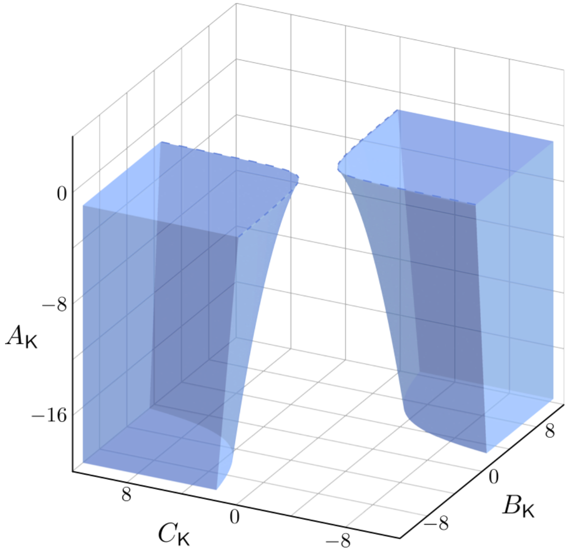
$$\begin{aligned}\mathcal{C}_{\text{full}} &= \left\{ \mathsf{K} = \begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \in \mathbb{R}^{2\times 2} \,\middle|\, \begin{bmatrix} A & BC_{\mathsf{K}} \\ B_{\mathsf{K}}C & A_{\mathsf{K}} \end{bmatrix} \text{ is stable} \right\} \\ &= \left\{ \mathsf{K} = \begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \in \mathbb{R}^{2\times 2} \,\middle|\, A_{\mathsf{K}} < -1,\ B_{\mathsf{K}}C_{\mathsf{K}} < A_{\mathsf{K}} \right\}.\end{aligned}$$

- **Two path-connected components**

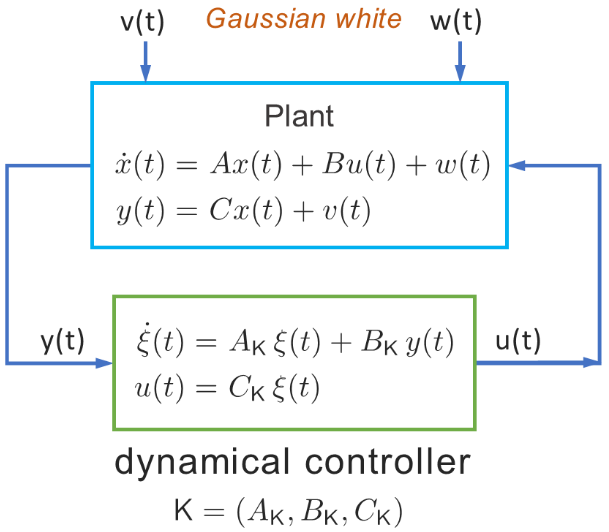
$$\mathcal{C}_1^+ := \left\{ \mathsf{K} = \begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \in \mathbb{R}^{2\times 2} \,\middle|\, A_{\mathsf{K}} < -1,\ B_{\mathsf{K}}C_{\mathsf{K}} < A_{\mathsf{K}},\ B_{\mathsf{K}} > 0 \right\},$$

$$\mathcal{C}_1^- := \left\{ \mathsf{K} = \begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix} \in \mathbb{R}^{2\times 2} \,\middle|\, A_{\mathsf{K}} < -1,\ B_{\mathsf{K}}C_{\mathsf{K}} < A_{\mathsf{K}},\ B_{\mathsf{K}} < 0 \right\}.$$

Disconnected feasible region



Policy Optimization formulation



Policy optimization for LQG

$$\min_{\mathsf{K}} \quad J(\mathsf{K})$$
$$\text{s.t.} \quad \mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \in \mathcal{C}_{\text{full}}$$

- **Q1: Connectivity of the feasible region $\mathcal{C}_{\text{full}}$**
  - Is it connected? **No**
  - If not, how many connected components can it have? **Two**
- **Q2: Structure of stationary points of $J(\mathsf{K})$**
  - Are there spurious (strictly suboptimal, saddle) stationary points?
  - How to check if a stationary point is globally optimal?

Non-convex Landscape Analysis

Outline

- ❑ LQG problem Setup
- ❑ Connectivity of the Set of Stabilizing Controllers
- ❑ **Structure of Stationary Points of the LQG cost**

Structure of Stationary Points

Simple observations

1) $J(\mathsf{K})$ is a real analytic function over its domain (smooth, infinitely differentiable)
2) $J(\mathsf{K})$ has **non-unique** and **non-isolated** global optima

$$\dot{\xi}(t) = A_{\mathsf{K}}\,\xi(t) + B_{\mathsf{K}}\,y(t)$$
$$u(t) = C_{\mathsf{K}}\,\xi(t)$$

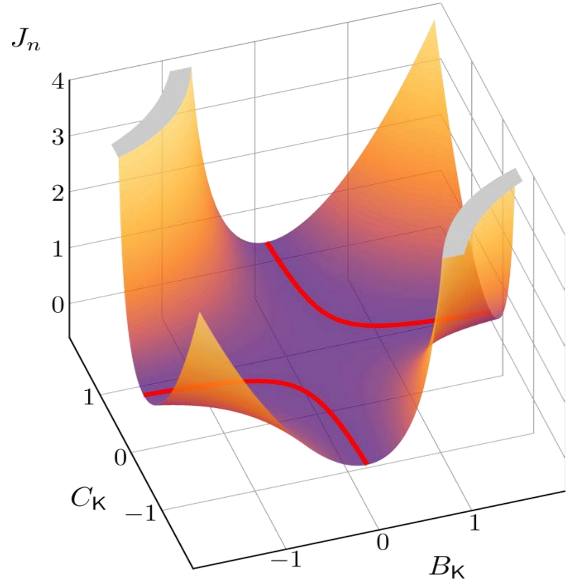
Similarity transformation

$$(A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \mapsto (TA_{\mathsf{K}}T^{-1}, TB_{\mathsf{K}}, C_{\mathsf{K}}T^{-1})$$

- $J(\mathsf{K})$ is invariant under similarity transformations.
- It has many stationary points, unlike the LQR with a unique stationary point

Policy optimization for LQG

$$\begin{aligned} \min_{\mathsf{K}} \quad & J(\mathsf{K}) \\ \text{s.t.} \quad & \mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \in \mathcal{C}_{\text{full}} \end{aligned}$$



Structure of Stationary Points

Gradient computation

Lemma 2: For every $\mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \in \mathcal{C}_{\text{full}}$, we have

$$\frac{\partial J(\mathsf{K})}{\partial A_{\mathsf{K}}} = 2\left(Y_{12}^{\mathsf{T}} X_{12} + Y_{22} X_{22}\right),$$
$$\frac{\partial J(\mathsf{K})}{\partial B_{\mathsf{K}}} = 2\left(Y_{22} B_{\mathsf{K}} V + Y_{22} X_{12}^{\mathsf{T}} C^{\mathsf{T}} + Y_{12}^{\mathsf{T}} X_{11} C^{\mathsf{T}}\right),$$
$$\frac{\partial J(\mathsf{K})}{\partial C_{\mathsf{K}}} = 2\left(R C_{\mathsf{K}} X_{22} + B^{\mathsf{T}} Y_{11} X_{12} + B^{\mathsf{T}} Y_{12} X_{22}\right),$$

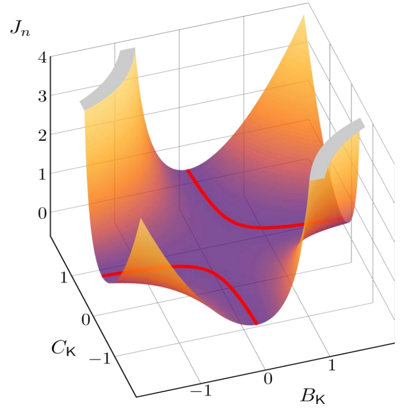
where $X_{\mathsf{K}} = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^{\mathsf{T}} & X_{22} \end{bmatrix}, \quad Y_{\mathsf{K}} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^{\mathsf{T}} & Y_{22} \end{bmatrix}$

are the unique positive semidefinite solutions to two Lyapunov equations.

How does the set of Stationary Points look like?

$$\left\{ \mathsf{K} \in \mathcal{C}_{\text{full}} \;\middle|\; \frac{\partial J(\mathsf{K})}{\partial A_{\mathsf{K}}} = 0, \; \frac{\partial J(\mathsf{K})}{\partial B_{\mathsf{K}}} = 0, \; \frac{\partial J(\mathsf{K})}{\partial C_{\mathsf{K}}} = 0, \right\}$$

- Non-unique, non-isolated
- Local minimum, local maximum, saddle points, or globally minimum?



Structure of Stationary Points

❑ Main Result: existences of strict saddle points

Theorem 4: Consider any open-loop stable plant. The zero controller with any stable A_{K}

$$\mathsf{K} = (A_{\mathsf{K}}, 0, 0) \in \mathcal{C}_{\text{full}}$$

is a stationary point. Furthermore, the corresponding hessian is either indefinite (**strict saddle point**) or equal to zero (**high-order saddle or else**).

Example:

$$\dot{x}(t) = -x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$

$$Q = 1, R = 1, V = 1, W = 1$$

Stationary point: $\mathsf{K}^\star = \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} \in \mathbb{R}^{2\times 2}, \qquad \text{with } a < 0$

- **Cost function:** $J\left(\begin{bmatrix} 0 & C_{\mathsf{K}} \\ B_{\mathsf{K}} & A_{\mathsf{K}} \end{bmatrix}\right) = \dfrac{A_{\mathsf{K}}^2 - A_{\mathsf{K}}(1 + B_{\mathsf{K}}^2 C_{\mathsf{K}}^2) - B_{\mathsf{K}} C_{\mathsf{K}}(1 - 3B_{\mathsf{K}} C_{\mathsf{K}} + B_{\mathsf{K}}^2 C_{\mathsf{K}}^2)}{2(-1 + A_{\mathsf{K}})(A_{\mathsf{K}} + B_{\mathsf{K}} C_{\mathsf{K}})}.$

- **Hessian:** $\left.\begin{bmatrix} \frac{\partial J^2(\mathsf{K})}{\partial A_{\mathsf{K}}^2} & \frac{\partial J^2(\mathsf{K})}{\partial A_{\mathsf{K}} \partial B_{\mathsf{K}}} & \frac{\partial J^2(\mathsf{K})}{\partial A_{\mathsf{K}} \partial C_{\mathsf{K}}} \\ \frac{\partial J^2(\mathsf{K})}{\partial B_{\mathsf{K}} A_{\mathsf{K}}} & \frac{\partial J^2(\mathsf{K})}{\partial B_{\mathsf{K}}^2} & \frac{\partial J^2(\mathsf{K})}{\partial B_{\mathsf{K}} \partial C_{\mathsf{K}}} \\ \frac{\partial J^2(\mathsf{K})}{\partial C_{\mathsf{K}} A_{\mathsf{K}}} & \frac{\partial J^2(\mathsf{K})}{\partial C_{\mathsf{K}} B_{\mathsf{K}}} & \frac{\partial J^2(\mathsf{K})}{\partial \partial C_{\mathsf{K}}^2} \end{bmatrix}\right|_{\mathsf{K}^\star = \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix}} = \dfrac{1}{2(1-a)} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$

Indefinite with eigenvalues:

$0 \text{ and } \pm\frac{1}{2(1-a)}$

Structure of Stationary Points

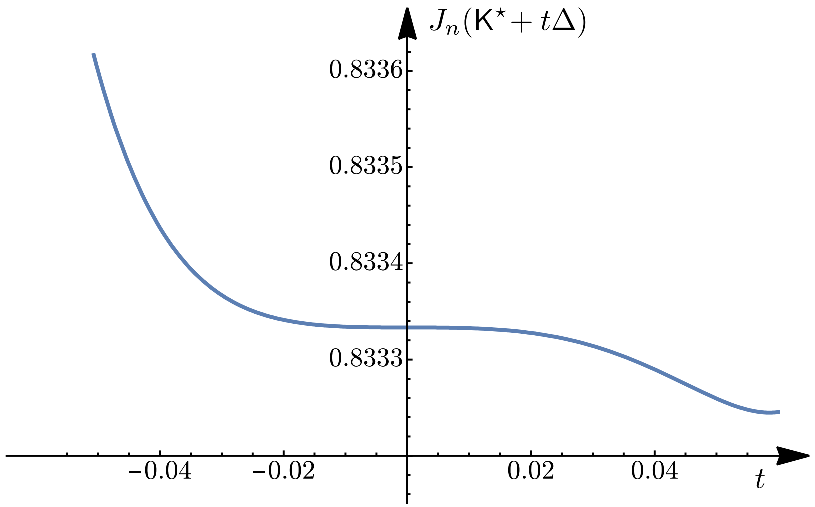
❑ Main Result: existences of strict saddle points

Theorem 4: Consider any open-loop stable plant. The zero controller with any stable A_{K}

$$\mathsf{K} = (A_{\mathsf{K}}, 0, 0) \in \mathcal{C}_{\text{full}}$$

is a stationary point. Furthermore, the corresponding hessian is either indefinite (**strict saddle point**) or equal to zero (**high-order saddle or else**).

Another example with zero Hessian



How does the set of Stationary Points look like?

$$\left\{ \mathsf{K} \in \mathcal{C}_{\text{full}} \;\middle|\; \begin{array}{l} \dfrac{\partial J(\mathsf{K})}{\partial A_{\mathsf{K}}} = 0, \\ \dfrac{\partial J(\mathsf{K})}{\partial B_{\mathsf{K}}} = 0, \\ \dfrac{\partial J(\mathsf{K})}{\partial C_{\mathsf{K}}} = 0, \end{array} \right\}$$

- ❑ **Non-unique, non-isolated**
- ❑ **Strictly suboptimal points; Strict saddle points**
- ❑ **All bad stationary points correspond to non-minimal controllers**

Structure of Stationary Points

❑ Main Result

Theorem 5: All stationary points corresponding to controllable and observable controllers are globally optimum.

$$\left\{ \mathsf{K} \in \mathcal{C}_{\text{full}} \;\middle|\; \frac{\partial J(\mathsf{K})}{\partial A_{\mathsf{K}}} = 0, \frac{\partial J(\mathsf{K})}{\partial B_{\mathsf{K}}} = 0, \frac{\partial J(\mathsf{K})}{\partial C_{\mathsf{K}}} = 0, \right\}$$

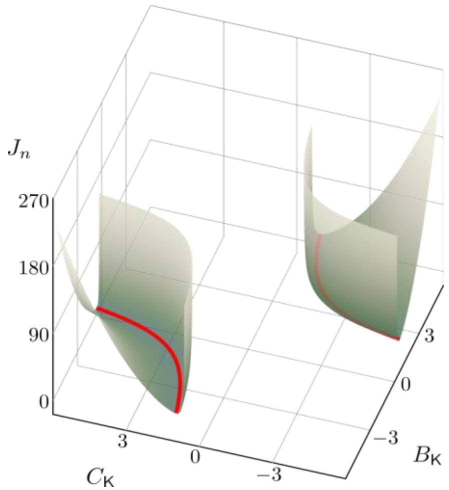
Local Zero Gradient + Structural Information → **Global Optimality Certificate**

Particularly, given a stationary point that is a **minimal** controller

1) It is **globally optimal**, and the set of all global optima forms a manifold with 2 connected components.

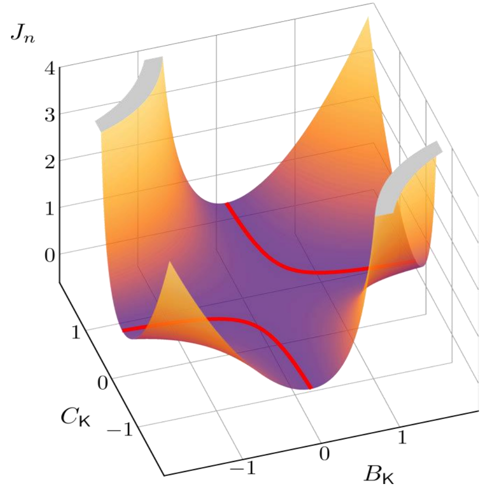
Example: open-loop unstable system

$$\dot{x}(t) = x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$



Example: open-loop stable system

$$\dot{x}(t) = -x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$



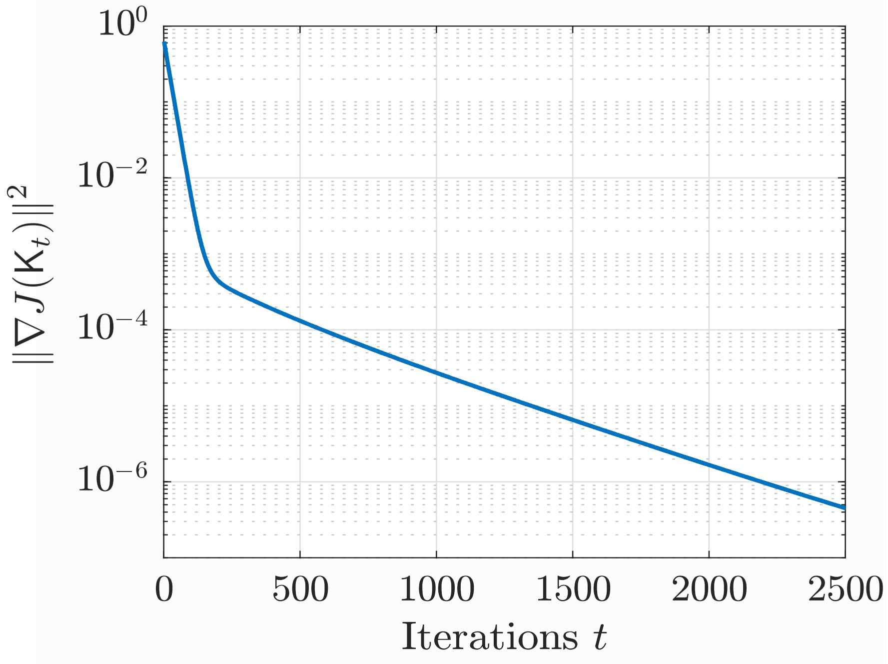
Structure of Stationary Points

❑ Implication

Corollary: Consider gradient descent iterations

$$\mathsf{K}_{t+1} = \mathsf{K}_t - \alpha \nabla J(\mathsf{K}_t)$$

If the iterates converge to a minimal controller, then this minimal controller is a global optima.



More questions:

- ✓ **Escaping saddle points**?
- ✓ Convergence conditions?
- ✓ Convergence speed?
- ✓ Alternative model-free parameterization?

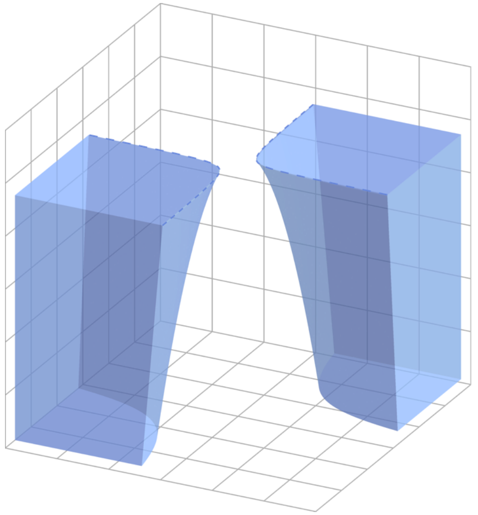
Comparison with LQR

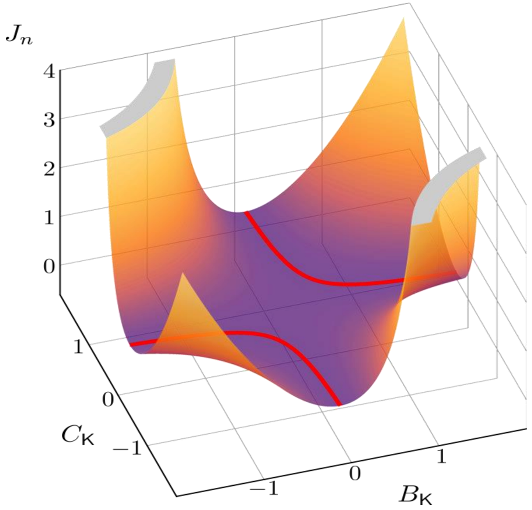
	Policy optimization for LQR $$\min_{K} \ J(K) \quad \text{s.t.} \ K \in \mathcal{K}$$	**Policy optimization for LQG** $$\min_{\mathsf{K}} \ J(\mathsf{K}) \quad \text{s.t.} \ \mathsf{K} = (A_{\mathsf{K}}, B_{\mathsf{K}}, C_{\mathsf{K}}) \in \mathcal{C}_{\text{full}}$$
Connectivity of feasible region	❖ Always connected	❖ Disconnected, but at most 2 connected comp. ❖ They are almost identical to each other
Stationary points	❖ Unique	❖ Non-unique, non-isolated stationary points ❖ Spurious stationary points (strict saddle, nonminimal controller) ❖ **All mini. stationary points are globally optimal**
Gradient Descent	❖ Gradient dominance ❖ Global fast convergence (like strictly convex)	❖ No gradient dominance ❖ Local convergence/speed (**unknown**) ❖ **Many open questions**
References	Fazel et al., ICML, 2018; Malik et al., 2019; Mohammadi et al., IEEE TAC, 2020; Li et al., 2019; K. Zhang, B. Hu, and T. Başar, 2021; Furieri et al., 2019; Feiran Zhao & Keyou You, 2021, and many others	Zheng*, Tang*, Li. 2021, link (* equal contribution)

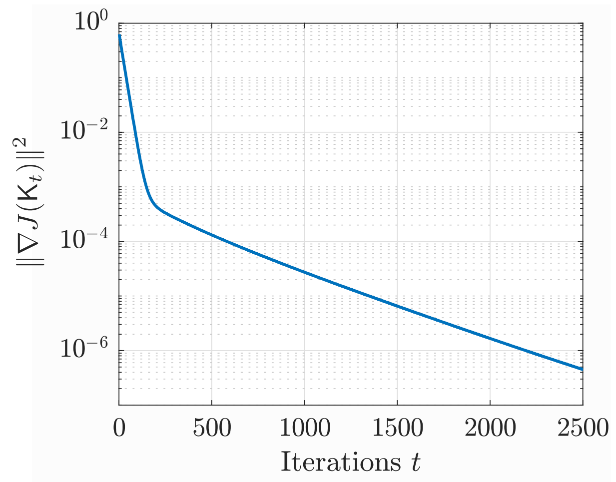
Conclusions

Policy optimization for LQG control

- Much richer and more complicated than LQR
- Disconnected, but at most 2 connected components
- Non-unique, non-isolated stationary points, strict saddle points
- Minimal (controllable and observable) stationary points are globally optimal







Ongoing and Future work

- ❑ How to certify the optimality of a non-minimal stationary point
- ❑ Perturbed policy gradient (PGD) for escaping saddle points
- ❑ Quantitative analysis of PGD algorithms for LQG
- ❑ Alternative model-free parametrization of dynamical controllers (e.g., Makdah & Pasqualetti, 2023; Zhao, Fu &You, 2022.)
  - ✓ Better optimization landscape structures, smaller dimension
- ❑ Nonconvex Landscape of Hinf dynamical output feedback control (Tang & Zheng, 2023 https://arxiv.org/abs/2304.00753;)

Analysis of the Optimization Landscape of Linear Quadratic Gaussian (LQG) Control

Thank you for your attention!

Q & A

1. Y. Tang*, Y. Zheng*, and N. Li, "Analysis of the optimization landscape of Linear Quadratic Gaussian (LQG) control," Mathematical Programming, 2023. Available: https://arxiv.org/abs/2102.04393 *Equal contribution
2. B. Hu and Y. Zheng, "Connectivity of the feasible and sublevel sets of dynamic output feedback control with robustness constraints," IEEE Control Systems Letters, 2022.
3. Y. Zheng*, Y. Sun*, M. Fazel, and N. Li. "Escaping High-order Saddles in Policy Optimization for Linear Quadratic Gaussian (LQG) Control." CDC, 2022 https://arxiv.org/abs/2204.00912. *Equal contribution

Role of convex parameterization

Message : favorable landscape properties for nonconvex J can be obtained *from* the convex parameterization under appropriate conditions on the mapping
[Sun, F.,'21 ; Umenberger et al.'22 ; Hu et al.'23 survey]

Warm up : convex formulation for continuous-time LQR

$$\min_{Z,L,P} Tr(QP + ZR)$$
$$\text{s.t. } AP + PA^T + BL + LB^T + \Sigma = 0,$$
$$P \succ 0, \qquad \begin{bmatrix} Z & L \\ L^\top & P \end{bmatrix} \succeq 0$$

$$\Rightarrow \qquad \min_{Z,L,P} \quad f(L,P,Z) \qquad \text{s.t.} \quad (L,P,Z) \in \mathcal{S}$$

where $K^* = L^*(P^*)^{-1}$.

- further, $K = LP^{-1}$ parameterizes *all* stabilizing $K \in \mathcal{K}$
- also see [Mohammadi et al.'19]

Assumptions on parameterization map

$$\min_{K} \quad J(K) \qquad\qquad \Rightarrow \qquad\qquad \min_{Z,L,P} \quad f(L,P,Z)$$
$$\text{s.t.} \quad K \in \mathcal{K} \qquad\qquad\qquad\qquad \text{s.t.} \quad (L,P,Z) \in \mathcal{S}$$

Assumptions :

1. $\mathcal{S}$ is convex, $f(L,P,Z)$ is convex, bounded, differentiable on $\mathcal{S}$.
2. we can express $J(K)$ as

$$J(K) = \min_{L,P,Z} f(L,P,Z), \ \text{s.t.}\ (L,P,Z) \in \mathcal{S},\ K = LP^{-1}.$$

more generally, $K = LP^{-1}$ can be replaced by a surjective map $K = \Phi(L,P)$ with "nicely behaved" first-order derivatives.

[Sun, F.,'21], [Umenberger et al.,'22]

Role of convex parameterization

$$\min_{K} \quad J(K) \qquad \text{s.t.} \quad K \in \mathcal{K}$$

$$\min_{Z,L,P} \quad f(L,P,Z), \qquad \text{s.t.} \quad (L,P,Z) \in \mathcal{S}$$

Theorem (simplified) [Sun & F.,'21]

Under assumptions 1 and 2,

$$\nabla J(K) = 0 \iff K = K^*.$$

Also,

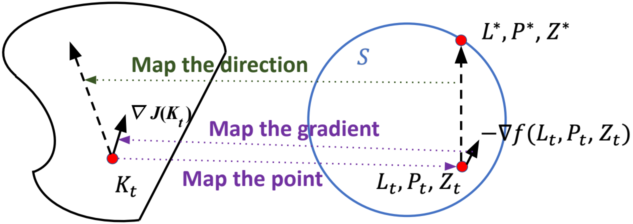
- If f is convex, $\|\nabla J(K)\|_F \gtrsim J(K) - (K^*)$.
- If f is μ-strongly convex, $\|\nabla J(K)\|_F \gtrsim (\mu(J(K) - J(K^*)))^{1/2}$.

($\gtrsim$ hides instance-dependent constants ; depend on system parameters & initial point K_0)

Proof picture

$$\min_{K\in\mathcal{K}} \quad J(K)$$

$$\min_{Z,L,P} \quad f(L,P,Z),$$
$$\text{s.t.,} \quad (L,P,Z)\in\mathcal{S}$$



Role of convex parameterization

A general version that applies to non-smooth $J(K)$ as well :

Theorem [Hu et al.,'23]

Suppose $J(K)$ is differentiable or subdifferentially regular, Assumptions 1, 2 hold. For any K satisfying $J(K) > J(K^*)$, there exists non-zero V in the descent cone of $\mathcal{K}$ at K, such that

$$0 < J(K) - J(K^*) \leq -J'(K, V),$$

so any stationary point of J is a global minimum.

$J'(K, V)$ denotes directional derivative of $J(K)$ along direction V. When J is differentiable, $J'(K, V) = Tr(V^T \nabla J(K))$.

Example : Continuous time LQR

$$\min_{Z,L,P} \; f(L,P,Z) := Tr(QP) + Tr(ZR)$$

$$\text{s.t.,} \; \mathcal{A}(P) + \mathcal{B}(L) + \Sigma = 0, \;\; G \succ 0,$$

$$\begin{bmatrix} Z & L^\top \\ L & G \end{bmatrix} \succeq 0$$

Question : $K = LP^{-1}$, is P always invertible ? (yes, if initial x_0 has full-rank covariance)

L, P, P^{-1} are bounded in the sublevel set $\{K : \; J(K) \leq a\}$.

then : $a \geq J(K) = Tr(QP) + Tr(LP^{-1}L^\top R)$.

Example : Continuous time LQR

L, P, P^{-1} are bounded in the sublevel set $\{K : \ (K) < a\}$.

Define

$$\nu = \frac{\lambda_{\min}^2(\Sigma)}{4} \left(\sigma_{\max}(A)\lambda_{\min}^{-1/2}(Q) + \sigma_{\max}(B)\lambda_{\min}^{-1/2}(R)\right)^{-2},$$

then

$$\|J(K)\| \leq -C_1(J(K) - J(K^*))$$

where

$$C_1 = \frac{\nu\lambda_{\min}^{1/2}(Q)\lambda_{\min}^{1/2}(R)}{4a^4} \cdot \min\left\{a^2, \ \nu\lambda_{\min}(Q)\right\}.$$

Related Results

Many other landscape results rely on connections to LMIs

$\mathcal{H}_\infty$ **landscape : Clarke stationary is global** [Guo et al., 2022]

Dynamic filtering : Differentiable convex lifting [Umenberger et al., 2022]

LQG : Connectivity [Y. Zheng, 2023]

Output-feedback $\mathcal{H}_\infty$ **: Connectivity** [Hu et al., 2022]

A general tool for landscape study. More study is needed for output-feedback problems !

Future Work

The last section of our survey article lists several directions :

- Further connections between optimization and control theory, e.g. complexity of escaping saddles for output feedback problems
- Advanced regularization for stability, robustness, and safety
- Nonlinear systems, deep RL, and perception-based control
- Multi-agent systems and decentralized control
- Integration of model-based and model-free methods
- New PO formulations from machine learning

And many more which are not listed in our article !