On the Robustness and Convergence of Policy Optimization in Continuous-Time Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Stochastic Control

Lekan Molu

Microsoft Research

New York City, NY 10012

Presented by **Lekan Molu** (Lay-con Mo-lu)

November 2, 2023

Talk Outline and Overview

- Policy Optimization and Stochastic Linear Control
  - Connections to risk-sensitive control;
  - Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control theory.
- The case for convergence analysis in stochastic PO.
  - Kleinman's algorithm, *redux.*
  - Kleiman's algorithm in an iterative best response setting;
  - PO Convergence in best response settings.
- Robustness margins in model- and sampling- settings.
  - PO as a discrete-time nonlinear system;
  - Kleiman and input-to-state-stability;
  - Robust policy optimization as a small-input stable state optimization algorithm

Research Significance

- (Deep) RL and modern AI
  - Robotic manipulation (Levine et al., 2016), text-to-visual processing (DALL-E), Atari games (Mnih et al., 2013), e.t.c.
  - Policy optimization (PO) is fundamental to modern AI algorithms' success.
  - Major success story: functional mapping of observations to policies.
  - But how does it work?

Policy Optimization – General Framework

- PO encapsulates policy gradients (Kakade, 2001) or PG, actor-critic methods (Vrabie and Lewis, 2011), trust region PO Schulman et al. (2015), and proximal PO methods (Schulman et al., 2017).
- PG particularly suitable for complex systems.

$$\begin{aligned} &\min J(K) \\ &\text{subject to } K \in \mathcal{K} \end{aligned} \tag{1}$$

where $\mathcal{K} = \{K_1, K_2, \cdots, K_n\}$.

- $J(K)$ could be tracking error, safety assurance, goal-reaching measure of performance e.t.c. required to be satisfied.

Continuous-time RL control applications

- A little randomness in a system's mathematical model coefficients?
  - Population growth model: $dN/dt = a(t)N(t), \; N(0) = N_0$; growth rate $a(t)$ subject to random effects e.g. $a(t) = r(t) +$ "noise".
  - We only know the distribution of "noise".
- Filtering and state estimation problems where the nature of the noise is unknown, but it is observed via sensor measurements.
  - Kalman + Bucy Filters – aerospace (Apollo, Mariner etc.).

Continuous-time RL control applications

- Semielliptic P.D.E.s with Dirichlet boundary value problems e.g. slender flexible rods, Cosserat dynamics etc: $\Delta q = \sum_{i=1}^{n} \dfrac{\partial^2 q}{\partial \xi_i^2} = 0 \in \Omega,\ q = q_{\rightarrow}$ on $\partial\Omega,\ \Omega \subset \mathbb{R}^n$
- An economic portfolio problem where the price, $p(t)$, of a stock satisfies a stochastic differential equation e.g. $dp/dt = (a + \alpha \cdot \text{``}noise\text{''})p$ for $a > 0$, $\alpha \in reline$.
- Call options pricing: The *Black-Scholes option price formula*.

Policy Optimization – Open questions

- Gradient-based data-driven methods: prone to divergence from true system gradients.
  - Challenge I: Optimization occurs in non-convex objective landscapes.
    - Get performance certificates as a mainstay for control design: Coerciveness property (Hu et al., 2023).
  - Challenge II: Taming PG's characteristic high-variance gradient estimates (REINFORCE, NPG, Zeroth-order approx.).
    - Hello, (linear) robust ($\mathcal{H}_\infty$-synthesis) control!

Policy Optimization – Open questions

- Challenge III: Under what circumstances do we have convergence to a desired equilibrium in RL settings?
- Challenge IV: Stochastic control, not deterministic control settings.
  - models involving round-off error computations in floating point arithmetic calculations; the stock market; protein kinetics.
- Challenge V: Continuous-time RL control.
  - Very little theory. Lots of potential applications encompassing rigid and soft robotics, aerospace or finance engineering, protein kinetics.

$\mathcal{H}_\infty$-Control Under Model Mismatch

$$
\begin{aligned}
\mathrm{d}x(t) &= Ax(t)\mathrm{d}t + Bu(t)\mathrm{d}t + D\mathrm{d}w(t), \\
z(t) &= Cx(t) + Eu(t), \;\; \alpha > 0;
\end{aligned}
$$

Algorithm 1 Search for the closed-loop $\mathcal{H}_\infty$-norm

Given a user-defined step size $\eta > 0$
Set the initial upper bound on γ as $\gamma_{ub} = \infty$.
Initialize a buffer for possible $\mathcal{H}_\infty$ norms for each K_1 to be found, $\Gamma_{buf} = \{\}$.
Initialize ordered poles $\mathcal{P} = \{p_i \in Re(s) < 0 \,|\, i = 1, 2, \}$ $\triangleright$ $p_1 < p_2 < \cdots$
for $p_i \in \mathcal{P}$ **do**
 Place p_i on (2); $\triangleright$ (Tits and Yang, 1996)
 Compute stabilizing $K_1^{p_i}$
 Find lower bound γ_{lb} for $H(\gamma, K_1^{p_i})$; $\triangleright$ using (22)
 $\Gamma_{buf}(i) =$ get_hinf_norm$(T_{zw}, \gamma_{lb}, K_1^{p_i})$.
end for
function get_hinf_norm$(T_{zw}, \gamma_{lb}, K_1^{p_i})$
 while $\gamma_{ub} = \infty$ **do**
  $\gamma := (1 + 2\eta)\, \gamma_{lb}$;
  Get $\lambda_i(H(\gamma, K_1^{p_i}))$ $\triangleright$ c.f. (14)
  if Re$(\Lambda) \neq \emptyset$ for $\Lambda = \{\lambda_1, \cdots \lambda_n\}$ **then**
   Set $\gamma_{ub} = \gamma$; exit
  else
   Set buffer $\Gamma_{lb} = \{\}$
   for $\lambda_k \in \{\text{Imag}(\Lambda)_{:p-1}\}$ **do** $\triangleright$ $k = 1$ to K
    Set $m_k = \frac{1}{2}(\omega_k + \omega_{k+1})$
    Set $\Gamma_{lb}(k) = \max\{\sigma\,[T_{zw}(jm_k)]\}$;
   end for
   $\gamma_{lb} = \max(\Gamma_{lb})$
  end if
  Set $\gamma_{ub} = \frac{1}{2}(\gamma_{lb} + \gamma_{ub})$.
 end while
 return γ_{ub}
end function

Tools: Complexity, Convergence, Robustness.

- Risk-sensitive $\mathcal{H}_\infty$-control (Glover, 1989) and discrete- and continuous-time mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design (Khargonekar et al., 1988; Hu et al., 2023):
    - min. upper bound on $\mathcal{H}_2$ cost subject to satisfying a set of risk-sensitive (often $\mathcal{H}_\infty$) constraints (Basar, 1990):

$$min_{K \in \mathcal{K}} J(K) := Tr(P_K D D^\top) \tag{2}$$
$$\text{subject to } \mathcal{K} := \{K | \rho(A - BK) < 1, \|T_{zw}(K)\|_\infty < \gamma\}$$

    - P_K: solution to the generalized algebraic Riccati equation (GARE);
    - A, B, D, K: standard closed-loop system matrices;
    - $\|T_{zw}(K)\|_\infty$: $\mathcal{H}_\infty$-norm of the closed-loop transfer function from a disturbance input w to output z.

Tools: Complexity, Convergence, Robustness.

Infinite-horizon

- discrete-time deterministic LQR settings (Fazel et al., 2018):

$$\min_{K \in \mathcal{K}} \mathbb{E} \sum_{t=0}^{\infty} (x_t^\top Q x_t + u_t^\top R u_t) \text{ s.t. } x_{t+1} = A x_t + B u_t, x_0 \sim \mathcal{P}_0$$

- discrete-time LQ problems under multiplicative noise (Gravell et al., 2021):
  $\min_{\pi \in \Pi} \mathbb{E}_{x_0, \{\delta_i\}, \{\gamma_i\}\}} \sum_{t=0}^{\infty} (x_t^\top Q x_t + u_t^\top R u_t)$
  subject to $x_{t+1} = (A + \sum_{i=1}^{p} \delta_{ti} A_i) x_t + (B + \sum_{i=1}^{q} \gamma_{ti} B_i) u_t;$

(Non-exhaustive) Lit. Landscape on PO Theory

Literature landscape	Cont. time (Kalman '61, Luenberger '63)	Stochastic. LQR (Kalman '60)	Cont. Phase	LEQG or Mixed H_2/H_∞	Finite/Infinite Horizon
Fazel (2018)	No	No	Yes	No	Finite-horizon
Mohammadi (TAC -- 2020)	Yes	No	Yes	No	Finite-Horizon
Zhang (2019)	Yes	Yes (Gaussian)	Yes	Yes	Inf-horizon
Gravell (2021)	No	Multiplicative	Yes	No	Inf-horizon
Zhang (2020)	No	No	Yes	Yes	Rand-horizon
Molu (2022)	Yes	Yes (Brownian)	Yes	Yes	Inf-Horizon
Cui & Molu (2023)	Yes	Yes (Brownian)	Yes	Yes	Inf-Horizon

Mainstay

- Continuous-time infinite-dimensional linear systems.
  - Disturbances enter additively as random stochastic Wiener processes.
  - Many natural systems admit uncertain additive Brownian noise as diffusion processes.
    - Theoretical analysis machinery: Îto's stochastic calculus.
- Goal: keep controlled process, z, small i.e.

$$\|z\|_2 = \left(\int |z(t)|^2 dt \right)^{1/2},$$

  - Under a minimizing $u(x(t)) \in \mathcal{U}$ in spite of unforeseen $w(t) \in \mathcal{W} \subseteq \mathbb{R}^q$.

Minimization Objective and Risk-Sensitive Control

- Risk-sensitive linear exponential quadratic Gaussian objective functional (Jacobson, 1973):

$$\min_{u \in \mathcal{U}} \mathcal{J}_{exp}(x_0, u, w) = \mathbb{E}\Big|_{x_0 \in \mathcal{P}_0} \exp\left[\frac{\alpha}{2}\int_0^\infty z^\top(t) z(t) \mathrm{d}t\right],$$

$$\text{subject to } \mathrm{d}x(t) = Ax(t)\mathrm{d}t + Bu(t)\mathrm{d}t + D\mathrm{d}w(t), \quad z(t) = Cx(t) + Eu(t), \;\; \alpha > 0; \tag{3}$$

- where $dw/dt = \mathcal{N}(0, W)$, $x_0 = \mathcal{N}(0, \mu)$, and $(x_0, w(t)) \subseteq (\Omega, \mathcal{F}, \mathcal{P})$.

Minimization Objective and Risk-Sensitive Control

- A Taylor series expansion of (3) reveals:

$$\mathcal{J}_{exp}(x_0, u, w) =$$

$$\lim_{T \to \infty} \mathbb{E}\Big|_{x_0 \in \mathcal{P}_0} \left[\frac{\alpha}{2} \sum_{t=0}^{T} z^\top(t) z(t)\right] + \frac{\alpha^2}{4} var \left[\sum_{t=0}^{T} z^\top(t) z(t)\right]. \tag{4}$$

- Consider the variance term $\frac{\alpha^2}{4} var \left[\sum_{t=0}^{T} z^\top(t) z(t)\right] \to \epsilon$.
  - α a measure of risk-propensity if $\alpha > 0$;
  - α a measure of risk-aversion if $\alpha < 0$;
  - $\alpha = 0$ implies solving a classic LQP.

RL PO as a Risk-Sensitive Control Problem

- RL (via PG) computes high-variance gradient estimates from Monte-Carlo trajectory roll-outs and bootstrapping.
- If we set $\alpha > 0$ in the LEQG problem (3), we have a controlled setting where we can study the theoretical properties of RL-based PO.
- Framework: an ADP policy iteration (PI) in a continuous PO setting.
- LEQG also interprets as a risk-attenuation algorithm.

Contributions

- A two-loop iterative alternating best-response procedure for computing the optimal mixed-design policy;
- Rigorous convergence analyses follow for the model-based loop updates;
- In the absence of exact system models, we provide an input-to-state-stable hybrid robust stabilization scheme.

Transition Slide

This page is left blank intentionally.

Problem Setup

For $\alpha > 0$, the cost $\mathcal{J}_{exp}(x_0, u) = \mathbb{E}\Big|_{x_0 \in \mathcal{P}_0} \exp\left[\frac{\alpha}{2}\int_0^\infty z^\top(t) z(t) \mathrm{d}t\right]$, becomes

$$\mathbb{E}\Big|_{x_0 \in \mathcal{P}_0} \exp\left\{\frac{\alpha}{2}\int_0^\infty \left[x^\top(t) Q x(t) + u^\top(t) R u(t)\right] \mathrm{d}t\right\}, \tag{5}$$

with the associated closed loop transfer function,

$$T_{zw}(K) = (C - EK)(sI - A + BK)^{-1} D. \tag{6}$$

Nonconvexity and Coercivity in PG

- Coercivity: iterates remain feasible and strictly separated from the infeasible set as the cost decreases.

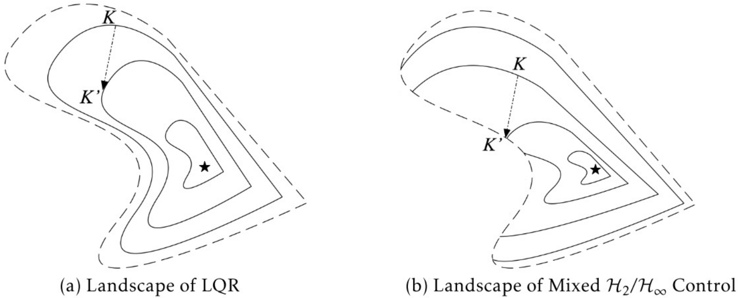


Figure: Coercivity property of PG on LQR and in mixed-design settings. Credit: (Zhang et al., 2019).

Assumptions

- $C^\top C = Q \succ 0,\ E^T(C,\ E) = (0,\ R)$ for some $R \succ 0$.
- Coercivity satisfaction: (A, B) is stabilizable;
- Optimization satisfaction: $(\sqrt{Q}, A)$ is detectable.

Transition Slide

This page is left blank intentionally.

PO and Dynamic Games: Finite-horizon Gain

- Coercivity: feasibility set of optimization iterates

$$\mathcal{K} = \{\, K : \; \lambda_i(A - B_1 K) < 0, \; \|T_{zw}(K)\|_\infty < \gamma\}. \quad (7)$$

- Finite-horizon optimization $u^\star(t) = -K^\star_{leqg}\hat{x}(t)$.
- $K^\star_{leqg} = R^{-1}B^\top P_\tau$, and P_τ is the unique, symmetric, positive definite solution to the algebraic Riccati equation (ARE)

$$A^\top P_\tau + P_\tau A - P_\tau(BR^{-1}B^\top - \alpha^{-2}DD^\top)P_\tau = -Q. \quad (8)$$

(Cui and Molu, 2023, Proposition I), (Duncan, 2013) .

- ∞-horizon case: $P^\star \triangleq P_\infty = \lim_{\tau\to\infty} P_\tau$, and $K^\star_{leqg} \triangleq K_\infty = \lim_{\tau\to\infty} K_\tau$[Theorem on limit of monotonic operators (Kan, 1964)].

Solving the LEQG Problem

- Directly solving the LEQG problem (3) in policy-gradient frameworks incurs biased gradient estimates during iterations;
- Affects risk-sensitivity preservation in infinite-horizon LTI settings (see (Zhang et al., 2021; Zhang et al., 2019));
- Workaround: an equivalent dynamic game formulation to the stochastic LQ PO problem.

Two-Player Zero-Sum Game and LEQG

- An equivalent closed-loop two-player game connection (Cui and Molu, 2023, Lemma 1):

$$\min_{u \in \mathcal{U}} \max_{\xi \in W} \bar{\mathcal{J}}_\gamma(x_0, u, \xi)$$
$$\text{subject to } dx(t) = Ax(t)dt + Bu(t)dt + Ddw(t),$$
$$z(t) = Cx(t) + Eu(t) \tag{9}$$

$$\bar{\mathcal{J}}_\gamma(x_0, u, \xi) = \mathbb{E}_{x_0 \sim \mathcal{P}_0,\, \xi(t)} \int_0^\infty \left[x^\top(t)Qx(t) + u^\top(t)Ru(t)\right] dt$$
$$-\mathbb{E}_{x_0 \sim \mathcal{P}_0,\, \xi(t)} \int_0^\infty \left[\gamma^2 \xi^\top(t)\xi(t)\right] dt$$

, $\xi(\equiv dw) \sim \mathcal{N}(0, \Sigma)$, and $\gamma \equiv \alpha$.

Proof Sketch (Cui and Molu, 2023, Lemma 1)

- If a non-negative definite (n.n.d) GARE (8)'s solution exists, then a minimal realization $P^\star$ must exist.
  - Existence: the bounded real Lemma (Zhou et al., 1996).
- If $(A, Q^{\frac{1}{2}})$ is observable, then every n.n.d solution of (8), $i.e.$ $P^\star$, is positive definite.
- For a n.n.d $P^\star$, we essentially have a Nash (equivalently a Saddle) equilibrium with $\bar{\mathcal{J}}_\gamma = \underline{\mathcal{J}}_\gamma$.

Proof Sketch (Cui and Molu, 2023, Lemma 1)

- If $\bar{\mathcal{J}}_\gamma$ is finite for some $\gamma = \hat{\gamma} > 0$, then $\bar{\mathcal{J}}_\gamma$ is bounded (if and only if the pair (A, B) is stabilizable).
- For a bounded $\bar{\mathcal{J}}_\gamma$ for some $\gamma = \hat{\gamma}$ and for optimal $K^\star = R^{-1}B^\top P_{K,L}$, $L^\star = \gamma^{-2}D^\top P_{K,L}$ and all $\gamma > \hat{\gamma}$, $\bar{\mathcal{J}}_\gamma$ admits the closed-loop matrices

$$A_K^\star = A - BK^\star,\ A_{K,L}^\star = A_K^\star + DL^\star. \tag{10}$$

- Whence, the saddle-point optimal controllers are

$$u^\star(x(t)) = -K^\star x(t),\ \xi^\star(x(t)) = L^\star x(t). \tag{11}$$

Transition Slide

This page is left blank intentionally.

Model-based PO

- Define $\{p, q\}_{p=1,q=1}^{\bar{p},\bar{q}}$ where $(\bar{p}, \bar{q}) \in \mathbb{N}_+$ as nested iteration indices for a gain K_p (in an outer loop) and an alternating gain $L_q(K_p)$ (in an inner-loop).

Problem 1 (Model-Based Policy Iteration)

Given system matrices A, B, C, D, E, find the optimal controller gains K_p, $L_q(K_p)$ that robustly stabilizes (3) such that the controller gains do not leave the set of all suboptimal controllers denoted by

$$\check{\mathcal{K}} = \{(K_p, L_q(K_p)) : \lambda_i(A_K^p) < 0, \lambda_i(A_{K,L}^{p,q}) < 0, \\ \|T_{zw}(K_p, L_q(K_p))\|_\infty < \gamma \text{ for all } (p, q) \in \mathbb{N}\}. \tag{12}$$

Model-based Policy Optimization

- Further, define the following closed-loop matrix identities

$$A_K^p = A - BK_p, \quad A_{K,L}^{p,q} = A_K^p + DL_q(K_p),$$
$$Q_K^p = Q + K_p^\top RK_p, \ A_K^\gamma = A_K^p + \gamma^{-2}DD^\top P_K^p. \tag{13}$$

- Equation (13) informs the value iterations of the Riccati equations for the outer and inner loops.

$$A_K^{p\top} P_K^p + P_K^p A_K^p + Q_K^p + \gamma^{-2} P_K^p DD^\top P_K^p = 0, \tag{14a}$$
$$K_{p+1} = R^{-1}B^\top P_K^p. \tag{14b}$$

$$A_{K,L}^{(p,q)^\top} P_{K,L}^{p,q} + P_{K,L}^{p,q} A_{K,L}^{p,q} + Q_K^p - \gamma^2 L_q^\top(K_p) L_q(K_p) = 0 \tag{15a}$$
$$K_{p+1} = R^{-1}B^\top P_K^{p,q}, \ L_{q+1}(K_p) = \gamma^{-2} D^\top P_{K,L}^{p,q}. \tag{15b}$$

Kleinman's Algorithm

- An iterative algorithm for solving infinite-time Riccati equations (Kleinman, 1968).
- Based on a successive substitution method.
- For a *deterministic LTI system's* cost matrix P_d, the value iterations of P_d^k are monotonically convergent to $P_d^\star$.
- Kleinman's algorithm as policy iteration
  - Choose a stabilizing control gain K_0, and let $p = 0$.
  - (Policy evaluation) Evaluate the performance of K_p from the GARE's solution.
  - (Policy improvement) Improve the policy: $K_p = -R^{-1}B^\top P_d^p$.
  - Advance iteration $p \leftarrow p + 1$.

Model-based Policy Iteration

Algorithm 1: (Model-Based) PO via Policy Iteration

Input: Max. outer iteration $\bar{p}$, $q = 0$, and an $\epsilon > 0$;
Input: Desired risk attenuation level $\gamma > 0$;
Input: Minimizing player's control matrix $R \succ 0$.

Compute $(K_0, L_0) \in \mathcal{K}$; ▷ From [24, Alg. 1];
Set $P_{K,L}^{0,0} = Q_K^0$; ▷ See equation (9);
for $p = 0, \ldots, \bar{p}$ **do**
  Compute Q_K^p and A_K^p ▷ See equation (9);
  Obtain P_K^p by evaluating K_p on (10);
  while $\|P_K^p - P_{K,L}^{p,q}\|_F \leq \epsilon$ **do**
    Compute $L_{q+1}(K_p) := \gamma^{-2} D^\top P_{K,L}^{p,q}$;
    Solve (11) until $\|P_K^p - P_{K,L}^{p,q}\|_F \leq \epsilon$;
    $\bar{q} \leftarrow q + 1$
  end
  Compute $K_{p+1} = R^{-1} B^\top P_{K,L}^{p,\bar{q}}$ ▷ See (11b) ;
end

Transition Slide

This page is left blank intentionally.

Convergence Analyses: Outer Loops

Lemma 1

Under our assumptions and for the ARE (14), if $K_0 \in \mathcal{K}$, then for any $p \in \mathbb{N}_+$, we must have the following conditions for the optimal $K^\star$ and $P^\star$,

(1) $K_p \in \mathcal{K}$;

(2) $P_K^0 \succeq P_K^1 \succeq \cdots P_K^p \succeq \cdots \succeq P^\star$;

(3) $\lim_{p\to\infty}\|K_p - K^*\|_F = 0$, $\lim_{p\to\infty}\|P_K^p - P^*\|_F = 0$.

Proof Sketch: The Bounded Real Lemma

Under our standard stabilizability and observability assumptions, for a stabilizing gain K, the following conditions are equivalent

- $$\|\mathcal{T}(K)\|_\infty < \gamma;$$
- The Riccati equation
$$A_K^\top P_K + P_K A_K + C^\top C + K^\top R K + \gamma^{-2} P_K D D^\top P_K = 0, \tag{16}$$
admits a unique positive definite solution $P_K \succeq 0$ for a Hurwitz matrix $(A_K + \gamma^{-2} D D^\top P_K)$;
- There exists $P_K \succ 0$ such that
$$A_K^\top P_K + P_K A_K + Q + K^\top R K + \gamma^{-2} P_K D D^\top P_K \prec 0. \tag{17}$$

Stabilizing Proof Sketch

- At an iteration 0, find a K_0 that is stabilizing (Molu, 2023, Alg. 1), so that $K_0 \in \mathcal{K}$ by the bounded real Lemma.
- For $p > 0$, set $Q_K^{p+1} = C^\top C + K_{p+1}^\top R K_{p+1}$, the outer loop GARE is

$$A_K^{(p+1)^\top} P_K^p + P_K^p A_K^{(p+1)} + \gamma^{-2} P_K^p D D^\top P_K^p + C^\top C \tag{A.2}$$
$$+ K_{p+1}^\top R K_{p+1} + (K_{p+1} - K_p)^\top R (K_{p+1} - K_p) = 0.$$

Thus, for a stabilizing $K_{p+1} (\neq K_p)$ we must have $(K_{p+1} - K_p)^\top R (K_{p+1} - K_p) \succ 0$ so that

$$A_K^{(p+1)^\top} P_K^p + P_K^p A_K^{(p+1)} + \gamma^{-2} P_K^p D D^\top P_K^p + Q_K^{p+1} \prec 0. \tag{A.3}$$

- For $p > 1$, $K_p \in \mathcal{K}$. Rest: completion of squares, the bounded real Lemma, and the theorem on the "limit of monotonic operators." (Kan, 1964).

Convergence Analysis

- In (Zhang et al., 2019, Theorem A.7 and A.8), the authors showed that this controller update in the outer-loop has a global sub-linear and local quadratic convergence rates.
- We now show that the outer-loop iteration has a global linear convergence rate.

Transition Slide

This page is left blank intentionally.

Convergence Analysis: Outer Loop

Lemma 2

Let $\Psi = (K_{p+1} - K_p)^\top R(K_{p+1} - K_p)$; *and* $\Psi = \Psi^\top \succeq 0$. *Furthermore, let* $\Phi \in \mathbb{R}^{n\times n}$ *be Hurwitz so that* $\Theta = \int_0^\infty e^{(\Phi^\top t)}\Psi e^{(\Phi t)}dt$ *and define* $c(\Phi) = \log(5/4)\|\Phi\|^{-1}$. *Then,* $\|\Theta\| \geq \frac{1}{2}c(\Phi)\|\Psi\|$.

Convergence Analysis: Outer Loop

Remark 1

For $A_K = A - BK$, we know from the bounded real Lemma (Zhang et al., 2019, Lemma A.1) that the Riccati equation

$$A_K^\top P_K + P_K A_K + Q_K + \gamma^{-2} P_K D D^\top P_K = 0 \tag{18}$$

admits a unique positive definite solution $P_K \succ 0$ with a Hurwitz $(A_K + \gamma^{-2} D D^\top P_K)$.

Transition Slide

This page is left blank intentionally.

Optimality of the Iteration

Lemma 3 (Optimality of the iteration)

Consider any $K \in \mathcal{K}$, let $K' = R^{-1}B^\top P_K$ (where P_K is the solution to (18)*, and $\Psi_K = (K - K')^\top R(K - K')$. If $\Psi_K = 0$, then $K = K^\star$.*

Proof.

Since $R \succ 0$, $\Psi_K = 0$ implies $K = K'$. Therefore at $\Psi_K = 0$, we must have $K = K'$ which implies that $P_K = P'_K$. If $K = K'$ and $P_K = P'_K$, it suffices to conclude that $K' = K \triangleq K^\star$ where $K^\star = R^{-1}B^\top P^\star$. Hence, $\Psi_K = 0$ is tantamount to $P_K = P^\star$ and $K = K^\star$. □

Bound on Cost Difference Matrix

Lemma 4 (Bound on Cost Difference Matrix)

For any $h > 0$, define $\mathcal{K}_h := \{K \in \mathcal{K} | Tr(P_K^p - P^\star) \le h\}$. For any $K \in \mathcal{K}_h$, let $K' := R^{-1}B^\top P_K^p$, where P_K^p is the p'th iterate's solution to (18), and $\Psi_{K_p} = (K_p - K_p')^\top R(K_p - K_p')$. Then, there exists $b(h) > 0$, such that $\|P_K^p - P^\star\|_F \le b(h)\|\Psi_{K_p}\|_F$.

Bound on Cost Difference Matrix

- For $A^\star = A - BR^{-1}B^\top P^\star + \gamma^{-2}DD^\top P^\star$, rewrite the closed-loop Riccati equation as

$$
\begin{aligned}
&A^{\star\top} P_K^p + P_K^p A^\star + Q_{K_p} + (K^\star - K_p)^\top R K_p' \\
&+ K_p'^\top R(K^\star - K_p) - \gamma^{-2} P^\star DD^\top P_K^p - \gamma^{-2} P_K^p DD^\top P^\star \\
&+ \gamma^{-2} P_K^p DD^\top P_K^p = 0.
\end{aligned} \tag{19}
$$

- Then do completion of squares so that

$$
\begin{aligned}
&A^{\star\top}(P_K^p - P^\star) + (P_K^p - P^\star)A^\star + \Psi_{K_p} \\
&+ \gamma^{-2}(P_K^p - P^\star)DD^\top(P_K^p - P^\star) \\
&- (K_p' - K^\star)^\top R(K_p' - K^\star) = 0.
\end{aligned} \tag{20}
$$

Proof

- Implicit function theorem: $P_K^p = f(K_p \in \mathcal{K})$, $f(\cdot) \in \mathcal{C}^n$.
- There exists a ball $\mathcal{B}_\delta(K^\star) := \{K \in \mathcal{K} | \|K - K^\star\|_F \leq \delta\}$, such that $\mathcal{A}(K)$ is invertible for any $K \in \mathcal{K}_h \cap \mathcal{B}_\delta(K^\star)$.
  - $\mathcal{A}(K_p) = I_n \otimes A^{\star\top} + (A - BR^{-1}B^\top P_K^p + \gamma^{-2}DD^\top P_K^p)^\top \otimes I_n$.
- Therefore, for any $K \in \mathcal{K}_h \cap \mathcal{B}_\delta(K^\star)$,
  - $\|\tilde{P}_K^p\|_F \leq \underline{\sigma}^{-1}(\mathcal{A}(K_p))\|\Psi_{K_p}\|_F$.
- Similarly, for any $K \in \mathcal{K}_h \cap \mathcal{B}_\delta^c(K^\star)$, where $\mathcal{B}^c$ is a complement of $\mathcal{B}$, $\Psi_{K_p} \neq 0$ and there exists a constant $b_1 > 0$ such that $\|\Psi_{K_p}\| \geq b_1$.
- Set $b_2 = \max_{K \in \mathcal{K}_h \cap \mathcal{B}_\delta(K^\star)} \underline{\sigma}^{-1}(\mathcal{A}(K))$ and $b(h) = \max\{b_2, \frac{h + Tr(P^\star)}{b_1}\}$, then the proof follows immediately.

Transition Slide

This page is left blank intentionally.

Outer Loop Convergence: Exponential Stability of P_K^p

Theorem 2

For any $h > 0$ and $K_0 \in \mathcal{K}_h$, there exists $\alpha(h) \in \mathbb{R}$ such that $Tr(P_K^{p+1} - P^\star) \le \alpha(h) Tr(P_K^p - P^\star)$. That is, $P^\star$ is an exponentially stable equilibrium.

Transition Slide

This page is left blank intentionally.

Convergence Analysis: Inner Loop

- Now, we analyze the monotonic convergence rate of the inner loop.
- Given arbitrary gains $K_p \in \mathcal{K}$ and $L_q(K_p) \in \mathcal{L}$, and $P_{K,L}^{p,q} \succ 0$ solution of the inner-loop Lyapunov equation, the cost matrix $P_{K,L}^{p,q}$ monotonically converges to the solution of (15).

$$A_{K,L}^{(p,q)^\top} P_{K,L}^{p,q} + P_{K,L}^{p,q} A_{K,L}^{p,q} + Q_K^p - \gamma^2 L_q^\top(K_p) L_q(K_p) = 0 \tag{21a}$$

$$K_{p+1} = R^{-1} B^\top P_K^{p,q}, \;\; L_{q+1}(K_p) = \gamma^{-2} D^\top P_{K,L}^{p,q}. \tag{21b}$$

Convergence Analysis: Inner Loop I

Lemma 5

Suppose that $L_0(K_0)$ is stabilizing, then for any $q \in \mathbb{N}_+$ (with $P_{K,L}^{p,\bar{q}}$ as the solution to (15)), i.e.

$$A_{K,L}^{(p,q)^\top} P_{K,L}^{p,q} + P_{K,L}^{p,q} A_{K,L}^{p,q} + Q_K^p - \gamma^2 L_q^\top(K_p) L_q(K_p) = 0 \quad \text{(22a)}$$

$$K_{p+1} = R^{-1} B^\top P_K^{p,q}, \; L_{q+1}(K_p) = \gamma^{-2} D^\top P_{K,L}^{p,q}. \quad \text{(22b)}$$

Then, the following statements hold

1. $A_{K,L}^{p,q}$ *is Hurwitz;*
2. $P_{K,L}^{p,\bar{q}} \succeq \cdots \succeq P_K^{(p,q+1)} \succeq P_K^{p,q} \succeq \cdots \succeq P_{K,L}^{p,0}$*; and*
3. $\lim_{q\to\infty} \| P_{K,L}^{p,q} - P_{K,L}^{p,\bar{q}} \|_F = 0.$

Convergence Rate – Inner Loop

Lemma 6 (Monotonic Convergence of the Inner-Loop)

For any $K \in \mathcal{K}$, let $L(K)$ be the control gain for the player w such that $A_K + DL(K)$ is Hurwitz. Let P_K^L be the solution of

$$(A_K + DL(K))^\top P_K^L + P_K^L (A_K + DL(K)) + Q_K - \gamma^2 L(K)^\top L(K) = 0. \tag{23}$$

Let $L'(K) = \gamma^{-2} D^\top P_K^L$ and $\Psi_K^L = \gamma^{-2}(L'(K) - L(K))^\top (L'(K) - L(K))$. Then, for a $c(K) = Tr\left(\int_0^\infty e^{(A_K + DL(K^\star))t} e^{(A_K+DL(K^\star))^\top t} \mathrm{d}t\right)$, the following inequality holds $Tr(P_K - P_K^L) \leq \|\Psi_K^L\| c(K)$.

Convergence of the Inner Loop Iteration

Theorem 3

For a $K \in \check{\mathcal{K}}$, and for any $(p, q) \in \mathbb{N}_+$, there exists $\beta(K) \in \mathbb{R}$ such that

$$Tr(P_K^p - P_{K,L}^{p,q+1}) \le \beta(K) Tr(P_K^p - P_{K,L}^{p,q}). \tag{24}$$

Remark 2

As seen from Lemma 5, $P_K^p - P_{K,L}^{p,q} \succeq 0$. By the norm on a matrix trace (Cui and Molu, 2023, Lemma 13) and the result of Theorem 3, we have $\|P_K - P_{K,L}^{p,q}\|_F \le Tr(P_K - P_{K,L}^{p,q}) \le \beta(K) Tr(P_K)$, i.e. $P_{K,L}^{p,q}$ exponentially converges to P_K in the Frobenius norm.

Algorithm as a Policy Iteration Scheme

- Choosing a stabilizing K_p we first evaluate u's performance by solving (14).
  - This is the policy evaluation step in PI.
- The policy is then improved in a following iteration by solving for the cost matrix in (15b);
  - This is the policy improvement step.
- Essentially, a policy iteration algorithm whereupon
  - Performance of an initial control gain K_p is first evaluated against a cost function.
  - A newer evaluation of the cost matrix $P_{K,L}^{p,q}$ is then used to improve the controller gain K_{p+1} in the outer loop.

Transition Slide

This page is left blank intentionally.

Sampling-based PO Scheme

- A, B, C, D, E are often unavailable so that the policy evaluation step will result in biased estimates.
- There is the possibility for a divergence from the stability-robustness feasibility set $\breve{\mathcal{K}}$:
  - When errors are present from I/O or state data;
  - Residuals from early termination of numerically solving Riccati equations;
  - Using an approximate cost function owing to inexact values of Q and R;
  - Since the inner loop is computed in a finite number of steps;
  - In a data sampling scheme, we must guarantee the stability and robustness of the closed-loop system.

Sampling-based PO: Statement of the Problem

Problem 4 (Sampling-based Policy Optimization)

If A, B, C, D, E, P are all replaced by approximate matrices $\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{E}, \hat{P}$, under what conditions will the sequences $\{\hat{P}_{K,L}^{p,q}\}_{(p,q)=1}^{(p,q)=\infty}$, $\{\hat{K}_p\}_{p=0}^{\infty}$, $\{\hat{L}_q\}_{q=0}^{\infty}$ converge to a small neighborhood of the optimal values $\{P_{K,L}^\star\}_{(p,q)=0}^{(p,q)=\infty}$, $\{K_p^\star\}_{p=0}^{\infty}$, and $\{L_q^\star\}_{q=0}^{\infty}$?

Discrete-Time Nonlinear System Interpretation

- From assumptions, a $P_K^0 \in \mathbb{S}^n$ exists such that when applied to find a K_0 such a K_0 will be stabilizing.
- Approximation errors between the nested iteration steps yield a hybrid of a continuous-time policy gain pair $(\hat{K}_p, \hat{L}_q(\hat{K}_p))$ and a learning scheme.
  - This learning scheme is essentially a discrete sampled data from a nonlinear system (owing to errors from various sources).
- Task: under inexact loop updates, lump iterates of gain errors into system inputs to the online PO scheme;

Transition Slide

This page is left blank intentionally.

Discrete-Time Nonlinear System Interpretation

- How do we converge to the optimal solution and preserve closed-loop dynamic stability?
- What does input-to-state stability (ISS) Sontag (2008) have to do with it?

Online Model-free Reparameterization

- Suppose that $\hat{P}_K^0 \in \mathbb{S}^n$ is chosen following the controllability and stabilizability assumptions.
  - Then $\hat{K}_k^1 = R^{-1} B^\top \hat{P}_K^0$ will be stabilizing since $\tilde{K}_k^1 = \hat{K}_k^1 - K_k^1 \triangleq 0$.
- Ditto argument for L_1.

Problem 5

For $(p, q) > 0$, show that for $\tilde{K}_k^p = \hat{K}_k^p - K_k^p \triangleq 0$ so that the sequence $\{P_{K,L}^{p,q}\}_{(p,q)=0}^{\infty}$ converges to the locally exponentially stable $\hat{P}_{K,L}^\star$.

Transition Slide

This page is left blank intentionally.

Hybrid System Reparameterization

- Lump estimate errors as an input into the gain terms to be computed in the PO algorithm.
- With inexact outer loop update, K_{p+1} becomes biased so that the inexact outer-loop GARE value iteration involves the recursions

$$\hat{A}_K^{p\top} \hat{P}_K^p + \hat{P}_K^p \hat{A}_K^p + \hat{Q}_K^p + \gamma^{-2} \hat{P}_K^p D D^\top \hat{P}_K^p = 0, \tag{25a}$$

$$\hat{K}_{p+1} = R^{-1} B^\top \hat{P}_K^p + \tilde{K}_{p+1} \triangleq \bar{K}_{p+1} + \tilde{K}_{p+1}, \tag{25b}$$

  - NB: $\hat{A}_K^p = A - B\hat{K}_p$ and $\hat{Q}_K^p = Q + \hat{K}_p^\top R \hat{K}_p$.

Discrete-Time System Closed-loop System

- Same argument for the inner-loop inexact GARE value iteration updates:

$$\hat{A}_{K,L}^{p,q\top} \hat{P}_{K,L}^{p,q} + \hat{P}_{K,L}^{p,q} \hat{A}_{K,L}^{p,q} + \hat{Q}_K^p - \gamma^2 \hat{L}_q^\top \hat{L}_q(\hat{K}_p) = 0 \tag{26a}$$

$$\hat{K}_{p+1} = R^{-1} B^\top \hat{P}_K^{p,q} + \tilde{K}_p, \tag{26b}$$

$$\hat{L}_{q+1}(\hat{K}_p) = \gamma^{-2} D^\top \hat{P}_{K,L}^{p,q} + \tilde{L}_{q+1}(\tilde{K}_p) \tag{26c}$$

$$\triangleq \bar{L}_{q+1}(\bar{K}_p) + \tilde{L}_{q+1}(\tilde{K}_p). \tag{26d}$$

- Rewrite the infinite-dimensional stochastic differential equation as the discrete-time system (for iterates $(p,q) > 0$):

$$\mathrm{d}x = [\hat{A}_{K,L}^{p,q} x + B(\hat{K}_p x - D\hat{L}_q(K_p) + u)]\mathrm{d}t + D\mathrm{d}w. \tag{27}$$

Transition Slide

This page is left blank intentionally.

System Trajectories from HJB Interpretation

- On a time interval $[s, s+\delta s]$, it follows from Itô's stochastic calculus and the Hamilton-Jacobi-Bellman equation that

$$d\left[x^\top(s+\delta s)\hat{P}_{K,L}^{p,q}x(s+\delta s) - x^\top(s)\hat{P}_{K,L}^{p,q}x(s)\right] = (\mathrm{d}x)^\top\hat{P}_{K,L}^{p,q}x + x^\top\hat{P}_{K,L}^{p,q}\mathrm{d}x + (\mathrm{d}x)^\top\hat{P}_{K,L}^{p,q}(\mathrm{d}x). \tag{28}$$

- Along the trajectories of equation (27) and using the gains in (15), *i.e.*

$$K_{p+1} = R^{-1}B^\top P_K^{p,q}, \;\; L_{q+1}(K_p) = \gamma^{-2}D^\top P_{K,L}^{p,q}.$$

System Trajectories

- The r.h.s. in (28) becomes

$$x^\top \left[\hat{A}_{K,L}^{p,q\top} \hat{P}_{K,L}^{p,q} + \hat{P}_{K,L}^{p,q} \hat{A}_{K,L}^{p,q}\right] x\mathrm{d}t + 2x^\top \hat{P}_{K,L}^{p,q} D\mathrm{d}w \tag{29}$$

$$+ 2x^\top \hat{P}_{K,L}^{p,q} B(K_p x - D\hat{L}_q(K_p) + u)dt + Tr(D^\top PD),$$

$$= -x^\top \hat{Q}_K^p x\mathrm{d}t - \gamma^{-2} x^\top \hat{P}_{K,L}^{p,q} DD^\top \hat{P}_{K,L}^{p,q} xdt + Tr(D^\top \hat{P}_{K,L}^{p,q} D) + 2x^\top \hat{P}_{K,L}^{p,q} B\left[\hat{K}_p x - D\hat{L}_q(K_p) + u\right]\mathrm{d}t + 2x^\top \hat{P}_{K,L}^{p,q} D\mathrm{d}w \tag{30}$$

- So that

$$
\begin{aligned}
&x^\top(s+\delta s)\hat{P}_{K,L}^{p,q}(s+\delta s) - x^\top(s)\hat{P}_{K,L}^{p,q}x(s) \\
&= \int_s^{s+\delta s} \left[(-x^\top \hat{Q}_K^p x - \gamma^2 w^\top w)\mathrm{d}t + 2\gamma^2 x^\top \hat{L}_{q+1}^\top(K_p)\mathrm{d}w\right] \\
&+ \int_s^{s+\delta s} 2x^\top \hat{K}_{p+1}^\top R\left[\hat{K}_p x - D\hat{L}_q(\hat{K}_p) + u\right]\mathrm{d}t \\
&\qquad + \int_s^{s+\delta s} Tr(D^\top \hat{P}_{K,L}^{p,q} D)\mathrm{d}t. \qquad (31)
\end{aligned}
$$

Transition Slide

This page is left blank intentionally.

Input To State System Interpretation

- System matrices $\hat{A}_{K,L}^{p,q}, B, C, D$ now embedded within input and state terms: $\hat{Q}_K^p$, $\hat{K}_{p+1}$, and $\hat{L}_{q+1}$;
- Retrievable via online measurements.
- We essentially end up with an input-to-state system!
- The price that we pay is that the noise feedthrough matrix D must be known precisely.
  - No marvel: in many linear stochastic system with Brownian motion, D is identity (Duncan et al., 2011; Duncan and Pasik-Duncan, 2010).

Sampling-based Scheme

- Explore system model until we achieve exact equality in $\hat{A}^{p,q}_{K,L} \equiv A^{p,q}_{K,L}$, $\hat{P}^{p,q}_{K,L}$, $\hat{K}_{p+1} \equiv K_{p+1}$, and $\hat{L}_{q+1}(K_p) \equiv L_{q+1}(K_p)$.
  - Choose $u = -K_0 x + \eta_p$ and $w = -L_0 x + \eta_q$ where (η_p, η_q) is drawn uniformly at random over matrices with a Frobenium norm r similar to (Gravell et al., 2021; Fazel et al., 2018).

Sampled System Parameterization

- Consider the identities

$$
\begin{aligned}
&x^\top \hat{Q}_K^p x = (x^\top \otimes x^\top)\,\mathrm{vec}(\hat{Q}_K^p),\\
&\gamma^2 w^\top w = \gamma^2 (w^\top \otimes w^\top)\ \mathrm{vec}(I_v),\\
&2\gamma^2 x^\top \hat{L}_{q+1}^\top(\hat{K}_p)\mathrm{d}w = 2\gamma^2 (I_n \otimes x^\top)\mathrm{d}w\ \mathrm{vec}(\hat{L}_{q+1}^\top(\hat{K}_p)),\\
&2x^\top \hat{K}_{p+1}^\top R \hat{K}_p x = 2(x^\top \otimes x^\top)(I_n \otimes \hat{K}_p^\top)\,\mathrm{vec}(\hat{K}_{p+1}^\top R),\\
&2x^\top \hat{K}_{p+1}^\top R D \hat{L}_q(\hat{K}_p) = 2(\hat{L}_q^\top(\hat{K}_p) D^\top \otimes x^\top)\,\mathrm{vec}(\hat{K}_{p+1}^\top R),\\
&2x^\top \hat{K}_{p+1}^\top R u = 2(u^\top \otimes x^\top)\,\mathrm{vec}(\hat{K}_{p+1}^\top R),\\
&Tr(D^\top \hat{P}_{K,L}^{p,q} D) = \mathrm{vec}^\top(D)\,\mathrm{vec}(\hat{P}_{K,L}^{p,q} D).
\end{aligned}
\tag{32}
$$

Sampled System Parameterization I

- Let $\Delta_{xx} \in \mathbb{R}^{\frac{n(n+1)}{2}l}$, $\Delta_{ww} \in \mathbb{R}^{\frac{v(v+1)}{2}l}$, $I_{xx} \in \mathbb{R}^{l \times n^2}$, and $I_{ux} \in \mathbb{R}^{l \times mn}$ for $l \in \mathbb{N}_+$
- It follows that

$$\Delta_{xx} = \left[\texttt{vecv}(x_1), \ldots, \texttt{vecv}(x_l)\right]^\top, \; x_l = x_{l+1} - x_l,$$

$$\Delta_{ww} = \left[\texttt{vecv}(w_1), \ldots, \texttt{vecv}(w_l)\right]^\top, \; w_l = w_{l+1} - w_l,$$

$$I_{xx} = \left[\int_{s_0}^{s_1} x \otimes x \, dt, \ldots, \int_{s_{l-1}}^{s_l} x \otimes x \, dt\right]^\top,$$

Transition Slide

This page is left blank intentionally.

Sampled System Parameterization

$$
\begin{aligned}
I_{xw} &= \left[\int_{s_0}^{s_1} (I_n \otimes x) \mathrm{d}w, \ldots, \int_{s_{l-1}}^{s_l} (I_n \otimes x) \mathrm{d}w\right]^\top, \\
I_{ux} &= \left[\int_{s_0}^{s_1} u \otimes x \, \mathrm{d}t, \ldots, \int_{s_{l-1}}^{s_l} u \otimes x \, \mathrm{d}t\right]^\top. \qquad (33)
\end{aligned}
$$

Next, set

$$
\begin{aligned}
\Theta_{K,L}^{p,q} &= \Big[\Delta_{xx}, -2I_{xx}(I_n \otimes \hat{K}_p^\top) + 2(\hat{L}_q^\top(\hat{K}_p) D^\top \otimes x^\top) \\
&\quad -2I_{ux}, -2\gamma^2 I_{xw}, -\mathtt{vec}^\top(D)\mathtt{vec}(\hat{P}_{K,L}^{p,q} D)\Big], \qquad (34a) \\
\Upsilon_{K,L}^{p,q} &= \Big[-I_{xx}\mathtt{vec}(\hat{Q}_K^p), \; -\gamma^2 I_{ww}\mathtt{vec}(I_v)\Big]. \qquad (34b)
\end{aligned}
$$

Sampled System Parameterization

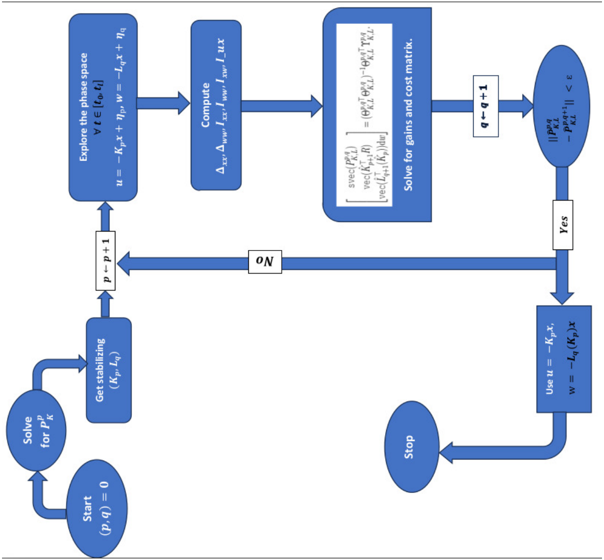
Define $\mathbf{1}_{q^2}$ as a one-vector with dimension q^2. Thus,

$$\Theta_{K,L}^{p,q}\left[\texttt{svec}(P_{K,L}^{p,q}) \quad \texttt{vec}(\hat{K}_{p+1}^{\top}R) \quad \texttt{vec}(\hat{L}_{q+1}^{\top}(\hat{K}_p)) \quad \mathbf{1}_{q^2}\right]^{\top} = \Upsilon_{K,L}^{p,q}. \tag{35}$$

Suppose that $\Theta_{K,L}^{p,q}$ is of full rank, then we can retrieve the unknown matrices via least squares estimation *i.e.*

$$\begin{bmatrix} \texttt{svec}(P_{K,L}^{p,q}) \\ \texttt{vec}(\hat{K}_{p+1}^{\top}R) \\ \texttt{vec}(\hat{L}_{q+1}^{\top}(\hat{K}_p))\mathrm{d}w \\ \mathbf{1}_{q^2} \end{bmatrix} = (\Theta_{K,L}^{p,q\top}\Theta_{K,L}^{p,q})^{-1}\Theta_{K,L}^{p,q\top}\Upsilon_{K,L}^{p,q}. \tag{36}$$

Sampling-based Algorithm



Transition Slide

This page is left blank intentionally.

Robustness Analyses

- Define $\tilde{P} = P_K - \hat{P}_K$ and $\tilde{K} = K - \hat{K}$.
- Keep $|\tilde{K}| < \epsilon$, start with a $K \in \mathcal{K}$: iterates stay in $\mathcal{K}$.

Lemma 7 (Lemma 10, C&M, '23)

For any $K \in \mathcal{K}$, there exists an $e(K) > 0$ such that for a perturbation $\tilde{K}$, $K + \tilde{K} \in \mathcal{K}$, as long as $\|\tilde{K}\| < e(K)$.

Theorem 6

The inexact outer loop is small-disturbance ISS. That is, for any $h > 0$ and $\hat{K}_0 \in \mathcal{K}_h$, if $\|\tilde{K}\| < f(h)$, there exist a $\mathcal{KL}$-function $\beta_1(\cdot,\cdot)$ and a $\mathcal{K}_\infty$-function $\gamma_1(\cdot)$ such that

$$\|P^p_{\hat{K}} - P^\star\| \le \beta_1(\|P^0_{\hat{K}} - P^*\|, p) + \gamma_1(\|\tilde{K}\|). \tag{37}$$

ISS Outer Loop Robustness Proof

- Prelim result (Lemma 12, C&M, '23): For any $h > 0$ and $K \in \mathcal{K}_h$, let $K' = R^{-1}B^\top P_K$, where P_K is the solution of (18), and $\hat{K}' = K' + \tilde{K}$. Then, there exists $f(h) > 0$, such that $\hat{K}' \in \mathcal{K}_h$ as long as $\|\tilde{K}\| < f(h)$.
- Therefore, $\hat{K}_K^p \in \mathcal{K}_h$ for any $p \in \mathbb{N}_+$.
- Let

$$f_1(\hat{K}') = \frac{\log(5/4)b(h)}{2n\|A^\star_{\hat{K}'}\|}, f_2(\hat{K}') = Tr\left(\int_0^\infty e^{A^{\star\top}_{\hat{K}'}t} e^{A^\star_{\hat{K}'}t} \mathrm{d}t\right).$$

ISS Outer Loop Robustness Proof

- $$\underline{f}_1(h) = \inf_{\hat{K}' \in \mathcal{K}_h} f_1(\hat{K}') > 0, \bar{f}_2(h) = \sup_{\hat{K}' \in \mathcal{K}_h} f_2(\hat{K}') < \infty. \tag{38}$$

- This implies
$$Tr(P_{\hat{K}}^p - P^\star) \le [1 - \underline{f}_1(h)]\, Tr(P_{\hat{K}}^{p-1} - P^\star) + \bar{f}_2(h)\|R\|\|\tilde{K}_K^p\|^2. \tag{39}$$

- Repeating (39) for $p, p-1, \cdots, 1$,
$$Tr[P_{\hat{K}}^p - P^\star] \le (1 - \underline{f}_1)^p\, Tr(P_{\hat{K}}^1 - P^\star) + \frac{\bar{f}_2\|R\|\|\tilde{K}\|_\infty^2}{\underline{f}_1(h)}. \tag{40}$$

Outer Loop Robustness

It follows from (40) and (Mori, 1988, Theorem 2) that

$$\|P_{\hat{K}}^{p} - P^{\star}\|_F \le (1 - \underline{f}_1)^p \sqrt{n} \|P_{\hat{K}}^{1} - P^{\star}\|_F + \frac{\bar{f}_2 \|R\| \|\tilde{K}\|_{\infty}^2}{\underline{f}_1}. \tag{41}$$

As $p \to \infty$, $P_{\hat{K}}^{p} \to P^{\star}$. Whence, a radius of $P^{\star}$'s neighbor is proportional to $\|\tilde{K}\|_{\infty}^2$.

Inner Loop Robustness

The perturbed inner-loop iteration (26) has inexact matrix $\hat{A}^{p,q}_{K,L}$, and sequences $\{\hat{L}_{q+1}(K_p)\}_{q=0}^{\infty}$, and $\{\hat{P}^{p,q}_{K,L}\}_{q=0}^{\infty}$.

Lemma 8 (Stability of the Inner-Loop's System Matrix)

Given $K \in \check{\mathcal{K}}$, there exists a $g \in \mathbb{R}_+$, such that if $\|\tilde{L}_{q+1}(K_p)\|_F \leq g$, $\hat{A}^{p,q}_{K,L}$ is Hurwitz for all $q \in \mathbb{N}_+$.

Inner Loop Robustness

Theorem 7

Assume $\|\tilde{L}_q(K_p)\| < e$ for all $q \in \mathbb{N}_+$. There exists $\hat{\beta}(K) \in [0, 1)$, and $\lambda(\cdot) \in \check{\mathcal{K}}_\infty$, such that

$$\|\hat{P}_{K,L}^{p,q} - P_{K,L}^{p,q}\|_F \leq \hat{\beta}^{q-1}(K)\, Tr(P_{K,L}^{p,q}) + \lambda(\|\tilde{L}\|_\infty). \tag{42}$$

- From Theorem 7, as $q \to \infty$, $\hat{P}_{K,L}^{p,q}$ approaches the solution P_K and enters the ball centered at $P_{K,L}^{p,q}$ with radius proportional to $\|\tilde{L}\|_\infty$.
- The proposed inner-loop iterative algorithm well approximates $P_{K,L}^{p,q}$.

Transition Slide

This page is left blank intentionally.

Numerical Results – Car Cruise Control System

- (Åström and Murray, 2021, §3.1):

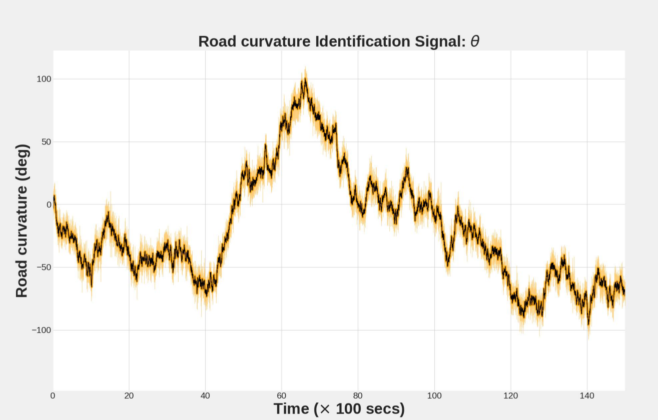
$$m\frac{dv}{dt} = \alpha_n u \tau(\alpha_n v) - mgC_r sgn(u) - \frac{1}{2}\rho C_d A|v|v - mg\sin\theta \tag{43}$$

- $u(x(t)) = [u_1(t), u_2(t)]$ must maintain a constant velocity v (the state), whilst automatically adjusting the car's throttle, $u_1(t), t \in [0, T]$
  - despite disturbances characterized by road slope changes ($u_3 = \theta$),
  - rolling friction (F_r), and
  - aerodynamic drag forces (F_d).

Numerical Results – Car Cruise Control System

- Well-suited to our robust control formulation because
  - the disturbances and state variables are separable and can be lumped into the form of the stochastic differential equations;
  - it is a multiple-input (throttle, gear, vehicle speed) single-output (vehicle acceleration) system that introduces modeling challenges;
  - the entire operating range of the system is nonlinear though there is a reasonable linear bandwidth that characterize the input/output (I/O) system as we will see shortly.

Road (Disturbance) Profile



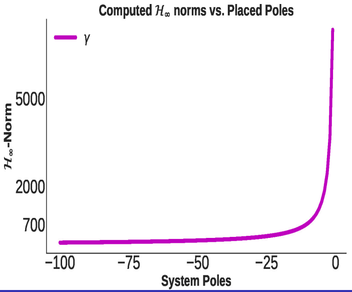
Search for initial stabilizing gain and $\mathcal{H}_\infty$-norm bound.

Proposition 1

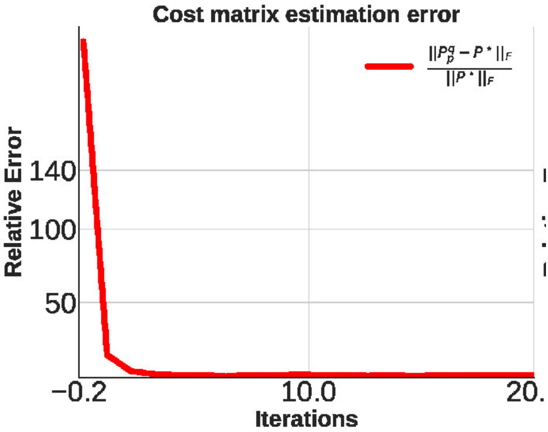
(Bruinsma and Steinbuch, 1990) For all $\omega_p \in \mathbb{R}$, we have that $j\omega_p$ is an eigenvalue of the Hamiltonian $H(\gamma_1)$ if and only if γ_1 is a singular value of $T_{zw}(j\omega_p)$.

Algorithm 1 Search for the closed-loop $\mathcal{H}_\infty$-norm

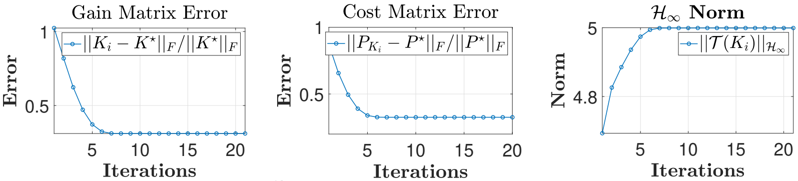
Given a user-defined step size $\eta > 0$
Set the initial upper bound on γ as $\gamma_{ub} = \infty$.
Initialize a buffer for possible $\mathcal{H}_\infty$ norms for each K_1 to be found, $\Gamma_{buf} = \{\}$.
Initialize ordered poles $\mathcal{P} = \{p_i \in Re(s) < 0 \,|\, i = 1, 2, \}$ $\quad \triangleright\ p_1 < p_2 < \cdots$
for $p_i \in \mathcal{P}$ **do**
    Place p_i on (2); $\quad \triangleright$ (Tits and Yang, 1996)
    Compute stabilizing $K_1^{p_i}$
    Find lower bound γ_{lb} for $H(\gamma, K_1^{p_i})$; $\quad \triangleright$ using (22)
    $\Gamma_{buf}(i) =$ `get_hinf_norm`$(T_{zw}, \gamma_{lb}, K_1^{p_i})$.
end for
function `get_hinf_norm`$(T_{zw}, \gamma_{lb}, K_1^{p_i})$
    while $\gamma_{ub} = \infty$ **do**
        $\gamma := (1 + 2\eta)\,\gamma_{lb}$;
        Get $\lambda_i(H(\gamma, K_1^{p_i}))$ $\quad \triangleright$ c.f. (14)
        if $\texttt{Re}(\Lambda) \neq \emptyset$ for $\Lambda = \{\lambda_1, \cdots \lambda_n\}$ **then**
            Set $\gamma_{ub} = \gamma$; exit
        else
            Set buffer $\Gamma_{lb} = \{\}$
            for $\lambda_k \in \{\texttt{Imag}(\Lambda)_{:p-1}\}$ **do** $\quad \triangleright\ k = 1$ to K
                Set $m_k = \frac{1}{2}(\omega_k + \omega_{k+1})$
                Set $\Gamma_{lb}(k) = \max\{\sigma\,[T_{zw}(jm_k)]\}$;
            end for
            $\gamma_{lb} = \max(\Gamma_{lb})$



Cost Matrix and Gains Convergence

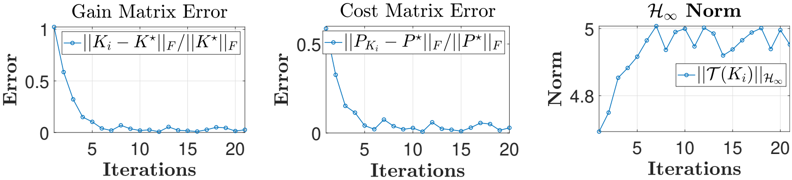


Pendulums Experiment – Comparison to NPG



Model-free design: $\|\tilde{K}\|_\infty = 0.15.$

Pendulums Experiment – Comparison to NPG



Model-based design: $\|\tilde{K}\|_\infty = 0.15$.

Double Pendulum and Acrobot Experiment – Comparison to NPG

Table: Computational Time: Model-based PO vs. Model-free PO vs. NPG.

Policy Optimization Computational time (secs)					
Double Inverted Pendulum			Triple Inverted Pendulum		
Model-based	Model-free	NPG	Model-based	Model-free	NPG
0.0901	0.3061	2.1649	0.1455	0.7829	2.3209

References I

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-End Training of Deep Visuomotor Policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Draguna Vrabie and Frank Lewis. Adaptive dynamic programming for online solution of a zero-sum differential game. *J. Contr. Theory Appl.*, 9:353–360, 08 2011. doi: 10.1007/s11768-011-0166-4.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123–158, 2023.

K. Glover. Minimum entropy and risk-sensitive control: the continuous time case. In *Proceedings of the 28th IEEE Conference on Decision and Control,*, pages 388–391 vol.1, 1989.

P.P. Khargonekar, I.R. Petersen, and M.A. Rotea. $\mathcal{H}_\infty$ optimal control with state-feedback. *IEEE Transactions on Automatic Control*, 33(8):786–788, 1988. doi: 10.1109/9.1301.

Tamer Basar. Minimax disturbance attenuation in ltv plants in discrete time. In *1990 American Control Conference*, pages 3112–3113. IEEE, 1990.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.

References II

Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11): 5283–5298, 2021.

D. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2):124–131, 1973. doi: 10.1109/TAC.1973.1100265.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy Optimization for $\mathcal{H}_2$ Linear Control with $\mathcal{H}_\infty$ Robustness Guarantee: Implicit Regularization and Global Convergence. *arXiv e-prints*, art. arXiv:1910.09496, oct 2019.

Leilei Cui and Lekan Molu. Robust Policy Optimization in Continuous-time Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Stochastic Control. 2023. URL https://scriptedonachip.com/downloads/Papers/h2hinf.pdf.

Tyrone E. Duncan. Linear-Exponential-Quadratic Gaussian control. *IEEE Transactions on Automatic Control*, 58(11):2910–2911, 2013. doi: 10.1109/TAC.2013.2257610.

Functional Analysis in Normed Spaces. New York: MacMillan, 1964.

Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Basar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34:2949–2964, 2021.

Kemin Zhou, John Comstock Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice hall Upper Saddle River, NJ, 1996.

David Z. Kleinman. On an iterative technique for riccati equation computations. *IEEE Transactions on Automatic Control*, 13:114–115, 1968.

Lekan Molu. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ policy synthesis. In *The International Federation of Automatic Control, 22nd World Congress*, page arXiv:2302.08846, July 2023.

References III

Eduardo D. Sontag. *Input to State Stability: Basic Concepts and Results*, pages 163–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Tyrone E Duncan, B Maslowski, and Bozenna Pasik-Duncan. Control of some linear stochastic systems in a hilbert space with fractional brownian motions. In *2011 16th International Conference on Methods & Models in Automation & Robotics*, pages 107–110. IEEE, 2011.

Tyrone E Duncan and Bozenna Pasik-Duncan. Stochastic linear-quadratic control for systems with a fractional brownian motion. In *49th IEEE Conference on Decision and Control (CDC)*, pages 6163–6168. IEEE, 2010.

T. Mori. Comments on "a matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equation" by J. M. Saniuk and I.B. Rhodes. *IEEE Transactions on Automatic Control*, 33(11): 1088–, 1988. doi: 10.1109/9.14428.

Karl Johan Åström and Richard M Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2021.

NA Bruinsma and M Steinbuch. A Fast Algorithm to Compute the H_∞-norm of a Transfer Function Matrix. *Systems & Control Letters*, 14:287–293, 1990.