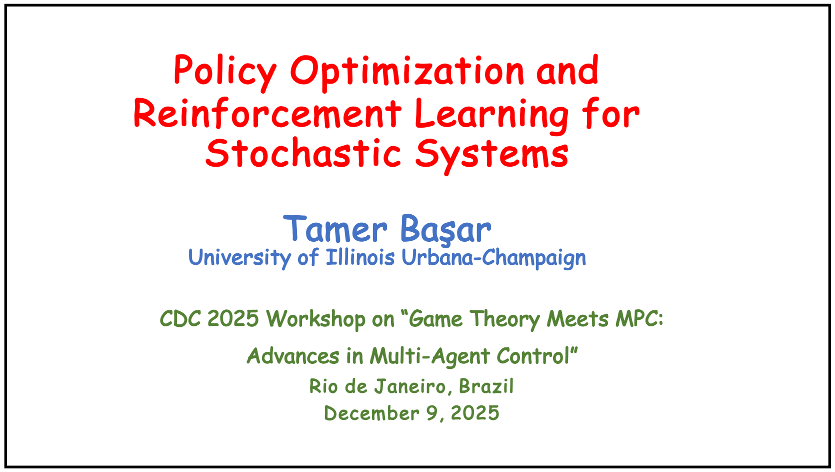
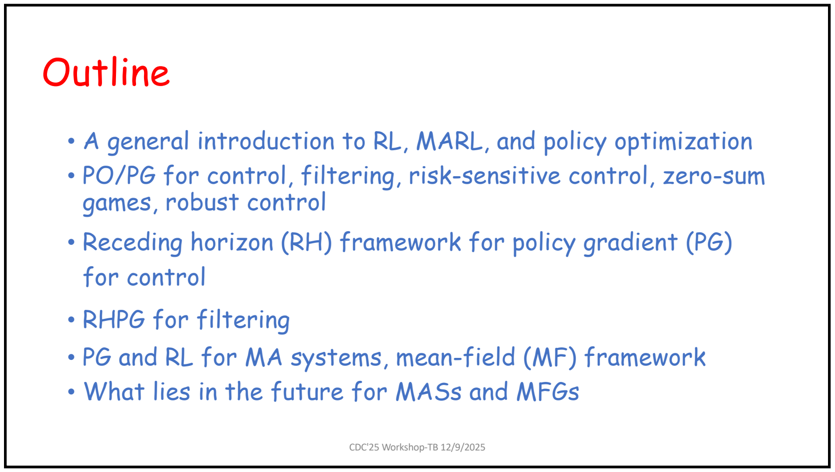


Policy Optimization and Reinforcement Learning for Stochastic Systems

Tamer Başar
University of Illinois Urbana-Champaign

CDC 2025 Workshop on "Game Theory Meets MPC:
Advances in Multi-Agent Control"
Rio de Janeiro, Brazil
December 9, 2025

2

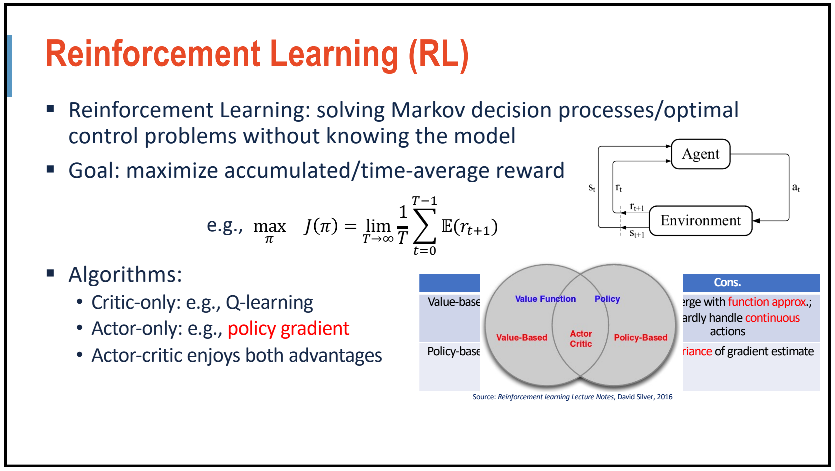


Outline

- A general introduction to RL, MARL, and policy optimization
- PO/PG for control, filtering, risk-sensitive control, zero-sum games, robust control
- Receding horizon (RH) framework for policy gradient (PG) for control
- RHPG for filtering
- PG and RL for MA systems, mean-field (MF) framework
- What lies in the future for MASs and MFGs

CDC'25 Workshop-TB 12/9/2025

3



Reinforcement Learning (RL)

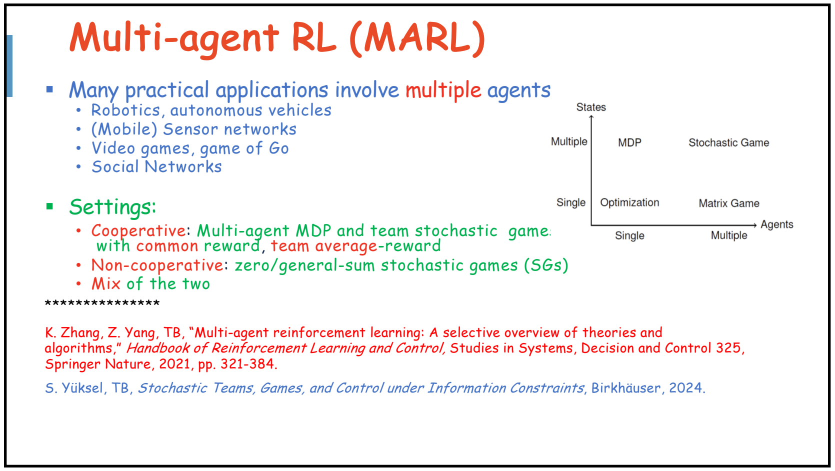
- Reinforcement Learning: solving Markov decision processes/optimal control problems without knowing the model
- Goal: maximize accumulated/time-average reward

$$\text{e.g.,}\quad \max_{\pi}\quad J(\pi) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(r_{t+1})$$

- Algorithms:
  - Critic-only: e.g., Q-learning
  - Actor-only: e.g., policy gradient
  - Actor-critic enjoys both advantages

Source: *Reinforcement learning Lecture Notes*, David Silver, 2016

4



Multi-agent RL (MARL)

- Many practical applications involve multiple agents
  - Robotics, autonomous vehicles
  - (Mobile) Sensor networks
  - Video games, game of Go
  - Social Networks
- Settings:
  - Cooperative: Multi-agent MDP and team stochastic games with common reward, team average-reward
  - Non-cooperative: zero/general-sum stochastic games (SGs)
  - Mix of the two

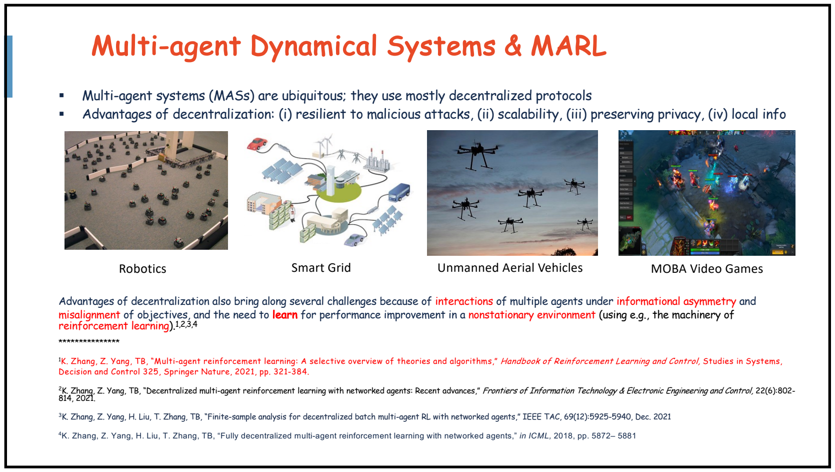
States
Multiple
Single
MDP
Stochastic Game
Optimization
Matrix Game
Agents
Single
Multiple

****

K. Zhang, Z. Yang, TB, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control,* Studies in Systems, Decision and Control 325, Springer Nature, 2021, pp. 321-384.

S. Yüksel, TB, *Stochastic Teams, Games, and Control under Information Constraints*, Birkhäuser, 2024.

5



Multi-agent Dynamical Systems & MARL

- Multi-agent systems (MASs) are ubiquitous; they use mostly decentralized protocols
- Advantages of decentralization: (i) resilient to malicious attacks, (ii) scalability, (iii) preserving privacy, (iv) local info

Robotics

Smart Grid

Unmanned Aerial Vehicles

MOBA Video Games

Advantages of decentralization also bring along several challenges because of interactions of multiple agents under informational asymmetry and misalignment of objectives, and the need to **learn** for performance improvement in a nonstationary environment (using e.g., the machinery of reinforcement learning).[1,2,3,4]

****

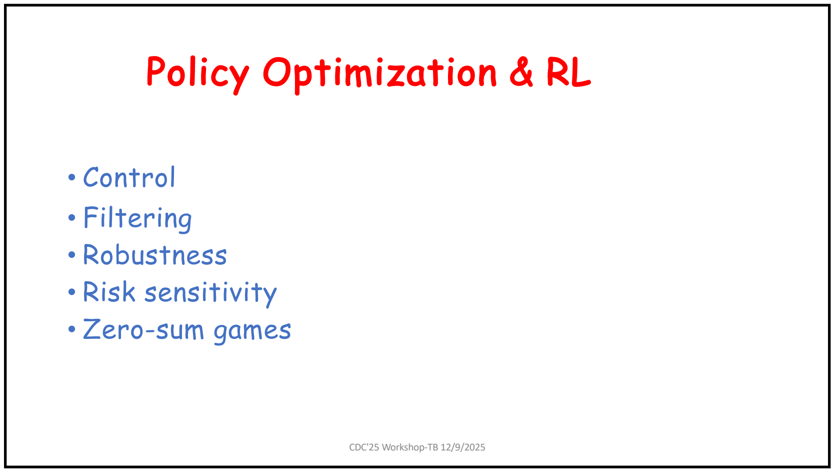
[1] K. Zhang, Z. Yang, TB, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control,* Studies in Systems, Decision and Control 325, Springer Nature, 2021, pp. 321-384.

[2] K. Zhang, Z. Yang, TB, "Decentralized multi-agent reinforcement learning with networked agents: Recent advances," *Frontiers of Information Technology & Electronic Engineering and Control,* 22(6):802-814, 2021.

[3] K. Zhang, Z. Yang, H. Liu, T. Zhang, TB, "Finite-sample analysis for decentralized batch multi-agent RL with networked agents," IEEE TAC, 69(12):5925-5940, Dec. 2021

[4] K. Zhang, Z. Yang, H. Liu, T. Zhang, TB, "Fully decentralized multi-agent reinforcement learning with networked agents," *in ICML*, 2018, pp. 5872– 5881

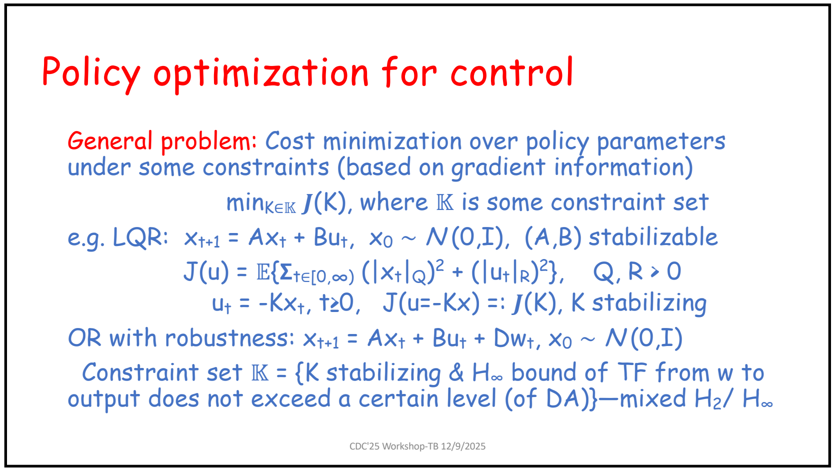
6



Policy Optimization & RL

- Control
- Filtering
- Robustness
- Risk sensitivity
- Zero-sum games

7



Policy optimization for control

General problem: Cost minimization over policy parameters under some constraints (based on gradient information)

$\min_{K\in\mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is some constraint set

e.g. LQR: $x_{t+1} = Ax_t + Bu_t$, $x_0 \sim \mathcal{N}(0,I)$, (A,B) stabilizable

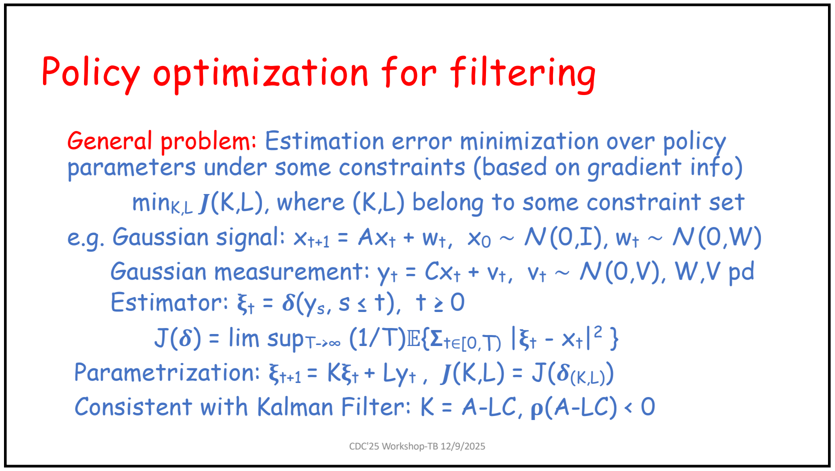
$J(u) = \mathbb{E}\{\Sigma_{t\in[0,\infty)} (|x_t|_Q)^2 + (|u_t|_R)^2\}$, $Q, R > 0$

$u_t = -Kx_t$, $t\geq 0$, $J(u=-Kx) =: \mathcal{J}(K)$, K stabilizing

OR with robustness: $x_{t+1} = Ax_t + Bu_t + Dw_t$, $x_0 \sim \mathcal{N}(0,I)$

Constraint set $\mathbb{K}$ = {K stabilizing & H_∞ bound of TF from w to output does not exceed a certain level (of DA)}—mixed H_2/ H_∞

8



Policy optimization for filtering

General problem: Estimation error minimization over policy parameters under some constraints (based on gradient info)

$\min_{K,L} \mathcal{J}(K,L)$, where (K,L) belong to some constraint set

e.g. Gaussian signal: $x_{t+1} = Ax_t + w_t$, $x_0 \sim \mathcal{N}(0,I)$, $w_t \sim \mathcal{N}(0,W)$

Gaussian measurement: $y_t = Cx_t + v_t$, $v_t \sim \mathcal{N}(0,V)$, W,V pd

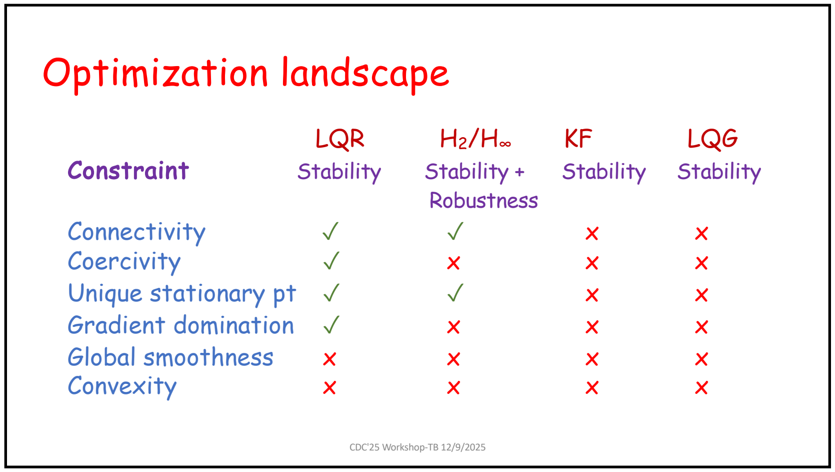
Estimator: $\xi_t = \delta(y_s, s \leq t)$, $t \geq 0$

$J(\delta) = \limsup_{T\to\infty} (1/T)\mathbb{E}\{\Sigma_{t\in[0,T)} |\xi_t - x_t|^2 \}$

Parametrization: $\xi_{t+1} = K\xi_t + Ly_t$, $\mathcal{J}(K,L) = J(\delta_{(K,L)})$

Consistent with Kalman Filter: $K = A-LC$, $\rho(A-LC) < 0$

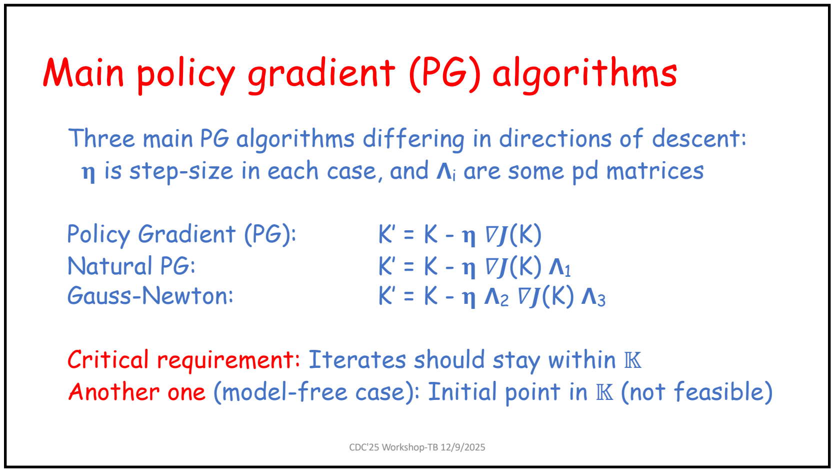
9



Optimization landscape

Constraint	LQR Stability	H_2/H_∞ Stability + Robustness	KF Stability	LQG Stability
Connectivity	✓	✓	×	×
Coercivity	✓	×	×	×
Unique stationary pt	✓	✓	×	×
Gradient domination	✓	×	×	×
Global smoothness	×	×	×	×
Convexity	×	×	×	×

10



Main policy gradient (PG) algorithms

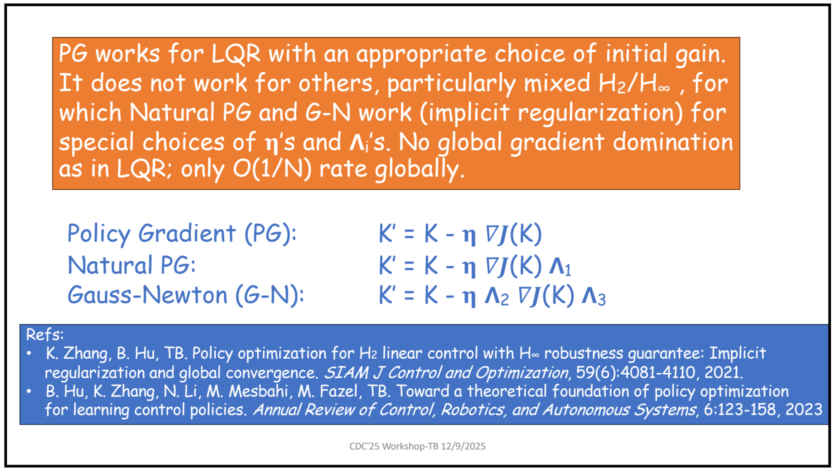
Three main PG algorithms differing in directions of descent: η is step-size in each case, and Λ_i are some pd matrices

Policy Gradient (PG):	$K' = K - \eta\, \nabla \mathcal{J}(K)$
Natural PG:	$K' = K - \eta\, \nabla \mathcal{J}(K)\, \Lambda_1$
Gauss-Newton:	$K' = K - \eta\, \Lambda_2\, \nabla \mathcal{J}(K)\, \Lambda_3$

Critical requirement: Iterates should stay within $\mathbb{K}$

Another one (model-free case): Initial point in $\mathbb{K}$ (not feasible)

11



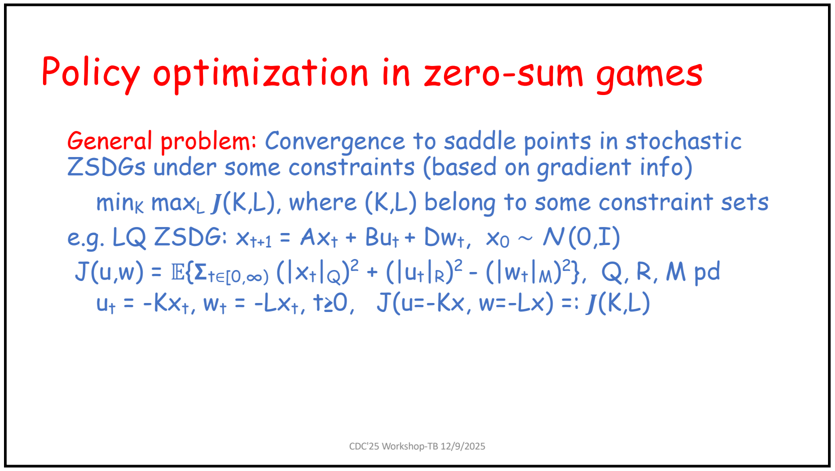
PG works for LQR with an appropriate choice of initial gain. It does not work for others, particularly mixed H_2/H_∞ , for which Natural PG and G-N work (implicit regularization) for special choices of η's and Λ_i's. No global gradient domination as in LQR; only O(1/N) rate globally.

Policy Gradient (PG):	$K' = K - \eta\, \nabla \mathcal{J}(K)$
Natural PG:	$K' = K - \eta\, \nabla \mathcal{J}(K)\, \Lambda_1$
Gauss-Newton (G-N):	$K' = K - \eta\, \Lambda_2\, \nabla \mathcal{J}(K)\, \Lambda_3$

Refs:

- K. Zhang, B. Hu, TB. Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence. *SIAM J Control and Optimization*, 59(6):4081-4110, 2021.
- B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, TB. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123-158, 2023

12



Policy optimization in zero-sum games

General problem: Convergence to saddle points in stochastic ZSDGs under some constraints (based on gradient info)

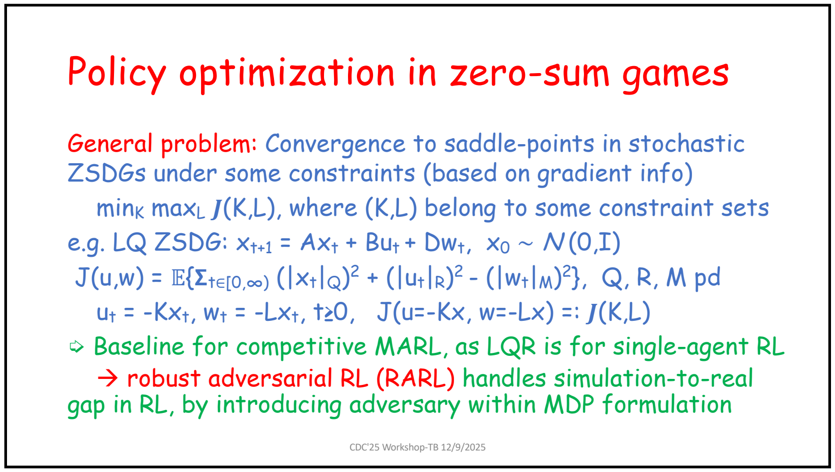
$\min_K \max_L \mathcal{J}(K,L)$, where (K,L) belong to some constraint sets

e.g. LQ ZSDG: $x_{t+1} = Ax_t + Bu_t + Dw_t$, $x_0 \sim \mathcal{N}(0,I)$

$J(u,w) = \mathbb{E}\{\Sigma_{t\in[0,\infty)} (|x_t|_Q)^2 + (|u_t|_R)^2 - (|w_t|_M)^2\}$, Q, R, M pd

$u_t = -Kx_t$, $w_t = -Lx_t$, $t\geq 0$, $J(u=-Kx, w=-Lx) =: \mathcal{J}(K,L)$

13



Policy optimization in zero-sum games

General problem: Convergence to saddle-points in stochastic ZSDGs under some constraints (based on gradient info)

$\min_K \max_L \mathcal{J}(K,L)$, where (K,L) belong to some constraint sets

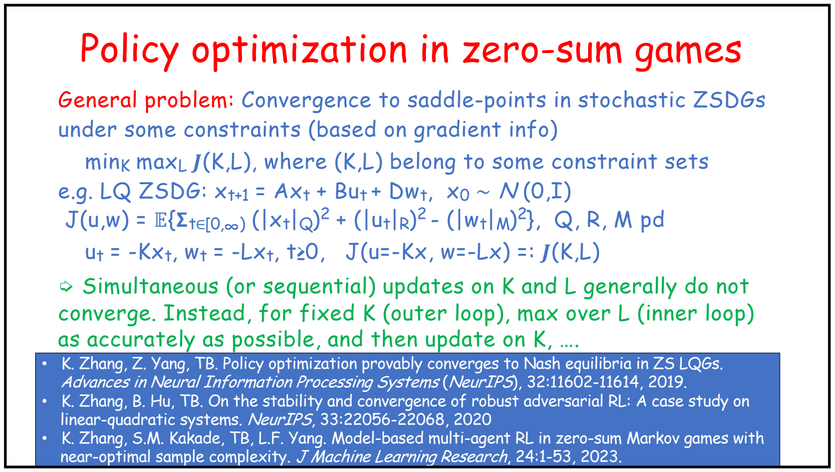
e.g. LQ ZSDG: $x_{t+1} = Ax_t + Bu_t + Dw_t$, $x_0 \sim \mathcal{N}(0,I)$

$J(u,w) = \mathbb{E}\{\Sigma_{t\in[0,\infty)} (|x_t|_Q)^2 + (|u_t|_R)^2 - (|w_t|_M)^2\}$, Q, R, M pd

$u_t = -Kx_t$, $w_t = -Lx_t$, $t\geq 0$, $J(u=-Kx, w=-Lx) =: \mathcal{J}(K,L)$

➪ Baseline for competitive MARL, as LQR is for single-agent RL → robust adversarial RL (RARL) handles simulation-to-real gap in RL, by introducing adversary within MDP formulation

14



Policy optimization in zero-sum games

General problem: Convergence to saddle-points in stochastic ZSDGs under some constraints (based on gradient info)

$\min_K \max_L \mathcal{J}(K,L)$, where (K,L) belong to some constraint sets

e.g. LQ ZSDG: $x_{t+1} = Ax_t + Bu_t + Dw_t$, $x_0 \sim \mathcal{N}(0,I)$

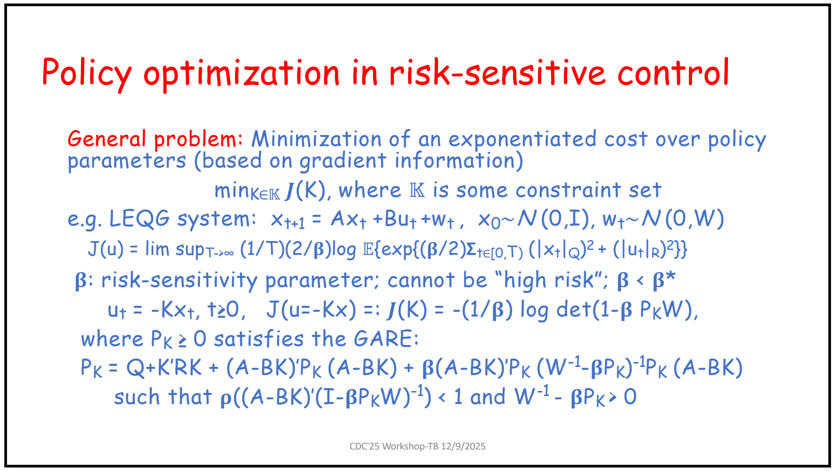
$J(u,w) = \mathbb{E}\{\Sigma_{t\in[0,\infty)} (|x_t|_Q)^2 + (|u_t|_R)^2 - (|w_t|_M)^2\}$, Q, R, M pd

$u_t = -Kx_t$, $w_t = -Lx_t$, $t\geq 0$, $J(u=-Kx, w=-Lx) =: \mathcal{J}(K,L)$

➪ Simultaneous (or sequential) updates on K and L generally do not converge. Instead, for fixed K (outer loop), max over L (inner loop) as accurately as possible, and then update on K, ….

- K. Zhang, Z. Yang, TB. Policy optimization provably converges to Nash equilibria in ZS LQGs. *Advances in Neural Information Processing Systems* (*NeurIPS*), 32:11602-11614, 2019.
- K. Zhang, B. Hu, TB. On the stability and convergence of robust adversarial RL: A case study on linear-quadratic systems. *NeurIPS*, 33:22056-22068, 2020
- K. Zhang, S.M. Kakade, TB, L.F. Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *J Machine Learning Research*, 24:1-53, 2023.

15



Policy optimization in risk-sensitive control

General problem: Minimization of an exponentiated cost over policy parameters (based on gradient information)

$\min_{K\in\mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is some constraint set

e.g. LEQG system: $x_{t+1} = Ax_t + Bu_t + w_t$, $x_0 \sim \mathcal{N}(0,I)$, $w_t \sim \mathcal{N}(0,W)$

$J(u) = \limsup_{T\to\infty} (1/T)(2/\beta)\log \mathbb{E}\{\exp\{(\beta/2)\Sigma_{t\in[0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2\}\}$

β: risk-sensitivity parameter; cannot be "high risk"; $\beta < \beta^*$

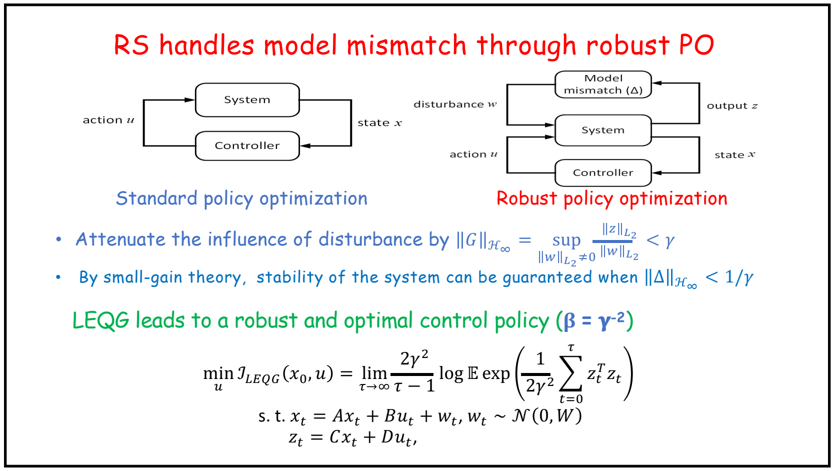
$u_t = -Kx_t$, $t\geq 0$, $J(u=-Kx) =: \mathcal{J}(K) = -(1/\beta) \log \det(1-\beta P_K W)$,

where $P_K \geq 0$ satisfies the GARE:

$P_K = Q+K'RK + (A-BK)'P_K (A-BK) + \beta(A-BK)'P_K (W^{-1}-\beta P_K)^{-1}P_K (A-BK)$

such that $\rho((A-BK)'(I-\beta P_K W)^{-1}) < 1$ and $W^{-1} - \beta P_K > 0$

16



RS handles model mismatch through robust PO

Standard policy optimization

Robust policy optimization

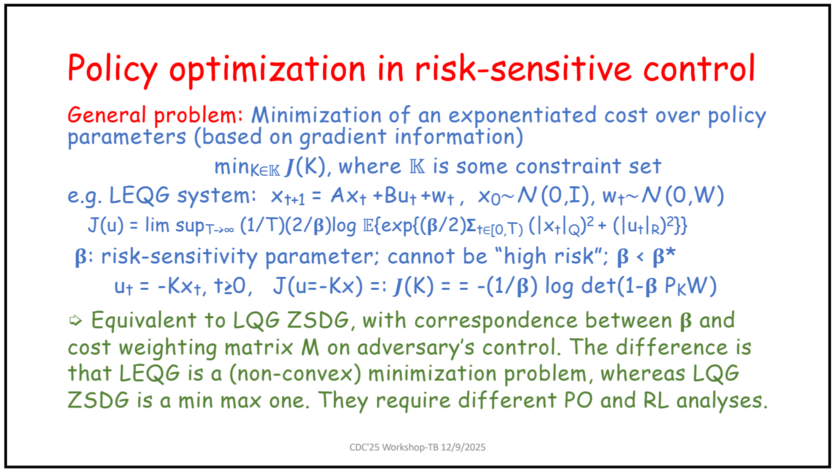
- Attenuate the influence of disturbance by $\|G\|_{\mathcal{H}_\infty} = \sup_{\|w\|_{L_2}\neq 0} \frac{\|z\|_{L_2}}{\|w\|_{L_2}} < \gamma$
- By small-gain theory, stability of the system can be guaranteed when $\|\Delta\|_{\mathcal{H}_\infty} < 1/\gamma$

LEQG leads to a robust and optimal control policy ($\beta = \gamma^{-2}$)

$$\min_u \mathcal{J}_{LEQG}(x_0,u) = \lim_{\tau\to\infty} \frac{2\gamma^2}{\tau-1} \log \mathbb{E} \exp\left(\frac{1}{2\gamma^2}\sum_{t=0}^{\tau} z_t^T z_t\right)$$

$$\text{s. t. } x_t = Ax_t + Bu_t + w_t,\ w_t \sim \mathcal{N}(0,W)$$
$$z_t = Cx_t + Du_t,$$

17



Policy optimization in risk-sensitive control

General problem: Minimization of an exponentiated cost over policy parameters (based on gradient information)

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is some constraint set

e.g. LEQG system: $x_{t+1} = Ax_t + Bu_t + w_t$, $x_0 \sim \mathcal{N}(0,I)$, $w_t \sim \mathcal{N}(0,W)$

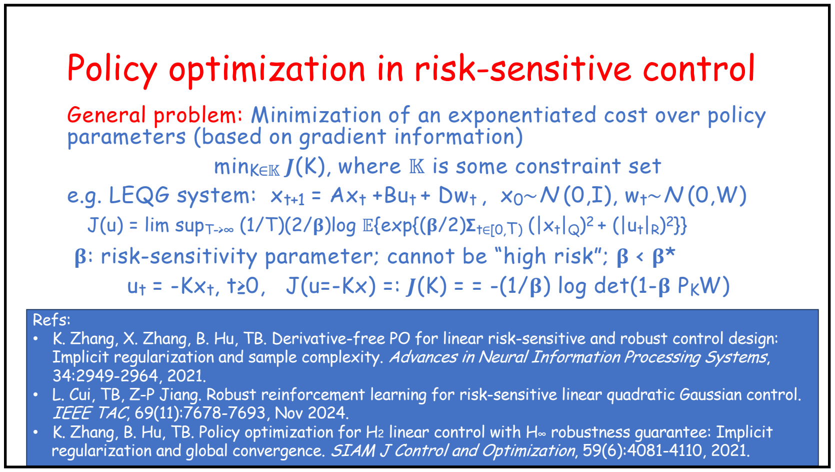
$J(u) = \limsup_{T \to \infty} (1/T)(2/\beta) \log \mathbb{E}\{\exp\{(\beta/2)\Sigma_{t\in[0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2\}\}$

β: risk-sensitivity parameter; cannot be “high risk”; $\beta < \beta^*$

$u_t = -Kx_t$, $t \geq 0$, $J(u=-Kx) =: \mathcal{J}(K) = = -(1/\beta) \log \det(1-\beta\, P_K W)$

➩ Equivalent to LQG ZSDG, with correspondence between β and cost weighting matrix M on adversary’s control. The difference is that LEQG is a (non-convex) minimization problem, whereas LQG ZSDG is a min max one. They require different PO and RL analyses.

18



Policy optimization in risk-sensitive control

General problem: Minimization of an exponentiated cost over policy parameters (based on gradient information)

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is some constraint set

e.g. LEQG system: $x_{t+1} = Ax_t + Bu_t + Dw_t$, $x_0 \sim \mathcal{N}(0,I)$, $w_t \sim \mathcal{N}(0,W)$

$J(u) = \limsup_{T \to \infty} (1/T)(2/\beta) \log \mathbb{E}\{\exp\{(\beta/2)\Sigma_{t\in[0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2\}\}$

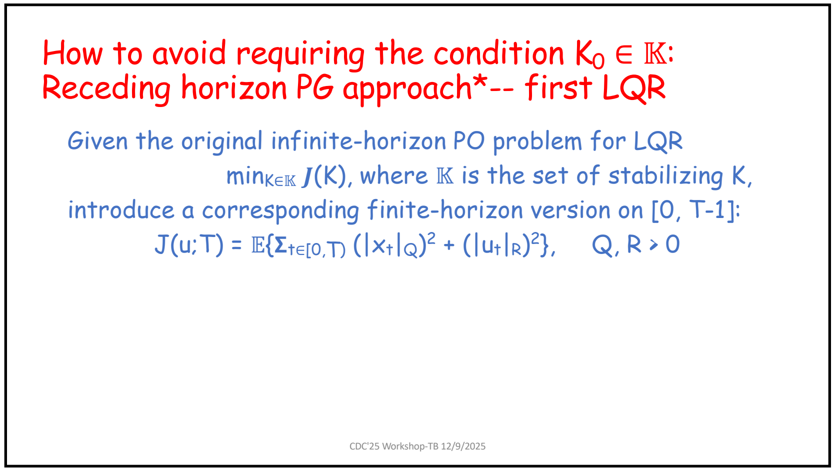
β: risk-sensitivity parameter; cannot be “high risk”; $\beta < \beta^*$

$u_t = -Kx_t$, $t \geq 0$, $J(u=-Kx) =: \mathcal{J}(K) = = -(1/\beta) \log \det(1-\beta\, P_K W)$

Refs:

- K. Zhang, X. Zhang, B. Hu, TB. Derivative-free PO for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34:2949-2964, 2021.
- L. Cui, TB, Z-P Jiang. Robust reinforcement learning for risk-sensitive linear quadratic Gaussian control. *IEEE TAC*, 69(11):7678-7693, Nov 2024.
- K. Zhang, B. Hu, TB. Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence. *SIAM J Control and Optimization*, 59(6):4081-4110, 2021.

19



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

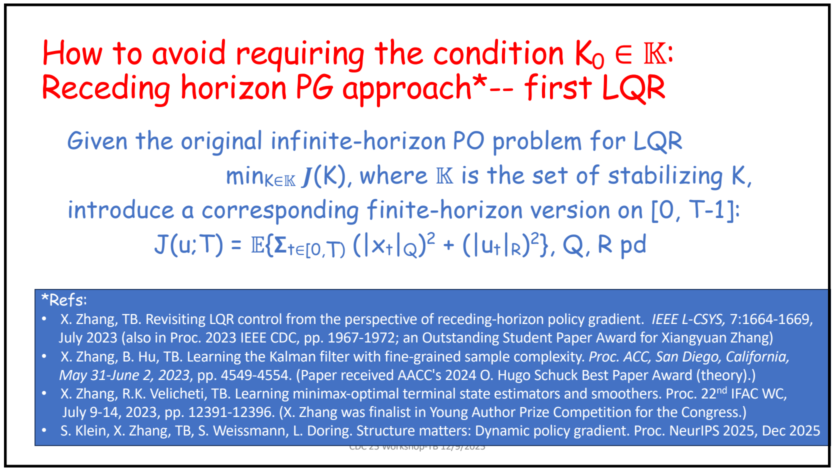
Given the original infinite-horizon PO problem for LQR

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is the set of stabilizing K,

introduce a corresponding finite-horizon version on [0, T-1]:

$J(u;T) = \mathbb{E}\{\Sigma_{t\in[0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2\}$, $Q, R > 0$

20



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Given the original infinite-horizon PO problem for LQR

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is the set of stabilizing K,

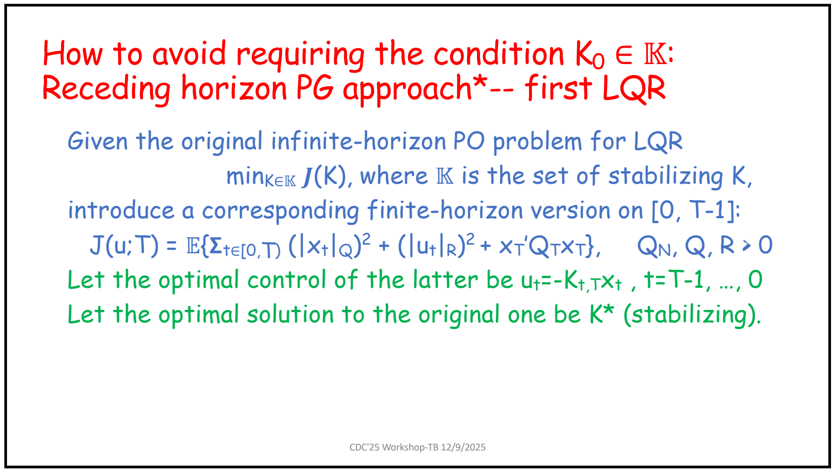
introduce a corresponding finite-horizon version on [0, T-1]:

$J(u;T) = \mathbb{E}\{\Sigma_{t\in[0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2\}$, Q, R pd

*Refs:

- X. Zhang, TB. Revisiting LQR control from the perspective of receding-horizon policy gradient. *IEEE L-CSYS,* 7:1664-1669, July 2023 (also in Proc. 2023 IEEE CDC, pp. 1967-1972; an Outstanding Student Paper Award for Xiangyuan Zhang)
- X. Zhang, B. Hu, TB. Learning the Kalman filter with fine-grained sample complexity. *Proc. ACC, San Diego, California, May 31-June 2, 2023*, pp. 4549-4554. (Paper received AACC's 2024 O. Hugo Schuck Best Paper Award (theory).)
- X. Zhang, R.K. Velicheti, TB. Learning minimax-optimal terminal state estimators and smoothers. Proc. 22nd IFAC WC, July 9-14, 2023, pp. 12391-12396. (X. Zhang was finalist in Young Author Prize Competition for the Congress.)
- S. Klein, X. Zhang, TB, S. Weissmann, L. Doring. Structure matters: Dynamic policy gradient. Proc. NeurIPS 2025, Dec 2025

21



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Given the original infinite-horizon PO problem for LQR

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is the set of stabilizing K,

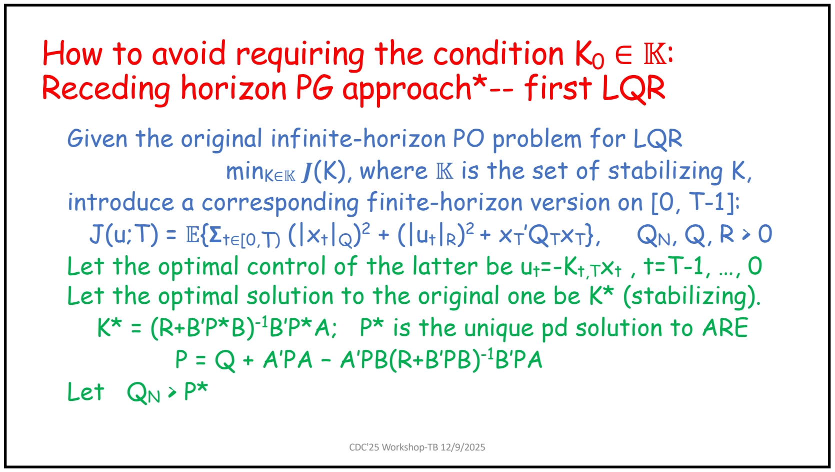
introduce a corresponding finite-horizon version on [0, T-1]:

$J(u;T) = \mathbb{E}\{\Sigma_{t \in [0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2 + x_T' Q_T x_T\}$, $Q_N, Q, R > 0$

Let the optimal control of the latter be $u_t = -K_{t,T} x_t$, t=T-1, …, 0

Let the optimal solution to the original one be K* (stabilizing).

22



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Given the original infinite-horizon PO problem for LQR

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is the set of stabilizing K,

introduce a corresponding finite-horizon version on [0, T-1]:

$J(u;T) = \mathbb{E}\{\Sigma_{t \in [0,T)} (|x_t|_Q)^2 + (|u_t|_R)^2 + x_T' Q_T x_T\}$, $Q_N, Q, R > 0$

Let the optimal control of the latter be $u_t = -K_{t,T} x_t$, t=T-1, …, 0

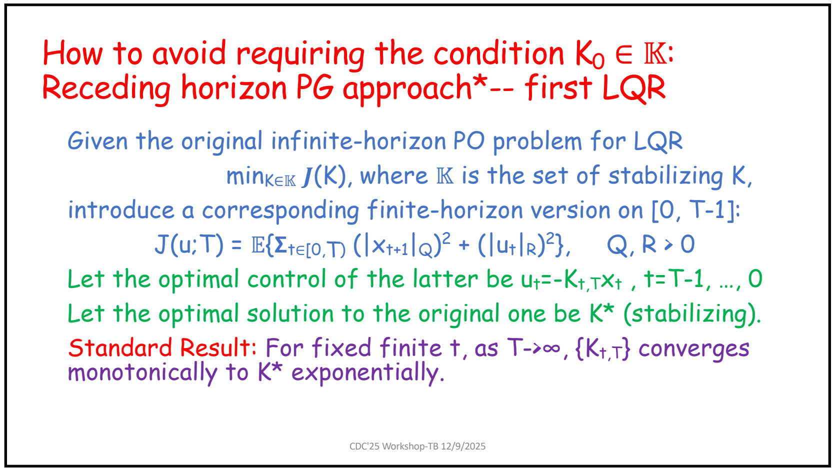
Let the optimal solution to the original one be K* (stabilizing).

$K^* = (R + B'P^*B)^{-1} B'P^*A$; P^* is the unique pd solution to ARE

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

Let $Q_N > P^*$

23



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Given the original infinite-horizon PO problem for LQR

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is the set of stabilizing K,

introduce a corresponding finite-horizon version on [0, T-1]:

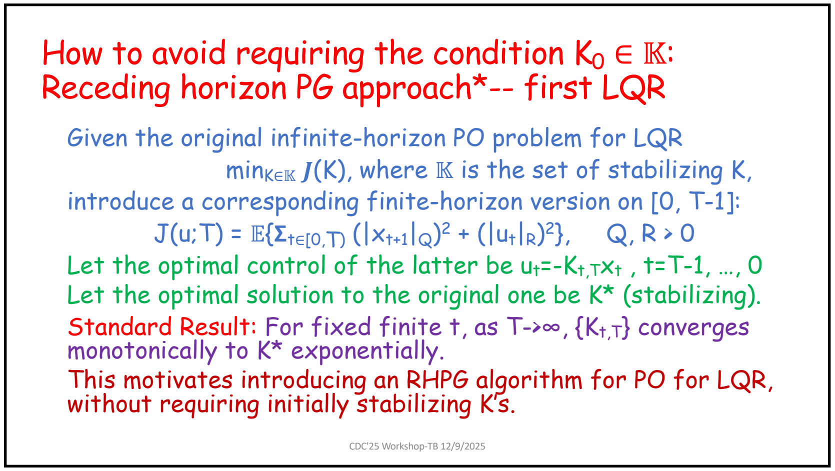
$J(u;T) = \mathbb{E}\{\Sigma_{t \in [0,T)} (|x_{t+1}|_Q)^2 + (|u_t|_R)^2\}$, $Q, R > 0$

Let the optimal control of the latter be $u_t = -K_{t,T} x_t$, t=T-1, …, 0

Let the optimal solution to the original one be K* (stabilizing).

Standard Result: For fixed finite t, as T->∞, $\{K_{t,T}\}$ converges monotonically to K* exponentially.

24



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Given the original infinite-horizon PO problem for LQR

$\min_{K \in \mathbb{K}} \mathcal{J}(K)$, where $\mathbb{K}$ is the set of stabilizing K,

introduce a corresponding finite-horizon version on [0, T-1]:

$J(u;T) = \mathbb{E}\{\Sigma_{t \in [0,T)} (|x_{t+1}|_Q)^2 + (|u_t|_R)^2\}$, $Q, R > 0$

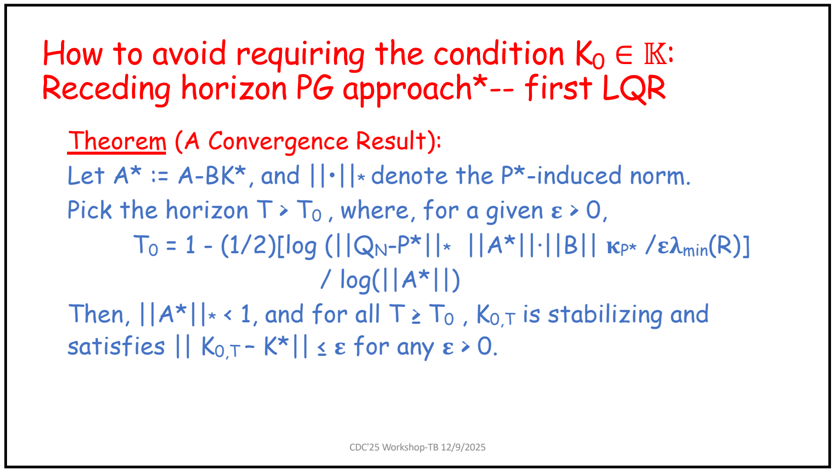
Let the optimal control of the latter be $u_t = -K_{t,T} x_t$, t=T-1, …, 0

Let the optimal solution to the original one be K* (stabilizing).

Standard Result: For fixed finite t, as T->∞, $\{K_{t,T}\}$ converges monotonically to K* exponentially.

This motivates introducing an RHPG algorithm for PO for LQR, without requiring initially stabilizing K's.

25



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Theorem (A Convergence Result):

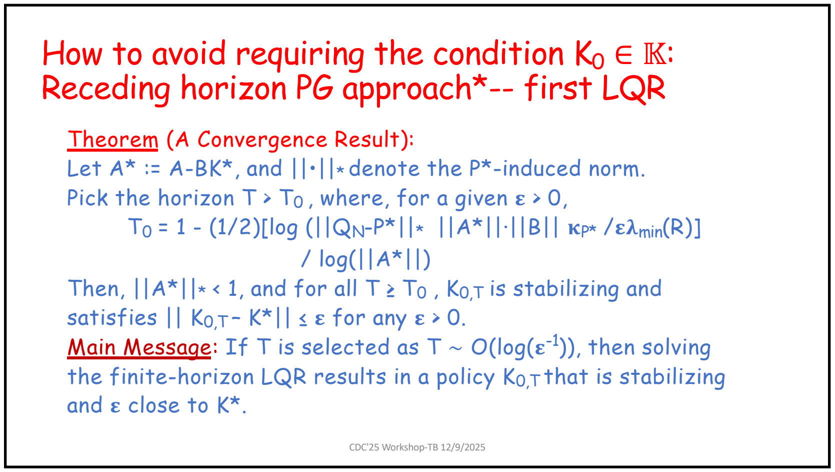
Let $A^* := A-BK^*$, and $||\cdot||_*$ denote the P*-induced norm.

Pick the horizon $T > T_0$, where, for a given $\varepsilon > 0$,

$$T_0 = 1 - (1/2)[\log(||Q_N-P^*||_* \ ||A^*||\cdot||B|| \ \kappa_{P^*} / \varepsilon\lambda_{\min}(R)] \ / \log(||A^*||)$$

Then, $||A^*||_* < 1$, and for all $T \geq T_0$, $K_{0,T}$ is stabilizing and satisfies $|| K_{0,T} - K^*|| \leq \varepsilon$ for any $\varepsilon > 0$.

26



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Theorem (A Convergence Result):

Let $A^* := A-BK^*$, and $||\cdot||_*$ denote the P*-induced norm.

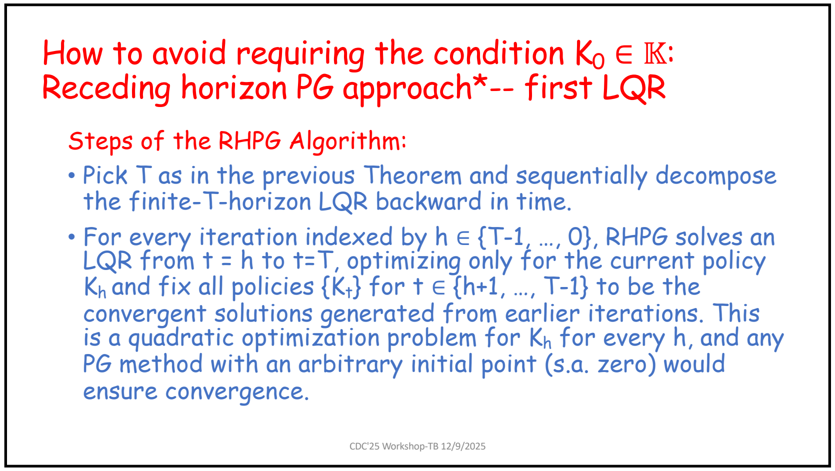
Pick the horizon $T > T_0$, where, for a given $\varepsilon > 0$,

$$T_0 = 1 - (1/2)[\log(||Q_N-P^*||_* \ ||A^*||\cdot||B|| \ \kappa_{P^*} / \varepsilon\lambda_{\min}(R)] \ / \log(||A^*||)$$

Then, $||A^*||_* < 1$, and for all $T \geq T_0$, $K_{0,T}$ is stabilizing and satisfies $|| K_{0,T} - K^*|| \leq \varepsilon$ for any $\varepsilon > 0$.

Main Message: If T is selected as $T \sim O(\log(\varepsilon^{-1}))$, then solving the finite-horizon LQR results in a policy $K_{0,T}$ that is stabilizing and ε close to K^*.

27

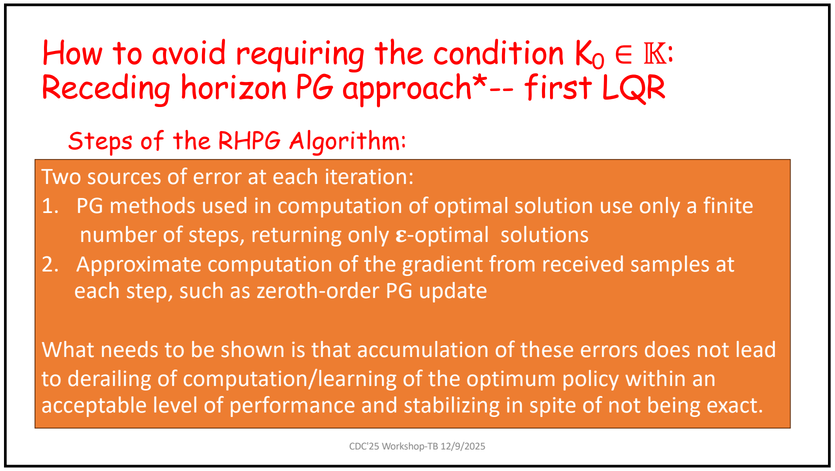


How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Steps of the RHPG Algorithm:

- Pick T as in the previous Theorem and sequentially decompose the finite-T-horizon LQR backward in time.
- For every iteration indexed by $h \in \{T-1, \ldots, 0\}$, RHPG solves an LQR from $t = h$ to $t=T$, optimizing only for the current policy K_h and fix all policies $\{K_t\}$ for $t \in \{h+1, \ldots, T-1\}$ to be the convergent solutions generated from earlier iterations. This is a quadratic optimization problem for K_h for every h, and any PG method with an arbitrary initial point (s.a. zero) would ensure convergence.

28



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

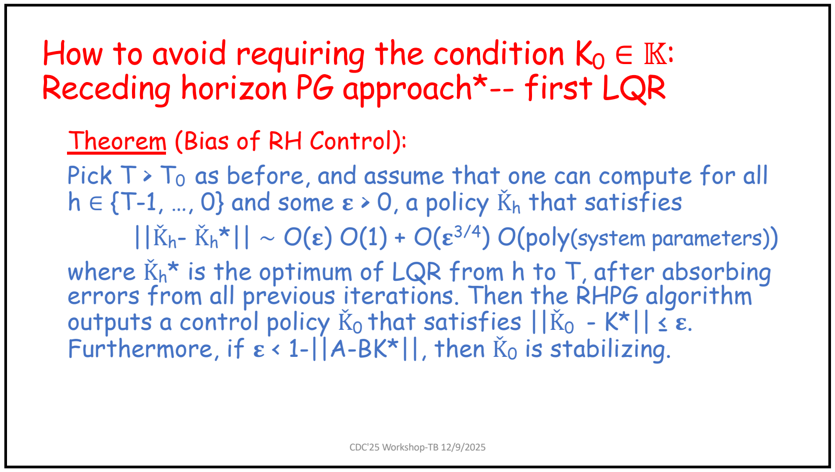
Steps of the RHPG Algorithm:

Two sources of error at each iteration:

1. PG methods used in computation of optimal solution use only a finite number of steps, returning only **ε**-optimal solutions
2. Approximate computation of the gradient from received samples at each step, such as zeroth-order PG update

What needs to be shown is that accumulation of these errors does not lead to derailing of computation/learning of the optimum policy within an acceptable level of performance and stabilizing in spite of not being exact.

29



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

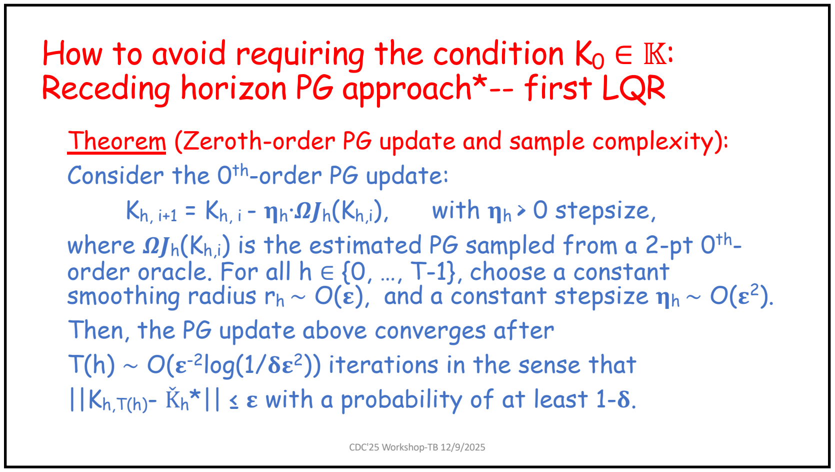
Theorem (Bias of RH Control):

Pick $T > T_0$ as before, and assume that one can compute for all $h \in \{T-1, \dots, 0\}$ and some $\varepsilon > 0$, a policy $\check{K}_h$ that satisfies

$$||\check{K}_h - \check{K}_h^*|| \sim O(\varepsilon)\, O(1) + O(\varepsilon^{3/4})\, O(\text{poly(system parameters)})$$

where $\check{K}_h^*$ is the optimum of LQR from h to T, after absorbing errors from all previous iterations. Then the RHPG algorithm outputs a control policy $\check{K}_0$ that satisfies $||\check{K}_0 - K^*|| \leq \varepsilon$. Furthermore, if $\varepsilon < 1-||A-BK^*||$, then $\check{K}_0$ is stabilizing.

30



How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Theorem (Zeroth-order PG update and sample complexity):

Consider the 0th-order PG update:

$$K_{h,\, i+1} = K_{h,\, i} - \eta_h \cdot \Omega J_h(K_{h,i}), \qquad \text{with } \eta_h > 0 \text{ stepsize,}$$

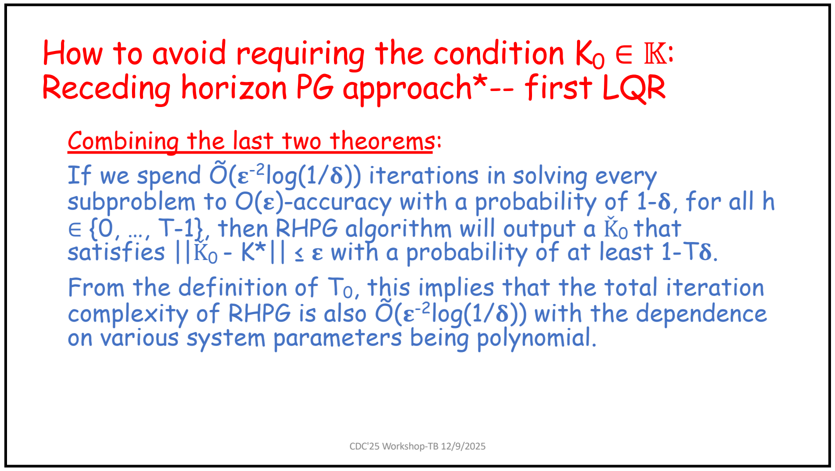
where $\Omega J_h(K_{h,i})$ is the estimated PG sampled from a 2-pt 0th-order oracle. For all $h \in \{0, \dots, T-1\}$, choose a constant smoothing radius $r_h \sim O(\varepsilon)$, and a constant stepsize $\eta_h \sim O(\varepsilon^2)$.

Then, the PG update above converges after

$T(h) \sim O(\varepsilon^{-2}\log(1/\delta\varepsilon^2))$ iterations in the sense that

$||K_{h,T(h)} - \check{K}_h^*|| \leq \varepsilon$ with a probability of at least $1-\delta$.

31



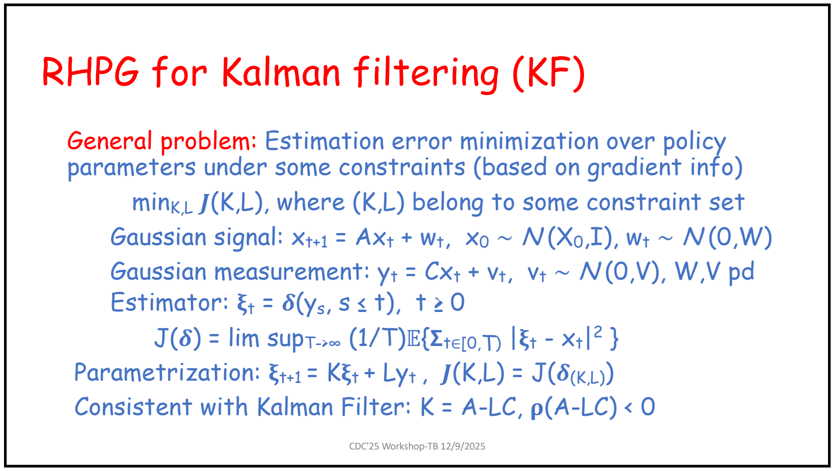
How to avoid requiring the condition $K_0 \in \mathbb{K}$: Receding horizon PG approach*-- first LQR

Combining the last two theorems:

If we spend $\tilde{O}(\varepsilon^{-2}\log(1/\delta))$ iterations in solving every subproblem to $O(\varepsilon)$-accuracy with a probability of $1-\delta$, for all $h \in \{0, \dots, T-1\}$, then RHPG algorithm will output a $\check{K}_0$ that satisfies $||\check{K}_0 - K^*|| \leq \varepsilon$ with a probability of at least $1-T\delta$.

From the definition of T_0, this implies that the total iteration complexity of RHPG is also $\tilde{O}(\varepsilon^{-2}\log(1/\delta))$ with the dependence on various system parameters being polynomial.

32



RHPG for Kalman filtering (KF)

General problem: Estimation error minimization over policy parameters under some constraints (based on gradient info)

$\min_{K,L} J(K,L)$, where (K,L) belong to some constraint set

Gaussian signal: $x_{t+1} = Ax_t + w_t,\ \ x_0 \sim N(X_0, I),\ w_t \sim N(0,W)$

Gaussian measurement: $y_t = Cx_t + v_t,\ \ v_t \sim N(0,V)$, W,V pd

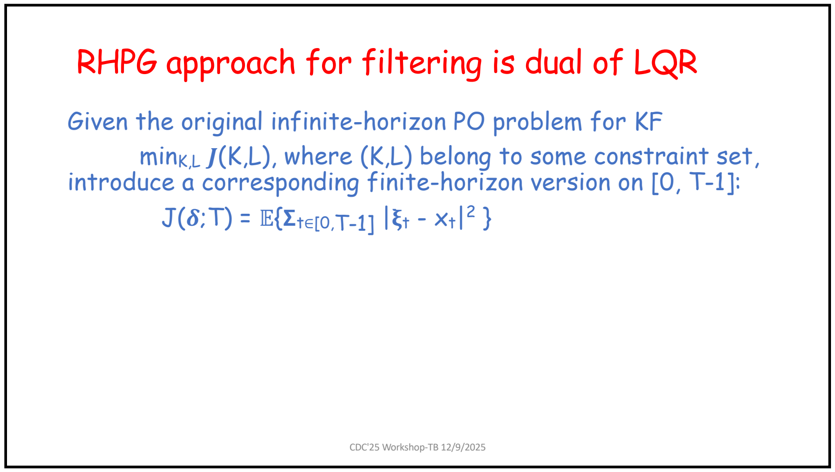
Estimator: $\xi_t = \delta(y_s, s \leq t),\ \ t \geq 0$

$$J(\delta) = \limsup_{T\to\infty} (1/T)\mathbb{E}\{\Sigma_{t\in[0,T)} |\xi_t - x_t|^2\}$$

Parametrization: $\xi_{t+1} = K\xi_t + Ly_t,\ \ J(K,L) = J(\delta_{(K,L)})$

Consistent with Kalman Filter: $K = A-LC$, $\rho(A-LC) < 0$

33



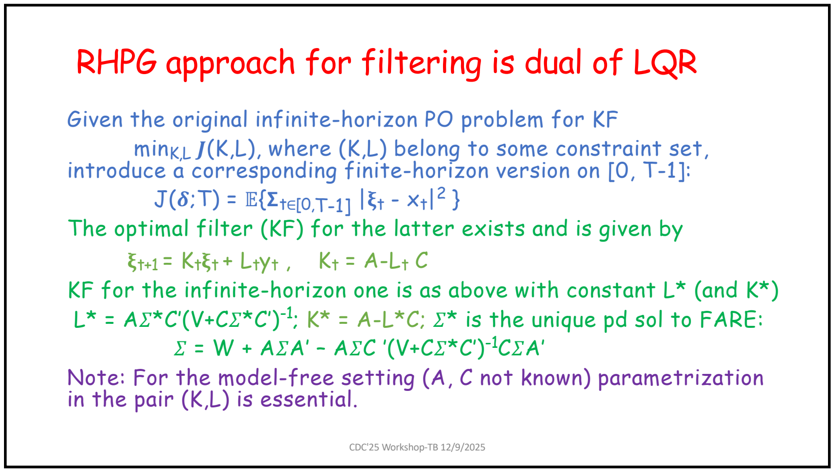
RHPG approach for filtering is dual of LQR

Given the original infinite-horizon PO problem for KF

$\min_{K,L} J(K,L)$, where (K,L) belong to some constraint set, introduce a corresponding finite-horizon version on $[0, T-1]$:

$$J(\delta;T) = \mathbb{E}\{\Sigma_{t\in[0,T-1]} |\xi_t - x_t|^2\}$$

34



RHPG approach for filtering is dual of LQR

Given the original infinite-horizon PO problem for KF

$\min_{K,L} J(K,L)$, where (K,L) belong to some constraint set, introduce a corresponding finite-horizon version on $[0, T-1]$:

$$J(\delta;T) = \mathbb{E}\{\Sigma_{t\in[0,T-1]} |\xi_t - x_t|^2\}$$

The optimal filter (KF) for the latter exists and is given by

$$\xi_{t+1} = K_t\xi_t + L_t y_t , \quad K_t = A-L_t C$$

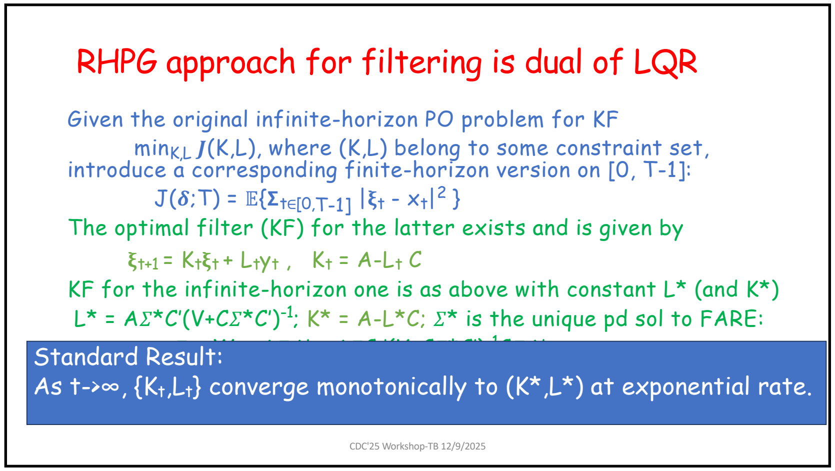
KF for the infinite-horizon one is as above with constant L^* (and K^*)

$L^* = A\Sigma^*C'(V+C\Sigma^*C')^{-1}$; $K^* = A-L^*C$; Σ^* is the unique pd sol to FARE:

$$\Sigma = W + A\Sigma A' - A\Sigma C'(V+C\Sigma^*C')^{-1}C\Sigma A'$$

Note: For the model-free setting (A, C not known) parametrization in the pair (K,L) is essential.

35



RHPG approach for filtering is dual of LQR

Given the original infinite-horizon PO problem for KF

$\min_{K,L} J(K,L)$, where (K,L) belong to some constraint set, introduce a corresponding finite-horizon version on $[0, T-1]$:

$$J(\delta;T) = \mathbb{E}\{\Sigma_{t\in[0,T-1]} |\xi_t - x_t|^2\}$$

The optimal filter (KF) for the latter exists and is given by

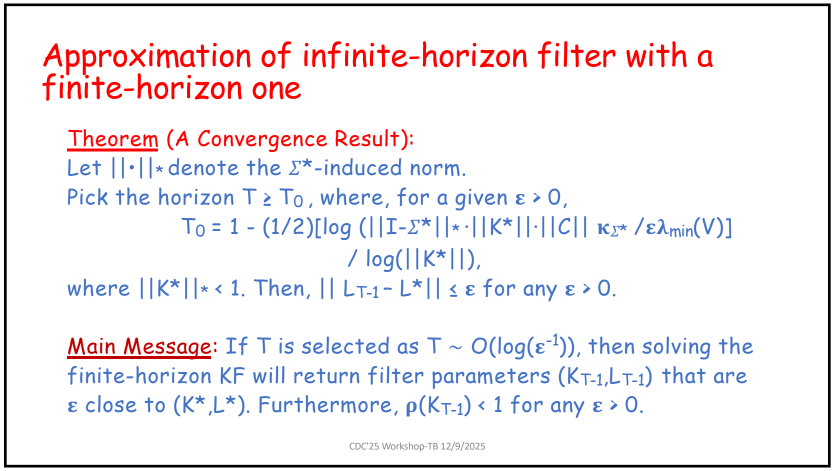
$$\xi_{t+1} = K_t\xi_t + L_t y_t , \quad K_t = A-L_t C$$

KF for the infinite-horizon one is as above with constant L^* (and K^*)

$L^* = A\Sigma^*C'(V+C\Sigma^*C')^{-1}$; $K^* = A-L^*C$; Σ^* is the unique pd sol to FARE:

Standard Result:
As $t \to \infty$, $\{K_t,L_t\}$ converge monotonically to (K^*,L^*) at exponential rate.

36



Approximation of infinite-horizon filter with a finite-horizon one

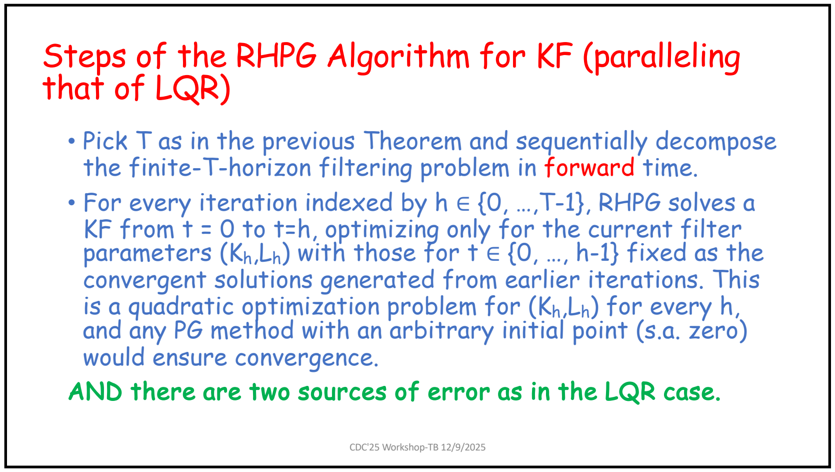
Theorem (A Convergence Result):
Let $||\cdot||_*$ denote the Σ^*-induced norm.
Pick the horizon $T \geq T_0$, where, for a given $\varepsilon > 0$,

$$T_0 = 1 - (1/2)[\log (||I-\Sigma^*||_* \cdot ||K^*|| \cdot ||C|| \, \kappa_{\Sigma^*} / \varepsilon\lambda_{\min}(V)] \, / \log(||K^*||),$$

where $||K^*||_* < 1$. Then, $|| L_{T-1} - L^*|| \leq \varepsilon$ for any $\varepsilon > 0$.

Main Message: If T is selected as $T \sim O(\log(\varepsilon^{-1}))$, then solving the finite-horizon KF will return filter parameters (K_{T-1},L_{T-1}) that are ε close to (K^*,L^*). Furthermore, $\rho(K_{T-1}) < 1$ for any $\varepsilon > 0$.

37

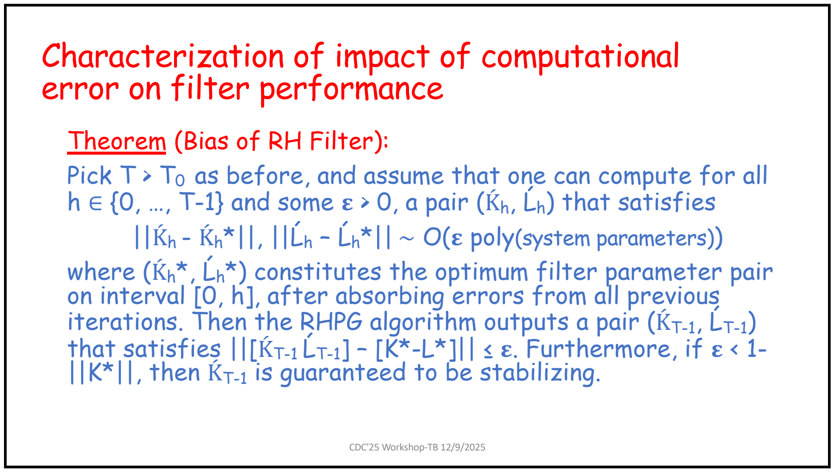


Steps of the RHPG Algorithm for KF (paralleling that of LQR)

- Pick T as in the previous Theorem and sequentially decompose the finite-T-horizon filtering problem in forward time.
- For every iteration indexed by $h \in \{0, \ldots, T-1\}$, RHPG solves a KF from $t = 0$ to $t=h$, optimizing only for the current filter parameters (K_h, L_h) with those for $t \in \{0, \ldots, h-1\}$ fixed as the convergent solutions generated from earlier iterations. This is a quadratic optimization problem for (K_h, L_h) for every h, and any PG method with an arbitrary initial point (s.a. zero) would ensure convergence.

AND there are two sources of error as in the LQR case.

38



Characterization of impact of computational error on filter performance

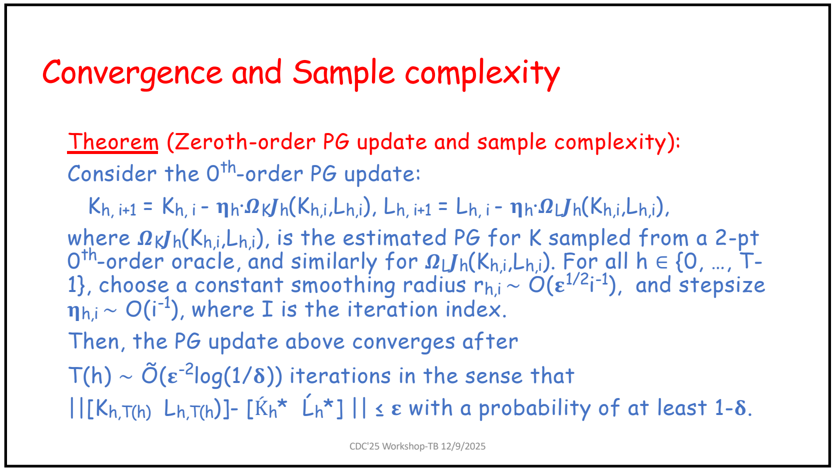
Theorem (Bias of RH Filter):

Pick $T > T_0$ as before, and assume that one can compute for all $h \in \{0, \ldots, T-1\}$ and some $\varepsilon > 0$, a pair $(\acute{K}_h, \acute{L}_h)$ that satisfies

$$||\acute{K}_h - \acute{K}_h^*||,\ ||\acute{L}_h - \acute{L}_h^*|| \sim O(\varepsilon \text{ poly(system parameters)})$$

where $(\acute{K}_h^*, \acute{L}_h^*)$ constitutes the optimum filter parameter pair on interval $[0, h]$, after absorbing errors from all previous iterations. Then the RHPG algorithm outputs a pair $(\acute{K}_{T-1}, \acute{L}_{T-1})$ that satisfies $||[\acute{K}_{T-1}\ \acute{L}_{T-1}] - [K^*-L^*]|| \leq \varepsilon$. Furthermore, if $\varepsilon < 1-||K^*||$, then $\acute{K}_{T-1}$ is guaranteed to be stabilizing.

39



Convergence and Sample complexity

Theorem (Zeroth-order PG update and sample complexity):

Consider the 0th-order PG update:

$$K_{h,\, i+1} = K_{h,\, i} - \eta_h \cdot \Omega_K J_h(K_{h,i}, L_{h,i}),\ L_{h,\, i+1} = L_{h,\, i} - \eta_h \cdot \Omega_L J_h(K_{h,i}, L_{h,i}),$$

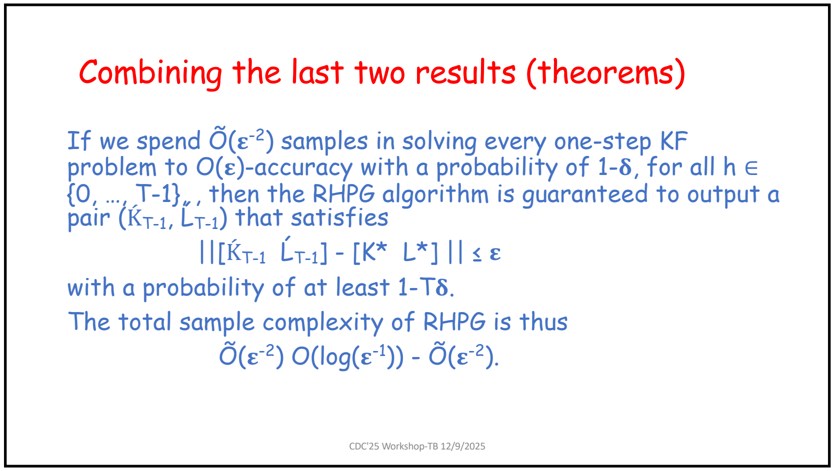
where $\Omega_K J_h(K_{h,i}, L_{h,i})$, is the estimated PG for K sampled from a 2-pt 0th-order oracle, and similarly for $\Omega_L J_h(K_{h,i}, L_{h,i})$. For all $h \in \{0, \ldots, T-1\}$, choose a constant smoothing radius $r_{h,i} \sim O(\varepsilon^{1/2} i^{-1})$, and stepsize $\eta_{h,i} \sim O(i^{-1})$, where I is the iteration index.

Then, the PG update above converges after

$T(h) \sim \tilde{O}(\varepsilon^{-2} \log(1/\delta))$ iterations in the sense that

$||[K_{h,T(h)}\ \ L_{h,T(h)}] - [\acute{K}_h^*\ \ \acute{L}_h^*]\ || \leq \varepsilon$ with a probability of at least $1-\delta$.

40



Combining the last two results (theorems)

If we spend $\tilde{O}(\varepsilon^{-2})$ samples in solving every one-step KF problem to $O(\varepsilon)$-accuracy with a probability of $1-\delta$, for all $h \in \{0, \ldots, T-1\}$, , then the RHPG algorithm is guaranteed to output a pair $(\acute{K}_{T-1}, \acute{L}_{T-1})$ that satisfies

$$||[\acute{K}_{T-1}\ \ \acute{L}_{T-1}] - [K^*\ \ L^*]\ || \leq \varepsilon$$

with a probability of at least $1-T\delta$.

The total sample complexity of RHPG is thus

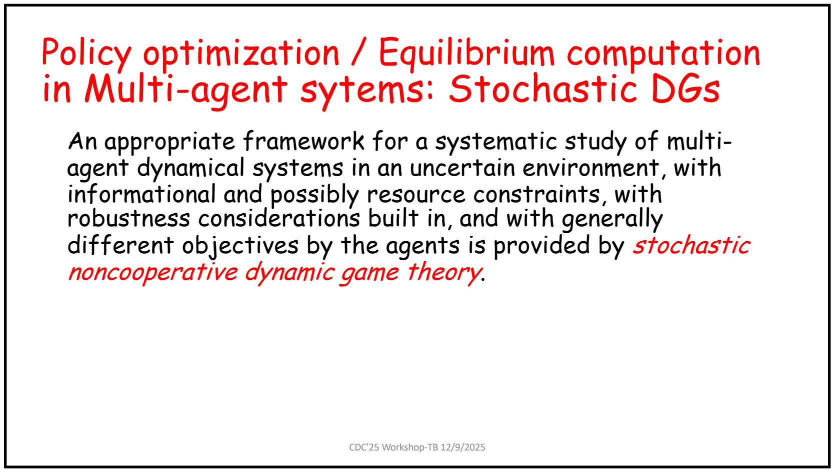
$$\tilde{O}(\varepsilon^{-2})\ O(\log(\varepsilon^{-1})) - \tilde{O}(\varepsilon^{-2}).$$

41

Multi-Agent Systems – PO, Eqm Computation, RL & Mean-Field Games

CDC'25 Workshop-TB 12/9/2025

42

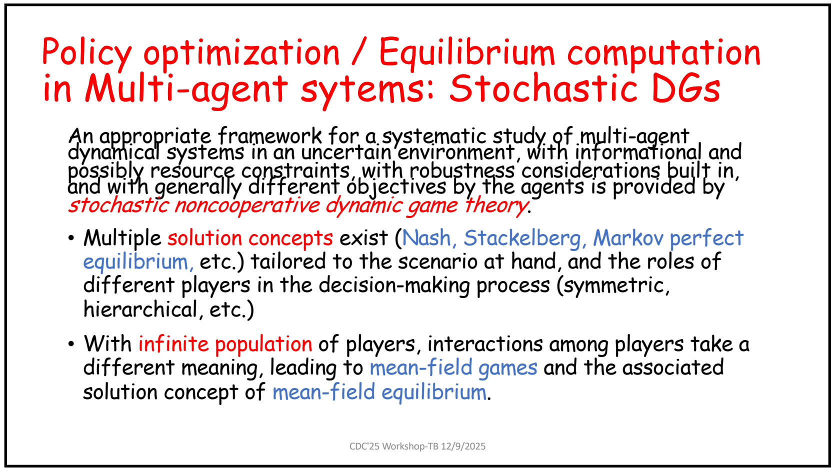


Policy optimization / Equilibrium computation in Multi-agent sytems: Stochastic DGs

An appropriate framework for a systematic study of multi-agent dynamical systems in an uncertain environment, with informational and possibly resource constraints, with robustness considerations built in, and with generally different objectives by the agents is provided by *stochastic noncooperative dynamic game theory*.

CDC'25 Workshop-TB 12/9/2025

43



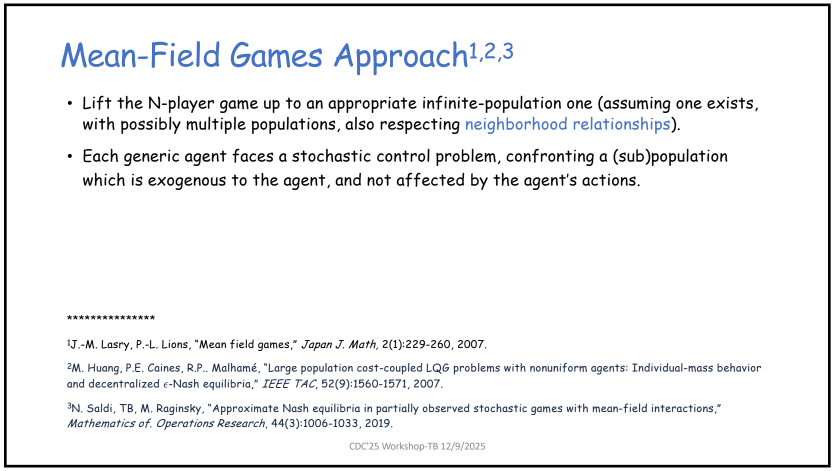
Policy optimization / Equilibrium computation in Multi-agent sytems: Stochastic DGs

An appropriate framework for a systematic study of multi-agent dynamical systems in an uncertain environment, with informational and possibly resource constraints, with robustness considerations built in, and with generally different objectives by the agents is provided by *stochastic noncooperative dynamic game theory*.

- Multiple solution concepts exist (Nash, Stackelberg, Markov perfect equilibrium, etc.) tailored to the scenario at hand, and the roles of different players in the decision-making process (symmetric, hierarchical, etc.)
- With infinite population of players, interactions among players take a different meaning, leading to mean-field games and the associated solution concept of mean-field equilibrium.

CDC'25 Workshop-TB 12/9/2025

44



Mean-Field Games Approach[1,2,3]

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.

****

[1] J.-M. Lasry, P.-L. Lions, "Mean field games," *Japan J. Math*, 2(1):229-260, 2007.

[2] M. Huang, P.E. Caines, R.P.. Malhamé, "Large population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-Nash equilibria," *IEEE TAC*, 52(9):1560-1571, 2007.

[3] N. Saldi, TB, M. Raginsky, "Approximate Nash equilibria in partially observed stochastic games with mean-field interactions," *Mathematics of. Operations Research*, 44(3):1006-1033, 2019.

CDC'25 Workshop-TB 12/9/2025

45

Mean-Field Games Approach[1,2,3]

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE).

****

[1] J.-M. Lasry, P.-L. Lions, "Mean field games," *Japan J. Math,* 2(1):229-260, 2007.

[2] M. Huang, P.E. Caines, R.P.. Malhamé, "Large population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-Nash equilibria," *IEEE TAC*, 52(9):1560-1571, 2007.

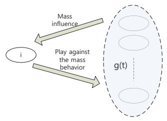
[3] N. Saldi, TB, M. Raginsky, "Approximate Nash equilibria in partially observed stochastic games with mean-field interactions," *Mathematics of Operations Research*, 44(3):1006-1033, 2019.

46

Mean-Field Games Approach

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE).

Schematically for a single population with exogenous process g:



- Stochastic control problem for generic agent leads to an optimal policy, say μ^*, that depends on g (and only local information for the agent)
- Use that policy in the state equation of the generic agent, and find g so that it is consistent with the emerging state process (FP)—g^*
- (μ^*, g^*) constitutes the MFE

47

Mean-Field Games Approach

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE).
- It is possible to build in robustness through a risk-sensitive formulation (working with exponentiated loss functions for the agents)—connection to introducing an adversary agent, with generic agent now facing a zero-sum stochastic dynamic game.[4,5,6]

****

[4] H. Tembine, Q. Zhu, TB, "Risk-sensitive mean-field games," *IEEE TAC, 59*(4):835-850, April 2014.

[5] J. Moon, TB, "Linear-quadratic risk-sensitive and robust mean-field games," *IEEE TAC*, 62(3):1062-1077, March 2017; --"Risk-sensitive mean field games via the stochastic maximum principle," *Dynamic Games and Applications*, 9:1100-1125, 2019.

[6] N. Saldi, TB, M. Raginsky, "Approximate Markov-Nash equilibria for discrete-time risk-sensitive mean-field games," *Mathematics of Operations Research*, 45(4):1596-1620, Nov 2020.

48

Mean-Field Games Approach

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE)
- Finally, study the relationship between finite N and infinite N solutions—leading to ε-NE, thus resolving the formidable task of obtaining approximate NE for games with asymmetric information (such as local measurements only), where $\varepsilon \rightarrow 0$ as $N \rightarrow \infty$.

49

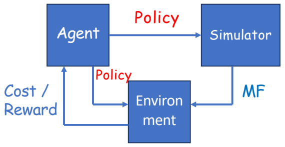
Computational Aspects

- MFE-based approximate NE policy is scalable
- Computation of the mean field at NE requires solution of a fixed-point equation, which requires full modelling knowledge
- One way around this is for each agent to interact with a central coordinator (simulator) who collects state values and/or policies of the agents, computes the mean field, and broadcasts to all agents, who then update their policies based on the received MF, … and so on. With a finite number, L, of different populations of agents, L different MFs are computed.

50

Computational Aspects

- MFE-based approximate NE policy is scalable
- Computation of the mean field at NE requires solution of a fixed-point equation, which requires full modelling knowledge
- One way around this is for each agent to interact with a central coordinator (simulator) who collects state values and/or policies of the agents, computes the mean field, and broadcasts to all agents, who then update their policies based on the received MF, … and so on. With a finite number, L, of different populations of agents, L different MFs are computed.



Single population schematic: Generic Agent (A) interacts with the Simulator (S) and the Environment (E), feeding policy and/or state values. S computes the MF, feeding it to E, where cost/reward of A is generated and sent to A, who updates its policy based on some optimization algorithm.

51

Computational Aspects with Learning

- What if the agents do not know their own models? Then bring in RL for each agent into the iterative/learning process
- Parametrize the policies and optimize over the parameters, using e.g., policy gradient, respecting also computation and communication constraints[1,2,3]
- When explicit form of the agent's objective function is not available, its gradient can be computed only approximately, using e.g., **zero-order stochastic optimization (ZSO)**
- This will require further study of finite sample guarantees for the underlying algorithms

****

[1] T. Li, G. Peng, Q. Zhu, TB, "The confluence of networks, games, and learning: A game-theoretic framework for multiagent decision making over networks," *IEEE Control Systems Magazine*, 42(4):35-67, August 2022.

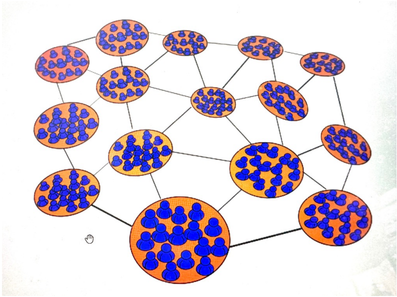
[2] T. Chen, K. Zhang, G.B. Giannakis, TB, "Communication-efficient policy gradient methods for distributed reinforcement learning," *IEEE TCNS*, 9(2):917-929, June 2022.

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

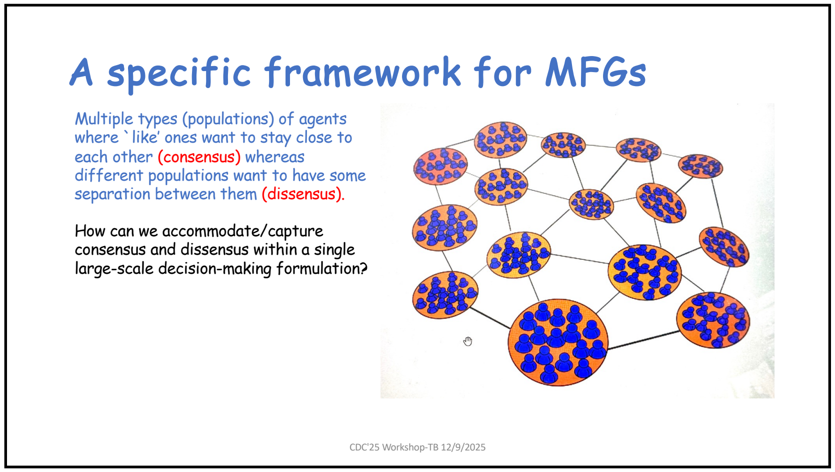
52

A specific framework for MFGs

Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).



53



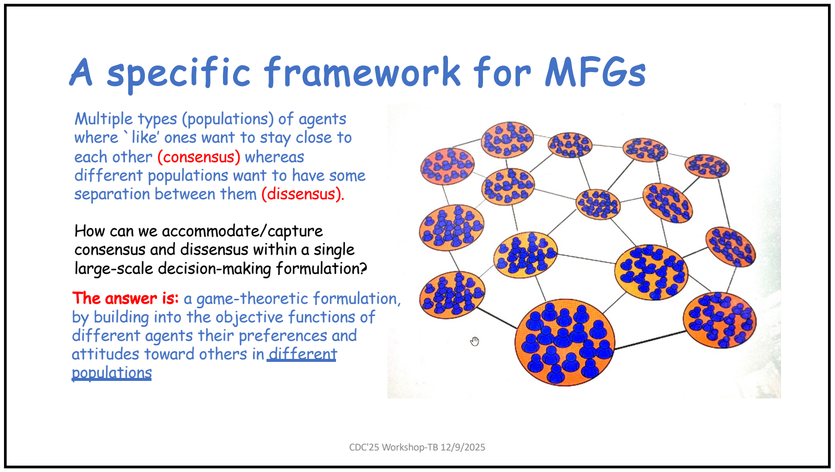
A specific framework for MFGs

Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).

How can we accommodate/capture consensus and dissensus within a single large-scale decision-making formulation?

CDC'25 Workshop-TB 12/9/2025

54



A specific framework for MFGs

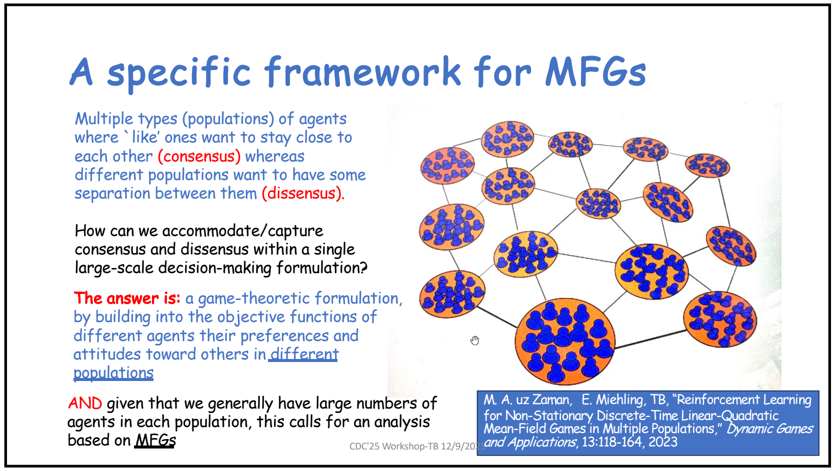
Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).

How can we accommodate/capture consensus and dissensus within a single large-scale decision-making formulation?

The answer is: a game-theoretic formulation, by building into the objective functions of different agents their preferences and attitudes toward others in different populations

CDC'25 Workshop-TB 12/9/2025

55



A specific framework for MFGs

Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).

How can we accommodate/capture consensus and dissensus within a single large-scale decision-making formulation?

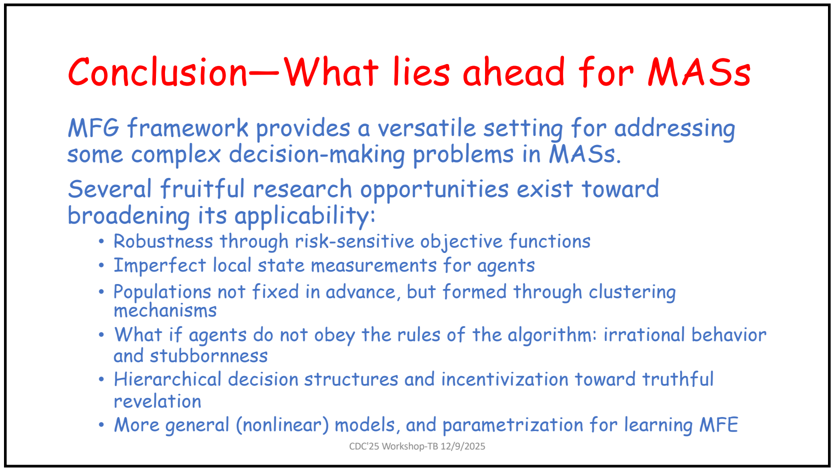
The answer is: a game-theoretic formulation, by building into the objective functions of different agents their preferences and attitudes toward others in different populations

AND given that we generally have large numbers of agents in each population, this calls for an analysis based on MFGs

M. A. uz Zaman, E. Miehling, TB, "Reinforcement Learning for Non-Stationary Discrete-Time Linear-Quadratic Mean-Field Games in Multiple Populations," *Dynamic Games and Applications*, 13:118-164, 2023

CDC'25 Workshop-TB 12/9/20

56



Conclusion—What lies ahead for MASs

MFG framework provides a versatile setting for addressing some complex decision-making problems in MASs.

Several fruitful research opportunities exist toward broadening its applicability:

- Robustness through risk-sensitive objective functions
- Imperfect local state measurements for agents
- Populations not fixed in advance, but formed through clustering mechanisms
- What if agents do not obey the rules of the algorithm: irrational behavior and stubbornness
- Hierarchical decision structures and incentivization toward truthful revelation
- More general (nonlinear) models, and parametrization for learning MFE

CDC'25 Workshop-TB 12/9/2025

57