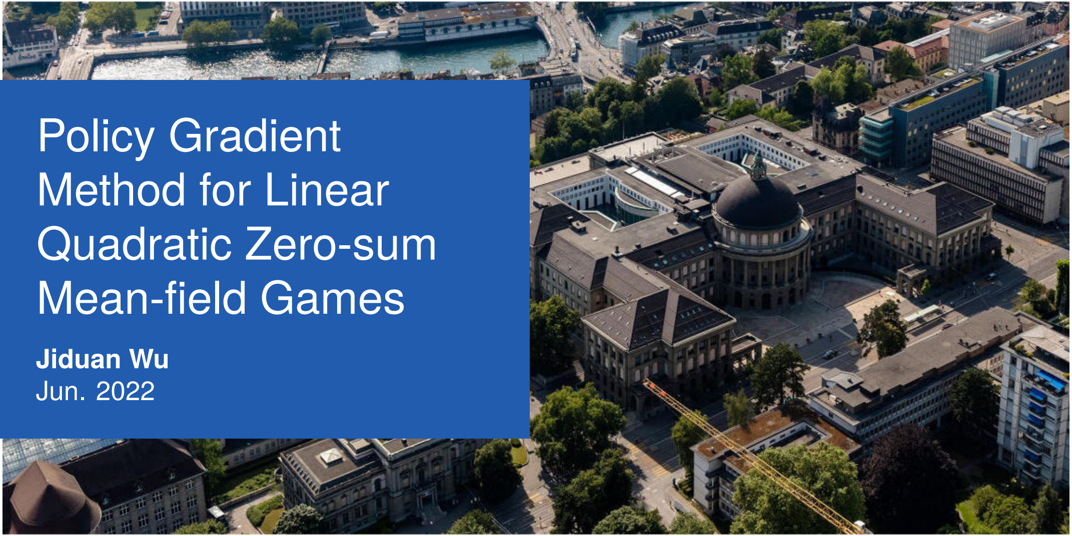
ETH zürich

Optimization & Decision Intelligence Group



Policy Gradient Method for Linear Quadratic Zero-sum Mean-field Games

Jiduan Wu

Jun. 2022

Outline

1. Introduction

2. Model-based Nested Policy Gradient Algorithm

3. Model-free Nested Policy Gradient Algorithm

Outline

1. Introduction

2. Model-based Nested Policy Gradient Algorithm

3. Model-free Nested Policy Gradient Algorithm

Linear-Quadratic Regulator (LQR) Problem Setting

Linear-Quadratic Regulator (LQR) Problem

Consider the case of the linear system

$$x_{t+1} = Ax_t + Bu_t, \quad t = 0, 1, \cdots$$

where x_t is the state, u_t is the control.

Consider the utility function

$$C(\{u_t\}_{t\geq 0}) = \mathbb{E}_{x_0}\left[\sum_{k=0}^{\infty} x_k^T Q x_k + u_t^T R u_t\right]$$

$$\min_{\{u_t\}_{t\geq 0}} C(\{u_t\}_{t\geq 0})$$

where $Q, R \succ 0$. The minimizing policy is static and linear

$$u_t^* = K^* x_t^*$$

Linear-Quadratic Zero-sum Mean-field Games

Problem Setting [Carmona et al., 2020, 2019]

We consider the evolution of the system state x_t

$$x_{t+1} = Ax_t + \bar{A}\bar{x}_t + B_1 u_{1,t} + \bar{B}_1 \bar{u}_{1,t} + B_2 u_{2,t} + \bar{B}_2 \bar{u}_{2,t} + \epsilon^0_{t+1} + \epsilon^1_{t+1}$$

with initial condition $\epsilon^0_0 + \epsilon^1_0$

- $(\epsilon^0_t)_{t\geq 0}$ and $(\epsilon^1_t)_{t\geq 0}$ are common and idiosyncratic noise respectively.
- $u_{1,t}, u_{2,t}$ are controls of two controllers at time t.

Consider utility function, $u_1 := \{u_{1,t}\}_{t\geq 0}, u_2 := \{u_{2,t}\}_{t\geq 0}$

$$C(u_1, u_2) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t c_t\right]$$

$$\begin{aligned} c_t = {} & (x_t - \bar{x}_t)^T Q (x_t - \bar{x}_t) + \bar{x}_t^T (Q + \bar{Q}) \bar{x}_t + (u_{1,t} - \bar{u}_{1,t})^T R_1 (u_{1,t} - \bar{u}_{1,t}) + \bar{u}_{1,t}^T (R_1 + \bar{R}_1) \bar{u}_{1,t} \\ & - (u_{2,t} - \bar{u}_{2,t})^T R_2 (u_{2,t} - \bar{u}_{2,t}) - \bar{u}_{2,t}^T (R_2 + \bar{R}_2) \bar{u}_{2,t} \end{aligned}$$

where $Q, Q + \bar{Q}, R_i, R_i + \bar{R}_i \succ 0$, $i = 1, 2$. Try to find the Nash equilibrium

$$C(u_1^*, u_2^*) = \inf_{u_1} \sup_{u_2} C(u_1, u_2)$$

Differences

- Mean-field terms: A workaround for the curse of dimensionality of the multi-agent system.

$$e.g. \quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x^i$$

- Idiosyncratic noises: $\bar{x}_t$ and $\bar{u}_t$ are the conditional mean of x_t and u_t given common noises, $(\epsilon_s^0)_{s=0,\cdots,t}$ [Carmona et al., 2019].

$$\bar{x}_t = \mathbb{E}[x_t|(\epsilon_s^0)_{0\leq s\leq t}]$$
$$\bar{u}_t = \mathbb{E}[u_t|(\epsilon_s^0)_{0\leq s\leq t}]$$

  Idiosyncratic noises are *necessary* for mean-field case.

- Discounted coefficient γ: make optimal cost finite.

Decomposition [Carmona et al., 2020]

Reparametrization trick

for every $t \geq 0$, let

$$y_t = x_t - \bar{x}_t, \quad z_t = \bar{x}_t$$

$$u_{1,t}^{(y)} := u_{1,t} - \bar{u}_{1,t} = -K_1 y_t, \qquad u_{2,t}^{(y)} := u_{2,t} - \bar{u}_{2,t} = K_2 y_t$$

$$u_{1,t}^{(z)} := \bar{u}_{1,t} = -L_1 z_t, \qquad u_{2,t}^{(z)} := \bar{u}_{2,t} = L_2 z_t$$

- Rewrite dynamics

$$y_{t+1} = A y_t + B_1 u_{1,t}^{(y)} + B_2 u_{2,t}^{(y)} + \epsilon_{t+1}^1, \quad y_0 \sim \epsilon_0^1, \quad z_{t+1} = \bar{A} z_t + \bar{B}_1 u_{1,t}^{(z)} + \bar{B}_2 u_{2,t}^{(z)} + \epsilon_{t+1}^0, \quad z_0 \sim \epsilon_0^0$$

- Decouple the utility function $C(K_1, K_2, L_1, L_2) = C_y(K_1, K_2) + C_z(L_1, L_2)$

$$C_y(K_1, K_2) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (y_t)^T (Q + K_1^T R_1 K_1 - K_2^T R_2 K_2) y_t\right]$$

$$C_z(L_1, L_2) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (z_t)^T (Q + \bar{Q} + L_1^T (R_1 + \bar{R}_1) L_1 - L_2^T (R_2 + \bar{R}_2) L_2) z_t\right]$$

Key Properties

- Connected Stabilizing Region: We consider control pairs on $\mathcal{S} := \mathcal{S}_1 \times \mathcal{S}_2$ s.t. $C(K_1, K_2, L_1, L_2) < \infty$.

$$\mathcal{S}_1 := \{(K_1, K_2) | \gamma\rho(A - B_1K_1 + B_2K_2) < 1\}$$
$$\mathcal{S}_2 := \{(L_1, L_2) | \gamma\rho(A + \bar{A} - (B_1 + \bar{B}_1)L_1 + (K_2 + \bar{K}_2)L_2 < 1\}$$

Theorem

The stabilizing region $\mathcal{S}_1$ ($\mathcal{S}_2$) is connected. Hence $\mathcal{S}$ is also connected.

This property justifies the policy gradient methods.

- Non-convexity: $\mathcal{S}_1, \mathcal{S}_2$ are not convex [Zhang et al., 2019].

$$K_1 = \begin{bmatrix} 1 & 0 & -10 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad K_1' = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

$A = B_1 = \frac{1}{\gamma} I_3, K_2 = K_2' = 0.$ $\rho(A - B_1K_1 - B_2K_2) < 1$, $\rho(A - B_1K_1' - B_2K_2') < 1$
$\rho(A - B_1(\frac{K_1+K_1'}{2}) - B_2(\frac{K_2+K_2'}{2})) > 1$

Gradient Expression

- Value matrix $P^y_{K_1,K_2}$ for stabilizing (K_1, K_2).

$$P^y_{K_1,K_2} = \sum_{t=0}^{\infty} \gamma^t ((A - B_1K_1 + B_2K_2)^t)^T (Q + K_1^T R_1 K_1 - K_2^T R_2 K_2)(A - B_1K_1 + B_2K_2)^t$$

We can rewrite the cost function [Carmona et al., 2020]

$$C_y(K_1, K_2) = \mathbb{E}_{y_0}\left[y_0^T P_{K_1,K_2} y_0\right] + \alpha^y_{K_1,K_2}$$
$$\alpha^y_{K_1,K_2} = \frac{\gamma}{1-\gamma}\mathbb{E}\left[(\epsilon_1^1)^T P^y_{K_1,K_2} \epsilon_1^1\right]$$

- Covariance matrix Σ_{K_1,K_2^y} for stabilizing (K_1, K_2).

$$\Sigma^y_{K_1,K_2} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t y_t y_t^T\right] = \sum_{t=0}^{\infty} \gamma^t (A - B_1K_1 + B_2K_2)^t (\Sigma_0^y + \frac{\gamma}{1-\gamma}\Sigma^1)((A - B_1K_1 + B_2K_2)^t)^T$$

where $\Sigma_0^y := \mathbb{E}_{y_0}[y_0 y_0^T]$. $\Sigma^1 := \mathbb{E}[(\epsilon_1^1)(\epsilon_1^1)^T]$

Gradient Expressions [Carmona et al., 2020]

Explicit Gradient Expression

By solving two Lyapunov equations

$$P^y_{K_1,K_2} = Q + K_1^T R_1 K_1 - K_2^T R_2 K_2 + \gamma(A - B_1K_1 + B_2K_2)^T P^y_{K_1,K_2}(A - B_1K_1 + B_2K_2)$$
$$\Sigma^y_{K_1,K_2} = \Sigma^y_0 + \frac{\gamma}{1-\gamma}\Sigma^1 + \gamma(A - B_1K_1 + B_2K_2)\Sigma^y_{K_1,K_2}(A - B_1K_1 + B_2K_2)^T$$

we have explicit expressions for the gradient of the utility function

$$\nabla_{K_j} C(K_1, K_2, L_1, L_2) = 2E^{y,j}_{K_1,K_2}\Sigma^y_{K_1,K_2}$$

where $E^{y,j}_{K_1,K_2} = poly(A, B_1, B_2, P^y_{K_1,K_2}, K_1, K_2)$.

Outline

1. Introduction
2. Model-based Nested Policy Gradient Algorithm
3. Model-free Nested Policy Gradient Algorithm

Model-based Nested Policy Gradient Algorithm

- Inner-loop: A new single-agent LQR problem. We assume access to

$$K_1(K_2) := \arg\min_{K_1} C_y(K_1, K_2) \quad or \quad K_2(K_1) := \arg\max_{K_2} C_y(K_1, K_2)$$

- Outer-loop:
  - Gradient descent (GD) [Zhang et al., 2019]

$$K_{2,t+1} = \mathbb{P}_{\Omega}^{GD}[K_{2,t} + \eta \nabla_{K_2} C_y(K_1(K_2), K_2)]$$

  - Natural gradient descent (NGD) [Bu et al., 2019]

$$\begin{aligned} & K_{2,t+1} = K_{2,t} + \eta \nabla_{K_2} C_y(K_1(K_2), K_2)(\Sigma_{K_1,K_2}^y)^{-1} \\ or \quad & K_{1,t+1} = K_{1,t} - \eta \nabla_{K_1} C_y(K_1, K_2(K_1))(\Sigma_{K_1,K_2}^y)^{-1} \end{aligned}$$

  - Quasi-Newton's Method [Bu et al., 2019] and Gauss-Newton Method [Zhang et al., 2019].

Assumptions for Model-based Nested Gradient Methods

Assumptions [Zhang et al., 2019, Bu et al., 2019]

A1. Existence and Uniqueness of NE: $\exists$ a minimal solution $P^* \succ 0$ to the generalized algebraic Riccati equation:

$$P^* = \gamma A^T P^* A + Q - \begin{bmatrix} \gamma A^T P^* B_1 & -\gamma A^T P^* B_2 \end{bmatrix} \begin{bmatrix} R_1 + \gamma B_1^T P^* B_1 & -\gamma B_1^T P^* B_2 \\ -\gamma B_2^T P^* B_1 & -R^v + \gamma B_2^T P^* B_2 \end{bmatrix}^{-1} \begin{bmatrix} \gamma B_1^T P^* A \\ -\gamma B_2^T P^* A \end{bmatrix}$$

where L^* satisfies $Q - (L^*)^T R^v L^* > 0$

A2. Stationary point$\rightarrow$ saddle point: $\Sigma_0 + \frac{\gamma}{1-\gamma}\Sigma^1 \succ 0$ and

$$R_1 + \gamma B_1^T P^* B_1 \succ 0, \quad R_2 - \gamma B_2^T X_* B_2 + \gamma^2 B_2^T P^* B_1 (R_1 + \gamma B_1^T P^* B_1)^{-1} B_1 P^* B_2 \succ 0$$

$$-R_2 + \gamma B_2^T P^* B_2 \prec 0, \quad R_1 + \gamma B_1^T P^* B_1 - \gamma^2 B_1^T P^* B_2 (-R_2 + \gamma B_2^T P^* B_2)^{-1} B_2^T P^* B_1 \succ 0$$

Convergence Rate for Model-based Nested Gradient Methods

Theorem

With gradient descent or natural gradient descent updates, (K_t, L_t) will converge to the Nash equilibrium sublinearly globally with $\frac{1}{N}\sum_{i=1}^{N} \|\nabla_{K_j} C(K_1, K_2, L_1, L_2)\|^2 \sim \mathcal{O}(\frac{1}{N})$, $\frac{1}{N}\sum_{i=1}^{N} \|\nabla_{L_j} C(K_1, K_2, L_1, L_2)\|^2 \sim \mathcal{O}(\frac{1}{N})$ (or gradient mappings) and linearly locally (when near the NE) with $C(K_{1,N}, K_{2,N}, L_{1,N}, L_{2,N}) - C(K_1^*, K_2^*, L_1^*, L_2^*) \sim \mathcal{O}(c_0^N)$

Key insights

- Cost difference lemma

$$C_y(K_1', K_2', y_0 = y) - C_y(K_1, K_2, y_0 = y) = \sum_{t \geq 0} A_{K_1,K_2}(y_t', K_1' y_t', K_2' y_t')$$

And we observe that noise terms are cancelled.

$$\begin{aligned} A_{K_1,K_2}(y, K_1'y, K_2'y) &= y^T[Q + (K_1')^T R_1 K_1' - (K_2')^T R_2 (K_2')]y \\ &+\gamma y^T(A - B_1K_1' + B_2K_2')^T P_{K_1,K_2}^y (A - B_1K_1' + B_2K_2')y + \frac{\gamma}{1-\gamma}\mathbb{E}[(\epsilon_1^1)^T P_{K_1,K_2}^y \epsilon_1^1] \\ &- y^T P_{K_1,K_2}^y y - \frac{\gamma}{1-\gamma}\mathbb{E}[(\epsilon_1^1)^T P_{K_1,K_2}^y \epsilon_1^1] \end{aligned}$$

Outline

1. Introduction
2. Model-based Nested Policy Gradient Algorithm
3. Model-free Nested Policy Gradient Algorithm

Gradient Estimation

- Only access to $C(K_1, K_2, L_1, L_2)$ instead of C_y, C_z

$$\nabla_{K_1} C_\tau^y(K_1, K_2) = \frac{d}{\tau^2}\mathbb{E}_{V_1}\left[C_y(K_1+V_1, K_2)V_1\right] = \frac{d}{\tau^2}\mathbb{E}_{V_1,U_1}\left[C(K_1+V_1, K_2, L_1+U_1, L_2)V_1\right]$$

where $V_1, U_1 \sim \mu_{\mathbb{S}_\tau} := Unif(\mathbb{S}_\tau = \partial\mathbb{B}_\tau)$.

Algorithm 1 Model-free MKV-Based Gradient Estimation

Input: Parameter (K_1, L_1); number of perturbations M; length T; radius τ
Output: A biased estimator for the gradient $(\nabla_{K_1} C(K_1, K_2, L_1, L_2), \nabla_{K_2} C(K_1, K_2, L_1, L_2))$
for $i = 1, 2, \cdots, M$ **do**
Sample v_{1i}, v_{2i} i.i.d. $\sim \mu_{\mathbb{S}_\tau}$
Set $(K^i, L^i) = (K_1 + v_{1i}, L_1 + v_{2i})$
Sample $\tilde{C}^i = \sum_{t=0}^{T-1} c_t^i$ using $\mathcal{S}_{MKV}^T(K^i, L^i, K_2, L_2)$
end for
Set $\tilde{\nabla}_{K_1} C(K_1, K_2, L_1, L_2) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^M \tilde{C}^i v_{1i}$, and $\tilde{\nabla}_{L_1} C(K_1, K_2, L_1, L_2) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^M \tilde{C}^i v_{2i}$

Convergence Results for Model-free Nested Algorithms

Consider model-free nested GD and Natural GD

$$K_{2,t+1} = \mathbb{P}_{\Omega}^{GD}\left[K_{2,t} + \eta\tilde{\nabla}_{K_2}C(\tilde{K}_1(K_{2,t}), K_{2,t})\right]$$

$$K_{2,t+1} = \mathbb{P}_{\Omega}^{NG}\left[K_{2,t} + \eta\tilde{\nabla}_{K_2}C(\tilde{K}_1(K_{2,t}), K_{2,t})\tilde{\Sigma}^{-1}_{\tilde{K}_1(K_{2,t},K_{2,t})}\right]$$

Theorem

Under the same assumptions, for any $\varepsilon > 0$, if length T, perturbation times M is large enough and the perturbation radius τ small enough. With iteration number $N \sim \mathcal{O}(\frac{1}{\varepsilon})$, we have

$$\frac{1}{N}\sum_{t=0}^{N-1}\frac{\|\mathbb{P}_{\Omega}^{GD}\left[K_{2,t} + \eta\tilde{\nabla}_{K_2}C(\tilde{K}_1(K_{2,t}), K_{2,t})\right] - K_{2,t}\|^2}{\eta} \leq \frac{C(y_0)}{N}$$

And M, T, τ have at most polynomial growth or decrease in $\|K_{1,t}\|, \|\tilde{K}_{2,t}(K_{1,t})\|, \|\tilde{L}_{1,t}(L_{2,t})\|, \|L_{2,t}\|$, and $C(\tilde{K}_1(K_{2,t}), K_{2,t}, \tilde{L}_1(L_{2,t}), L_{2,t})$.

- Important clarification: the bounds are well defined since $(\tilde{K}_1(K_{2,t}), K_{2,t}, \tilde{L}_1(L_{2,t}), L_{2,t})$, $C(\tilde{K}_1(K_{2,t}), K_{2,t}, \tilde{L}_1(L_{2,t}), L_{2,t})$ are indeed bounded.

Summary

	Convergence Guarantee	Algorithms	Assumptions	Noises
LQR [Fazel et al., 2018]	Linear $(C(\theta) - C(\theta^*))$	GD Natural GD Gauss-Newton GD	$\mathbb{E}_{x_0 \sim \mathcal{D}} x_0 x_0^T \succ 0$ $Q, R \succ 0$ $C(\theta_0)$ finite $\|x_0\| \leq L$ a.s. $x_0 \sim \mathcal{D}$	$\times$
MF-LQR [**?**]	Linear $(C(\theta) - C(\theta^*))$	GD	$Q, Q + \bar{Q}, R, R + \bar{R} \succeq 0$ Bounded noise variances	$\checkmark$
Zero-Sum LQR [Zhang et al., 2019]	Globally sublinear $(\frac{1}{N} \sum_{i=1}^{N} \|G_i\|^2)$ Locally linear $(C(\theta) - C(\theta^*))$	Projected Nested GD Projected Natural Nested GD Projected Gauss-Newton GD	GARE solvable NE $\exists!$	$\times$
Zero-Sum MF-LQR [Carmona et al., 2020]	Sublinearly & Linearly ?	Nested GD Alternating GD, GDA	Stabilizing params (Finite costs)	$\checkmark$

Take-home Message

- We can decouple the zero-sum mean-field LQR problem into two zero-sum LQR problem.
- $\exists$ model-based & model-free nested policy gradient algorithms achieve sublinear/linear convergence for quadratic zero-sum mean-field games.
- For single-loop algorithms:
  - $\{(K_1, K_2) | \rho(A - B_1 K_1 + B_2 K_2) < 1\}$ is open, connected.
  - ! But tricky to find a “nice” trajectory: satisfies conditions such as 2-sided PL condition.

2-Sided PL Condition

$$\max_y f(x,y) - f(x,y)$$
$$\leq \frac{1}{4}\|\Sigma_{K_1,K_2(K_1)}\|\|\Sigma^{y}_{K_1,K_2}{}^{-1}\|^2\|(R_2 - \gamma B_2^T P_{K_1,K_2} B_2)^{-1}\| Tr(\nabla_{K_2} C_y(K_1,K_2)^T \nabla_{K_2} C_y(K_1,K_2))$$

where we already know that $\|\Sigma^{y}_{K_1,K_2}{}^{-1}\|$ can be bounded by $\sigma_{\min}(\Sigma^0)$. On the other side

$$f(x,y) - \min_x f(x,y)$$
$$\leq \frac{1}{4}\|\Sigma_{K_1(K_2),K_2}\|\|\Sigma^{-1}_{K_1,K_2}\|^2\|(R_1 + \gamma B_1^T P_{K_1,K_2} B_1)^{-1}\| Tr(\nabla_{K_1} C_y(K_1,K_2)^T \nabla_{K_1} C_y(K_1,K_2))$$

Jingjing Bu, Lillian J. Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games, 2019. URL https://arxiv.org/abs/1911.04672.

René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: Convergence of policy gradient methods, 2019. URL https://arxiv.org/abs/1910.04295.

René Carmona, Kenza Hamidouche, Mathieu Laurière, and Zongjun Tan. Linear-quadratic zero-sum mean-field type games: Optimality conditions and policy optimization, 2020. URL https://arxiv.org/abs/2009.00578.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator, 2018. URL https://arxiv.org/abs/1801.05039.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games, 2019. URL https://arxiv.org/abs/1906.00729.