SCIENCE CHINA
Information Sciences

April 2024, Vol. 67, Iss. 4, 140203:1–140203:2
https://doi.org/10.1007/s11432-023-3962-0

• LETTER •
Special Topic: Analysis and Control of Stochastic Systems

# An online value iteration method for linear-quadratic mean field social control with unknown dynamics

Bing-Chang WANG, Shumei LI* & Ying CAO

*School of Control Science and Engineering, Shandong University, Jinan* 250000*, China*

Received 1 November 2023/Revised 2 January 2024/Accepted 29 January 2024/Published online 27 March 2024

Mean field (MF) models have been widely applied to economics, control theory, and other fields. Its prominent feature is that the individual influence on the overall population is negligible and the impact of the entire system to the single agent is significant and cannot be ignored. As a classical problem in control theory, linear quadratic (LQ) control for MF models has been widely investigated (e.g., [1, 2]). In addition to noncooperative cases, MF social optimum problems have also drawn much attention. All agents cooperate to optimize the cost function which is defined as the sum of the individual costs, which is a kind of team decision-making problem. In recent years, some progress has been made in the study of LQ MF social optimization. For instance, uniform stabilization and social optimality are analyzed by decoupling forward-backward stochastic differential equations [3].

For a standard LQ control problem, we can obtain the optimal control laws and the corresponding cost function depending on the solution to Riccati equations. However, when the system knowledge is not completely known, it is not feasible to attain the optimal control laws. An online value iteration (VI) algorithm is presented to tackle the optimal control problem and further a robust VI algorithm is developed for LQ systems with disturbances [4]. For LQ MF games and control, there has been prosperous interest on both discrete-time and continuous-time settings. Ref. [5] computed a set of decentralized strategies for LQ games by using a model-free policy iteration (PI) approach. Ref. [6] investigated a PI reinforcement learning method to solve MF LQ problems over an infinite horizon.

This study develops an online VI algorithm for MF LQ social control with ergodic cost functions. By employing the technique of completing squares, we obtain the social optimal control law depending on two algebraic Riccati equations (AREs). Firstly, we introduce an offline VI algorithm, which requires all the parameters of the system. Then we design two online model-free learning VI algorithms, where only a system trajectory is required. After iterating the required vector sequences using the aboved robust VI learning algorithms, we can obtain the solutions of two AREs, and then obtain the optimal control law. Meanwhile, by employing the converse Lyapunov theorem, we can prove the convergence of the online algorithm.

Compared with the existing studies, this work is characterized by the following features:

(1) Instead of the expectation-type cost function in [5,6], this study considers the MF model with stochastic ergodic costs, which has a more practical and physicial meaning. By exploiting the structure of the social cost, we propose a model-free VI online learning algorithm.

(2) For the proposed algorithm, only a system trajectory is required, and hence it is easier to implement than the algorithm under the expectation-type cost function.

*Problem formulation.* Consider a stochastic linear system with $N$ agents. Agent $i, 1 \leqslant i \leqslant N$ satisfies the stochastic differential equation

$$\mathrm{d}x_i(t) = \big(Ax_i(t) + Bu_i(t) + Gx^{(N)}(t)\big)\mathrm{d}t + D\mathrm{d}W_i(t), \quad (1)$$

where $x_i(t) \in \mathbb{R}^n$ and $u_i(t) \in \mathbb{R}^m$ are the state and control input of the $i$th agent, respectively. $x^{(N)}(t) \triangleq \frac{1}{N}\sum_{i=1}^N x_i(t)$ is the mean field term. $\{W_i(t), 1 \leqslant i \leqslant N\}$ is a sequence of independent $d$-dimensional Brownian motions. $D \in \mathbb{R}^{n\times d}$ is the noise intensity constant matrix.

The cost function of agent $i$ has the ergodic form

$$J_i(u) = \limsup_{T\to\infty} \frac{1}{T}\int_0^T \Big\{\|x_i(t) - \Gamma x^{(N)}(t)\|_Q^2 + \|u_i(t)\|_R^2\Big\}\mathrm{d}t, \quad (2)$$

where $Q = Q^{\mathrm{T}} \geqslant 0, R = R^{\mathrm{T}} > 0$. The social cost for the system (1) and (2) is defined as $J_{\mathrm{soc}}^{(N)}(u) = \sum_{i=1}^N J_i(u)$. The admissible control set is given by

$$\begin{aligned}\mathcal{U}_{ad} = \Big\{&(u_1, \ldots, u_N) | u_i \in \sigma(x_i(s), x^{(N)}(s), s \leqslant t),\\ &1 \leqslant i \leqslant N,\ \|x_i(T)\| = o(\sqrt{T}),\\ &\int_0^T \|x_i(t)\|^2\mathrm{d}t = O(T) \text{ a.s., } T \to \infty\Big\}.\end{aligned}$$

The purpose of the study is to design an online model-free algorithm to solve the following social control problem.

(**P**): minimize $J_{\mathrm{soc}}^{(N)}(u)$ over $u \in \mathcal{U}_d$.

We assume

(**A1**) The system $(A, B)$ is stabilizable, and the system $(A + G, B)$ is stabilizable.

(**A2**) The system $(A, \sqrt{Q})$ is observable, and the system $(A + G, \sqrt{Q}(I - \Gamma))$ is observable.

* Corresponding author (email: shumeili@mail.sdu.edu.cn)

 

Define the following AREs:

$$A^{\mathrm{T}}P + PA - PBR^{-1}B^{\mathrm{T}}P + Q = 0, \tag{3}$$

$$(A+G)^{\mathrm{T}}\Pi + \Pi(A+G) - \Pi BR^{-1}B^{\mathrm{T}}\Pi + Q - \Xi = 0, \tag{4}$$

where $\Xi \triangleq \Gamma^{\mathrm{T}}Q + Q\Gamma - \Gamma^{\mathrm{T}}Q\Gamma$.

First, we propose a model-based optimal control law.

**Proposition 1.** Assume (**A1**) and (**A2**) hold. Then for Problem (**P**), the optimal control law is given by

$$\begin{aligned}&\hat{u}_i(t) = -R^{-1}B^{\mathrm{T}}[Px_i(t) + (\Pi - P)x^{(N)}(t)],\\ &t \geqslant 0,\ i = 1, \ldots, N,\end{aligned} \tag{5}$$

where $P$ and $\Pi$ satisfy (3) and (4), respectively.

*Proof.* See Appendix A.

As nonlinear matrix equations, the AREs (3) and (4) are difficult to obtain directly. Due to the appearance of random disturbance, we have a robust ADP algorithm (i.e., Algorithm B1 in Appendix B) [4]. This algorithm provides robust numerical solutions to AREs (3) and (4), but has the limitation that the system dynamics is completely known, otherwise it cannot be implemented further.

Note the integrand of the cost function (2) has the decomposition form $r = r_1 + r_2$, where $r_1(z) = \|\tilde{x}_i\|_Q^2 + \|\tilde{u}_i\|_R^2$, $r_2(y) = \|x^{(N)}\|_{Q-\Xi}^2 + \|u^{(N)}\|_R^2$. The cost function is called separable, if $r_1$ and $r_2$ are known, respectively; otherwise inseparable. Now we derive a model-free algorithm to solve the separable social control problem.

Let $z^i = [\tilde{x}_i^{\mathrm{T}}, \tilde{u}_i^{\mathrm{T}}, 1]^{\mathrm{T}}$, $y = [(x^{(N)})^{\mathrm{T}}, (u^{(N)})^{\mathrm{T}}, 1]^{\mathrm{T}}$, $\phi(y) = [y_1^2, 2y_1y_2, \ldots, 2y_1y_{n+m+1}, y_2^2, \ldots, y_{n+m+1}^2]^{\mathrm{T}}$, $\psi^i(z^i) = [(z_1^i)^2, 2z_1^iz_2^i, \ldots, 2z_1^iz_{n+m+1}^i, (z_2^i)^2, \ldots, (z_{n+m+1}^i)^2]^{\mathrm{T}}$.

(**A3**) There exist $t_0 > 0$, $\beta_0 > 0$, $\bar{\beta} > 0$, such that for all $1 \leqslant i \leqslant N$, $t > t_0$, the inequalities

$$\frac{1}{t}\int_0^t \psi^i(\psi^i)^{\mathrm{T}}\mathrm{d}s \geqslant \beta_0 I,$$

$$\frac{1}{t}\int_0^t \phi\phi^{\mathrm{T}}\mathrm{d}s \geqslant \bar{\beta} I$$

hold with probability 1.

Let

$$\hat{\theta}^i(P, t_k) = \left(\int_0^{t_k} \psi^i(\psi^i)^{\mathrm{T}}\mathrm{d}s\right)^{-1}\left(\int_0^{t_k} \psi^i \mathrm{d}\left(\tilde{x}_i^{\mathrm{T}}P_k\tilde{x}_i\right) + \int_0^{t_k} \psi^i r_1 \mathrm{d}s\right),$$

$$\hat{\alpha}(\Pi, t_k) = \left(\int_0^{t_k} \phi\phi^{\mathrm{T}}\mathrm{d}s\right)^{-1}\left(\int_0^{t_k} \phi \mathrm{d}\left((x^{(N)})^{\mathrm{T}}\Pi_k x^{(N)}\right) + \int_0^{t_k} \phi r_2 \mathrm{d}s\right),$$

$$\mathcal{T}(\theta) = A^{\mathrm{T}}P + PA - PBR^{-1}B^{\mathrm{T}}P + Q,$$

$$\Lambda(\alpha) = (A+G)^{\mathrm{T}}\Pi + \Pi(A+G) - \Pi BR^{-1}B^{\mathrm{T}}\Pi + Q - \Xi.$$

Then we have Algorithm 1, and see more details involved in Appendix B.

**Algorithm 1** An online robust learning algorithm

1. Initialize $P_0^i = (P_0^i)^{\mathrm{T}} \geqslant 0$, $\Pi_0 = \Pi_0^{\mathrm{T}} \geqslant 0$ $k, q \leftarrow 0$.
2. **Loop**
   $\hat{\theta}_k^i \leftarrow (\int_0^{t_k} \psi^i(\psi^i)^{\mathrm{T}}\mathrm{d}s)^{-1}(\int_0^{t_k} \psi^i \mathrm{d}(\tilde{x}_i^{\mathrm{T}}P_k^i\tilde{x}_i) + \int_0^{t_k} \psi^i r_1 \mathrm{d}s)$,
   $\hat{\alpha}_k \leftarrow (\int_0^{t_k} \phi\phi^{\mathrm{T}}\mathrm{d}s)^{-1}(\int_0^{t_k} \phi \mathrm{d}((x^{(N)})^{\mathrm{T}}\Pi_k x^{(N)}) + \int_0^{t_k} \phi r_2 \mathrm{d}s)$,
   $P_{k+1/2}^i \leftarrow P_k^i + h_k\mathcal{T}(\hat{\theta}_k^i)$,
   $\Pi_{k+1/2} \leftarrow \Pi_k + h_k\Lambda(\hat{\alpha}_k)$,
3. **if** $P_{k+1/2}^i > 0$ and $|P_{k+1/2}^i - P_k^i|/h_k < \bar{\varepsilon}$ **then**
   **return** $P_k^i$ as an approximation to $P^*$,
   **else if** $|P_{k+1/2}^i| > q$ or $P_{k+1/2}^i \leqslant 0$,
   **then** $P_{k+1}^i \leftarrow P_0^i$, $q \leftarrow q+1$,
   **else** $P_{k+1}^i \leftarrow P_{k+1/2}^i$, $k \leftarrow k+1$,
4. **if** $\Pi_{k+1/2} > 0$ and $|\Pi_{k+1/2} - \Pi_k|/h_k < \bar{\varepsilon}$ **then**
   **return** $\Pi_k$ as an approximation to $\Pi^*$,
   **else if** $|\Pi_{k+1/2}| > q$ or $\Pi_{k+1/2} \leqslant 0$,
   **then** $\Pi_{k+1} \leftarrow \Pi_0$, $q \leftarrow q+1$,
   **else** $\Pi_{k+1} \leftarrow \Pi_{k+1/2}$, $k \leftarrow k+1$.

**Remark 1.** The studies in [5, 6] considered mean field LQ games and control by using model-free PI algorithms. In contrast, we propose a model-free VI online learning algorithm to solve mean field LQ social control with ergodic costs. In the algorithm described above, only a system trajectory is required, and hence it is easier to implement than the system with expectation-type cost function.

**Theorem 1.** Under (**A1**)–(**A3**), for $1 \leqslant i \leqslant N$, we have $\lim_{k\to\infty} P_k^i = P^*$ and $\lim_{k\to\infty} \Pi_k = \Pi^*$ with Probability 1, where $\{P_k^i\}_{k=0}^{\infty}$ and $\{\Pi_k\}_{k=0}^{\infty}$ are obtained from Algorithm 1.

*Proof.* See Appendix C.

The convergence of Algorithm 1 is given above for the separable case. Also, we can tackle the case where $r$ is inseparable (see Appendix D). A numerical example is given in Appendix E.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant No. 62122043).

**Supporting information** Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

**References**

1 Huang M, Caines P E, Malhame R P. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized $\varepsilon$-nash equilibria. IEEE Trans Autom Control, 2007, 52: 1560–1571
2 Li T, Zhang J F. Asymptotically optimal decentralized control for large population stochastic multiagent systems. IEEE Trans Autom Control, 2008, 53: 1643–1660
3 Wang B C, Zhang H, Zhang J F. Mean field linear-quadratic control: uniform stabilization and social optimality. Automatica, 2020, 121: 109088
4 Bian T, Jiang Z P. Continuous-time robust dynamic programming. SIAM J Control Optim, 2019, 57: 4150–4174
5 Xu Z, Shen T, Huang M. Model-free policy iteration approach to NCE-based strategy design for linear quadratic Gaussian games. Automatica, 2023, 155: 111162
6 Li N, Li X, Xu Z Q. Policy iteration reinforcement learning method for continuous-time mean-field linear-quadratic optimal problem. 2023. ArXiv:2305.00424