Decentralized Policy Gradient Method for Mean-Field Linear Quadratic Regulator with Global Convergence

Lewis Liu LEWIS3Y@OUTLOOK.COM
Mila & DIRO, Université de Montréal, Montreal, QC, Canada

Zhuoran Yang ZY6@PRINCETON.EDU
Princeton University, Princeton, NJ, USA

Yuchen Lu LUYUCHEN.PAUL@GMAIL.COM
Mila & DIRO, Université de Montréal, Montreal, QC, Canada

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Northwestern University, Evanston, IL, USA

Abstract

The scalability of multi-agent reinforcement learning methods to a large number of population is drawing more and more attention in both practice and theory. We consider the basic yet important model, *i.e.*, linear quadratic regulator (LQR), in a mean-field approximation scheme against the curse of the action space dimensionality and the exponential growth of agent interactions. Several methods proposed in mean-field setting require a centralized controller, which is unrealistic in practice. In this paper, we present the first decentralized policy gradient method (MF-DPGM) for mean-field multi-agent reinforcement learning, where a large team of exchangeable agents communicate via a connected network. After a linear transformation of states and policies, we update the new local and mean-field policies by a decentralized gradient primal-dual algorithm respectively in a decoupling way, in order to achieve a global policy consensus. We also give a rigorous proof of the global convergence rate of MF-DPGM by studying the geometry of the problem and estimating one-step progress under decentralized scheme. In addition, extensive experiments are conducted to support our theoretical findings.

Keywords: Multi-Agent Reinforcement Learning, Decentralized Learning, Mean-Field Approximation, Global Convergence

1. Introduction

Recent years have witnessed a promising resurgence of multi-agent reinforcement learning (MARL) in data-driven and large population environments. Motivating applications span over multi-robotics systems (Corke et al., 2005), autonomous driving (Lo, 2012; Shalev-Shwartz et al., 2016), and sensor networks (Rabbat and Nowak, 2004; Cortes et al., 2004). MARL involves a set of agents learning to make decisions that minimize their accumulative cost by iterative interactions with a shared environment (Shoham et al., 2003, 2007; Bu et al., 2008). As a result, a fundamental difficulty in MARL is that changes in the policy of one agent will affect that of the others, and vice versa (Matignon et al., 2012). What is worse, large modern multi-agent systems result in an exponential growth of the capacity of the joint action space with the number of agents. Hence the classical MARL methods (Bowling and Veloso, 2002; Lipsa and Martins, 2011; Lessard and Lall, 2015) via either equilibrium-solving or few controllers stagger in large-scale applications. Additionally, although a central controller receiving costs and determining actions reduces MARL to

a classical MDP which can be solved by existing single-agent RL approaches, the central controller is usually costly to install and the communication overhead degrades the scalability and robustness.
Motivation of mean-field settings. In this paper, we consider homogeneous large-scale MARL systems with symmetry, where each agent has the same reward function and state transition rule. To address the complicated correlations in multi-agent systems, Foerster et al. (2018) and Panait and Luke (2005) consider accounting for the extra information from conjecturing the policies of other agents, while Lee et al. (2018) and Zhang et al. (2018a) study the decentralized actor-critic algorithm. On the other hand, the mean field approximation (Lasry and Lions, 2007; Huang et al., 2006) serves as an effective alternative to modelling strategic interactions for large populations with symmetry. To characterize the mean-field effect in finite-agent systems, Arabneydi and Mahajan (2016) shows that any exchangeable system, where exchanging any two agents does not affect the dynamics or costs, is equivalent to one where the dynamics and costs are coupled across agents through the mean field (empirical mean state). More importantly, compared to other formulations, the mean-field setting greatly alleviates the curse of action space dimensionality by a symmetric global optimal policy and a cost function for all agents while decoupling complex correlations in large interactive systems.

However, neither decentralized algorithms nor accompanied guarantees are studied for the well-behaved mean-field setting. To fill this void, we study the decentralized exchangeable multi-agent systems in the collaborative setting where each agent seeks the optimal policy that minimizes the accumulative global cost over all the agents, via neighborhood communications by a connected network. We propose the first decentralized policy learning scheme with smaller exploration space, less communication, and more robustness. Moreover, we study the non-convex problem geometry by several almost continuity results, which are combined with one-step iterate progress to establish a sublinear global convergence rate for the simple yet fundamental setting LQR. Our contributions are concluded as follows: (1) We formulate the policy gradients for MARL under the mean-field setting. (2) We proposed the first decentralized algorithm (MF-DPGM) to effectively learn the optimal policy for mean-field MARL. (3) We present a novel global convergence guarantee for MF-DPGM under mild assumptions and simulation results to justiify the performance of our algorithm.

Related Work. There has been a line of work on solving normal MARL problems. Based on the seminal work for the framework of Markov games Littman (1994), follow-up works, such as Lauer and Riedmiller (2000); Littman (2001); Hu and Wellman (2003), studied both collaborative and competitive relationships among agents. Recently, MARL with a large population Sandholm (2010); Lo (2012) becomes increasingly popular, such as urban transportation Shalev-Shwartz et al. (2016); Lo (2012), social dilemmas Leibo et al. (2017); Hughes et al. (2018), multi-robotics systems Corke et al. (2005), and power grids Callaway and Hiskens (2010), wherein the curse of dimensionality for learning and control Matignon et al. (2012) arises. Our work is in the line of collaborative settings, where a central controller can help solve MARL by existing single-agent algorithms Bradtke et al. (1994); Tu and Recht (2017); Malik et al. (2018). Nonetheless, due to high cost to set central controllers in large-scale applications Kulkarni and Venayagamoorthy (2010), a series of decentralized methods Wai et al. (2018); Zhang et al. (2018a); Lee et al. (2018) are developed following Zhang et al. (2018b) to learn optimal policies with local rewards and actions.

Mean-field approximation of the system stems from Stanley (1971), which is generalized to multi-agent scenarios by Huang et al. (2003); Lasry and Lions (2006, 2007); Huang et al. (2006). Our setting is also closely related to mean-filed control for exchangeable agents Madjidian and Mirkin (2014); Arabneydi and Mahajan (2016) but from a model-free reinforcement learning perspective. Although another independent simultanuous work (Carmona et al., 2019) investigate MARL for a mean-field case, only the centralized algorithm with infinite players is considered and directly reduced to single-agent LQR when doing variable transformation.

Notations. In this section, we introduce some frequently used notations through the paper, while others will be defined as they are needed. We denote by $[n]$ the set of integers $\{1, 2, ..., n\}$, by $\bar{\mathcal{N}}$ all of the nonnegative integers. We use $\langle \cdot, \cdot \rangle$ to denote the inner-product for vectors, matrices, tensors, and block-wise cases according to the context. For a matrix $X \in \mathbb{R}^{d\times d}$ and a vector $v \in \mathbb{R}^d$, we define $\|v\|_X^2 := v^\top X v$ as the squared norm of v under metric matrix X. When v is a matrix or tensor, $\|v\|$ refers to the norm of its vectorization. The *mode-n matrix product* of tensor $\mathbb{X} \in \mathbb{R}^{d_1\times\cdots\times d_p}$ and a matrix $X \in \mathbb{R}^{e\times d_n}$ is a tensor $X\mathbb{X}_{(n)} \in \mathbb{R}^{d_1\times\cdots\times d_{n-1}\times e\times\cdots\times d_p}$. Without specifying, $X\mathbb{X}$ denotes mode-1 matrix product. We denote by $\mathbf{1}$ and $\mathbf{0}$ the all-one and all-zero vector (resp. tensor) respectively by the context, by $\sigma_{\min}(X)$ the smallest eigenvalue of X.

2. Problem Formulation

We study discrete-time linear-quadratic MARL under mean-field settings with exchangeable finite n agents. The states of the system at time-step t is given by $\{x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(n)}\}$, where $x_t^{(i)} \in \mathbb{R}^d$ denotes the state vector of the i^{th} agent ($i \in [n]$). Then the system dynamics is described as follows,

$$x_{t+1}^{(i)} = Ax_t^{(i)} + Bu_t^{(i)} + \bar{A}\bar{x}_t + w_t^{(i)}, \quad \forall i \in [n], \tag{2.1}$$

where $u_t^{(i)} \in \mathbb{R}^m$ denotes the control (the action), a Gaussian noise $w_t^{(i)} \sim N(0, I_d)$ independent of each agent and time-step is added to the succeeding state, $\bar{x}_t = 1/n \sum_{i=1}^n x_t^{(i)}$ denotes the mean-field state of the system at time-step t. For such dynamics, we have a collective cost function of the distributed system at time-step t,

$$c_t = \sum_{i=1}^n {x_t^{(i)}}^\top Q x_t^{(i)} + {u_t^{(i)}}^\top R u_t^{(i)} + \bar{x}_t^\top \bar{Q} \bar{x}_t, \tag{2.2}$$

where we also define the cost for agent i as $c_t^{(i)} = {x_t^{(i)}}^\top Q x_t^{(i)} + {u_t^{(i)}}^\top R u_t^{(i)} + \bar{x}_t^\top \bar{Q}\bar{x}_t$ at time-step t. Furthermore, it is shown that optimal control for the i^{th} agent can be written as a linear combination of $x_t^{(i)}$ and $\bar{x}_t$, $u_t^{(i)} = Kx_t^{(i)} + L\bar{x}_t$, where the same matrices $\{K, L\}$ apply to all agents by the symmetry results in optimal control (Arabneydi and Mahajan, 2016). By plugging $u_t^{(i)}$ into (2.1), we can rewrite $x_{t+1}^{(i)}$ as $x_{t+1}^{(i)} = (A + BK)x_t^{(i)} + (\bar{A} + BL)\bar{x}_t + w_t^{(i)}$. Let $\Theta = (K; L) \in \mathbb{R}^{2\times m\times d}$, beginning with initial states $\{x_0^{(i)}\}_{i=1}^n$, our goal is to find the controls $\{u_t^{(i)}\}_{i=1}^n$ ($t \geq 0$) minimizing the long-term collective cost,

$$C(\Theta) = \mathbb{E}_{\mathbf{x}_0,\mathbf{w}}\left[\sum_{t=0}^\infty \gamma^t c_t\right] = \mathbb{E}_{\mathbf{x}_0,\mathbf{w}} \sum_{t=0}^\infty \gamma^t \sum_{i=1}^n \left({x_t^{(i)}}^\top Q x_t^{(i)} + {u_t^{(i)}}^\top R u_t^{(i)} + \bar{x}_t^\top \bar{Q}\bar{x}_t\right). \tag{2.3}$$

We denote by $\gamma \in (0, 1)$ the discounted factor in the infinite horizon. The expectation is taken *w.r.t.* all the initial states $\mathbf{x}_0 = \left({x_0^{(1)}}^\top, ..., {x_0^{(n)}}^\top\right)^\top$ and all independent noise terms $\mathbf{w} = \{w_t^{(i)}\}_{i\in[n],t\in\mathbb{N}}$. Therefore, our goal is to solve the following infinite horizon mean-field LQR problem,

$$\text{minimize } C(\Theta) \text{ s.t. } \quad x_{t+1}^{(i)} = Ax_t^{(i)} + Bu_t^{(i)} + \bar{A}\bar{x}_t + w_t^{(i)}, \quad x_0^{(i)} \sim \mathcal{D} \text{ for each } i \in [n]. \tag{2.4}$$

3. The Approach and Algorithm

To deal with the correlated states in the mean-field setting, we adopt the reparametrization trick in Section A to obtain the dynamics below,

$$u_t^{(i)} = Kx_t^{(i)} + L\bar{x}_t = M(x_t^{(i)} - \bar{x}_t) + N\bar{x}_t \triangleq My_t^{(i)} + N\bar{y}_t. \tag{3.1}$$

However, without a central controller the i^{th} agent possesses $M^{(i)}$ $(i \in [n])$, which is defined as the policy iterate of the i^{th} agent in decentralized scenarios, as its local policy before convergence and is restricted to communicating policies with neighbor agents over a network, although each agent has access to the global mean state. Similar problems emerge when updating $N^{(i)}$'s, which is the mean-field policy of each agent before finding the optimum. Hence, we resort to performing a decentralized optimization scheme with global consensus. As M and N are decoupled into two similar update processes, below we denote by $\widetilde{\Theta}^{(i)}$ either $M^{(i)}$ or $N^{(i)}$. We concatenate policy parameters together as a higher dimensional tensor, *i.e.*, $\widetilde{\mathbf{\Theta}} = [\widetilde{\Theta}^{(1)}; \widetilde{\Theta}^{(2)}; ...; \widetilde{\Theta}^{(n)}] \in \mathbb{R}^{n\times m\times d}$ to give a compact formulation of the decentralized optimization problem:

$$\min_{\widetilde{\mathbf{\Theta}}} \widetilde{C}(\widetilde{\mathbf{\Theta}}) := \frac{1}{n}\sum_{i=1}^{n} C(\widetilde{\Theta}^{(i)}), \quad \text{s.t.} \quad \widetilde{\Theta}^{(i)} = \widetilde{\Theta}^{(j)} \quad \text{for} \quad (i,j) \in \mathcal{E}, \tag{3.2}$$

where the communication network is considered as an *unweighted* and *undirected* graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Such a description gives a separable global objective function $\widetilde{C}(\widetilde{\mathbf{\Theta}})$ and linear constraints indicating the connectivity property of the communication network. In next section, we will introduce a more tractable formulation for algorithms by graph theory. Note that the summands in (3.2) are disjoint components of $\widetilde{\mathbf{\Theta}}$. Therefore, the gradient of the global objective should be expanded as: $\nabla \widetilde{C}(\widetilde{\mathbf{\Theta}}) = 1/n\big(\nabla C\big(\widetilde{\Theta}^{(1)}\big); ...; \nabla C\big(\widetilde{\Theta}^{(n)}\big)\big)$.

Communication Structure and Algerabic Representations. Below we capture the structure of $\mathcal{G}$ with $|\mathcal{V}| = n$ and $|\mathcal{E}| = e$ by some tools from spectral graph theory. Define the edge-associated (Ω_θ) and agent-associated (Γ_θ) parameters $\Omega_\theta = \mathrm{diag}(\sigma_1^\theta, \cdots, \sigma_e^\theta) \succ 0, \Gamma_\theta = \mathrm{diag}(\gamma_1^\theta, \cdots, \gamma_n^\theta) \succ 0$, where θ can be replaced by M or N to represent two sets of configurations since the policy parameters are decoupled into two optimization processes. Hereafter if we use symbols without specifying M or N, the statement will hold for both M and N, respectively. We will see how they serve as step-size for MF-DPGM in Section 3.1. We also use σ_{ij}, where $(i,j) \in \mathcal{E}$ is the k^{th} edge also denoted as $i \sim j$, as an alternative to σ_k. The quantity assigned to agent i is used for $M^{(i)}$ or $N^{(i)}$, or both by the context. Denoting the graph Laplacian matrix and its normalized version Chung and Graham (1997) as $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ respectively, we define $\mathcal{L}_a := |\widetilde{\mathcal{L}}| \in \mathbb{R}^{e\times n}$ by taking element-wise absolute value on $\widetilde{\mathcal{L}}$. By definitions we can find the following relation: $D = 1/2(\widetilde{\mathcal{L}}^\top \widetilde{\mathcal{L}} + \mathcal{L}_a^\top \mathcal{L}_a)$, where $D = \mathrm{diag}(d_1, ..., d_n)$ with d_i as the degree of vertex (agent) i. For further proof, by calculation we define $\widetilde{\Omega} := \mathrm{diag}\left(\left\{\sum_{j:j\sim i} \sigma_{ij}^2\right\}_{j\in\mathcal{G}}\right) = 1/2(\widetilde{\mathcal{L}}^\top \Omega^2 \widetilde{\mathcal{L}} + \mathcal{L}_a \Omega^2 \mathcal{L}_a)$, $H := \mathcal{L}_a{}^\top \Omega^2 \mathcal{L} + \Gamma^2$. , and $\eta_i^2 = 2\sum_{j:j\sim i} \sigma_{ij}^2 + \gamma_i^2$. With these parameters, we can rewrite (3.2) in a more tractable form for numerical algorithms: $\min_{\widetilde{\mathbf{\Theta}}} \widetilde{C}(\widetilde{\mathbf{\Theta}}) := 1/n \sum_{i=1}^{n} C(\widetilde{\Theta}^{(i)})$, s.t. $\widetilde{\mathcal{L}}\widetilde{\mathbf{\Theta}}_{(1)} = \mathbf{0}$. The idea of MF-DPGM relies on this formulation from a primal-dual view.

3.1 The MF-DPGM Algorithm

The overall MF-DPGM can be split into following steps.

Initialization: Similar to Prox-MM Wright (1990), our method first initialize the policy parameter $\widetilde{\Theta}_{-1}^{(i)}$ $\big(M_{-1}^{(i)}; N_{-1}^{(i)}\big)$ of agent i with zeros at iteration -1, and then endow policies at step 0 with $\widetilde{\Theta}_{-1}^{(i)}$'s. Such a setup benefits further updates by controlling the constants of theoretical bounds in Section 4.

Policy running and evaluation: At the beginning of the k^{th} iteration, we sample n_p trajectories of each agent i according to its own current policy $\widetilde{\Theta}_k^{(i)}$, to estimate the action-value function by $1/n_p \sum_{p=1}^{n_p} \widehat{Q}_{i,p}^{\pi}(x_t^{(i)}, u_t^{(i)})$ for policy evaluation in model-free settings[1]. Furthermore, to retain

1. Also, we can share part of noise with term $u_t^{(i)}$ in 2.1 to make stochastic policies. To highlight our main idea, we focus on theoretical results with exact gradients. See Yu (1994) for a straightforward statistical error.

Algorithm 1: Mean-Field Decentralized Policy Gradient Method (MF-DPGM)

Data: Agent dynamics $n, A, B, \bar{A}$; Cost parameters $Q, R, \bar{Q}$; Initial state distribution $\mathcal{D}$.
Input: Network $\mathcal{G}$; Ω, Γ; Length of horizon T; Number of sample paths n_p; Estimation error ϵ.
Output: Estimation of optimal control policies $\widehat{\mathbf{\Theta}}$.
Initialization: $\widetilde{\mathbf{\Theta}}_{-1} = \mathbf{0}$; $\widetilde{\Theta}_0^{(i)} = \nabla C^{(i)}(\mathbf{0})\eta_i^{-2}/n, \forall i \in [n]$; Iteration $k \leftarrow 1$.
while $\varepsilon(k) > \epsilon$ **do**
  for *path p = 1 to* n_p **do**
    for *t = 1 to T* **do**
      $u_t^{(i)} = M_k^{(i)} y_t^{(i)} + N_k^{(i)} \bar{y}_t$;
      $x_{t+1}^{(i)} \leftarrow Ax_t^{(i)} + Bu_t^{(i)} + \bar{A}\bar{x}_t + w_t^{(i)}$, for all $i \in [n]$ in paralell;
      $\widehat{Q}_{i,p}^{\pi}(x_t^{(i)}, u_t^{(i)}) \leftarrow \sum_{s=t}^{T} \gamma^{s-t} c_s^{(i)}$;
    end
  end
  Compute $\widehat{\nabla} C(\widetilde{\Theta}_k^{(i)}) \leftarrow 1/n_p \sum_{p=1}^{n_p} \sum_{t=1}^{T} \widehat{Q}_{i,p}^{\pi}(x_t^{(i)}, u_t^{(i)}) \cdot \nabla \log \pi_{\widetilde{\Theta}^{(i)}}(u_t^{(i)}|x_t^{(i)})$, for all $i \in [n]$
  Communication and update: For all $i \in [n]$,

$$\widetilde{\Theta}_{k+1}^{(i)} \leftarrow \widetilde{\Theta}_k^{(i)} - \frac{1}{(\eta_i^{\theta})^2} \left(\frac{1}{n} \left(\nabla \widehat{C}(\widetilde{\Theta}_k^{(i)}) - \nabla \widehat{C}(\widetilde{\Theta}_{k-1}^{(i)}) \right) \right.$$
$$\left. - 2 \sum_{j:j\sim i} (\sigma_{ij}^{\theta})^2 \widetilde{\Theta}_k^{(j)} + (\gamma_i^{\theta})^2 \left(\widetilde{\Theta}_{k-1}^{(i)} - \widetilde{\Theta}_k^{(i)} \right) + \sum_{j:j\sim i} (\sigma_{ij}^{\theta})^2 \left(\widetilde{\Theta}_{k-1}^{(j)} + \widetilde{\Theta}_{k-1}^{(i)} \right) \right) \tag{3.3}$$

  $\widehat{\mathbf{\Theta}} \leftarrow \widetilde{\mathbf{\Theta}}_k$, $k \leftarrow k + 1$
end

accuracy for large populations, we adopt gradient estimator in REINFORCE Sutton et al. (2000) for the cost function instead of the biased zeroth order method Fazel et al. (2018). In practice, the deterministic policy $u_t^{(i)}$ can be realized by a Gaussian policy $\pi_{\Theta^{(i)}}$ with a nearly zero variance.
Communication and update: After running policies at the k^{th} iteration, each agent i collects policies from neighbors and its own gradient estimators of the cost function at the k^{th} and $(k-1)^{th}$ iterations, and combines them linearly by Ω_θ and Γ_θ as a decentralized version of policy gradient method Baxter and Bartlett (2001) to update the local policy $\widetilde{\Theta}_k^{(i)}$. As a result, we can view tunable parameters Ω_θ and Γ_θ as step-sizes in MF-DPGM. See (3.3) for a detailed update. A practical choice (Gemulla et al., 2011; Shi et al., 2015) is $\Omega_\theta^2 := \alpha^2 I$ as a multiple of identity,

A full description of MF-DPGM is shown in Algorithm 1. Due to decentralized nature, our convergence measure, $\varepsilon(t) = \min_{s\in[t]} \left| 1/n \sum_{i=1}^{n} C(\widetilde{\Theta}_s^{(i)}) - \widetilde{C}(\widetilde{\mathbf{\Theta}}^*) \right| + \|\Omega \widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_k\}_{(1)}\|^2$ for $t \in \bar{\mathcal{N}}$, not only takes cost error into account, but also consider consensus error to guarantee an identical optimal policy in Section 2. See Appendix B for a connection with the primal-dual paradigm.

4. Theoretical Results and Analysis

As many other dencentralized algorithms, our convergence results also rely on some smoothness property (Lipschitzness of the gradient) of the objective. However, we notice that the optimization landscape for each agent is not strictly smooth due to unstability of $A+BM$ (resp. $A+\bar{A}+BN$). At the boundary between stable and unstable policies, the cost function rapidly becomes infinity, which violates the traditional smoothness conditions. To address this issue, we regulate the parameters Ω and Γ for a small step-size to set the next iterate $\widetilde{\Theta}_{t+1}^{(i)}$ sufficiently close to the current one, so that: $\Sigma_{\widetilde{\Theta}_{t+1}^i} \approx \Sigma_{\widetilde{\Theta}_t^i} + \mathcal{O}\left(\|\widetilde{\Theta}_{t+1}^i - \widetilde{\Theta}_t^i\|\right)$. Using tools from Fazel et al. (2018), we show in Lemma F.2-E.1

the continuity property of the cost functions, state trajectories, and gradients of cost corresponding to policy $\mathbb{M}$ and $\mathbb{N}$ within an adaptive area for each agent. Accordingly, we have the following condition on edge and agent associated parameters, *i.e.* Ω and Γ to guarantee the development of our decentralized update.

Condition 4.1 . The parameters of MF-DPGM are chosen to satisfy for any $k \geq 1$ and $i \in [n]$,

$$\left(\Omega_\theta + \Gamma_\theta^2\right)/2 \succeq (2c+1)\Phi_\theta/n + 4\kappa\Phi_\theta\Gamma_\theta^{-2}\Phi_\theta/n^2,\ c = \max\{1, 6\kappa\},\ \Gamma_\theta^2 \succeq \Phi_\theta\Gamma_\theta^{-2}\Phi_\theta/n^2, \quad (4.1)$$

$$\gamma_i^2\mathcal{F}_\theta(\theta_k^{(i)}) - \frac{n\gamma_i^2 + \beta_i}{n}\mathcal{F}_\theta(\theta_{k-1}^{(i)}) \geq \sum_{j:j\sim i}\sigma_{ij}^2\left(\|\theta_k^{(i)}\| + \|\theta_{k-1}^{(i)}\| + \mathcal{F}_\theta(\theta_{k-1}^{(j)}) - 2\mathcal{F}_\theta(\theta_k^{(i)})\right), \quad (4.2)$$

where $\kappa = 1/\lambda_{\min}\left(\Omega F H^{-1} F^T \Omega\right)$, (4.2) admits two sets of conditions when replacing θ by M or N exclusively, along with $\mathcal{F}_M(X) = \min\{\xi\sigma_{\min}(Q)/[4(\|A+BX\|+1)\cdot\|B\|\cdot C(X)], \|X\|\}$ and $\mathcal{F}_N(X) = \min\{\bar{\xi}\sigma_{\min}(Q+\bar{Q})/[4(\|A+BX\|+1)\cdot\|B\|\cdot C(X)], \|X\|\}$, Similarly $\Phi_M = \mathrm{diag}(\beta_1^M, ..., \beta_n^M) \otimes I_{md} \in \mathbb{R}^{nmd\times nmd}$ and $\Phi_N = \mathrm{diag}(\beta_1^N, ..., \beta_n^N) \otimes I_{md}$ are almost-smoothness constants specified in Lemma F.2.

We can see from proof that κ encodes the network structure and c helps construct the potential function indicating one-step progress of MF-DPGM. By properly choosing Ω_θ and Γ_θ, we can meet the condition, so that potential function (E.1) decreases and boundary condition (F.9) of two adjacent iterates is met by dynamically adjusting as the adaptive area in (F.9), which is illustrated in F.3. Below we layout the global convergence result.

Theorem 4.2. Given Condition 4.1, then for time-step $t \geq 1$, MF-DPGM gives

$$\min_{s\in[t]}\left|\frac{1}{n}\sum_{i=1}^n C(\widetilde{\mathbf{\Theta}}_s^{(i)}) - \widetilde{C}(\widetilde{\mathbf{\Theta}}^*)\right| + \|\Omega\widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_s\}_{(1)}\|^2 \leq \underbrace{8\alpha_g\mathcal{C}\mathcal{C}'/t}_{\text{cost error bound}} + \underbrace{20\mathcal{C}'/t}_{\text{consensus error bound}}, \quad (4.3)$$

where $\alpha_g = \max\{\alpha_g^M, \alpha_g^N\}$ with $\alpha_g^M = \|\Sigma_{M^*}\|/[\sigma_{\min}(\Sigma_0^{(i)})^2\sigma_{\min}(R)]$ and $\alpha_g^N = \|\Sigma_{N^*}\|/[\sigma_{\min}(\Sigma_0)^2\sigma_{\min}(R)]$ is a problem related constant for gradient domination also appearing in Lemma F.3, $\widetilde{\mathbf{\Theta}}^*$ indicates the optimal policy parameters, and

$$\mathcal{C}' := \widetilde{C}(\widetilde{\mathbf{\Theta}}_0) - \widetilde{C}(\widetilde{\mathbf{\Theta}}^*) + 2\nabla C(\mathbf{0})^\top\Phi^{-1}\nabla C(\mathbf{0})/n, \quad \mathcal{C} := 4\sum_{(i,j):i\sim j}\sigma_{ij}^2 + \sum_{i=1}^n\gamma_i^2. \quad (4.4)$$

Proof. See Appendix F.5 for a detailed proof. □

The main upshot is to characterize in detail the first global sublinear convergence rate for the overall error including the cost and consensus components. More precisely, MF-DPGM not only drives the average cost of agents to the optimal cost over time, but also guarantees heading for the same optimal policy. An alternative way to display the rates is distributedly showing for each agent with slightly different constants where $\mathcal{C}'$ is decomposed for each. Moreover, the optimization of $\widetilde{\mathbf{\Theta}}$ can be split into development of local and mean-field policy respectively, which is frequently observed in theoretical illustration and proof, while the difference sits between the only one mean-field state associated with $N^{(i)}$'s and corresponding multiple states for $M^{(i)}$'s.

5. Conclusion

In this paper, focusing on a simple yet fundamental model LQR under the mean-field setting, we propose the first decentralized policy gradient method where each agent update the local policy by combining policies from neighborhood with its own policy. which is applicable to complex mean-field models. In addition, we quantify the non-convex problem geometry by several almost continuity results, which is combined with one-step progresses for our algorithm to establish a sublinear global convergence rate for LQR. Additional experiments justify our theoretical results and show a promising performance.

References

Jalal Arabneydi and Aditya Mahajan. Linear quadratic mean field teams: Optimal and approximately optimal decentralized solutions. *arXiv preprint arXiv:1609.00056*, 2016.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Sergio Bittanti, Alan J Laub, and Jan C Willems. *The Riccati Equation*. Springer Science & Business Media, 2012.

Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

Steven J Bradtke, B Erik Ydstie, and Andrew G Barto. Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, pages 3475–3479. IEEE, 1994.

Lucian Bu, Robert Babu, Bart De Schutter, et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Duncan S Callaway and Ian A Hiskens. Achieving controllability of electric loads. *Proceedings of the IEEE*, 99(1):184–199, 2010.

René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

Peter Corke, Ron Peterson, and Daniela Rus. Networked robots: Flying robot navigation using a sensor net. In *Robotics research. The eleventh international symposium*, pages 234–243. Springer, 2005.

Jorge Cortes, Sonia Martinez, Timur Karatas, and Francesco Bullo. Coverage control for mobile sensing networks. *IEEE Transactions on robotics and Automation*, 20(2):243–255, 2004.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

Rainer Gemulla, Erik Nijkamp, Peter J Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–77. ACM, 2011.

Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1): 337–364, 2016.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Minyi Huang, Peter E Caines, and Roland P Malhamé. Individual and mass behaviour in large population stochastic wireless power control problems: centralized and nash equilibrium solutions. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, volume 1, pages 98–103. IEEE, 2003.

Minyi Huang, Roland P Malhamé, Peter E Caines, et al. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, pages 3326–3336, 2018.

Dušan Jakovetić, José MF Moura, and Joao Xavier. Linear convergence rate of a class of distributed augmented lagrangian algorithms. *IEEE Transactions on Automatic Control*, 60(4):922–936, 2014.

Raghavendra V Kulkarni and Ganesh Kumar Venayagamoorthy. Particle swarm optimization in wireless-sensor networks: A brief survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(2):262–267, 2010.

Jean-Michel Lasry and Pierre-Louis Lions. Games à middle field. ii - finite horizon and control ô the optimal. *Math é Math Accounts*, 343(10):679–684, 2006.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2 (1):229–260, 2007.

Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

Donghwan Lee, Hyungjin Yoon, and Naira Hovakimyan. Primal-dual algorithm for distributed reinforcement learning: distributed gtd. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1967–1972. IEEE, 2018.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

Laurent Lessard and Sanjay Lall. Optimal control of two-player systems with output feedback. *IEEE Transactions on Automatic Control*, 60(8):2129–2144, 2015.

Gabriel M Lipsa and Nuno C Martins. Remote state estimation with communication costs for first-order lti systems. *IEEE Transactions on Automatic Control*, 56(9):2013–2025, 2011.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Michael L Littman. Value-function reinforcement learning in markov games. *Cognitive Systems Research*, 2(1):55–66, 2001.

Steven KC Lo. A collaborative multi-agent message transmission mechanism in intelligent transportation system–a smart freeway example. *Information Sciences*, 184(1):246–265, 2012.

Daria Madjidian and Leonid Mirkin. Distributed control with low-rank coordination. *IEEE Transactions on Control of Network Systems*, 1(1):53–63, 2014.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*, 2018.

Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11(3):387–434, 2005.

James A Preiss, Sébastien MR Arnold, Chen-Yu Wei, and Marius Kloft. Analyzing the variance of policy gradient estimators for the linear-quadratic regulator. *arXiv preprint arXiv:1910.01249*, 2019.

Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27. ACM, 2004.

Dominic Richards and Patrick Rebeschini. Optimal statistical rates for decentralised non-parametric regression with linear speed-up. In *Advances in Neural Information Processing Systems*, pages 1214–1225, 2019.

R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976.

William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: a critical survey. *Web manuscript*, 2003.

Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

H Eugene Stanley. *Phase transitions and critical phenomena*. Clarendon Press, Oxford, 1971.

Haoran Sun and Mingyi Hong. Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 38–42. IEEE, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*, 2017.

Hirofumi Uzawa. Iterative methods for concave programming. *Studies in linear and nonlinear programming*, 6:154–165, 1958.

Hoi-To Wai, Zhuoran Yang, Princeton Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*, pages 9649–9660, 2018.

Stephen J Wright. Implementing proximal point methods for linear programming. *Journal of optimization Theory and Applications*, 65(3):531–554, 1990.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Networked multi-agent reinforcement learning in continuous spaces. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2771–2776. IEEE, 2018a.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018b.

Appendix A. Policy Gradient with Reparametrized States

In this section, we provide the details for Reparametrizing states in mean-field dynamics in Section 3, as direct policy gradients over the matrix parameters K and L lead to gradients of K (resp. L) containing the other parameter L (resp. K) due to the correlation mean-field term $\bar{x}_t$ across different agents, which is hard to apply analysis of standard policy gradients.

We characterize the optimal cost from a state going forward (see Eq. (A.6)) with (algebraic) Riccati equations (Bittanti et al., 2012) governed by policy parameters (*i.e.*, K, L). Note that in the formulation of the previous section, it is hard to directly obtain the optimal control $u_t^{(i)}$'s depending on both $x_t^{(i)}$'s and $\bar{x}_t$'s correlated in a single dynamics. Hence we adopt the following reparameterization method to derive decoupled Riccati equations.

$$u_t^{(i)} = Kx_t^{(i)} + L\bar{x}_t = M(x_t^{(i)} - \bar{x}_t) + N\bar{x}_t \triangleq My_t^{(i)} + N\bar{y}_t. \tag{A.1}$$

where we call $M = K$ the optimal *local policy* and $N = K + L$ the optimal *mean-field policy*. Below we show how to derive policy gradient to update these policy parameters separately. We rewrite the dynamic equations below:

$$y_{t+1}^{(i)} = (A + BM)y_t^{(i)} + w_t^{(i)} - \bar{w}_t, \quad \bar{y}_{t+1} = [A + \bar{A} + BN]\bar{y}_t + \bar{w}_t, \tag{A.2}$$

where $\bar{w}_t = 1/n \sum_{i=1}^n w_t^{(i)}$. Then we have

$$C(\Theta) = \mathbb{E}_{\mathbf{y}_0 \sim \mathcal{D}, \mathbf{w}} \sum_{t=0}^{\infty} \gamma^t \big(\sum_{i=1}^{n} {y_t^{(i)}}^\top (Q + M^\top RM) y_t^{(i)} + n\bar{y}_t^\top (Q + \bar{Q} + N^\top RN)\bar{y}_t \big) \tag{A.3}$$

$$= \sum_{i=1}^{n} \operatorname{Tr}[(Q + M^\top RM)\Sigma_M^{(i)}] + n \operatorname{Tr}[(\widetilde{Q} + N^\top RN)\Sigma_N],$$

where $\Sigma_M^{(i)} = \mathbb{E}_{\mathbf{y}_0 \sim \mathcal{D}, \mathbf{w}}[\sum_{t=0}^{\infty} \gamma^t y_t^{(i)} {y_t^{(i)}}^\top]$, $\Sigma_N = \mathbb{E}_{\bar{y}_0 \sim \bar{\mathcal{D}}, \mathbf{w}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{y}_t \bar{y}_t^\top\right]$, $\widetilde{Q} = Q + \bar{Q}$, $\mathcal{D}$ denotes the *i.i.d.* initial state distribution of $y_0^{(i)}$, and $\bar{\mathcal{D}}$ indicates the distribution of mean-field $\bar{y}_0$. We will omit the explicit distributions for expectations without ambiguity. With an abuse of notations, let $C(M^{(i)}) = \operatorname{Tr}[(Q + M^\top RM)\Sigma_M^{(i)}]$, $C(N) = n \operatorname{Tr}[(\widetilde{Q} + N^\top RN)\Sigma_N]$, $\underline{C}(M) = \min_{i \in [n]} \operatorname{Tr}[(Q + M^\top RM)\Sigma_M^{(i)}]$. From Bellman equation for cost functions, it follows that

$$C_\Theta(\mathbf{y}_0) = \sum_{i=1}^{n} {y_0^{(i)}}^\top (Q + M^\top RM) y_0^{(i)} + n\bar{y}_0 \bar{P}_N \bar{y}_0$$
$$+ \gamma \mathbb{E}_{w_0} C_\Theta(\operatorname{diag}(A + BM)\mathbf{y}_0 + \mathbf{w}_0 - \bar{\mathbf{w}}_0), \tag{A.4}$$

where $C_\Theta(y)$ is the value function with initial state y, $\operatorname{diag}(X)$ denotes the $nd \times nd$ block-diagonal matrix with matrix $X \in \mathbb{R}^d$ being the block elements. We assume that the value function takes a quadratic form $C_\Theta(\mathbf{y}_0) = \sum_{i=1}^{n} {y_0^{(i)}}^\top P_M y_0^{(i)} + n\bar{y}_0^\top \bar{P}_N \bar{y}_0 + \alpha_\Theta$, where $P_M, \bar{P}_N \in \mathbb{R}^{d \times d}$ and we ignore the cross terms as $\mathbb{E}[y_0^\top \bar{P}_N \bar{y}_0] = 0$. Let $A_y = A + BM$, $\bar{A}_y = A + \bar{A} + BN$, then by Bellman equation (A.4), we have

$$\sum_{i=1}^{n} {y_0^{(i)}}^\top P_M y_0^{(i)} + n\bar{y}_0^\top \bar{P}_N \bar{y}_0 + \alpha_\Theta = \sum_{i=1}^{n} {y_0^{(i)}}^\top [\gamma A_y^\top P_M A_y + Q + M^\top RM] y_0^{(i)}$$
$$+ n\bar{y}_0^\top [Q + \bar{Q} + \bar{N}RN + \gamma \bar{A}_y^\top \bar{P}_N \bar{A}_y]\bar{y}_0 + \gamma \frac{n+1}{n} \operatorname{Tr} P_M + \gamma \operatorname{Tr} \bar{P}_N + \gamma \alpha_\Theta, \tag{A.5}$$

which implies

$$\begin{aligned} P_M &= Q + M^\top RM + \gamma A_y^\top P_M A_y, \\ \alpha_\Theta &= \frac{\gamma}{1-\gamma}\left(\frac{n+1}{n}\operatorname{Tr} P_M + \bar{P}_N\right), \\ \bar{P}_N &= Q + \bar{Q} + N^\top RN + \gamma \bar{A}_y^\top \bar{P}_N \bar{A}_y. \end{aligned} \tag{A.6}$$

Hence M and N are decoupled for Riccati equations after transformation.The gradient with respect to M in (A.4) gives

$$\nabla_M C_\Theta(\mathbf{y}_0) = (2RM + 2\gamma B^\top P_M A_y)\sum_{i=1}^n y_0^{(i)} {y_0^{(i)}}^\top + \gamma \mathbb{E}\nabla_M C_\Theta(\mathbf{y}_1). \tag{A.7}$$

By applying recursion $y_{t+1} = (A + BM)y_t + w_t^{(i)} - \bar{w}_t$ $(t \geq 0)$ to (A.7) iteratively, we can finally obtain

$$\begin{aligned} \nabla_M C_\Theta(\mathbf{y}_0) &= 2\left(RM + \gamma B^\top P_M A_y\right)\sum_{i=1}^n \mathbb{E}_{\mathbf{y}_0,\mathbf{w}} \sum_{t=0}^\infty \gamma^t y_t^{(i)} {y_t^{(i)}}^\top \\ &= 2\left(RM + \gamma B^\top P_M A_y\right)\sum_{i=1}^n \Sigma_M^{(i)} \\ &\triangleq 2\Xi_M \sum_{i=1}^n \Sigma_M^{(i)}. \end{aligned} \tag{A.8}$$

Similarly, we have

$$\nabla_N C_\Theta(\mathbf{y}_0) = 2\left(RN + \gamma B^\top \bar{P}_N \bar{A}_y\right)\Sigma_N \triangleq 2\Xi_N \Sigma_N. \tag{A.9}$$

The Σ_N in $\nabla_N C_\Theta(\mathbf{y}_0)$ is a shared expectation term for all the agents resulted from mean-field states.

Appendix B. A primal-dual view on MF-DPGM

To see this, we demonstrate the neat connection between our algorithm and primal-dual paradigm commonly used for solving constrained programming. Let us introduce the augmented Lagrangian function $(\mathcal{A})$ in tensor variables:

$$\mathcal{A}_k := \mathcal{A}(\widetilde{\mathbf{\Theta}}_k, \Lambda_k) = \widetilde{C}(\widetilde{\mathbf{\Theta}}_k) + \langle \Lambda_k, \widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_k\}_{(1)}\rangle + \frac{1}{2}\left\|\Omega\left[\widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_k\}_{(1)}\right]_{(1)}\right\|^2, \tag{B.1}$$

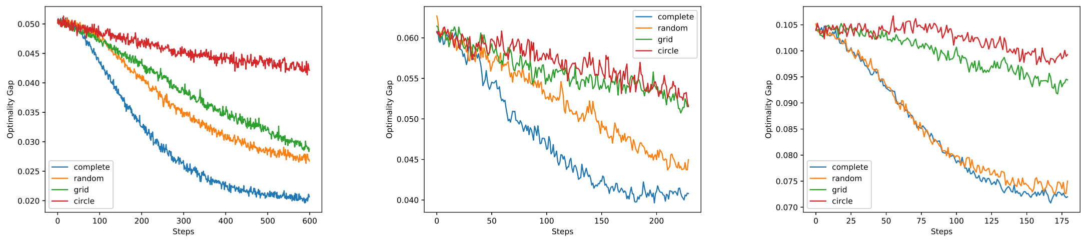
where $\Lambda_k \in \mathbb{R}^{e\times m\times d}$ is the dual variable at iteration k, which is updated as $\Lambda_{k+1} = \Lambda_k + \Omega\left[\widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_k\}_{(1)}\right]_{(1)}$(*). By plugging this into (3.3) and using definition of H we have

$$\nabla\widetilde{C}(\widetilde{\mathbf{\Theta}}_k) + H(\widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k)_{(1)} + \widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_k\}_{(1)} + \widetilde{\mathcal{L}}^\top \Omega^2 \widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_{k+1}\}_{(1)} = \mathbf{0}. \tag{B.2}$$

Viewing (B.2) as the optimal (first-order) condition of some function, we observe that update (3.3) is equivalent to

$$\begin{aligned} \widetilde{\mathbf{\Theta}}_{k+1} = \arg\min_{\widetilde{\mathbf{\Theta}}} &\left\langle \nabla C(\widetilde{\mathbf{\Theta}}_k) + \widetilde{\mathcal{L}}^\top \Lambda_{k(1)}, \widetilde{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}_k \right\rangle \\ &+ \frac{1}{2}\|\Omega\widetilde{\mathcal{L}}\widetilde{\mathbf{\Theta}}_{(1)}\|^2 + \frac{1}{2}\left\|\Omega\mathcal{L}_a\left(\widetilde{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}_k\right)_{(1)}\right\|^2 + \frac{1}{2}\left\|\Gamma\left(\widetilde{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}_k\right)_{(1)}\right\|^2 \end{aligned} \tag{B.3}$$

together with (*), where the third term of (B.3) encodes the network structure that we utilized for neighborhood average. Terms like $\|\Omega\widetilde{\mathcal{L}}\widetilde{\mathbf{\Theta}}_{(1)}\|^2$ show how close policies of agents to each other are with policy $\widetilde{\mathbf{\Theta}}$ by definitions in § **??**. Such a primal-dual interpretation is closely related to some classical comstrained optimization methods, such as the Uzawa method Uzawa



(a) 3-dimensional dynamics (b) 4-dimensional dynamics (c) 5-dimensional dynamics

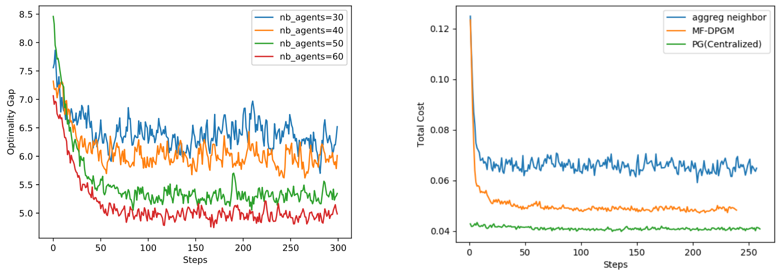
Figure 1: Simulation results (convergence curves) on complete (blue), random (orange), grid (green) and circle (red) networks, as well as different dynamics $d = 3, 4, 5$.

(1958) and the prox-MM Rockafellar (1976); Wright (1990). Following distributed constrained optimization approaches Jakovetić et al. (2014); Hong et al. (2016), we set the error $\varepsilon(t) = \min_{s\in[t]} \left|1/n \sum_{i=1}^{n} C(\widetilde{\Theta}_s^{(i)}) - \widetilde{C}(\widetilde{\boldsymbol{\Theta}}^*)\right| + \|\Omega\widetilde{\mathcal{L}}\{\widetilde{\boldsymbol{\Theta}}_k\}_{(1)}\|^2$ for $t \in \bar{\mathcal{N}}$ to monitor consensus.

Appendix C. Simulation and Result Analysis

We conduct simulation experiments shown in Figure 1(a)-1(c), where we synthesize three different multi-agent systems by specifying dynamics parameters $A, B, \bar{A}$ and cost parameters $Q, R, \bar{Q}$. under the LQR setting each with $d = 3, 4, 5$, $n = 25$, and action space $m = d$; See Appendix D for more setup details. We can see from the numerical results that when fixing the dynamics, different topologies of communication networks lead to diverse performances. The convergence rate ranges from the best to the worst over complete, random, grid, and circle graphs respectively. This reveals the influence of communication structure in control for large coupled systems, where a decentralized algorithm achieves an performance improvement as more possible communication links among agents are established. Moreover, it is shown that the dynamics settings also have an impact on the number of iterations to achieve convergence and stability. More precisely, the system with lower dimension enjoys a faster and stable convergence, which corresponds to the constant α_g of the overall bound in Theorem 4.2.

Furthermore, We plot the convergence curves in Figure 2(a) for different numbers of agents to show the effectiveness of our method at different scales, where we also justify that the mean-field approximation works better as the size of the population grows by the increasing performance with the larger population, as the effect of mean-field phenomenon is more significant in larger systems. In addition, although it is unfair to compare against the centralized setting, where the controller has access to the costs of all agents and updates their policies simultaneously and identically, we plot the comparison in Figure 2(b) for identification with $n = 25$. We reproduce a baseline of decentralized policy gradient with gossip matrices (Richards and Rebeschini, 2019), which aggregates the updated policies from the neighborhood (aggreg. neighbor). Note that our algorithm compares favorably against this baseline and is competitive with the centralized one, which gives a better performance achievable.



(a) Convergence curves on different populations n over a circle graph.

(b) Comparison with baselines on a circle graph.

Figure 2: Simulation results with different size of populations and comparison with baselines.

Appendix D. Experiment Setup and Additional Details

In this section, We provide additional configuration details and analysis of the experimental results in Section C.

Experiment setup. In the experiement we intend to demonstrate the convergence performance of our algorithm under different graph structures and different dynamics. We consider policy learning with global consensus as follows:

$$\min_{\widetilde{\boldsymbol{\Theta}}} \widetilde{C}(\widetilde{\boldsymbol{\Theta}}) := \frac{1}{n}\sum_{i=1}^{n} C(\widetilde{\Theta}^{(i)}), \qquad \text{s.t.} \quad \widetilde{\Theta}^{(i)} = \widetilde{\Theta}^{(j)} \quad \text{for} \quad (i,j) \in \mathcal{E}, \tag{D.1}$$

where for all i,

$$x_{t+1}^{(i)} = Ax_t^{(i)} + Bu_t^{(i)} + \bar{A}\bar{x}_t + w_t^{(i)}, \tag{D.2}$$

$$c_t^{(i)} = x_t^{(i)\top} Q x_t^{(i)} + u_t^{(i)\top} R u_t^{(i)} + \bar{x}_t^\top \bar{Q} \bar{x}_t. \tag{D.3}$$

The state transition matrix $A, B, \bar{A}$ is generated by first sampling a random uniform matrix from 0 to 1 and then tossing a biased coin with probability 0.7 for setting each element zero to keep sparsity for computational efficiency. Also, such transition matrices may lead to smoother loss surfaces. For the reward function we adopt diagonal matrices with each element on the diagonal from 0 to 1 and sparse perturbation for off-diagonal entries. We test on four different graphs as the communication network $\mathcal{G}$: 1) complete graph 2) grid graph 3) circle graph and 4) random graph. The random graph is essentially an Erdos-Rényi graph generated with connectivity of 0.25. Alternatively, for each pair (i, j) with $i \in [n], j \in [n]$, we toss a coin to decide whether there will be an edge between agent i and agent j. As we toss coins two times for both (i, j) and (j, i), an equivalent random graph is obtained with 0.75 probability for each edge to vanish. We set $\Omega^2 = \Gamma^2 = I$ as identity matrices to better understand the popular average of neighborhood scheme. The results of convergence curves are presented in Figure 1(a)- 1(c), where y-axis denotes the error measure $\varepsilon(t) = \min_{s\in[t]} \left|1/n \sum_{i=1}^{n} C(\widetilde{\Theta}_s^{(i)}) - \widetilde{C}(\widetilde{\boldsymbol{\Theta}}^*)\right| + \|\Omega\widetilde{\mathcal{L}}\{\widetilde{\boldsymbol{\Theta}}_k\}_{(1)}\|^2$ for $t \in \bar{\mathcal{N}}$.

For the impact of the topology of the communication graphs on the convergence rate, when choosing Ω and Γ according to Section 4, we can easily verify that the condition (4.1) holds, so that constant $\mathcal{C}$ of the bound in Theorem 4.2 can be reparameterized as follows,

$$\mathcal{C} \leq \frac{320\max\{\sigma_{\max}(Z), 1\}}{\min\{\underline{\sigma}_{\min}(\mathcal{L}_G), 1\}} \sum_{i\sim j}\left(\frac{\sqrt{\beta_i\beta_j}}{\sqrt{d_i d_j} n} + \frac{\bar{\beta}}{4}\right), \tag{D.4}$$

where $\bar{\beta} = 1/n \sum_{i=1}^{n} \beta_i$;see more details in Sun and Hong (2018). Hence the global rate is connected to the algebraic summary, *i.e.*, spectral gap, that captures the connectivity of the communication network, which quantifies how the connectivity of four different graphs impacts convergence performance of MF-DPGM.

For the impact of the dimension of dynamics on the convergence, high dimensions have a higher chance to introduce high variance in $\Sigma_0^{(i)}$ depending on the initial random states, and different cost dynamics also account a change in α_g. In addition, we observe that performance gap between complete graph and other two deterministic graphs increases in higher dimension configuration, which implies an interplay between communication structure ($\mathcal{C}$) and system dynamics (α_g) in the convergence rate.

Appendix E. Proof Sketch

We sketch the proof of results in Section 4. Define $\Sigma_0^{(i)} := \mathbb{E} y_0^{(i)} {y_0^{(i)}}^\top$, $\Sigma_0 := \mathbb{E}\bar{y}_0 \bar{y}_0^\top$, $\xi_i := \sigma_{\min}(\Sigma_0^{(i)})$, $\bar{\xi} = \sigma_{\min}(\mathbb{E}_{y_0^{(i)} \sim \mathcal{D}} \bar{y}_0 \bar{y}_0^\top)$, where $\sigma_{\min}(X)$ denotes the smallest singular value of X. We denote by $\underline{\sigma}_{\min}(X)$ the second smallest eigenvalue of X.

Geometry of cost functions. As mentioned in Section 4, Theorem 4.2 requires moderate smoothness of the landscape of cost functions. Based on the almost Lipschitzness of positive definite matrix P_θ parameterizing the optimal cost from a state going forward, and almost Lipschitzness of the Σ_θ which plays an key role in cost function. Lemma F.2 for each cost function with an almost Lipschitz gradient and Lemma F.3 for each cost function almost dominated by the gradient are crucial in bounding the one-step differences in Lemma F.6 and F.7.

Progress of decentralized iterations. To estimate one-step progress of Algorithm 1, we construct the auxiliary potential function U to show certain monotonicity along the solution path in both primal ($\widetilde{\mathbf{\Theta}}$) and dual ($\Lambda$) variables:

$$U_{k+1} = U_c\left(\widetilde{\mathbf{\Theta}}_{k+1}, \widetilde{\mathbf{\Theta}}_k, \Lambda_{k+1}\right) := \mathcal{A}_{k+1} + \frac{2\kappa}{n^2}\left\|\Gamma^{-1}\Phi\left(\widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k\right)\right\|^2 \tag{E.1}$$
$$+ \frac{c}{2}\left(\left\|\Omega\widetilde{\mathcal{L}}\{\widetilde{\mathbf{\Theta}}_{k+1}\}_{(1)}\right\|^2 + \left\|\widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k\right\|^2_{H+\widetilde{\mathcal{L}}/n}\right),$$

where c is a constant chosen according to Condition 4.1 and $\mathcal{A}_{k+1}$ is the augmented Lagrangian defined in (B.1). The decrement of the potential function at each step is identified with a metric between two adjacent primal iterates in the following lemma.

Lemma E.1. When the parameters of MF-DPGM are chosen to satisfy (4.1) then it holds that

$$U_k - U_{k+1} \geq \tfrac{1}{4}\left\|\widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k\right\|^2_{\widetilde{\Omega}+\Gamma^2} + \kappa\left\|\mathbb{V}_{k+1}\right\|^2_H \tag{E.2}$$

for any $k \geq 0$, where $\mathbb{V}_{k+1} := \left(\widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k\right) - \left(\widetilde{\mathbf{\Theta}}_k - \widetilde{\mathbf{\Theta}}_{k-1}\right)$. Moreover, we have the following bounds:

$$U_k \leq U_0 \leq \widetilde{C}(\widetilde{\mathbf{\Theta}}_0) + \frac{2\nabla\widetilde{C}(\mathbf{0})^\top\Phi^{-1}\nabla\widetilde{C}(\mathbf{0})}{n},\ U_{k+1} \geq \underline{C} > -\infty \tag{E.3}$$

for any $k \geq 1$, where $\mathbf{0} \in \mathbb{R}^{nmd}$ is the all zero vector, and $\nabla\widetilde{C}(\mathbf{0}) = 1/n(\nabla C^{(1)}(\mathbf{0}), ..., \nabla C^{(n)}(\mathbf{0}))$.

Proof. Detailed proof can be found in the appendix of Sun and Hong (2018). □

Note that decreasing potential function in Lemma E.1 also tracks stability of decentralized LQR in the optimization process. We start by optimality condition (B.2) and derive upper bounds for objective gradient norms by the distances of primal variables. Then Lemma E.1 is applied reducing the bounds to differences of adjacent potential functions. Similarly, the consensus error is controlled using Lemma F.6 and F.7, where almost smoothness of costs (F.10) is involved, and processed by Lemma E.1 to keep the same difference terms as those of gradient norms. Finally, combining two bounds of similar structure with Lemma F.3 we establish Theorem 4.2. See F.5 for a detailed proof.

Appendix F. Detailed Proof of Main Results

In this section, we develop detailed proofs for the main result in Section 4 and give complementary details for theoretical claims.

In the sequel, due to the similarity between evolutions of local policies M, and mean-field policy N, we mainly focus on $M^{(i)}$'s to state and prove the results, where similar results are straightforward up to constants in $\bar{A}$, $\bar{Q}$, etc. In some cases, to stress on the discrepancy we formulate both illustrations, or to establish a unified higher-level convergence results, we use compact tensor / vectorization representations, such as $\widetilde{\mathbf{\Theta}}, \mathbf{M}$ for statements. Also, we omit superscript for agent i without confusion in a specific proof.

We first define the following operators on symmetric matrix X,

$$\mathcal{H}_M(X) = \sum_{t=0}^{\infty} \gamma^t (A + BM)^t X [(A + BM)^\top]^t,$$
$$\mathcal{H}_N(X) = \sum_{t=0}^{\infty} \gamma^t (A + \bar{A} + BN)^t X [(A + \bar{A} + BN)^\top]^t. \tag{F.1}$$

We also recall that for $X \in \mathbb{R}^{d\times m}$, we define

$$\mathcal{F}_M(X) = \min \left\{\xi\sigma_{\min}(Q)/[4(\|A + BX\| + 1) \cdot \|B\| \cdot C(X)], \|X\|\right\}, \tag{F.2}$$
$$\mathcal{F}_N(X) = \min\left\{\bar{\xi}\sigma_{\min}(Q + \bar{Q})/[4(\|A + BX\| + 1) \cdot \|B\| \cdot C(X)], \|X\|\right\}, \tag{F.3}$$

which are frequently used notations to simplify our conditions and proof. For notation convenience, define $\xi := \inf_{i\in[n]}[\sigma_{\min}(\mathbb{E}_{y_0^{(i)}\sim\mathcal{D}} y_0^{(i)} {y_0^{(i)}}^\top)]$, $\bar{\xi} := \sigma_{\min}(\mathbb{E}_{y_0^{(i)}\sim\mathcal{D}} \bar{y}_0 \bar{y}_0^\top)$, where $\sigma_{\min}(X)$ refers to the smallest singular value of matrix X. In addition, we define

$$\Sigma_0^{(i)} \triangleq \mathbb{E} y_0^{(i)} {y_0^{(i)}}^\top, \qquad \|\mathcal{H}_M\| \triangleq \sup_X \frac{\|\mathcal{H}_M(X)\|}{\|X\|}. \tag{F.4}$$

It follows that the operator norms are bounded by the composite cost and the extremal singular values of cost matrices.

Lemma F.1 (Upper bounds of operators $\mathcal{H}_M$ and $\mathcal{H}_N$)**.** It holds that

$$\|\mathcal{H}_M\| \le \frac{C(M)}{\xi\sigma_{min}(Q)}, \qquad \|\mathcal{H}_N\| \le \frac{C(N)}{\bar{\xi}\sigma_{min}(Q + \bar{Q})}. \tag{F.5}$$

Using the operator norm bounds and definitions, we show the continuity property of the cost functions, state trajectories, and gradients of cost corresponding to policy $\mathbb{M}$ and $\mathbb{N}$ with an adaptive area for each agent. Lemma F.4-F.2 quantify the problem geometry from different perspectives in distributed settings.

Proof. By the definition in (F.4) and $\Sigma_0 = \mathbb{E}\bar{y}_0\bar{y}_0^\top$, we can obtain for any $i \in [n]$,

$$\mathcal{H}_M(\Sigma_0^{(i)}) = \Sigma_M^{(i)}, \qquad \mathcal{H}_N(\Sigma_0) = \Sigma_N \tag{F.6}$$

By the definition of the operator norm, for $x \in \mathbb{R}^d$ of unit vector norm and matrix X of unit spectral norm, for any $i \in [n]$ we have

$$\begin{aligned}
x^\top(\mathcal{H}_M(X))x &= \sum_{t=0}^{\infty} \mathrm{Tr}([(A+BM)^\top]^t x x^\top (A+BM)^t X) \\
&= \sum_{t=0}^{\infty} \mathrm{Tr}({\Sigma_0^{(i)}}^{1/2}[(A+BM)^\top]^t x x^\top (A+BM)^t {\Sigma_0^{(i)}}^{1/2}{\Sigma_0^{(i)}}^{-1/2} X {\Sigma_0^{(i)}}^{-1/2}) \\
&\leq \sum_{t=0}^{\infty} \mathrm{Tr}({\Sigma_0^{(i)}}^{1/2}[(A+BM)^\top]^t x x^\top (A+BM)^t {\Sigma_0^{(i)}}^{1/2}) \|{\Sigma_0^{(i)}}^{-1/2} X {\Sigma_0^{(i)}}^{1/2}\| \\
&= x^\top \mathcal{H}_M(\Sigma_0^{(i)}) \|{\Sigma_0^{(i)}}^{-1/2} X {\Sigma_0^{(i)}}^{1/2}\| \\
&\overset{(a)}{\leq} \frac{\|\mathcal{H}_M(\Sigma_0^{(i)})\|}{\sigma_{min}(\mathbb{E} x_0^{(i)} {x_0^{(i)}}^\top)} \\
&= \frac{\|\Sigma_M^{(i)}\|}{\xi}, \qquad \text{(F.7)}
\end{aligned}$$

where (a) uses the property $\|\Sigma_0^{(i)}\| \geq \sigma_{\min}(\Sigma_0^{(i)})$. On the other hand, we can derive a upper bound on $\|\Sigma_M^{(i)}\|$ as follows

$$\begin{aligned}
\|\Sigma_M^{(i)}\| \leq \mathrm{Tr}(\Sigma_M^{(i)}) &\leq \frac{\mathrm{Tr}(\Sigma_M^{(i)})\sigma_{\min}(Q)}{\sigma_{\min}(Q)} \\
&\leq \frac{\mathrm{Tr}(\Sigma_M^{(i)}(Q + M^\top R M))}{\sigma_{\min}(Q)} \\
&= \frac{C^{(i)}(M)}{\sigma_{\min}(Q)}. \qquad \text{(F.8)}
\end{aligned}$$

Combining (F.17) and (F.18) and applying uniform lower bound of $C^{(i)}(M)$'s we have $\|\mathcal{H}_M\| \leq \frac{C(M)}{\xi \sigma_{min}(Q)}$. Similar computation gives the upper bound for the norm of $\mathcal{H}_N$. □

F.1 Main Lemmas for the Geometry of Cost Functions

According to Section E, appropriate smoothness of the landscape of cost functions is studied to characterize convergence rates. The following lemma for each cost function with an almost Lipschitz gradient, which is a result of the almost Lipschitzness of positive definite matrix P_θ and almost Lipschitzness of the Σ_θ, shows significance in bounding the one-step differences in Lemma F.6 and F.7.

Lemma F.2. (Almost β-smoothness of private cost functions) Assume that for each $i \in [n]$ and any $\widehat{M}^{(i)}, M^{(i)} \in \mathbb{R}^{m\times d}$ it holds that

$$\begin{aligned}
\|\widehat{M}^{(i)} - M^{(i)}\| &\leq \mathcal{F}_M(M^{(i)}), \\
\|\widehat{N}^{(i)} - N^{(i)}\| &\leq \mathcal{F}_N(N^{(i)}), \qquad \text{(F.9)}
\end{aligned}$$

then for the i-th agent, we have

$$\begin{aligned}
\|\nabla C(\widehat{M}^{(i)}) - \nabla C(M^{(i)})\| &\leq \beta_i^M \|\widehat{M}^{(i)} - M^{(i)}\|, \\
\|\nabla C(\widehat{N}^{(i)}) - \nabla C(N^{(i)})\| &\leq \beta_i^N \|\widehat{N}^{(i)} - N^{(i)}\|, \qquad \text{(F.10)}
\end{aligned}$$

where $\beta_i^M = \text{poly}\left(\mathcal{B}, \mathbb{E}\|y_0^{(i)}\|^2, \frac{C(M_0^{(i)})}{\xi_i \sigma_{\min}(Q)}\right)$, $\beta_i^N = \text{poly}\left(\mathcal{B}, \mathbb{E}\|\bar{y}_0\|^2, \frac{C(N_0^{(i)})}{\bar{\xi}\sigma_{\min}(Q+\bar{Q})}\right)$ denotes the almost smoothness constants, and $\mathcal{B} = \{\|A\|, \|B\|, \|R\|, \sigma_{\min}^{-1}(R)\}$.

Proof. See Lemma 6 in Fazel et al. (2018) for a detailed proof. □

Another crucial property to guarantee the global convergence of MF-DPGM is the gradient domination condition, where the difference of the current cost and optimal cost is bounded by the current gradient norm. We conclude this landscape in $\mathbb{M}$ and $\mathbb{N}$ for each agent in the lemma below.

Lemma F.3. (Gradient domination of cost functions) Suppose $(M^*; N^*)$ is the optimal policy for each agent, and $\Sigma_0^{(i)}$ is full rank. Then $C(M^{(i)}), C(N^{(i)})$ is gradient dominated for each i, that is,

$$C(M^{(i)}) - C(M^*) \leq \alpha_g^M \|\nabla C(M^{(i)})\|^2,$$
$$C(N^{(i)}) - C(N^*) \leq \alpha_g^N \|\nabla C(N^{(i)})\|^2, \tag{F.11}$$

where α_g^M and α_g^N are geometry-dependent coefficients specified in Theorem 4.2.

Proof. See Appendix F.4 for a detailed proof. □

As $\Sigma_M^{(i)} \succeq \Sigma_0^i$, the full-rank condition essentially prevents the denominator of α_g^M from going to zero, so that a stationary point ($\nabla C(M^{(i)}) = 0$) on the R.H.S. of (F.11) implies an optimal policy $M^{(i)}$. Although $\Sigma_{N^{(i)}} \succeq \Sigma_0$, the difference from single-agent setting is the absence of the assumption for Σ_0. In fact, we have $\Sigma_0 = 1/n^2 \mathbb{E}\left(\sum y_0^{(i)}\right)\left(\sum y_0^{(i)}\right)^\top = 1/n^2 \mathbb{E}\sum_{i,j} y_0^{(i)} y_0^{(j)\top} = \mathbb{E} y_0^{(i)} y_0^{(i)\top}$, where *i.i.d.* initial state distributions and linearity of expectation are used. Hence, the only condition in Lemma F.3 suffices to guarantee gradient domination for both local and mean-field policies. Such detail reveals additional advantages of condition relaxation from mean-field symmetry besides dimensionality reduction.

F.2 Lemmas for Almost-Smoothness of Cost Functions

In this subsection, we present two almost continuity lemmas on which Lemma F.2 is based. One of them is the following almost Lipschitz continuity of P_θ parameterizing the cost functions.

Lemma F.4 (Almost Lipschitzness of P_θ (value function)). For any two $M^{(i)}$ and $M^{(i)\prime}$ close enough to each other, that is,

$$\|M^{(i)} - M^{(i)\prime}\| \leq \min\left\{\frac{\xi_i \sigma_{\min}(Q)}{4(\|A + BM^{(i)}\| + 1)\|B\|C(M^{(i)})}, \|M^{(i)}\|\right\}, \tag{F.12}$$

then

$$\|P_{M^{(i)\prime}} - P_{M^{(i)}}\| \leq \frac{6\|M^{(i)}\|\|R\|}{\xi_i^2 \sigma_{\min}^2(Q)}(\|M^{(i)}\|\|B\|\|A + BM^{(i)}\| + \|M^{(i)}\|\|B\| + 1) \cdot \|M^{(i)\prime} - M^{(i)}\|. \tag{F.13}$$

Similarly, if

$$\|N^{(i)\prime} - N^{(i)}\| \leq \min\left\{\frac{\bar{\xi}\sigma_{\min}(Q + \bar{Q})}{4(\|A + \bar{A} + BN^{(i)}\| + 1)\|B\|C(N^{(i)})}, \|N^{(i)}\|\right\}, \tag{F.14}$$

then we have

$$\|P'_{N^{(i)}} - P_{N^{(i)}}\| \leq \frac{6\|N^{(i)}\|\|R\|}{\bar{\xi}^2 \sigma_{\min}^2(Q + \bar{Q})}(\|N^{(i)}\|\|B\|\|A + \bar{A} + BN^{(i)}\| + \|N^{(i)}\|\|B\| + 1)\|N^{(i)\prime} - N^{(i)}\|.$$

Proof. We first define the following operators on symmetric matrix X,

$$\mathcal{H}_M(X) = \sum_{t=0}^{\infty} \gamma^t (A+BM)^t X[(A+BM)^\top]^t,$$
$$\mathcal{H}_N(X) = \sum_{t=0}^{\infty} \gamma^t (A+\bar{A}+BN)^t X[(A+\bar{A}+BN)^\top]^t,$$
$$\mathcal{J}_M(X) = \gamma^t (A+BM)X(A+BM)^\top,$$
$$\mathcal{J}_N(X) = \gamma^t (A+\bar{A}+BN)X(A+\bar{A}+BN)^\top. \tag{F.15}$$

Then we can rewrite the difference between P_M and $P_{M'}$ as

$$\begin{aligned}
&\|P_{M'} - P_M\| \\
&= \|\mathcal{H}_{M'}(Q + M'^\top RM') - \mathcal{H}_M(Q + M^\top RM)\| \\
&\le \left\|\mathcal{H}_{M'}(Q + M'^\top RM') - \mathcal{H}_M(Q + M'^\top RM') - \left(\mathcal{H}_M(Q + M^\top RM) - \mathcal{H}_M(Q + M'^\top RM')\right)\right\| \\
&\le 2\|\mathcal{H}_M\|^2\|\mathcal{J}_M - \mathcal{J}_{M'}\|\|(M')^\top RM'\| + \|\mathcal{H}_M\|\|M^\top RM - (M')^\top RM'\| \\
&\overset{(a)}{\le} \|\mathcal{H}_M\| \left(\|(M')^\top RM' - M^\top RM\| + 2\|\mathcal{H}_M\|\|\mathcal{J}_M - \mathcal{J}_{M'}\|\|M^\top RM\|\right) \\
&+ \|\mathcal{H}_M\|\|M^\top RM - (M')^\top RM'\| \\
&= 2\|\mathcal{H}_M\|^2\|\mathcal{J}_M - \mathcal{J}_{M'}\|\|M^\top RM\| + 2\|\mathcal{H}_M\|\|(M')^\top RM' - M^\top RM\|,
\end{aligned} \tag{F.16}$$

where (a) uses the triangle inequality for ℓ_2-norm, and the assumption $\|\mathcal{H}_M\|\|\mathcal{J}_M - \mathcal{J}_{M'}\| \le 1/2$ for the coefficient of $\|(M')^\top RM' - M^\top RM\|$. To bound the first term in (F.16), letting $\delta = M - M'$, we take the following decomposition for $\|\mathcal{J}_M - \mathcal{J}_{M'}\|$ for each matrix X,

$$\begin{aligned}
\|(\mathcal{J}_M - \mathcal{J}_{M'})(X)\| &= \|(A+BM)X(B\delta)^\top + (B\delta)X(A+BM)^\top - (B\delta)X(B\delta)^\top\| \\
&\le 2\|(A+BM)\|\|X\|\|B\|\|\delta\| + \|B\|^2\|\delta\|^2\|X\|.
\end{aligned} \tag{F.17}$$

According to the definition of the spectral norm and the assumed condition on $\|M - M'\|$ (F.15), we are able to bound the first term as below.

$$\begin{aligned}
&2\|\mathcal{H}_M\|^2\|\mathcal{J}_M - \mathcal{J}_{M'}\|\|M^\top RM\| \\
&\le 2\|\mathcal{H}_M\|^2(2\|(A+BM)\|\|B\|\|M - M'\| + \|B\|^2\|M - M'\|^2)\|M^\top RM\| \\
&\le 4\|\mathcal{H}_M\|^2\|B\|\|M - M'\| \left(\|(A+BM)\| + \frac{\sigma_{min}(Q)\xi}{8C(\Theta)(\|A+BM\|+1)}\right)\|M^\top RM\| \\
&\le 4\|\mathcal{H}_M\|^2\|B\|(\|(A+BM)\|+1)\|M^\top RM\|\|M - M'\|.
\end{aligned} \tag{F.18}$$

Note that $\|M' - M\| \le \|M\|$, the second term in (F.16) can be bounded as By plugging (F.17) and (F.18) into (F.16), we can finally obtain the almost Lipschitzness result for P_M. Similarly, we can also derive an argument for P_N as (F.16) below: Then applying a slightly different upper bound for $\|\mathcal{H}_N\|$ lead to the result. □

The next lemma quantifies a Lipschitz continuity condition for $\Sigma_\theta^{(i)}$. Due to the policy gradient structure, it plays an important role in bounding a part of the gradient difference of cost functions.

Lemma F.5 (Almost-Lipschitzness of Σ_θ)**.** For each $i \in [n]$, if the following holds

$$\|M^{(i)} - M'^{(i)}\| \le \left\{\frac{\sigma_{min}(Q)\xi_i}{4C(M^{(i)})\|B\|(\|A+BM^{(i)}\|+1)}, \|M^{(i)}\|\right\}, \tag{F.19}$$

it follows that

$$\left\|\Sigma_{M'}^{(i)} - \Sigma_{M}^{(i)}\right\| \leq 4\left(\frac{C(M^{(i)})}{\sigma_{\min}(Q)}\right)^2 \frac{\|B\|(\|A+BM^{(i)}\|+1)}{\xi_i}\left\|M^{(i)} - M'^{(i)}\right\|. \tag{F.20}$$

Also, when

$$\|N^{(i)} - N'^{(i)}\| \leq \left\{\frac{\sigma_{min}(Q+\bar{Q})\bar{\xi}}{4C(N^{(i)})\|B\|(\|A+\bar{A}+BN^{(i)}\|+1)}, \|N^{(i)}\|\right\}, \tag{F.21}$$

we have

$$\left\|\Sigma_{N'}^{(i)} - \Sigma_{N}^{(i)}\right\| \leq 4\left(\frac{C(N^{(i)})}{\sigma_{\min}(Q+\bar{Q})}\right)^2 \frac{\|B\|(\|A-BN^{(i)}\|+1)}{\bar{\xi}}\left\|N^{(i)} - N'^{(i)}\right\|. \tag{F.22}$$

Proof. See Fazel et al. (2018) for a detailed proof. □

F.3 Adaptive Choice of Parameters Ω and Γ

In this section, we provide guidance to choose edge associated parameter Ω and agent associated parameter Γ in MF-DPGM algorithm, in order to meet the conditions of the adaptive area of (F.12) and (F.14). Let $\beta_i := \max\{\beta_i^M, \beta_i^N\}$ in Lemma F.2. According to the communication and update step in Algorithm 1, for $M_t^{(i)}$ we have

$$\begin{aligned}
\|M_{t+1}^{(i)} - M_t^{(i)}\| = &\frac{1}{2\sum_{j:j\sim i}\sigma_{ij}^2 + \gamma_i^2}\Bigg\|\frac{1}{n}\left(\nabla C(M_t^{(i)}) - \nabla C(M_{t-1}^{(i)})\right) - 2\sum_{j:j\sim i}\sigma_{ij}^2 M_t^{(j)} \\
&+ \gamma_i^2\left(M_{t-1}^{(i)} - M_t^{(i)}\right) + \sum_{j:j\sim i}\sigma_{ij}^2\left(M_{t-1}^{(j)} + M_{t-1}^{(i)}\right)\Bigg\| \\
\leq &\frac{1}{2\sum_{j:j\sim i}\sigma_{ij}^2 + \gamma_i^2}\Bigg(\frac{1}{n}\|\nabla C(M_t^{(i)}) - \nabla C(M_{t-1}^{(i)})\| + \sum_{j:j\sim i}\sigma_{ij}^2\|M_t^{(j)} - M_{t-1}^{(j)}\| \\
&+ \sum_{j:j\sim i}\sigma_{ij}^2\|M_t^{(j)}\| + \gamma_i^2\|M_t^{(i)} - M_{t-1}^{(i)}\| + \sum_{j:j\sim i}\sigma_{ij}^2\|M_{t-1}^{(i)}\|\Bigg).
\end{aligned} \tag{F.23}$$

Given the almost smoothness is met by iterates t and $t-1$, we proceed with

$$\begin{aligned}
&\|M_{t+1}^{(i)} - M_t^{(i)}\| \\
&\leq \frac{1}{2\sum_{j:j\sim i}\sigma_{ij}^2 + \gamma_i^2}\Bigg(\frac{n\gamma_i^2 + \beta_i}{n}\|M_t^{(i)} - M_{t-1}^{(i)}\| + \sum_{j:j\sim i}\sigma_{ij}^2\|M_t^{(j)} - M_{t-1}^{(j)}\| \\
&+ \sum_{j:j\sim i}\sigma_{ij}^2\|M_t^{(j)}\| + \sum_{j:j\sim i}\sigma_{ij}^2\|M_{t-1}^{(i)}\|\Bigg) \qquad (F.24) \\
&\overset{(b)}{\leq} \frac{1}{2\sum_{j:j\sim i}\sigma_{ij}^2 + \gamma_i^2}\left(\frac{n\gamma_i^2 + \beta_i}{n}\mathcal{F}_M(M_{t-1}^{(i)}) + \sum_{j:j\sim i}\sigma_{ij}^2(\|M_t^{(j)}\| + \|M_{t-1}^{i}\|) + \mathcal{F}_M(M_{t-1}^{(j)})\right).
\end{aligned}$$

Where (b) uses the condition that $M_t^{(i)}$ and $M_{t-1}^{(i)}$ have already stayed in the required adaptive area of (F.12) and $\mathcal{F}_M(X) = \min\left\{\frac{\xi\sigma_{\min}(Q)}{4(\|A+BX\|+1)\|B\|C(X)}, \|X\|\right\}$. Therefore, as long as we have

$$\frac{n\gamma_i^2 + \beta_i}{n}\mathcal{F}_M(M_{t-1}^{(i)}) + \sum_{j:j\sim i}\sigma_{ij}^2\left(\|M_t^{(i)}\| + \|M_{t-1}^{(i)}\| + \mathcal{F}_M(M_{t-1}^{(j)}) - 2\mathcal{F}_M(M_t^{(i)}))\right)$$
$$-\gamma_i^2\mathcal{F}_M(M_t^{(i)}) \leq 0, \tag{F.25}$$

we can always meet the conditions of almost smoothness in Lemma F.2, leading to establishment of one-step progress lemmas (F.6, F.7), and finally providing global convergence theorem.

F.4 Proof of Lemma F.2

Now we proceed to prove the gradient domination lemma based on last two almost Lipschitzness results for cost functions, which is essential to control the one-step progress of MF-DPGM in both primal and dual variables.

Proof. By (A.8) we can split the left-hand-side of (F.10) into two terms

$$\begin{aligned}\|\nabla C(\widehat{M}^{(i)}) - \nabla C(M^{(i)})\|_F &= \|2\Xi_{\widehat{M}^{(i)}}\Sigma_{M^{(i)}} - 2\Xi_{M^{(i)}}\Sigma_{M^{(i)}}\|_F \\ &\leq 2\|\Xi_{M^{(i)}}(\Sigma_{\widehat{M}^{(i)}} - \Sigma_{M^{(i)}})\| + 2\|(\Xi_{\widehat{M}^{(i)}} - \Xi_{M^{(i)}})\Sigma_{\widehat{M}^{(i)}}\|. \end{aligned} \tag{F.26}$$

Let $\widehat{y}_t$ and $\widehat{u}_t$ be the sequence induced by $\widehat{M}^{(i)}$. Note that $C(M^{*(i)}) \leq C(\widehat{M}^{(i)})$. Let $V_M(y) = \mathbb{E}_{\mathbf{w}} y^\top P_M y$, $Q_M(y,u) = y^\top Qy + u^\top Ru + V_M((A+BM)y + \widehat{w})$, $A_M(y,u) = Q_M(y,u) - V_M(y)$. By using cost difference lemma Fazel et al. (2018) and Lemma F.4 and F.5 we have

$$C(M) - C(M^*) \geq C(M) - C\left(\widehat{M}\right) \tag{F.27}$$
$$= -\mathbb{E}\sum_t A_M(\widehat{y}_t, \widehat{u}_t) \tag{F.28}$$
$$= \mathbb{E}\sum_t \mathrm{Tr}\left(\widehat{y}_t\widehat{y}_t^\top \Xi_M^\top \left(R + B^\top P_M B\right)^{-1}\Xi_M\right) \tag{F.29}$$
$$\geq \mathrm{Tr}\left(\Sigma_{\widehat{M}}\Xi_M^\top\left(R + B^\top P_M B\right)^{-1}\Xi_M\right) \tag{F.30}$$
$$\geq \frac{\xi}{\|R + B^\top P_M B\|}\mathrm{Tr}\left(\Xi_M^\top \Xi_M\right). \tag{F.31}$$

Hence we have the norm bound

$$\|\Xi_M\|_F^2 \leq \frac{\|R + B^\top P_M B\|}{\xi}\left(C(M) - C(M^*)\right). \tag{F.32}$$

Then, for the first term in (F.26), we adopt Lemma F.5 to derive the upper bound. For the latter term, we note that $\|\Sigma_{\widehat{M}}\| \leq \|\Sigma_M\| + \frac{C(M)}{\sigma_{\min}(Q)}$ due to small norm of $\widehat{M} - M$. Combining with Lemma F.4 we can obtain the final bound in $\|\widehat{M} - M\|$. □

Next we turn to formulate the one-step progress of dual variable controlling the consensus error in the following lemma.

Lemma F.6 (One-step progress of dual variable)**.** For any $k \in \mathbb{N}$, it holds that

$$\|\Lambda_{k+1} - \Lambda_k\| \le 2\kappa \left(\frac{\left\| \Gamma^{-1}\Phi \left(\widetilde{\mathbf{\Theta}}_k - \widetilde{\mathbf{\Theta}}_{k-1} \right) \right\|^2}{n^2} + \|\mathbb{V}_{k+1}\|_H^2 \right), \tag{F.33}$$

where

$$\begin{aligned} \kappa &:= \frac{1}{\lambda_{\min}\left(\Omega F H^{-1} F^T \Omega\right)}, \\ \mathbb{V}_{k+1} &:= \left(\widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k \right) - \left(\widetilde{\mathbf{\Theta}}_k - \widetilde{\mathbf{\Theta}}_{k-1} \right). \end{aligned} \tag{F.34}$$

Proof. See Sun and Hong (2018) for a detailed proof. □

By this lemma, we are able to transform difference norms of dual variables into primal varibles. $\mathbb{V}$ also keep a second order difference corresponding to the requirement of past gradients and policies by MF-DPGM. Then, we introduce the progress of augmented Lagrangian, which captures the dynamics of both primal and dual variables.

Lemma F.7 (One-step progress of augmented Lagrangian function)**.** For all $k \ge 0$, the iterates in MF-DPGM gives

$$\begin{aligned} \mathcal{A}_{k+1} - \mathcal{A}_k &\le -\frac{1}{2} \left\| \widetilde{\mathbf{\Theta}}_{k+1} - \widetilde{\mathbf{\Theta}}_k \right\|^2_{\widetilde{\Omega} + 2\Gamma^2 - \Phi/n} \\ &+ \kappa \left(\frac{2}{n^2} \left\| \Gamma^{-1}\Phi \left(\widetilde{\mathbf{\Theta}}_k - \widetilde{\mathbf{\Theta}}_{k-1} \right) \right\|^2 + 2 \|\mathbb{V}_{k+1}\|_H^2 \right). \end{aligned} \tag{F.35}$$

Proof. See Sun and Hong (2018) for a detailed proof. □

Again the progress bound of augmented Lagrangian is parameterized by primal first-order differences and second order differences. Converting to such uniform differences is helpful in the proof of the main theorem.

With all the lemmas above in place, now we are ready to prove the global convergence result of our novel decentralized MARL algorithm.

F.5 Proof of Theorem 4.2

Proof. Let $\mathbf{M} \in \mathbb{R}^{nmd}$ be the vectorization of tensor parameter $\mathbb{M}$, $\mathbf{1}$ be the block all one vector with block vectors in $\mathbb{R}^{md}$. From the update of the algorithm, we have the following optimality condition that for any $k \ge -1$, it holds that

$$\langle \mathbf{1}, \nabla \widetilde{C}(\mathbf{M}_k) \rangle + \langle \mathbf{1}, H(\mathbf{M}_{k+1} - \mathbf{M}_k) \rangle = 0. \tag{F.36}$$

By taking the square of both sides, we can obtain

$$\|\frac{1}{n} \sum_{i=1}^{n} \nabla C(\mathbf{M}_k^{(i)})\|^2 = |\mathbf{1}^\top H(\mathbf{M}_{k+1} - \mathbf{M}_k)|^2. \tag{F.37}$$

Using Cauchy-Schwarz inequality under metric matrix H, we have

$$\begin{aligned} |\mathbf{1}^\top H(\mathbf{M}_{k+1} - \mathbf{M}_k)|^2 &\le \|\mathbf{M}_{k+1} - \mathbf{M}_k\|_H^2 \|\mathbf{1}\|_H^2 \\ &= (\mathbf{M}_{k+1} - \mathbf{M}_k)^\top H (\mathbf{M}_{k+1} - \mathbf{M}_k) \cdot \mathbf{1}^\top H \mathbf{1} \\ &\overset{(b)}{\le} \left(4 \sum_{(i,j), i \sim j} \sigma_{ij}^2 + \sum_{i=1}^{n} \gamma_i^2 \right) \|\mathbf{M}_{k+1} - \mathbf{M}_k\|_H^2, \end{aligned} \tag{F.38}$$

when (b) result from the definition of H. Then we combine Lemma E.1, (F.37), and (F.38) to get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\nabla C(M_k^{(i)})\|_F^2 &= \|\frac{1}{n}\sum_{i=1}^{n}\nabla C(\mathbf{M}_k^{(i)})\|^2 \\ &\leq \|\mathbf{M}_{k+1}-\mathbf{M}_k\|_H^2\left(4\sum_{(i,j),i\sim j}\sigma_{ij}^2+\sum_{i=1}^{n}\gamma_i^2\right) \\ &\leq 8\left(4\sum_{(i,j),i\sim j}\sigma_{ij}^2+\sum_{i=1}^{n}\gamma_i^2\right)(U_k-U_{k+1}),\end{aligned} \tag{F.39}$$

where we also use the fact that $H \preceq 2(\widetilde{\Omega}+\Gamma^2)$.

For the error caused by constraint violence (inexact consensus), we know from Lemma F.6 and parameter setting of Γ in Lemma E.1 that

$$\begin{aligned}\|\Omega\widetilde{\mathcal{L}}\mathbf{M}_{k+1}\|^2 &\leq \kappa\left(2\mathbf{V}_{k+1}^\top H\mathbf{V}_{k+1}+\frac{2}{n^2}\|\Gamma^{-1}\Phi(\mathbf{M}_{k+1}-\mathbf{M}_k)\|^2\right) \\ &\leq 4\kappa\left(\frac{1}{n^2}\|\Gamma^{-1}\Phi(\mathbf{M}_{k+1}-\mathbf{M}_k)\|^2+2\|\mathbf{V}_{k+1}\|_H^2\right).\end{aligned} \tag{F.40}$$

Then, we look one step back and using Jensen's inequality to bound similar term for step k,

$$\begin{aligned}\|\Omega\widetilde{\mathcal{L}}\mathbf{M}_k\|^2 &\leq 2\left(\|\Omega\widetilde{\mathcal{L}}\mathbf{M}_{k+1}\|^2+\|\Omega\widetilde{\mathcal{L}}(\mathbf{M}_{k+1}-\mathbf{M}_k)\|^2\right) \\ &\leq 8\kappa\left(\frac{1}{n^2}\|\Gamma^{-1}\Phi(\mathbf{M}_{k+1}-\mathbf{M}_k)\|^2+2\|\mathbf{V}_{k+1}\|_H^2\right)+2\|\Omega\widetilde{\mathcal{L}}(\mathbf{M}_{k+1}-\mathbf{M}_k)\|^2 \\ &= 8\kappa\left(\frac{1}{n^2}\|\mathbf{M}_{k+1}-\mathbf{M}_k\|_{\Phi\Gamma^{-2}\Phi}^2+2\|\mathbf{V}_{k+1}\|_H^2\right)+2\|\mathbf{M}_{k+1}-\mathbf{M}_k\|_{\widetilde{\mathcal{L}}\Omega^2\widetilde{\mathcal{L}}}^2.\end{aligned} \tag{F.41}$$

From the constraints in (4.1), we have

$$\Phi\Gamma^{-2}\Phi \preceq \frac{n^2}{8\kappa}(\widetilde{\Omega}+\Gamma^2). \tag{F.42}$$

Meanwhile, the definition of $\widetilde{\Omega}$ gives

$$\widetilde{\mathcal{L}}\Omega^2\widetilde{\mathcal{L}} \preceq 2\widetilde{\Omega}. \tag{F.43}$$

Again using the step improvement in potential function U combined with (F.42) and (F.43), we have

$$\begin{aligned}\|\Omega\widetilde{\mathcal{L}}\mathbf{M}_k\|^2 &\leq \|\mathbf{M}_{k+1}-\mathbf{M}_k\|_{\widetilde{\Omega}+\Gamma^2}^2+16\kappa\|\mathbf{V}_{k+1}\|_H^2+4\|\mathbf{M}_{k+1}-\mathbf{M}_k\|_{\widetilde{\Omega}}^2 \\ &\leq 5\|\mathbf{M}_{k+1}-\mathbf{M}_k\|_{\widetilde{\Omega}+\Gamma^2}^2+16\kappa\|\mathbf{V}_{k+1}\|_H^2 \\ &\leq 20(U_k-U_{k+1}).\end{aligned} \tag{F.44}$$

On the other hand, according to Lemma F.3 and the measure (left hand side of (4.3)) for convergence rate, we can derive the following inequality on the cost error,

$$\begin{aligned}
t \cdot \min_{k\in[t]} \left| \frac{1}{n}\sum_{i=1}^{n} C(M_k^{(i)}) - C(M^*) \right| &\leq \sum_{k=1}^{t} \left| \frac{1}{n}\sum_{i=1}^{n} C(M_k^{(i)}) - C(M^*) \right| \\
&\leq \sum_{k=1}^{t} \frac{1}{n} \left(\sum_{i=1}^{n} \left| C(M_k^{(i)}) - C(M^*) \right| \right) \\
&\overset{(F.11)}{\leq} \sum_{k=1}^{t} \frac{\|\Sigma_{M^*}\|}{n\sigma_{\min}(\Sigma)^2\sigma_{\min}(R)} \sum_{i=1}^{n} \|\nabla C(M_k^{(i)})\|_F^2 \qquad \text{(F.45)} \\
&\overset{(F.39)}{\leq} \sum_{k=1}^{t} \frac{8\|\Sigma_{M^*}\|(U_k - U_{k+1})}{\sigma_{\min}(\Sigma)^2\sigma_{\min}(R)} \left(4 \sum_{(i,j),i\sim j} \sigma_{ij}^2 + \sum_{i=1}^{n} \gamma_i^2 \right).
\end{aligned}$$

We note that the summation indexed by k only operates on difference terms of adjacent potential functions $(U_k - U_{k+1})$'s. Combining the above results with the upper bound and lower bound of the potential function in Lemma E.1, we have

$$\begin{aligned}
\min_{k\in[t]} \left| \frac{1}{n}\sum_{i=1}^{n} C(M_k^{(i)}) - C(M^*) \right| &\leq \frac{8\|\Sigma_{M^*}\|(U_1 - U_{t+1})}{t\sigma_{\min}(\Sigma)^2\sigma_{\min}(R)} \left(4 \sum_{(i,j),i\sim j} \sigma_{ij}^2 + \sum_{i=1}^{n} \gamma_i^2 \right) \\
&\leq \frac{8\|\Sigma_{M^*}\|(U_0 - \inf_{\mathbb{M}} \widetilde{C}(\mathbb{M}))}{t\sigma_{\min}(\Sigma)^2\sigma_{\min}(R)} \left(4 \sum_{(i,j),i\sim j} \sigma_{ij}^2 + \sum_{i=1}^{n} \gamma_i^2 \right). \qquad \text{(F.46)}
\end{aligned}$$

Similarly, we can directly derive the consensus error bound in constants and t from (F.44) as follows,

$$\begin{aligned}
\min_{k\in[t]} \|\Omega\widetilde{\mathcal{L}}\mathbf{M}_k\|^2 &\leq \frac{1}{t} \cdot \sum_{k=1}^{t} \|\Omega\widetilde{\mathcal{L}}\mathbf{M}_k\|^2 \\
&\leq \frac{20}{t} \sum_{k=1}^{t} (U_k - U_{k+1}) \\
&\leq \frac{20(U_0 - U_{t+1})}{t} \\
&\overset{(E.3)}{\leq} \frac{20}{t} \left(\widetilde{C}(\mathbb{M}_0) - \inf_{\mathbb{M}} \widetilde{C}(\mathbb{M}) + \frac{2\nabla C(\mathbf{0})^\top \Phi^{-1} \nabla C(\mathbf{0})}{n} \right). \qquad \text{(F.47)}
\end{aligned}$$

Therefore, we have attained the cost error and consensus error bound respectively in some problem coonstants. As the first inequalities of both derivations come from the same argument, we can finally obtain the overall convergence rate as

$$\min_{s\in[t]}\left|\frac{1}{n}\sum_{i=1}^{n}C(M_s^{(i)})-C(M^*)\right|+\|\Omega\widetilde{\mathcal{L}}\mathbf{M}_k\|^2$$
$$\leq \underbrace{\frac{8\|\Sigma_{M^*}\|U_0-\inf_{\mathbb{M}}\widetilde{C}(\mathbb{M}))}{\sigma_{\min}(\Sigma)^2\sigma_{\min}(R)}\left(4\sum_{(i,j),i\sim j}\sigma_{ij}^2+\sum_{i=1}^{n}\gamma_i^2\right)}_{\text{cost error bound}}$$
$$+\underbrace{\frac{20}{t}\left(\widetilde{C}(\mathbb{M}_0)-\inf_{\mathbb{M}}\widetilde{C}(\mathbb{M})+\frac{2\nabla C(\mathbf{0})^\top\Phi^{-1}\nabla C(\mathbf{0})}{n}\right)}_{\text{consensus error bound}}$$
$$\leq \frac{2\mathcal{C}'}{t}(5+4\alpha_g\mathcal{C}). \tag{F.48}$$

Since $\alpha_g=\max\{\alpha_g^M,\alpha_g^N\}$ and we choose the commonly applied constant when using inequalities for $\mathbf{M}$ and $\mathbf{N}$, we complete the proof for overall variables. □

Analysis of model-free policy gradient estimator. Our main analysis is based on the exact policy gradient $\nabla C(\mathbb{M})$, while in practice we adopt an empirical version $\widehat{\nabla}C(\mathbb{M})$ for updates. As mentioned in Section 3.1, our proof can be easily tweaked to include the variance incurred by $\widehat{\nabla}C(\mathbb{M})$. Specifically, the objective on the left-hand side of (F.45) is changed into $C(\widehat{M}_k^{(i)})$ where $\widehat{M}_k^{(i)}$ is the iterate obtained by unbiased estimated policy gradient, which gives

$$t\cdot\min_{k\in[t]}\left|\frac{1}{n}\sum_{i=1}^{n}C(\widehat{M}_k^{(i)})-C(M^*)\right|$$
$$\leq t\cdot\min_{k\in[t]}\left|\frac{1}{n}\sum_{i=1}^{n}\left(C(\widehat{M}_k^{(i)})-C(M_k^{(i)})\right)\right|+t\cdot\min_{k\in[t]}\left|\frac{1}{n}\sum_{i=1}^{n}C(M_k^{(i)})-C(M^*)\right|, \tag{F.49}$$

where the first term can be bounded according to the almost Lipschitzness of the cost function by the $\|\widehat{M}_k^{(i)}-M_k^{(i)}\|=\|\widehat{M}_k^{(i)}-\mathbb{E}\widehat{M}_k^{(i)}\|=\mathbf{\Pi}\cdot\|\widehat{\nabla}C(M_k^{(i)})-\mathbb{E}\widehat{\nabla}C(M_k^{(i)})\|$, where Π denotes the step-size of the policy gradient in Algorithm 1. Such standard variance can be further bounded by the variance of the REINFORCE estimator in Preiss et al. (2019). On the other hand, the second term can still be bounded following the proof above. Combining these bounds we obtain the error bound with estimated policy gradients.

Appendix G. Analysis of Computation and Communication Complexities

In this section, we briefly conclude the storage and computation resources and communication overhead during running MF-DPGM.

Firstly, for each agent the method requires previously computed gradients from its own and past simulated states from the neighborhood. Therefore, each agent needs to store two gradient tensors and two policy tensors to avoid the computational overhead brought by re-evaluating trajectory-based policy gradients and states, which is more space efficient than another policy evaluation method Wai et al. (2018). To conclude, the whole system is supposed to store $(2nmd+2nmd)$ real numbers at any iteration. From a view of the update step for each agent, each step inlvolves summation of $m\times d$ matrices by number of neighbors, leading to an $\mathcal{O}(d_imd)$ computation complexity for agent i. On the other hand, according to the information exchanging round in the communication and update step, MF-DPGM requires $\mathcal{O}(2e)$ communications at each round and each exchange delivers md real numbers. Although the overhead is greatly alleviated compared to centralized scheme, it still cost much bandwidth when the policy is extremely complicated to parameterize. We are trying to infer state information from part of the neighborhood to further reduce communication as the future work.