Statistical Inference for Distributed Contextual Multi-armed Bandit

Wuwenqing Yan and Yongchao Liu
School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China
(ywwq@mail.dlut.edu.cn (Yan), lyc@dlut.edu.cn (Liu))

Abstract. In this paper, we study the online statistical inference of distributed contextual multi-armed bandit problems, where the agents collaboratively learn an optimal policy by exchanging their local estimates of the global parameters with neighbors over a communication network. We propose a distributed online decision making algorithm, which balances the exploration and exploitation dilemma via the ε-greedy policy and updates the policy online by incorporating the distributed stochastic gradient descent algorithm. We establish the pivotal limiting distribution for the estimator of reward model parameter as a stochastic process and then employ the random scaling method to construct its asymptotic confidence interval. We also establish the asymptotic normality of the online inverse probability weighted value estimator and construct an asymptotic confidence interval of the value by plug-in method. The proposed algorithm and theoretical results are tested by simulations and a real data application to a warfarin drug dosing problem.

Key words. Distributed contextual multi-armed bandit, confidence intervals, random scaling, ε-greedy, distributed stochastic gradient descent

1 Introduction

Contextual multi-armed bandit (CMAB) is an online sequential decision making process where the agent observes a feature of users (referred to as context) at each round and then selects an action, in return only receiving a scalar reward corresponding to the chosen action. The goal of the agent is to learn an optimal policy that maximizes the expected cumulative reward through balancing the trade-off between exploitation and exploration. Many practical problems, such as healthcare [5, 18], recommend system [6, 27] and dialogue system [30], can be modeled as CMAB. In the past decades, efficient algorithms have been proposed for CMAB, including forced sampling method [5, 20], ε-greedy method [33, 38], upper confidence bound method [4, 27] and Thompson sampling [2, 8]. We refer interested readers to the survey [39] for a more comprehensive overview on the development of contextual bandit problems.

In many real-world applications, the sequential decision making involves multiple agents and is distributed by nature. For example, multiple physically separated hospitals may conduct the same clinical trial simultaneously; E-commerce platforms or streaming media recommendation systems utilize multiple servers to provide personalized product or content recommendations to a large number of users. This motivates people to expand the traditional single-agent bandit problems to accommodate multi-agents frameworks [17, 22, 26, 36]. Wang et al. [36] consider the linear contextual bandit with a central server, where the agents are given access to the same bandit model and communicate with the server via packet-based transmission. The authors propose two algorithms: DELB and DisLinUCB. DELB is an elimination-based protocol designed for fixed action sets, which divides the time horizon into phases and eliminates suboptimal arms at the end of each phase through global confidence bounds. DisLinUCB is a linear upper confidence bound (UCB) based algorithm for time-varying action sets, where the agents update the

local confidence set based on new observations in each step, and upload the accumulation of the local data to the central server in the step that the volume of the local confidence ellipsoid varies greatly. DELB achieves a regret bound of $O(p\sqrt{nTlogT})$ and DisLinUCB achieves a regret bound of $O(p\sqrt{nT}log^2T)$, where n is the number of agents, T is the time horizon and p is the dimension of the feature vector. Different from [36], Huang et al. [22] focus on a federated linear contextual bandit under heterogeneous data. The authors propose Fed-PE, a federated phased elimination algorithm where agents only upload their local estimators of the global parameters to the central server. Fed-PE may achieve a regret bound of $O(\sqrt{pnTlog(KnT)})$, where K is the number of arms. Li and Wang [26] extend the linear models [36] to federated generalized linear contextual bandit problems and propose the FedGLB-UCB, which updates local parameters via online newton iteration in each step. On the other hand, FedGLB-UCB updates global parameters offline via accelerated gradient descent method in the step that an indicator determined by the time interval and the local covariance matrix exceeds a preset threshold value. FedGLB-UCB achieves a regret bound of $O(p\sqrt{nTlog(nT)})$. Dubey and Pentland [17] study distributed linear contextual bandits, and propose FedUCB for both centralized and decentralized (peer-to-peer) settings, which uploads the noisy parameter estimators in each round of communication to ensure the (ϵ, δ)-federated differential privacy for the agents. FedUCB may achieve a regret bound of $\tilde{O}(p^{\frac{3}{4}}\sqrt{nT/\epsilon})$ in the centralized setting and a regret bound of $\tilde{O}(p^{\frac{3}{4}}\sqrt{diam(G)nT/\epsilon})$ in the decentralized setting, where $diam(G)$ is the diameter of communication network G.

Existing literature on distributed contextual bandit problems often focuses on maximizing the expected cumulative reward outcomes, with less attention paid to statistical inference. In real-world applications, it is important to have reliable uncertainty quantification of the learned policy, as it may provide guidance for policy interventions and indicate potential risks in recommendations [11]. Indeed, the statistical inference of contextual bandit problems has been widely studied [9, 10, 11, 21], where the majority of the works focus on the ε-greedy based algorithm. Chen et al. [9] study the two-armed contextual bandit problem with linear reward using ε-greedy policy, where the authors provide the online ordinary least squares estimators for parameters and use an inverse propensity weighting (IPW) estimator of the value function to correct the bias induced by the ε-greedy exploration. The asymptotic normality for the estimators of parameter and value is established, and then the corresponding confidence intervals are constructed in an offline setting. As a following work of [9], Chen et al. [10] propose a ε-greedy policy and stochastic gradient descent (SGD) based algorithm for contextual bandits, where the asymptotic normality of the IPW estimators of the reward model parameter and the value is established by the martingale central limit theorem. For constructing the confidence interval, the authors provide the online plug-in estimator of the asymptotic variance for the parameters and the value. Chen et al. [11] study ε-greedy policy and SGD based algorithm for contextual bandits by providing a generalized-weighting method, which allows multiple weighting schemes including IPW, sqrt-IPW and vanilla. The authors establish the asymptotic normality for the estimator of reward model parameter and construct the confidence interval by plug-in method. More recently, Han et al. [21] study ε-greedy policy and SGD based algorithm for low-rank matrix contextual bandit problems, where SGD has to accommodate the low-rank structure of the model parameter. The authors introduce an doubly-debiasing inference procedure, which may correct the bias induced by both low-rankness and data adaptivity simultaneously, and then establish the asymptotic normality of the proposed online doubly-debiased estimators and

construct the confidence intervals for the true matrix.

Motivated by the works [9, 10, 11, 21], we focus on the statistical inference of distributed contextual bandit problems with n agents, where the individual agent is associated with local private datasets and they collaboratively learn a mapping from a feature vector to an optimal action over a communication network. As far as we are concerned, the contribution of the paper can be summarized as follows.

- We propose an ε-greedy policy and distributed stochastic gradient descent based algorithm for the contextual bandits (ε-DSGDCB). Each agent in ε-DSGDCB builds a local update by first performing a consensus step — communicating its local estimates of global reward model parameters with neighbors to seek a common parameter over a connected network — followed by a corrected gradient-based update. The distributed structure of ε-DSGDCB eliminates the dependency on the central server and does not have to share their local feature vectors or raw observations, which inherently reduces the cost of data transmission and enhances the privacy of personal information. We show that ε-DSGDCB may achieve the expected regret bound of $O(T^{3/4})$.

- We derive inferential results of the policy produced by ε-DSGDCB. We establish the pivotal limiting distribution for the estimator of reward model parameter as a stochastic process by the functional central limit theorem, where the asymptotic distribution is free of any unknown nuisance parameters. Then, we leverage the random scaling method to construct an asymptotically pivotal statistic by normalizing the estimator of parameter with its random transformation. Compared with the plug-in technique in [10, 11, 21], the random scaling method has lower computational and storage costs as it only uses a solution path for inference and does not need to estimate the asymptotic variance. Another advantage of the random scaling method for distributed contextual bandit is the reduction of communication costs as it does not need to transmit gradients and Hessian matrices among neighbors. Furthermore, the asymptotic normality and the confidence interval of the value are also studied.

- We illustrate the numerical performance of our methods in simulations and a real data application to the warfarin drug dosing problem. In the real-world application, ε-DSGDCB makes 21 physically separated research groups to jointly learn an optimal dosing policy to maximize overall patient remission rates without private data sharing. The resulting bandit policy significantly outperforms the physicians' benchmark policy in practice.

The rest of the paper is organized as follows. In Section 2, we introduce the model of distributed contextual multi-armed bandit and ε-DSGDCB, where the regret bound of the proposed algorithm is discussed. Section 3 establishes the pivotal limiting distribution for the estimators of parameter and constructs its asymptotic confidence interval by the random scaling method. Moreover, inferential results for the estimators of value are discussed. Finally, empirical results on simulated data and a real application on the warfarin drug dosing problem are presented in Section 4.

Throughout the paper, $\mathbb{R}^p$ denotes the p-dimension Euclidean space endowed with norm $\|\beta\| = \sqrt{\langle \beta, \beta \rangle}$. $[a]$ denotes the largest integer less than or equal to a. $\mathbf{1} \in \mathbb{R}^p$ denotes the vector

of all 1s. $\beta_{[i]}$ denotes the i-th element of β. e_t denotes the unit vector with the t-th coordinate as 1 and the other coordinates as 0. For matrices, $\|\cdot\|$ represents the Frobenius norm. $\mathbf{I}_p \in \mathbb{R}^{p\times p}$ denotes the identity matrix. $A \otimes B$ denotes the Kronecker product of matrix A and B. For a sequence of random vectors $\{\xi_t\}$ and a random vector ξ, $\xi_t \Rightarrow \xi$ denotes the weak convergence, $\xi_t \xrightarrow{d} \xi$ denotes the convergence in distribution and $\mathrm{Cov}(\xi)$ denotes the covariance matrix of random vector ξ. The communication network is denoted by an undirected connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \cdots, n\}$ denotes the set of agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the edge set of connecting agents, and the corresponding communication matrix is $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{n\times n}$ with $m_{ij} \geq 0$.

2 Problem Setting and Algorithm

Consider a distributed contextual bandit problem involving n agents deployed on a communication network. For $t \in \mathcal{T} := \{1, 2, \cdots, T\}$, each agent i observes a p-dimensional individual-specific feature $x_{i,t} \in \mathbb{R}^p$, and then pulls an arm $a_{i,t} \in \mathcal{A} := \{1, 2, \cdots, K\}$ and receive a reward $y_{i,t} \in [0, 1]$. The task of agents is to learn a common optimal policy $d^* : \mathbb{R}^{Kp} \mapsto \mathcal{A}$ via exchanging information only with their neighbors privately.

Note that it is often difficult to obtain a closed form of the conditional expected reward or compute its value numerically, we consider the case that conditional expectation reward is a parametric function for agent i,

$$\mathbb{E}(Y_i \mid A_i, X_i) = \mu_i(A_i, X_i; \beta^*),$$

where $\beta^* \in \mathbb{R}^{Kp}$ is an unknown parameter. Let $O_i = (X_i, A_i, Y_i)$ be the observed triples [1] and the true reward Y_i is generated by $\mu_i(A_i, X_i; \beta^*) + E_i$, where E_i is a random error and independent of A_i and X_i. We consider the case that for each agent, the random policy selects each action with equal probability and then the expected loss is

$$\min_{\beta\in\mathbb{R}^{Kp}} L(\beta) = \sum_{i=1}^{n} L_i(\beta), \tag{2.1}$$

where

$$L_i(\beta) = \iiint \ell_i(\beta; x_i, a_i, y_i) d\mathcal{P}_{Y_i|X_i,A_i}(y_i|x_i, a_i) d\mathcal{P}^r_{A_i}(a_i) d\mathcal{P}_{X_i}(x_i). \tag{2.2}$$

Here, $\ell_i(\beta; O_i)$ denotes the i-th loss function, which is a function of the reward model $\mu_i(A_i, X_i; \beta)$, $\mathcal{P}_{X_i}$ is the distribution of X_i, $\mathcal{P}^r_{A_i}$ is a discrete uniform distribution with probability $1/K$ and $\mathcal{P}_{Y_i|X_i,A_i}$ is the conditional distribution of Y_i given X_i, A_i

Next, we introduce ε-DSGDCB.

Algorithm 1 provides a pseudo-code of ε-DSGDCB. In line 6, the inverse probability weighting gradient $g_i(\beta_{i,t-1}; o_{i,t})$ is used to correct the bias induced by the ε-greedy exploration. In line 7, the parameter of loss function is updated by the distributed stochastic gradient descent, where

[1] We use uppercase letters "$O = (X, A, Y)$" to denote the random variables representing the triples including the context, action and reward, while their lowercase counterparts "$o = (x, a, y)$" represent the corresponding realizations.

Algorithm 1 ε-greedy policy and distributed stochastic gradient descent based algorithm for the contextual bandits(ε-DSGDCB): At each node $i \in \mathcal{V} = \{1, 2, \ldots, n\}$

1: **Input:** initial value $\beta_{i,0} \in \mathbb{R}^{Kp}$, $\pi_{i,0} = 1/K$, learning rate $\gamma_t > 0$, exploration rate $\varepsilon_t \in [0, 1]$, communication matrix $\mathbf{M} = [m_{ij}]$.
2: **for** $t = 1$ to T **do**
3: Observe a context $x_{i,t}$.
4: Choose an arm $a_{i,t}$ with probability $\pi_{i,t-1}(x_{i,t})$.
5: Observe a reward $y_{i,t}$.
6: Calculate the inverse probability weighting gradient with $o_{i,t} = (x_{i,t}, a_{i,t}, y_{i,t})$

$$g_i(\beta_{i,t-1}; o_{i,t}) = \frac{\nabla \ell_i(\beta_{i,t-1}; o_{i,t})}{K \pi_{i,t-1}(x_{i,t})}.$$

7: State update

$$\beta_{i,t} = \sum_{j=1}^{n} m_{ij} \beta_{j,t-1} - \gamma_t g_i(\beta_{i,t-1}; o_{i,t}). \tag{2.3}$$

8: Update $\bar{\beta}_{i,t} = \frac{t-1}{t} \bar{\beta}_{i,t-1} + \frac{1}{t} \beta_{i,t}$.
9: Update $\pi_{i,t}(x_{i,t}) = \begin{cases} \varepsilon_t/K + 1 - \varepsilon_t, & \text{if } a_{i,t} = \operatorname{argmax}_{a \in \mathcal{A}} \mu_i\left(a, x_{i,t}; \bar{\beta}_{i,t}\right), \\ \varepsilon_t/K, & \text{otherwise.} \end{cases}$
10: **end for**

the first term on the right hand side of (2.3) represents the communication of parameter with neighbors through the communication matrix $\mathbf{M}$ for agreement and the second term represents the current parameter adjustment along the direction of the corrected gradient $g_i(\beta_{i,t-1}; o_{i,t})$. In line 9, the ε-greedy policy is used to pull the optimal arm with high probability $\varepsilon_t/K + 1 - \varepsilon_t$ and the suboptimal arm with low probability ε_t/K.

We next record the assumptions that will be used throughout the paper.

Assumption 2.1. *(i) $L(\beta)$ is μ-strongly convex ($\mu > 0$) in β, that is,*

$$L(\gamma) \geq L(\beta) + \langle \nabla L(\beta), \gamma - \beta \rangle + \frac{\mu}{2} \|\beta - \gamma\|^2, \quad \forall \beta, \gamma \in \mathbb{R}^{Kp}.$$

(ii) $\nabla^2 L(\beta^*)$ is positive definite and there exists $c > 0$ such that

$$\left\| \nabla L(\beta) - \nabla^2 L(\beta^*)(\beta - \beta^*) \right\| \leq c \|\beta - \beta^*\|^2, \quad \forall \beta \in \mathbb{R}^{Kp},$$

where β^* is the optimal solution to problem (2.1).

Assumption 2.2. *For $\forall i \in \mathcal{V}$, (i) there exists a positive random variable $c_{L,i}(O_i)$ such that $\mathbb{E}\left[c_{L,i}^2(O_i)\right] < \infty$ and*

$$\|\nabla \ell_i(\beta; O_i) - \nabla \ell_i(\gamma; O_i)\| \leq c_{L,i}(O_i) \|\beta - \gamma\|, \quad \forall \beta, \gamma \in \mathbb{R}^{Kp};$$

(ii) there exists a constant $c_f > 0$ such that

$$\mathbb{E}_{\mathcal{P}_O^r}\left[\|\nabla \ell_i(\beta^*; O_i)\|^2\right] \leq c_f,$$

where $\mathcal{P}_O^r$ is the joint distribution of O_i when A_i follows $\mathcal{P}_A^r$.

Assumption 2.3. *The communication matrix* $\mathbf{M}$ *is doubly stochastic, i.e.,* $\mathbf{M1} = \mathbf{1}$ *and* $\mathbf{1}^T\mathbf{M} = \mathbf{1}^T$. *There exists a constant* $\rho \in (0, 1]$ *such that* $\left\|\mathbf{M} - \frac{\mathbf{11}^T}{n}\right\| \leq 1 - \rho$.

Assumption 2.4. *There exists a constant* $c_l > 0$ *such that for* $\forall i \in \mathcal{V}$,

$$\|L_i(\beta) - L_i(\gamma)\| \leq c_l\|\beta - \gamma\|, \quad \forall \beta, \gamma \in \mathbb{R}^{Kp}.$$

Assumption 2.1 (i) guarantees the uniqueness of the optimal parameter of the loss function (2.1). Assumption 2.1 (ii) is the standard condition for studying the asymptotic normality of stochastic approximation based algorithm [31]. Assumption 2.2 (i) implies the Lipschitz continuity of $\nabla L_i(\cdot)$, i.e.,

$$\|\nabla L_i(\beta) - \nabla L_i(\gamma)\| \leq c_L\|\beta - \gamma\|,$$

where $c_L = \max_{1\leq i\leq n} \mathbb{E}\left[c_{L,i}\left(O_i\right)\right]$. Assumption 2.2 (ii) ensures the boundedness of covariance matrix of $\nabla \ell_i(\beta^*; O_i)$. Assumption 2.3 is the standard conditions in distributed stochastic gradient descent literature [7], which implies that $\left(\frac{1}{n}\mathbf{11^T}\right)\mathbf{M} = \mathbf{M}\left(\frac{1}{n}\mathbf{11^T}\right) = \frac{1}{n}\mathbf{11}^T$. Assumption 2.4 presents the Lipschitz continuity of $L_i(\cdot)$, which will be used to analyze the regret bound of ε-DSGDCB.

For ease of the notation, we define the filtration $\mathcal{F}_t := \sigma\{o_{i,k} : i \in \mathcal{V}, 1 \leq k \leq t\}$, which is the σ-algebra generated by $\{o_{i,k} : i \in \mathcal{V}, 1 \leq k \leq t\}$, and denote $\mathcal{P}_O^\pi$ as the joint distribution of O_i under the ε-greedy policy in ε-DSGDCB,

$$\begin{aligned}
\beta_t &:= \left[\beta_{1,t}^T, \beta_{2,t}^T, \cdots, \beta_{n,t}^T\right]^T, \\
G_t &:= \left[g_1\left(\beta_{1,t}; o_{1,t+1}\right)^T, g_2\left(\beta_{2,t}; o_{2,t+1}\right)^T, \cdots, g_n\left(\beta_{n,t}; o_{n,t+1}\right)^T\right]^T, \\
\hat{\beta}_t &:= \left(\frac{\mathbf{1}^T}{n} \otimes \mathbf{I}_{Kp}\right)\beta_t, \quad \hat{G}_t := \left(\frac{\mathbf{1}^T}{n} \otimes \mathbf{I}_{Kp}\right)G_t,
\end{aligned}$$

Here, β_t and G_t are formed by stacking all agents' parameters and corrected gradients, $\hat{\beta}_t$ and $\hat{G}_t$ are the averages of all agents' parameters and corrected gradients.

The following two technical lemmas characterize the solution error and consensus error of ε-DSGDCB.

Lemma 2.1 (Solution Error)**.** *Let* $c_1 = \frac{c_L}{2\delta}, c_2 = 4c_L^2\left(\frac{8}{\varepsilon_\infty} + 1\right)$ *and* $\tilde{\mu} = \frac{\mu}{n} - \frac{\delta}{n}c_L$ *with* $\delta \in (0, \frac{\mu}{c_L})$. *Suppose that (i) Assumptions 2.1-2.3 hold, (ii) the exploration rate* $\varepsilon_t \to \varepsilon_\infty > 0$, *(iii) the diminishing learning rate* $\gamma_t \to 0$ *satisfies* $\gamma_0 \leq \frac{\tilde{\mu}}{c_2}$. *Then, for any* $t \geq 0$,

$$\mathbb{E}_{\mathcal{P}_O^\pi}\left[\|\Delta_{t+1}\|^2 \mid \mathcal{F}_t\right] \leq (1 - \tilde{\mu}\gamma_{t+1})\|\Delta_t\|^2 + \left[c_1\frac{\gamma_{t+1}}{n} + c_2\frac{\gamma_{t+1}^2}{n}\right]\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|^2 + \frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2,$$

where $\Delta_t = \hat{\beta}_t - \beta^*$.

Proof. The proof is provided in Section 5.2 of the Appendix. □

Lemma 2.2 (Consensus Error)**.** *Suppose that (i) Assumptions 2.1-2.3 hold, (ii) the exploration rate* $\varepsilon_t \to \varepsilon_\infty > 0$, *(iii) the diminishing learning rate* $\gamma_t \to 0$ *satisfies* $\gamma_0 \leq \rho/\sqrt{2c_3}$, *where*

$c_3 = 6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)$. *Then, for any* $t \geq 0$,

$$\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\beta_{t+1} - \mathbf{1}_n \otimes \hat{\beta}_{t+1}\right\|^2 \mid \mathcal{F}_t\right] \leq (1-\rho/2)\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|^2 + \frac{\gamma_{t+1}^2}{\rho} 6nc_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\|\Delta_t\|^2 + 3n\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\frac{\gamma_{t+1}^2}{\rho}.$$

Proof. The proof is provided in Section 5.3 of the Appendix. □

With Lemmas 2.1 and 2.2 in hand, we have the following convergence result of parameters.

Theorem 2.1. *Let* $c_1 = \frac{c_L}{2\delta}, c_2 = 4c_L^2\left(\frac{8}{\varepsilon_\infty}+1\right), c_3 = 6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)$ *and* $\tilde{\mu} = \left(\frac{\mu}{n} - \frac{\delta}{n}c_L\right)$ *with* $\delta \in (0, \frac{\mu}{c_L})$. *Suppose that (i) Assumptions 2.1-2.3 hold, (ii) the exploration rate* $\varepsilon_t \to \varepsilon_\infty > 0$, *(iii) the diminishing learning rate* $\gamma_t \to 0$ *satisfies*

$$\frac{\gamma_t}{\gamma_{t+1}} \leq \min\left\{\sqrt{1+(\widetilde{\mu}/4)\gamma_{t+1}^2}, \sqrt[3]{1+(\widetilde{\mu}/4)\gamma_{t+1}^3}, 1+\rho/(4-2\rho)\right\}, \quad \forall t \geq 0$$

and $\gamma_0 \leq \min\left\{\frac{\widetilde{\mu}}{c_2}, \frac{\rho}{\sqrt{2c_3}}, \sqrt{\frac{\rho^2\widetilde{\mu}}{96c_1\left(\frac{16}{\varepsilon_\infty}+3\right)c_L^2}}, \frac{3\rho c_1}{4\widetilde{\mu}c_1+\rho c_2}\right\}$. *Then,* $\forall t \geq 0$,

$$\mathbb{E}\left\|\Delta_{t+1}\right\|^2 \leq \prod_{i=1}^{t+1}\left(1-\frac{\widetilde{\mu}\gamma_i}{2}\right)\mathrm{D} + \frac{96c_1\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)}{\rho^2\widetilde{\mu}}\gamma_{t+1}^2 + \frac{16c_f}{\widetilde{\mu}\varepsilon_\infty}\gamma_{t+1},$$

$$\mathbb{E}\left[\left\|\beta_{t+1} - \mathbf{1}_n \otimes \hat{\beta}_{t+1}\right\|^2\right] \leq \left(1-\frac{\rho}{2}\right)^{t+1}\left\|\beta_0 - \mathbf{1}_n \otimes \hat{\beta}_0\right\|^2 + \frac{4n\left[6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\bar{\Delta} + 3\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\right]}{\rho^2}\gamma_{t+1}^2,$$

where $\mathrm{D} = \|\Delta_0\|^2 + \frac{8\gamma_1 c_1}{\rho n}\left\|\beta_0 - \mathbf{1}_n \otimes \hat{\beta}_0\right\|^2$ *is the initial error and* $\bar{\Delta} = \mathrm{D} + 1 + \frac{16c_f}{c_2\varepsilon_\infty}$.

Proof. Define an auxiliary sequence

$$\mathcal{L}_{t+1} := \mathbb{E}\left[\|\Delta_{t+1}\|^2 + \gamma_{t+1}\frac{8c_1}{\rho n}\left\|\beta_{t+1} - \mathbf{1}_n \otimes \hat{\beta}_{t+1}\right\|^2\right], \quad \forall t \geq 0. \tag{2.4}$$

By Lemmas 2.1 and 2.2,

$$\begin{aligned}
\mathcal{L}_{t+1} \leq & \underbrace{(1-\widetilde{\mu}\gamma_{t+1})\,\mathbb{E}\|\Delta_t\|^2 + \left[c_1\frac{\gamma_{t+1}}{n} + c_2\frac{\gamma_{t+1}^2}{n}\right]\mathbb{E}\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|^2 + \frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2}_{Lemma\ 2.1} + \gamma_{t+1}\frac{8c_1}{\rho n} \\
& \underbrace{\left[\left(1-\frac{\rho}{2}\right)\mathbb{E}\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|^2 + \frac{\gamma_{t+1}^2}{\rho}6nc_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\mathbb{E}\|\Delta_t\|^2 + 3n\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\frac{\gamma_{t+1}^2}{\rho}\right]}_{Lemma\ 2.2} \\
= & \left[1 - \tilde{\mu}\gamma_{t+1} + \frac{48c_1}{\rho^2}\left(\frac{16}{\varepsilon_\infty}+3\right)c_L^2\gamma_{t+1}^3\right]\mathbb{E}\|\Delta_t\|^2 + \frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2 + \frac{24c_1}{\rho^2}\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\gamma_{t+1}^3 \\
& + \gamma_t\frac{8c_1}{\rho n}\left[\frac{\gamma_{t+1}}{\gamma_t}\left(1-\frac{\rho}{2}+\frac{\rho}{8}+\frac{c_2\rho}{8c_1}\gamma_{t+1}\right)\right]\mathbb{E}\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|^2.
\end{aligned}$$

Note that $\gamma_0 \leq \min\left\{\sqrt{\frac{\rho^2\widetilde{\mu}}{96c_1\left(\frac{16}{\varepsilon_\infty}+3\right)c_L^2}}, \frac{3\rho c_1}{4\widetilde{\mu}c_1+\rho c_2}\right\}$ and the diminishing learning rate $\gamma_t \to 0$, then

$$1-\tilde{\mu}\gamma_{t+1}+\frac{48c_1}{\rho^2}\left(\frac{16}{\varepsilon_\infty}+3\right)c_L^2\gamma_{t+1}^3 \leq 1-\frac{\tilde{\mu}\gamma_{t+1}}{2}$$
$$\frac{\gamma_{t+1}}{\gamma_t}\left(1-\frac{\rho}{2}+\frac{\rho}{8}+\frac{c_2\rho}{8c_1}\gamma_{t+1}\right) \leq 1-\frac{\tilde{\mu}\gamma_{t+1}}{2},$$

which implies

$$\begin{aligned}
\mathcal{L}_{t+1} &\leq (1-\widetilde{\mu}\gamma_{t+1}/2)\,\mathcal{L}_t + \frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2 + \frac{24c_1}{\rho^2}\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\gamma_{t+1}^3 \\
&\leq \prod_{i=1}^{t+1}\left(1-\frac{\widetilde{\mu}\gamma_i}{2}\right)\mathrm{D} + \sum_{s=1}^{t+1}\prod_{i=s+1}^{t+1}(1-\widetilde{\mu}\gamma_i/2)\left(\frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2 + \frac{24c_1}{\rho^2}\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\gamma_s^3\right) \\
&\leq \prod_{i=1}^{t+1}\left(1-\frac{\widetilde{\mu}\gamma_i}{2}\right)\mathrm{D} + \frac{96c_1\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)}{\rho^2\widetilde{\mu}}\gamma_{t+1}^2 + \frac{16c_f}{\widetilde{\mu}\varepsilon_\infty}\gamma_{t+1},
\end{aligned}$$

where $\mathrm{D} = \|\Delta_0\|^2 + \frac{8\gamma_1 c_1}{\rho n}\left\|\beta_0 - \mathbf{1}_n\otimes\hat{\beta}_0\right\|^2$, the last inequality follows from the fact that the learning rate $\frac{\gamma_{t-1}}{\gamma_t} \leq \min\left\{\sqrt{1+(\widetilde{\mu}/4)\gamma_t^2}, \sqrt[3]{1+(\widetilde{\mu}/4)\gamma_t^3}\right\}$ and Lemma 5.1 in Appendix. Note that $\mathcal{L}_{t+1} \geq \mathbb{E}\,\|\Delta_{t+1}\|^2$,

$$\mathbb{E}\,\|\Delta_{t+1}\|^2 \leq \prod_{i=1}^{t+1}\left(1-\frac{\widetilde{\mu}\gamma_i}{2}\right)\mathrm{D} + \frac{96c_1\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)}{\rho^2\widetilde{\mu}}\gamma_{t+1}^2 + \frac{16c_f}{\widetilde{\mu}\varepsilon_\infty}\gamma_{t+1}.$$

Note also that $\gamma_0 \leq \min\left\{\sqrt{\frac{\rho^2\widetilde{\mu}}{96c_1\left(\frac{16}{\varepsilon_\infty}+3\right)c_L^2}}, \frac{\widetilde{\mu}}{c_2}\right\}$ and $\gamma_t \downarrow 0$,

$$\sup_{t\geq 0}\mathbb{E}\,\|\Delta_t\|^2 \leq \mathrm{D} + \frac{96c_1\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)}{\rho^2\widetilde{\mu}}\gamma_1^2 + \frac{16c_f}{\widetilde{\mu}\varepsilon_\infty}\gamma_1 \leq \mathrm{D}+1+\frac{16c_f}{c_2\varepsilon_\infty}. \tag{2.5}$$

Substitute (2.5) into Lemma 2.2,

$$\begin{aligned}
&\mathbb{E}\left\|\beta_{t+1}-\mathbf{1}_n\otimes\hat{\beta}_{t+1}\right\|^2 \\
\leq&(1-\frac{\rho}{2})\mathbb{E}\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2 + \frac{n}{\rho}\left[6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\bar{\Delta}+3\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\right]\gamma_{t+1}^2 \\
\leq&\left(1-\frac{\rho}{2}\right)^{t+1}\left\|\beta_0-\mathbf{1}_n\otimes\hat{\beta}_0\right\|^2 + \frac{n}{\rho}\left[6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\bar{\Delta}+3\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\right]\sum_{s=1}^{t+1}\left(1-\frac{\rho}{2}\right)^{t+1-s}\gamma_s^2 \\
\leq&\left(1-\frac{\rho}{2}\right)^{t+1}\left\|\beta_0-\mathbf{1}_n\otimes\hat{\beta}_0\right\|^2 + \frac{4n}{\rho^2}\left[6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\bar{\Delta}+3\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\right]\gamma_{t+1}^2,
\end{aligned}$$

where $\bar{\Delta} = \mathrm{D}+1+\frac{16c_f}{c_2\varepsilon_\infty}$, the last inequality follows from Lemma 5.2 in Appendix and the fact that the learning rate $\gamma_t/\gamma_{t+1} \leq 1+\rho/(4-2\rho), t\geq 0$. The proof is complete. □

Theorem 2.1 shows that the averaged parameter $\hat{\beta}_t$ converges to the optimal parameter β^* at rate $O(\gamma_t)$ and the consensus error $\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|^2$ converges to 0 at rate $O(\gamma_t^2)$, where γ_t is a diminishing learning rate of ε-DSGDCB. Theorem 2.1 paves the way to study the regret bound of ε-DSGDCB and construct the asymptotic distribution of the estimators of parameter and value.

At the end of this section, we present the regret bound of ε-DSGDCB. For simplicity, we denote $\ell_{i,a} \in [0,1]$ as the loss incurred by the selected action a for agent i.

Theorem 2.2. *Let*

$$Reg(T) := \sum_{t=1}^{T}\sum_{i=1}^{n} \ell_{i,a_{i,t}} - \sum_{t=1}^{T}\sum_{i=1}^{n} \ell_{i,a_{i,t}^*}$$

be the regret at round T, where $a_{i,t}^*$ is the optimal arm for agent i at time t. Under the conditions of Theorem 2.1 and Assumption 2.4,

$$\mathbb{E}\left[Reg(T)\right] \leq O\left(\sum_{t=1}^{T} \frac{n\sqrt{\gamma_t}}{\varepsilon_t} + n\sum_{t=1}^{T}\varepsilon_t\right).$$

Proof. Recall that $\mathcal{P}_O^r$ is the joint distribution of O_i when A_i follows $\mathcal{P}_A^r$, for any $i \in \mathcal{V}$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^r}\left[\ell_{i,a_{i,t}} - \ell_{i,a_{i,t}^*} \mid \mathcal{F}_t\right] \\
=&\underbrace{\mathbb{E}_{\mathcal{P}_O^r}\left[\ell_{i,a_{i,t}} - \ell_i\left(\beta_{i,t}; o_{i,t+1}\right) \mid \mathcal{F}_t\right]}_{\textit{Exploration error}} + \underbrace{\mathbb{E}_{\mathcal{P}_O^r}\left[\ell_i\left(\beta_{i,t}; o_{i,t+1}\right) - \ell_i\left(\beta^*; o_{i,t+1}\right) \mid \mathcal{F}_t\right]}_{\textit{Estimation error}} \\
\leq&\varepsilon_t + L_i(\beta_{i,t}) - L_i(\beta^*).
\end{aligned}$$

Next, we focus on the cumulative estimation error. Recall that $\mathcal{P}_O^\pi$ is the joint distribution of O_i under the ε-greedy policy in ε-DSGDCB,

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=1}^{n}\left[L(\beta_{i,t}) - L(\beta^*)\right] &\leq \mathbb{E}_{\mathcal{P}_O^\pi}\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\left\langle g\left(\beta_{i,t}; o_{i,t+1}\right), \beta_{i,t} - \beta^*\right\rangle \mid \mathcal{F}_t\right] \\
&\leq \mathbb{E}_{\mathcal{P}_O^\pi}\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{c_l}{\varepsilon_t}\left\|\beta_{i,t} - \beta^*\right\| \mid \mathcal{F}_t\right] \\
&\leq \mathbb{E}_{\mathcal{P}_O^\pi}\left[n\sum_{t=1}^{T}\frac{c_l}{\varepsilon_t}\left\|\Delta_t\right\| + \sum_{t=1}^{T}\frac{c_l}{\varepsilon_t}\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\| \mid \mathcal{F}_t\right] \\
&= O\left(\sum_{t=1}^{T}\frac{n\sqrt{\gamma_t}}{\varepsilon_t}\right),
\end{aligned} \tag{2.6}$$

where the first inequality follows from Assumption 2.1 (i), the second inequality follows from Assumption 2.4 and the last equality follows from Theorem 2.1. Then, we have

$$\mathbb{E}\left[Reg(T)\right] \leq O\left(\sum_{t=1}^{T} \frac{n\sqrt{\gamma_t}}{\varepsilon_t} + n\sum_{t=1}^{T}\varepsilon_t\right).$$

The proof is complete. □

When the learning rate $\gamma_t = m_0/(t+m_1)^\kappa$ with $\kappa \in \left(\frac{1}{2}, 1\right)$, $m_0, m_1 > 0$, and the exploration rate $\varepsilon_t = t^{-\kappa/4}$, Theorem 2.2 shows that ε-DSGDCB attains a regret bound of $O(T^{1-\frac{\kappa}{4}})$. Moreover, if the time horizon T is known a priori, ε-DSGDCB may achieve the regret bound of $O(T^{\frac{3}{4}})$ with the learning rate $\gamma_t = 1/T$. Recently, Arya and Sriperumbudur [3] propose a kernel ε-greedy algorithm for contextual bandits, which achieves the theoretically optimal regret bound of $O(T^{2/3})$. The proposed algorithm in [3] updates the parameter β_t through solving a sample average approximation optimization problem, where the optimal convergence rate of $\beta_t \to \beta^*$ is $O(1/t)$. On the other hand, ε-DSGDCB updates the parameters $\beta_{i,t}$ by a stochastic gradient descent iteration (2.3) and the optimal convergence rate of $\beta_{i,t} \to \beta^*$ is $O(1/\sqrt{t})$.

3 Statistical Inference

The asymptotic normality of stochastic approximation method can be traced back to the 1950s [15, 19]. Recently, many works have leveraged the asymptotic normality to construct the confidence interval of the optimal solution, such as plug-in method [13], batch-means method [13, 40] and random scaling method [24]. In this section, we employ the random scaling method to study the statistical inference of parameters for distributed contextual bandit in Section 3.1 and the plug-in method to study the statistical inference of value in Section 3.2.

3.1 Parameter Inference

In this subsection, we study the pivotal limiting distribution for the estimator of parameter and employ the random scaling method to establish an asymptotically pivotal statistic for the distributed contextual bandit. The idea of random scaling is borrowed from the time-series literature on fixed bandwidth heteroskedasticity and autocorrelation robust inference [23]. Lee et al. [24] first utilize the random scaling method for the statistical inference of the iteration points of stochastic approximation. We refer the interested readers to the works [12, 14, 25, 29] for recent advances in the statistical inference of random scaling method.

For establishing the pivotal limiting distribution for the estimator of parameter, the following conditions are needed.

Assumption 3.1. [24] *(i) The learning rate $\gamma_t = m_0/(t+m_1)^\kappa$ with $\kappa \in \left(\frac{1}{2}, 1\right)$ and $m_0, m_1 > 0$. (ii) For $\forall i \in \mathcal{V}$, there exists a constant $p \geq 1/(1-\kappa)$ such that $\mathbb{E}\left\|\nabla L_i(\hat{\beta}_t) - g_j(\hat{\beta}_t; o_{i,t+1})\right\|^{2p}$ is bounded.*

Assumption 3.1 (i) is a standard condition on the learning rate of distributed stochastic gradient descent. Assumption 3.1 (ii) is the moment condition to enhance the results for uniform convergence, which will be used to establish the functional central limit theorem.

Theorem 3.1. *Under the conditions of Theorem 2.1 and Assumption 3.1, for $\forall i \in \mathcal{V}$,*

$$\frac{\sqrt{t}w^T\left(\bar{\beta}_{i,t} - \beta^*\right)}{\sqrt{w^T R_{i,t} w}} \xrightarrow{d} W(1)\left[\int_0^1 \left(W(r) - rW(1)\right)^2 dr\right]^{-1/2}, \quad \forall w \in \mathbb{R}^p,$$

where

$$R_{i,t} = \frac{1}{t}\sum_{s=1}^{t}\left[\frac{1}{\sqrt{t}}\sum_{k=1}^{s}\left(\beta_{i,k} - \bar{\beta}_{i,t}\right)\right]\left[\frac{1}{\sqrt{t}}\sum_{k=1}^{s}\left(\beta_{i,k} - \bar{\beta}_{i,t}\right)\right]^T, \tag{3.7}$$

$\bar{\beta}_{i,t} = \frac{1}{t}\sum_{k=1}^{t}\beta_{i,k}$ *and* $W(\cdot)$ *is a standard one-dimensional Wiener processes.*

Proof. Denote

$$\Sigma = H^{-1}SH^{-1}, \quad H = \frac{1}{n}\nabla^2 L(\beta^*), \quad S = \mathrm{Cov}\left(\sum_{j=1}^{n} g_j\left(\beta^*; o_{j,k}\right)\right).$$

Next, we show that

$$\frac{1}{\sqrt{t}}\sum_{k=1}^{[tr]}(\beta_{i,k} - \beta^*) \Rightarrow \Sigma^{1/2}W(r), \quad r \in [0,1].$$

By Theorem 2.1 and the fact that $\sum_{t=0}^{\infty}\frac{\gamma_t}{\sqrt{t}} < \infty$, for $r \in [0,1]$,

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{k=0}^{[tr]}(\hat{\beta}_k - \beta^*) - \frac{1}{\sqrt{t}}\sum_{k=0}^{[tr]}(\beta_{i,k} - \beta^*)\right\|\right] \le \frac{1}{\sqrt{t}}\sum_{k=0}^{[tr]}\sqrt{\mathbb{E}\left[\left\|\beta_k - 1\otimes\hat{\beta}_k\right\|^2\right]} \le \frac{1}{\sqrt{t}}\sum_{k=0}^{[tr]}O\left(\gamma_k\right),$$

where the last term converges to 0 by Kronecker lemma. Then, we just need to verify

$$\frac{1}{\sqrt{t}}\sum_{k=0}^{[tr]}(\hat{\beta}_k - \beta^*) \Rightarrow \Sigma^{1/2}W(r), \quad r \in [0,1].$$

By the recursion (2.3) in ε-DSGDCB,

$$\begin{aligned}
&\hat{\beta}_{t+1} - \beta^* \\
=&\hat{\beta}_t - \gamma_{t+1}\hat{G}_t - \beta^* \\
=&\left(\mathbf{I}_{Kp} - \gamma_{t+1}\frac{1}{n}\nabla^2 L(\beta^*)\right)(\hat{\beta}_t - \beta^*) - \gamma_{t+1}\left(\frac{1}{n}\sum_{j=1}^{n} g_j(\beta_{j,t}; o_{j,t+1}) - \frac{1}{n}\sum_{j=1}^{n} g_j(\hat{\beta}_t; o_{j,t+1})\right) \quad (3.8)\\
&- \gamma_{t+1}\left(\frac{1}{n}\nabla L(\hat{\beta}_t) - \frac{1}{n}\nabla^2 L(\beta^*)(\hat{\beta}_t - \beta^*)\right) - \gamma_{t+1}\left(\frac{1}{n}\sum_{j=1}^{n} g_j(\hat{\beta}_t; o_{j,t+1}) - \frac{1}{n}\nabla L(\hat{\beta}_t)\right).
\end{aligned}$$

Denote

$$U = \frac{1}{n}\nabla^2 L(\beta^*), \quad \zeta_t = \frac{1}{n}\nabla L(\hat{\beta}_t) - \frac{1}{n}\sum_{j=1}^{n} g_j(\hat{\beta}_t; o_{j,t+1})$$

and

$$\eta_t = -\left(\frac{1}{n}\nabla L(\hat{\beta}_t) - \frac{1}{n}\nabla^2 L(\beta^*)(\hat{\beta}_t - \beta^*)\right) - \left(\frac{1}{n}\sum_{j=1}^{n} g_j(\beta_{j,t}; o_{j,t+1}) - \frac{1}{n}\sum_{j=1}^{n} g_j(\hat{\beta}_t; o_{j,t+1})\right).$$

Then, we may reformulate (3.8) as

$$\begin{aligned}
\Delta_{t+1} =&(\mathbf{I}_{Kp} - \gamma_{t+1}U)\Delta_t + \gamma_{t+1}(\zeta_t + \eta_t) \\
=&\prod_{j=1}^{t+1}(\mathbf{I}_{Kp} - \gamma_j U)\Delta_0 + \sum_{j=1}^{t+1}\prod_{i=j+1}^{t+1}(\mathbf{I}_{Kp} - \gamma_i U)\gamma_j(\zeta_{j-1} + \eta_{j-1}).
\end{aligned}$$

Next, for $r \in [0,1]$, we introduce a partial sum process

$$\bar{\Delta}_t(r) := \frac{1}{t}\sum_{k=0}^{[tr]} \Delta_k.$$

Then,

$$\begin{aligned}\bar{\Delta}_t(r) =& \frac{1}{t}\sum_{j=0}^{[tr]}\prod_{i=1}^{j}(\mathbf{I}_{Kp}-\gamma_i U)\Delta_0 + \frac{1}{t}\sum_{j=0}^{[tr]}\left[\sum_{k=j}^{[tr]}\prod_{i=j+1}^{k}(\mathbf{I}_{Kp}-\gamma_i U)\right]\gamma_j(\zeta_{j-1}+\eta_{j-1}) \\ =& \frac{1}{t\gamma_0}\alpha_{[tr]}\Delta_0 + \frac{1}{t}\sum_{j=1}^{[tr]} U^{-1}(\zeta_{j-1}+\eta_{j-1}) + \frac{1}{t}\sum_{j=1}^{[tr]} w_j^{[tr]}(\zeta_{j-1}+\eta_{j-1}),\end{aligned} \tag{3.9}$$

where $\alpha_j^{[tr]} = \gamma_j \sum_{k=j}^{[tr]}\prod_{i=j+1}^{k}(\mathbf{I}_{Kp}-\gamma_i U)$, $\alpha_{[tr]} = \alpha_0^{[tr]}$ and $w_j^{[tr]} = \alpha_j^{[tr]} - U^{-1}$.

By (3.9), we have

$$\sqrt{t}\bar{\Delta}_t(r) = I^{(1)}(r) + I^{(2)}(r) + I^{(3)}(r) + I^{(4)}(r) + I^{(5)}(r),$$

where

$$\begin{aligned} I^{(1)}(r) &= \frac{1}{\sqrt{t}\gamma_0}\alpha_{[tr]}\Delta_0, \qquad I^{(2)}(r) = \frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]} U^{-1}\zeta_{j-1}, \\ I^{(3)}(r) &= \frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]} U^{-1}\eta_{j-1}, \quad I^{(4)}(r) = \frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]} w_j^{[tr]}\zeta_{j-1}, \\ I^{(5)}(r) &= \frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]} w_j^{[tr]}\eta_{j-1}. \end{aligned}$$

Note that $\|\alpha_{[tr]}\| \le C$ and $\frac{1}{t}\sum_{j=1}^{t}\|w_j^t\| \to 0$ by [31, Lemma 1], we have $\sup_r \|I^{(1)}(r)\| = o_p(1)$. By Assumption 2.1 (ii), the Lipschitz continuity of $\nabla\ell_j(\cdot, O_j)$ and Theorem 2.1,

$$\begin{aligned}\mathbb{E}\left[\|\eta_t\|\right] =& \mathbb{E}\left[\left\| -\left(\frac{1}{n}\nabla L(\hat{\beta}_t) - \frac{1}{n}\nabla^2 L(\beta^*)(\hat{\beta}_t - \beta^*)\right)\right.\right. \\ & \left.\left. - \left(\frac{1}{n}\sum_{j=1}^{n} g_j(\beta_{j,t}; o_{j,t+1}) - \frac{1}{n}\sum_{j=1}^{n} g_j(\hat{\beta}_t; o_{j,t+1})\right)\right\|\right] \\ \le& \frac{c}{n}\mathbb{E}\left[\left\|\hat{\beta}_t - \beta^*\right\|^2\right] + \frac{c_L}{n\varepsilon_\infty}\mathbb{E}\left[\left\|\beta_t - \mathbf{1}_n \otimes \hat{\beta}_t\right\|\right] = O\left(\gamma_t\right).\end{aligned} \tag{3.10}$$

Then, we have $\mathbb{E}\left[\sup_r \left\|\frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]}\eta_j\right\|\right] = o(1)$, which implies $\mathbb{E}\left[\sup_r \|I^{(3)}(r)\|\right] = o(1)$. By the similar analysis in the proof of [24, Theorem 1, Page 7], we have $\mathbb{E}\left[\sup_r \left\|I^{(4)}(r)\right\|^p\right] = o(1)$. Moreover, the boundedness of $\left\|w_j^{[tr]}\right\|$ and (3.10) imply that $\mathbb{E}\left[\sup_r \left\|I^{(5)}(r)\right\|\right] = o(1)$.

We establish the asymptotic properties of $I^{(2)}(r)$ by martingale functional central limit theorem [37, Theorem 2.3.9]. We decompose the martingale difference sequence ζ_t into the

following two parts,

$$\zeta_t^{(1)} := \frac{1}{n}\sum_{j=1}^{n} g_j(\beta^*; o_{j,t+1}),\ \zeta_t^{(2)} := \frac{1}{n}\sum_{j=1}^{n} g_j(\hat{\beta}_t; o_{j,t+1}) - \frac{1}{n}\sum_{j=1}^{n} g_j(\beta^*; o_{j,t+1}) - \frac{1}{n}\nabla L(\hat{\beta}_t).$$

By the definition of $\zeta_t^{(2)}$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left[g_i(\hat{\beta}_t; o_{i,t+1}) - g_i(\beta^*; o_{i,t+1}) + \nabla L_i(\beta^*) - \nabla L_i(\hat{\beta}_t)\right]\right\|^2 \mid \mathcal{F}_t\right] \\
\leq &\mathbb{E}_{\mathcal{P}_O^\pi}\left[\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{\left\|\nabla \ell_i\left(\hat{\beta}_t; o_{i,t+1}\right) - \nabla \ell_i\left(\beta^*; o_{i,t+1}\right)\right\|^2 1_{\{a_{i,t+1}=j\}}}{K^2\left\{\pi_{i,t}\left(x_{i,t+1}\right)\right\}^2} \mid \mathcal{F}_t\right] + \frac{2}{n}\sum_{i=1}^{n}\left\|\nabla L_i\left(\hat{\beta}_t\right) - \nabla L_i\left(\beta^*\right)\right\|^2 \\
\leq &\frac{1}{\varepsilon_\infty}\mathbb{E}_{\mathcal{P}_O^r}\left[\frac{2}{n}\sum_{i=1}^{n}\left\|\nabla \ell_i\left(\hat{\beta}_t; o_{i,t+1}\right) - \nabla \ell_i\left(\beta^*; o_{i,t+1}\right)\right\|^2\right] + \frac{2}{n}\sum_{i=1}^{n}\left\|\nabla L_i\left(\hat{\beta}_t\right) - \nabla L_i\left(\beta^*\right)\right\|^2.
\end{aligned}$$

By Assumption 2.2 (i) and Theorem 2.1, we have

$$\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\zeta_t^{(2)}\right\|^2 \mid \mathcal{F}_t\right] \leq \frac{4c_L^2}{\varepsilon_\infty}\|\hat{\beta}_t - \beta^*\|^2 \to 0.$$

Next, we focus on martingale difference sequence $\zeta_t^{(1)}$. Denote

$$\xi_{t,k} = \frac{\zeta_k^{(1)}}{\sqrt{t}}, \quad S_{t,k} = \mathbb{E}\left[\xi_{t,k}\xi_{t,k}^T \mid \mathcal{F}_{t,k-1}\right], \quad S_{[tr]} = \sum_{k=1}^{[tr]} S_{t,k}, \quad r \in [0,1],$$

where the filtration $\mathcal{F}_{t,k} := \sigma\{\xi_{t,k} : 1 \leq k \leq t\}$. Note that $\{\zeta_k^{(1)}\}$ is a martingale difference sequence, $\mathbb{E}\left[\xi_{t,k} \mid \mathcal{F}_{t,k-1}\right] = 0$.

By Assumption 2.2 (ii), we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\|\xi_{t,k}\|^2 \mid \mathcal{F}_{t,k-1}\right] \\
=&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{\frac{1}{n}\sum_{i=1}^{n} g_i(\beta^*; o_{i,k+1})}{\sqrt{t}}\right\|^2 \mid \mathcal{F}_{t,k-1}\right] \\
\leq &\mathbb{E}_{\mathcal{P}_O^\pi}\left[\frac{1}{nt}\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{\|\nabla \ell_i\left(\beta^*; o_{i,k+1}\right)\|^2 1_{\{a_{i,k+1}=j\}}}{K^2\left\{\pi_{i,k}\left(x_{i,k+1}\right)\right\}^2} \mid \mathcal{F}_{t,k-1}\right] \\
=&\mathbb{E}\left[\frac{1}{nt}\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{\mathbb{E}\left\{\|\nabla \ell_i\left(\beta^*; x_{i,k+1}, j, y_{i,k+1}\right)\|^2 \mid \mathcal{F}_{t,k-1}, x_{i,k+1}\right\}}{K^2\pi_{i,k}\left(x_{i,k+1}\right)} \mid \mathcal{F}_{t,k-1}\right] \\
\leq &\frac{1}{\varepsilon_\infty}\mathbb{E}_{\mathcal{P}_O^r}\left[\frac{1}{nt}\sum_{i=1}^{n}\|\nabla \ell_i\left(\beta^*; o_{i,k+1}\right)\|^2\right] \leq \frac{c_f}{t\varepsilon_\infty}.
\end{aligned}$$

Then,

$$\lim_{t\to\infty}\sum_{k=0}^{t-1}\mathbb{E}_{\mathcal{P}_O^\pi}\left[\|\xi_{t,k}\|^2 \mid \mathcal{F}_{t,k-1}\right] \leq \lim_{t\to\infty}\frac{tc_f}{t\varepsilon_\infty} = \frac{c_f}{\varepsilon_\infty}.$$

Note that $S_{t,k} = \frac{1}{tn^2} \operatorname{Cov}\left(\sum_{j=1}^{n} g_j\left(\beta^*; o_{j,k+1}\right)\right) = \frac{1}{tn^2} S$, for $r \in [0,1]$,

$$S_{[tr]} = \sum_{k=1}^{[tr]} S_{t,k} \Rightarrow \frac{r}{n^2} \operatorname{Cov}\left(\sum_{j=1}^{n} g_j\left(\beta^*; o_{j,k+1}\right)\right),$$

which implies condition (3.24) in [37, Theorem 2.3.9]. Then, we just need to verify the Lindeberg condition. For any $\delta > 0$, by the Hölder inequality,

$$\begin{aligned}
\mathbb{E}\left[\left\|\xi_{t,k}\right\|^2 1_{\left\{\left\|\xi_{t,k}\right\| \geq \delta\right\}}\right] &\leq \left(\mathbb{E}\left[\left\|\xi_{t,k}\right\|^{2(p/2)}\right]\right)^{2/p} \left(\mathbb{E}\left[1^q_{\left\{\left\|\xi_{t,k}\right\| > \delta\right\}}\right]\right)^{1/q} \\
&= \left(\mathbb{E}\left[\left\|\xi_{t,k}\right\|^p\right]\right)^{2/p} P^{1/q}\left(\left\|\xi_{t,k}\right\| > \delta\right) \\
&\leq \left(\mathbb{E}\left[\left\|\xi_{t,k}\right\|^p\right]\right)^{2/p} \left(\frac{\mathbb{E}\left[\left\|\xi_{t,k}\right\|\right]}{\delta}\right)^{1/q} \\
&\leq \frac{c_f^{1+1/(2q)}}{t^{1+1/(2q)} \varepsilon_\infty^{1+1/(2q)} \delta^{1/q}},
\end{aligned}$$

where $p > 2$ and $q > 0$ such that $2/p + 1/q = 1$, the second inequality follows from Markov inequality, the third inequality follows from

$$\mathbb{E}\left[\left\|\xi_{t,k}\right\|^p\right] = \mathbb{E}\left[\left\|\frac{\frac{1}{n}\sum_{j=1}^{n} g_j\left(\beta^*; o_{j,k}\right)}{\sqrt{t}}\right\|^p\right] \leq \frac{\sum_{j=1}^{n} \mathbb{E}\left[\left\|g_j\left(\beta^*; o_{j,k}\right)\right\|^p\right]}{nt^{p/2}} \leq \frac{c_f^{p/2}}{t^{p/2} \varepsilon_\infty^{p/2}}.$$

Then, for $r \in [0,1]$,

$$\lim_{t\to\infty} \sum_{k=1}^{[tr]} \mathbb{E}\left[\left\|\xi_{t,k}\right\|^2 1_{\left\{\left\|\xi_{t,k}\right\| \geq \delta\right\}}\right] \leq \lim_{t\to\infty} \frac{trc^{1+1/(2q)}}{t^{1+1/(2q)} \varepsilon_\infty^{1+1/(2q)} \delta^{1/q}} = 0.$$

Therefore, all the conditions of [37, Theorem 2.3.9] hold and we have

$$I^{(2)}(r) = \frac{1}{\sqrt{t}} \sum_{j=1}^{[tr]} U^{-1} \zeta_j \Rightarrow \Sigma^{1/2} W(r), \quad r \in [0,1].$$

Summarizing above, for any $i \in \mathcal{V}$,

$$\frac{1}{\sqrt{t}} \sum_{k=1}^{[tr]} \left(\beta_{i,k} - \beta^*\right) \Rightarrow \Sigma^{1/2} W(r), \quad r \in [0,1]. \tag{3.11}$$

Denote $C_t(r) := w^T \frac{1}{\sqrt{t}} \sum_{k=1}^{[tr]} \left(\beta_{i,k} - \beta^*\right)$. Then, for $\forall w \in \mathbb{R}^p$,

$$C_t(r) \Rightarrow \left(w^T \Sigma w\right)^{1/2} W(r), \quad r \in [0,1].$$

Furthermore, $w^T\left(\bar{\beta}_{i,t} - \beta^*\right) = \frac{1}{\sqrt{t}} C_t(1)$ and $w^T R_{i,t} w = \frac{1}{t}\sum_{s=1}^{t}\left[C_t\left(\frac{s}{t}\right) - \frac{s}{t} C_t(1)\right]^2$. By continuous mapping theorem,

$$\frac{\sqrt{t} w^T\left(\bar{\beta}_{i,t} - \beta^*\right)}{\sqrt{w^T R_{i,t} w}} \xrightarrow{d} W(1)\left[\int_0^1 \left(W(r) - rW(1)\right)^2 dr\right]^{-1/2}.$$

The proof is complete. □

Theorem 3.1 establishes the asymptotic distribution of parameters of distributed contextual bandit framework as a stochastic process and constructs an asymptotically pivotal statistic by the random scaling method. Compared with the previous works on statistical inference of contextual bandit problem [9, 10, 11, 21], the random scaling method does not need to estimate the asymptotic variance but studentize $\sqrt{t}\left(\bar{\beta}_{i,t}-\beta^{*}\right)$ via the following random scaling matrix

$$R_{i,t}=\frac{1}{t}\sum_{s=1}^{t}\left[\frac{1}{\sqrt{t}}\sum_{k=1}^{s}\left(\beta_{i,k}-\bar{\beta}_{i,t}\right)\right]\left[\frac{1}{\sqrt{t}}\sum_{k=1}^{s}\left(\beta_{i,k}-\bar{\beta}_{i,t}\right)\right]^{T},$$

which can be updated online as follows,

$$\begin{aligned}R_{i,t}&=\frac{1}{t^{2}}\sum_{s=1}^{t}s^{2}\left(\bar{\beta}_{i,s}-\bar{\beta}_{i,t}\right)\left(\bar{\beta}_{i,s}-\bar{\beta}_{i,t}\right)^{T}\\&=\frac{1}{t^{2}}\sum_{s=1}^{t}s^{2}\bar{\beta}_{i,s}\bar{\beta}_{i,s}^{T}-\frac{1}{t^{2}}\bar{\beta}_{i,t}\sum_{s=1}^{t}s^{2}\bar{\beta}_{i,s}^{T}-\frac{1}{t^{2}}\left(\sum_{s=1}^{t}s^{2}\bar{\beta}_{i,s}\right)\bar{\beta}_{i,t}^{T}+\frac{1}{t^{2}}\sum_{s=1}^{t}s^{2}\bar{\beta}_{i,t}\bar{\beta}_{i,t}^{T},\end{aligned}$$

where $\bar{\beta}_{i,s}=\frac{1}{s}\sum_{k=1}^{s}\beta_{i,k}$.

3.2 Value Inference

In this subsection, we study the statistical inference of the value of the optimal policy for distributed contextual bandit problems.

For any $i\in\mathcal{V}$, recall the optimal policy $d^{*}(X_i)=\arg\max_{a\in\mathcal{A}}\ \mu_i\left(a,X_i;\beta^{*}\right)$, then the corresponding expected value is

$$V_i=\mathbb{E}\left[\mathbb{E}\left[Y_i\mid d^{*}(X_i),X_i\right]\right]=\int\mu_i\left(d^{*}(X_i),X_i;\beta^{*}\right)d\mathcal{P}_{X_i}$$

and the cumulative expected value is

$$V=\sum_{i=1}^{n}V_i=\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[Y_i\mid d^{*}(X_i),X_i\right]\right]=\sum_{i=1}^{n}\int\mu_i\left(d^{*}(X_i),X_i;\beta^{*}\right)d\mathcal{P}_{X_i}.$$

Following the idea outlined in [10], we use a local inverse probability weighting estimator

$$V_{i,t}=\frac{1}{t}\sum_{s=1}^{t}w_{i,s}y_{i,s},\quad\forall i\in\mathcal{V}\tag{3.12}$$

for agent i's expected value V_i, and a distributed variant of agent i's inverse probability weighted value estimator

$$\hat{V}_{i,t}=\frac{t-1}{t}\sum_{j=1}^{n}m_{ij}\hat{V}_{j,t-1}+\frac{n}{t}w_{i,t}y_{i,t},\quad\forall i\in\mathcal{V}\tag{3.13}$$

for cumulative expected value V, where

$$w_{i,t}=\frac{1_{\left\{a_{i,t}=d_{t-1}^{*}(x_{i,t})\right\}}}{P\left(1_{\left\{a_{i,t}=d_{t-1}^{*}(x_{i,t})\right\}}\mid\mathcal{F}_{t-1},x_{i,t}\right)}=\begin{cases}\frac{1}{1+\frac{\varepsilon_{t-1}}{K}-\varepsilon_{t-1}}, & \text{if } a_{i,t}=d_{t-1}^{*}\left(x_{i,t}\right),\\0, & \text{otherwise,}\end{cases}$$

and $d_{t-1}^{*}\left(x_{i,t}\right) = \text{argmax}_{a\in\mathcal{A}}\ \mu_i\left(a, x_{i,t}; \bar{\beta}_{i,t-1}\right)$.

In (3.12), the inverse probability weighting method corrects the bias of reward induced by the ε-greedy policy, where we select the weighted largest reward $w_{i,t}y_{i,t}$ when $a_{i,t}$ is the estimated optimal decision $d_{t-1}^{*}\left(x_{i,t}\right)$, and 0 otherwise. Notice that each agent only has access to local information in distributed bandit setting, the cumulative expected value is estimated by exchanging the agent's local estimates of this value with neighbors across a communication network in (3.13).

The following conditions are made to establish the asymptotic normality for the estimator of value.

Assumption 3.2. [10] *For $\forall i \in \mathcal{V}$, (i) the features vector X_i satisfies $\mathbb{E}\|X_i\| < \infty$, the second moment of reward exists for any given features and action, that is,*

$$\theta_i^2(A_i, X_i) = \mathbb{E}\left[Y_i^2 \mid A_i, X_i\right] < \infty;$$

(ii) there exists a constant $C \geq 0$ such that for all a_i and a_j in $\mathcal{A}$ where $a_i \neq a_j$,

$$P\left\{0 < \left|\mu_i\left(a_i, X_i; \beta^*\right) - \mu_i\left(a_j, X_i; \beta^*\right)\right| \leq \kappa\right\} \leq C\kappa \text{ for all } \kappa \in \mathbb{R}^+.$$

Assumption 3.2 (i) guarantees the boundedness of the observed data. Assumption 3.2 (ii) is a margin condition widely adopted by contextual bandit literature [3, 5, 9, 10], which ensures a sufficient gap between the rewards for different arms.

Theorem 3.2. *Under the conditions of Theorem 2.1 and Assumption 3.2, for $\forall i \in \mathcal{V}$,*

$$\sqrt{t}\left\{V_{i,t} - V_i\right\} \xrightarrow{d} \mathcal{N}\left(0, \eta_i^2\right) \tag{3.14}$$

and

$$\eta_{i,t}^2 \xrightarrow{d} \eta_i^2,$$

where

$$\eta_i^2 = \frac{K}{K - (K-1)\varepsilon_\infty} \int \theta_i^2\left(d^*(X_i), X_i\right) d\mathcal{P}_{X_i} - V_i^2$$

and

$$\eta_{i,t}^2 = \frac{K}{K - (K-1)\varepsilon_t} \frac{1}{t} \sum_{s=1}^{t} w_{i,s} y_{i,s}^2 - V_{i,t}^2. \tag{3.15}$$

Proof. The proof is similar to the proof of [10, Theorem 4.1]. □

Theorem 3.2 shows the asymptotic normality of the local value estimator of the optimal policy (3.14) and the efficiency of plug-in method (3.15) for each agent.

Next, we move to study the statistical inference of the cumulative value of the optimal policy for distributed contextual bandit.

Theorem 3.3. *Under the conditions of Theorem 2.1 and Assumption 3.2, for $\forall i \in \mathcal{V}$,*

$$\sqrt{t}\left(\hat{V}_{i,t} - V\right) \xrightarrow{d} \mathcal{N}\left(0, \eta^2\right) \tag{3.16}$$

and

$$\hat{\eta}_{i,t}^2 \xrightarrow{d} \eta^2,$$

where

$$\eta^2 = \sum_{i=1}^{n} \left[\frac{K}{K-(K-1)\varepsilon_\infty} \int \theta_i^2 \left(d^*(X_i), X_i\right) d\mathcal{P}_{X_i} - V_i^2 \right],$$

$$\hat{\eta}_{i,t}^2 = \frac{K}{K-(K-1)\varepsilon_t} \hat{\eta}_{i,t}^{(1)} - \hat{\eta}_{i,t}^{(2)}, \tag{3.17}$$

$$\hat{\eta}_{i,t}^{(1)} = \frac{t-1}{t} \sum_{j=1}^{n} m_{ij} \hat{\eta}_{j,t-1}^{(1)} + \frac{n}{t} w_{i,t} y_{i,t}^2 \tag{3.18}$$

and

$$\hat{\eta}_{i,t}^{(2)} = \sum_{j=1}^{n} m_{ij} \hat{\eta}_{j,t-1}^{(2)} + \left(\frac{1}{t} \sum_{s=1}^{t} w_{i,s} y_{i,s} \right)^2 - \left(\frac{1}{t-1} \sum_{s=1}^{t-1} w_{i,s} y_{i,s} \right)^2. \tag{3.19}$$

Proof. We finish the proof in two steps: (i) Establish the asymptotic normality of $\sqrt{t}\left(\hat{V}_{i,t} - V\right)$. (ii) Show the consistency of distributed variant of plug-in estimator $\hat{\eta}_{i,t}^2$.

Step (i). Denote $\bar{V}_t = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_{i,t}$, then

$$\left\| \sqrt{t}\left(\hat{V}_{i,t} - V\right) \right\| \leq \left\| \sqrt{t}\left(\hat{V}_{i,t} - \bar{V}_t\right) \right\| + \left\| \sqrt{t}\left(\bar{V}_t - V\right) \right\|.$$

Next, we focus on the convergence rate of $\left\|\hat{V}_{i,t} - \bar{V}_t\right\|$. Denote

$$\hat{V}_t := \left[\hat{V}_{1,t}, \hat{V}_{2,t}, \cdots, \hat{V}_{n,t}\right]^T,$$

(3.13) can be rewritten compactly as

$$\hat{V}_t = \frac{t-1}{t} \mathbf{M} \hat{V}_{t-1} + \frac{1}{t} W_t,$$

where $W_t = n\left[w_{1,t}y_{1,t}, w_{2,t}y_{2,t}, \ldots, w_{n,t}y_{n,t}\right]^T$.

Then, it is sufficient to study the convergence rate of $\left\|\hat{V}_t - \mathbf{1}_n \bar{V}_t\right\|$.

Let $w_{max} = \max_{1\leq i\leq n} \mathbb{E}\left[w_i\right]$. By the recursion of $\hat{V}_t$ and the definition of $\bar{V}_t$,

$$\begin{aligned}
\left\|\hat{V}_t - \mathbf{1}_n \bar{V}_t\right\| =& \left\| \frac{t-1}{t} \mathbf{M} \hat{V}_{t-1} + \frac{1}{t} W_t - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \left(\frac{t-1}{t} \mathbf{M} \hat{V}_{t-1} + \frac{1}{t} W_t \right) \right\| \\
=& \left\| \frac{t-1}{t} \left(\mathbf{M} - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right) \left(\hat{V}_{t-1} - \mathbf{1}_n \bar{V}_{t-1} \right) + \frac{1}{t} \left(\mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right) W_{t-1} \right\| \\
\leq& \frac{t-1}{t} \left\| \mathbf{M} - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right\| \left\| \hat{V}_{t-1} - \mathbf{1}_n \bar{V}_{t-1} \right\| + \frac{1}{t} \left\| \mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right\| \|W_{t-1}\| \\
\leq& \frac{t-1}{t} (1-\rho) \left\| \hat{V}_{t-1} - \mathbf{1}_n \bar{V}_{t-1} \right\| + \frac{1}{t} \left\| \mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right\| n^{\frac{3}{2}} w_{max},
\end{aligned} \tag{3.20}$$

where the last inequality follows from Assumption 2.3. Then, we have

$$\begin{aligned}\left\|\hat{V}_t - \mathbf{1}_n\bar{V}_t\right\| \leq & \frac{1}{t}\left((1-\rho)^t\left\|\hat{V}_0 - \mathbf{1}_n\bar{V}_0\right\| + \sum_{s=1}^{t}(1-\rho)^{t-s}\left\|\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n}{n}\right\| n^{\frac{3}{2}}w_{max}\right)\\ = & \frac{1}{t}\sum_{s=1}^{t}(1-\rho)^{t-s}\left\|\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n}{n}\right\| n^{\frac{3}{2}}w_{max} \qquad (3.21)\\ \leq & \frac{1}{t}\frac{\left\|\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n}{n}\right\| n^{\frac{3}{2}}w_{max}}{\rho}.\end{aligned}$$

Note the boundedness of $\frac{\left\|\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n}{n}\right\| n^{\frac{3}{2}}w_{max}}{\rho}$, $\sqrt{t}\left\|\hat{V}_t - \mathbf{1}_n\bar{V}_t\right\| = O\left(\frac{1}{\sqrt{t}}\right)$.

Next, we focus on the asymptotic normality of $\sqrt{t}\left(\bar{V}_t - V\right)$.

By the definition of $\bar{V}_t$ and recursion in (3.19),

$$\begin{aligned}\bar{V}_t = \frac{\mathbf{1}_n^T}{n}\hat{V}_t &= \frac{\mathbf{1}_n^T}{n}\left(\frac{t-1}{t}\mathbf{M}\hat{V}_{t-1} + \frac{1}{t}W_t\right)\\ &= \frac{t-1}{t}\bar{V}_{t-1} + \frac{1}{t}\sum_{i=1}^{n}w_{i,t}y_{i,t}\\ &= \frac{1}{t}\sum_{t=1}^{t}\sum_{i=1}^{n}w_{i,t}y_{i,t}.\end{aligned}$$

The rest proof is similar to the proof of centralized asymptotic normality in [10, Theorem 4.1].

Step (ii). We just study (3.18) as the proof of (3.19) is similar.

Denote $\bar{\eta}_t^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{i,t}^{(1)}$, then

$$\left\|\hat{\eta}_{i,t}^{(1)} - \sum_{i=1}^{n}\int\theta_i^2\left(d^*(X_i), X_i\right)d\mathcal{P}_{X_i}\right\| \leq \left\|\hat{\eta}_{i,t}^{(1)} - \bar{\eta}_t^{(1)}\right\| + \left\|\bar{\eta}_t^{(1)} - \sum_{i=1}^{n}\int\theta_i^2\left(d^*(X_i), X_i\right)d\mathcal{P}_{X_i}\right\|.$$

Denote $\hat{\eta}_t^{(1)} = \left[\hat{\eta}_{1,t}^{(1)}, \hat{\eta}_{2,t}^{(1)}, \cdots, \hat{\eta}_{n,t}^{(1)}\right]^T$, (3.18) can be rewritten compactly as

$$\hat{\eta}_t^{(1)} = \frac{t-1}{t}\mathbf{M}\hat{\eta}_{t-1}^{(1)} + \frac{1}{t}H_t,$$

where $H_t = n\left[w_{1,t}y_{1,t}^2, w_{2,t}y_{2,t}^2, \ldots, w_{n,t}y_{n,t}^2\right]^T$. Then, it is sufficient to study the convergence rate of $\left\|\hat{\eta}_t^{(1)} - \mathbf{1}_n\bar{\eta}_t^{(1)}\right\|$.

By the similar analysis on (3.20)-(3.21),

$$\begin{aligned}\left\|\hat{\eta}_t^{(1)} - \mathbf{1}_n\bar{\eta}_t^{(1)}\right\| &= \left\|\frac{t-1}{t}\mathbf{M}\hat{\eta}_{t-1}^{(1)} + \frac{1}{t}H_t - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\left(\frac{t-1}{t}\mathbf{M}\hat{\eta}_{t-1}^{(1)} + \frac{1}{t}H_t\right)\right\|\\ &\leq \frac{t-1}{t}(1-\rho)\left\|\hat{\eta}_{t-1}^{(1)} - \mathbf{1}_n\bar{\eta}_{t-1}^{(1)}\right\| + \frac{1}{t}\left\|\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\right\| n^{\frac{3}{2}}w_{max}\\ &\leq \frac{1}{t}\frac{\left\|\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n}{n}\right\| n^{\frac{3}{2}}w_{max}}{\rho} = O\left(\frac{1}{t}\right).\end{aligned}$$

Next, we show that $\left\|\bar{\eta}_t^{(1)} - \sum_{i=1}^n \int \theta_i^2 \left(d^*(X_i), X_i\right) d\mathcal{P}_{X_i}\right\| \xrightarrow{d} 0$. By the definition of $\bar{\eta}_t^{(1)}$ and recursion in (3.18),

$$\begin{aligned}\bar{\eta}_t^{(1)} = \frac{\mathbf{1}_n^T}{n}\hat{\eta}_t^{(1)} &= \frac{\mathbf{1}_n^T}{n}\left(\frac{t-1}{t}\mathbf{M}\hat{\eta}_{t-1}^{(1)} + \frac{1}{t}H_t\right) \\ &= \frac{t-1}{t}\bar{\eta}_{t-1}^{(1)} + \frac{1}{t}\sum_{i=1}^n w_{i,t}y_{i,t}^2 \\ &= \frac{1}{t}\sum_{t=1}^t\sum_{i=1}^n w_{i,t}y_{i,t}^2.\end{aligned}$$

The rest proof of the consistency of $\bar{\eta}_t^{(1)}$ is similar to the proof of [10, Theorem 4.2]. □

Compared with Theorem 3.2, Theorem 3.3 establishes the asymptotic normality of the value estimator for the cumulative expected value (3.16) and presents the consistency of the asymptotic variance estimator by the distributed variant of the plug-in method (3.17). Notice that individual agent is only associated with local private value, Theorem 3.3 needs to communicate the weighted value to neighbors over the communication network.

4 Numerical Results

In this section, we report some preliminary numerical results on the individual confidence intervals of the parameters and value for the distributed contextual bandit problems.

Under Theorem 3.1, when $w = e_j$ for $j = 1, 2, ..., Kp$, we have the following asymptotically pivotal distribution in each coordinate j,

$$\frac{\sqrt{t}\left(\bar{\beta}_{i,t[j]} - \beta_{[j]}^*\right)}{\sqrt{R_{i,t[jj]}}} \xrightarrow{d} W(1)\left[\int_0^1 \left(W(r) - rW(1)\right)^2 dr\right]^{-1/2}, \quad \forall i \in \mathcal{V},$$

and then the approximate $(1-\alpha/2)$ confidence interval for the j-th component of the parameter β^* is

$$\left\{\beta_{[j]}^* : \bar{\beta}_{i,t[j]} - c_{1-\alpha/2}\sqrt{\frac{R_{i,t[j,j]}}{t}} \le \beta_{[j]}^* \le \bar{\beta}_{i,t[j]} + c_{1-\alpha/2}\sqrt{\frac{R_{i,t[j,j]}}{t}}\right\},$$

where $\bar{\beta}_{i,t[j]}$ and $R_{i,t[j,j]}$ are the j-th and (j,j)-th components of $\bar{\beta}_{i,t}$ and $R_{i,t}$ respectively, and the critical value $c(1-\alpha/2)$ satisfies $P\left(U \le c_{1-\alpha/2}\right) = 1 - \alpha/2$ for the mixed normal random variable U. Additionally, we set significance level of $\alpha = 0.05$ and the corresponding critical value $c_{1-\alpha/2} = 6.747$ [1, Table I].

Similarly, the approximate $(1-\alpha/2)$ confidence intervals for the local value V_i and the cumulative value V under Theorems 3.2 and 3.3 are

$$\left\{V_i : V_{i,t} - t^{-\frac{1}{2}} z_{\alpha/2}\eta_{i,t} \le V_i \le V_{i,t} + t^{-\frac{1}{2}} z_{\alpha/2}\eta_{i,t}\right\}$$

and

$$\left\{V : \hat{V}_{i,t} - t^{-\frac{1}{2}} z_{\alpha/2}\hat{\eta}_{i,t} \le V \le \hat{V}_{i,t} + t^{-\frac{1}{2}} z_{\alpha/2}\hat{\eta}_{i,t}\right\}$$

respectively, where $z_{\alpha/2}$ satisfies $P\left(U \leq z_{\alpha/2}\right) = 1-\alpha/2$ for the standard normal random variable U.

Next, we report the empirical performance of the proposed methods on synthetic data in Section 4.1 and the warfarin dosing problem using a real patient dataset in Section 4.2, respectively.

4.1 Synthetic Data

Consider a distributed contextual bandit problem with a linear regression loss function,

$$\min_{\beta} \sum_{i=1}^{n} \left(\mathbb{E}\left[Y_i - \mu_i(A_i, X_i; \beta_i) \right]^2 \right),$$

where the number of agents is $n = 20$, the agents make decisions in a binary action space with linear reward models

$$\mu_i(A_i, X_i; \beta_i) = (1 - A_i) X_i^T \beta_{i[1:3]} + A_i X_i^T \beta_{i[4:6]}.$$

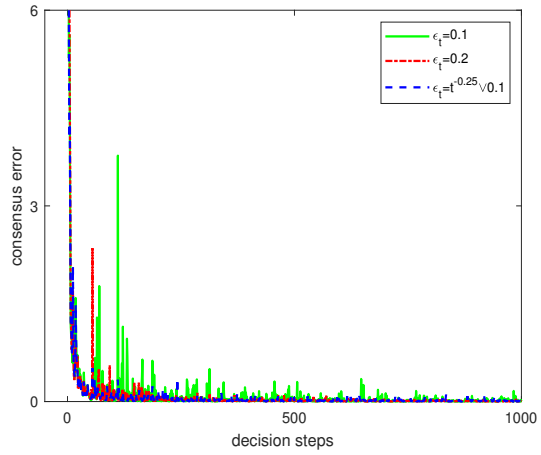
Here, $X = (1, X_2, X_3)$ is the random input vector, X_2 and X_3 independently follow the standard normal distribution, $Y = \mu(A, X; \beta^*) + E$ is the response variable with random error $E \sim \mathcal{N}(0, 1)$ and the true parameter $\beta^* = (0.5, 1.8, -2.2, 0.8, 0.7, -1.3)^T$ is randomly generated. We generate an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by adding random links to a ring network, where a link exists between any two nonadjacent nodes with a probability $prob > 0.7$ [32]. The corresponding weights m_{ji} are defined by the Metropolis rule [35]:

$$m_{ji} = \begin{cases} \frac{1}{\max\{n_i, n_j\}+1} & j \in \mathcal{N}_i, \\ 1 - \sum_{j \in \mathcal{N}_i} m_{ji} & j = i, \\ 0 & \text{otherwise}, \end{cases}$$

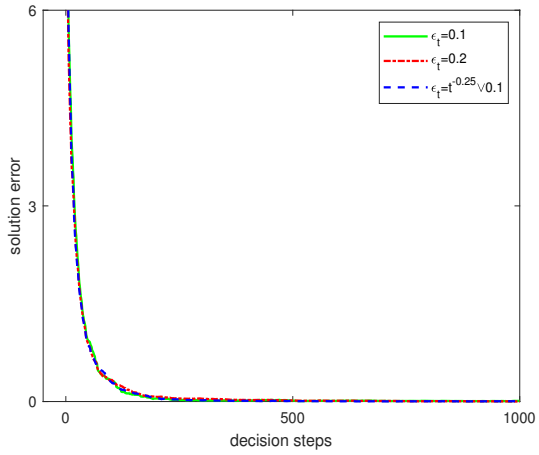
where $\mathcal{N}_i := \{j \mid (i, j) \in \mathcal{V}, j \neq i\}$ is agent i 's set of neighbors, $n_i := |\mathcal{N}_i|$ is the cardinality of $\mathcal{N}_i$. In this experiment, the learning rate is $\alpha_t = 0.1/(t+1)^{0.505}$ and the initial point is 0_{120}, where $0_{120} \in \mathbb{R}^{120}$ is the vector of all 0s. Moreover, we set the first 50 steps as the initial random exploration time.

We record the convergence of ε-DSGDCB in Figure 1. We can observe from Figures 1 (a) and (b) that the consensus error $\left\| \beta_t - \mathbf{1}_n \otimes \hat{\beta}_t \right\|^2$ and the solution error $\|\Delta_t\|^2$ converge to 0 with the increase of decision steps t, which implies the convergence of the parameter estimators in Theorem 2.1. Figure 1 (c) presents the convergence of the local value estimators for Agent 1[2], which verifies the result of Theorem 3.2. Figures 1 (d) and (e) illustrate the convergence of the cumulative value estimators, which verify the result of Theorem 3.3. Moreover, Figure 1 (e) depicts the value estimator of the cumulative expected value for Agent 1 almost coincides with the average of all agents as t increases, which indicates the agreement of the cumulative value estimators. Furthermore, we may conclude the robustness of ε-DSGDCB with respect to the exploration rates as the parameters and values exhibit less variation relative to the exploration rate with increasing t in Figures 1 (a)-(d).

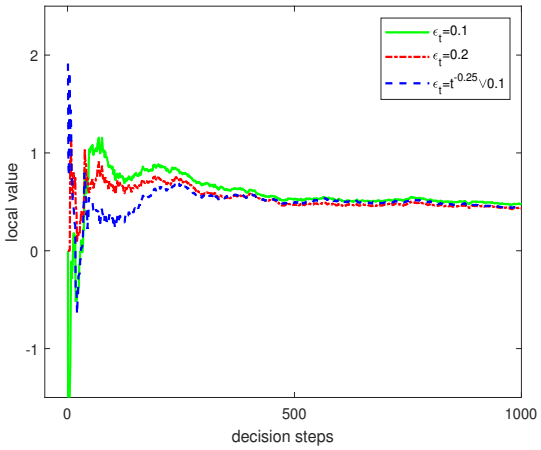
[2]Without loss of generality, we present the results of Agent 1 as representative of the entire agent group.



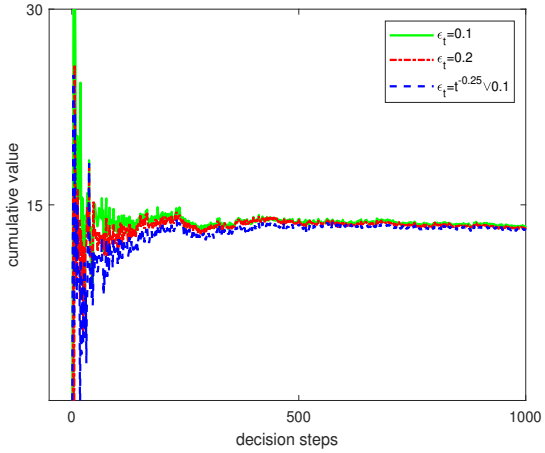
(a) Consensus error



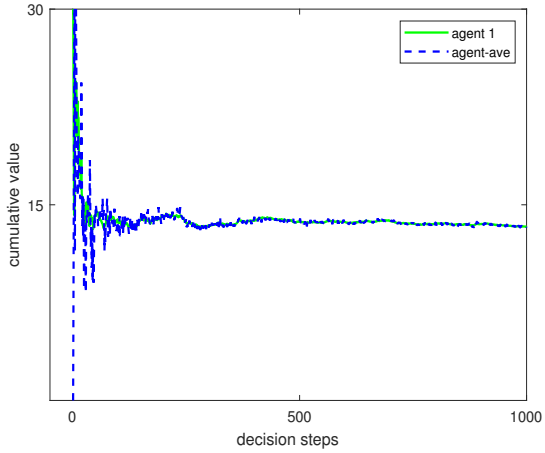
(b) Solution error



(c) Local value estimator for Agent 1

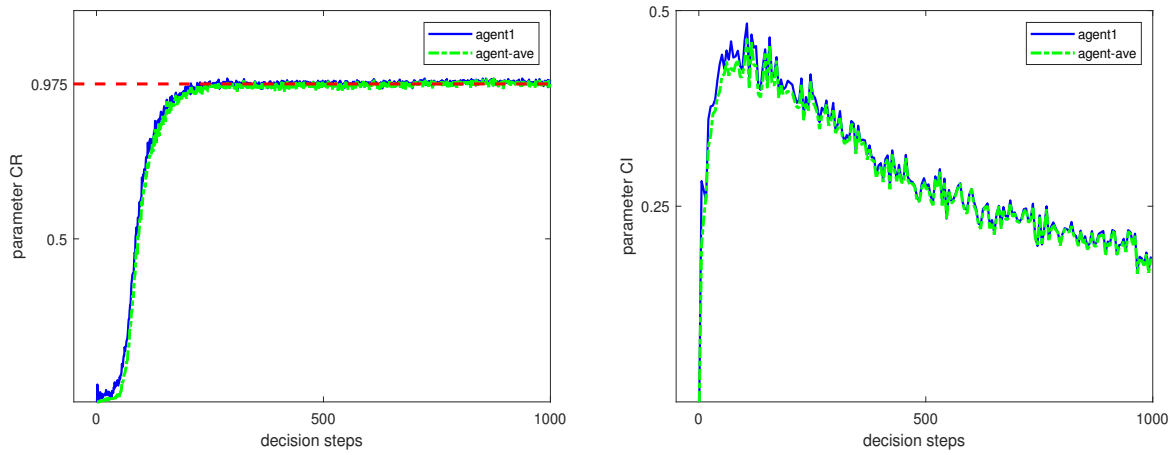


(d) Cumulative value estimator for Agent 1



(e) Cumulative value estimator with $\varepsilon_t = 0.1$

Figure 1: The convergence of the parameter and value estimators



(a) Parameter CR (b) Parameter CI

Figure 2: Statistical inference of the parameter estimators with $\varepsilon_t = 0.1$

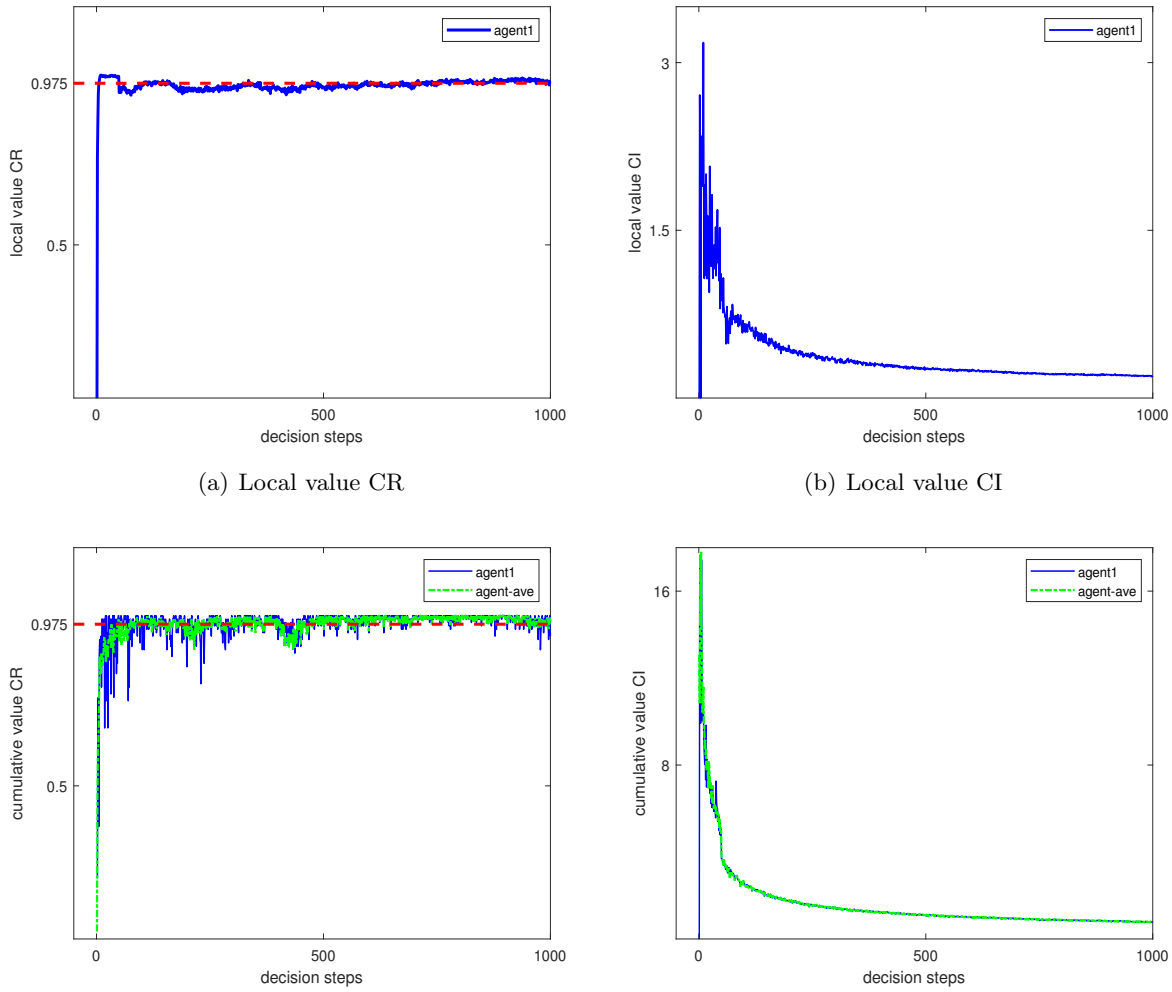
In Figure 2, we record the coverage probability and average length of 97.5% confidence interval for the estimators of parameter, where blue solid line and green dashed line denote the statistical inference for Agent 1 and the average of all agents respectively. Figure 2 shows that the average length of the confidence intervals shrinks and the coverage probabilities are getting closer to the nominal level 97.5% with the increasing of the time horizon T, which verifies the result of Theorem 3.1. Moreover, Figure 2 indicates the agreement of ε-DSGDCB as the coverage probability, the average length of the confidence intervals of parameter estimators for Agent 1 and the average of all agents almost coincide.

Figure 3 records the coverage probability and average length of 97.5% confidence interval for the value estimators. Similar to Figure 2, we may observe the reduction of the average length of the confidence intervals and the convergence of coverage probabilities to the nominal level 97.5% with increasing time horizon T, which verifies the results of Theorems 3.2 and 3.3. Moreover, we may conclude the agreement of the distributed variant of plug-in cumulative value estimators in Figures 3 (c) and (d).

4.2 Real Data Analysis

Warfarin is one of the most widely prescribed anticoagulant drugs worldwide, but the determination of the appropriate dose varies significantly among individuals, and improper dosing can lead to severe health consequences, such as stroke or internal bleeding [34]. Consequently, the development of individualized dosing policy has emerged as a critical priority in precision medicine. In this section, we use distributed contextual bandit to learn a warfarin optimal dosing policy based on the real patient dataset[3], which is collected by 21 research groups spanning 9 countries and 4 continents in PharmGKB for 5700 patients treated with warfarin. We formulate the problem as a distributed contextual bandit problem with $n = 21$ agents by classifying the drug dose into three arms: low dosage, less than 3mg/day; medium dosage, ranging from

[3]The complete data set of genotypes and clinical variables, as well as the full genotype quality-control data, is available to registered PharmGKB users at www.pharmgkb.org (full data set accession number, PA162355460).



(a) Local value CR

(b) Local value CI

(c) Cumulative value CR

(d) Cumulative value CI

Figure 3: Statistical inference of the value estimators with $\varepsilon_t = 0.1$

3mg/day to 7mg/day; and high dosage, more than 7mg/day.[4] The distributed structure of ε-DSGDCB may protect patient privacy and reduce the cost of data transfer by sharing local estimators of global parameters across institutions rather than individual patient information.

In the experiment, we use 5528 patient samples with a known true patient-specific optimal warfarin dose, which is randomly divided into training (75%) and testing (25%) sample subsets. Moreover, the reward is defined as the therapeutic daily dose of warfarin, and the same linear regression loss function configuration as in subsection 4.1 is preserved. We encode and normalize the raw user features, which include demographics, diagnosis, pre-existing diagnoses, medications and genetics, and the resulting user features consist of $p = 27$ covariates ranging between 0 and 1. In ε-DSGDCB, the exploration rate is $\varepsilon_t = 0.1$, the learning rate is $\alpha_t = 0.1/(t+1)^{0.7}$, the initial point is $\beta = 0_{1701}$ and the same communication matrix $\mathbf{M}$ configuration is preserved as subsection 4.1.

Table 1: Percentage of warfarin dosage assigned to patients with $T = 10000$

True	Bandit policy assigned dosage (%)			Physician policy assigned dosage (%)			% of
dosage	Low	Medium	High	Low	Medium	High	patient
Low	**44**	56	<u>0</u>	0	100	0	29
Medium	6	**90**	4	0	100	0	57
High	<u>1</u>	68	**31**	0	100	0	14

Table 1 reports a comparison between the bandit policy and the physicians' fixed-dose policy [16] for the percentage of warfarin doses assigned to patients, where the bold numbers indicate the fraction of patients recommended an appropriate dose, the underlined numbers denote the fraction of patients recommended a significantly worse dose, and other numbers represent the fraction of patients recommended an unsuitable dose. We can observe from Table 1 that the bandit policy increases the risk of incorrect dosing for a small number of patients in exchange for a large improvement in average dosing accuracy. Specifically, patients experience significantly suboptimal dosing under the bandit policy with only a 0.1% weighted probability. The bandit policy demonstrates superior performance to physician decisions, correctly recommending dosages for 44% of low-dose patients and 31% of high-dose patients. This contrasts with physician policies resulting in suboptimal treatment outcomes for these critical subgroups, which constitute half of the patient population.

Table 2 lists the 97.5% asymptotic confidence intervals length for the estimators of parameter and value, where the 'Avglen' and 'Medlen' denote the average and the median of confidence interval length for the parameter estimators, 'loc-Vallen' and 'cum-Vallen' denote the confidence interval length of the local and cumulative value estimators. As we can observe from Table 2, for a fixed decision step, the confidence interval length for the parameters of Agent 1 is similar to that of the average of all agents, which implies the agreement of ε-DSGDCB in Theorem 2.1. Additionally, the confidence interval length for the cumulative value estimator of Agent 1 and the average of all agents almost coincide, which indicates the agreement of the distributed

[4]The classifications of the drug dose follow the 'clinically relevant' criteria suggested in [16].

Table 2: 97.5% confidence intervals for the estimators of parameter and optimal value with different decision steps

Decision	Agent 1				Agent-ave			
steps	Avglen	Medlen	loc-Vallen	cum-Vallen	Avglen	Medlen	loc-Vallen	cum-Vallen
$T = 1000$	0.0938	0.0433	0.2119	1.4989	0.0928	0.0444	0.2069	1.4998
$T = 5000$	0.0875	0.0376	0.0909	0.6723	0.0873	0.0362	0.0899	0.6723
$T = 10000$	0.0634	0.0311	0.0626	0.4776	0.0694	0.0318	0.0636	0.4776

variant of the plug-in method in Theorem 3.3. Furthermore, the length of the confidence interval shrinks when the number of decision step T grows larger, which are consistent with those obtained from numerical simulations in synthetic data.

Acknowledgements

The research is supported by by National Key R&D Program of China No. 2022YFA1004000, NSFC #12471283 and Fundamental Research Funds for the Central Universities DUT24LK001.

References

[1] Karim M. Abadir and Paolo Paruolo. Two mixed normal densities from cointegration analysis. *Econometrica*, 65(3):671–680, 1997.

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 127–135, 2013.

[3] Sakshi Arya and Bharath K Sriperumbudur. Kernel *epsilon*-greedy for contextual bandits. *arXiv preprint arXiv:2306.17329*, 2023.

[4] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[5] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[6] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *Neural Information Processing: 19th International Conference*, volume 7665, pages 324–331. Springer, 2012.

[7] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

[8] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

[9] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.

[10] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.

[11] Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for contextual bandits via stochastic gradient descent. *arXiv preprint arXiv:2212.14883*, 2022.

[12] Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for stochastic optimization via kiefer-wolfowitz methods. *Journal of the American Statistical Association*, pages 1–24, 2024.

[13] Xi Chen, Jason Lee, Xin Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

[14] Xiaohong Chen, Sokbae Lee, Yuan Liao, Myung Hwan Seo, Youngki Shin, and Myunghyun Song. SGMM: Stochastic approximation to generalized method of moments. *Journal of Financial Econometrics*, pages 1–40, 2023.

[15] K. Chung. On a stochastic approximation method. *Annals of Mathematical Statistics*, 25:463–483, 1954.

[16] International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[17] Abhimanyu Dubey and AlexSandy' Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33:6003–6014, 2020.

[18] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

[19] Vaclav Fabian. On asymptotic normality in stochastic approximation. *Annals of Mathematical Statistics*, 39:1327–1332, 1968.

[20] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[21] Qiyu Han, Will Wei Sun, and Yichen Zhang. Online statistical inference for matrix contextual bandit. *arXiv preprint arXiv:2212.11385*, 2022.

[22] Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in neural information processing systems*, 34:27057–27068, 2021.

[23] Nicholas M Kiefer, Timothy J Vogelsang, and Helle Bunzel. Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714, 2000.

[24] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022.

[25] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast inference for quantile regression with tens of millions of observations. *Journal of Econometrics*, page 105673, 2024.

[26] Chuanhao Li and Hongning Wang. Communication efficient federated learning for generalized linear bandits. *Advances in Neural Information Processing Systems*, 35:38411–38423, 2022.

[27] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[28] Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems*, 35:38449–38460, 2022.

[29] Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and inference via local SGD in federated learning. *arXiv preprint arXiv:2109.01326*, 2021.

[30] Bing Liu, Tong Yu, Ian Lane, and Ole Mengshoel. Customized nonlinear bandits for online response selection in neural conversation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5245–5252, 2018.

[31] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[32] Shi Pu. A robust gradient tracking method for distributed optimization over directed networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2335–2341. IEEE, 2020.

[33] Wei Qian and Yuhong Yang. Kernel estimation and model combination in a bandit problem with covariates. *The Journal of Machine Learning Research*, 17(1):5181–5217, 2016.

[34] Allan E Rettie and Guoying Tai. The pharmocogenomics of warfarin: closing in on personalized medicine. *Molecular Interventions*, 6(4):223–227, 2006.

[35] Ali H Sayed et al. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.

[36] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2020.

[37] Ward Whitt. *Stochastic-process limits : an introduction to stochastic-process limits and their application to queues*. Springer series in operations research. Springer-Verlag, 2002.

[38] Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

[39] Li Zhou. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*, 2015.

[40] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–12, 2021.

5 Appendix

5.1 Useful Inequalities

Lemma 5.1. [28, Lemma 6] *Let a sequence of non-negative, non-increasing learning rate* $\{\gamma_t\}_{t\geq 1}$, $a>0, p\in\mathbb{Z}_+$ *and* $\gamma_0<2/a$. *If* $\gamma_t^p/\gamma_{t+1}^p\leq 1+(a/2)\gamma_{t+1}^p$ *for any* $t\geq 1$, *then*

$$\sum_{j=1}^{t}\gamma_j^{p+1}\prod_{i=j+1}^{t}(1-\gamma_i a)\leq\frac{2}{a}\gamma_t^p,\quad\forall t\geq 1.$$

Lemma 5.2. [28, Lemma 7] *Let a sequence of non-negative, non-increasing learning rate* $\{\gamma_t\}_{t\geq 1}$, $p\in\mathbb{Z}^+$. *If* $\sup_{t\geq 1}\gamma_t^p/\gamma_{t+1}^p\leq 1+\frac{p}{4-2p}$, *then for any* $t\geq 0$, *it holds that*

$$\sum_{i=1}^{t+1}\left(1-\frac{\rho}{2}\right)^{t+1-i}\gamma_i^p\leq\frac{4}{\rho}\gamma_{t+1}^p.$$

5.2 Proof of Lemma 2.1

Proof. By the definition of $\hat{\beta}_t$,

$$\hat{\beta}_{t+1}=\left(\frac{\mathbf{1}^T}{n}\otimes\mathbf{I}_{Kp}\right)\left(\tilde{\mathbf{A}}\beta_t-\gamma_{t+1}G_t\right)=\hat{\beta}_t-\gamma_{t+1}\hat{G}_t,$$

where $\tilde{\mathbf{A}}=\mathbf{A}\otimes\mathbf{I}_{Kp}$. Recall that $\Delta_t=\hat{\beta}_t-\beta^*$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{P}_O^\pi}\left[\|\Delta_{t+1}\|^2\mid\mathcal{F}_t\right]=&\|\Delta_t\|^2-\frac{2\gamma_{t+1}}{n}\left\langle\Delta_t,\sum_{i=1}^{n}\nabla L_i\left(\beta_{i,t}\right)\right\rangle\\&+\frac{\gamma_{t+1}^2}{n^2}\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\sum_{i=1}^{n}g_i\left(\beta_{i,t};O_{i,t+1}\right)\right\|^2\mid\mathcal{F}_t\right].\end{aligned}\tag{5.22}$$

For the second term on the right hand of (5.22),

$$\begin{aligned}\frac{1}{n}\left\langle\Delta_t,\sum_{i=1}^{n}\nabla L_i\left(\beta_{i,t}\right)\right\rangle=&\frac{1}{n}\sum_{i=1}^{n}\left\langle\Delta_t,\nabla L_i\left(\beta_{i,t}\right)-\nabla L_i(\hat{\beta}_t)\right\rangle+\frac{1}{n}\sum_{i=1}^{n}\left\langle\Delta_t,\nabla L_i(\hat{\beta}_t)-\nabla L_i\left(\beta^*\right)\right\rangle\\\geq&\frac{\mu}{n}\|\Delta_t\|^2-\frac{c_L}{n}\sum_{i=1}^{n}\|\Delta_t\|\left\|\beta_{i,t}-\hat{\beta}_t\right\|,\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality, Assumptions 2.1 (i) and 2.2 (i). By the Young's inequality, for $\delta\in(0,\frac{\mu}{c_L})$,

$$\frac{\mu}{n}\|\Delta_t\|^2-\frac{c_L}{n}\sum_{i=1}^{n}\|\Delta_t\|\left\|\beta_{i,t}-\hat{\beta}_t\right\|\geq\left(\frac{\mu}{n}-\frac{\delta}{n}c_L\right)\|\Delta_t\|^2-\frac{c_L}{4n\delta}\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2.\tag{5.23}$$

For the third term on the right hand of (5.22), we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}g_i\left(\beta_{i,t};O_{i,t+1}\right)\right\|^2=\left\|\frac{1}{n}\sum_{i=1}^{n}\left[g_i\left(\beta_{i,t};O_{i,t+1}\right)-\nabla L_i\left(\beta_{i,t}\right)+\nabla L_i\left(\beta_{i,t}\right)-\nabla L_i\left(\beta^*\right)\right]\right\|^2,$$

which implies

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i\left(\beta_{i,t};O_{i,t+1}\right)\right\|^2 \mid \mathcal{F}_t\right]\\
\leq& 2\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left[g_i\left(\beta_{i,t};O_{i,t+1}\right)-\nabla L_i\left(\beta_{i,t}\right)\right]\right\|^2 \mid \mathcal{F}_t\right]+\frac{2}{n}\sum_{i=1}^n\left\|\nabla L_i\left(\beta_{i,t}\right)-\nabla L_i\left(\beta^*\right)\right\|^2.
\end{aligned}$$

Denote $\phi_t=\frac{1}{n}\sum_{i=1}^n\left[g_i\left(\beta_{i,t};O_{i,t+1}\right)-\nabla L_i\left(\beta_{i,t}\right)\right]$, we decompose the martingale difference sequence ϕ_t into the following two parts,

$$\begin{aligned}
\phi_t^{(1)}:=&\frac{1}{n}\sum_{i=1}^n g_i(\beta^*;O_{i,t+1}),\\
\phi_t^{(2)}:=&\frac{1}{n}\sum_{i=1}^n\left[g_i(\beta_{i,t};O_{i,t+1})-g_i(\beta^*;O_{i,t+1})+\nabla L_i(\beta^*)-\nabla L_i(\beta_{i,t})\right].
\end{aligned}$$

For $\phi_t^{(1)}$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\phi_t^{(1)}\right\|^2 \mid \mathcal{F}_t\right]=&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i(\beta^*;O_{i,t+1})\right\|^2 \mid \mathcal{F}_t\right]\\
\leq&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^K\frac{\left\|\nabla\ell_i\left(\beta^*;O_{i,t+1}\right)\right\|^2 1_{\{A_{i,t+1}=j\}}}{K^2\left\{\pi_{i,t}\left(X_{i,t+1}\right)\right\}^2} \mid \mathcal{F}_t\right]\\
=&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^K\frac{\mathbb{E}\left\{\left\|\nabla\ell_i\left(\beta^*;X_{i,t+1},j,Y_{i,t+1}\right)\right\|^2 \mid \mathcal{F}_t,X_{i,t+1}\right\}}{K^2\pi_{i,t}\left(X_{i,t+1}\right)} \mid \mathcal{F}_t\right]\\
\leq&\frac{1}{\varepsilon_\infty}\mathbb{E}_{\mathcal{P}_O^r}\left[\frac{1}{n}\sum_{i=1}^n\left\|\nabla\ell_i\left(\beta^*;O_{i,t+1}\right)\right\|^2\right].
\end{aligned}\tag{5.24}$$

By Assumption 2.2 (ii), we have

$$\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\phi_t^{(1)}\right\|^2 \mid \mathcal{F}_t\right]\leq\frac{1}{\varepsilon_\infty}\mathbb{E}_{\mathcal{P}_O^r}\left[\frac{1}{n}\sum_{i=1}^n\left\|\nabla\ell_i\left(\beta^*;O_{i,t+1}\right)\right\|^2\right]\leq\frac{c_f}{\varepsilon_\infty}.$$

For $\phi_t^{(2)}$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left[g_i(\beta_{i,t};O_{i,t+1})-g_i(\beta^*;O_{i,t+1})+\nabla L_i(\beta^*)-\nabla L_i(\beta_{i,t})\right]\right\|^2 \mid \mathcal{F}_t\right]\\
\leq&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\frac{2}{n}\sum_{i=1}^n\sum_{j=1}^K\frac{\left\|\nabla\ell_i\left(\beta_{i,t};O_{i,t+1}\right)-\nabla\ell_i\left(\beta^*;O_{i,t+1}\right)\right\|^2 1_{\{A_{i,t+1}=j\}}}{K^2\left\{\pi_{i,t}\left(X_{i,t+1}\right)\right\}^2} \mid \mathcal{F}_t\right]\\
&+\frac{2}{n}\sum_{i=1}^n\left\|\nabla L_i\left(\beta_{i,t}\right)-\nabla L_i\left(\beta^*\right)\right\|^2\\
\leq&\frac{1}{\varepsilon_\infty}\mathbb{E}_{\mathcal{P}_O^r}\left[\frac{2}{n}\sum_{i=1}^n\left\|\nabla\ell_i\left(\beta_{i,t};O_{i,t+1}\right)-\nabla\ell_i\left(\beta^*;O_{i,t+1}\right)\right\|^2\right]+\frac{2}{n}\sum_{i=1}^n\left\|\nabla L_i\left(\beta_{i,t}\right)-\nabla L_i\left(\beta^*\right)\right\|^2.
\end{aligned}$$

By Assumption 2.2 (i),

$$\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\phi_t^{(2)}\right\|^2 \mid \mathcal{F}_t\right] \le \frac{4c_L^2}{n\varepsilon_\infty}\sum_{i=1}^n \|\beta_{i,t}-\beta^*\|^2, \tag{5.25}$$

which implies

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i\left(\beta_{i,t};O_{i,t+1}\right)\right\|^2 \mid \mathcal{F}_t\right] \\
\le&\frac{4c_f}{\varepsilon_\infty}+\frac{16c_L^2}{n\varepsilon_\infty}\sum_{i=1}^n\|\beta_{i,t}-\beta^*\|^2+\frac{2}{n}\sum_{i=1}^n c_L^2\|\beta_{i,t}-\beta^*\|^2 \\
\le&\frac{4c_f}{\varepsilon_\infty}+c_2\|\Delta_t\|^2+\frac{c_2}{n}\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2,
\end{aligned} \tag{5.26}$$

where $c_2 = 4c_L^2\left(\frac{8}{\varepsilon_\infty}+1\right)$.

Substitute (5.23) and (5.26) into (5.22), we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\|\Delta_{t+1}\|^2 \mid \mathcal{F}_t\right] \\
\le&\|\Delta_t\|^2-2\gamma_{t+1}\left[\left(\frac{\mu}{n}-\frac{\delta}{n}c_L\right)\|\Delta_t\|^2-\frac{c_L}{4n\delta}\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2\right] \\
&+\gamma_{t+1}^2\left[\frac{4c_f}{\varepsilon_\infty}+c_2\|\Delta_t\|^2+\frac{c_2}{n}\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2\right] \\
=&\left(1-2\tilde{\mu}\gamma_{t+1}+c_2\gamma_{t+1}^2\right)\|\Delta_t\|^2+\left[c_1\frac{\gamma_{t+1}}{n}+c_2\frac{\gamma_{t+1}^2}{n}\right]\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2+\frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2 \\
\le&\left(1-\tilde{\mu}\gamma_{t+1}\right)\|\Delta_t\|^2+\left[c_1\frac{\gamma_{t+1}}{n}+c_2\frac{\gamma_{t+1}^2}{n}\right]\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2+\frac{4c_f}{\varepsilon_\infty}\gamma_{t+1}^2,
\end{aligned}$$

where the last inequality follows from the condition $\gamma_{t+1}\le\widetilde{\mu}/c_2$. □

5.3 Proof of Lemma 2.2

Proof. By the definition of $\hat{\beta}_t$,

$$\beta_{t+1}-\mathbf{1}_n\otimes\hat{\beta}_{t+1}=\left(\tilde{\mathbf{A}}-\frac{\mathbf{1}\mathbf{1}^T}{n}\otimes\mathbf{I}_{Kp}\right)\left(\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right)-\gamma_{t+1}\left(\mathbf{I}_{n\times Kp}-\frac{\mathbf{1}\mathbf{1}^T}{n}\otimes\mathbf{I}_{Kp}\right)G_t.$$

Then, for any $\alpha>0$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\beta_{t+1}-\mathbf{1}_n\otimes\hat{\beta}_{t+1}\right\|^2 \mid \mathcal{F}_t\right] \\
\le&(1+\alpha)(1-\rho)^2\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2 \\
&+\left(1+\frac{1}{\alpha}\right)\gamma_{t+1}^2\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\left(\mathbf{I}_{n\times Kp}-\frac{\mathbf{1}\mathbf{1}^T}{n}\otimes\mathbf{I}_{Kp}\right)G_t\right\|^2 \mid \mathcal{F}_t\right] \\
\le&(1-\rho)\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2+\frac{\gamma_{t+1}^2}{\rho}\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\left(\mathbf{I}_{n\times Kp}-\frac{\mathbf{1}\mathbf{1}^T}{n}\otimes\mathbf{I}_{Kp}\right)G_t\right\|^2 \mid \mathcal{F}_t\right],
\end{aligned} \tag{5.27}$$

where the first inequality follows from Assumption 2.3 and the second inequality is obtained by setting $\alpha = \frac{\rho}{1-\rho}$.

For the last term on the right hand of (5.27),

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\left(\mathbf{I}_{n\times Kp} - \frac{\mathbf{1}\mathbf{1}^T}{n}\otimes \mathbf{I}_{Kp}\right) G_t\right\|^2 \mid \mathcal{F}_t\right] \\
=&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\sum_{i=1}^n \left(g_i\left(\beta_{i,t}; O_{i,t+1}\right) - \frac{1}{n}\sum_{j=1}^n g_j\left(\beta_{j,t}; O_{j,t+1}\right)\right)\right\|^2 \mid \mathcal{F}_t\right] \\
\leq& 6\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\sum_{i=1}^n \left(g_i\left(\beta_{i,t}; O_{i,t+1}\right) - \nabla L_i\left(\beta_{i,t}\right)\right)\right\|^2 \mid \mathcal{F}_t\right] + 3\sum_{i=1}^n \left\|\nabla L_i\left(\beta_{i,t}\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\beta_{j,t}\right)\right\|^2 \\
\leq& 12n\left(\frac{c_f}{\varepsilon_\infty} + \frac{4c_L^2}{n\varepsilon_\infty}\sum_{i=1}^n \left\|\beta_{i,t} - \beta^*\right\|^2\right) + 3\sum_{i=1}^n \left\|\nabla L_i\left(\beta_{i,t}\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\beta_{j,t}\right)\right\|^2 \\
\leq& 12n\left(\frac{c_f}{\varepsilon_\infty} + \frac{8c_L^2}{\varepsilon_\infty}\left\|\Delta_t\right\|^2 + \frac{8c_L^2}{n\varepsilon_\infty}\left\|\beta_t - \mathbf{1}_n\otimes\hat{\beta}_t\right\|^2\right) \\
&+ 3\sum_{i=1}^n \left\|\nabla L_i\left(\beta_{i,t}\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\beta_{j,t}\right)\right\|^2,
\end{aligned} \tag{5.28}$$

where the second inequality follows from the analysis of (5.24)-(5.25). For the last term on the right hand of (5.28),

$$\begin{aligned}
&\sum_{i=1}^n \left\|\nabla L_i\left(\beta_{i,t}\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\beta_{j,t}\right)\right\|^2 \\
\leq& 6\sum_{i=1}^n \left\|\nabla L_i\left(\beta_{i,t}\right) - \nabla L_i\left(\hat{\beta}_t\right)\right\|^2 + 3\sum_{i=1}^n \left\|\nabla L_i\left(\hat{\beta}_t\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\hat{\beta}_t\right)\right\|^2.
\end{aligned}$$

By Assumption 2.2 (ii),

$$\begin{aligned}
&\sum_{i=1}^n \left\|\nabla L_i\left(\hat{\beta}_t\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\hat{\beta}_t\right)\right\|^2 \\
\leq& 3\sum_{i=1}^n \left\|\nabla L_i\left(\beta^*\right)\right\|^2 + 3\sum_{i=1}^n \left\|\nabla L_i\left(\hat{\beta}_t\right) - \nabla L_i\left(\beta^*\right)\right\|^2 + \frac{3}{n}\sum_{i=1}^n \left\|\nabla L\left(\beta^*\right) - \nabla L\left(\hat{\beta}_t\right)\right\|^2 \\
\leq& 3nc_f^2 + 6nc_L^2\left\|\Delta_t\right\|^2,
\end{aligned}$$

which implies that

$$\sum_{i=1}^n \left\|\nabla L_i\left(\beta_{i,t}\right) - \frac{1}{n}\sum_{j=1}^n \nabla L_j\left(\beta_{j,t}\right)\right\|^2 \leq 6c_L^2\left\|\beta_t - \mathbf{1}_n\otimes\hat{\beta}_t\right\|^2 + 3nc_f^2 + 6nc_L^2\left\|\Delta_t\right\|^2. \tag{5.29}$$

Substitute (5.29) into (5.28), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\left(\mathbf{I}_{n\times Kp}-\frac{\mathbf{1}\mathbf{1}^T}{n}\otimes\mathbf{I}_{Kp}\right)G_t\right\|^2\mid\mathcal{F}_t\right]\\
\leq&12n\left(\frac{c_f}{\varepsilon_\infty}+\frac{8c_L^2}{\varepsilon_\infty}\left\|\Delta_t\right\|^2+\frac{8c_L^2}{n\varepsilon_\infty}\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2\right)+9nc_f^2+18c_L^2\left(\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2+n\left\|\Delta_t\right\|^2\right) \quad (5.30)\\
\leq&3n\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)+6nc_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\left\|\Delta_t\right\|^2+6c_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2.
\end{aligned}
$$

Then, substitute the resulting expression (5.30) into (5.27), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{P}_O^\pi}\left[\left\|\beta_{t+1}-\mathbf{1}_n\otimes\hat{\beta}_{t+1}\right\|^2\mid\mathcal{F}_t\right]\\
\leq&(1-\rho)\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2+\frac{\gamma_{t+1}^2}{\rho}\left[6nc_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\left\|\Delta_t\right\|^2+c_3\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2\right]+3n\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\frac{\gamma_{t+1}^2}{\rho}\\
\leq&(1-\rho/2)\left\|\beta_t-\mathbf{1}_n\otimes\hat{\beta}_t\right\|^2+\frac{\gamma_{t+1}^2}{\rho}6nc_L^2\left(\frac{16}{\varepsilon_\infty}+3\right)\left\|\Delta_t\right\|^2+3n\left(\frac{4c_f}{\varepsilon_\infty}+3c_f^2\right)\frac{\gamma_{t+1}^2}{\rho},
\end{aligned}
$$

where the last inequality follows from $\gamma_t\leq\rho/\sqrt{2c_3}$. The proof is complete. □