MOVE: LEARNING MARKOV PERSUASION FOR ONLINE FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Online Federated Learning (OFL) trains a shared model from sequentially arriving clients without centralizing raw data. Sustaining participation at low payment is challenging because participation affects future states, only the server observes communication conditions, and their distribution and state dynamics are unknown to the server. Existing methods often neglect future effects, assume known environments, or design rewards and disclosure separately, leading to excessive payments or insufficient participation. We propose MOVE, an online Markov persuasion framework for participation control in unknown action dependent environments. MOVE coordinates rewards and disclosure while learning the condition distribution and state transitions. We formulate participation control through occupancy measures under Bayes plausibility, Bayes benefit, and participation requirements, and prove that posteriors supported on at most two communication conditions suffice for optimal signaling. This structure enables recovery of a signaling policy consistent with Bayes' rule and finite optimistic planning that accounts for estimation uncertainty. With high probability, MOVE achieves sublinear payment pseudo regret against the best fixed feasible mechanism and sublinear cumulative constraint violations. Experiments show average reductions of 2.66% in payment per arrival versus w/o Signaling, 66.94% in participation violations versus w/o Markov, and 23.47% in target payment versus existing baselines.

1 INTRODUCTION

Online Federated Learning (OFL) updates a global model through sequential client arrivals to adapt to evolving data and network conditions (Chen et al., 2020; Zhu et al., 2023). Fluctuating link quality and network load affect clients' model upload costs (Hou et al., 2024). Through network monitoring, the server can observe these conditions before an arriving client can accurately assess its upload cost, creating information asymmetry (Xin et al., 2025; Luo et al., 2025). Disclosed information shapes the client's cost estimate, which combines with the offered reward to determine its participation decision (Hu et al., 2022; Agrawal et al., 2023). Participation updates the global model and generates traffic that can alter subsequent network conditions (Ding et al., 2023). Heterogeneous workloads and evolving networks make condition distributions and state transitions affected by participation difficult to know in advance. The server must therefore learn to coordinate rewards and disclosure to sustain sufficient participation at low payment while accounting for their effects on future states.

Existing research follows three main directions. First, online Bayesian persuasion learns disclosure policies when the sender privately observes the state, with extensions to action only feedback and unknown state priors Castiglioni et al. (2020; 2021); Bernasconi et al. (2023); Bacchiocchi et al. (2024). These studies omit participation rewards and the effects of current actions on future states. Second, FL studies use monetary incentives to encourage client participation and data sharing (Wei et al., 2023), without incorporating strategic disclosure of communication conditions observed by the server. Third, Markov persuasion models action effects on subsequent states and supports signaling policy learning under unknown transitions (Wu et al., 2022; Bacchiocchi et al., 2025). However, these methods do not optimize participation rewards that shape clients' responses to disclosure. Thus, jointly learning reward and disclosure policies remains challenging in OFL with unknown communication condition distributions and participation dependent transitions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Closing this gap in an unknown, action dependent environment raises three technical challenges. *(i) Intertemporal state dependence.* The server's reward and disclosure decisions influence participation, thereby affecting subsequent state visitation and participation opportunities within the episode (Gan et al., 2025). *(ii) Endogenous environment learning.* The server estimates the unknown condition distribution and transition dynamics from trajectories generated by its policies, while subsequent planning must account for estimation uncertainty (Stradi et al., 2024). *(iii) Infinite mechanism space.* Continuous rewards and posterior beliefs yield infinitely many mechanisms, requiring a finite, tractable representation that preserves Bayesian feasibility and controls approximation error.

To address these challenges, we propose MOVE, an online Markov persuasion framework for sustaining client participation at low payment in OFL. MOVE coordinates rewards and disclosure while learning the unknown condition distribution and participation dependent transitions from observations generated by its policies. It models OFL participation control as episodic Markov persuasion through an occupancy formulation under Bayes plausibility, Bayes benefit, and participation requirements. MOVE plans through occupancy measures using a compact representation of pairwise posteriors. By discretizing rewards and pairwise posteriors and using confidence bounds for optimistic planning, MOVE reduces the infinite mechanism design problem in an unknown environment to a sequence of finite linear programs with provable guarantees. Our main contributions are:

- We characterize a compact posterior structure for the constrained formulation by proving that posteriors supported on at most two communication conditions suffice for optimality, enabling closed form recovery of a signaling policy consistent with Bayes' rule.
- We prove that MOVE achieves, with high probability, sublinear payment pseudo regret and sublinear cumulative violations of participation and Bayes benefit constraints.
- We validate MOVE through controlled experiments and federated learning evaluations, with scalability tests demonstrating its computational practicality. On average, MOVE reduces payment per arrival by 2.66% versus w/o Signaling, participation violations by 66.94% versus w/o Markov, and target payment by 23.47% versus existing baselines.

2 Related Work

Client Participation in FL. Prior studies used payment and client selection policies to encourage FL participation by resource constrained clients (Murhekar et al., 2023). Existing methods typically relied on observable or reported information (Pang et al., 2023; Ding et al., 2025). When decision relevant information was private, they generally required its distribution to be known (Hu et al., 2024). OFL methods combined signaling with adaptive rewards under independent client conditions and a known communication condition prior, while learning the client computation resource distribution online (Xin et al., 2025). This formulation omitted participation effects on subsequent network states. MOVE learns the condition distribution and participation dependent transitions while coordinating disclosure with rewards.

Table 1: Comparison of representative studies.

Work	Unknown prior	Unknown transitions	Controlled transitions	Payments & signals
Pang et al. (2023)	✗	✗	✗	✗
Xin et al. (2025)	✗	✗	✗	✓
Bernasconi et al. (2023)	✗	✗	✗	✗
Bacchiocchi et al. (2024)	✓	✗	✗	✗
Agrawal et al. (2023)	✗	✗	✗	✓
Wu et al. (2022)	✓	✓	✓	✗
Bacchiocchi et al. (2025)	✓	✓	✓	✗
MOVE (Ours)	✓	✓	✓	✓

Markov Persuasion. Bayesian persuasion designed disclosure to influence receiver decisions (Kamenica & Gentzkow, 2011). Online methods learned signaling policies for single or multiple receivers from limited information (Castiglioni et al., 2020; 2021; Bernasconi et al., 2023). Further work learned signaling policies under unknown priors (Bacchiocchi et al., 2024). Side information was incorporated while the private state prior remained known (Balcan et al., 2026). Pricing and disclosure were jointly learned across repeated buyer interactions (Agrawal et al., 2023). Markov persuasion modeled action effects on subsequent states (Wu et al., 2022) and supported signaling policy learning from partial feedback in unknown environments (Bacchiocchi et al., 2025). These methods did not optimize participation rewards affecting receiver responses and future states. MOVE coor-

dinates rewards with disclosure under minimum participation and Bayes benefit constraints, using pairwise posterior reduction for finite optimistic planning. Table 1 summarizes this comparison.

3 OFL Participation Control

3.1 System Model

Figure 1 illustrates how rewards and signaling affect client participation and subsequent network states in OFL. The server operates over T episodes indexed by $t \in [T]$, each with a random number $N_t \in [N]$ of sequential client arrivals, capped at N. Each episode starts at public network state x_0 and follows six steps. (i) At the start of episode t, the server commits to a reward $\gamma_t \in [\underline{\gamma}, \overline{\gamma}]$ per participant and a signaling policy for each state x. (ii) When a client arrives in state x, the server observes through network monitoring a communication condition $\tau \sim \lambda(\cdot \mid x)$ from the m possible conditions $\Omega := \{\tau_i : i \in [m]\}$, where λ is unknown. The client does not directly observe τ before its participation decision. (iii) Given τ, the server draws and sends a signal $s \in \mathcal{S}$ according to its committed signaling policy $\rho_{t,x}(\cdot \mid \tau)$, where $\mathcal{S}$ is the signal space. (iv) Upon receiving s, the client forms a posterior over τ and evaluates its expected cost using the known function $\kappa : \Omega \to \mathbb{R}_+$. It then chooses $a \in \{0, 1\}$, where 1 denotes participation and 0 nonparticipation. (v) If $a = 1$, OFL local and global updates occur, and the client receives γ_t. Otherwise, no payment or model update occurs. (vi) The next public state x' is drawn from the unknown controlled transition kernel $P(\cdot \mid x, \tau, a)$, and the server records the transition (x, τ, a, x').

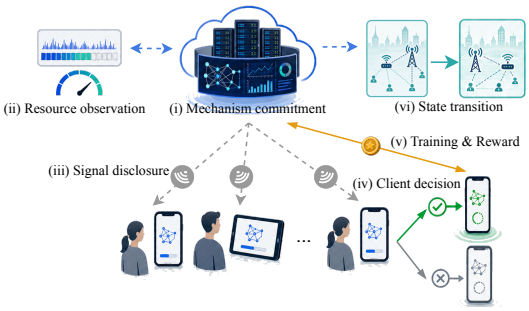


Figure 1: Server and client interaction in OFL.

3.2 Problem Formulation

Under signaling rule $\rho_{t,x}$, signal s has probability $\phi_{t,x}(s) := \sum_{j=1}^{m} \lambda(\tau_j \mid x)\rho_{t,x}(s \mid \tau_j)$. For $\phi_{t,x}(s) > 0$, the posterior is $\mu_{t,x}(\tau_i \mid s) := \lambda(\tau_i \mid x)\rho_{t,x}(s \mid \tau_i)/\phi_{t,x}(s)$. Following the Bayes plausibility principle in Bayesian persuasion (Kamenica & Gentzkow, 2011), we formalize the requirement that the induced posteriors average to the prior as follows.

Definition 1. *$\rho_{t,x}$ satisfies Bayes plausibility if* $\sum_{s\in\mathcal{S}} \phi_{t,x}(s)\mu_{t,x}(\tau_i \mid s) = \lambda(\tau_i \mid x)$, $\forall \tau_i \in \Omega$.

Let $\Pi := (\pi_t)_{t\in[T]}$ denote the server's online policy, where $\pi_t = (\gamma_t, \rho_t)$. Following the no disclosure benchmark in Bayesian persuasion, we compare payments with and without signaling. Let $c(\Pi)$ denote the expected cumulative payment under Π and $c(\emptyset)$ the corresponding payment without signaling, evaluated under the same rewards and state visitation as Π. Without signaling, participation is based only on the prior. We define Bayes benefit (Kamenica & Gentzkow, 2011) as follows.

Definition 2. *Π satisfies the Bayes benefit if* $c(\Pi) \le c(\emptyset)$.

Let $a_{t,n}$ denote the participation action at the $(n+1)$-st client arrival of episode t. $v(\varepsilon)$ denotes the server's valuation of achieving target accuracy ε within T episodes. For fixed $v(\varepsilon)$, maximizing its net utility $v(\varepsilon) - c(\Pi)$ is equivalent to minimizing $c(\Pi)$:

$$
\begin{aligned}
\mathbf{P1}: \quad \min_{\Pi} \quad & c(\Pi) := \mathbb{E}_{\Pi}\Big[\sum_{t\in[T]} \gamma_t \sum_{n=0}^{N_t-1} a_{t,n}\Big], \\
\text{s.t.} \quad & \sum_{s\in\mathcal{S}} \phi_{t,x}(s)\mu_{t,x}(\tau_i \mid s) = \lambda(\tau_i \mid x), \\
& c(\Pi) \le c(\emptyset), \\
& \mathbb{E}_{\Pi}\Big[\sum_{t\in[T]} \sum_{n=0}^{N_t-1} a_{t,n}\Big] \ge q_{\min}\mathbb{E}_{\Pi}\Big[\sum_{t\in[T]} N_t\Big].
\end{aligned}
\tag{1}
$$

P1 minimizes expected payment subject to Bayes plausibility, Bayes benefit, and minimum participation $q_{\min}$. Direct optimization is challenging because rewards and posteriors are continuous, participation affects future states, and the unknown λ and P must be learned from policy generated observations. We next reformulate **P1** using occupancy measures to capture this state dependence.

4 Online Framework for Learning Markov Persuasion

This section develops an online framework for learning Markov persuasion in OFL through occupancy based optimization and environment learning.

4.1 Markov Persuasion Formulation

Let $\mathcal{X}_n$ denote the finite set of public network states x at the $(n+1)$-st client arrival, with $\mathcal{X}_0 = \{x_0\}$, where x_0 is the initial state. The terminal state $x_\dagger$ marks the end of client arrivals. At stage n, the public state belongs to $\mathcal{X}_n \cup \{x_\dagger\}$. The expected participation cost given signal s and state x is

$$\bar{\kappa}_{t,x}(s) := \sum_{i=1}^{m} \mu_{t,x}(\tau_i \mid s)\kappa(\tau_i). \tag{2}$$

Given signal s at state x, the client participates with probability $p(\gamma_t, \bar{\kappa}_{t,x}(s))$, where p is the known participation probability function. Let $q_t(x, \tau_i, s, a, x')$ denote the joint occupancy probability, with marginals $q_t(x) := \sum_{i,s,a,x'} q_t(x, \tau_i, s, a, x')$ and $q_t(x, \tau_i, a) := \sum_{s,x'} q_t(x, \tau_i, s, a, x')$, where $q_t(x_0) = 1$. The joint occupancy is

$$q_t(x, \tau_i, s, a, x') := q_t(x)\lambda(\tau_i \mid x)\rho_{t,x}(s \mid \tau_i)P(x' \mid x, \tau_i, a)p(\gamma_t, \bar{\kappa}_{t,x}(s))^a[1 - p(\gamma_t, \bar{\kappa}_{t,x}(s))]^{1-a}. \tag{3}$$

Given λ, P, and $\pi_t = (\gamma_t, \rho_t)$, define the expected arrival and participation counts as $A(\lambda, P, \pi_t) := \sum_{n=0}^{N-1} \sum_{x \in \mathcal{X}_n} q_t(x)$ and $J(\lambda, P, \pi_t) := \sum_{n=0}^{N-1} \sum_{x \in \mathcal{X}_n} \sum_{i,s,x'} q_t(x, \tau_i, s, 1, x')$, respectively. The expected payment is $C(\lambda, P, \pi_t) := \gamma_t J(\lambda, P, \pi_t)$. Under the same state occupancy, the payment difference from the no disclosure reference is $B(\lambda, P, \pi_t) := C(\lambda, P, \pi_t) - \gamma_t \sum_{n=0}^{N-1} \sum_{x \in \mathcal{X}_n} q_t(x) p(\gamma_t, \sum_i \lambda(\tau_i \mid x)\kappa(\tau_i))$. Since $\mathbb{E}_\Pi[\sum_{t \in [T]} B(\lambda, P, \pi_t)] = c(\Pi) - c(\emptyset)$, we rewrite **P1** as

$$\begin{aligned}
\mathbf{P2}: \quad \min_{\Pi} \quad & c(\Pi) := \mathbb{E}_\Pi[\sum_{t \in [T]} C(\lambda, P, \pi_t)], && \text{(4a)}\\
\text{s.t.} \quad & \sum_{s,a,x'} q_t(x, \tau_i, s, a, x') = q_t(x)\lambda(\tau_i \mid x), && \text{(4b)}\\
& \mathbb{E}_\Pi[\sum_{t \in [T]} B(\lambda, P, \pi_t)] \le 0, && \text{(4c)}\\
& \mathbb{E}_\Pi[\sum_{t \in [T]} (J(\lambda, P, \pi_t) - q_{\min} A(\lambda, P, \pi_t))] \ge 0. && \text{(4d)}
\end{aligned}$$

The objective sums expected payments across episodes. The first constraint enforces Bayes plausibility by matching the state and condition marginal to $q_t(x)\lambda(\tau_i \mid x)$. The remaining constraints retain **P1**'s Bayes benefit and minimum participation requirements. We require $q_{\min} \le p(\overline{\gamma}, \sum_i \lambda(\tau_i \mid x)\kappa(\tau_i))$ at every client arrival state x. This service attainability condition ensures feasibility with maximum reward and no disclosure.

4.2 Markov Signaling Design

Given the prior, a signaling rule determines the client's posterior after observing a signal. We identify a sufficient posterior structure and recover the signaling rule from the occupancy measure. To control how changes in the reward and posterior expected participation cost affect the participation probability, we impose the following regularity condition.

Assumption 1. *There exists a constant L_p such that $|p(\gamma, \bar{\kappa}) - p(\gamma', \bar{\kappa}')| \le L_p(|\gamma - \gamma'| + |\bar{\kappa} - \bar{\kappa}'|)$.*

Assumption 1 is consistent with bounded derivative conditions in generalized linear bandit models (Li et al., 2017). Logistic bandit models (Faury et al., 2020) satisfy this condition when applied to reward minus expected cost. More generally, continuously differentiable responses satisfy it on the compact reward and cost domain. Under this assumption, $\mathbf{P2}$ admits an optimal solution.

Lemma 1. *Under Assumption 1,* $\mathbf{P2}$ *admits an optimal solution for given* λ *and* P*.*

Building on this existence result, we next characterize the posterior support sufficient for optimality. This characterization identifies the structure needed to recover ρ from the occupancy measure.

Lemma 2. *Under Assumption 1,* $\mathbf{P2}$ *admits an optimal solution whose induced posteriors in every episode satisfy* $|\{i \in [m] : \mu_{t,x}(\tau_i \mid s) > 0\}| \leq 2$*.*

Each posterior thus involves at most two communication conditions, enabling Proposition 1 to recover the signaling rule from the occupancy measure.

Proposition 1. *In episode* t*, for a signal* s *at state* x *with posterior support* $\{\tau_i, \tau_j\}$*, where* $\kappa(\tau_i) < \kappa(\tau_j)$*, the corresponding signaling probabilities are*

$$\begin{cases} \rho_{t,x}(s \mid \tau_i) = \dfrac{[\kappa(\tau_j) - \bar{\kappa}_{t,x}(s)] \sum_{r,a,x'} q_t(x, \tau_r, s, a, x')}{q_t(x)\lambda(\tau_i \mid x)[\kappa(\tau_j) - \kappa(\tau_i)]}, \\ \rho_{t,x}(s \mid \tau_j) = \dfrac{[\bar{\kappa}_{t,x}(s) - \kappa(\tau_i)] \sum_{r,a,x'} q_t(x, \tau_r, s, a, x')}{q_t(x)\lambda(\tau_j \mid x)[\kappa(\tau_j) - \kappa(\tau_i)]}. \end{cases} \tag{5}$$

Lemma 2 restricts each posterior's support rather than the communication condition space Ω. Different signals may use different condition pairs, allowing signaling over all m conditions.

4.3 Online Environment Estimation

Before episode t, the server uses past observed interactions to estimate the prior $\lambda(\cdot \mid x)$ and controlled transition kernel $P(\cdot \mid x, \tau, a)$. Let $\mathsf{N}_t(x)$ count visits to state x, $\mathsf{N}_t(x, \tau_i)$ count observations of condition τ_i at x, $\mathsf{N}_t(x, \tau_i, a)$ count occurrences of action a under (x, τ_i), and $\mathsf{N}_t(x, \tau_i, a, x')$ count transitions to x' under (x, τ_i, a). The terminal state $x_\dagger$ contributes no count. The prior and transition estimators for $x \in \mathcal{X}_n$ with $n < N - 1$ are

$$\widehat{\lambda}_t(\tau_i \mid x) := \frac{\mathsf{N}_t(x, \tau_i) + 1}{\mathsf{N}_t(x) + m}, \qquad \widehat{P}_t(x' \mid x, \tau_i, a) := \frac{\mathsf{N}_t(x, \tau_i, a, x')}{\max\{1, \mathsf{N}_t(x, \tau_i, a)\}}. \tag{6}$$

Adding one to each condition count and m to the total ensures that $\widehat{\lambda}_t(\cdot \mid x)$ is normalized and strictly positive. The server publishes $\widehat{\lambda}_t(\cdot \mid x)$ to clients in episode t. The prior and transition estimation errors are bounded as follows.

Lemma 3. *With probability at least* $1 - \delta/2$*, the following bounds hold simultaneously:*

$$\|\lambda(\cdot \mid x) - \widehat{\lambda}_t(\cdot \mid x)\|_1 \leq \eta_t(x), \qquad \|P(\cdot \mid x, \tau_i, a) - \widehat{P}_t(\cdot \mid x, \tau_i, a)\|_1 \leq \epsilon_t(x, \tau_i, a), \tag{7}$$

where $\eta_t(x) := \min\{2, \sqrt{2m \log(16T|\mathcal{X}|m/\delta)/\max\{1, \mathsf{N}_t(x)\}} + 2m/(\mathsf{N}_t(x) + m)\}$ *and* $\epsilon_t(x, \tau_i, a) := \min\{2, \sqrt{2(|\mathcal{X}_{n+1}| + 1)\log(32T|\mathcal{X}|m/\delta)/\max\{1, \mathsf{N}_t(x, \tau_i, a)\}}\}$.

The square root terms control sampling uncertainty, the additive term in $\eta_t(x)$ accounts for add-one smoothing, and the clipping value two is the maximum L_1 distance between probability distributions. Let $\mathcal{P}_t$ be the set of stochastic transition kernels $\widetilde{P}$ satisfying

$$\|\widetilde{P}(\cdot \mid x, \tau_i, a) - \widehat{P}_t(\cdot \mid x, \tau_i, a)\|_1 \leq \epsilon_t(x, \tau_i, a). \tag{8}$$

Lemma 3 ensures $P \in \mathcal{P}_t$ on the same event. At $x \in \mathcal{X}_{N-1}$, set $\widehat{P}_t(x_\dagger \mid x, \tau_i, a) = 1$ and $\epsilon_t(x, \tau_i, a) = 0$, since termination is deterministic. $\widetilde{P}_t$ is used only for optimistic planning.

4.4 Online Learning for Markov Persuasion

Under the same policy π_t, let q_t and $\widehat{q}_t$ denote the benchmark and planning occupancies generated by (λ, P) and $(\widehat{\lambda}_t, \widetilde{P}_t)$, respectively. Continuous rewards and posterior expected participation costs

yield an infinite decision space. For $h > 0$, let Γ_h be the uniform h-grid on $[\underline{\gamma}, \overline{\gamma}]$ and $\mathcal{Q}_h$ the set of pairwise posteriors with expected participation costs on the corresponding h-grid. These grids yield a finite posterior occupancy program. The following result bounds the occupancy error and quantifies the constraint relaxations due to grid approximation.

Lemma 4. *Under Assumption 1, for any $\delta \in (0,1)$, with probability at least $1-\delta/2$, we have:*

(i) *The benchmark and estimated occupancies satisfy*

$$\|q_t - \widehat{q}_t\|_1 \le \mathfrak{b}_t(\widehat{q}_t), \tag{9}$$

where $\mathfrak{b}_t(\widehat{q}) := N(1+2L_p\Delta_\kappa)\sum_x \widehat{q}(x)\eta_t(x) + 2N\sum_{x,i,a}\widehat{q}(x,\tau_i,a)\epsilon_t(x,\tau_i,a)$ and $\Delta_\kappa := \max_i \kappa(\tau_i) - \min_i \kappa(\tau_i)$.

(ii) *If a benchmark $\pi = (\gamma, \rho)$ satisfies Eq. (4c), its grid approximation under $(\widehat{\lambda}_t, \widetilde{P})$ satisfies:*

$$B(\widehat{\lambda}_t, \widetilde{P}, \pi) \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}) + [2N + \overline{\gamma}(NL_p(8N+1) + K_{\mathfrak{b}})]\, h, \tag{10}$$

where $K_{\mathfrak{b}} := 8N^3L_p(3+2L_p\Delta_\kappa)$.

(iii) *If a benchmark $\pi = (\gamma, \rho)$ satisfies Eq. (4d), its grid approximation under $(\widehat{\lambda}_t, \widetilde{P})$ satisfies:*

$$J(\widehat{\lambda}_t, \widetilde{P}, \pi) - q_{\min} A(\widehat{\lambda}_t, \widetilde{P}, \pi) \ge -(1+q_{\min})[\mathfrak{b}_t(\widehat{q}) + (4N^2L_p + K_{\mathfrak{b}})h]. \tag{11}$$

The terms involving $\mathfrak{b}_t(\widehat{q})$ account for estimation errors, while those proportional to h account for reward and posterior discretization. The factors $1 + q_{\min}$ and $\overline{\gamma}$ reflect the scales of the participation residual and participation payment, respectively. To approximate **P2** under the estimated environment, we formulate an optimistic linear program for each reward $\gamma \in \Gamma_h$:

$$\mathbf{P3}: \quad \min_{\widehat{q},\, \widetilde{P} \in \mathcal{P}_t} \quad \gamma \sum_{n,x,i,s,x'} \widehat{q}(x,\tau_i,s,1,x') - \overline{\gamma}\mathfrak{b}_t(\widehat{q}) \quad \text{s.t.} \quad \text{Eqs. (4b), (10), and (11).} \tag{12}$$

Under the estimated prior $\widehat{\lambda}_t$, **P3** optimizes the planning occupancy $\widehat{q}$ and transition kernel $\widetilde{P} \in \mathcal{P}_t$. In the marginal consistency constraint, (q_t, λ) is replaced by $(\widehat{q}, \widehat{\lambda}_t)$, and sums over n restrict x to $\mathcal{X}_n$. Proposition 1 recovers the signaling rule from $\widehat{q}$. Lemma 4 justifies the optimistic bonus $\overline{\gamma}\mathfrak{b}_t(\widehat{q})$ and constraint relaxations.

Algorithm 1 initializes four observation counters before the episode loop (line 1). Each episode starts with environment estimation, confidence set construction, and finite grids based on pairwise posteriors (lines 3–4). For each $\gamma \in \Gamma_h$, the server solves P3, recovers the signaling rule, and selects the feasible candidate with the minimum objective (lines 5–7). It then publishes $\widehat{\lambda}_t$ and resets the public state to x_0 (line 8). Execution stops after N arrivals or at $x_\dagger$ (lines 9–10). At each arrival, the server observes the condition and signals the client, who then chooses an action (line 11). Participation triggers local and global OFL updates. The server records the observed transition and updates the public state (lines 12–14), then updates the four counters using the recorded interactions (line 15). Estimation and confidence set construction require $O(m|\mathcal{X}|^2)$ operations per episode. The grid sizes satisfy $|\Gamma_h| = O(1 + (\overline{\gamma} - \underline{\gamma})/h)$ and $|\mathcal{Q}_h| = O(m^2(1 + \Delta_\kappa/h))$, so the algorithm solves $T|\Gamma_h| = O(T(1 + (\overline{\gamma} - \underline{\gamma})/h))$ instances of P3. Each instance is a polynomial time solvable linear program with $O(|\mathcal{X}||\mathcal{Q}_h| + m|\mathcal{X}|^2)$ variables and constraints. Record processing and counter updates require $O(TN)$ operations.

Algorithm 1: Online Markov Persuasion for OFL

Input: T, N, δ, h
1 Initialize $\mathsf{N}_1(x)$, $\mathsf{N}_1(x,\tau_i)$, $\mathsf{N}_1(x,\tau_i,a)$, and $\mathsf{N}_1(x,\tau_i,a,x')$ to 0;
2 **for** $t \in [T]$ **do**
3 Compute $(\widehat{\lambda}_t, \widehat{P}_t, \eta_t, \epsilon_t, \mathcal{P}_t)$ using Eqs. (6)–(8);
4 Construct Γ_h and $\mathcal{Q}_h$ using Lemma 2;
5 **for** $\gamma \in \Gamma_h$ **do**
6 Solve P3 in Eq. (12) over $\mathcal{Q}_h$ and recover ρ from $\widehat{q}$ using Eq. (5);
7 Select a feasible (γ_t, ρ_t) that minimizes the P3 objective;
8 Publish $\widehat{\lambda}_t$ and set $x \leftarrow x_0$;
9 **for** $n = 0, \dots, N-1$ **do**
10 If $x = x_\dagger$, stop the episode;
11 Observe τ_i, send $s \sim \rho_{t,x}(\cdot \mid \tau_i)$, and observe a using Eqs. (2) and (3);
12 Execute the OFL update if $a = 1$;
13 Observe $x' \sim P(\cdot \mid x, \tau_i, a)$ using Eq. (3);
14 Record (x, τ_i, a, x') and set $x \leftarrow x'$;
15 Update $\mathsf{N}_{t+1}(x)$, $\mathsf{N}_{t+1}(x,\tau_i)$, $\mathsf{N}_{t+1}(x,\tau_i,a)$, and $\mathsf{N}_{t+1}(x,\tau_i,a,x')$;

5 THEORETICAL ANALYSIS

We first analyze the signaling benchmark and its finite approximation, and then establish the online learning guarantee by controlling the estimation and occupancy errors.

5.1 GRID APPROXIMATION GUARANTEE

Given λ and P, let C^* denote the minimum expected per-episode cost $C(\lambda, P, \pi)$ over episode-independent mechanisms $\pi = (\gamma, \rho)$ satisfying Bayes plausibility, Bayes benefit, and minimum participation in each episode. We next relate C^* to $\mathbf{P3}$ through a feasible grid approximation.

Theorem 1. *Under Assumption 1, with probability at least $1 - \delta/2$, every episode admits a reward $\gamma \in \Gamma_h$ for which $\mathbf{P3}$ has a feasible solution with objective value at most $C^* + O(h)$.*

This result shows that finite planning preserves a feasible approximation of the best fixed mechanism in $\mathbf{P3}$. Its optimistic objective is at most $O(h)$ above C^*, so refining the grid improves the approximation. The guarantee connects the continuous optimization benchmark to finite computation and provides the foundation for MOVE's online performance analysis.

5.2 ONLINE LEARNING GUARANTEE

The occupancies q_t and $\widehat{q}_t$ correspond to the same mechanism π_t under (λ, P) and $(\widehat{\lambda}_t, \widetilde{P}_t)$, respectively. Algorithm 1 operates under (λ, P), while clients form posteriors using the published prior $\widehat{\lambda}_t$. Let $\bar{q}_t$ denote the resulting executed occupancy. The following result bounds $\|\bar{q}_t - \widehat{q}_t\|_1$ in each episode and the cumulative uncertainty bonus.

Lemma 5. *Under Assumption 1, with probability at least $1 - \delta$, Algorithm 1 satisfies $\|\bar{q}_t - \widehat{q}_t\|_1 \le \mathfrak{b}_t(\widehat{q}_t)$, with the cumulative uncertainty bonus bounded by*

$$\sum_{t=1}^{T} \mathfrak{b}_t(\widehat{q}_t) \le \widetilde{O}(N^2|\mathcal{X}|\sqrt{Tm\log(1/\delta)}). \tag{13}$$

Using the executed occupancy $\bar{q}_t$, define $\bar{A}_t, \bar{J}_t, \bar{C}_t, \bar{B}_t$ as the corresponding executed quantities of A, J, C, B, with the no disclosure response evaluated under $\widehat{\lambda}_t$. These quantities define three measures of cumulative performance over T episodes. The payment pseudo regret $R_T := \sum_{t=1}^{T} \bar{C}_t - TC^*$ measures cumulative excess expected payment relative to the optimal episode-independent mechanism. The participation violation $V_T^{\mathrm{P}} := [\sum_{t=1}^{T}(q_{\min}\bar{A}_t - \bar{J}_t)]_+$ is the positive part of the cumulative participation deficit. Similarly, the Bayes benefit violation $V_T^{\mathrm{B}} := [\sum_{t=1}^{T} \bar{B}_t]_+$ is the positive part of the cumulative Bayes benefit residual. Combining the grid approximation guarantee in Theorem 1 with the cumulative error bound in Lemma 5 yields the following bounds on these three measures.

Theorem 2. *Under Assumption 1, $\forall\, h > 0$, with probability at least $1 - \delta$, Algorithm 1 satisfies*

$$\max\{R_T, V_T^{\mathrm{P}}, V_T^{\mathrm{B}}\} \le \widetilde{O}(N^2|\mathcal{X}|\sqrt{Tm\log(1/\delta)}) + O(Th). \tag{14}$$

Choosing $h = T^{-1/2}$ gives a high probability upper bound of $C^* + \widetilde{O}(T^{-1/2})$ on MOVE's average expected payment, while cumulative participation and Bayes benefit violations divided by T vanish at the same rate. Thus, MOVE learns the unknown environment with diminishing average payment overhead while satisfying participation and Bayes benefit constraints in the long run.

6 EXPERIMENTS

We assess the contributions of signaling and Markov planning through ablation studies, compare MOVE with existing methods, and evaluate its scalability. Additional ablation, robustness, and sensitivity results are provided in Appendix A.9.

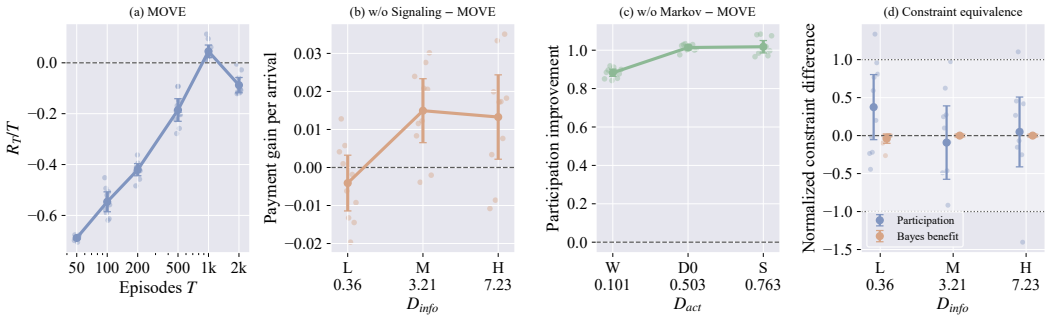


Figure 2: Online performance and ablation studies. (a) Payment pseudo regret per episode R_T/T. (b) Payment gain from signaling. (c) Participation improvement from Markov planning. (d) Constraint equivalence: MOVE versus *w/o Signaling*.

Table 2: Target payment and participation across federated datasets.

Method	FEMNIST		CIFAR-10		Shakespeare	
	Target payment	Participation J/A	Target payment	Participation J/A	Target payment	Participation J/A
MOVE	1108.2 ± 51.7	0.582 ± 0.003	209.4 ± 27.6	0.585 ± 0.004	95.8 ± 9.0	0.582 ± 0.009
DaringFed	1816.7 ± 60.0	0.930 ± 0.000	369.6 ± 74.1	0.931 ± 0.000	162.2 ± 22.5	0.950 ± 0.000
Dynamic Pricing	1124.5 ± 25.0	0.530 ± 0.002	223.6 ± 41.0	0.550 ± 0.019	102.8 ± 27.6	0.570 ± 0.038
L2I	1344.8 ± 59.5	0.726 ± 0.002	291.3 ± 37.5	0.786 ± 0.009	132.4 ± 22.7	0.833 ± 0.017

6.1 Experimental Setup

We evaluate MOVE in simulated network environments with stochastic participation and state transitions affected by participation. Federated experiments use FEMNIST, CIFAR-10, and Shakespeare with CNN, ResNet-18, and LSTM models, respectively. FEMNIST and Shakespeare use natural LEAF partitions, while CIFAR-10 uses Dirichlet partitions across 20 clients with $\alpha = 0.3$. Each participating client performs one local epoch followed by a server update. Methods share initialization, data partitions, client sampling, environments, and episode budgets within each seed. Target payment is measured at the first target accuracy crossing, and participation and payment metrics use conditional expected executed occupancy. Appendix A.9.1 provides experimental details.

6.2 Ablation Studies

We compare MOVE with two structural ablations using 10 seeds. *w/o Markov* uses an action-independent transition kernel, while *w/o Signaling* uses one uninformative signal. Translucent points are individual seeds and error bars are 95% confidence intervals. Figure 2(a) shows R_T/T approaching zero, consistent with the regret bound in Theorem 2. In Figure 2(b), D_{info} measures hidden cost differences. The payment difference per arrival is *w/o Signaling* minus MOVE, so positive values favor MOVE. MOVE reduces mean payment per arrival by 5.26% and 4.82% relative to *w/o Signaling* at Medium and High, respectively, while the gain at Low remains unresolved. In Figure 2(c), D_{act} measures how strongly participation affects the next state distribution. W, D0, and S represent weak, default, and strong action coupling. MOVE reduces mean participation violation per episode by approximately 61.7%, 69.8%, and 69.3% relative to *w/o Markov* in W, D0, and S, respectively. In Figure 2(d), all confidence intervals for participation and Bayes benefit differences between MOVE and *w/o Signaling* lie within $[-1, 1]$, supporting comparable constraint performance. Detailed statistics appear in Appendix A.9.2.

6.3 Baseline Comparisons

All baselines are adapted to the common participation setting. We compare MOVE with *DaringFed* (Xin et al., 2025), *Dynamic Pricing* (Wang et al., 2024), and *L2I* (Wang et al., 2026) over 5 seeds on FEMNIST (Caldas et al., 2018), CIFAR-10 (Krizhevsky & Hinton, 2009), and Shakespeare (McMahan et al., 2017), with respective accuracy targets of 80%, 70%, and 50%. Table 2 shows that MOVE achieves the lowest mean target payment across all three datasets, with participation close to $q_{\min} = 0.5$. Figure 3 shows the corresponding accuracy versus payment trajectories, with seed

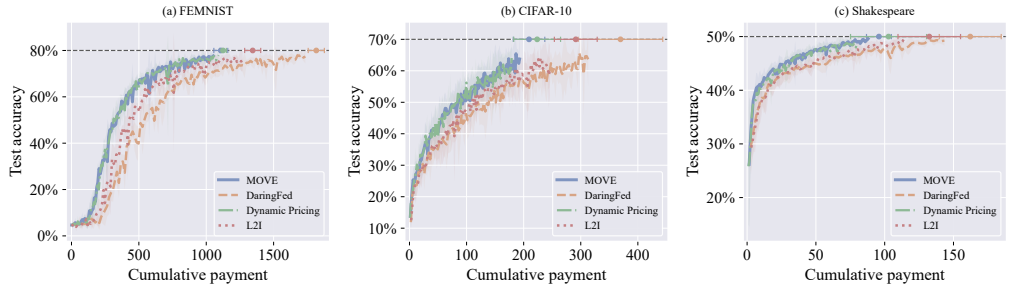


Figure 3: Test accuracy versus cumulative payment on (a) FEMNIST, (b) CIFAR-10, and (c) Shakespeare. Markers indicate the first target accuracy crossing.

Table 3: P3 scalability.

N	Time ↑	Atoms —	Rewards —	P3 size ↑ / ↑ / ↑	$\lvert\mathcal{X}\rvert$	Time ↑	Atoms —	Rewards —	P3 size ↑ / ↑ / ↑
2	0.0547 ± 0.0004	26	19	276/64/170	3	0.0490 ± 0.0002	26	19	171/48/92
3	0.0940 ± 0.0006	26	19	414/96/254	6	0.0944 ± 0.0013	26	19	414/96/254
4	0.1453 ± 0.0017	26	19	552/128/338	9	0.1668 ± 0.0012	26	19	729/144/488
5	0.1953 ± 0.0021	26	19	690/160/422	12	0.2904 ± 0.0023	26	19	1116/192/794
m	Time ↑	Atoms ↑	Rewards —	P3 size ↑ / ↑ / ↑	h	Time ↓	Atoms ↓	Rewards ↓	P3 size ↓/—/—
2	0.0613 ± 0.0005	14	19	258/66/170	0.02	0.2667 ± 0.0024	55	42	588/96/254
3	0.0934 ± 0.0006	26	19	414/96/254	0.03	0.1649 ± 0.0013	39	30	492/96/254
4	0.1381 ± 0.0012	41	19	588/126/338	0.04	0.1102 ± 0.0008	28	22	426/96/254
5	0.1896 ± 0.0019	61	19	792/156/422	0.07	0.0622 ± 0.0005	19	14	372/96/254
7	0.3348 ± 0.0032	110	19	1254/216/590	0.10	0.0418 ± 0.0004	14	10	342/96/254

means, 95% confidence bands, and markers indicating target payments. Across FEMNIST, CIFAR-10, and Shakespeare, MOVE reduces mean target payment by 39.00%, 43.35%, and 40.93% relative to *DaringFed* and by 17.59%, 28.11%, and 27.65% relative to *L2I*, respectively. Mean reductions versus *Dynamic Pricing* are 1.45%, 6.37%, and 6.81%, but paired differences remain statistically unresolved. *DaringFed* ignores participation effects on future states and guards against the worst resource cost, inducing excess participation. *L2I* uses a fixed reward schedule that cannot adapt to disclosure or evolving states. *Dynamic Pricing* compensates cumulative participation shortfall without using disclosed information. MOVE uses disclosure and accounts for participation effects on future states to reduce payments while sustaining participation.

6.4 Scalability Analysis

We evaluate scalability over 10 seeds and 200 planning episodes per setting, varying N, $|\mathcal{X}|$, m, or h individually. The m sweep varies the number of equally spaced costs over a fixed interval. Table 3 shows that larger N, $|\mathcal{X}|$, or m increases problem size and planning time, whereas larger h reduces planning time through coarser grids with fewer posterior atoms and candidate rewards. All 180 runs completed planning without fallback. Times report planning only as means ± 95% confidence interval half widths. **P3** size lists variables, equalities, and inequalities.

7 Conclusion

We proposed MOVE, an online Markov persuasion framework for OFL participation control under asymmetric information and unknown dynamics. Its core idea is to coordinate rewards with disclosure while accounting for how participation shapes future states and the observations available for learning. Through an occupancy formulation, we establish that posteriors supported on at most two communication conditions suffice for optimality. This structure enables finite optimistic linear programs that account for environment uncertainty. With high probability, MOVE achieves sublinear payment pseudo regret relative to the best fixed feasible mechanism and sublinear cumulative participation and Bayes benefit violations. Experiments support the benefits of disclosure and dynamic planning for payment efficiency and participation, alongside computational practicality.

AI USE STATEMENT

Generative AI helped us improve the language and presentation of the manuscript by refining wording, correcting grammar, and removing repetition. We reviewed all suggested changes and made the final decisions on the paper's content.

ETHICS STATEMENT

This study uses benchmark datasets and simulated incentives. It does not collect new data from human participants or involve real payments. Any practical deployment should protect client data, ensure informed participation, and follow applicable regulations.

REPRODUCIBILITY STATEMENT

The main paper describes the assumptions and algorithm. Appendices A.1–A.8 contain the complete proofs, while Appendix A.9 provides the experimental details. The supplementary material provides the implementation and experimental records needed to reproduce the reported results.

REFERENCES

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online Bayesian Persuasion. In *Advances in Neural Information Processing Systems 33*, 2020.

Matteo Castiglioni, Alberto Marchesi, Andrea Celli, and Nicola Gatti. Multi-Receiver Online Bayesian Persuasion. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1314–1323, 2021.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Optimal Rates and Efficient Algorithms for Online Bayesian Persuasion. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2164–2183, 2023.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential Information Design: Markov Persuasion Process and its Efficient Reinforcement Learning. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 471–472, 2022.

Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Markov Persuasion Processes: Learning to Persuade from Scratch. In *Advances in Neural Information Processing Systems 38*, 2025.

Maria-Florina Balcan, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Keegan Harris, and Zhiwei Steven Wu. Nearly-Optimal Bandit Learning in Stackelberg Games with Side Information. In *International Conference on Learning Representations*, 2026.

Jinlong Pang, Jieling Yu, Ruiting Zhou, and John C. S. Lui. An Incentive Auction for Heterogeneous Client Selection in Federated Learning. *IEEE Transactions on Mobile Computing*, 22(10):5733–5750, 2023.

Aniket Murhekar, Zhuowen Yuan, Bhaskar Ray Chaudhury, Bo Li, and Ruta Mehta. Incentives in Federated Learning: Equilibria, Dynamics, and Mechanisms for Welfare Maximization. In *Advances in Neural Information Processing Systems 36*, 2023.

Ningning Ding, Lin Gao, and Jianwei Huang. Incentive Mechanism Design for Federated Learning with Dynamic Network Pricing. *IEEE Transactions on Mobile Computing*, 24(8):7206–7222, 2025.

Miao Hu, Wenzhuo Yang, Zhenxiao Luo, Xuezheng Liu, Yipeng Zhou, Xu Chen, and Di Wu. AutoFL: A Bayesian Game Approach for Autonomous Client Participation in Federated Edge Learning. *IEEE Transactions on Mobile Computing*, 23(1):194–208, 2024.

Yun Xin, Jianfeng Lu, Shuqin Cao, Gang Li, Haozhao Wang, and Guanghui Wen. DaringFed: A Dynamic Bayesian Persuasion Pricing for Online Federated Learning under Two-Sided Incomplete Information. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6687–6695, 2025.

Yujing Chen, Yue Ning, Martin Slawski, and Huzefa Rangwala. Asynchronous Online Federated Learning for Edge Devices with Non-IID Data. In *2020 IEEE International Conference on Big Data*, pages 15–24, 2020.

Hongbin Zhu, Yong Zhou, Hua Qian, Yuanming Shi, Xu Chen, and Yang Yang. Online Client Selection for Asynchronous Federated Learning with Fairness Consideration. *IEEE Transactions on Wireless Communications*, 22(4):2493–2506, 2023.

Shipra Agrawal, Yiding Feng, and Wei Tang. Dynamic Pricing and Learning with Bayesian Persuasion. In *Advances in Neural Information Processing Systems 36*, 2023.

Miao Hu, Di Wu, Yipeng Zhou, Xu Chen, and Min Chen. Incentive-Aware Autonomous Client Participation in Federated Learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(10):2612–2628, 2022.

Ningning Ding, Lin Gao, and Jianwei Huang. Joint Participation Incentive and Network Pricing Design for Federated Learning. In *IEEE INFOCOM 2023–IEEE Conference on Computer Communications*, 2023.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Asynchronous Federated Optimization. *arXiv preprint arXiv:1903.03934*, 2019.

Xiaoli Tang, Han Yu, and Xiaoxiao Li. A Reinforcement Learning-Based Bidding Strategy for Data Consumers in Auction-Based Federated Learning. In *Advances in Neural Information Processing Systems 38*, 2025.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online Bayesian Persuasion without a Clue. In *Advances in Neural Information Processing Systems 37*, 2024.

Jiajun Wang, Jianxiong Guo, Juncheng Wang, Xingjian Ding, Deying Li, and Weili Wu. Learning to Incentivize: Convergence-Guaranteed Federated Learning via Client Quality Discovery. *IEEE Transactions on Mobile Computing*, 25(5):6767–6783, 2026.

Xuehe Wang, Shensheng Zheng, and Lingjie Duan. Dynamic Pricing for Client Recruitment in Federated Learning. *IEEE/ACM Transactions on Networking*, 32(2):1273–1286, 2024.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv preprint arXiv:1812.01097*, 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Wenjing Hou, Hong Wen, Ning Zhang, Wenxin Lei, Haojie Lin, Zhu Han, and Qiang Liu. Adaptive Training and Aggregation for Federated Learning in Multi-Tier Computing Networks. *IEEE Transactions on Mobile Computing*, 23(5):4376–4388, 2024.

Shuai Luo, Qiming Wan, Hongrui Wang, Tianchun Xiang, Yang Wang, Xin He, and Wei Zhang. Data-Oriented Optimized Nonuniform Quantization for CR-Enhanced Communication Efficiency in Federated Learning. *EURASIP Journal on Wireless Communications and Networking*, 2025:40, 2025.

Zhepei Wei, Chuanhao Li, Haifeng Xu, and Hongning Wang. Incentivized Communication for Federated Bandits. In *Advances in Neural Information Processing Systems 36*, 2023.

Jiarui Gan, Rupak Majumdar, Debmalya Mandal, and Goran Radanovic. Stochastic Principal-Agent Problems: Computing and Learning Optimal History-Dependent Policies. In *Advances in Neural Information Processing Systems 38*, 2025.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online Learning in CMDPs: Handling Stochastic and Adversarial Constraints. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46692–46721, 2024.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, pp. 2071–2080, 2017.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR, vol. 119, pp. 3052–3060, 2020.

A Appendix

A.1 Proof of Lemma 1

Set $\rho_x(s_0 \mid \tau_i) = 1$ and $\gamma = \overline{\gamma}$ in every episode. The posterior at state x is then $\lambda(\cdot \mid x)$, and

$$J - q_{\min} A = \sum_x q(x) \left[p\left(\overline{\gamma}, \sum_i \lambda(\tau_i \mid x) \kappa(\tau_i) \right) - q_{\min} \right] \geq 0. \tag{15}$$

The service attainability condition following $\mathbf{P2}$ gives the inequality above. No disclosure also gives $B = 0$, so this policy is feasible. For fixed λ and P, a signaling rule can be represented by a probability measure on the posterior simplex with barycenter $\lambda(\cdot \mid x)$ at each state x. The set of such measures is compact under weak convergence, as is the reward interval. Under Assumption 1, the finite horizon occupancy recursion makes the objective and constraints continuous in these choices. Hence the feasible set is nonempty and compact, and the minimum is attained.

A.2 Proof of Lemma 2

Fix $\bar{\kappa}$ and consider

$$\mathcal{M}_\kappa(\bar{\kappa}) := \left\{ \mu \in \Delta_m : \sum_i \mu_i \kappa(\tau_i) = \bar{\kappa} \right\}. \tag{16}$$

The two equalities defining $\mathcal{M}_\kappa(\bar{\kappa})$ fix total probability and expected cost. A point with at least three positive coordinates admits a nonzero perturbation preserving both equalities, so it is not extreme. The same holds for two positive coordinates with equal costs. Thus every extreme point has one or two positive coordinates, with distinct costs in the latter case. Since $\mathcal{M}_\kappa(\bar{\kappa})$ is a compact polytope, every posterior in it is a convex combination of these points.

Take a signal of mass ω and posterior μ at state x, and write $\mu = \sum_k \alpha_k \mu^k$ using extreme points of $\mathcal{M}_\kappa(\bar{\kappa})$. Replace it with signals of masses $\omega\alpha_k$ and posteriors μ^k. For each condition τ_i, their combined mass remains $\omega \sum_k \alpha_k \mu_i^k = \omega \mu_i$, including when its prior probability is zero. Every μ^k has expected cost $\bar{\kappa}$, so the participation probability is unchanged. The condition and action masses, and therefore the state transitions and termination probabilities, remain unchanged. Repeating the refinement at each stage preserves the occupancy measures and A, C, J, B. Applying it to an optimizer yields an equally good mechanism whose posteriors each support at most two conditions.

A.3 Proof of Proposition 1

Fix episode t, state x, condition τ_i, and signal s. Define $y_{t,x,i,s} := \sum_{a,x'} q_t(x, \tau_i, s, a, x')$, and reserve $s_1, \dots, s_m$ for singleton posterior labels. The signaling rule recovered from the occupancy measure is

$$\rho_{t,x}(s \mid \tau_i) = \begin{cases} \dfrac{y_{t,x,i,s}}{q_t(x)\lambda(\tau_i \mid x)}, & q_t(x)\lambda(\tau_i \mid x) > 0, \\ 1, & q_t(x)\lambda(\tau_i \mid x) = 0,\ s = s_i, \\ 0, & q_t(x)\lambda(\tau_i \mid x) = 0,\ s \neq s_i. \end{cases} \tag{17}$$

Feasibility gives $y_{t,x,i,s} \geq 0$ and $\sum_s y_{t,x,i,s} = q_t(x)\lambda(\tau_i \mid x)$. Thus the first branch is a probability distribution when its denominator is positive. When the denominator is zero, every $y_{t,x,i,s}$ is zero, and the singleton branch is a probability distribution. In either case,

$$q_t(x)\lambda(\tau_i \mid x)\rho_{t,x}(s \mid \tau_i) = \sum_{a,x'} q_t(x, \tau_i, s, a, x'). \tag{18}$$

For a signal with positive mass at a state with $q_t(x) > 0$, Bayes' rule and the identity above give

$$\frac{\lambda(\tau_i \mid x)\rho_{t,x}(s \mid \tau_i)}{\sum_j \lambda(\tau_j \mid x)\rho_{t,x}(s \mid \tau_j)} = \frac{y_{t,x,i,s}}{\sum_j y_{t,x,j,s}}, \tag{19}$$

For $q_t(x) > 0$, the induced signal probabilities satisfy $\phi_{t,x}(s) = \sum_j y_{t,x,j,s}/q_t(x)$. Therefore,

$$\sum_s \phi_{t,x}(s)\mu_{t,x}(\tau_i \mid s) = \frac{\sum_s y_{t,x,i,s}}{q_t(x)} = \lambda(\tau_i \mid x). \tag{20}$$

Conversely, multiplying the posterior averaging identity by $q_t(x)$ recovers Eq. (4b). For a signal supported on τ_i and τ_j with $\kappa(\tau_i) < \kappa(\tau_j)$, we also have $\mu_{t,x}(\tau_j \mid s) = 1 - \mu_{t,x}(\tau_i \mid s)$. Solving $\bar{\kappa}_{t,x}(s) = \mu_{t,x}(\tau_i \mid s)\kappa(\tau_i) + \mu_{t,x}(\tau_j \mid s)\kappa(\tau_j)$ gives

$$\mu_{t,x}(\tau_i \mid s) = \frac{\kappa(\tau_j) - \bar{\kappa}_{t,x}(s)}{\kappa(\tau_j) - \kappa(\tau_i)}, \qquad \mu_{t,x}(\tau_j \mid s) = \frac{\bar{\kappa}_{t,x}(s) - \kappa(\tau_i)}{\kappa(\tau_j) - \kappa(\tau_i)}. \tag{21}$$

Substituting these posterior probabilities into the recovery rule yields the formula in Proposition 1. Since a feasible occupancy satisfies the action response and transition constraints, the recovered channel reproduces its action and successor state distributions. The recovered signaling rule, together with the corresponding rewards, therefore implements an optimal solution to $\mathbf{P2}$.

A.4 Proof of Lemma 3

Fix x and condition on its successive visits. The observed communication conditions have conditional law $\lambda(\cdot \mid x)$. For a positive count r, let $\bar{\lambda}_r$ be their empirical distribution. The sign vector identity $\|z\|_1 = \max_{v \in \{-1,1\}^m} v^\top z$ and Hoeffding's inequality give

$$\Pr(\|\bar{\lambda}_r - \lambda\|_1 \geq u) \leq 2^m \exp(-ru^2/2). \tag{22}$$

Taking $u = \sqrt{2m \log(16T|\mathcal{X}|m/\delta)/r}$ and applying a union bound over counts, episodes, and states yields failure probability at most $\delta/8$. For $r \geq 1$,

$$\widehat{\lambda}_t(\cdot \mid x) = \frac{r}{r+m}\bar{\lambda}_r + \frac{m}{r+m}(1/m, \dots, 1/m), \qquad \|\widehat{\lambda}_t(\cdot \mid x) - \bar{\lambda}_r\|_1 \leq \frac{2m}{r+m}. \tag{23}$$

Combining the empirical bound with the second inequality above gives the stated radius after clipping at two. When $r = 0$, the estimate is uniform and the radius is two. Thus the bound also allows zero probabilities in the true prior. The prior confidence event fails with probability at most $\delta/4$.

For a transition row observed $r \geq 1$ times, the same concentration argument applies to its successor states. A union bound over rows and visit counts gives the stated transition radius. When $r = 0$, the estimate is the zero vector, whose distance from the true stochastic row is one and is covered by the radius two. At the final stage, both rows terminate with probability one, so their distance is zero. Early termination only reduces visit counts. The prior and transition confidence events each fail with probability at most $\delta/4$, giving the joint probability in Lemma 3.

A.5 Proof of Lemma 4

Write $\kappa_{\min} := \min_i \kappa(\tau_i)$, $\kappa_{\max} := \max_i \kappa(\tau_i)$, $\Delta_\kappa := \kappa_{\max} - \kappa_{\min}$, and $c_\kappa := 1 + 2L_p\Delta_\kappa$. Lemma 3 ensures that both confidence bounds hold simultaneously with probability at least $1-\delta/2$. Fix t, $\widetilde{P} \in \mathcal{P}_t$, and π. For a prior ν, define $z^\nu_{i,s} := \nu(\tau_i \mid x)\rho_x(s \mid \tau_i)$ and $w^\nu_s := \sum_i z^\nu_{i,s}$. A posterior expected cost attached to a zero probability signal may be assigned any value in $[\kappa_{\min}, \kappa_{\max}]$. With $c_0 := (\kappa_{\max} + \kappa_{\min})/2$, signal mass weighting gives, whenever $w^{\nu'}_s > 0$,

$$\begin{aligned} & w^\nu_s |\bar\kappa^\nu_x(s) - \bar\kappa^{\nu'}_x(s)| \\ & \quad \le \left| \sum_i (z^\nu_{i,s} - z^{\nu'}_{i,s})(\kappa(\tau_i) - c_0) \right| + |w^\nu_s - w^{\nu'}_s| \frac{\Delta_\kappa}{2} \\ & \quad \le \Delta_\kappa \sum_i |z^\nu_{i,s} - z^{\nu'}_{i,s}|. \end{aligned} \tag{24}$$

If $w^{\nu'}_s = 0$, the same bound follows directly from $w^\nu_s |\bar\kappa^\nu_x(s) - \bar\kappa^{\nu'}_x(s)| \le \Delta_\kappa w^\nu_s$. Summing over signals and using that every channel row sums to one yields

$$\sum_s w^\nu_s |\bar\kappa^\nu_x(s) - \bar\kappa^{\nu'}_x(s)| \le \Delta_\kappa \|\nu - \nu'\|_1. \tag{25}$$

For a positive mass signal under published prior b, let $\bar\kappa^b_x(s) := \sum_i b(\tau_i \mid x)\rho_x(s \mid \tau_i)\kappa(\tau_i) / \sum_j b(\tau_j \mid x)\rho_x(s \mid \tau_j)$ and $\chi^b_a(x,s) := p(\gamma, \bar\kappa^b_x(s))^a [1 - p(\gamma, \bar\kappa^b_x(s))]^{1-a}$. Define $G^{\nu,b}_x(i,s,a) := \nu(\tau_i \mid x)\rho_x(s \mid \tau_i)\chi^b_a(x,s)$ as the outcome signal action kernel when the outcome law is ν and the published prior is b. For the benchmark, decompose

$$G^{\lambda,\lambda}_x - G^{\widehat\lambda_t,\widehat\lambda_t}_x = \big(G^{\lambda,\lambda}_x - G^{\widehat\lambda_t,\lambda}_x\big) + \big(G^{\widehat\lambda_t,\lambda}_x - G^{\widehat\lambda_t,\widehat\lambda_t}_x\big). \tag{26}$$

The first difference changes only the outcome law and has L_1 norm at most $\eta_t(x)$. The second difference is weighted by the signal masses of $\widehat\lambda_t$. The L_1 distance between two Bernoulli action laws is twice the difference between their participation probabilities. Applying Assumption 1 and the preceding posterior sensitivity inequality with $\widehat\lambda_t$ as its first prior gives

$$\begin{aligned} & \|G^{\widehat\lambda_t,\lambda}_x - G^{\widehat\lambda_t,\widehat\lambda_t}_x\|_1 \\ & \quad \le 2L_p \sum_s w^{\widehat\lambda_t}_s |\bar\kappa^\lambda_x(s) - \bar\kappa^{\widehat\lambda_t}_x(s)| \le 2L_p\Delta_\kappa\eta_t(x). \end{aligned} \tag{27}$$

Consequently,

$$\|G^{\lambda,\lambda}_x - G^{\widehat\lambda_t,\widehat\lambda_t}_x\|_1 \le c_\kappa \eta_t(x). \tag{28}$$

On the confidence event, $P, \widetilde{P} \in \mathcal{P}_t$, and hence $\|P(\cdot \mid x,\tau_i,a) - \widetilde{P}(\cdot \mid x,\tau_i,a)\|_1 \le 2\epsilon_t(x,\tau_i,a)$. Applying Eq. (28) together with this transition bound in the layerwise simulation recursion and using planning visitation probabilities on the right gives

$$\|q_t - \widehat{q}_t\|_1 \le N \left[c_\kappa \sum_x \widehat{q}_t(x)\eta_t(x) + 2 \sum_{x,i,a} \widehat{q}_t(x,\tau_i,a)\epsilon_t(x,\tau_i,a) \right] = \mathfrak{b}_t(\widehat{q}_t). \tag{29}$$

This proves part (i) of Lemma 4.

For the Bayes benefit comparison, define the no disclosure response $r_b(x) := p(\gamma, \sum_i b(\tau_i \mid x)\kappa(\tau_i))$. Since probability-vector differences sum to zero,

$$|r_\lambda(x) - r_{\widehat\lambda_t}(x)| \le \frac{L_p\Delta_\kappa}{2} \|\lambda(\cdot \mid x) - \widehat\lambda_t(\cdot \mid x)\|_1. \tag{30}$$

For the benchmark, the signal induced participation term changes by at most $(\frac{1}{2} + L_p\Delta_\kappa)\eta_t(x)$: the outcome-law replacement contributes at most $\eta_t(x)/2$, while the published prior replacement contributes at most $L_p\Delta_\kappa\eta_t(x)$. Adding Eq. (30) gives coefficient $\frac{1}{2} + \frac{3}{2}L_p\Delta_\kappa \le c_\kappa$. Applying the

layerwise performance-difference recursion directly to the bounded stage residual $\gamma(a - r_b(x))$ and the controlled transition comparison therefore yields

$$|B(\lambda, P, \pi) - B(\widehat{\lambda}_t, \widetilde{P}, \pi)| \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}_t). \tag{31}$$

For part (ii), fix a benchmark feasible in the episode under consideration and write $B^* := B(\lambda, P, \pi) \le 0$. Let $\widehat{B}^0, \widehat{q}^0$ denote the planning functional and occupancy of the same mechanism before grid rounding. Eq. (31) gives

$$\widehat{B}^0 \le B^* + \overline{\gamma}\mathfrak{b}_t(\widehat{q}^0) \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}^0). \tag{32}$$

Let $B^h, \widehat{q}^h$ denote the planning functional and occupancy after pairwise refinement and grid rounding. Then

$$B^h \le \widehat{B}^0 + K_{\mathrm{ben}}h \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}^h) + (K_{\mathrm{ben}} + \overline{\gamma}K_{\mathfrak{b}})h. \tag{33}$$

The grid error bounds in Appendix A.6 account for reward rounding, occupancy perturbation, and the no disclosure response. Substituting K_{ben} and the bonus bound from Eq. (39) gives Eq. (10).

For part (iii), fix a benchmark satisfying $J^* - q_{\min}A^* \ge 0$ in the episode under consideration, where $J^* := J(\lambda, P, \pi)$ and $A^* := A(\lambda, P, \pi)$. Evaluate the same mechanism in the planning model before grid rounding and denote its functionals and occupancy by $\widehat{J}^0, \widehat{A}^0, \widehat{q}^0$. Part (i) gives

$$\widehat{J}^0 - q_{\min}\widehat{A}^0 \ge -(1 + q_{\min})\mathfrak{b}_t(\widehat{q}^0). \tag{34}$$

Pairwise refinement preserves these planning functionals. Let $J^h, A^h, \widehat{q}^h$ denote the functionals and occupancy after grid rounding. The grid approximation bound gives

$$\left|(\widehat{J}^0 - q_{\min}\widehat{A}^0) - (J^h - q_{\min}A^h)\right| \le K_{\mathrm{part}}h. \tag{35}$$

Combining this inequality with Eq. (39) yields

$$J^h - q_{\min}A^h \ge -(1 + q_{\min})\mathfrak{b}_t(\widehat{q}^h) - \big(K_{\mathrm{part}} + (1 + q_{\min})K_{\mathfrak{b}}\big)h. \tag{36}$$

Substituting $K_{\mathrm{part}} = 4(1 + q_{\min})N^2L_p$ proves Eq. (11). □

A.6 Proof of Theorem 1

For the grid construction, fix $i, j \in [m]$ with $\kappa(\tau_i) < \kappa(\tau_j)$ and $\bar{\kappa} \in [\kappa(\tau_i), \kappa(\tau_j)]$. The binary posterior is

$$\mu_i = \frac{\kappa(\tau_j) - \bar{\kappa}}{\kappa(\tau_j) - \kappa(\tau_i)}, \qquad \mu_j = \frac{\bar{\kappa} - \kappa(\tau_i)}{\kappa(\tau_j) - \kappa(\tau_i)}. \tag{37}$$

Equal cost pairs add no new extreme posteriors. Place an h grid, including its endpoints, on each remaining cost interval. Split each posterior whose expected cost is not on the grid between its adjacent grid atoms. Because the posterior coordinates are affine in expected cost, this preserves the posterior barycenter and Bayes plausibility. Rounding reward and expected cost changes each by at most h. Assumption 1 then bounds the change in each action probability by $2L_ph$, and the corresponding Bernoulli laws differ by at most $4L_ph$ in L_1 distance. Coupling the layers gives an occupancy error of at most K_qh, where $K_q := 4N^2L_p$. Rounding γ upward to Γ_h changes it by at most h. The functional errors are bounded by K_Ch, $K_{\mathrm{part}}h$, and $K_{\mathrm{ben}}h$, where

$$\begin{cases} K_C := N + \overline{\gamma}K_q, \\ K_{\mathrm{part}} := (1 + q_{\min})K_q, \\ K_{\mathrm{ben}} := 2N + 2\overline{\gamma}K_q + \overline{\gamma}NL_p, \\ K_{\mathfrak{b}} := 2N(3 + 2L_p\Delta_\kappa)K_q. \end{cases} \tag{38}$$

K_C accounts for reward rounding and occupancy error. K_{ben} also accounts for the no disclosure response. Its expected cost is unchanged because the grid preserves the prior barycenter. Since $\eta_t, \epsilon_t \le 2$ and the bonus is linear in occupancy, the planning occupancies before and after rounding, $\widehat{q}^0$ and $\widehat{q}^h$, satisfy

$$|\mathfrak{b}_t(\widehat{q}^0) - \mathfrak{b}_t(\widehat{q}^h)| \le K_{\mathfrak{b}}h. \tag{39}$$

To express $\mathbf{P3}$ as a finite posterior occupancy program, fix $\gamma \in \Gamma_h$ and enumerate the pairwise atoms in $\mathcal{Q}_h$ by k. Let q_x, $y_{x,k}$, $v_{x,i,a}$, and $z_{x,i,a,x'}$ be the nonnegative state, signal atom, outcome action, and transition flow marginals, respectively. Set $q_{x_0} = 1$. With atom posterior μ_i^k and response $p_k = p(\gamma, \sum_i \mu_i^k \kappa(\tau_i))$, impose

$$\begin{cases} \displaystyle\sum_k y_{x,k}\mu_i^k = q_x \widehat{\lambda}_t(\tau_i \mid x), \\ v_{x,i,1} = \displaystyle\sum_k y_{x,k}\mu_i^k p_k, \\ v_{x,i,0} = q_x \widehat{\lambda}_t(\tau_i \mid x) - v_{x,i,1}, \\ \displaystyle\sum_{x'} z_{x,i,a,x'} = v_{x,i,a}, \\ q_{x'} = \displaystyle\sum_{x,i,a} z_{x,i,a,x'}. \end{cases} \tag{40}$$

For $x \in \mathcal{X}_n$, the transition sum is over $x' \in \mathcal{X}_{n+1} \cup \{x_\dagger\}$ when $n < N-1$ and $x' = x_\dagger$ when $n = N-1$. The last flow equality applies to $x' \in \mathcal{X}_{n+1}$ when $n < N-1$. To encode an optimistic transition row, introduce nonnegative deviations $e_{x,i,a,x'}$ satisfying

$$\begin{cases} e_{x,i,a,x'} \geq \left| z_{x,i,a,x'} - \widehat{P}_t(x' \mid x, \tau_i, a) v_{x,i,a} \right|, \\ \displaystyle\sum_{x'} e_{x,i,a,x'} \leq \epsilon_t(x, \tau_i, a) v_{x,i,a}. \end{cases} \tag{41}$$

For $v_{x,i,a} > 0$, the row $\widetilde{P}(x' \mid x, \tau_i, a) := z_{x,i,a,x'}/v_{x,i,a}$ belongs to $\mathcal{P}_t$ by the perspective constraint. If $v_{x,i,a} = 0$, the transition row does not affect the occupancy and can be chosen from $\mathcal{P}_t$. Conversely, every feasible $(\widehat{q}, \widetilde{P})$ admits this representation by setting $z_{x,i,a,x'} := v_{x,i,a}\widetilde{P}(x' \mid x, \tau_i, a)$ and taking $e_{x,i,a,x'}$ as the corresponding absolute deviation. Thus, the extended representation is equivalent to the occupancy and transition variables in $\mathbf{P3}$. The objective, relaxed constraints, flow equations, and uncertainty bonus are linear, so the fixed reward problem is a linear program.

The number of posterior atoms is at most $m + \binom{m}{2}(\lceil (\max_i \kappa(\tau_i) - \min_i \kappa(\tau_i))/h \rceil + 1)$, and the program has $O(|\mathcal{X}||\mathcal{Q}_h| + |\mathcal{X}|^2 m)$ variables and constraints. Enumerating $O(1 + (\overline{\gamma} - \underline{\gamma})/h)$ rewards therefore terminates and is polynomial in the displayed finite dimensions and $1/h$ for a fixed bounded cost range, provided that p can be evaluated to the required precision.

It remains to construct the fixed comparator. The known environment comparator is

$$\begin{aligned} C^* := \min_{\gamma, \rho} \quad & C(\lambda, P, \pi), \\ \text{s.t. } & \gamma \in [\underline{\gamma}, \overline{\gamma}], \\ & J(\lambda, P, \pi) \geq q_{\min} A(\lambda, P, \pi), \\ & B(\lambda, P, \pi) \leq 0, \end{aligned} \tag{42}$$

where ρ ranges over state dependent stochastic channels. By Lemma 1 and the service attainability condition, the fixed mechanism feasible set is nonempty and compact. Hence C^* is attained.

Let π^* attain Eq. (42). On the simultaneous confidence event, $P \in \mathcal{P}_t$. Evaluate π^* in the planning model with published prior $\widehat{\lambda}_t$, and denote its occupancy before grid rounding by $\widehat{q}_t^0$. Part (i) of Lemma 4 bounds the difference between the benchmark and planning payments. Pairwise refinement under $\widehat{\lambda}_t$ preserves the planning objective and constraints exactly. Let $(\gamma_t^h, \widehat{q}_t^h)$ be the comparator obtained by rounding the reward and posterior expected costs, and define $K_{\mathrm{grid}} := K_C + \overline{\gamma} K_{\mathfrak{b}}$. Eqs. (38) and (39) give

$$\gamma_t^h \sum_{n,x,i,s,x'} \widehat{q}_t^h(x, \tau_i, s, 1, x') - \overline{\gamma}\mathfrak{b}_t(\widehat{q}_t^h) \leq C^* + K_{\mathrm{grid}} h. \tag{43}$$

Parts (ii) and (iii) of Lemma 4 show that the rounded mechanism satisfies the two relaxed constraints and is therefore feasible for $\mathbf{P3}$.

A.7 Proof of Lemma 5

For the executed mechanism, the published prior is $\widehat{\lambda}_t$ in both the true and planning models, so the posterior response is unchanged and

$$\|G_x^{\lambda,\widehat{\lambda}_t} - G_x^{\widehat{\lambda}_t,\widehat{\lambda}_t}\|_1 = \|\lambda(\cdot \mid x) - \widehat{\lambda}_t(\cdot \mid x)\|_1 \le \eta_t(x). \tag{44}$$

Combining this inequality with the transition comparison in the layerwise simulation recursion and using planning visitation probabilities on the right gives

$$\|\bar{q}_t - \widehat{q}_t\|_1 \le \mathfrak{b}_t(\widehat{q}_t).$$

The online case has the same no disclosure response on both sides. Let $\widehat{B}_t$ denote the planning counterpart of $\bar{B}_t$ induced by $\widehat{q}_t$. Applying the same performance-difference recursion to the bounded stage residual gives

$$|\bar{B}_t - \widehat{B}_t| \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}_t). \tag{45}$$

For the cumulative bound, write

$$H_t(q) := c_\kappa \sum_x q(x)\eta_t(x) + 2\sum_{x,i,a} q(x,\tau_i,a)\epsilon_t(x,\tau_i,a), \qquad \mathfrak{b}_t(q) = NH_t(q). \tag{46}$$

Reversing the decomposition in the layerwise simulation recursion weights every local error by the true executed occupancy and gives

$$\|\bar{q}_t - \widehat{q}_t\|_1 \le NH_t(\bar{q}_t). \tag{47}$$

We first bound $\sum_t H_t(\bar{q}_t)$, then transfer the bound to the planning occupancy. Let $\mathcal{H}_{t-1}$ be the history before episode t. Let $Y_t(x)$ and $Y_t(x,\tau_i,a)$ indicate whether x and (x,τ_i,a), respectively, are visited in episode t. The layered process visits each such category at most once per episode, including under early termination. Their conditional expectations satisfy

$$\mathbb{E}[Y_t(x) \mid \mathcal{H}_{t-1}] = \bar{q}_t(x), \qquad \mathbb{E}[Y_t(x,\tau_i,a) \mid \mathcal{H}_{t-1}] = \bar{q}_t(x,\tau_i,a). \tag{48}$$

The following count bound applies to visits to states and to tuples (x,τ_i,a). Let $\mathcal{I}$ be a finite category set in one layer, write $Y_t(j) \in \{0,1\}$, $p_t(j) := \mathbb{E}[Y_t(j) \mid \mathcal{H}_{t-1}]$, and $\mathsf{N}_t(j) := \sum_{u<t} Y_u(j)$. For $k \ge 0$, define the predictable dyadic bin

$$\mathcal{T}_{j,k} := \{t : 2^k \le 1 \vee \mathsf{N}_t(j) < 2^{k+1}\}, \qquad P_{j,k} := \sum_{t\in\mathcal{T}_{j,k}} p_t(j), \qquad Z_{j,k} := \sum_{t\in\mathcal{T}_{j,k}} Y_t(j). \tag{49}$$

Because $\mathbb{E}[e^{-Y_t(j)} \mid \mathcal{H}_{t-1}] \le \exp(-(1-e^{-1})p_t(j))$, the corresponding stopped exponential process is a supermartingale. Markov's inequality and a union bound over j and the at most $1+\lceil \log_2 T\rceil$ nonempty bins imply, with probability at least $1-\alpha$,

$$P_{j,k} \le 2Z_{j,k} + 2\log\left(\frac{2|\mathcal{I}|(1+\lceil\log_2 T\rceil)}{\alpha}\right). \tag{50}$$

Moreover, $Z_{j,k} \le 2^{k+1}$. Dividing Eq. (50) by $2^{k/2}$ and summing the geometric series gives a universal constant $c > 0$ such that

$$\sum_{t=1}^{T} \frac{p_t(j)}{\sqrt{1\vee \mathsf{N}_t(j)}} \le c\sqrt{1+\mathsf{N}_{T+1}(j)} + c\log\left(\frac{2|\mathcal{I}|(1+\lceil\log_2 T\rceil)}{\alpha}\right). \tag{51}$$

Summing over categories and applying Cauchy–Schwarz, together with $\sum_j \mathsf{N}_{T+1}(j) \le T$, yields

$$\sum_{j\in\mathcal{I}}\sum_{t=1}^{T} \frac{p_t(j)}{\sqrt{1\vee \mathsf{N}_t(j)}} \le \widetilde{O}\Big(\sqrt{|\mathcal{I}|T\log(1/\alpha)} + |\mathcal{I}|\log(1/\alpha)\Big). \tag{52}$$

The left hand side is also at most T. In each application below, taking the smaller of this trivial bound and Eq. (52) absorbs the additive category term into the displayed final rate.

For the prior radius, with $\beta_\lambda := \log(16T|\mathcal{X}|m/\delta)$,

$$\eta_t(x) \le \frac{\sqrt{2m\beta_\lambda} + 2\sqrt{m}}{\sqrt{1\vee \mathsf{N}_t(x)}}, \tag{53}$$

where $m/(r+m) \le \sqrt{m/(1 \vee r)}$ for every count $r \ge 0$. Applying Eq. (52) with $\alpha = \delta/(4N)$ to $\mathcal{I} = \mathcal{X}_n$ in each layer, taking a union bound over layers, and using $\sum_n \sqrt{|\mathcal{X}_n|} \le \sqrt{N \sum_n |\mathcal{X}_n|} \le |\mathcal{X}|$ gives

$$\sum_{t,x} \bar{q}_t(x)\eta_t(x) \le \widetilde{O}\Big(|\mathcal{X}|\sqrt{Tm\log(1/\delta)}\Big). \tag{54}$$

For transitions from layer $n < N-1$, take $\mathcal{I}_n = \mathcal{X}_n \times \Omega \times \{0,1\}$, so $|\mathcal{I}_n| = 2m|\mathcal{X}_n|$, and $d_n := |\mathcal{X}_{n+1}| + 1$. Applying Eq. (52) with $\alpha = \delta/(4N)$ and the factor $\sqrt{d_n}$ in ϵ_t, followed by a union bound over layers, gives

$$\sum_{t,x,i,a} \bar{q}_t(x,\tau_i,a)\epsilon_t(x,\tau_i,a) \le \widetilde{O}\left(\sum_{n=0}^{N-2} \sqrt{mT|\mathcal{X}_n|d_n\log(1/\delta)}\right) \le \widetilde{O}\Big(|\mathcal{X}|\sqrt{Tm\log(1/\delta)}\Big). \tag{55}$$

The leading term uses $\sqrt{|\mathcal{X}_n|(|\mathcal{X}_{n+1}|+1)} \le (|\mathcal{X}_n| + |\mathcal{X}_{n+1}| + 1)/2$ before summing over layers. For the additive term in Eq. (52), write $a_n := |\mathcal{X}_n|$. The case split at $T = m(a_n + d_n)^2$ gives $\min\{T, ma_n\sqrt{d_n}\} \le \sqrt{mT}(a_n + d_n)$; summing over layers proves the second inequality up to logarithmic factors. The final-layer radius is zero because its transition is deterministic. Assigning failure probability $\delta/4$ to each confidence and count event, Eqs. (54) and (55) give

$$\sum_{t=1}^{T} H_t(\bar{q}_t) \le \widetilde{O}\Big(|\mathcal{X}|\sqrt{Tm\log(1/\delta)}\Big). \tag{56}$$

To transfer the confidence-radius weights from the executed occupancy to the planning occupancy, use $\eta_t, \epsilon_t \le 2$ and the fact that marginalization contracts the L_1 distance. These properties give

$$|H_t(\widehat{q}_t) - H_t(\bar{q}_t)| \le 2(c_\kappa + 2)\|\widehat{q}_t - \bar{q}_t\|_1. \tag{57}$$

Define $K_{\mathrm{tr}} := 1 + 2N(c_\kappa + 2) = 1 + 2N(3 + 2L_p\Delta_\kappa)$. Combining the preceding inequality with Eq. (47) yields

$$\mathfrak{b}_t(\widehat{q}_t) = NH_t(\widehat{q}_t) \le NK_{\mathrm{tr}}H_t(\bar{q}_t). \tag{58}$$

Summing over episodes gives the claimed cumulative bound, with the N^2 factor arising from NK_{tr}. The two confidence events and two martingale events have total failure probability at most δ.

A.8 Proof of Theorem 2

For the mechanism selected in episode t, let $\widehat{q}$ denote its planning occupancy. Let (A, J, C, B) and $(\widehat{A}, \widehat{J}, \widehat{C}, \widehat{B})$ denote its executed and planning functionals, respectively. On the event in Lemma 5, the occupancy bound and Eq. (45) give

$$\begin{cases} |C - \widehat{C}| \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}), \\ |A - \widehat{A}| \le \mathfrak{b}_t(\widehat{q}), \\ |J - \widehat{J}| \le \mathfrak{b}_t(\widehat{q}), \\ |B - \widehat{B}| \le \overline{\gamma}\mathfrak{b}_t(\widehat{q}). \end{cases} \tag{59}$$

The participation residual therefore changes by at most $(1 + q_{\min})\mathfrak{b}_t(\widehat{q})$. Theorem 1 provides a feasible comparator for $\mathbf{P3}$ with objective at most $C^* + O(h)$. Let $\widehat{q}_t$ and $\widehat{C}_t$ denote the selected mechanism's planning occupancy and payment. Optimality gives

$$\widehat{C}_t - \overline{\gamma}\mathfrak{b}_t(\widehat{q}_t) \le C^* + O(h). \tag{60}$$

The payment comparison then gives

$$\bar{C}_t - C^* \le 2\overline{\gamma}\mathfrak{b}_t(\widehat{q}_t) + O(h). \tag{61}$$

The relaxed constraints and the remaining functional comparisons give

$$q_{\min}\bar{A}_t - \bar{J}_t \le 2(1+q_{\min})\mathfrak{b}_t(\widehat{q}_t) + \big(K_{\mathrm{part}} + (1+q_{\min})K_{\mathfrak{b}}\big)h, \tag{62}$$

$$\bar{B}_t \le 2\overline{\gamma}\mathfrak{b}_t(\widehat{q}_t) + (K_{\mathrm{ben}} + \overline{\gamma}K_{\mathfrak{b}})h. \tag{63}$$

Summing over episodes and invoking Lemma 5 proves Eq. (14) with discretization constant $\max\{K_C + \overline{\gamma}K_{\mathfrak{b}}, K_{\mathrm{part}} + (1+q_{\min})K_{\mathfrak{b}}, K_{\mathrm{ben}} + \overline{\gamma}K_{\mathfrak{b}}\}$. These bounds apply to conditional expected payments and constraint residuals.

A.9 Additional Experiments

A.9.1 Experimental Setup

Dataset Setup. FEMNIST and Shakespeare use natural LEAF partitions by writer and speaking role, respectively. CIFAR-10 is partitioned among 20 clients using a classwise Dirichlet distribution with $\alpha = 0.3$. All methods share the initialization, partition, client sampling rule, environment, and maximum episode budget within each dataset and seed. The FEMNIST CNN uses two 5×5 convolutional layers with 32/64 channels and max pooling, a 256-unit hidden layer, and 62 outputs. CIFAR-10 uses ResNet-18 trained from scratch with a 3×3 stem of stride one, no initial max pooling, and random crop and horizontal flip augmentation. Shakespeare uses an 80-symbol vocabulary, a 64-dimensional embedding, and one LSTM layer with 128 hidden units. For FEMNIST, CIFAR-10, and Shakespeare, respectively, episode budgets are 3000/500/200, batch sizes are 32/64/32, SGD learning rates are $0.01/0.01/0.50$, and server mixing weights are $1/0.5/1$. Clients are sampled uniformly and independently of communication state. Each participation triggers one local epoch and a sequential server update. We evaluate the full test set after each episode and measure payment at the first crossing of target accuracies 80%/70%/50% over 5 seeds. DaringFed stops at this crossing; other trajectories are truncated retrospectively.

Simulation Setup. The default environment uses $N = 3$, $|\mathcal{X}| = 6$, $m = 3$, and condition costs $(0.2, 0.5, 0.8)$. Participation favors transitions to states with lower costs, with a nonterminal termination probability of 0.05. The dynamic study uses 10 seeds at $T = 1000$ under default D0, weak action coupling W, strong action coupling S, and high termination H. D0, W, and S have an early termination probability of 0.0975 and expected episode length 2.8525, compared with 0.5100 and 2.1900 in H. We use the logit response $p(\gamma, \bar{\kappa}) = [1 + \exp(-(\gamma - \bar{\kappa})/\tau_{\mathrm{temp}})]^{-1}$ (Faury et al., 2020). Federated training uses $\tau_{\mathrm{temp}} = 0.12$, rewards in $[0.5, 1]$, and $q_{\min} = 0.5$. Theory experiments use $\tau_{\mathrm{temp}} = 0.10$, rewards in $[0.1, 1]$, $q_{\min} = 0.55$, and 10 seeds at $T \in \{50, 100, 200, 500, 1000, 2000\}$. Both use reward and posterior grid spacing $h = T^{-1/2}$, with $\delta = 0.05$ for federated training. Pseudo regret uses a fixed feasible numerical oracle with reward and posterior grid spacing 0.02. The signaling study fixes priors and transitions within each seed and varies condition costs across Low $(0.4, 0.5, 0.6)$, Medium $(0.2, 0.5, 0.8)$, and High $(0.05, 0.5, 0.95)$ disparity. The corresponding $D_{\mathrm{info}} = \mathbb{E}_x \operatorname{Var}_{\lambda(\cdot|x)}[\kappa(\tau)]/\tau_{\mathrm{temp}}^2$ values are 0.357, 3.213, and 7.230. Final states terminate deterministically, and maximum rewards satisfy the participation requirement.

Baseline Setup. *DaringFed* uses a single uninformative signal and the lowest reward satisfying $q_{\min}$ at the largest resource cost. This conservative implementation guarantees participation for any condition prior and has zero Bayes residual without modeling transitions. *Dynamic Pricing* maps its increasing reward schedule to the common reward interval and compensates accumulated expected participation shortfall under a fixed prior. *L2I* uses a fixed exploration and exploitation reward schedule without signaling, client quality selection, or Stackelberg optimization. Neither accesses the true prior or transition matrix. *w/o Markov* shares one estimated transition kernel across both actions and sets transition confidence to zero, while retaining prior learning and its optimism. *w/o Optimism* retains signaling and Markov planning but sets both confidence radii to zero. *w/o Signaling* retains learning and optimistic Markov planning but allows only one uninformative signal.

Evaluation Protocol. Payment and constraint metrics use conditional expected executed occupancy. Payment per arrival, participation, and Bayes residual per arrival are $\sum_t \bar{C}_t / \sum_t \bar{A}_t$, $\sum_t \bar{J}_t / \sum_t \bar{A}_t$, and $\sum_t \bar{B}_t / \sum_t \bar{A}_t$, respectively. Signaling gains are measured as *w/o Signaling* minus MOVE, using payment per arrival and normalized violations V^{P}/A and V^{B}/A. We compare

Table 4: Ablation with w/o Signaling.

Information disparity	Payment gain	Reduction (%)	Participation difference	Bayes difference
Low	-0.00410 ± 0.00733	-2.10	0.00750 ± 0.00857	-0.000203 ± 0.000302
Medium	$(0.01493 \pm 0.00841)^*$	5.26	-0.00184 ± 0.00967	$(0 \pm 0)^\dagger$
High	$(0.01327 \pm 0.01110)^*$	4.82	0.00096 ± 0.00919	$(0 \pm 0)^\dagger$

Table 5: Ablation with w/o Markov.

Env.	D_{act}	Participation improvement
W	0.101	$(0.88238 \pm 0.01880)^*$
D0	0.503	$(1.01389 \pm 0.01278)^*$
S	0.763	$(1.01793 \pm 0.03241)^*$
S−W contrast	—	$(0.13555 \pm 0.02618)^*$

Table 6: MOVE robustness across dynamic environments.

Env.	R_T/T	V_T^{P}/T	V_T^{B}/T
D0	0.0171 ± 0.0223	0.4379 ± 0.0134	0.0003 ± 0.0006
W	-0.1141 ± 0.0182	0.5485 ± 0.0165	0.0008 ± 0.0007
S	-0.0324 ± 0.0331	0.4508 ± 0.0303	0.0005 ± 0.0006
H	-0.1117 ± 0.0409	0.3364 ± 0.0277	0.0044 ± 0.0022

methods using the same random seeds and report average differences, 95% confidence intervals, and effect sizes. We account for multiple comparisons using Holm correction. Constraint violations are considered equivalent only when the adjusted equivalence tests pass and the full confidence intervals lie within ± 0.02 for participation and ± 0.005 for Bayes benefit. In Figure 3, mean curves cover the payment range shared by all 5 seeds, while target markers are computed separately.

Limitations. MOVE assumes a known participation response function and stationary condition distributions and transition kernels over a finite state and condition space. Extending the framework to learn client responses and handle nonstationary dynamics and continuous spaces remains an important direction for practical deployment.

A.9.2 Ablation Studies

Table 4 shows payment reductions of 5.26% at Medium and 4.82% at High relative to the mean payment per arrival of *w/o Signaling*, with equivalent normalized constraint violations under the margins defined above. The Low effect remains unresolved. This pattern is consistent with disclosure making rewards more effective when hidden costs are sufficiently separated. The High gain exceeds the Low gain by 0.01736 with 95% CI $[0.00048, 0.03425]$ and Holm corrected $p = 0.02420$, but Medium exceeds High numerically, so the gain is not monotonic across all levels.

Table 5 reports the participation violation per episode of *w/o Markov* minus MOVE. Improvements are significant in all three environments and are greater in S than W. Stronger action coupling makes future state occupancy more sensitive to current participation, increasing the value of accounting for these effects in planning. The results establish smaller participation shortfalls under the same $q_{\min}$, not lower payment at matched participation. Complete statistics remain in the supplementary records.

A.9.3 Robustness Analysis

Table 6 shows smaller participation shortfalls in D0 and S than in W, consistent with stronger action coupling allowing participation to improve subsequent conditions more effectively. Bayes benefit violations remain small in these three environments. H has a larger Bayes benefit violation but a

Table 7: MOVE sensitivity to $q_{\min}$.

$q_{\min}\uparrow$	Payment ↑	Participation ↑	V_T^{P}/T —	$V_T^{\mathrm{B}}/T\downarrow$
0.15	0.124 ± 0.015	0.343 ± 0.025	0.0000	0.0115
0.75	0.586 ± 0.013	0.842 ± 0.008	0.0000	0.0000
0.90	0.764 ± 0.021	0.936 ± 0.005	0.0000	0.0000

smaller participation shortfall per episode, partly reflecting fewer arrival opportunities under earlier termination. Participation shortfalls remain at the tested horizon.

A.9.4 Sensitivity Analysis

Calibration uses 5 seeds and a predefined grid on $[0.10, 0.90]$ with spacing 0.05. A fixed rule selects three participation targets spanning different constraint regimes, evaluated on 10 additional seeds without retuning. Table 7 shows increasing payment per arrival and participation as $q_{\min}$ rises, with zero reported participation violation in all three settings. These results illustrate the payment cost of meeting stricter participation requirements. Bayes benefit violations are reported as zero at the two higher targets. Payment and participation are reported as means $\pm$ 95% confidence interval half widths. All episode planning problems were solved successfully.