Scheduling Training-Inference Co-Location in Demand Response for Sustainable Edge AI

Konglin Zhu, Siyuan Wei, Xuan'er Wu, Lei Jiao, Jin Dong, Lin Zhang

Abstract—In the pursuit of data privacy and reduced latency, the adoption of edge intelligence has surged. Meanwhile, the enormous increase in AI has resulted in significant energy consumption. Edge intelligence plays a crucial role in Energy Demand Response (EDR). However, existing edge intelligence falls short of meeting the demands of co-locating training and inference tasks while satisfying EDR. Specifically, the intertwinement between balancing energy consumption, system delay and model accuracy, and uncertain future inputs adds to the challenge of designing an online sustainable system for co-located training and inference tasks. To address these challenges, we propose a novel two-timescale system for co-locating training and inference EDR. Our approach satisfies EDR by strategically planning training schedules on macro-timescales and migrating inference requests between heterogeneous edges on micro-timescales while minimizing long-term cost. We introduce a novel online polynomial time algorithm that first breaks down the problem into two subproblems, which are subsequently solved using an online-learning-based fractional algorithm and a randomized roun ding algorithm, respectively. Rigorous analysis demonstrates that our approach achieves both sublinear dynamic regret and sublinear dynamic fit. Extensive trace-driven evaluations validate the practical superiority of our approach over multiple existing methods, highlighting its effectiveness in real-world scenarios.

Index Terms—Demand response, federated learning, edge computing, two-timescale, online learning, co-location

1 Introduction

An *Edge Artificial Intelligence (AI)* system simultaneously handles training and inference tasks on geo-distributed infrastructures such as cellular base stations and WiFi access points equipped with micro datacenters [1], [2]. These systems benefit from proximity to data sources and end users, offering low latency and privacy protection [3], [4]. However, their growing scale has led to increasing energy consumption [5], [6], making them promising candidates for participating in Emergency Demand Response (EDR) programs [7]–[9]. Under EDR, edge systems are incentivized to reduce energy use below time-varying caps set by the grid.

Efficiently orchestrating both training and inference tasks under EDR constraints presents key challenges. First, tasks arrive online and must be scheduled immediately. While inference workloads can be flexibly migrated to energy-efficient servers [10], [11], training workloads are tightly coupled to large volumes of data, limiting mobility.

- *This work was supported in part by the National Key Research and Development Program of China (2023YFB2704500), the Beijing Natural Science Foundation (4222033), the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, and the U.S. National Science Foundation (CNS-2047719 and CNS-2225949). (Corresponding author: Lei Jiao.)*

- *Konglin Zhu is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China, and with the Purple Mountain Laboratories, Nanjing, China (e-mail: klzhu@bupt.edu.cn).*
- *Siyuan Wei and Lin Zhang are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China (e-mail: {weisy, zhanglin}@bupt.edu.cn).*
- *Xuan'er Wu is with the Institute of Artificial Intelligence (TeleAI), China Telecom (e-mail: wuxn5@chinatelecom.cn).*
- *Lei Jiao is with the Center for Cyber Security and Privacy, University of Oregon, Eugene, OR 97403, USA (e-mail: ljiao2@uoregon.edu).*
- *Jin Dong is with the Beijing Academy of Blockchain and Edge Computing, Beijing 100093, China (e-mail: dongjin@baec.org.cn).*

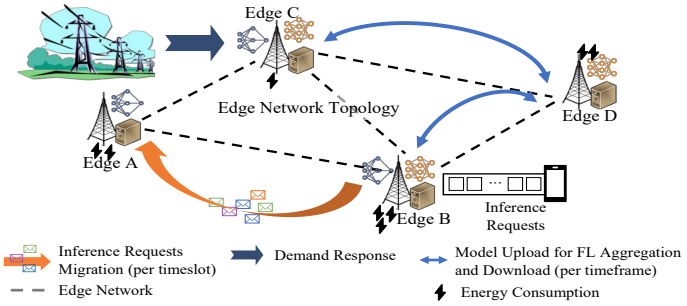


Fig. 1: System Scenario

Balancing the two while ensuring convergence and accuracy under varying energy caps is non-trivial.

Second, training and inference operate on different time scales. Inference is short and delay-sensitive, while training is longer and more delay-tolerant [12], [13]. This leads to asynchronous decision-making, where inference scheduling must respect prior training decisions that remain fixed during longer macro intervals. The challenge is compounded by the unpredictability of future task arrivals.

Prior work on EDR in edge computing [14]–[17] does not address the specific nature of AI tasks, while existing AI task management studies [18]–[22] overlook EDR considerations. Recent research on training-inference co-location [13], [23]–[25] focuses on resource sharing, but lacks attention to energy sustainability and the multi-timescale nature of scheduling. Section 5 provides a detailed comparison.

In this work, targeting *Federated Learning (FL)* as the training tasks, we model a two-timescale total cost optimization problem for the training-inference co-located edge AI systems. Our formulation jointly controls FL training iterations across *time frames*, which we call the *macro-timescale*,

and inference migration across *time slots* within each time frame, which we call the *micro-timescale*, minimizing energy cost and inference delay while respecting FL model convergence, EDR energy caps, and edge resource capacities. Our formulation allows arbitrary temporal and spacial fluctuations and heterogeneities in the inputs, including energy caps, the energy prices, the network delay, and the training and inference task arrivals. Our problem turns out to be a mixed-integer program, which is unsurprisingly NP-hard.

We then design a novel polynomial-time online algorithmic framework for this problem. Central to our framework is an online learning approach that incorporates the long-term constraint into the objective via Lagrange multipliers [26], splits it into each individual time frame with regularization, and utilizes alternate ascent-descent iterations to make the control decisions on the fly without needing foreknowledge of future inputs. Our approach decomposes our original problem into two sub-problems: macro-timescale training scheduling and micro-timescale inference scheduling. While employing online learning as mentioned for the macro timescale with a randomized rounding algorithm to convert the obtained fractional solutions into integers, we propose to allocate the expected energy evenly on the micro timescale, so that we just divide the micro-timescale problem into a series of single-time-slot problems and solve them sequentially by standard convex optimization techniques.

We characterize the performance of our proposed algorithmic framework by rigorously proving that our approach can achieve sub-linear dynamic regret and sub-linear dynamic fit compared to the series of single-time-frame offline optimums, as [27], [28]. That said, compared to the optimal objectives in the setting which splits the original problem into individual time frames assuming full knowledge of the inputs in each time frame, the time-averaged difference between the objectives incurred online by our approach and those optimums vanishes as time goes; and the time-averaged violation of the long-term constraint due to absorbing it into the objective also vanishes as time goes.

We conduct a thorough evaluation of our approach using a dynamic range of FL tasks, based on real-world training data [29]–[31] that arrive over a 168-hours period [15], [18], and inference requests [32] occurring every 5 minutes. Our assessment incorporates diverse real-world data, including heterogeneous edge servers [32], [33], EDR events [34], electricity prices [35], etc., ensuring a robust validation of our system's practical performance. We observe multiple results: (i) In meeting EDR requirements, our approach outperforms alternatives like the greedy method, the state-of-the-art `OASDR` algorithm [36], Lyapunov algorithm [37] and reinforcement-learning (RL)-based algorithm, reducing the total cost by approximately 58.8%, 13.4%, 20.8% and 3.6% respectively; (ii) Our approach demonstrates sub-linear growth in the dynamic regret and dynamic fit; (iii) Our approach executes highly efficiently, completing our major control logic within milliseconds.

2 Models and Problem Formulation

2.1 System Models

We summarize the major notations in Table 1.

TABLE 1: Notations

Inputs	Meaning
$\mathcal{T}$	Set of time frames
$\mathcal{D}$	Set of time slots of each time frame
$\mathcal{J}$	Set of edges
$\mathcal{I}$	Set of FL tasks
$\mathcal{N}$	Set of inference tasks
R^t	Energy cap at time frame t
r^t	Electricity price of the grid at time frame t
t_i^{in}	Time frame when the FL task i arrives at the system.
t_i^{out}	Time frame when the FL task i quit the system.
ϵ_i	The convergence accuracy of the global model that is to be trained of the FL task i
E_i	The number of local iterations to be executed in each global iteration for FL task i
α_{ij}	Whether the edge j is chosen for aggregation for FL task i
$\mathcal{A}_i$	The set of edges taking part in the local training of FL task i
$\mathcal{D}_{ij}$	The set of the training data located at edge j for the FL task i
H_n	The size of the inference request for inference task n
Q_i	The total number of global iterations required for FL task i
A_i	The amount of FLOPS consumed for each local iteration of FL task i
B_i	The amount of FLOPS consumed for each global aggregation of FL task i
β_{jn}	Whether edge j has the capability to handle inference task n
$SLO_{n\tau}^t$	The SLO for inference task n at the time slot τ of inference task t.
R_{jk}^t	The data transmission rate from edge j to edge k in the time frame t.
φ_i	The amount of FLOPS per aggregation for FL task i
$\varphi_i^{'}$	The amount of FLOPS consumed to train one unit of data samples during per local iteration for FL task i
γ_j	Energy consumption per unit computation on edge j
E_{ij}^t	Energy consumption per global iteration for FL task i on edge j in the time frame t
$E_{ij}^{'}$	Energy consumption per aggregation for FL task i on edge j
$E_{ij}^{t''}$	Energy consumption for transmitting the model between an edge j performing local training for the FL task i and the designated edge for aggregation in the time frame t.
$E_{ij}^{t'''}$	Energy consumption for local training during per global iteration for FL task i on edge j in the time frame t
M_i	The size of model for FL task i
$w_{jn\tau}^t$	The amount of inference requests of inference task n arriving at the edge j at the time slot τ in the time frame t
d_{jn}	Energy consumption per inference request of inference task n on edge j cost
S_n	The amount of FLOPS consumed for each inference request of inference task n
C_j	The total computational resource (FLOPS) available at edge node j for each time slot
$q_{jk\tau}^t$	Propagation delay from edge j to edge k in the time slot τ of the time frame t
$\omega_{jkn\tau}^t$	Transmission delay for an inference request of inference task n migrated from the edge j to the edge k in the time slot τ of the time frame t.
ν_{ijk}^t	Transmission delay for transferring models between edge j and edge k in FL task i in the time frame t.
δ_{ij}	Whether the edge j is a client the FL task i
b_{jk}^t	Computation delay for inference task n in the time frame t at the edge k
f_k	Computation speed at the edge k

Outputs	Meaning
x_i^t	The number of global iterations that FL task i conducts in time frame t
$y_{jkn\tau}^t$	Amount of inference requests of inference task n migrated from the edge j to the edge k in the time slot τ of the time frame t
z_j^t	The amount of FLOPS consumed for FL tasks at edge j at time frame t
$z_{j\tau}^t$	The amount of FLOPS consumed for FL tasks at edge j at time slot τ of time frame t

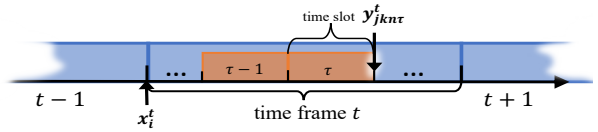


Fig. 2: Two-timescale Decision Making

Edge Infrastructures: We consider an edge system consisting of a set $\mathcal{J} \triangleq \{1, ..., J\}$ of distributed edges, where an "edge" refers to a cellular base station or a WiFi access point equipped with a server cluster or a micro datacenter. These edges are interconnected to one another via backhaul networks, and are accessed by the end users via wireless networks. All edges operate over a shared wireless spectrum with partial frequency reuse. Specifically, a fraction ρ of edges are allowed to reuse the same frequency band in each time slot, while others operate on orthogonal channels. As a result, wireless transmissions between edges may experience both noise and controlled co-channel interference.

Dual Timescales: As shown in Fig. 2, we divide the entire time horizon into T time frames, indexed by $t \in \mathcal{T} \triangleq \{1, 2, ..., T\}$, which we refer to as the macro-timescales. We further divide each time frame into D time slots, indexed by $\tau \in \mathcal{D} \triangleq \{1, 2, ..., D\}$, which we refer to as the micro-timescales. The duration of each single time frame and that of each single time slot are pre-specified, reflecting the desired appropriate time granularity to update the control decisions for the training tasks and the inference tasks, respectively.

Demand Response: The edge system is powered by the power grid and is enrolled in the Emergency Demand Response (EDR) program. Upon receiving an EDR signal, the edge system is notified of an energy cap for each time frame, denoted as R^t, $\forall t \in \mathcal{T}$. The edge system must reduce its energy consumption to stay under the cap for each corresponding time frame. EDR periods may occur intermittently throughout the time horizon $\mathcal{T}$, and we deem $R^t = +\infty$ when no EDR occurs. The energy cost at the time frame t is $r^t e^t$, where r^t is the electricity price of the grid and e^t is the amount of electricity consumed by the edge system at the time frame t.

FL Training Tasks: We consider a set of Federated Learning (FL) tasks $\mathcal{I} = \{1, ..., I\}$. Each FL task $i \in \mathcal{I}$ is characterized by $\Phi_i = \{t_i^{in}, t_i^{out}, \epsilon_i, E_i, \mathcal{A}_i, \alpha_{ij}, \{\mathcal{D}_{ij}, j \in \mathcal{A}_i\}\}$. Here, t_i^{in} denotes the time frame when the FL task i enters the system; t_i^{out} is the deadline required for completing FL task i, and the FL task i quit the system when reaching the deadline no matter whether it is completed; ϵ_i represents the convergence accuracy of the global model that is to be trained for the FL task i; E_i indicates the number of "local iterations" required for each "global iteration" during the training process, which will be further explained later; $\mathcal{A}_i$ is the set of edges that participate in the "local training" of FL task i (all the edges in $\mathcal{A}_i$ are the clients for FL task i and one of them is selected as a server), and we use $\delta_{ij} \in \{0, 1\}$ to denote whether edge j serves as a client for FL task i; $\alpha_{ij} \in \{0, 1\}$ denotes whether edge j is selected for global aggregation of FL task i; $\mathcal{D}_{ij}$ refers to the set of training data located at edge j ($j \in \mathcal{A}_i$) for the FL task i, where each data sample $m \in \mathcal{D}_{ij}$ is associated with a loss function $f_m(\cdot)$.

Each FL task i aims to train a model denoted as $\boldsymbol{w}_i$. The local loss for FL task i at edge j ($j \in \mathcal{A}_i$) is $F_{ij}(\boldsymbol{w}_i) = \frac{1}{|\mathcal{D}_{ij}|} \sum_{m \in \mathcal{D}_{ij}} f_m(\boldsymbol{w}_i)$. The global loss is defined as $F_i(\boldsymbol{w}_i) = \frac{1}{\sum_j |\mathcal{D}_{ij}|} \sum_j (|\mathcal{D}_{ij}| \cdot F_{ij}(\boldsymbol{w}_i))$. The objective of FL task i is to find the optimal model parameters $\boldsymbol{w}_i^*$ that minimize the global loss $F_i(\boldsymbol{w}_i)$, ensuring the convergence criterion $F_i(\boldsymbol{w}_i) - F_i(\boldsymbol{w}_i^*) \leq \epsilon_i$ is satisfied.

The training process for each FL task i is as follows. After the FL task i enters the system, as time goes on, at the beginning of each time frame t $(t \geq t_i^{in})$, we decide whether the FL task i is trained and the number of global iterations x_i^t to perform in the time frame t. Then in each global iteration χ_i^t, on each client $j \in \mathcal{A}_i$, the model is locally trained, started with initial value $\boldsymbol{w}_{i,j,\chi_i^t,0}^t = \boldsymbol{w}_{i,\chi_i^t-1}^t$ and ending with value $\boldsymbol{w}_{i,j,\chi_i^t}^t = \boldsymbol{w}_{i,j,\chi_i^t,E_i}^t$, by employing the mini-batch Stochastic Gradient Descent (SGD) approach as $\boldsymbol{w}_{i,j,\chi_i^t,\kappa_i}^t = \boldsymbol{w}_{i,j,\chi_i^t,\kappa_i-1}^t - \frac{\eta^t}{|\xi_{i,j,\chi_i^t,\kappa_i}^t|} \sum_{m \in \xi_{i,j,\chi_i^t,\kappa_i}^t} \nabla f_m(\boldsymbol{w}_{i,j,\chi_i^t,\kappa_i-1}^t)$ through E_i iterations. We highlight that to achieve $F_i(\boldsymbol{w}_i) - F_i(\boldsymbol{w}_i^*) \leq \epsilon_i$, the number of global iterations $\sum_{t \in \mathcal{T}} x_i^t$, must satisfy $\sum_{t \in \mathcal{T}} x_i^t \geq Q_i$, with $Q_i \geq \frac{2Lv_i}{\mu^2 \epsilon_i} - \gamma$, $v_i = \sum_{j \in \mathcal{J}} \frac{\sigma_j^2}{J^2} + 8(E_i - 1)^2 G^2 + \frac{4(J - |\mathcal{A}_i|)}{|\mathcal{A}_i|(J-1)} E_i^2 G^2 + 8LG^2$ [38], given E_i, where σ_j^2 and G are upper bounds of the variance and the squared norm of the stochastic gradient in each device, the loss function $F_{ij}(\cdot)$ is L-smooth and μ-strongly convex, $\gamma = \max\{\frac{L}{\mu}, E_i\}$.

In the edge system, there are multiple FL tasks, and we assume that each edge node can serve as either an FL client or an FL server for each FL task. For example, edge node 0 can be a client for local training in FL task 1 and FL task 2 simultaneously. In addition, each node can simultaneously take the roles of both client and server for an FL task, and each FL task only has one server for global aggregation.

Inference Tasks: We consider a set of inference tasks, represented as $\mathcal{N} = \{1, ..., N\}$. For each inference task n, the inference requests are dynamically arriving at the edge node in an online manner, and we use $w_{jn\tau}^t$ to denote the amount of requests of inference task n arriving in the time slot τ of the time frame t. The inference requests are delay-sensitive, and are required to meet specific service-level objectives (SLOs) in terms of latency. More precisely, each inference task n has a latency requirement $SLO_{n\tau}^t$ in time slot τ of the time frame t. The SLOs for inference tasks may vary, but they all require completion within a single time slot. If the any inference request cannot meet the latency requirement, the quality of service (QoS) for that task is considered violated. Each inference request arriving at edge node j can be migrated to another edge node if needed, to ensure timely completion and to reduce energy consumption.

In this paper, we assume that FL tasks and inference tasks are independent. That is, the models generated in the FL process are not immediately available for online inference services. This is due to the fact that the models intended for inference require post-training fine-tuning and compression to ensure they deliver consistent and high-performance results. Here, we use $\beta_{jn} \in \{0, 1\}$ to denote whether the edge j is able to serve the inference requests of task n.

Example Algorithm of FL for Task i

▷ **Input:** $\mathcal{A}_i$, t_i^{in}, E_i, $\{\mathcal{D}_{ij}\}$;
▷ **Initialization on server**
Initialize the model $\boldsymbol{w}_i$, $\boldsymbol{w}_i^0 = \boldsymbol{w}_i$;
for $t \in \{t_i^{in}, t_i^{in}+1, ..., t_i^{out}\}$ **do**
  if $x_i^t = 0$, $\boldsymbol{w}_i^t = \boldsymbol{w}_i^{t-1}$
  else
  $\boldsymbol{w}_{i,0}^t = \boldsymbol{w}_i^{t-1}$
  for $\chi_i^t \in \{1, 2, ..., x_i^t\}$ **do**
    ▷ **Local training on each edge** $j \in \mathcal{A}_i$**:**
    download $\boldsymbol{w}_{i,\chi_i^t-1}^t$ from server edge;
    $\boldsymbol{w}_{i,j,\chi_i^t,0}^t = \boldsymbol{w}_{i,\chi_i^t-1}^t$;
    for $\kappa_i = 1, ..., E_i$ **do**
    $\boldsymbol{w}_{i,j,\chi_i^t,\kappa_i}^t = \boldsymbol{w}_{i,j,\chi_i^t,\kappa_i-1}^t - \frac{\eta^t}{|\xi_{i,j,\chi_i^t,\kappa_i}^t|} \sum_{m \in \xi_{i,j,\chi_i^t,\kappa_i}^t} \nabla f_m(\boldsymbol{w}_{i,j,\chi_i^t,\kappa_i-1}^t)$;
    $\boldsymbol{w}_{i,j,\chi_i^t}^t = \boldsymbol{w}_{i,j,\chi_i^t,E_i}^t$
    upload $\boldsymbol{w}_{i,j,\chi_i^t}^t$ to the server edge;
    ▷ **Global aggregation on server edge:**
    $\boldsymbol{w}_{i,\chi_i^t}^t = \frac{1}{|\mathcal{A}_i|} \sum_{j \in \mathcal{A}_i} \boldsymbol{w}_{i,j,\chi_i^t}^t$;

Wireless Channels Model: The data transmission rate from the edge j to the edge k in the time frame t is modeled as $R_{jk}^t = B \log_2 \left(1 + \frac{p_j^t h_{jk}^t}{N_0 B + I_k^t}\right)$, where B is the allocated bandwidth; p_j^t is the transmit power of edge j; N_0 is the noise power spectral density; h_{jk}^t denotes the effective channel gain from edge j to edge k; and I_k^t represents the aggregate co-channel interference observed at edge k in time frame t. Under the partial frequency reuse scheme, the interference term I_k^t is given by $I_{jk}^t = \sum_{\ell \in \mathcal{J}_t^{\text{co}}(k) \setminus \{j,k\}} p_\ell^t h_{\ell k}^t$, where $\mathcal{J}_t^{\text{co}}(k)$ denotes the set of edges that reuse the same frequency band as edge k in time frame t, and $h_{\ell k}^t$ denotes the effective channel gain from the interfering edge ℓ to the edge k.

FL Training Energy: First, for the global aggregations on the selected edge j for FL task i, the energy consumed per aggregation is $E_{ij}^{'} = \varphi_i \gamma_j \alpha_{ij}$, where φ_i denotes the amount of floating point operations (FLOPS) consumed per aggregation for FL task i, and γ_j is the energy consumption per FLOP on the edge j (This can be obtained either from normalized TDP-based estimations (e.g., TDP divided by peak FLOPS), or through empirical profiling.). Second, the energy consumption for transmitting the model between an edge j performing local training and the designated edge k for aggregation is $E_{ij}^{t''} = \frac{M_i p_j^t}{R_{jk}^t} + \frac{M_i p_k^t}{R_{kj}^t}$ [39], where M_i is the size of the model for FL task i; p_j^t is the transmission power of edge j; B is the allocated bandwidth; h_j^t is the wireless channel gain; and N_0 stands for the background noise of the wireless channel. The first term, $\frac{M_i p_j^t}{R_{jk}^t}$, represents the energy required for uploading the model from edge j to edge k, while the second term, $\frac{M_i p_j^t}{R_{kj}^t}$, corresponds to the energy consumed for downloading the aggregated model from edge k back to edge j. Third, for the local training at edge j, we use $E_{ij}^{t'''} = \varphi_i^{'} E_i \gamma_j \left|\xi_{i,j}^t\right|$ to refer to the energy consumption for the local training during each global iteration, where $\varphi_i^{'}$ is the amount of FLOPS consumed to train one unit of data samples during per local iteration for FL task i, $\left|\xi_{i,j}^t\right|$ is the size of the mini-batch data samples for each local iteration. All things considered, the energy consumption at the edge j per global iteration in the time frame t is $E_{ij}^t = E_{ij}^{'} + E_{ij}^{t''} + E_{ij}^{t'''}$.

FL Communication Cost: We consider the communication cost incurred by transferring models among edges during federated learning. For each FL task i in a global iteration, the per-round communication cost between an edge j acting as a client for local training and an edge k acting as a server for model aggregation is $\nu_{ijk}^t + q_{jk}^t + \nu_{ikj}^t + q_{kj}^t$, where ν_{ijk}^t denotes the transmission delay for transferring the model of FL task i from edge j to edge k in time frame t, and q_{jk}^t denotes the corresponding propagation delay. The transmission delay is given by $\nu_{ijk}^t = \frac{M_i}{R_{jk}^t}$, where M_i is the model size of FL task i and R_{jk}^t is the achievable transmission rate between edges j and k. The propagation delay may vary across time slots within a time frame t; therefore, we approximate it by the average propagation delay $q_{jk}^t = \overline{q_{jk\tau}^t}$, where $q_{jk\tau}^t$ denotes the propagation delay from edge j to edge k in the time slot τ of the time frame t. The factor of 2 accounts for the uplink and downlink model transmissions in each federated learning round.

Inference Energy: For the inference task n, we denote the energy consumption per inference request cost $d_{jn} = \gamma_j S_n$, where S_n is the amount of FLOPS required for each inference request of inference task n.

Inference Delay: The inference delay consists of the time for migrating inference requests among edges, the computation time, and the time for returning the inference result.

For each inference request of task n, if it is migrated from edge j to edge k, the inference delay is composed of the following components.

1) *Request migration delay*, including the propagation delay and the transmission delay from edge j to edge k, given by $q_{jk\tau}^t + \omega_{jkn\tau}^t$, where $\omega_{jkn\tau}^t = \frac{H_n}{R_{jk}^t}$ is the transmission delay, H_n is the size of the inference request dominated by input data (e.g., images, text, audio).
2) *Computation delay* at edge k, given by $b_{nk}^t = \frac{S_n}{f_k}$, where f_k denotes the computation speed of edge k.
3) *Result return delay*. Similar to some other works [40], since the inference output is typically small, we ignore the transmission delay for sending the output back to edge j. However, the propagation delay $q_{jk\tau}^t$ is still accounted for.

Capacity for FL tasks and Inference tasks: We assume that both FL tasks and inference tasks compete for a shared pool of heterogeneous computational resources, including CPUs and GPUs, available on each edge node. Let C_j denote the total computational resource (FLOPS) available at edge node j for each time slot. This parameter is derived from the theoretical FLOPS capability of the underlying hardware (e.g., GPUs or CPUs) as stated in the manufacturer's specifications. For FL task i, each local iteration consumes A_i FLOPS of computation, and each global aggregation consumes B_i FLOPS. Inference tasks are also scheduled under the same resource constraint, and the combined computational load of all tasks must not exceed the available capacity C_j at any time slot.

Control Decisions: The system operator makes the following decisions in a two-timescale manner: (1) $x_i^t \in \mathbb{N}$,

representing the number of global iterations for FL task i to be performed in the time frame t. This is macro-timescale FL task scheduling. (2) $y^t_{jkn\tau} \geq 0$, denoting the number of inference requests of inference task n migrated from the edge j to the edge k in the time slot τ of the time frame t. This is micro-timescale inference request scheduling.

Total Cost: (1) Energy cost: The energy cost for conducting the FL tasks and inference tasks is $r^t \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E^t_{ij} x^t_i + r^t \sum_{t\in\mathcal{T}} \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} d_{kn} y^t_{jkn\tau}$. (2) FL communication cost: $\sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} \sum_{k\in\mathcal{A}_i} \left(\nu^t_{ijk} + q^t_{jk} + \nu^t_{ikj} + q^t_{kj}\right) x^t_i \alpha_{ik}$, (3) Inference delay: $\sum_{t\in\mathcal{T}} \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} \left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{jk} + \omega^t_{jkn\tau}\right) y^t_{jkn\tau}$

2.2 Problem Formulation, Challenges, and Goal

Problem Formulation: The total cost is the sum of the energy cost and the delay cost of the edge system. We formulate the total cost minimization problem $\mathbb{P}$ as follows.

$$
\begin{aligned}
\mathbb{P}:\ \min\quad & r^t \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E^t_{ij}\, x^t_i \\
& + r^t \sum_{t\in\mathcal{T}} \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} d_{kn}\, y^t_{jkn\tau} \\
& + \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} \sum_{k\in\mathcal{A}_i} (\nu^t_{ijk} + q^t_{jk} + \nu^t_{ikj} + q^t_{kj}) x^t_i \alpha_{ik} \\
& + \sum_{t\in\mathcal{T}} \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} (q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}) y^t_{jkn\tau} \\
\text{s.t.}\quad & \sum_{t\in\mathcal{T}:\, t^{in}_i \leq t \leq t^{out}_i} x^t_i \geq Q_i, \qquad \forall i, && \text{(1a)} \\
& \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} d_{kn}\, y^t_{jkn\tau} \\
& \quad + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{A}_i} E^t_{ij}\, x^t_i \leq R^t, \qquad \forall t, && \text{(1b)} \\
& S_n\, y^t_{jkn\tau} \leq \beta_{kn} C_k, \qquad \forall j,k,n,\tau,t, && \text{(1c)} \\
& \sum_{k\in\mathcal{J}} y^t_{jkn\tau} \geq w^t_{jn\tau}, \qquad \forall j,n,\tau,t, && \text{(1d)} \\
& y^t_{jkn\tau} \left[q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau} - SLO^t_{n\tau} \right]^+ = 0, \\
& \quad \forall j,k,n,\tau,t, && \text{(1e)} \\
& z^t_j = \sum_{i\in\mathcal{I}_t} \left(\delta_{ij} E_i x^t_i A_i + \alpha_{ij} x^t_i B_i\right), \qquad \forall j,t, && \text{(1f)} \\
& z^t_j = \sum_{\tau\in\mathcal{D}} z^t_{j\tau}, \qquad \forall j,t, && \text{(1g)} \\
& z^t_{k\tau} + \sum_{n\in\mathcal{N}} \sum_{j\in\mathcal{J}} S_n\, y^t_{jkn\tau} \leq C_k, \qquad \forall k,\tau,t, && \text{(1h)} \\
\text{var.}\quad & x^t_i \in \mathcal{N}, \quad y^t_{jkn\tau} \geq 0, \quad z^t_j \geq 0, \quad z^t_{j\tau} \geq 0.
\end{aligned}
$$

The objective (1) minimizes the total cost. Constraint (1a) ensures that sufficient global iterations for FL task i are conducted to preserve the specified convergence. Constraint (1b) reflects the energy cap at each time frame to train the FL tasks and process inference requests. Constraint (1c) ensures only the edges that has the capability to handle inference task n can process the corresponding inference requests. Constraint (1d) guarantees that all inference requests for inference task n arriving at each edge in each time slot are migrated or served locally. Constraint (1e) means that the inference delay of each inference task cannot exceed its SLO. Constraint (1f) defines the total computational demand z^t_j for FL tasks at edge j at time frame t, which includes both the local iteration and global aggregation. Constraint (1g) shows the decomposition of z^t_j into different time slots. Constraint (1h) enforces computation demand for FL tasks and inferences must not exceed the node's resources available C_k at each time slot τ.

Challenges: It is non-trivial to solve the optimization problem due to the following challenges. First, the long-term constraints (1a) and (1b) couple future control decisions and system information. Blind to the unpredictable future EDR and electricity price, it is difficult to efficiently solve the problem online, without relying on prior knowledge of future system information. Second, the problem is NP-hard even in the offline setting with full prior knowledge of system information, which can be reduced to a minimum-cost knapsack problem by only retaining constraints (1a) and (1b) and removing terms with $y^t_{jkn\tau}$. It is intractable to solve this problem in an offline setting, not to mention we want to solve it online in an online setting. Third, the training tasks and the inference requests are scheduled in different timescales. The macro-timescale decisions for the global iterations of FL tasks and the micro-timescale decisions for migrating the inference requests are coupled in the long-term constraints, which adds to the complexity of the problem.

Algorithmic Goal: The goal is to design a polynomial-time algorithm to solve the cost-minimization problem in an online manner. Furthermore, our algorithm should achieve sub-linear dynamic regret and sub-linear dynamic fitness when compared to an offline algorithm, which is discussed in section 3.

3 Online Algorithm and Analysis

For solving our formulated problem, we design an online algorithm to schedule the FL tasks and inference requests in an online two-timescale manner, while addressing the aforementioned challenges. Then we analyze the dynamic regret and the dynamic fit of the algorithm and finally we rigorously prove their sublinearity.

3.1 Overview

To solve the cost minimization problem in an online manner, we proposed our online scheduling scheme in this section. The technical roadmap is to first decompose the original problem $\mathbb{P}$ into a macro-timescale FL task scheduling problem $\mathbb{P}_1$ and a series of micro-timescale inference requests scheduling problems $\{\mathbb{P}^t_2, \forall t\}$. $\mathbb{P}_1$ and $\{\mathbb{P}^t_2, \forall t\}$ are solved in a macro-timescale manner and in a micro-timescale manner respectively.

We propose Algorithm 0 as our overall online control algorithm. Algorithm 1 and 2 are the Macro-Timescale Fraction Algorithm and the Randomized Rounding Algorithm for $\mathbb{P}_1$, respectively. Algorithm 3 is the Micro-Timescale Fraction Algorithm for $\{\mathbb{P}^t_2, \forall t\}$. At the beginning of each

time frame t, Algorithm 0 invokes Algorithms 1 and 2 to obtain x_i^t for scheduling global iterations for FL tasks in the time frame t. Algorithm 1 provides the fractional decisions $\widetilde{x}_i^t$ and Algorithm 2 converts them into the integer decisions. Within each time frame t, at each time slot τ, it invokes Algorithm 3 to determine the inference requests migration $\left\{y_{jkn\tau}^t\right\}$. Fig. 3 illustrates the workflow of our approach.

3.2 Problem Decomposition

We decompose the problem $\mathbb{P}$ into $\mathbb{P}_1$ for FL task scheduling and a series of problems $\{\mathbb{P}_2^t, \forall t\}$ for inference requests scheduling. The objective function of the problem is split as $\mathbb{P}$ as $\mathbb{P} = \mathbb{P}_1 + \sum_{t\in\mathcal{T}} \mathbb{P}_2^t$, and the Constraint (1c) can also be split as (2c) in $\mathbb{P}_1$ and (3a) in $\{\mathbb{P}_2^t, \forall t\}$.

Denote $\mathcal{I}_t$ as $\mathcal{I}_t = \left\{i \in \mathcal{I}, t_i^{in} \le t \le t_i^{out}\right\}$. The formulations of $\mathbb{P}_1$ and $\mathbb{P}_2^t, \forall t$ are as follows.

$$\mathbb{P}_1: \text{Min} \quad r^t \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E_{ij}^t x_i^t + \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} \sum_{k\in\mathcal{A}_i} \left(\nu_{ijk}^t + q_{jk}^t + \nu_{ikj}^t + q_{kj}^t\right) x_i^t \alpha_{ik} \tag{2}$$

$$\text{s.t.} \quad \sum_{t\in\mathcal{T}: t_i^{in}\le t\le t_i^{out}} x_i^t \ge Q_i, \forall i, \tag{2a}$$

$$\sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E_{ij}^t x_i^t \le R^t, \tag{2b}$$

$$z_j^t = \sum_{i\in\mathcal{I}_t} \left(\delta_{ij} E_i x_i^t A_i + \alpha_{ij} x_i^t B_i\right), \forall j, t \tag{2c}$$

$$\text{var.} \quad x_i^t \in \mathbb{N}, z_j^t \ge 0.$$

$$\mathbb{P}_2^t: \text{Min} \quad r^t \sum_{t\in\mathcal{T}} \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} d_{kn} y_{jkn\tau}^t + \sum_{t\in\mathcal{T}} \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} (q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t) y_{jkn\tau}^t \tag{3}$$

$$\text{s.t.} \quad \sum_{\tau\in\mathcal{D}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{J}} \sum_{n\in\mathcal{N}} d_{kn} y_{jkn\tau}^t \le R^t - \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E_{ij}^t x_i^t, \tag{3a}$$

$$S_n y_{jkn\tau}^t \le \beta_{kn} C_k, \forall j, k, n, \tau, t, \tag{3b}$$

$$\sum_{k\in\mathcal{J}} y_{jkn\tau}^t \ge w_{jn\tau}^t, \forall j, n, \tau, \tag{3c}$$

$$y_{jkn\tau}^t \left[q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t - SLO_{n\tau}^t \right]^+ = 0, \quad \forall j, k, n, \tau, t, \tag{3d}$$

$$\sum_{\tau\in D} z_{j\tau}^t = z_j^t, \qquad \forall j, t \tag{3e}$$

$$z_{k\tau}^t + \sum_{n\in\mathcal{N}} \sum_{j\in\mathcal{J}} S_n y_{jkn\tau}^t \le C_k, \forall k, \tau, t \tag{3f}$$

$$\text{var.} \quad y_{jkn\tau}^t \ge 0.$$

We have split the objective function of $\mathbb{P}$ into $\mathbb{P}_1$ and $\{\mathbb{P}_2^t, \forall t\}$. The Constraints (1a)∼(1g) are split into Constraints (2a)∼(2c) for $\mathbb{P}_1$ and Constraints (3a)∼(3f) for $\{\mathbb{P}_2^t, \forall t\}$ respectively.

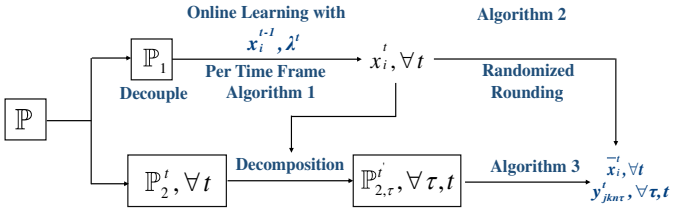


Fig. 3: Structure of the Proposed Approach

Algorithm 0: Overall Online Control Framework

```
for t ∈ {1, 2, ..., T} do
    ▷ Macro-timescale FL Tasks Scheduling
    Invoke Algorithm 1 to get fractional solutions X̃^t ;
    Invoke Algorithm 2 to get integral solutions X̄^t ;
    Calculating Z^t based on X̄^t;
    ▷ Micro-timescale Inference Requests Scheduling
    for τ ∈ {1, 2, ..., D} do
        Invoke Algorithm 3 to get solutions Y_τ^t;
```

3.3 Algorithms for Macro-Timescale FL Task Scheduling

To solve the macro-timescale FL Task scheduling problem $\mathbb{P}_1$, we relax it to $\widetilde{\mathbb{P}}_1$ in the continuous domain and adopt a concise representation of the problem. We denote the decision variables of $\mathbb{P}_1$ as $\mathbf{X}^t = [x_1^t, ..., x_I^t]$, the fractional decisions of $\widetilde{\mathbb{P}}_1$ and $\mathbf{Z}^t = [z_1^t, ..., z_J^t]$. as $\widetilde{\mathbf{X}}^t = [\widetilde{x}_1^t, ..., \widetilde{x}_I^t]$.

Then the objective function of $\widetilde{\mathbb{P}}_1$ can be expressed as $f^t\left(\widetilde{\mathbf{X}}^t\right) = r^t \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E_{ij}^t \widetilde{x}_i^t + \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} \sum_{k\in\mathcal{A}_i} \left(\nu_{ijk}^t + q_{jk}^t + \nu_{ikj}^t + q_{kj}^t\right) \widetilde{x}_i^t \alpha_{ik}$. We also introduce some new notations: $\widetilde{g}_i^t = \frac{Q_i}{t_i^{out} - t_i^{in}} - \widetilde{x}_i^t$; $g^t\left(\widetilde{\mathbf{X}}^t\right) = [\widetilde{g}_1^t, ..., \widetilde{g}_I^t]^{\mathrm{T}}$, and $h^t\left(\widetilde{\mathbf{X}}^t\right) = R^t - \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{A}_i} E_{ij}^t \widetilde{x}_i^t$. Here δ_{ij} is omitted for constraint tightening, for the tractability of randomized rounding. Thus, we can represent the relaxed problem of $\widetilde{\mathbb{P}}_1$ as

$$\text{Min} \quad \sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right) \tag{4}$$

$$\text{s.t.} \quad \sum_{t\in\mathcal{T}} g^t\left(\widetilde{\mathbf{X}}^t\right) \le \mathbf{0} \tag{4a}$$

$$h^t(\widetilde{\mathbf{X}}^t) \ge 0 \tag{4b}$$

$$\text{Var.} \quad \widetilde{\mathbf{X}}^t \in \widetilde{\mathcal{X}} = \{\widetilde{\mathbf{X}}^t | x_i^t \ge 0, \forall i \in \mathcal{I}\}.$$

According to the Lagrange-Dual method, we can solve the above problem (4) via solving an equivalent convex-concave problem which is formulated as

$$\text{Min}_{\widetilde{\mathbf{X}}^t} \text{Max}_{\boldsymbol{\lambda}^t} \mathcal{L}^t(\widetilde{\mathbf{X}}^t, \boldsymbol{\lambda}) \quad \text{s.t. } h^t(\widetilde{\mathbf{X}}^t) \ge 0, \widetilde{\mathbf{X}}^t \in \mathcal{X}, \tag{5}$$

Algorithm 1: Macro-Timescale Fractional Algorithm

Input: Fractional solution $\widetilde{\mathbf{X}}^t$; dual solution $\boldsymbol{\lambda}^t$
Output: Fractional solution $\widetilde{\mathbf{X}}^{t+1}$
Calculate $\boldsymbol{\lambda}^{t+1}$ according to (6);
Calculate $\widetilde{\mathbf{X}}^{t+1}$ by solving the problem (7).

where $\mathcal{L}^t(\widetilde{\mathbf{X}}^t, \boldsymbol{\lambda}) = f^t(\widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda}^{t\top} g^t(\widetilde{\mathbf{X}}^t)$, and $\boldsymbol{\lambda}^t$ is the Lagrange multiplier at t. Note that, through this transformation, the long-term constraint has been eliminated.

To solve the problem (5), we design the following online algorithm using an alternating primal-dual method. At each time frame $t+1$, we first update the dual variable $\boldsymbol{\lambda}^{t+1}$ via a standard dual ascent step:

$$\boldsymbol{\lambda}^{t+1} = [\boldsymbol{\lambda}^t + \mu g^t(\widetilde{\mathbf{X}}^t)]^+, \tag{6}$$

where μ is the positive step size, and $g^t(\widetilde{\mathbf{X}}^t) = \nabla_{\boldsymbol{\lambda}} \mathcal{L}^t(\widetilde{\mathbf{X}}^t, \boldsymbol{\lambda}^t)$ is the gradient of $\mathcal{L}^t(\widetilde{\mathbf{X}}, \boldsymbol{\lambda})$ given $\boldsymbol{\lambda} = \boldsymbol{\lambda}^t$. Next, we adopt a modified descent step to minimize $\mathcal{L}^t(\widetilde{\mathbf{X}}, \boldsymbol{\lambda}^{t+1})$. We obtain the fractional primal solution $\widetilde{\mathbf{X}}^{t+1}$ by solving the following problem:

$$\begin{aligned} \text{Min } & \nabla f^t(\widetilde{\mathbf{X}}^t)^\top (\mathbf{X} - \widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda}^{t+1\top} g^t(\mathbf{X}) + \frac{\|\mathbf{X} - \widetilde{\mathbf{X}}^t\|^2}{2\alpha} \\ \text{s.t.} \quad & h^t(\mathbf{X}) \geq 0, \mathbf{X} \in \widetilde{\mathcal{X}}, \end{aligned} \tag{7}$$

where α is a predefined constant; $\nabla f^t(\widetilde{\mathbf{X}}^t)$ is the gradient of $f^t(\mathbf{X})$ at $\mathbf{X} = \widetilde{\mathbf{X}}^t$; and $\frac{1}{2\alpha}\left\|\mathbf{X} - \widetilde{\mathbf{X}}^t\right\|^2$ is a regularization term. Note that via such ascent and descent steps, $\widetilde{\mathbf{X}}^{t+1}$ and $\boldsymbol{\lambda}^{t+1}$ can be solved using only the information known so far, rather than the future unknown information. This is the key of *online learning*. This algorithm is shown as Algorithm 1. Algorithm 1 is polynomial-time, because the problem (7) can be solved using standard optimization solvers such as CasADi [41], which finds an ϵ-accurate optimal solution in $O(I^2 \log(1/\epsilon))$ iterations [42] via the interior point method.

We design Algorithm 2 to convert the fractional solution $\{\widetilde{x}_i^t\}$ from Algorithms 1, into integers $\{\overline{x}_i^t\}$ in a randomized manner, ensuring the following objectives: (1) each fraction $\widetilde{x}_i^t$ is rounded to an integer; (2) no violation of constraint (2b) after rounding; (3) the expectation of the integer equal to the corresponding fraction, that is, $E(\bar{x}_i^t) = \widetilde{x}_i^t, \forall i \in \mathcal{I}$. Achieving these aims is important for deriving our theoretical performance analysis later.

In each iteration, Algorithm 2 chooses a pair of fractions to round at least one of them into an integer in a randomized manner, while ensuring that the weighted sum of the two values remains unchanged after rounding. Algorithm 2 has a time complexity of $O(I)$ for the macro-timescale rounding.

Lemma 1. *Algorithm 3 satisfies the aforementioned requirements (2) and (3): (2) no violation of constraint (2b) after rounding; (3) the expectation of the integer equal to the corresponding fraction, that is, $E(\bar{x}_i^t) = \widetilde{x}_i^t, \forall i \in \mathcal{I}$.*

Proof. If $\widetilde{v'}_{i_2}^t$ is an integer, then $E(\overline{v}_{i_2}^t) = \frac{\gamma_2}{\gamma_1+\gamma_2}(\widetilde{v}_{i_2}^t - q\gamma_1) + \frac{\gamma_1}{\gamma_1+\gamma_2}(\widetilde{v}_{i_2}^t + q\gamma_2) = \widetilde{v}_{i_2}^t, \forall \widetilde{v}_{i_2}^t \in \mathcal{I}_t \backslash \mathcal{I}_t'$, thus (3) is satisfied; so as the situation that $\widetilde{v'}_{i_1}^t$ is an integer. This equation demonstrates that the expectation of the integer is preserved.

Algorithm 2: Randomized Rounding Algorithm

Input: Fractional solution $\widetilde{\mathbf{X}}^t$
Output: Integral solution $\overline{\mathbf{X}}^t$
Define $\widetilde{\mathbf{V}} = \{\widetilde{v}_i^t = \widetilde{x}_i^t - \lfloor \widetilde{x}_i^t \rfloor, \forall i \in \mathcal{I}_t\}$;
Define $\mathcal{I}_t' = \mathcal{I}_t \backslash \{i | \widetilde{v}_i^t \in \{0, 1\}\}$;
while $\mathcal{I}_t' \neq \emptyset$ **do**
  if $|\mathcal{I}_t'| = 1$ **then** Set $\overline{x}_i^t = \lfloor \widetilde{x}_i^t \rfloor$ for the only $i \in \mathcal{I}_t'$;
  else
    Select $i_1, i_2 \in \mathcal{I}_t', i_1 \neq i_2$, define $q = \frac{\sum_{j \in \mathcal{A}_{i_1}} E_{i_1 j}^t}{\sum_{j \in \mathcal{A}_{i_2}} E_{i_2 j}^t}$;
    $\gamma_1 = \min\{\lceil \widetilde{v}_{i_1}^t \rceil - \widetilde{v}_{i_1}^t, \frac{1}{q} \widetilde{v}_{i_2}^t\}$,
    $\gamma_2 = \min\{\frac{1}{q}\left(\lceil \widetilde{v}_{i_2}^t \rceil - \widetilde{v}_{i_2}^t\right), \widetilde{v}_{i_1}^t\}$;
    With probability $\frac{\gamma_2}{\gamma_1+\gamma_2}$,
    set $\widetilde{v'}_{i_1}^t = \widetilde{v}_{i_1}^t + \gamma_1$, $\widetilde{v'}_{i_2}^t = \widetilde{v}_{i_2}^t - q\gamma_1$;
    With probability $\frac{\gamma_1}{\gamma_1+\gamma_2}$,
    set $\widetilde{v'}_{i_1}^t = \widetilde{v}_{i_1}^t - \gamma_2$, $\widetilde{v'}_{i_2}^t = \widetilde{v}_{i_2}^t + q\gamma_2$;
    if $\widetilde{v'}_{i_1}^t \in \{0, 1\}$, **then** $\overline{x}_i^t = \lfloor \widetilde{x}_i^t \rfloor + \widetilde{v'}_{i_1}^t, \mathcal{I}_t' = \mathcal{I}_t' \backslash \{i_1\}$;
    else set $\widetilde{v}_{i_1}^t = \widetilde{v'}_{i_1}^t$;
    if $\widetilde{v'}_{i_2}^t \in \{0, 1\}$, **then** $\overline{x}_i^t = \lfloor \widetilde{x}_i^t \rfloor + \widetilde{v'}_{i_2}^t, \mathcal{I}_t' = \mathcal{I}_t' \backslash \{i_2\}$;
    else set $\widetilde{v}_{i_2}^t = \widetilde{v'}_{i_2}^t$;

To show no violation of Constraint (2b), we have $\sum_{j \in A_{i_1}} E_{i_1 j}^t \widetilde{v'}_{i_1} + \sum_{j \in A_{i_2}} E_{i_2 j}^t \widetilde{v'}_{i_2} = \sum_{j \in A_{i_1}} E_{i_1 j}^t (\widetilde{v}_{i_1} + \gamma_1) + \sum_{j \in A_{i_2}} E_{i_2 j}^t (\widetilde{v}_{i_2} - q\gamma_1) = \sum_{j \in A_{i_1}} E_{i_1 j}^t (\widetilde{v}_{i_1} - \gamma_2) + \sum_{j \in A_{i_2}} E_{i_2 j}^t (\widetilde{v}_{i_2} + q\gamma_2) = \sum_{j \in A_{i_1}} E_{i_1 j}^t \widetilde{v}_{i_1} + \sum_{j \in A_{i_2}} E_{i_2 j}^t \widetilde{v}_{i_2}$ in each iteration, whether choosing Line 9 or Line 11. □

3.4 Algorithm for Micro-Timescale Inference Requests Scheduling

The problem $\mathbb{P}_2^t$ can be naturally split into a series of one-slot problems $\mathbb{P}_{2,\tau}^{t\prime}, \forall \tau$, corresponding to individual time slots. Specifically, to address the long-term constraint (3a), we propose to evenly partition (3a) into a set of constraints corresponding to individual time slots. That is, (3a) can be reformulated as

$$\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{J}} \sum_{n \in \mathcal{N}} d_{kn} y_{jkn\tau}^t \leq \frac{R^t - \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{A}_i} E_{ij}^t \widetilde{x}_i^t}{D}, \forall \tau. \tag{8}$$

The constraint (3d) is difficult to handle using standard solvers. Therefore, in each time slot τ of the time frame t, we precompute the set of QoS-feasible assignments $\mathcal{E}_{n\tau}^{t,QoS}$, as

$$\mathcal{E}_{n\tau}^{t,QoS} = \left\{(j,k) \,|q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t \leq SLO_{n\tau}^t\right\} \tag{9}$$

and simply enforce $y_{jkn\tau}^t = 0$ for $(j,k) \notin \mathcal{E}_{n\tau}^{t,QoS}$.

Denote $\mathbf{Z}_\tau^t = \left\{z_{j\tau}^t, \forall j\right\}$. We evenly partition z_j^t into different time slots in the time frame t, as $z_{j\tau}^t = \frac{1}{D} z_j^t$. Thus (3e) is reformulated as

$$\frac{1}{D} z_k^t + \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{J}} S_n y_{jkn\tau}^t \leq C_k, \forall k, \tau, t \tag{10}$$

Algorithm 3: Micro-Timescale Fractional Algorithm

Input: $\left\{q_{jk\tau}^t\right\}$, $\{b_{nk}^t\}$, $\left\{\omega_{jkn\tau}^t\right\}$,$\{d_{jn}\}$, $\{S_n\}$,$\{\beta_{kn}\}$, $\left\{E_{ij}^t\right\}$, $\left\{w_{jn\tau}^t\right\}$, $\{C_j\}$, $\{SLO_{n\tau}^t\}$, $\{\overline{x}_i^t\}$, $\left\{z_j^t\right\}$

Output: Fractional solution $\mathbf{Y}_\tau^t$

Solve the one-shot instance of $\mathbb{P}_{2,\tau}^{t'}$ at time slot τ using a solver (e.g., the interior-point method);

Thus $\mathbb{P}_{2,\tau}^{t\prime}, \forall\tau$ is formulated as

$$\begin{aligned}
\mathbb{P}_{2,\tau}^{t\prime}: \text{Min} \quad & f_\tau^t\left(\mathbf{Y}_\tau^t\right) && (11)\\
\text{s.t.} \quad & (3b), (3c), (8), (10),\\
& f_\tau^t\left(\mathbf{Y}_\tau^t\right) = r^t \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn} y_{jkn\tau}^t\\
& + \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} (q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t) y_{jkn\tau}^t\\
\text{var.} \quad & y_{jkn\tau}^t \geqslant 0.
\end{aligned}$$

Note that r^t is unknown until at the end of the time frame t, and here we estimate r^t as $\widehat{r}^t = r^{t-1}$ for tractability. Algorithm 3 is executed at each time slot τ in each time frame t to determine the amount of workload that needs to be migrated across heterogeneous edges. We define the solution set of problem $\mathbb{P}_{2,\tau}^{t'}$ as $\mathbf{Y}_\tau^t = \left\{y_{jkn\tau}^t, \forall j,k,n\right\}$. $\mathbb{P}_{2,\tau}^{t'}$ is a standard linear program and can be solved optimally by existing optimization solvers in polynomial time [43]. The time complexity for Algorithm 3, the complexity is $O(N^2J^4\log(1/\epsilon))$, since for each problem $\mathbb{P}_{2,\tau}^{t\prime}, \forall\tau$ we have NJ^2 decision variables.

3.5 Regret and Fit Analysis

We introduce "dynamic regret" and "dynamic fit" [27], [28] as the performance metrics, and rigorously analyze these metrics for our algorithms.

To ease of expression for the definition of dynamic regret and dynamic fit, we represent the original problem $\mathbb{P}$ in a more concise form as follows.

$$\begin{aligned}
\mathbb{P}: \text{Min} \quad & \sum_{t\in\mathcal{T}} f_o^t\left(\mathbf{W}^t\right) && (12)\\
\text{s.t.} \quad & \sum_{t\in\mathcal{T}} g_o^t\left(\mathbf{W}^t\right) \leq \mathbf{0}, && (12a)\\
& h_o^t\left(\mathbf{W}^t\right) \geq \mathbf{0}, \forall t, && (12b)\\
& h_5^t\left(\mathbf{W}^t\right) = \mathbf{0}, \forall t && (12c)\\
& f_o^t\left(\mathbf{W}^t\right) := \quad r^t \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} E_{ij}^t x_i^t\\
& + r^t \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn} y_{jkn\tau}^t\\
& + \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i} \left(\nu_{ijk}^t + q_{jk}^t + \nu_{ikj}^t + q_{kj}^t\right) x_i^t \alpha_{ik}\\
& + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} (q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t) y_{jkn\tau}^t, && (12d)\\
& g_o^t\left(\mathbf{W}^t\right) := \left[g_1^t, ..., g_I^t\right]^{\mathrm{T}} && (12e)\\
& h_o^t\left(\mathbf{W}^t\right) := \left[h_1^t, \mathbf{h}_2^t, \mathbf{h}_3^t, \mathbf{h}_4^t\right]^{\mathrm{T}} && (12f)\\
& h_5^t\left(\mathbf{W}^t\right) := \left[h_{5,1}^t, ..., h_{5,J}^t\right]^{\mathrm{T}} && (12g)\\
& h_1^t := R^t - \sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} E_{ij}^t x_i^t\\
& - \sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn} y_{jkn\tau}^t, && (12i)\\
& \mathbf{h}_2^t := \left\{h_{2,jkn\tau}^t = S_n y_{jkn\tau}^t - \beta_{kn} C_k, \forall j,k,n,\tau\right\}, && (12j)\\
& \mathbf{h}_3^t = \left\{h_{3,jn\tau}^t = \sum_{k\in\mathcal{J}} y_{jkn\tau}^t - w_{jn\tau}^t, \forall j,n,\tau\right\}, && (12k)\\
& \mathbf{h}_4^t = \left\{h_{4,k\tau}^t = C_k - z_{k\tau}^t - \sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{J}} S_n y_{jkn\tau}^t, \forall k,\tau\right\}, && (12l)\\
& h_{t,j}^t = z_j^t - \sum_{\tau\in D} z_{j\tau}^t, \forall j && (12m)\\
& \mathbf{W}^t = \left[\mathbf{X}^t, \mathbf{Z}^t, \mathbf{Y}_1^t, \mathbf{Y}_2^t, ..., \mathbf{Y}_D^t, \mathbf{Z}_1^t, \mathbf{Z}_2^t, ..., \mathbf{Z}_D^t\right] && (12n)\\
\text{var.} \quad & x_i^t \in \mathbb{N}, y_{jkn\tau}^t \geq 0, z_j^t \geq 0, z_{j\tau}^t \geq 0
\end{aligned}$$

We use $\overline{\mathbf{W}}^t = [\overline{x}_1^t, \overline{x}_2^t, ..., \overline{x}_i^t, \mathbf{Y}_1^t, \mathbf{Y}_2^t, ..., \mathbf{Y}_D^t]$ to denote the online decisions with integral $x_i^t \in \mathbb{N}$ and $\overline{\mathbf{W}}^{t*}$ to denote the offline optimal decisions with integral $x_i^t \in \mathbb{N}$ for the single time-frame problem at t. Analogously, we can introduce the notations $\widetilde{\mathbf{W}}^t$ and $\widetilde{\mathbf{W}}^{t*}$ for the optimal decisions for the relaxed problems in the continuous domain.

Dynamic Regret: The dynamic regret measures the difference between the objective function value evaluated with the online solutions and that evaluated with the offline single-time-frame optimal solutions. The dynamic regret for our original problem $\mathbb{P}$ and the dynamic regret for the relaxed problem $\widetilde{\mathbb{P}}$ are as follows, respectively:

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{T}} &:= E\left[\sum_{t\in\mathcal{T}} f_o^t\left(\overline{\mathbf{W}}^t\right)\right] - \sum_{t\in\mathcal{T}} f_o^t\left(\overline{\mathbf{W}}^{t*}\right)\\
\overline{\mathbf{W}}^{t*} &= \underset{x_i^t\in\mathbb{N}, y_{jkn\tau}^t\geq 0}{\arg\min} f_o^t\left(\mathbf{W}^t\right), \text{s.t. } g_o^t\left(\mathbf{W}^t\right) \leq \mathbf{0}, h_o^t(\mathbf{W}^t) \geq \mathbf{0},\\
\widetilde{\mathrm{Reg}}_{\mathcal{T}} &:= \sum_{t\in\mathcal{T}} f_o^t\left(\widetilde{\mathbf{W}}^t\right) - \sum_{t\in\mathcal{T}} f_o^t\left(\widetilde{\mathbf{W}}^{t*}\right)\\
\widetilde{\mathbf{W}}^{t*} &= \underset{x_i^t\geq 0, y_{jkn\tau}^t\geq 0}{\arg\min} f_o^t\left(\mathbf{W}^t\right), \text{s.t. } g_o^t\left(\mathbf{W}^t\right) \leq \mathbf{0}, h_o^t(\mathbf{W}^t) \geq \mathbf{0},
\end{aligned}$$

Dynamic Fit: The dynamic fit measures the cumulative violation of the long-term constraints evaluated with the online solutions. The dynamic fit for our original problem $\mathbb{P}$ and the dynamic fit for the relaxed problem $\mathbb{P}$ are as follows, respectively, where $[\cdot]^+ = \max\{\cdot, 0\}$:

$$\begin{aligned}
Fit_{\mathcal{T}} &:= \left\| \left[E\left[\sum_{t\in\mathcal{T}} g_o^t\left(\overline{\mathbf{W}}^t\right)\right]\right]^+ \right\|,\\
\widetilde{Fit}_{\mathcal{T}} &:= \left\| \left[\sum_{t\in\mathcal{T}} g_o^t\left(\widetilde{\mathbf{W}}^t\right)\right]^+ \right\|.
\end{aligned}$$

Regret and Fit Analysis: We prove that, for our original problem $\mathbb{P}$, the dynamic regret and the dynamic fit evaluated with the online integral solutions produced by our

algorithms only grow sublinearly with the length of time horizon.

To derive this proof, we leverage several assumptions which are very commonly made in a wide range of similar problems [27], [28], [44], [45]:

Assumption 1: $f_o^t\left(\widetilde{\mathbf{W}}^t\right)$ has bounded gradients, i.e., $\|\nabla f_o^t(\widetilde{\mathbf{W}}^t)\| \leq F$, $g_o^t(\widetilde{\mathbf{W}}^t)$ is also bounded, i.e., $\|g_o^t(\widetilde{\mathbf{W}}^t)\| \leq U, \forall t$.

Assumption 2: there exists a constant $\varepsilon > 0$ and an interior point $\widehat{\mathbf{W}}^t$, such that $g_o^t(\widehat{\mathbf{W}}^t) \preceq -\epsilon \mathbf{1}, \forall t$.

Assumption 3: the slack constant ε in *Assumption 2* is larger than the point-wise maximal variation of the consecutive constraints, i.e., $\varepsilon > \overline{\mathcal{V}}(g)$, where $\overline{\mathcal{V}}(g) = \max_t \mathcal{V}_{\mathbf{g^t}}$ and $\mathcal{V}_{\mathbf{g^t}} = \max_{\widetilde{\mathbf{W}}} \|[g_o^{t+1}(\widetilde{\mathbf{W}}^t) - g_o^t(\widetilde{\mathbf{W}}^t)]^+\|$.

Based on all these assumptions, we derive the following results for the dynamic regret and the dynamic fit:

Theorem 1. *We have* $\mathrm{Reg}_{\mathcal{T}} \leq \widetilde{\mathrm{Reg}}_{\mathcal{T}}$ *and* $\mathrm{Fit}_{\mathcal{T}} \leq \widetilde{\mathrm{Fit}}_{\mathcal{T}} + M + \mathcal{K}\varkappa_{\varpi}^{\varpi}$*, where* $\mathcal{K}$ *and* $\varkappa_{\varpi}^{\varpi}$ *are constants from the Jensen Gap [46].*

Proof. See in Appendix B, based on Lemma 2 in Appendix A. □

Theorem 2. *Under previous assumptions and the dual variable initialization of* $\boldsymbol{\lambda}^1 = 0$*, the integral dynamic fit is upper-bounded:*

$$\mathrm{Fit}_{\mathcal{T}} \leq \widetilde{\mathrm{Fit}}_{\mathcal{T}} + \mathcal{K}\varkappa_{\varpi}^{\varpi} \leq \frac{\|\bar{\boldsymbol{\lambda}}\|}{\mu} + \mathcal{K}\varkappa_{\varpi}^{\varpi}. \tag{13}$$

Proof. See in Appendix D, based on Lemma 3 in Appendix C. □

Theorem 3. *Under previous assumptions and the dual variable initialization of* $\boldsymbol{\lambda}^1 = 0$*, the integral dynamic regret is upper-bounded:*

$$\mathrm{Reg}_{\mathcal{T}} \leq \widetilde{\mathrm{Reg}}_{\mathcal{T}} \leq \mathcal{A}_T + \mathcal{L}, \tag{14}$$

where $\mathcal{A}_T$*, and* $\mathcal{L}$ *are given as*

$$\mathcal{A}_T = \frac{\mu U^2(T+1)}{2} + \frac{\alpha F^2 T}{2} + \|\overline{\boldsymbol{\lambda}}\|\mathcal{V}_{\mathbf{g^t}}^T + \frac{R\mathcal{V}_{\widetilde{\mathbf{X}}^{t*}}^T}{\alpha} + \frac{R^2}{2\alpha}, \tag{15}$$

$$\begin{aligned}\mathcal{L} = & \max_{i,j,k,t}\left\{\frac{\left(\nu_{ijk}^t + q_{jk}^t + \nu_{ikj}^t + q_{kj}^t\right)\alpha_{ik}}{r^t E_{ij}^t}\right\} J^2 \cdot \max_t\left\{r^t R^t\right\} \cdot T \\ & + \max_t\left(r^t R^t\right) \cdot T + \frac{\max_{k,n}\left(d_{kn} C_k\right)}{\min_n S_n} \cdot TDJ \\ & + \frac{\max_{j,k,n,\tau,t}\left(\left(q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t\right) C_k\right)}{\min_n S_n} \cdot TDJ \\ & - 2TJ^2 \min_{i,j,k,t}\left\{\left(\nu_{ijk}^t + q_{jk}^t\right)\alpha_{ik}\right\} - IJT \cdot \min_{i,j,t}\left(r^t E_{ij}^t Q_i\right) \\ & - \min_{j,k,n,\tau,t}\left(\left(q_{jk\tau}^t + q_{kj\tau}^t + b_{nk}^t + \omega_{jkn\tau}^t\right) w_{jn\tau}^t\right) \cdot TDNJ \\ & - \min_{j,k,n,\tau,t}\left(d_{kn} w_{jn\tau}^t\right) \cdot TDNJ \end{aligned} \tag{16}$$

Proof. See in Appendix G, based on Lemma 4 and Lemma 5 in Appendix E and Appendix F, respectively. □

Corollary 1. *Under Assumptions 1 ∼ 3 and previous initialization, dynamic regret and dynamic fitness are bounded by controlling step sizes:*

$$\begin{aligned} \alpha = \mu &= \max\left\{\sqrt{\frac{\mathcal{V}_{\widetilde{\mathbf{X}}_t^*}^T}{T}}, \sqrt{\frac{\mathcal{V}_{\mathbf{g}_t}^T}{T}}\right\}, \\ \mathrm{Reg}_{\mathcal{T}} &\leq \mathcal{O}\left(\max\left\{\sqrt{\mathcal{V}_{\widetilde{\mathbf{X}}_t^*}^T T}, \sqrt{\mathcal{V}_{\mathbf{g}_t}^T T}\right\}\right) + \mathcal{L}, \\ \mathrm{Fit}_{\mathcal{T}} &\leq \mathcal{O}\left(\max\left\{\frac{T}{\mathcal{V}_{\widetilde{\mathbf{X}}_t^*}^T}, \frac{T}{\mathcal{V}_{\mathbf{g}_t}^T}\right\}\right) + M + \mathcal{K}\varkappa_{\varpi}^{\varpi}. \end{aligned} \tag{17}$$

Based on the corollary, if we further set $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{3}})$, the dynamic regret and the dynamic fit can be expressed respectively as: $\mathrm{Reg}_{\mathcal{T}} \leq \mathcal{O}\left(\max\left\{\mathcal{V}_{\widetilde{\mathbf{X}}_t^*}^T T^{\frac{1}{3}}, \mathcal{V}_{\mathbf{g_t}}^T T^{\frac{1}{3}}, T^{\frac{2}{3}}\right\}\right) + \mathcal{L}$, $\mathrm{Fit}_{\mathcal{T}} \leq \mathcal{O}\left(T^{\frac{2}{3}}\right) + M + \mathcal{K}\varkappa_{\varpi}^{\varpi}$.

3.6 Complexity Analysis

The time complexity of our overall control framework Algorithm 0 is analyzed as follows. In each time frame t, Algorithm 1, and Algorithm 2 are invoked once respectively, and Algorithm 3 is invoked for D times. According to previous analysis, Algorithm 1 takes $O(I^2 \log(1/\epsilon))$ to find an ϵ-accurate optimal solution. Algorithm 2 takes $O(I)$. Algorithm 3 takes $O(N^2 J^4 \log(1/\epsilon))$ to find an ϵ-accurate optimal solution. Therefore, the time complexity of our entire algorithmic approach is $O\left(T\left(I^2 \log(1/\epsilon) + I + DN^2 J^4 \log(1/\epsilon)\right)\right)$. Here, I is the number of training tasks; N is the number of inference tasks; J is the number of edge nodes under consideration; D is the number of time slots per time frame; T is the number of time frames of the entire time horizon.

3.7 Adaptation for Large Scale Deployment

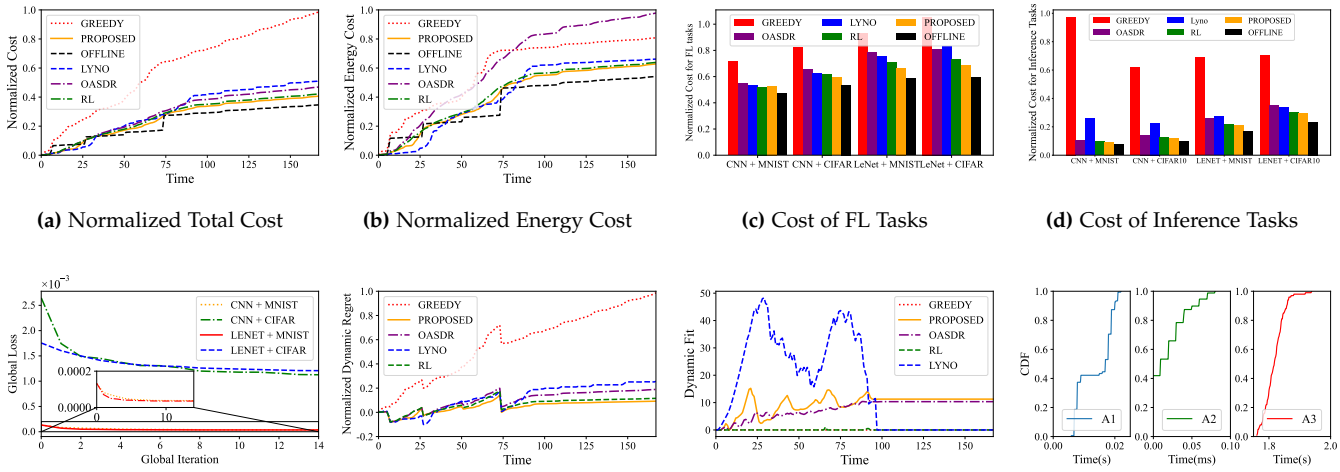
Although our algorithm achieves efficient performance under the evaluated scale (e.g., 50 edges and up to 500 tasks), scaling to much larger systems (e.g., thousands of nodes) may introduce computational bottlenecks. In particular, the inference task scheduling component involves migration decisions across all node pairs, which can become increasingly burdensome as the number of edge nodes grows.

To address this, we propose a practical extension via a cluster-based scheduling scheme. Specifically, the entire edge network can be partitioned into multiple clusters based on geographic or network proximity, and inference tasks are only allowed to migrate within each cluster. Each cluster independently solves a smaller-scale optimization problem, thereby significantly reducing the computation overhead and communication complexity.

4 Experimental Study

4.1 Experimental Settings

Edge System: We adopt the EUA dataset which contains 95,562 edge sites [47], and we select the first 1000 edge sites. The geographical distance between two stations is used to approximate the propagation delay between the edges [33]. Considering possible congestion and other situations, in each time slot, we randomly set the delay between edges



(a) Normalized Total Cost **(b)** Normalized Energy Cost **(c)** Cost of FL Tasks **(d)** Cost of Inference Tasks

(e) Training Loss **(f)** Dynamic Regret **(g)** Dynamic Fit **(h)** Execution Time

Fig. 4: Evaluation Results

to $[0.7, 1.3]$ times the original delay. We study the system for $T = 168$ hours and the length of each time frame is set as one hour [18] and each time frame would be divided into 12 time slots where each time slot continues for 5 minutes [32]. We set the energy cost per unit computation γ_j randomly from $[0.04, 0.08]$ mWh [48]; the transmission power p_j^t is from $[2, 20]$ dBm [48]; the channel gain h_j^t is from $[-90, -95]$ dB [49]; the noise power spectral density is $N_0 = -174$ dBm/Hz [45] and the bandwidth B is 25 kHz [48]. For the partial frequency reuse, we set $\rho \in [0.1, 0.2]$.

FL Tasks: In this system, we focus on FL tasks for image classification, using the MNIST dataset [30] and CIFAR-10 dataset [29]. We consider two models: the LeNet-5 [31] and a Convolutional Neural Network (CNN) with two 3×3 convolutional layers (where the first layer has 16 channels and the second layer has 32 channels), with each of them followed by ReLU activation and 2×2 max pooling, a fully-connected layer, and a softmax output layer. Combining these datasets and models results in four distinct types of FL tasks. We consider $100 \sim 500$ FL tasks in total, with each type comprising one-fourth of the total FL tasks. The number of FL tasks arriving in each time frame is set in proportion to the dynamic job arrival trace of Google clusters [50], where the last FL task arrives no later than the 100th time frame to ensure successful training. The number of edges $|\mathcal{A}_i|$ for conducting training for each FL task is 30 [51] based on Google data as well. Without loss of generality, the number of local iterations per global iteration is set to $E_i = 5$ and the target accuracy is set to $\epsilon_i = 0.01, \forall i$. For the other parameters, we set $\sigma_j = 1, \forall j$, $G = 0.01$, $\mu = 1$ and $L = 1$ [52].

Inference Tasks: We use the dynamic passenger numbers in each station to represent the corresponding inference requests for each edge. The maximum processing capacity of each edge is set as the maximum request amount arriving at each edge. We consider $N = 4$ inference tasks and the requests for each task are evenly divided into the number of tasks being served in that time slot. The four inference tasks are image classification tasks, corresponding to the four types of FL tasks, with each type occupying a quarter of all inference requests. We consider the edge set for each inference task takes in the range $[200, 300]$.

Energy Demand Response: We set the energy caps $\{R^t, \forall t\}$ using the real-world EDR events of Elia from June 18, 2023 through June 24, 2023 [34]. We set the electricity price $\{r^t, \forall t\}$ based on the hourly real-time pricing data of ComEd from June 18, 2023 through June 24, 2023 [35].

Algorithms and Implementation: We implement and compare multiple different approaches: (1) `PROPOSED` refers to our proposed online algorithms; (2) `GREEDY` completes the FL tasks as soon as possible at each time frame and migrates inference requests to the most effective edge at each time slot while satisfying the EDR constraint. (3) `OASDR` [36], the online algorithm which aims only at minimizing the delay while setting a long-term constraint for long-term accumulative energy cost. (4) `LYNO` [37], Lyapunov Optimization Algorithm, which solves time-coupling stochastic optimization problems using virtual queues and drift-plus-penalty algorithm. (5) `RL`, a reinforcement learning-based algorithm which solves the problem $\mathbb{P}_1$ based on deep Q-learning framework and solves the problem $\mathbb{P}_2^t, \forall t$ based on DDPG. (6) `OFFLINE`, which solves the original problem $\mathbb{P}$ via the Gurobi optimization solver [53], knowing all the inputs over the entire time horizon in advance. The nonlinear mixed-integer program requires an unacceptably long time to solve, even with advanced solvers. Therefore, we employ the offline optimal fractional solutions as a lower bound to approximate the offline optimal mixed-integer solutions for our problem.

We conduct experiments of all the above algorithms on a commodity laptop in our lab, equipped with a 2.6-GHz 12-core Intel(R) Core(TM) i7 CPU and 16-GB memory.

4.2 Evaluation Results

Total Cost: Fig. 4 (a) presents a comparison of the normalized total cost across the entire time horizon for various algorithms. It is evident that our proposed online algorithm consistently yields a lower total cost compared to GREEDY

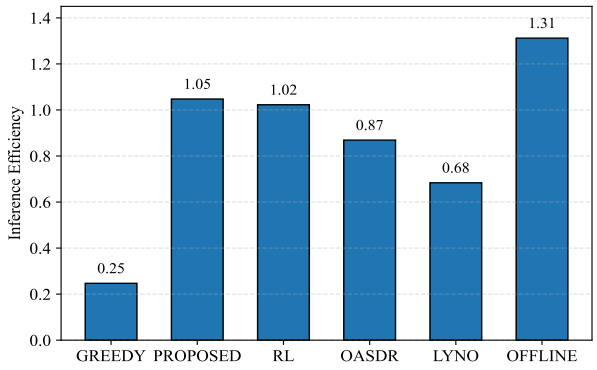


Fig. 5: Inference Efficiency

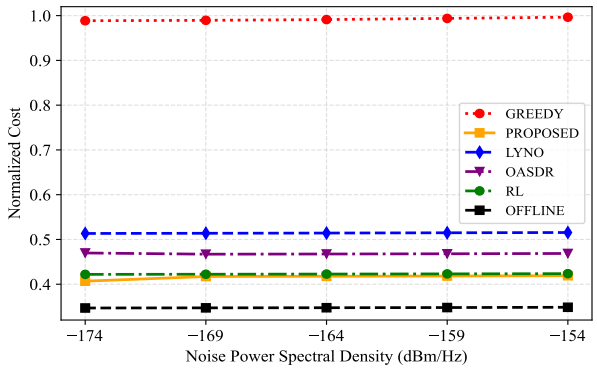


Fig. 6: Total Cost Under Varying Noise Conditions

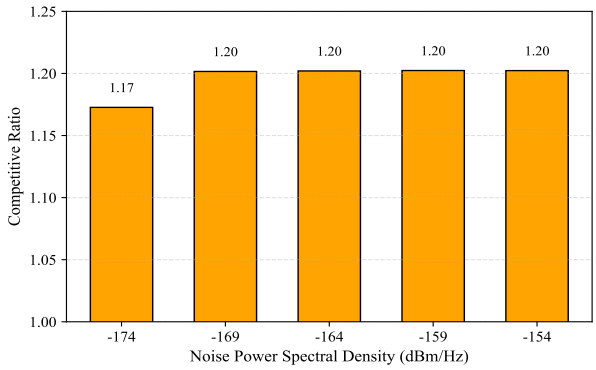


Fig. 7: Competitive Ratio Under Varying Noise Conditions

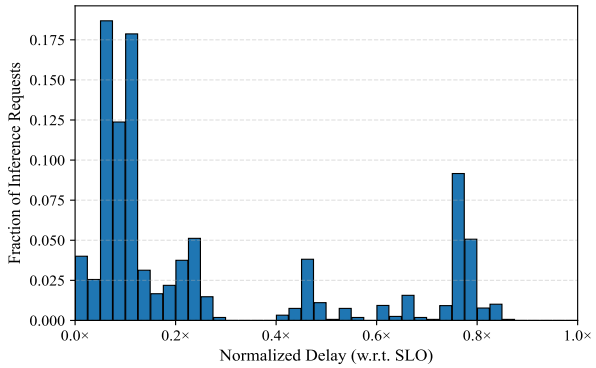


Fig. 8: QoS Performance

and OASDR. Our approach achieves 58.8% less total cost than GREEDY, 13.4% less total cost than OASDR, 20.8% less total cost than LYNO and 3.6% less total cost than RL.

Energy Cost: Fig. 4 (b) illustrates the cumulative energy cost of various algorithms across continuous time slots. Our proposed online algorithm has lower energy cost compared to GREEDY, OASDR, LYNO, and RL and it closely approaches the performance of the offline optimum.

Total cost for FL tasks and Training performance: We present our experimental results on the total cost and training performance for four kinds of different FL tasks, as well as the total cost for four types of inference tasks. Fig. 4 (c) and Fig. 4 (d) demonstrate that our algorithm can perform a lower cost compared to GREEDY, OASDR, LYNO and RL for different kinds of FL tasks and inference tasks. Fig. 4 (e) presents how the global loss is minimized in the FL tasks and how satisfied training loss can be achieved.

Dynamic Regret and Fitness: Fig. 4 (f) depicts the dynamic regret of each algorithm as the length of the entire time horizon varies. Our proposed online algorithm outperforms GREEDY, OASDR, LYNO and RL, exhibiting a slow, stepwise growth pattern. This figure also shows that the regret increases sublinearly over time in practice, consistent with our theoretical analysis. Fig. 4 (g) shows that the dynamic fit also has sub-linear growth.

Algorithm Running time: Fig. 4 (h) depicts the cumulative distribution of the execution time for each of our proposed algorithms 1-3. Algorithms 1 and 2 can be executed and finished within milliseconds. Algorithm 3 can be completed within 2 seconds. Hence, our proposed online algorithms are practically and computationally efficient.

Inference Efficiency: We define inference efficiency as $I_{eff} = \frac{\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} y^t_{jkn\tau}\eta^t_{jkn\tau}}{r^t \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}}\left(d_{kn}+q^t_{jk\tau}+q^t_{kj\tau}+b^t_{nk}+\omega^t_{jkn\tau}\right)y^t_{jkn\tau}}$, where $\eta^t_{jkn\tau} \in [0,1]$ is the inference accuracy of the inference task n on the edge k in the timeslot τ of timeframe t. Fig. 5 compares the inference efficiency of different methods. OFFLINE achieves the highest efficiency (1.31). PROPOSED and RL obtain comparable performance (1.05 and 1.02). OASDR and LYNO achieve lower efficiency (0.87 and 0.68), while GREEDY performs the worst (0.25). These results imply that our approach achieves near-optimal inference efficiency in dynamic environments.

System Performance and Robust Testness Under Varying Noise Condition: We evaluate the impact of background noise by varying the noise power spectral density N_0 from -174 dBm/Hz to -154 dBm/Hz. Fig. 6 presents the system performance under different noise levels, while Fig. 7 reports the corresponding competitive ratio. As shown in Fig. 6, the system cost exhibits only marginal variations with increasing noise. This is mainly because co-channel transmissions introduce strong aggregated interference, leading to an interference-limited regime where thermal noise plays a secondary role. Consequently, variations in N_0 have limited impact on the overall system cost. Fig. 7 further shows that the competitive ratio, defined as the ratio between the cost of the proposed online method and that of the offline optimal solution under the same noise condition, remains stable, ranging from 1.17 to 1.20 across all noise levels. This

indicates that the proposed method consistently maintains a near-optimal performance gap to the offline solution, demonstrating strong robustness against noise variations.

QoS Performance: Fig. 8 illustrates the SLO compliance. It can be observed that most inference requests are completed with normalized delays below $0.3\times$ SLO, while a secondary concentration appears around $0.8\times$ SLO. All measured delays remain within $1.0\times$ SLO, indicating that 100% of tasks satisfy the latency requirement. These results demonstrate that our proposed framework can effectively control inference latency under dynamic workloads while ensuring strict QoS guarantees.

5 Related Work

Edge Systems Demand Response: Liu *et al.* [14] focused on the interaction between intermittent renewable generations and variable computation under unknown EDR requirements. Wang *et al.* [15] proposed to utilize an auction mechanism to incentive EVs to power edge systems in EDR. Cui *et al.* [16] studied a two-stage game-theoretical approach to handle EDR in mobile edge systems under resource constraints. Wang *et al.* [17] proposed a bi-level optimization framework with a virtual region decomposition method to meet EDR needs.

These studies concentrate on the EDR issues in edge systems. None have considered AI/ML tasks, not to mention considering the characteristics of AI/ML tasks in the edge system, including the accuracy of the model, and the computational characteristics of AI/ML tasks.

AI/ML Tasks in Edge System: Wang *et al.* [18] proposed an auction-based approach to scheduling FL tasks in the edge system, while balancing the energy consumption and model accuracy. Wu *et al.* [19] studied an experience-driven deep reinforcement learning algorithm for sustainable FL while ensuring long-term economic properties. Li *et al.* [20] proposed a D2D-assisted FL system based on maximum weight matching in auxiliary graphs aiming to minimize global loss in the resource-constrained environment. Zhao *et al.* [21] worked out an online DNN model selection and placement solution to achieve a trade-off between inference accuracy, latency, and resource cost. Zhao *et al.* [22] designed a framework for adaptive distributed execution of DNN inference based on DNN partitions.

These works considered the features of AI/ML tasks when designing the system, but did not consider EDR. In addition, no consideration was given to the challenges of collocated inference and training tasks in an edge system.

Training and Inference Co-location: Chen [13] studied a dynamic scheduling system that flexibly allocates resources for training and inference tasks in a GPU cluster environment. Chen [23] proposed a two-tier training-inference colocation solution that provides explicit inference latency guarantees. Mobin [24] studied model-sharing between training and inference jobs in GPU memory-constrained environments. These studies have made significant progress in improving resource utilization and optimizing system performance, but fall short due to the lack of consideration for environmental sustainability explicitly, such as carbon emissions, energy consumption, and EDR.

In response to these limitations, Su *et al.* [25] proposed a carbon-aware edge system that colocates learning and inference tasks. But they did not consider the objective factor, that is, the difference between the execution time of the two tasks is so large that they cannot be decided at the same frequency. In addition, they fall short due to a lack of EDR and fine-grained energy control.

6 Conclusion

As edge intelligence grows, the energy demand escalates, highlighting the need for efficient EDR strategies. Our study proposes a two-timescale system for co-locating training and inference tasks in EDR systems and navigating the complex interplay between energy consumption, system delay, training model accuracy, and the uncertainty of future inputs through strategic macro-timescale training schedules and micro-timescale inference request migrations across heterogeneous edges. Our innovative online algorithm, which employs online learning and randomized rounding, achieves sublinear performance metrics and surpasses existing methods in practical evaluations. Future research will focus on enhancing these strategies for the evolving landscape of sustainable edge computing.

References

[1] D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, and P. Hui, "Edge intelligence: Empowering intelligence to the edge of network," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1778–1837, 2021.

[2] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020.

[3] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[4] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[5] Y. Zhang, H. Huang, L.-X. Yang, Y. Xiang, and M. Li, "Serious challenges and potential solutions for the industrial internet of things with edge intelligence," *IEEE Network*, vol. 33, no. 5, pp. 41–45, 2019.

[6] R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo, "Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning," *Sustainable Computing: Informatics and Systems*, vol. 38, p. 100857, 2023.

[7] S. Chen, L. Jiao, F. Liu, and L. Wang, "Edgedr: An online mechanism design for demand response in edge clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 2, pp. 343–358, 2022.

[8] G. Cui, Q. He, X. Xia, F. Chen, T. Gu, H. Jin, and Y. Yang, "Demand response in noma-based mobile edge computing: A two-phase game-theoretical approach," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1449–1463, 2021.

[9] Z. Zhou, F. Liu, S. Chen, and Z. Li, "A truthful and efficient incentive mechanism for demand response in green datacenters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 1, pp. 1–15, 2018.

[10] Z. Song, R. Zhou, S. Zhao, S. Qin, J. C. Lui, and Z. Li, "Edge emergency demand response control via scheduling in cloudlet cluster," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 394–399.

[11] L. Zeng, X. Chen, Z. Zhou, L. Yang, and J. Zhang, "Coedge: Cooperative dnn inference with adaptive workload partitioning over heterogeneous edge devices," *IEEE/ACM Transactions on Networking*, vol. 29, no. 2, pp. 595–608, 2021.

[12] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, J. Law, K. Lee, J. Lu, P. Noordhuis, M. Smelyanskiy, L. Xiong, and X. Wang, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018, pp. 620–629.

[13] Z. Chen, X. Zhao, C. Zhi, and J. Yin, "Deepboot: Dynamic scheduling system for training and inference deep learning tasks in gpu cluster," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 9, pp. 2553–2567, 2023.

[14] Y. Liu, S. Xie, Q. Yang, and Y. Zhang, "Joint computation offloading and demand response management in mobile edge network with renewable energy sources," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 720–15 730, 2020.

[15] F. Wang, L. Jiao, K. Zhu, and L. Zhang, "Online edge computing demand response via deadline-aware v2g discharging auctions," *IEEE Transactions on Mobile Computing*, pp. 1–14, 2022.

[16] G. Cui, Q. He, X. Xia, F. Chen, T. Gu, H. Jin, and Y. Yang, "Demand response in noma-based mobile edge computing: A two-phase game-theoretical approach," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1449–1463, 2023.

[17] J. Wang and Y. Peng, "Distributed optimal dispatching of multi-entity distribution network with demand response and edge computing," *IEEE Access*, vol. 8, pp. 141 923–141 931, 2020.

[18] F. Wang, L. Jiao, K. Zhu, X. Lin, and L. Li, "Toward sustainable ai: Federated learning demand response in cloud-edge systems via auctions."

[19] L. Wu, S. Guo, Y. Liu, Z. Hong, Y. Zhan, and W. Xu, "Sustainable federated learning with long-term online vcg auction mechanism," in *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, 2022, pp. 895–905.

[20] Y. Li, W. Liang, J. Li, X. Cheng, D. Yu, A. Y. Zomaya, and S. Guo, "Energy-constrained d2d assisted federated learning in edge computing," in *Proceedings of the 25th International ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems*, 2022, pp. 33–37.

[21] K. Zhao, Z. Zhou, X. Chen, R. Zhou, X. Zhang, S. Yu, and D. Wu, "Edgeadaptor: Online configuration adaption, model selection and resource provisioning for edge dnn inference serving at scale," *IEEE Transactions on Mobile Computing*, pp. 1–16, 2022.

[22] Z. Zhao, K. M. Barijough, and A. Gerstlauer, "Deepthings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2348–2359, 2018.

[23] G. Chen, S. Subramaniyan, and X. Wang, "Latency-guaranteed co-location of inference and training for reducing data center expenses," in *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, 2024, pp. 473–484.

[24] J. Mobin, A. Maurya, and M. M. Rafique, "Colti: Towards concurrent and co-located dnn training and inference," in *Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing*, 2023, pp. 309–310.

[25] S. Su, Z. Zhou, T. Ouyang, R. Zhou, and X. Chen, "Learning to be green: Carbon-aware online control for edge intelligence with colocated learning and inference," in *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2023, pp. 567–578.

[26] T. Chen, Q. Ling, and G. B. Giannakis, "An online convex optimization approach to proactive network resource allocation," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6350–6364, 2017.

[27] T. Chen, Q. Ling, Y. Shen, and G. B. Giannakis, "Heterogeneous online learning for "thing-adaptive" fog computing in iot," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4328–4341, 2018.

[28] T. Chen, Q. Ling, and G. B. Giannakis, "An online convex optimization approach to proactive network resource allocation," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6350–6364, 2017.

[29] "Cifar-10 dataset." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[30] "Mnist database." [Online]. Available: http://yann.lecun.com/exdb/mnist/

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[32] "London underground passenger counts data," https://tfl.gov.uk/info-for/open-data-users/.

[33] L. Jiao, L. Pu, L. Wang, X. Lin, and J. Li, "Multiple granularity online control of cloudlet networks for edge computing," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2018, pp. 1–9.

[34] "Demand response activity on january 7-8, 2014." [Online]. Available: http://www.pjm.com

[35] "Hourly pricing." [Online]. Available: https://hourlypricing.comed.com/live-prices/

[36] Y. Shi, C. Yi, R. Wang, Q. Wu, B. Chen, and J. Cai, "Service migration or task rerouting: A two-timescale online resource optimization for mec," *IEEE Transactions on Wireless Communications*, vol. 23, no. 2, pp. 1503–1519, 2024.

[37] Q. Tang, R. Xie, Z. Fang, T. Huang, T. Chen, R. Zhang, and F. R. Yu, "Joint service deployment and task scheduling for satellite edge computing: A two-timescale hierarchical approach," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 5, pp. 1063–1079, 2024.

[38] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2019.

[39] Y. Yuan, L. Jiao, K. Zhu, and L. Zhang, "Incentivizing federated learning under long-term energy constraint via online randomized auctions," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5129–5144, 2022.

[40] Y. Chen, Z. Liu, Y. Zhang, Y. Wu, X. Chen, and L. Zhao, "Deep reinforcement learning-based dynamic resource management for mobile edge computing in industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4925–4934, 2020.

[41] "Build efficient optimal control software, with minimal effort." https://web.casadi.org/.

[42] S. Mizuno and F. Jarre, "Global and polynomial-time convergence of an infeasible-interior-point algorithm using inexact computation." *Mathematical Programming*, vol. 84, no. 1, 1999.

[43] Q. Sun, C. Wu, Z. Li, and S. Ren, "Colocation demand response: Joint online mechanisms for individual utility and social welfare maximization," *IEEE Journal on Selected Areas in Communications*, vol. PP, no. 12, pp. 1–1, 2016.

[44] E. C. Hall and R. M. Willett, "Online convex optimization in dynamic environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 647–662, 2015.

[45] Y. Jin, L. Jiao, Z. Qian, S. Zhang, S. Lu, and X. Wang, "Resource-efficient and convergence-preserving online participant selection in federated learning," in *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2020, pp. 606–616.

[46] X. Gao, M. Sitharam, and A. E. Roitberg, "Bounds on the jensen gap, and implications for mean-concentrated distributions," *arXiv preprint arXiv:1712.05267*, 2017.

[47] P. Lai, "Eua dataset," https://github.com/PhuLai/eua-dataset, 2020, accessed: Jan. 2026.

[48] Y. Yuan, L. Jiao, K. Zhu, and L. Zhang, "Incentivizing federated learning under long-term energy constraint via online randomized auctions," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5129–5144, 2021.

[49] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2020.

[50] J. Wilkes, "More Google cluster data," Google research blog, Mountain View, CA, USA, Nov. 2011, posted at http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html.

[51] M. Kim, W. Saad, M. Mozaffari, and M. Debbah, "On the tradeoff between energy, precision, and accuracy in federated quantized neural networks," in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 2194–2199.

[52] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác, "Sgd and hogwild! convergence without the bounded gradients assumption," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3750–3758.

[53] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com

Konglin Zhu received the master's degree in computer Science from the University of California, Los Angeles, CA, USA, and the Ph.D. degree from the University of Göttingen, Germany, in 2009 and 2014, respectively. He is now an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China. His research interests include Internet of Vehicles, Edge Computing and Distributed Learning.

Siyuan Wei received the B.S. degree in Electronic Information Science and Technology from Beijing University of Posts and Telecommunications, China, in 2022. He is currently a Ph.D. student in School of Artificial Intelligence in Beijing University of Posts and Telecommunications. His research interests are in the areas of edge computing, blockchain, and trustworthy AI.

Xuan'er Wu received the B.S. and M.S. degrees in Information and Communication Engineering from Beijing University of Posts and Telecommunications, Beijing, China. She is currently with TeleAI, where her research interests include AI infrastructure, reinforcement learning systems, and edge computing.

Lei Jiao received the Ph.D. degree in computer science from the University of Göttingen, Germany and is currently a faculty member at the University of Oregon, USA. He researches machine learning systems, cloud computing, edge computing, network optimization, and network economics. He has published 90 papers mainly in leading journals such as IEEE Journal on Selected Areas in Communications, IEEE Transactions on Networking, IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems, and IEEE Transactions on Services Computing and conferences such as INFOCOM, MOBIHOC, ICDCS, SECON, and ICNP. He is a U.S. National Science Foundation CAREER awardee, and is also a recipient of the Ripple Faculty Fellowship, the Alcatel-Lucent Bell Labs UK and Ireland Recognition Award, and the Best Paper Awards of IEEE CNS 2019 and IEEE LANMAN 2013. He has been on the program committees as a track chair for ICDCS and as a member for INFOCOM, MOBIHOC, ICDCS, and IWQoS, and has also served as the program chair of multiple workshops with INFOCOM and ICDCS.

Jin Dong is currently the general director with the Beijing Academy of Blockchain and Edge Computing. He was the director with Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing. His research interests include blockchain, artificial intelligence, and low-power chip design.

Lin Zhang received the B.S. and the Ph.D. degrees in 1996 and 2001, both from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently a Professor with the same university. His current research interests include mobile cloud computing and Internet of Things.

APPENDIX

A: Proof of Lemma 2

Lemma 2. *We have* $\sum_{t\in\mathcal{T}} f^t\left(E\left[\overline{\mathbf{X}}^t\right]\right) = \sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right)$ *and* $\| \sum_{t\in\mathcal{T}} g^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)\| = \| \sum_{t\in\mathcal{T}} g^t\left(\widetilde{\mathbf{X}}^t\right)\|$.

Proof. We first derive the relationship between $\sum_{t\in\mathcal{T}} f^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)$ and $\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right)$ as

$$\begin{aligned}
&\sum_{t\in\mathcal{T}} f^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)\\
=&r^t\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} E^t_{ij}E\left[\overline{x}^t_i\right]\\
+&\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i} 2\left(v^t_{ijk}+q^t_{jk}\right)E\left[\overline{x}^t_i\right]\alpha_{ik}\\
=&r^t\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} E^t_{ij}\widetilde{x}^t_i\\
+&\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i} 2\left(v^t_{ijk}+q^t_{jk}\right)\widetilde{x}^t_i\alpha_{ik}\\
=&\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right)
\end{aligned}\tag{18}$$

The relationship between $\| \sum_{t\in\mathcal{T}} g^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)\|$ and $\| \sum_{t\in\mathcal{T}} g^t\left(\widetilde{\mathbf{X}}^t\right)\|$ can be describe as

$$\begin{aligned}
\| \sum_{t\in\mathcal{T}} g^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)\| &= \| \sum_{i\in\mathcal{I}}\left(Q_i - \sum_{t\in\mathcal{T}} E\left[\overline{x}^t_i\right]\right)\|\\
&= \| \sum_{i\in\mathcal{I}}\left(Q_i - \sum_{t\in\mathcal{T}} \widetilde{x}^t_i\right)\| = \| \sum_{t\in\mathcal{T}} g^t\left(\widetilde{\mathbf{X}}^t\right)\|.
\end{aligned}\tag{19}$$

Then we have $\| \sum_{t\in\mathcal{T}} g^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)\| = \| \sum_{t\in\mathcal{T}} g^t\left(\widetilde{\mathbf{X}}^t\right)\|$ □

B: Proof of Theorem 1

We derive the relationship between $\mathrm{Reg}_{\mathcal{T}}$ and $\widetilde{\mathrm{Reg}_{\mathcal{T}}}$ as

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{T}} &= E\left[\sum_{t\in\mathcal{T}} f^t\left(\overline{\mathbf{X}}^t\right)\right] + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\overline{\mathbf{W}}^{t*}\right)\\
&\overset{20(a)}{=} \sum_{t\in\mathcal{T}} f^t\left(E\left[\overline{\mathbf{X}}^t\right]\right) + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\overline{\mathbf{W}}^{t*}\right)\\
&+ \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
&\overset{20(b)}{\leq} \sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right) + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\overline{\mathbf{W}}^{t*}\right)\\
&+ \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
&\overset{20(c)}{\leq} \sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right) + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
&= \widetilde{\mathrm{Reg}_{\mathcal{T}}},
\end{aligned}\tag{20}$$

where 20(a) holds by the linearity of $f^t(\mathbf{X}^t)$, 20(b) holds due to Lemma 2. 20(c) holds due to $\sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right) \leqslant \sum_{t\in\mathcal{T}} f^t_o\left(\overline{\mathbf{W}}^{t*}\right)$.

The relationship between $\mathrm{Fit}_{\mathcal{T}}$ and $\widetilde{\mathrm{Fit}_{\mathcal{T}}}$ can be derived as

$$\begin{aligned}
\mathrm{Fit}_{\mathcal{T}} &= \left\| \left[E\left[\sum_{t\in\mathcal{T}} g^t_o\left(\overline{\mathbf{W}}^t\right)\right]\right]^+\right\|,\\
&= \left\| \left[E\left[\sum_{t\in\mathcal{T}} g^t\left(\overline{\mathbf{X}}^t\right)\right]\right]^+\right\| \overset{21(a)}{\leq} \left\| E\left[\sum_{t\in\mathcal{T}} g^t\left(\overline{\mathbf{X}}^t\right)\right]\right\|\\
&\overset{21(b)}{=} \left\| \sum_{t\in\mathcal{T}} g^t\left(E\left[\overline{\mathbf{X}}^t\right]\right)\right\| = \left\| \sum_{t\in\mathcal{T}} g^t\left(\widetilde{\mathbf{X}}^t\right)\right\| = \widetilde{\mathrm{Fit}_{\mathcal{T}}}.
\end{aligned}\tag{21}$$

C:Proof of Lemma 3

Lemma 3. *Under previous assumptions and the dual variable initialization of* $\boldsymbol{\lambda}^1 = 0$*, we have the following:*

$$\frac{\left(\|\boldsymbol{\lambda}^{t+1}\|^2 - \|\boldsymbol{\lambda}^t\|^2\right)}{2} \leq \mu\boldsymbol{\lambda}^{t\top} g^t\left(\widetilde{\mathbf{X}}^t\right) + \frac{\mu^2}{2}\|g^t\left(\widetilde{\mathbf{X}}^t\right)\|^2, \tag{22}$$

$$\forall t, \|\boldsymbol{\lambda}^t\| \leq \|\bar{\boldsymbol{\lambda}}\| = \frac{\frac{\mu U^2}{2} + 2FR + \frac{R^2}{2\alpha}}{\varepsilon - \bar{\mathcal{V}}(g)} + \mu U. \tag{23}$$

Proof. Updating $\boldsymbol{\lambda}$ by using the equation in (6), we have:

$$\begin{aligned}
\|\boldsymbol{\lambda}^{t+1}\|^2 &= \|[\boldsymbol{\lambda}^t + \mu g^t(\widetilde{\mathbf{X}}^t)]^+\|^2 \overset{24(a)}{\leq} \|\boldsymbol{\lambda}^t + \mu g^t(\widetilde{\mathbf{X}}^t)\|^2\\
&= \|\boldsymbol{\lambda}^t\|^2 + 2\mu\boldsymbol{\lambda}^{t\top} g^t(\widetilde{\mathbf{X}}^t) + \mu^2\|g^t(\widetilde{\mathbf{X}}^t)\|^2,
\end{aligned}\tag{24}$$

where inequality 24(a) holds because applying $[]^+$ on each dimension would only decrease the absolute value. Since $\widetilde{\mathbf{X}}^{t+1}$ is the optimum for objective in (7), by using the interior point $\hat{\mathbf{X}}^t$ mentioned in Assumption 2, we have

$$\begin{aligned}
&\nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) + \frac{\|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|^2}{2\alpha}\\
&\leq \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda}^{t+1\top} g^t(\hat{\mathbf{X}}^t) + \frac{\|\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t\|^2}{2\alpha}\\
&\overset{25(a)}{\leq} \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t) - \varepsilon\boldsymbol{\lambda}^{t+1\top}\mathbf{1} + \frac{\|\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t\|^2}{2\alpha}\\
&\overset{25(b)}{\leq} \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t) - \varepsilon\|\boldsymbol{\lambda}^{t+1}\| + \frac{\|\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t\|^2}{2\alpha},
\end{aligned}\tag{25}$$

where inequality 25(a) holds due to Assumption 2, and inequality holds because $\|\boldsymbol{\lambda}^{t+1}\|$ is less or equal to $\boldsymbol{\lambda}^{t+1\top}\mathbf{1}$. Then we have:

$$\begin{aligned}
&\boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) \leq \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t) - \varepsilon\|\boldsymbol{\lambda}^{t+1}\|\\
&- \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t) + \frac{\|\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t\|^2 - \|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|^2}{2\alpha}\\
&\leq \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t) - \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t)\\
&- \varepsilon\|\boldsymbol{\lambda}^{t+1}\| + \frac{R^2}{2\alpha}\\
&\overset{26(a)}{\leq} \nabla\|f^t(\widetilde{\mathbf{X}}^t)\|(\|\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t\| + \|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|) - \varepsilon\|\boldsymbol{\lambda}^{t+1}\| + \frac{R^2}{2\alpha}\\
&\leq 2FR - \varepsilon\|\boldsymbol{\lambda}^{t+1}\| + \frac{R^2}{2\alpha} \overset{def}{=} \Psi^{t+1},
\end{aligned}\tag{26}$$

where inequality 26(a) holds by using Cauchy-Schwartz twice on the first two terms. After plugging inequality (26) into (22), we have

$$\begin{aligned}
&\Delta\left(\boldsymbol{\lambda}^{t+1}\right) = \frac{\left(\|\boldsymbol{\lambda}^{t+1}\|^2 - \|\boldsymbol{\lambda}^t\|^2\right)}{2}\\
&\leq \mu\boldsymbol{\lambda}^{t+1\top} g^{t+1}(\widetilde{\mathbf{X}}^{t+1}) + \frac{\mu^2}{2}\|g^{t+1}(\widetilde{\mathbf{X}}^{t+1})\|^2\\
&\overset{27(a)}{\leq} \mu\boldsymbol{\lambda}^{t+1\top}\left(g^{t+1}\left(\widetilde{\mathbf{X}}^{t+1}\right) - g^t\left(\widetilde{\mathbf{X}}^{t+1}\right)\right) + \frac{\mu^2T^2}{2} + \mu\Psi^{t+1}\\
&\overset{27(b)}{\leq} \mu\boldsymbol{\lambda}^{t+1\top}\left[g^{t+1}\left(\widetilde{\mathbf{X}}^{t+1}\right) - g^t\left(\widetilde{\mathbf{X}}^{t+1}\right)\right]^+ + \frac{\mu^2T^2}{2} + \mu\Psi^{t+1}\\
&\overset{27(c)}{\leq} \mu\bar{V}(g)\|\boldsymbol{\lambda}^{t+1}\| + \frac{\mu^2T^2}{2} + \mu(2FR - \varepsilon\|\boldsymbol{\lambda}^{t+1}\| + \frac{R^2}{2\alpha}),
\end{aligned}\tag{27}$$

where inequality 27(a) holds by adding two complementary terms to the right sides, i.e., $\pm\mu g^t\left(\widetilde{\mathbf{X}}^{t+1}\right)$, and using the upper bound of g^t; inequality 27(b) holds due to the non-negative property of $\boldsymbol{\lambda}^{t+1}$; inequality 27(c) holds due to Assumption 3.

Next, we show the correctness of inequality (23) by contradiction. Without loss of generation, we suppose that $t+2$ is the first time index that breaks inequality (23), namely:

$$\|\boldsymbol{\lambda}^{t+1}\| \leq \|\bar{\boldsymbol{\lambda}}\| \leq \|\boldsymbol{\lambda}^{t+2}\|. \tag{28}$$

However, by using the equation in (6), the relationship can be obtained on $\boldsymbol{\lambda}$ between consecutive epochs as follows:

$$\begin{aligned}\|\boldsymbol{\lambda}^{t+1}\| &\overset{29(a)}{\geq} \|\boldsymbol{\lambda}^{t+2}\| - \|\boldsymbol{\lambda}^{t+2} - \boldsymbol{\lambda}^{t+1}\| \\ &= \|\boldsymbol{\lambda}^{t+2}\| - \|[\boldsymbol{\lambda}^{t+1} + \mu g^{t+1}(\widetilde{\boldsymbol{X}}^{t+1})]^+ - \boldsymbol{\lambda}^{t+1}\| \\ &\overset{29(b)}{\geq} \|\boldsymbol{\lambda}^{t+2}\| - \|\boldsymbol{\lambda}^{t+1} + \mu g^{t+1}(\widetilde{\boldsymbol{X}}^{t+1}) - \boldsymbol{\lambda}^{t+1}\| \\ &= \|\boldsymbol{\lambda}^{t+2}\| - \|\mu g^{t+1}(\widetilde{\boldsymbol{X}}^{t+1})\| \overset{29(c)}{>} \|\bar{\boldsymbol{\lambda}}\| - \mu U, \end{aligned} \tag{29}$$

where 29(a) holds due to the triangle inequality; 29(b) holds because of the non-expansive property of the projection; 29(c) holds by using the hypothesis from (28). Then we plug (29) into (27), we have $\Delta\left(\boldsymbol{\lambda}^{t+1}\right) < 0$, leading to $\|\boldsymbol{\lambda}^{t+2}\| < \|\boldsymbol{\lambda}^{t+1}\|$, which contradicts (28). So inequality (23) holds.

□

D:Proof of Theorem 2

As $\boldsymbol{\lambda}^t$ is updated by equation (6), we have

$$\left[\boldsymbol{\lambda}^T + \mu g^T\left(\widetilde{\boldsymbol{X}}^T\right)\right]^+ \geq \ldots \geq \boldsymbol{\lambda}^1 + \textstyle\sum_{t=1}^T \mu g^t\left(\widetilde{\boldsymbol{X}}^t\right). \tag{30}$$

As $\boldsymbol{\lambda}^1 = 0$, we have

$$\textstyle\sum_{t=1}^T \mu g^t\left(\widetilde{\boldsymbol{X}}^t\right) \leq \dfrac{\boldsymbol{\lambda}^{T+1} - \boldsymbol{\lambda}^1}{\mu} \leq \dfrac{\boldsymbol{\lambda}^{T+1}}{\mu}. \tag{31}$$

Thus,

$$\widetilde{Fit}_{\mathcal{T}} = \left\| \left[\sum_{t\in\mathcal{T}} g^t\left(\widetilde{\boldsymbol{X}}^t\right)\right]^+ \right\| \leq \left\| \frac{\boldsymbol{\lambda}^{T+1}}{\mu} \right\| \leq \left\| \frac{\bar{\boldsymbol{\lambda}}}{\mu} \right\|. \tag{32}$$

By plugging (32) into (21), the proof is completed.

E:Proof of Lemma 4

Lemma 4. *Under previous assumptions and the dual variable initialization of $\boldsymbol{\lambda}^1 = 0$, the upper bound of $\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right) - \sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)$ can be derived as:*

$$\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right) - \sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right) \leq \mathcal{A}_T, \tag{33}$$

where

$$\mathcal{A}_T = \frac{\mu U^2(T+1)}{2} + \frac{\alpha F^2 T}{2} + \|\bar{\boldsymbol{\lambda}}\|\mathcal{V}_{\mathbf{g^t}}^T + \frac{R\mathcal{V}_{\widetilde{\mathbf{X}}^{t^*}}^T}{\alpha} + \frac{R^2}{2\alpha}, \tag{34}$$

$$\mathcal{V}_{\mathbf{g^t}}^T = \sum_{t\in\mathcal{T}} \mathcal{V}_{\mathbf{g^t}}, \mathcal{V}_{\widetilde{\mathbf{X}}^{t^*}}^T = \sum_{t\in\mathcal{T}} \|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^{t-1^*}\|. \tag{35}$$

Proof. The objective in (7) is $\frac{1}{\alpha}$-strongly convex with respect to $\widetilde{\mathbf{X}}$, denoted by $J^t(\widetilde{\mathbf{X}})$, i.e., $\forall \mathbf{a}, \mathbf{b} \in \widetilde{\mathcal{X}}$:

$$J^t(\mathbf{b}) \geq J^t(\mathbf{a}) + \nabla J^t(\mathbf{a})^\top(\mathbf{b} - \mathbf{a}) + \frac{\|\mathbf{b} - \mathbf{a}\|^2}{2\alpha}. \tag{36}$$

Since $\widetilde{\mathbf{X}}^{t+1}$ is the optimal solution for $\min_{\widetilde{\mathbf{X}}\in\widetilde{\mathcal{X}}} J^t(\widetilde{\mathbf{X}})$, then we have the condition:

$$\nabla J^t(\widetilde{\mathbf{X}}^{t+1})^\top(\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^{t+1}) \geq 0. \tag{37}$$

Thus, by setting $\mathbf{a} = \widetilde{\mathbf{X}}^{t+1}$, $\mathbf{b} = \widetilde{\mathbf{X}}^{t^*}$, and plugging inequality (37) into inequality (36), we have

$$J^t(\widetilde{\mathbf{X}}^{t^*}) \geq J^t(\widetilde{\mathbf{X}}^{t+1}) + \frac{\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^{t+1}\|^2}{2\alpha}. \tag{38}$$

Then we add $f^t(\widetilde{\mathbf{X}}^t)$ on both sides, and expand $J^t(\cdot)$ according to its definition, then we have

$$\begin{aligned} &f^t(\widetilde{\mathbf{X}}^t) + \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda^{t+1}}^\top g^t(\widetilde{\mathbf{X}}^{t+1}) \\ &+ \frac{\|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|^2}{2\alpha} \\ &\overset{39(a)}{\leq} f^t(\widetilde{\mathbf{X}}^{t^*}) + \boldsymbol{\lambda^{t+1}}^\top g^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^t\|^2}{2\alpha} \\ &- \frac{\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^{t+1}\|^2}{2\alpha} \\ &\overset{39(b)}{\leq} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^t\|^2}{2\alpha} - \frac{\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^{t+1}\|^2}{2\alpha}, \end{aligned} \tag{39}$$

where inequality 39(a) holds because of the property of convex function that $f^t(\widetilde{\mathbf{X}}^{t^*}) \geq f^t(\widetilde{\mathbf{X}}^t) + \nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^t)$ and inequality 39(b) comes from that $\boldsymbol{\lambda^{t+1}} \leq \mathbf{0}$ because the optimal solution $\widetilde{\mathbf{X}}^{t^*}$ is feasible in every time frame, and $g^t(\widetilde{\mathbf{X}}^{t^*}) \leq \mathbf{0}$, so $\boldsymbol{\lambda^{t+1}} g^t(\widetilde{\mathbf{X}}^{t^*}) \leq 0$. Then we focus on the gradient term

$$\begin{aligned} &-\nabla f^t(\widetilde{\mathbf{X}}^t)^\top(\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t) \overset{40(a)}{\leq} \|\nabla f^t(\widetilde{\mathbf{X}}^t)\|\|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\| \\ &\overset{40(b)}{\leq} \frac{\|\nabla f^t(\widetilde{\mathbf{X}}^t)\|^2}{2\beth} + \frac{\beth}{2}\|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|^2 \overset{40(c)}{\leq} \frac{F^2}{2\beth} + \frac{\beth}{2}\|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|^2, \end{aligned} \tag{40}$$

where $\beth$ is an arbitrary positive constant. Inequality 40(a) holds because of the property of 2-norms; inequality 40(b) holds because of Cauchy inequality; and inequality 40(c) comes from the assumption of bounded gradient of $f^t(\cdot)$. Plugging (40) into (39), we have

$$\begin{aligned} &f^t(\widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda^{t+1}}^\top g^t(\widetilde{\mathbf{X}}^{t+1}) \leq f^t(\widetilde{\mathbf{X}}^{t^*}) + (\frac{\beth}{2} - \frac{1}{2\alpha})\|\widetilde{\mathbf{X}}^{t+1} - \widetilde{\mathbf{X}}^t\|^2 \\ &+ \frac{1}{2\alpha}(\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^t\|^2 - \|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^{t+1}\|^2) + \frac{F^2}{2\beth}, \\ &\overset{41(a)}{=} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right) + \frac{1}{2\alpha}(\|\widetilde{\mathbf{X}}^{t^*} - \widetilde{\mathbf{X}}^t\|^2 - \|\widetilde{\mathbf{X}}^{t*} - \widetilde{\mathbf{X}}^{t+1}\|^2) + \frac{\alpha F^2}{2}, \end{aligned} \tag{41}$$

where inequality 41(a) holds because $\beth$, i.e., $\beth = \frac{1}{\alpha}$, such that $(\frac{\beth}{2} - \frac{1}{2\alpha}) = 0$. By applying (41) into (22), we have

$$\begin{aligned} &\frac{\triangle(\boldsymbol{\lambda}^{t+1})}{\mu} + f^t(\widetilde{\mathbf{X}}^t) \overset{42(a)}{\leq} \boldsymbol{\lambda}^{t+1\top} g^{t+1}(\widetilde{\mathbf{X}}^{t+1}) + \frac{\mu}{2}\|g^{t+1}(\widetilde{\mathbf{X}}^{t+1})\|^2 \\ &+ f^t(\widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) - \boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) \\ &\overset{42(b)}{=} f^t(\widetilde{\mathbf{X}}^t) + \boldsymbol{\lambda}^{t+1\top} g^{t+1}(\widetilde{\mathbf{X}}^{t+1}) + \frac{\mu}{2}\|g^{t+1}(\widetilde{\mathbf{X}}^{t+1})\|^2 \\ &+ \boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) - \boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) \\ &\overset{42(c)}{\leq} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{1}{2\alpha}(\|\mathbf{X}^{t^*} - \mathbf{X}^t\|^2 - \|\mathbf{X}^{t^*} - \mathbf{X}^{t+1}\|^2) + \frac{\alpha F^2}{2} \\ &+ \frac{\mu}{2}\|g^{t+1}(\widetilde{\mathbf{X}}^{t+1})\|^2 + \boldsymbol{\lambda}^{t+1\top} g^t(\widetilde{\mathbf{X}}^{t+1}) - g^t(\widetilde{\mathbf{X}}^{t+1}) \\ &\overset{42(d)}{\leq} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{1}{2\alpha}(\|\mathbf{X}^{t^*} - \mathbf{X}^t\|^2 - \|\mathbf{X}^{t^*} - \mathbf{X}^{t+1}\|^2) + \frac{\alpha F^2}{2} \\ &+ \frac{\mu U^2}{2} + \boldsymbol{\lambda}^{t+1\top} g^t\left(\widetilde{\mathbf{X}}^{t+1}\right) - g^t(\widetilde{\mathbf{X}}^{t+1}) \\ &\overset{42(e)}{\leq} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{1}{2\alpha}(\|\mathbf{X}^{t^*} - \mathbf{X}^t\|^2 - \|\mathbf{X}^{t^*} - \mathbf{X}^{t+1}\|^2) + \frac{\alpha F^2}{2} \\ &+ \frac{\mu U^2}{2} + \|\boldsymbol{\lambda}^{t+1}\|\mathcal{V}_{\mathbf{g^t}}, \end{aligned} \tag{42}$$

where inequality 42(a) holds because we add the term $f^t(\widetilde{\mathbf{X}}^t)$ on both two sides; equation 42(b) holds because we re-arrange the terms; inequality 42(c) holds due to the application of equality (42); inequality 42(d) holds due to the bounded value of g^{t+1} mentioned in the previous assumption. Then we consider the intermediate terms as follows:

$$\begin{aligned} &\|\mathbf{X}^{t*} - \mathbf{X}^t\|^2 \\ &\overset{43(a)}{=} \|\mathbf{X}^{t^*} - \mathbf{X}^t\|^2 - \|\mathbf{X}^t - \mathbf{X}^{(t-1)^*}\|^2 + \|\mathbf{X}^t - \mathbf{X}^{(t-1)^*}\|^2 \\ &\overset{43(b)}{=} \|\mathbf{X}^{t^*} - \mathbf{X}^{(t-1)^*}\|^2\|\mathbf{X}^{t^*} - 2\mathbf{X}^t + \mathbf{X}^{(t-1)^*}\| + \|\mathbf{X}^t - \mathbf{X}^{(t-1)*}\|^2 \\ &\overset{43(c)}{\leq} 2R\|\mathbf{X}^{t^*} - \mathbf{X}^{(t-1)^*}\| + \|\mathbf{X}^t - \mathbf{X}^{(t-1)^*}\|^2, \end{aligned} \tag{43}$$

where equation 43(a) holds because we add two complementary terms; equation 43(b) holds because we apply difference of two squares on the first two terms; equation

43(c) holds due to triangle inequality for the bounded radius on domain. Applying inequality (43) to (42), we have

$$\begin{aligned}
&\frac{\triangle(\boldsymbol{\lambda}^{t+1})}{\mu} + f^t(\widetilde{\mathbf{X}}^t) \\
&\leq f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{\mu U^2}{2} + \frac{\alpha F^2}{2} + \|\boldsymbol{\lambda}^{t+1}\|\mathcal{V}_{\mathbf{g}^t} \\
&+ \frac{1}{2\alpha}(2R\|\mathbf{X}^{t^*} - \mathbf{X}^{t-1^*}\| + \|\mathbf{X}^t - \mathbf{X}^{t-1^*}\|^2 - \|\mathbf{X}^{t^*} - \mathbf{X}^{t+1}\|^2). \quad (44)
\end{aligned}$$

Summing up previous inequality over $t = 1$ to T, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(\frac{\triangle(\boldsymbol{\lambda}^{t+1})}{\mu}) + \sum_{t=1}^{T} f^t(\widetilde{\mathbf{X}}^t) \\
&\leq \sum_{t=1}^{T} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{\mu U^2 T}{2} + \frac{\alpha F^2 T}{2} + \|\boldsymbol{\lambda}^{t+1}\|\mathcal{V}^T_{\mathbf{g}^t} \\
&+ \frac{1}{2\alpha}\sum_{t=1}^{T}(\|\mathbf{X}^t - \mathbf{X}^{t-1^*}\|^2 - \|\mathbf{X}^{t^*} - \mathbf{X}^{t+1}\|^2) + \frac{R\mathcal{V}^T_{\widetilde{\mathbf{X}}^{t*}}}{\alpha} \\
&\overset{45(a)}{\leq} \sum_{t=1}^{T} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{\mu U^2 T}{2} + \frac{\alpha F^2 T}{2} + \|\overline{\boldsymbol{\lambda}}\|\mathcal{V}^T_{\mathbf{g}^t} \\
&+ \frac{1}{2\alpha}(\|\mathbf{X}^1 - \mathbf{X}^{0^*}\|^2 - \|\mathbf{X}^{T^*} - \mathbf{X}^{T+1}\|^2) + \frac{R\mathcal{V}^T_{\widetilde{\mathbf{X}}^{t*}}}{\alpha} \\
&\leq \sum_{t=1}^{T} f^t(\widetilde{\mathbf{X}}^{t^*}) + \frac{\mu U^2 T}{2} + \frac{\alpha F^2 T}{2} + \|\overline{\boldsymbol{\lambda}}\|\mathcal{V}^T_{\mathbf{g}^t} \\
&+ \frac{1}{2\alpha}\|\mathbf{X}^1 - \mathbf{X}^{0^*}\|^2 + \frac{R\mathcal{V}^T_{\widetilde{\mathbf{X}}^{t*}}}{\alpha}, \quad (45)
\end{aligned}$$

where inequality 45(a) holds due to the definition of $\|\overline{\boldsymbol{\lambda}}\|$. Then,

$$\begin{aligned}
&f^t(\widetilde{\mathbf{X}}^t) - f^t(\widetilde{\mathbf{X}}^{t^*}) \leq \frac{\mu U^2 T}{2} + \frac{\alpha F^2 T}{2} + \|\overline{\boldsymbol{\lambda}}\|\mathcal{V}^T_{\mathbf{g}^t} \\
&+ \frac{R\mathcal{V}^T_{\widetilde{\mathbf{X}}^{t*}}}{\alpha} + \frac{1}{2\alpha}(\|\mathbf{X}^1 - \mathbf{X}^{0^*}\|^2 - \sum_{t=1}^{T}(\frac{\triangle(\boldsymbol{\lambda}^{t+1})}{\mu}) \\
&= \frac{\mu U^2 T}{2} + \frac{\alpha F^2 T}{2} + \|\overline{\boldsymbol{\lambda}}\|\mathcal{V}^T_{\mathbf{g}^t} + \frac{R\mathcal{V}^T_{\widetilde{\mathbf{X}}^{t*}}}{\alpha} \\
&+ \frac{1}{2\alpha}(\|\mathbf{X}^1 - \mathbf{X}^{0^*}\|^2) - \frac{\|\boldsymbol{\lambda}^{T+2}\|^2}{2\mu} + \frac{\|\boldsymbol{\lambda}^2\|^2}{2\mu}) \overset{46(a)}{\leq} \mathcal{A}_{\mathcal{T}}, \quad (46)
\end{aligned}$$

where inequality 46(a) holds because $\|\widetilde{\mathbf{X}}^1 - \widetilde{\mathbf{X}}^{0^*}\|^2$ has been bounded by R, and $\|\boldsymbol{\lambda}^{T+2}\|^2 \geq 0$, as well as $\|\boldsymbol{\lambda}^2\|^2 \leq \mu^2 F^2$ if $\boldsymbol{\lambda}_1 = 0$. □

F:Proof of Lemma 5

Lemma 5. *Under all the constraints, we have*

$$\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right) + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right) \leq \mathcal{L}, \quad (47)$$

where

$$\begin{aligned}
\mathcal{L} = &\max_{i,j,k,t}\left\{\frac{\left(\nu^t_{ijk} + q^t_{jk} + \nu^t_{ikj} + q^t_{kj}\right)\alpha_{ik}}{r^t E^t_{ij}}\right\} J^2 \cdot \max_t\left\{r^t R^t\right\} \cdot T \\
&+ \max_t\left(r^t R^t\right)\cdot T + \frac{\max_{k,n}\left(d_{kn}C_k\right)}{\min_n S_n}\cdot TDJ \\
&+ \frac{\max_{j,k,n,\tau,t}\left(\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)C_k\right)}{\min_n S_n}\cdot TDJ \\
&- 2TJ^2 \min_{i,j,k,t}\left\{\left(\nu^t_{ijk} + q^t_{jk}\right)\alpha_{ik}\right\} - IJT\cdot\min_{i,j,t}\left(r^t E^t_{ij} Q_i\right) \\
&- \min_{j,k,n,\tau,t}\left(\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)w^t_{jn\tau}\right)\cdot TDNJ \\
&- \min_{j,k,n,\tau,t}\left(d_{kn}w^t_{jn\tau}\right)\cdot TDNJ \quad (48)
\end{aligned}$$

Proof. We denote the optimum fractional solution of the objective function $f^t(\widetilde{\mathbf{X}}^{t^*})$ for problem $\mathbb{P}_1$ as $\left(\widetilde{x}^{t^*}_i\right)_1$ and the optimum fractional solution of $f^t_\tau(\mathbf{Y}^t_\tau)$ for $\mathbb{P}^t_2$ as $\left(\widetilde{y}^t_{jkn\tau}\right)_2$. The optimum solution for the problem $\mathbb{P}$ is $\widetilde{x}^{t^*}_i$ and $\widetilde{y}^t_{jkn\tau^*}$. We substitute these parameters into equation (47), then we have:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right) + \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right) - \sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right) \\
&= \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} r^t E^t_{ij}\left(\widetilde{x}^{t^*}_i\right)_1 + \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}\left(\nu^t_{ijk} + q^t_{jk} + \nu^t_{ikj} + q^t_{kj}\right)\left(\widetilde{x}^{t^*}_i\right)_1\alpha_{ik} \\
&+ \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}}\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)\left(\widetilde{y}^t_{jkn\tau}\right)_2 \\
&+ \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn}\left(\widetilde{y}^t_{jkn\tau}\right)_2 \\
&- \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} r^t E^t_{ij}\widetilde{x}^{t^*}_i - \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}\left(\nu^t_{ijk} + q^t_{jk} + \nu^t_{ikj} + q^t_{kj}\right)\widetilde{x}^{t^*}_i\alpha_{ik} \\
&- \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}}\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)\widetilde{y}^t_{jkn\tau*} \\
&- \sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn}\widetilde{y}^t_{jkn\tau_*} \quad (49)
\end{aligned}$$

According to the constraint (2b), we have:

$$\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} r^t E^t_{ij}\left(\widetilde{x}^{t^*}_i\right)_1 \leq \max_t\left(r^t R^t\right)\cdot T. \quad (50)$$

According to the constraint (2a), we have:

$$\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} r^t E^t_{ij}\widetilde{x}^{t^*}_i \geq \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i} r^t E^t_{ij} Q_i \geq IJT\cdot\min_{i,j,t}\left(r^t E^t_{ij} Q_i\right). \quad (51)$$

According to the constraint (2b), we have:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i} 2\left(\nu^t_{ijk} + q^t_{jk}\right)\left(\widetilde{x}^{t^*}_i\right)_1\alpha_{ik} \\
&= \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}\frac{2\left(\nu^t_{ijk} + q^t_{jk}\right)\alpha_{ik}}{r^t E^t_{ij}} r^t E^t_{ij}\left(\widetilde{x}^{t^*}_i\right)_1 \\
&\leq \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}\max_{i,j,k,t}\left\{\frac{2\left(\nu^t_{ijk} + q^t_{jk}\right)\alpha_{ik}}{r^t E^t_{ij}}\right\} r^t E^t_{ij}\left(\widetilde{x}^{t^*}_i\right)_1 \\
&\leq \max_{i,j,k,t}\left\{\frac{2\left(\nu^t_{ijk} + q^t_{jk}\right)\alpha_{ik}}{r^t E^t_{ij}}\right\}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i} r^t E^t_{ij}\left(\widetilde{x}^{t^*}_i\right)_1 \\
&\leq \max_{i,j,k,t}\left\{\frac{2\left(\nu^t_{ijk} + q^t_{jk}\right)\alpha_{ik}}{r^t E^t_{ij}}\right\}\cdot J^2\cdot\max_t\left\{r^t R^t\right\}\cdot T. \quad (52)
\end{aligned}$$

According to the constraint (2a), we have:

$$\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i} 2\left(\nu^t_{ijk} + q^t_{jk}\right)\widetilde{x}^{t^*}_i\alpha_{ik} \geq 2TJ^2\min_{i,j,k,t}\left\{\left(\nu^t_{ijk} + q^t_{jk}\right)\alpha_{ik}Q_i\right\}. \quad (53)$$

According to the constraints (3d), we have:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{j\in\mathcal{J}}\sum_{n\in\mathcal{N}}\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)\left(\widetilde{y}^t_{jkn\tau}\right)_2 \\
&\leq \frac{\max_{j,k,n,\tau,t}\left(\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)C_k\right)}{\min_n S_n}\cdot TDJ. \quad (54)
\end{aligned}$$

$$\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn}\left(\widetilde{y}^t_{jkn\tau}\right)_2 \leq \frac{\max_{k,n}\left(d_{kn}C_k\right)}{\min_n S_n}\cdot TDJ. \quad (55)$$

According to the constraint (3c), we have:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}}\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)\widetilde{y}^t_{jkn\tau*} \\
&\geq \min_{j,k,n,\tau,t}\left(q^t_{jk\tau} + q^t_{kj\tau} + b^t_{nk} + \omega^t_{jkn\tau}\right)\cdot TDNJ. \quad (56)
\end{aligned}$$

$$\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{J}}\sum_{n\in\mathcal{N}} d_{kn}\widetilde{y}^t_{jkn\tau *} \geq \min_{j,k,n,\tau,t}\left(d_{kn}w^t_{jn\tau}\right)\cdot TDNJ. \quad (57)$$

Now we plug the inequality (50)∼(56) into (49), then we have:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)+\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)-\sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
\leq&\max_{i,j,k,t}\left\{\frac{\left(\nu^t_{ijk}+q^t_{jk}+\nu^t_{ikj}+q^t_{kj}\right)\alpha_{ik}}{r^tE^t_{ij}}\right\}J^2\cdot\max_t\left\{r^tR^t\right\}\cdot T\\
&+\max_t\left(r^tR^t\right)\cdot T+\frac{\max\limits_{k,n}\left(d_{kn}C_k\right)}{\min\limits_n S_n}\cdot TDJ\\
&+\frac{\max\limits_{j,k,n,\tau,t}\left(\left(q^t_{jk\tau}+q^t_{kj\tau}+b^t_{nk}+\omega^t_{jkn\tau}\right)C_k\right)}{\min\limits_n S_n}\cdot TDJ\\
&-2TJ^2\min_{i,j,k,t}\left\{\left(\nu^t_{ijk}+q^t_{jk}\right)\alpha_{ik}\right\}-IJT\cdot\min_{i,j,t}\left(r^tE^t_{ij}Q_i\right)\\
&-\min_{j,k,n,\tau,t}\left(\left(q^t_{jk\tau}+q^t_{kj\tau}+b^t_{nk}+\omega^t_{jkn\tau}\right)w^t_{jn\tau}\right)\cdot TDNJ\\
&-\min_{j,k,n,\tau,t}\left(d_{kn}w^t_{jn\tau}\right)\cdot TDNJ. \quad (58)
\end{aligned}$$

□

G: Proof of Theorem 3

$\widetilde{Reg}_{\mathcal{T}}$ can be treated as

$$\begin{aligned}
\widetilde{Reg}_{\mathcal{T}}&=\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right)+\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)-\sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
&+\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)+\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)-\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)\\
&-\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)\\
&=\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right)-\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)+\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)\\
&-\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)+\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)+\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)\\
&-\sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
&=\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^t\right)-\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)+\sum_{t\in\mathcal{T}} f^t\left(\widetilde{\mathbf{X}}^{t^*}\right)\\
&+\sum_{t\in\mathcal{T}}\sum_{\tau\in\mathcal{D}} f^t_\tau\left(\mathbf{Y}^t_\tau\right)-\sum_{t\in\mathcal{T}} f^t_o\left(\widetilde{\mathbf{W}}^{t*}\right)\\
&\overset{59(a)}{\leq}\mathcal{A}_T+\mathcal{L}. \quad (59)
\end{aligned}$$

inequation 59(a) holds due to (33) and (47).