# Multi-Tenant Edge AI as a Virtual Power Plant via Online Auction-Driven Energy Scheduling

Hengdi Wang, Lei Jiao, Konglin Zhu, Lin Zhang, Jin Dong

***Abstract*—We innovatively propose that the modern edge computing infrastructure equipped with battery energy storage can actively operate as a Virtual Power Plant (VPP) to support grid stability through real-time energy dispatch. However, the edge infrastructure does not directly control the workload of the edge AI services that rent its resources, making it challenging to align energy flexibility with grid needs. To bridge this gap, we introduce an auction-driven online framework that incentivizes edge AI services to dynamically adapt their inference models in exchange for payment, thereby freeing battery capacity for grid support. Our approach tackles a novel NP-hard long-term social-cost minimization problem that jointly optimizes energy scheduling, battery degradation, inference accuracy, and model-switching overhead over time. Addressing the fundamental challenges of battery state transitions, complex switching cost across auctions, and economic guarantees for each single auction, we design a polynomial-time algorithmic framework that integrates switching-cost-aware control, online learning, randomized rounding, and a truthful payment mechanism. Theoretically, we rigorously prove that our online algorithms achieve an asymptotic competitive ratio, sub-linear regret and fit, and truthfulness and individual rationality. Extensive practical experiments using real-world traces demonstrate that our approach not only achieves a competitive ratio of 1.63 against the offline optimal solution while yielding verifiable sub-linear regret and fit curve, but also reduces social cost substantially compared to heuristics and state-of-the-arts and scales efficiently under dynamic workloads.**

***Index Terms*—Edge Computing, Edge AI, Virtual Power Plant, Battery Management, Online Algorithm, Auction.**

## I. Introduction

Contemporary 5G deployments, particularly those integrating Multi-access Edge Computing (MEC) servers, are increasingly equipped with energy storage systems (e.g., lithium-ion batteries) that function as managed grid assets [1], [2]. This represents a significant departure from traditional cellular base stations, which have long relied on passive (e.g., lead-acid) battery backups reserved solely for power grid failures. Modern systems enable dynamic energy arbitrage, provide ancillary grid services, and can participate in Virtual Power Plants (VPPs) by offering stored energy back to the power grid to enhance stability [3], [4]. This evolution presents a promising yet underexplored opportunity to co-optimize energy storage dispatch with edge computing workload management, improving both economic efficiency and grid support.

This work was supported in part by the Science and Technology Projects of Xizang Autonomous Region (Grant No. XZ202401ZY0027), the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, the Beijing Natural Science Foundation (2026FTZD001 and 4222033), the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance, and the U.S. National Science Foundation (CNS-2047719 and CNS-2225949).

(Corresponding authors: Lei Jiao, Konglin Zhu.)

Hengdi Wang and Lin Zhang are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: {wanghengdi, zhanglin}@bupt.edu.cn).

Lei Jiao is with the Center for Cyber Security and Privacy, University of Oregon, Eugene, OR 97403, USA (e-mail: ljiao2@uoregon.edu).

Konglin Zhu is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Purple Mountain Laboratories, Nanjing 210023, China (e-mail: klzhu@bupt.edu.cn).

Jin Dong is with the Beijing Academy of Blockchain and Edge Computing, Beijing 100131, China (e-mail: dongjin@baec.org.cn).

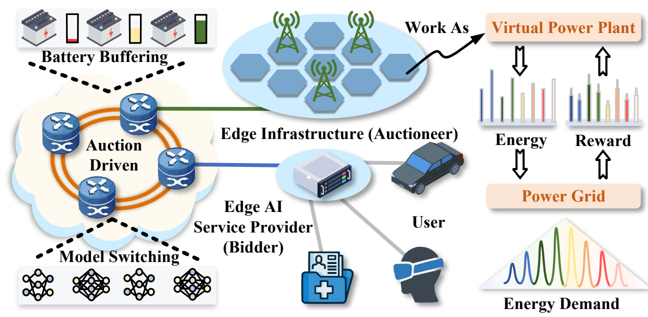



Fig. 1. System scenario

**Research Gap:** A key obstacle, however, stems from the separation of control between the *edge infrastructure provider* and the services deployed atop it [5]. Infrastructure providers own and maintain the physical assets (including servers and batteries) and typically offer computational resources on a pay-as-you-go basis. These resources are leased by independent *service providers*, who deploy and manage their own applications such as latency-sensitive *edge AI inference* services, with little regard for the underlying energy footprint. Consequently, while the edge infrastructure provider can modulate battery usage, it lacks direct influence over the workload that largely determines energy consumption, leading to potential inefficiencies in both energy storage management and grid integration.

To bridge this gap, we advocate an *auction* approach to enable the edge infrastructure provider to procure energy-responsive model adaptions from edge AI services, so that freed battery capacity can be more effectively used for grid support via VPPs, as shown in Fig. 1. However, designing such an auction framework confronts fundamental challenges.

**Challenge 1:** This is a sequence of auctions intertwined with the edge workload and battery. In each time slot, the auction outcome must be judiciously determined to shape the workload, while sufficient energy must be drawn from the grid and/or the battery to support that workload. Simultaneously, the battery may be discharged as part of a VPP to meet grid demand or charged for future use. This requires balancing trade-offs between energy cost (favoring charging at low price and discharging at high price) and battery degradation (favoring fewer charging/discharging cycles) [6], [7]. Moreover, battery dynamics introduces intertemporal coupling: the energy stored in the current time slot, plus any charging, constrains the maximum discharge in the next time slot. As future grid prices and demands are unknown [8], [9], deciding the charging amount in the current slot is non-trivial; prematurely discharging may force the edges to recharge from the grid at high prices later.

**Challenge 2:** Every auction is coupled with adjacent auctions via *switching cost*, preventing isolated optimization [10]. Each edge AI service operates a *default* model on its allocated edge servers—a model that may vary over time. The edge AI service may participate in an auction by submitting bids that include *proposed* models with reduced energy consumption and potentially lower inference accuracy. Deploying a new model incurs switching cost, i.e., deployment time. Depending on adjacent auctions, an edge AI service may replace a default model with a proposed one, revert to a default model, or switch between two proposed models. This results in a switching cost that depends on model identity and transition type, differing from most prior work [11], [12], [13], [14], [15], [16]. With unknown future outcomes, minimizing cumulative switching costs in the current auction is also challenging.

**Challenge 3:** Each auction must satisfy economic guarantees of *truthfulness* (ensuring no bid gains by misreporting its cost) and *individual rationality* (ensuring no bid incurs a net loss). While the classic Vickrey-Clarke-Groves (VCG) mechanism [17], [18] achieves these properties, it requires computing the exact allocation that minimizes the *social cost*, i.e., the total cost of the entire system. In our setting, this problem turns out NP-hard, rendering exact optimization intractable. Further, the social cost in our case is not merely a function of the selected bids, but a hybrid function that incorporates edge AI inference control and energy scheduling. This violates the separable allocation-centric cost structure for which VCG or its fractional variant [19], [20] is designed and proven truthful, necessitating a novel auction design.

To address these challenges, we conduct a rigorous modeling and algorithmic study on auction-driven model switching and battery management for VPP-enabled edge computing systems. We have the following four contributions:

**Contribution 1:** We model a series of consecutive auctions and formulate a corresponding long-term social cost minimization problem, capturing both the cost of the edge infrastructure (i.e., auctioneer) and the cost of the edge AI services (i.e., bidders). Our formulation integrates decision factors including winning-bid selection, grid energy usage, and battery charging and discharging orchestration to jointly optimize bidding cost, AI inference accuracy, model switching cost, energy cost, and battery degradation, while enabling the edge infrastructure to supply energy to the grid. Our problem makes almost no assumption on input heterogeneity and dynamism, and is a nonlinear mixed-integer program, provably NP-hard.

**Contribution 2:** We design a comprehensive Auction-Driven Energy Scheduling (ADES) algorithmic framework, where four polynomial-time algorithms operate together *in an online manner*. The key is a *switching-cost-aware control* algorithm that addresses our unique switching-cost structure by dynamically balancing the switching and non-switching cost on the fly without relying on future inputs. To address energy management, we convert battery state transitions into a long-term constraint for cumulative satisfaction. This way, we apply an *online learning* algorithm to pursue fractional decisions and transform them into integral decisions by a *randomized rounding* algorithm via an iterative rounding process. Meanwhile, we design a *payment allocation* algorithm for auctions based on our randomization procedure, leveraging bid winning probabilities and projected alternative bidding prices.

**Contribution 3:** We perform rigorous theoretical analysis for our algorithms. We prove that, under a positive per-slot basic cost, our approach leads to a constant *asymptotic competitive ratio*. Specifically, for sufficiently long time horizons, the long-term social cost incurred by our online algorithm is upper bounded by a constant times the offline optimum. We also prove the sub-linear *regret* and *fit* regarding online learning, i.e., the time-averaged difference between the cost incurred by our approach and that incurred by the series of per-time-slot optimums vanishes as time goes, and the long-run average violation of the long-term constraint also vanishes. Further, we prove that our payments in auctions attain truthfulness and individual rationality by satisfying the necessary and sufficient conditions required for the randomized auctions.

**Contribution 4:** We finally conduct extensive evaluations using multiple real-world traces, including edge infrastructure settings [21], edge AI service distributions [10], [22], battery configurations [1], [23], [24], AI models [25], [26], [27], [28], [29], user-generated workloads [7], [30], and VPP instances [31]. We briefly summarize our results: (i) Our approach reduces social cost by at least 31% cumulatively compared to state-of-the-arts, while leading to a competitive ratio of approximately 1.63; (ii) Our approach scales well with inputs and consistently outperforms others; (iii) Our online learning component achieves sub-linear growth in regret and fit; (iv) Our auction achieves truthfulness and individual rationality.

## II. Related Work

We review existing research extensively in multiple categories and highlight insufficiencies compared to our work.

**Virtual Power Plants:** Y. Qin *et al.* [3] focus on the integration of VPP and vehicle-to-grid techniques. C. Liu *et al.* [4] optimize the scheduling for VPPs in urban environments. H. Liang *et al.* [32] introduce a data-driven enclosing polyhedron method for energy scheduling of VPP in industrial systems. Z. Dai *et al.* [8] study a novel bi-level optimization approach to enable the coordination of distributed energy resources in VPPs. T. Morstyn *et al.* [33] integrate VPPs with peer-to-peer energy trades to coordinate distributed prosumers. L. Yan *et al.* [34] pioneer an adaptive orchestration framework that turns renewable uncertainty into market profit for VPPs.

Related research studies how to integrate distributed energy for the stable operation of VPPs, but none of them has considered edge infrastructures as a novel VPP. The unique features of edge infrastructures, including the online arrival of service requests and the trade-off between service quality and energy consumption, make the problem more intractable.

**Edge Battery Scheduling:** D. Zheng *et al.* [1] propose a deep reinforcement learning approach in coping with the dynamic power demand at edge-cloud. Y. Luo *et al.* [2] propose an energy harvesting model to enhance computational performance of edges in solar energy environment. Z. Meng *et al.* [6] design a learning-driven method in resource-constrained edge computing. Q. Li *et al.* [7] focus on pattern-aware online energy optimization to extend the lifespan of edge batteries. X. Deng *et al.* [35] study stochastic task in wireless powered mobile edge computing systems. M. Abner *et al.* [36] develop a multi-strategy power management framework to extend battery life in wireless sensor edge network.

These research studies edge battery scheduling, but only uses batteries as backup energy sources while energy shortage occurs. Furthermore, they neglect the fact that battery systems can jointly serve as a VPP to maintain the stability of the power grid, thus overlooking a significant opportunity to enhance grid reliability and provide valuable ancillary services.

**Switching-Cost-Aware Edge Management:** K. Zhao *et al.* [11] investigate adaptive model switching to balance latency and accuracy in edge inference services. C. Chen *et al.* [12] analyze the overhead of container switching linked to cold-start latency in caching-based serverless function orchestration across edges. K. Yang *et al.* [13] study the switching cost associated with deploying in-band network telemetry and monitoring within edge computing networks. Q. Hu *et al.* [14] focus on mitigating view-switching latency to enhance quality of experience in edge-driven free-viewpoint video streaming. X. Chen *et al.* [15] learn the cost of service migration and base station mode switching in dynamic edge computing environments. X. Liu *et al.* [16] study task reassignment overhead in live video transcoding systems, aiming to reduce switching frequency among edge transcoders.

These studies investigate issues related to switching cost in edge scheduling. However, they overlook model switching in the context of VPP energy scheduling, and neglect to treat model switching as a method to adjust energy consumption.

**Edge AI Auctions:** X. Chen *et al.* [19] develop a service auction that jointly optimizes resource allocation and quality of service for edge AI. J. Pang *et al.* [20] design an online auction scheduler for dynamic task and price allocation in distributed edge learning systems. R. Lu *et al.* [37] propose an auction-based clustered federated learning scheme to improve convergence and energy efficiency in the edge network. H. Liu *et al.* [38] introduce a privacy-preserving VCG auction with adaptive clustering for shared AI service deployment in edge networks. P. Wang *et al.* [39] propose a mobility-aware edge inference framework using distributed auction for adaptive model partitioning. W. Y. Lim *et al.* [40] construct a two-tier incentive mechanism combining evolutionary game and auction for hierarchical federated learning at the edge.

These studies focus on auction-based methods to incentivize scheduling in edge AI. However, they often ignore incentivizing each edge service provider to save energy in the VPP scenario, let alone the maximum and minimum limits of the edge batteries and their impact on energy incentives.

## III. Modeling and Formulation

### A. System Models

We summarize our notations and descriptions in Table I.

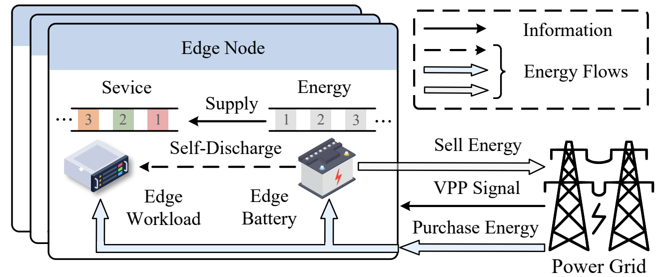



Fig. 2. Edge energy consumption

**Edge Infrastructure:** We target an edge computing infrastructure consisting of a group of distributed and potentially heterogeneous "edges", denoted as $\mathcal{N} \triangleq \{1, ..., N\}$, which are all owned and operated by an *Edge Infrastructure Provider*. An edge is typically a network access point, such as a base station or WiFi access point, that is augmented with co-located computing resources like a server or a micro data center. These edges are connected to one another via wired networks, and are connected to the end users via wired or wireless networks. The system is operated over a series of time slots $\mathcal{T} \triangleq \{1, ..., T\}$.

**Edge AI Services:** We consider a group of *Edge AI Service Providers*, denoted as $\mathcal{I} \triangleq \{1, ..., I\}$, which are the tenants renting resources on the edge infrastructure to offer their edge AI services to the end users. Without loss of generality, we envisage that each such service provider offers only one type of edge AI inference service. We use $\mathcal{I}_n \subseteq \mathcal{I}$, $\forall n \in \mathcal{N}$ to denote the set of the edge AI services that use the edge $n$. At the time slot $t \in \mathcal{T}$, for each edge $n$, the edge AI service $i \in \mathcal{I}_n$ always has a *default* model, which can be decided by the edge AI service itself and vary over time, to be deployed to serve its end users. We use $\lambda_{in}^t$ to denote the number of the inference requests that arrive at the edge $n$ at $t$ for the edge AI service $i$, and $\rho_{in}^t$ to denote the amount of energy consumed to serve one inference request, by the edge AI service $i$'s default model on the edge $n$ at $t$.

**Edge Energy Consumption:** As shown in Fig. 2, the entire edge infrastructure is connected to one external power grid, while each edge also has its local battery or battery groups installed. To support multi-tenant edge AI services, the edge infrastructure can draw energy from both the grid and the battery. For grid energy, we denote its purchasing price at the time slot $t$ as $s_-^t$. For battery energy, we denote the available battery energy at the edge $n$ at the time slot $t$ as $e_n^t$, and have $O_n \leq e_n^t \leq B_n$, where $O_n$ and $B_n$ are the desired minimum and maximum battery levels, respectively, and can be specified by the edge infrastructure provider or the battery manufacturer. For example, the lower and upper limits for the battery level can be set to 20% and 80% of full battery capacity. As every charge and discharge operation can harm the battery's lifespan, we consider the degradation cost [23] per unit energy charged or discharged as $d_n = M_n/(2K_n(B_n - O_n))$, where $M_n$ is the acquisition price of the battery, and $K_n$ refers to the number of cycles that the battery can be operated for. During energy transmission, the loss primarily stems from the grid-to-edge loss (usually borne by the power grid) and the battery charging/discharging inefficiency (usually borne by the edge infrastructure) [1], [24]. To precisely describe the system

TABLE I
NOTATIONS AND DESCRIPTIONS

| Input | Description |
|---|---|
| $\mathcal{I}$ | Set of edge AI service providers (bidders) |
| $\mathcal{N}$ | Set of edges |
| $\mathcal{T}$ | Set of time slots |
| $\lambda_{in}^t$ | Number of inference requests to edge $n$ for service $i$ at $t$ |
| $\rho_{in}^t$ | Energy per request by service $i$'s default model on edge $n$ at $t$ |
| $s_-^t$ | Energy purchasing price from power grid at $t$ |
| $s_+^t$ | Energy selling price to power grid at $t$ |
| $e_n^t$ | Available battery energy at edge $n$ at $t$ |
| $O_n$ | Minimum battery level for edge $n$ |
| $B_n$ | Maximum battery level for edge $n$ |
| $M_n$ | Acquisition price of battery for edge $n$ |
| $K_n$ | Number of cycles for which battery of edge $n$ can operate |
| $d_n$ | Degradation cost per unit energy charged/discharged at edge $n$ |
| $D^t$ | Energy demand of power grid at $t$ |
| $\eta_c$ | Battery charging efficiency |
| $\eta_d$ | Battery discharging efficiency |
| $\mathcal{V}_{in}^t$ | Bid issued by bidder $i$ for edge $n$ at $t$ |
| $r_{in}^t$ | Bidding price for bid of bidder $i$ for edge $n$ at $t$ |
| $\psi_{in}^t$ | Reduced energy per request of proposed model of bidder $i$ for edge $n$ compared to corresponding default model at $t$ |
| $\theta_{in}^t$ | Error rate of proposed model of bidder $i$ for edge $n$ at $t$ |
| $a_{in}^t$ | Whether default model for $t$ is same as proposed model for $t-1$ on edge $n$ for service $i$ |
| $b_{in}^t$ | Whether default model for $t-1$ is same as proposed model for $t$ on edge $n$ for service $i$ |
| $c_{in}^t$ | Whether proposed model for $t$ is same as proposed model for $t-1$ on edge $n$ for service $i$ |
| $m_{in}^t$ | Switching cost of default model on edge $n$ for service $i$ at $t$ |
| $l_{in}^t$ | Switching cost of proposed model on edge $n$ for service $i$ at $t$ |

| Decision | Description |
|---|---|
| $x_{in}^t$ | Whether bid from service $i$ for edge $n$ wins at $t$ |
| $y_n^t$ | Whether battery of edge $n$ charges from power grid at $t$ |
| $z_n^t$ | Whether battery of edge $n$ discharges to power grid at $t$ |
| $q_n^t$ | Energy drawn from power grid to workload of edge $n$ at $t$ |
| $u_n^t$ | Energy drawn from power grid to charge battery of edge $n$ at $t$ |
| $v_n^t$ | Energy discharged from battery of edge $n$ to power grid at $t$ |
| $w_n^t$ | Energy discharged from battery of edge $n$ to its workload at $t$ |
| $p_{in}^t$ | Payment made to service $i$ for its bid for edge $n$ at $t$ |

cost, we use $\eta_c$ and $\eta_d$ to represent the battery charging and discharging efficiency. That is, if the battery of the edge $n$ charges/discharges $J_n^t$ amount of energy, then the net charging energy is $J_n^t\eta_c$ and the net discharging energy is $J_n^t/\eta_d$.

**Virtual Power Plant (VPP):** The edge infrastructure can discharge batteries and release energy to the power grid as a virtual power plant. At any time slot $t$, the power grid can send its energy demand $D^t$; when receiving such energy demand, the edge infrastructure conducts an auction to incentivize the edge AI services to reduce their energy consumption so that the spared battery energy can be offered to the power grid at the price $s_+^t$, following their mutual agreements [31], to meet the energy demand from the power grid.

**VPP Auctions:** The edge infrastructure acts as the *auctioneer* and the edge AI services act as the *bidders* that submit bids. Each bid contains information about a *proposed* model, with the reduced energy consumed to serve one inference request compared to the default model. If a bid wins, then the corresponding proposed model is deployed and used to resolve end users' inference requests, so that the energy consumption is reduced; otherwise, the default model is deployed and used. There exists one auction in every time slot $t$, which proceeds as follows, also shown in Fig. 3.

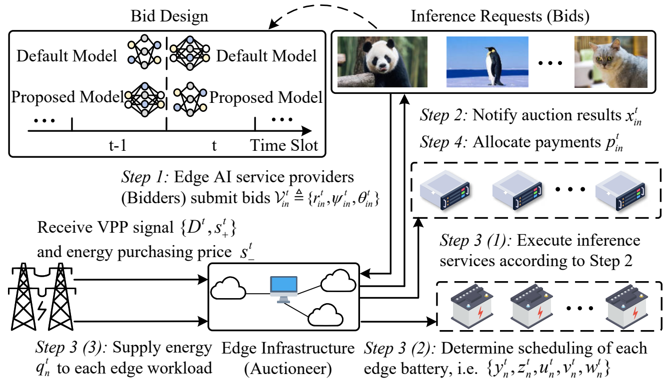



Fig. 3. Operation of a VPP auction

- *Step 1*: The auctioneer collects bids from bidders. The bid issued by the bidder $i$ for the edge $n$ is in the following format: $\mathcal{V}_{in}^t \triangleq \{r_{in}^t, \psi_{in}^t, \theta_{in}^t\}$, where $r_{in}^t$ is the bidding price, i.e., the amount of money that this bid is willing to charge from the auctioneer; $\psi_{in}^t$ is the reduced energy consumed to serve one inference request of the proposed new model of the bidder $i$ for the edge $n$, compared to the default model used by the bidder $i$ on the edge $n$; and $\theta_{in}^t$ is the error rate [5] of the proposed new model. We assume that $\psi_{in}^t$ and $\theta_{in}^t$ are verifiable by the edge infrastructure provider through post-deployment monitoring, thus the bidders have no incentive to misreport them. The truthfulness analysis (described in Section V-C) therefore focuses on the bidding price $r_{in}^t$.
- *Step 2*: Together with other input information that will be described next, the auctioneer decides the winning bids and the payment for each winning bid, where the payment unnecessarily equals the bidding price as elaborated later, with other control decisions including battery charging and discharging for each edge. The auctioneer notifies every bidder of the bid-selection outcome.
- *Step 3*: The bidders deploy and use their proposed model corresponding to each winning bid, deploy and use their default model corresponding to each bid that does not win. The auctioneer also conducts the energy control operations on each edge specified by other control decisions.
- *Step 4*: The auctioneer sends the payment to the bidders regarding each winning bid. Depending on any prior agreement, this step may happen before *Step 3*.

To decide the winning bids, the auctioneer solves the *social cost* minimization problem, which will be formulated next. Participation in an auction is voluntary. Any bidder can join one auction and does not join another, or vice versa. We do not penalize any bidder for this behavior. If a bid does not join or is not selected as a winning bid for some auction, and then if this bid wins in a later auction, this may incur the switching cost, which will be modeled and described as part of the social cost. Sending no bid to an auction can be equivalently considered as sending a virtual bid with an extremely high bidding price, so that the virtual bid will not be chosen as the winning bid and thus have no impact on social cost minimization.

**Model Switching:** For any edge AI service, on any of its edges, if the model used at the time slot $t$ differs from the

TABLE II
SWITCHING COST

| Indicators | | | Auction outcomes $\{x_{in}^{t-1}, x_{in}^t\}$ | | | |
|---|---|---|---|---|---|---|
| $a_{in}^t$ | $b_{in}^t$ | $c_{in}^t$ | $\{1,1\}$ | $\{1,0\}$ | $\{0,1\}$ | $\{0,0\}$ |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | $l_{in}^t$ | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | $l_{in}^t$ | 0 |
| 0 | 1 | 1 | $l_{in}^t$ | 0 | $l_{in}^t$ | 0 |
| 1 | 0 | 0 | 0 | $m_{in}^t$ | 0 | 0 |
| 1 | 0 | 1 | $l_{in}^t$ | $m_{in}^t$ | 0 | 0 |
| 1 | 1 | 0 | 0 | $m_{in}^t$ | $l_{in}^t$ | 0 |
| 1 | 1 | 1 | $l_{in}^t$ | $m_{in}^t$ | $l_{in}^t$ | 0 |

model used at the time slot $t-1$, then deploying the model at $t$ often incurs performance overhead, including the lead time of model loading and environment initialization, which we call the *switching cost*. Across $t-1$ and $t$, depending on the outcomes of the two auctions, the switching cost can occur between a default model and a proposed model, between a proposed model and a default model, or between two proposed models. Note that any switching cost between two default models is not incurred by the auctions and thus not considered.

We need further notations to express the model switching cost. At the time slot $t$, for the edge AI service $i$ on the edge $n$, we use $a_{in}^t \in \{0,1\}$ to indicate whether or not the default model for $t$ is the same as the proposed model for $t-1$, $b_{in}^t \in \{0,1\}$ to indicate whether or not the default model for $t-1$ is the same as the proposed model for $t$, and $c_{in}^t \in \{0,1\}$ to indicate whether or not the proposed model for $t$ is the same as the proposed model for $t-1$; also, we use $m_{in}^t$ to represent the switching cost incurred by deploying the default model for $t$, and $l_{in}^t$ to represent the switching cost incurred by deploying the proposed model for $t$. Note that these values are all known to the edge AI service at $t$, and also need to be reported to the edge infrastructure for the auction at $t$.

We are ready to formulate the model switching cost as

$$a_{in}^t m_{in}^t [x_{in}^{t-1} - x_{in}^t]^+ + b_{in}^t l_{in}^t [x_{in}^t - x_{in}^{t-1}]^+ + c_{in}^t l_{in}^t x_{in}^t x_{in}^{t-1},$$

where $[\cdot]^+ \triangleq \max\{\cdot, 0\}$. Depending on the values of the three indicators and the outcomes of the two auctions, we actually have $2^3 \times 2^2 = 32$ cases regarding the model switching cost, as elaborated in Table II. Note that our single formula in the above exactly captures these values altogether in this table.

**Control Decisions:** The edge infrastructure makes the following decisions at each $t$: (i) $x_{in}^t \in \{1,0\}$, denoting whether or not the bid from the edge AI service $i$ for the edge $n$ is chosen as a winning bid in the auction; (ii) $y_n^t \in \{1,0\}$, denoting whether or not the battery of the edge $n$ charges energy from the power grid; (iii) $z_n^t \in \{1,0\}$, denoting whether or not the battery of the edge $n$ discharges energy to the power grid; (iv) $q_n^t \geq 0$, denoting the amount of energy drawn from the power grid to support the workload of the edge $n$; (v) $u_n^t \geq 0$, denoting the amount of energy drawn from the power grid to charge the battery of the edge $n$; (vi) $v_n^t \geq 0$, denoting the amount of energy discharged from the battery of the edge $n$ to release to the power grid; (vii) $w_n^t \geq 0$, denoting the amount of energy discharged from the battery of the edge $n$ to support its workload; (viii) $p_{in}^t \geq 0$, denoting the payment made to the edge AI service $i$ for its bid for the edge $n$.

**Cost of Edge Infrastructure Provider:** The total cost of the edge infrastructure provider (i.e., auctioneer) consists of (i) cost of buying energy, (ii) revenue of selling energy (treated as negative cost), (iii) overhead of battery degradation, and (iv) payments to the bidders. That is $\sum_t \sum_n \big(s_-^t(u_n^t + w_n^t) - s_+^t v_n^t + d_n(\eta_c u_n^t + v_n^t/\eta_d + w_n^t/\eta_d) + \sum_{i\in\mathcal{I}_n} p_{in}^t\big)$.

**Cost of Edge AI Service Providers:** The total cost of the edge AI service providers (i.e., bidders) consists of (i) bidding cost, (ii) inference error rates, (iii) model switching cost, and (iv) payments from the auctioneer (treated as negative cost). That is $\sum_t \sum_n \sum_{i\in\mathcal{I}_n} \big(r_{in}^t x_{in}^t + \theta_{in}^t x_{in}^t + a_{in}^t m_{in}^t [x_{in}^{t-1} - x_{in}^t]^+ + b_{in}^t l_{in}^t [x_{in}^t - x_{in}^{t-1}]^+ + c_{in}^t l_{in}^t x_{in}^t x_{in}^{t-1} - p_{in}^t\big)$.

### B. Problem Formulation and Algorithmic Challenges

**Problem Formulation:** We formulate the *social cost minimization problem*, where the social cost refers to the sum of the total cost of the edge infrastructure provider and the total cost of the edge AI service providers. Note that the payments are canceled automatically, but still need to be decided carefully.

$$\begin{aligned}
\min\quad & \mathcal{P} \triangleq \textstyle\sum_t \sum_n \sum_{i\in\mathcal{I}_n} \big(r_{in}^t x_{in}^t + \theta_{in}^t x_{in}^t \\
& + a_{in}^t m_{in}^t [x_{in}^{t-1} - x_{in}^t]^+ + b_{in}^t l_{in}^t [x_{in}^t - x_{in}^{t-1}]^+ \\
& + c_{in}^t l_{in}^t x_{in}^t x_{in}^{t-1}\big) + \textstyle\sum_t \sum_n \big(s_-^t (q_n^t + u_n^t) \\
& - s_+^t v_n^t + d_n(\eta_c u_n^t + v_n^t/\eta_d + w_n^t/\eta_d)\big) & (1)\\
\text{s.t.}\quad & e_n^t = e_n^{t-1} + \eta_c u_n^t - v_n^t/\eta_d - w_n^t/\eta_d, \forall n, \forall t, & (1a)\\
& O_n \leq e_n^t \leq B_n, \forall n, \forall t, & (1b)\\
& y_n^t + z_n^t \leq 1, \forall n, \forall t, & (1c)\\
& \textstyle\sum_n v_n^t \leq D^t, \forall t, & (1d)\\
& u_n^t \leq y_n^t B_n, \forall n, \forall t, & (1e)\\
& v_n^t \leq z_n^t B_n, \forall n, \forall t, & (1f)\\
& \textstyle\sum_{i\in\mathcal{I}_n} \big(\lambda_{in}^t(\rho_{in}^t - \psi_{in}^t x_{in}^t)\big) \leq q_n^t + w_n^t, \forall n, \forall t, & (1g)\\
\text{var.}\quad & x_{in}^t \in \{0,1\}, y_n^t \in \{0,1\}, z_n^t \in \{0,1\}, q_n^t \geq 0, \\
& u_n^t \geq 0, v_n^t \geq 0, w_n^t \geq 0, \forall i \in \mathcal{I}_n, \forall n, \forall t. & (1h)
\end{aligned}$$

The objective function (1) is the social cost. Constraint (1a) ensures battery state transitions are correctly maintained under charging and discharging operations. Constraint (1b) ensures the battery energy is always within the desired limits. Constraint (1c) ensures the battery cannot be charged and discharged simultaneously. Constraint (1d) ensures the energy provided by the edge infrastructure to the power grid does not exceed the demand, for grid safety and stability. Constraints (1e) and (1f) ensure the amount of energy charged and discharged must not exceed the battery limit. Constraint (1g) ensures the energy provided jointly by the power grid and the battery is sufficient to support the edge AI services, considering the proposed models in the bids. All decision variables with their domains are specified as Constraint (1h).

**Algorithmic Challenges:** Solving this problem *in an online manner*, i.e., making the control decisions in real time for each time slot as time goes, faces multiple fundamental challenges.

*Challenge #1: Intractability.* Our problem is NP-hard:

**Theorem 1.** *The problem $\mathcal{P}$ is NP-hard, due to reduction from the minimum knapsack problem.*

TABLE III
TRANSFORMED VARIABLES

| Notations | Descriptions |
|---|---|
| **Decision Variables** | |
| $\{\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t\}$ | Decision variables $\{x_{in}^t, y_n^t, z_n^t\}$ omitting $i$ and $n$ |
| $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ | Decision variables $\{u_n^t, v_n^t, w_n^t, q_n^t\}$ omitting $n$ |
| $\mathrm{Z}_1^t$ for $\mathcal{P}_1$ | Decision variable $\mathrm{Z}_1^t = \{\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| $\mathrm{Z}_2^t$ for $\mathcal{P}_2$ | Decision variable $\mathrm{Z}_2^t = \{\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| $\mathrm{Z}_3^t$ for $\mathcal{P}_3$ | Decision variable $\mathrm{Z}_3^t = \{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| **Online Solutions** | |
| $\{\tilde{\boldsymbol{x}}^t, \tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t\}$ | Online fractional solutions of decisions $\{\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t\}$ |
| $\{\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t\}$ | Online integral solutions of decisions $\{\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t\}$ |
| $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ | Online solutions of decisions $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| $\bar{\mathrm{Z}}_1^t$ for $\mathcal{P}_1$ | Online solution $\bar{\mathrm{Z}}_1^t = \{\bar{\boldsymbol{x}}^t, \tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| $\bar{\mathrm{Z}}_2^t$ for $\mathcal{P}_2$ | Online solution $\bar{\mathrm{Z}}_2^t = \{\bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| $\bar{\mathrm{Z}}_3^t$ for $\mathcal{P}_3$ | Online solution $\bar{\mathrm{Z}}_3^t = \{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| $\bar{\mathrm{Z}}^t$ for $\mathcal{P}$ | Online solution $\bar{\mathrm{Z}}^t = \{\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ |
| **Offline Optimal Solutions** | |
| $\mathrm{Z}_{1,2,3}^{t\star}$ for $\mathcal{P}_{1,2,3}$ | Offline optimal solution at each time slot $t$ |
| $\mathrm{Z}^{t*}$ for $\mathcal{P}$ | Offline optimal solution over whole time horizon $T$ |

*Proof.* See Appendix A in our supplemental material. □

*Challenge #2: Unconventional Switching Cost.* Unlike many existing works, which typically have $[x_{in}^t - x_{in}^{t-1}]^+$ as the only switching-cost term, our switching-cost terms include $[x_{in}^t - x_{in}^{t-1}]^+$, $[x_{in}^{t-1} - x_{in}^t]^+$, and $x_{in}^t x_{in}^{t-1}$. At the time $t-1$, because the time slot $t$ has not arrived and $x_{in}^t$ remains unknown, it is non-trivial to decide $x_{in}^{t-1}$ to minimize our switching cost between $t-1$ and $t$. The complex form escalates this difficulty.

*Challenge #3: State Transitions.* Constraint (1a) connects $e_n^t$ at $t$ to $e_n^{t-1}$ at $t-1$, coupling every pair of adjacent time slots. Due to this decisive relation, whatever charging and discharging decisions we make at $t-1$ will restrict the corresponding decisions we could make at $t$ within the battery limits as in Constraint (1b). This in turn restricts how we can optimize our objective. Still, at $t-1$, the inputs for $t$ are not revealed yet; it is thus not easy to manage battery energy.

*Challenge #4: Economic Requirements.* For the individual auction at each time slot $t$, selecting winning bids and deciding payment for each winning bid entails satisfying truthfulness and individual rationality, as defined and elaborated later. Unlike standard social cost minimization, our social cost has a non-standard form and is intertwined with multiple challenges as stated, making it not straightforward to meet the two economic requirements while addressing those challenges.

## IV. ALGORITHM DESIGN

We first transform and decompose the problem $\mathcal{P}$ (Section IV-A). Then we propose Algorithm 1 (Section IV-B) which invokes Algorithm 2 (Section IV-C) to obtain the fractional solutions and Algorithm 3 (Section IV-D) for randomized rounding, while calling Algorithm 4 (Section IV-E) for payment allocation. The transformed variables are in TABLE III.

### A. Transformation and Decomposition

**Transformation:** For the objective function (1), we divide it into two parts in order to solve them collaboratively. At each time slot $t$, the switching cost is denoted as $f_S^t \triangleq \sum_n \sum_{i\in\mathcal{I}_n} \big(a_{in}^t m_{in}^t [x_{in}^{t-1} - x_{in}^t]^+ + b_{in}^t l_{in}^t [x_{in}^t - x_{in}^{t-1}]^+ + c_{in}^t l_{in}^t x_{in}^t x_{in}^{t-1}\big)$, which only refers to the cost caused by switching models. Then the non-switching cost at each time slot $t$ is defined as $f_{-S}^t \triangleq \sum_n \sum_{i\in\mathcal{I}_n} \big(r_{in}^t x_{in}^t + \theta_{in}^t x_{in}^t\big) + \sum_n \big(s_-^t(q_n^t + u_n^t) - s_+^t v_n^t + d_n(u_n^t \eta_c + v_n^t/\eta_d + w_n^t/\eta_d)\big)$, which indicates others in the social cost except switching cost.

For the constraints (1a) and (1b), we regard the energy scheduling for the battery of each edge $n$ as an "energy queue". Following this, we have $e_n^{T+1} = e_n^T + u_n^T\eta_c - v_n^T/\eta_d - w_n^T/\eta_d = ... = e_n^0 + \sum_{t=1}^T (u_n^t \eta_c - v_n^t/\eta_d - w_n^t/\eta_d)$, which leads to $\sum_{t=1}^t (u_n^t\eta_c - v_n^t/\eta_d - w_n^t/\eta_d) = e_n^{t+1} - e_n^0 \leq B_n - O_n$. Next, we sum the inequality above from $t=1$ to $t=T-1$ and convert the constraints (1a) and (1b) to a *long-term constraint*: $\sum_t g^t \triangleq \sum_t (T-t)(u_n^t\eta_c - v_n^t/\eta_d - w_n^t/\eta_d) - T(B_n - O_n) \leq 0$, $\forall n$. We emphasize: (i) This is a *relaxation*, any feasible solution of the original problem $\mathcal{P}$ satisfies $\sum_t g^t \leq 0$, but the converse does not necessarily hold. (ii) This relaxation is adopted to enable online primal-dual decomposition (Algorithm 2), as the original per-slot coupling would otherwise make the problem intractable without knowledge of future inputs. (iii) Theorem 2 proves that the cumulative violation of $\sum_t g^t \leq 0$ is sublinear in $T$, i.e., $\mathrm{Fit}^T \leq \mathcal{O}(T^\sigma), \sigma \in (0,1)$. Thus, the long-run average of the battery level $e_n^t$ asymptotically respects the battery capacity $[O_n, B_n]$. (iv) Fig. 15 shows that $e_n^t$ always remains within $[O_n, B_n]$, i.e., the levels of all batteries stay within their capacity limits in our experiments.

For the constraints (1c) to (1g), we consider them as *instantaneous constraints* and simplify their expressions: (1c) is denoted as $h_1^t \triangleq y_n^t + z_n^t - 1, \forall n, \forall t$; (1d) is defined as $h_2^t \triangleq u_n^t - y_n^t B_n, \forall n, \forall t$; (1e) is written as $h_3^t \triangleq v_n^t - z_n^t B_n, \forall n, \forall t$; (1f) is expressed as $h_4^t \triangleq \sum_n v_n^t - D^t, \forall t$; (1g) is notated as $h_5^t \triangleq \sum_{i\in\mathcal{I}_n} \big(\lambda_{in}^t(\rho_{in}^t - \psi_{in}^t x_{in}^t)\big) - q_n^t - w_n^t, \forall n, \forall t$. After that, we indicate $h^t \triangleq \{h_1^t, h_2^t, h_3^t, h_4^t, h_5^t\}$ as a set of them.

To intuitively illustrate the "online" solution, we omit the indices $i$ and $n$ and only use the index $t$ to represent the state of time. The problem $\mathcal{P}$ is now transformed as follows:

$$\begin{aligned}
\min\quad & \mathcal{P}' \triangleq \textstyle\sum_t f_S^t(\boldsymbol{x}^t) + \sum_t f_{-S}^t(\boldsymbol{x}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) && (2)\\
\text{s.t.}\quad & \textstyle\sum_t g^t(\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t) \leq 0, && (2a)\\
& h^t(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) \leq 0, \forall t, && (2b)\\
\text{var.}\quad & \boldsymbol{x}^t \in [0,1], \boldsymbol{y}^t \in [0,1], \boldsymbol{z}^t \in [0,1],\\
& \boldsymbol{u}^t \geq 0, \boldsymbol{v}^t \geq 0, \boldsymbol{w}^t \geq 0, \boldsymbol{q}^t \geq 0, \forall t, && (2c)
\end{aligned}$$

**Decomposition:** We denote $\{\tilde{\boldsymbol{x}}^t, \tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t\}$ as the fractional solutions and $\{\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t\}$ as the integral solutions. In addition, the values of $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ are always fractions. Based on the above, we decouple $\mathcal{P}'$ into three sub-problems:

$$\begin{aligned}
\min\quad & \mathcal{P}_1 \triangleq \textstyle\sum_t f_{-S}^t(\boldsymbol{x}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) && (3)\\
\text{s.t.}\quad & \textstyle\sum_t g^t(\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t) \leq 0, && (3a)\\
& h^t(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) \leq 0, \forall t, && (3b)\\
\text{var.}\quad & \boldsymbol{x}^t \in [0,1], \boldsymbol{y}^t \in [0,1], \boldsymbol{z}^t \in [0,1],\\
& \boldsymbol{u}^t \geq 0, \boldsymbol{v}^t \geq 0, \boldsymbol{w}^t \geq 0, \boldsymbol{q}^t \geq 0, \forall t. && (3c)
\end{aligned}$$

where the objective function (3) captures the non-switching cost, while Algorithm 1 jointly incorporates both switching

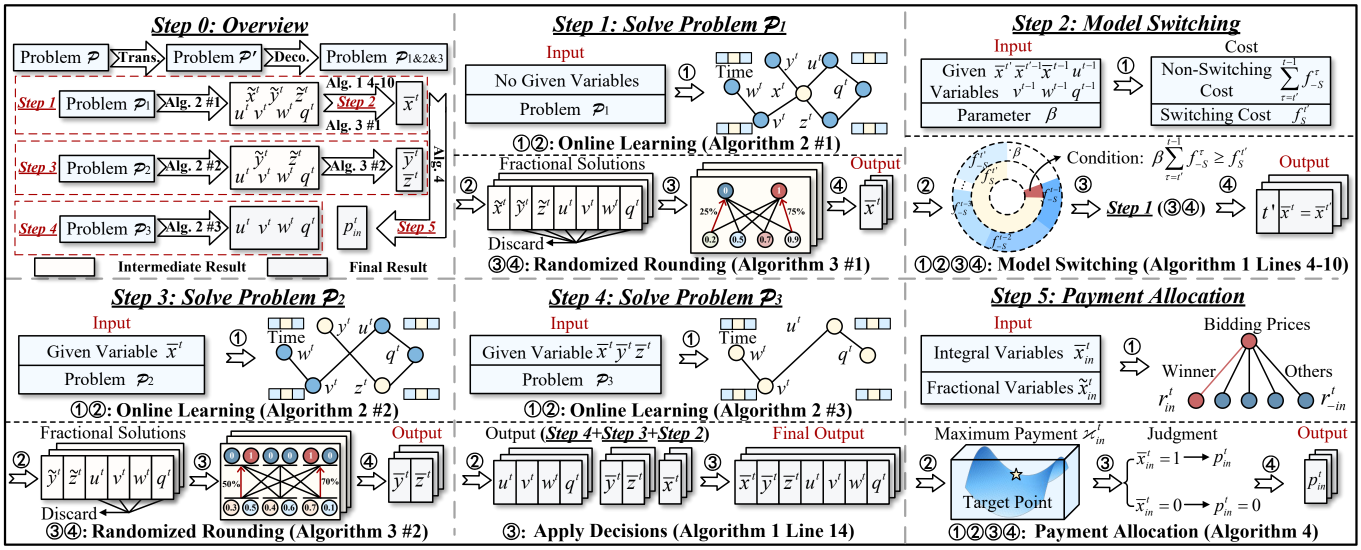



Fig. 4. Workflow of Auction-Driven Energy Scheduling (ADES) Algorithm which consists of 5 steps operating collaboratively to solve the problem.

and non-switching cost through its carefully designed *switching condition*. And the constraints (3a) (3b) (3c) are the same as the constraints (2a) (2b) (2c) in the problem $\mathcal{P}'$.

$$\begin{aligned}
\min \quad & \mathcal{P}_2 \triangleq \textstyle\sum_t f^t_{-S}(\bar{\boldsymbol{x}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) && (4)\\
\text{s.t.} \quad & \textstyle\sum_t g^t(\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t) \le 0, && (4a)\\
& h^t(\bar{\boldsymbol{x}}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) \le 0, \forall t, && (4b)\\
\text{var.} \quad & \boldsymbol{y}^t \in [0,1], \boldsymbol{z}^t \in [0,1],\\
& \boldsymbol{u}^t \ge 0, \boldsymbol{v}^t \ge 0, \boldsymbol{w}^t \ge 0, \boldsymbol{q}^t \ge 0, \forall t. && (4c)
\end{aligned}$$

where the integral solution $\bar{\boldsymbol{x}}^t$ of $\mathcal{P}_1$ (is given in Line 10 of Algorithm 1 described latter) is used as the input to solve $\mathcal{P}_2$.

$$\begin{aligned}
\min \quad & \mathcal{P}_3 \triangleq \textstyle\sum_t f^t_{-S}(\bar{\boldsymbol{x}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) && (5)\\
\text{s.t.} \quad & \textstyle\sum_t g^t(\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t) \le 0, && (5a)\\
& h^t(\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t) \le 0, \forall t, && (5b)\\
\text{var.} \quad & \boldsymbol{u}^t \ge 0, \boldsymbol{v}^t \ge 0, \boldsymbol{w}^t \ge 0, \boldsymbol{q}^t \ge 0, \forall t. && (5c)
\end{aligned}$$

where the integral solutions $\{\bar{\boldsymbol{x}}^t(\mathcal{P}_1), \bar{\boldsymbol{y}}^t(\mathcal{P}_2), \bar{\boldsymbol{z}}^t(\mathcal{P}_2)\}$ ($\bar{\boldsymbol{x}}^t$ is given in Line 10 of Algorithm 1 while $\bar{\boldsymbol{y}}^t$ and $\bar{\boldsymbol{z}}^t$ come from Line 12 of Algorithm 1) are used as the inputs to solve $\mathcal{P}_3$.

**Overview (*Step 0* of Fig. 4):** The original problem $\mathcal{P}$ and its variant $\mathcal{P}'$ are not solved directly; instead, the subproblems $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ and the payments $p^t_{in}$ are solved by the Algorithms 1∼4 . (i) *Step 1:* For $\mathcal{P}_1$, we get $\{\tilde{\boldsymbol{x}}^t, \tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ by Algorithm 2 *#1*. (ii) *Step 2:* We round $\tilde{\boldsymbol{x}}^t$ into $\bar{\boldsymbol{x}}^t$ by Lines 4-10 of Algorithm 1 (including Algorithm 3 *#1*). (iii) *Step 3:* For $\mathcal{P}_2$, taking $\bar{\boldsymbol{x}}^t$ in $\mathcal{P}_1$ as input, we get $\{\tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ by Algorithm 2 *#2* and round $\{\tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t\}$ into $\{\bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t\}$ by Algorithm 3 *#2*. (iv) *Step 4:* For $\mathcal{P}_3$, taking $\bar{\boldsymbol{x}}^t$ from $\mathcal{P}_1$ and $\{\bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t\}$ from $\mathcal{P}_2$ as inputs, we obtain $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ by Algorithm 2 *#3*. (v) *Step 5:* We compute $p^t_{in}$ from Algorithm 4. The above *Steps 1∼5* are also integrated into Algorithm 1 below.

**Insights:** The original problem couples model selection, battery control, and energy scheduling, making it hard to solve online due to time coupling and unknown future inputs. We relax the strict per-slot battery limits into a long-term average constraint, then decompose the problem into three sequential subproblems: first choose which models to use (balancing switching cost vs. non-switching cost), then decide battery charging/discharging/resting directions, and finally compute the exact energy amounts. This stepwise pipeline isolates each difficulty, enabling an efficient online algorithm while keeping the key interactions through carefully defined interfaces.

### *B. Auction-Driven Energy Scheduling Algorithm*

We display the workflow in Fig. 4, also detail below.

(0) *Step 0:* The overview. (i) *Step 1:* We solve the problem $\mathcal{P}_1$ by Algorithm 2 *#1* in Line 3, obtaining the fractional solutions. The integral solutions $\bar{\boldsymbol{x}}^t$ can be received via Algorithm 3 *#1* and confirmed after meeting the *switching condition* in Line 4. Here, we discard $\{\tilde{\boldsymbol{y}}^t, \tilde{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ of $\mathcal{P}_1$. (ii) *Step 2:* We balance the switching cost $f^t_S$ and the non-switching cost $f^t_{-S}$ using the pre-specified parameter $\beta$ in Line 4, where $t'$ records the most recent time slot that a *switching operation* occurs. If $\beta$ times the cumulative non-switching cost that occurred from the time slot $t'$ to $t-1$ exceeds the switching cost of the last *switching operation* at the time slot $t'$, we will apply the new decisions which can be computed in Lines 5-8. After that, the model switching operation is executed in Line 10. (iii) *Step 3:* We acquire the fractional solutions of the problem $\mathcal{P}_2$ via Algorithm 2 *#2* in Line 11, where $\bar{\boldsymbol{x}}^t$ is the input. Afterwards, the fractional solutions $\tilde{\boldsymbol{y}}^t$ and $\tilde{\boldsymbol{z}}^t$ are rounded to integer solutions $\bar{\boldsymbol{y}}^t$ and $\bar{\boldsymbol{z}}^t$ via Algorithm 3 *#2* in Line 12. Here, we discard $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ of $\mathcal{P}_2$ (iv) *Step 4:* The solutions of the problem $\mathcal{P}_3$ are calculated by Algorithm 2 *#3* in Line 13, where $\bar{\boldsymbol{x}}^t$, $\bar{\boldsymbol{y}}^t$ and $\bar{\boldsymbol{z}}^t$ are the inputs. Then a series of variables are added to $t$ as the final solutions in Line 14. (v) *Step 5:* We allocate payments $p^t_{in}$ using $\bar{x}^t_{in}$ and $\tilde{x}^t_{in}$ by invoking Algorithm 4 in Line 15.

**Insights:** Aside from its invocation of other algorithmic components, Algorithm 1 employs a pre-specified parameter $\beta$ to enforce that the cost of each previous switch between consecutive switching operations is no greater than $\beta$ times the accumulated non-switching costs. Intuitively, this method

**Algorithm 1** Auction-Driven Energy Scheduling Algorithm

```
Initialization: t' = 0, x̄^{-1} = 0, x̄^0 = ȳ^0 = z̄^0 = 0,
    u^0 = v^0 = w^0 = q^0 = 0, parameter β;
for t = 1, 2, ..., T do
    Get x̃^t, ỹ^t, z̃^t, u^t, v^t, w^t, q^t from P_1 via Algorithm 2 #1;
    if β Σ_{τ=t'}^{t-1} f^τ_{-S}(x̄^τ, u^τ, v^τ, w^τ, q^τ) // Switching condition
        ≥ f^{t'}_S(x̄^{t'}, x̄^{t'-1}) then
        Round x̃^t to x̄^t via Algorithm 3 #1;
        if x̄^t ≠ x̄^{t-1} then
            Set t' = t;  // Record model switching time slots
        end if
    end if
    Set x̄^t = x̄^{t'};   // Record model switching operations
    Get ỹ^t, z̃^t, u^t, v^t, w^t, q^t from P_2 via Algorithm 2 #2;
    Round ỹ^t, z̃^t to ȳ^t, z̄^t via Algorithm 3 #2;
    Get u^t, v^t, w^t, q^t from P_3 via Algorithm 2 #3;
    Apply decisions x̄^t, ȳ^t, z̄^t, u^t, v^t, w^t, q^t to t;
    Invoke Algorithm 4 to allocate payments for t;
end for
```

ensures that the total cost, including both switching and non-switching components, remains no greater than $\beta+1$ times the non-switching cost alone. While this method of managing the switching operations is passive rather than active in nature, the parameter $\beta$ serves as a precise regulative mechanism. Under this principle, we can effectively decouple the interdependent bid selection processes across sequential auctions.

### C. Online Fractional Algorithm

We note that $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ are in the form as follows:

$$\min \quad \textstyle\sum_t f^t_{-S}(\mathrm{Z}^t) \tag{6}$$
$$\text{s.t.} \quad \textstyle\sum_t g^t(\mathrm{Z}^t) \leq 0, h^t(\mathrm{Z}^t) \leq 0, \tag{6a}$$

where $\mathrm{Z}^t$ is the set of decision variables. Then we reformulate (6) to (7) equivalently with Lagrangian method:

$$\min_{\mathrm{Z}^t} \max_{\varphi^t} \quad \textstyle\sum_t \left(f^t_{-S}(\mathrm{Z}^t) + \varphi^t g^t(\mathrm{Z}^t)\right) \tag{7}$$
$$\text{s.t.} \qquad h^t(\mathrm{Z}^t) \leq 0, \tag{7a}$$

where $\varphi^t$ is the corresponding Lagrange multiplier at $t$ and we denote $\mathcal{L}^t(\mathrm{Z}^t, \varphi^t) \triangleq f^t_{-S}(\mathrm{Z}^t) + \varphi^t g^t(\mathrm{Z}^t)$. Based on these, we then design a primal-dual method to update the dual variable $\varphi^{t+1}$ in (8) and the primal variable $\mathrm{Z}^{t+1}$ in (9) alternately in an online manner as time goes. The expression of $\varphi^{t+1}$ is

$$\varphi^{t+1} = [\varphi^t + \mu g^t(\mathrm{Z}^t)]^+, \tag{8}$$

where $\mu$ is the positive step size, $g^t(\mathrm{Z}^t) \triangleq \nabla_\varphi \mathcal{L}^t(\mathrm{Z}^t, \varphi^t)$ is the gradient of $\mathcal{L}^t(\mathrm{Z}, \varphi)$ when $\varphi = \varphi^t$. Afterwards, we can obtain $\mathrm{Z}^{t+1}$ by solving the following problem:

$$\min \quad \nabla f^t_{-S}(\mathrm{Z}^t)(\mathrm{Z} - \mathrm{Z}^t) + \varphi^{t+1} g^t(\mathrm{Z}) + \tfrac{\|\mathrm{Z}-\mathrm{Z}^t\|^2}{2\alpha} \tag{9}$$
$$\text{s.t.} \quad h^t(\mathrm{Z}) \leq 0, \tag{9a}$$

where $\alpha$ is a predefined constant and $\nabla f^t_{-S}(\mathrm{Z}^t)$ is the gradient of $f^t_{-S}(\mathrm{Z})$ at $\mathrm{Z} = \mathrm{Z}^t$. And we approximate $\mathcal{L}^t(\mathrm{Z}, \varphi^t)$ to $\nabla f^t_{-S}(\mathrm{Z}^t)(\mathrm{Z}-\mathrm{Z}^t) + \varphi^{t+1} g^t(\mathrm{Z})$, where the regularization term $\frac{1}{2\alpha}\|\mathrm{Z}-\mathrm{Z}^t\|^2$ is a proximal term. Note that problem (9) can be solved using optimization tools by the interior point method.

Algorithm 2 adopts the above method. Through such alternate descent-ascent steps, we can obtain $\mathrm{Z}^{t+1}$ and $\varphi^{t+1}$ by applying the inputs before $t+1$ rather than the future

**Algorithm 2** Online Fractional Algorithm

```
Input: f^t_{-S}(·), g^t(·), h^t(·), Z^t
Output: Fractional solution Z^{t+1} and dual solution φ^{t+1}
    // #1: For P_1, Z^t = Z^t_1 = {x^t, y^t, z^t, u^t, v^t, w^t, q^t}
    // #2: For P_2, Z^t = Z^t_2 = {y^t, z^t, u^t, v^t, w^t, q^t}, input x̄^t
    // #3: For P_3, Z^t = Z^t_3 = {u^t, v^t, w^t, q^t}, input {x̄^t, ȳ^t, z̄^t}
Initialization: step size μ and constant α;
Calculate φ^{t+1} according to formula (8);
Calculate Z^{t+1} according to formula (9).
```

information. Note that Algorithm 2 is not a standard primal-dual method, but a modified online approach. Overall, the time complexity of Algorithm 2 is $\mathcal{O}(I^2N^2\log(1/\nu))$ while operating *situation #1* and $\mathcal{O}(N^2\log(1/\nu))$ while operating *situation #2* as well as *situation #3*. These results arise because the core optimization problem (9) is solvable using an standard convex optimization solver [41] which has the $\nu$-accurate optimal solution via the interior-point method.

**Insights:** (i) The long-term constraints are no longer issues, since they are incorporated into the objective function. Consequently, only instantaneous constraints remain, which naturally allows the problem to be decomposed and solved within each individual time slot. (ii) The posterior input such as inference error rate also cease to be a challenge, because when computing $\mathrm{Z}^t$, no information at or after $t$ is required or used, let alone any posterior information. In fact, this is why this process can be referred to as "online learning".

### D. Randomized Rounding Algorithm

We design Algorithm 3 to convert the fractional solutions $\widetilde{x}^t_{in}$, $\widetilde{y}^t_n$ and $\widetilde{z}^t_n$ obtained from Algorithm 2 into integers in two "randomized" manners without violating the instantaneous constraints while ensuring the expectation of each integer after rounding equals the corresponding fraction before rounding.

First, as shown in Lines 1-13, we apply a randomized *method #1* with two probabilities to round $\widetilde{x}^t_{in}$. In order to preserve the expectation, as indicated in Lines 4-8, we have $\mathbb{E}[\bar{x}^t_{i_1n}] = \frac{\omega_2}{\omega_1+\omega_2}(\widetilde{x}^t_{i_1n} + \omega_1) + \frac{\omega_1}{\omega_1+\omega_2}(\widetilde{x}^t_{i_1n} - \omega_2) = \widetilde{x}^t_{i_1n}$ as well as $\mathbb{E}[\bar{x}^t_{i_2n}] = \frac{\omega_2}{\omega_1+\omega_2}(\widetilde{x}^t_{i_2n} - \varrho\omega_1) + \frac{\omega_1}{\omega_1+\omega_2}(\widetilde{x}^t_{i_2n} + \varrho\omega_2) = \widetilde{x}^t_{i_2n}$. To ensure the instantaneous constraint (1g) of the problem $\mathcal{P}$, where $\widetilde{x}^t_{in}$ appears, has no violation after rounding, we either have $\lambda^t_{i_1n}\psi^t_{i_1n}(\widetilde{x}^t_{i_1n} + \omega_1) + \lambda^t_{i_2n}\psi^t_{i_2n}(\widetilde{x}^t_{i_2n} - \varrho\omega_1) = \lambda^t_{i_1n}\psi^t_{i_1n}\widetilde{x}^t_{i_1n} + \lambda^t_{i_2n}\psi^t_{i_2n}\widetilde{x}^t_{i_2n}$ in Line 7, or have $\lambda^t_{i_1n}\psi^t_{i_1n}(\widetilde{x}^t_{i_1n} - \omega_2) + \lambda^t_{i_2n}\psi^t_{i_2n}(\widetilde{x}^t_{i_2n} + \varrho\omega_2) = \lambda^t_{i_1n}\psi^t_{i_1n}\widetilde{x}^t_{i_1n} + \lambda^t_{i_2n}\psi^t_{i_2n}\widetilde{x}^t_{i_2n}$ in Line 8. That is, after rounding, the value of $\sum_{i\in\mathcal{I}_n}\lambda^t_{in}\psi^t_{in}\widetilde{x}^t_{in}$ does not change,thus (1g) still holds.

Second, Lines 14-20 design a randomized rounding *method #2*, simultaneously rounding the fractional values $\widetilde{y}^t_n$ and $\widetilde{z}^t_n$ with three carefully chosen probabilities. This procedure is designed to preserve the expectations and maintain the instantaneous constraint of the problem $\mathcal{P}$. The instantaneous constraint (1c) is not violated at each time slot $t$ since we have both $\bar{y}^t_n + \bar{z}^t_n = 1$ in Lines 17-18, and $\bar{y}^t_n + \bar{z}^t_n = 0$ in Line 19. Moreover, we preserve $\mathbb{E}[\bar{y}^t_n] = \widetilde{y}^t_n$ and $\mathbb{E}[\bar{z}^t_n] = \widetilde{z}^t_n$ in the meanwhile since the former can be calculated as $\mathbb{E}[\bar{y}^t_n] = \frac{\gamma_1}{\gamma_1+\gamma_2+\gamma_3}(\widetilde{y}^t_n + \gamma_2 + \gamma_3) + \frac{\gamma_2}{\gamma_1+\gamma_2+\gamma_3}(\widetilde{y}^t_n - \gamma_1) + \frac{\gamma_3}{\gamma_1+\gamma_2+\gamma_3}(\widetilde{y}^t_n - \gamma_1) = \widetilde{y}^t_n$, while the latter can be expressed as $\mathbb{E}[\bar{z}^t_n] = \frac{\gamma_1}{\gamma_1+\gamma_2+\gamma_3}(\widetilde{z}^t_n - \gamma_2) + \frac{\gamma_2}{\gamma_1+\gamma_2+\gamma_3}(\widetilde{z}^t_n + \gamma_1 + \gamma_3) + \frac{\gamma_3}{\gamma_1+\gamma_2+\gamma_3}(\widetilde{z}^t_n - \gamma_2) = \widetilde{z}^t_n$. In summary, both methods employ

**Algorithm 3** Randomized Rounding Algorithm

**Input:** Fractional solutions $\widetilde{x}_{in}^t$ or $\{\widetilde{y}_n^t, \widetilde{z}_n^t\}$
**Output:** Integral solutions $\bar{x}_{in}^t$ or $\{\bar{y}_n^t, \bar{z}_n^t\}$
// ***#1:*** Round fraction $\widetilde{x}_{in}^t$ to integer $\bar{x}_{in}^t$
**for** $n \in \mathcal{N}$ **do**
  Define $\mathcal{I}_n' = \mathcal{I}_n \backslash \{i | \widetilde{x}_{in}^t \in \{0,1\}\}$
  **while** $\mathcal{I}_n' \neq \emptyset$ **do**
    Select $i_1, i_2 \in \mathcal{I}_n'$, $i_1 \neq i_2$, define $\varrho = \frac{\lambda_{i_1 n}^t \psi_{i_1 n}^t}{\lambda_{i_2 n}^t \psi_{i_2 n}^t}$;
    $\omega_1 = \min\{\lceil \widetilde{x}_{i_1 n}^t \rceil - \widetilde{x}_{i_1 n}^t, \frac{1}{\varrho} \widetilde{x}_{i_2 n}^t\}$,
    $\omega_2 = \min\{\frac{1}{\varrho}(\widetilde{x}_{i_2 n}^t - \lfloor \widetilde{x}_{i_2 n}^t \rfloor), \widetilde{x}_{i_1 n}^t\}$;
    With probability $\frac{\omega_2}{\omega_1 + \omega_2}$,
    Set $\widetilde{x'}_{i_1 n}^t = \widetilde{x}_{i_1 n}^t + \omega_1, \widetilde{x'}_{i_2 n}^t = \widetilde{x}_{i_2 n}^t - \varrho \omega_1$;
    With probability $\frac{\omega_1}{\omega_1 + \omega_2}$,
    Set $\widetilde{x'}_{i_1 n}^t = \widetilde{x}_{i_1 n}^t - \omega_2, \widetilde{x'}_{i_2 n}^t = \widetilde{x}_{i_2 n}^t + \varrho \omega_2$;
    **if** $\forall i \in \{i_1, i_2\}$, $\widetilde{x'}_{in}^t \in \{0,1\}$ **then**
      Set $\bar{x}_{in}^t = \widetilde{x'}_{in}^t$, $\mathcal{I}_n' = \mathcal{I}_n' \setminus \{i\}$;
    **end if**
  **end while**
**end for**
// ***#2:*** Round fractions $\{\widetilde{y}_n^t, \widetilde{z}_n^t\}$ to integers $\{\bar{y}_n^t, \bar{z}_n^t\}$
Define $\mathcal{N}' = \mathcal{N} \backslash \{n | \widetilde{y}_n^t \in \{0,1\}, \widetilde{z}_n^t \in \{0,1\}\}$;
**for** $n \in \mathcal{N}'$ **do**
  $\gamma_1 = \widetilde{y}_n^t - \lfloor \widetilde{y}_n^t \rfloor, \gamma_2 = \widetilde{z}_n^t - \lfloor \widetilde{z}_n^t \rfloor, \gamma_3 = 1 - \widetilde{y}_n^t - \widetilde{z}_n^t$;
  With probability $\frac{\gamma_1}{\gamma_1 + \gamma_2 + \gamma_3}$,
  Set $\bar{y}_n^t = \widetilde{y}_n^t + \gamma_2 + \gamma_3, \bar{z}_n^t = \widetilde{z}_n^t - \gamma_2$;
  With probability $\frac{\gamma_2}{\gamma_1 + \gamma_2 + \gamma_3}$,
  Set $\bar{y}_n^t = \widetilde{y}_n^t - \gamma_1, \bar{z}_n^t = \widetilde{z}_n^t + \gamma_1 + \gamma_3$;
  With probability $\frac{\gamma_3}{\gamma_1 + \gamma_2 + \gamma_3}$,
  Set $\bar{y}_n^t = \widetilde{y}_n^t - \gamma_1, \bar{z}_n^t = \widetilde{z}_n^t - \gamma_2$;
**end for**

the concept of dependent rounding and the time complexity is $\mathcal{O}(IN)$ for *method #1* and $\mathcal{O}(N)$ for *method #2*.

**Insights:** (i) The core idea of the *method #1* is to select a pair of fractions in each iteration and round them in opposite directions to ensure that their (weighted) sum remains unchanged. (ii) The *method #2* adopts a three-probability rounding scheme, which is designed to satisfy the condition that $\bar{y}_n^t$ and $\bar{z}_n^t$ are not simultaneously equal to 1. (iii) Both methods maintains adherence to instantaneous constraints. They also preserve the expected value of the objective function and ensure the boundedness of long-term constraint violations and achieve desirable economic properties in expectation. These three aspects will be elaborated in detail later.

### E. Payment Allocation Algorithm

Algorithm 4 presents the payment design in our auction mechanism. The auctioneer pays the winning bids according to Line 3, while the other bids receive no payment, as shown in Line 6. We use $\widetilde{x}_{in}^t(r_{in}^t, r_{-in}^t)$ to denote the fractional solution obtained in Line 3 of Algorithm 1, where $r_{in}^t$ is the bidding price reported by the bid $i$ of the edge $n$ and $r_{-in}^t$ is the bidding price reported by all the other bids. And we set $\varkappa_{in}^t$ as the upper bound of the integration, which captures the maximum payment the auctioneer made to the corresponding winning bids. Note that $\varkappa_{in}^t$ is derived from (8) and (9) using an optimization method [42] to locate the target point. The time complexity of Algorithm 4 is $\mathcal{O}(I^3 N^3 \epsilon \log(1/\nu))$, where $\epsilon$ is the number of segments while dividing the range of $\varkappa_{in}^t - r_{in}^t$ for calculating the integral numerically.

**Algorithm 4** Payment Allocation Algorithm

**Input:** Integers solutions $\bar{x}_{in}^t$ and fractions solutions $\widetilde{x}_{in}^t$
**Output:** Payments $p_{in}^t$
1: **for** $i \in \mathcal{I}_n$, $n \in \mathcal{N}$ **do**
2:   **if** $\bar{x}_{in}^t = 1$ **then**
3:     Set $p_{in}^t = r_{in}^t \widetilde{x}_{in}^t \left(r_{in}^t, r_{-in}^t\right) + \int_{r_{in}^t}^{\varkappa_{in}^t} \widetilde{x}_{in}^t \left(r, r_{-in}^t\right) dr$, where $\varkappa_{in}^t = \frac{\widetilde{x}_{in}^t}{\alpha} + \frac{\lambda_{in}^t \psi_{in}^t \varphi^{t+1}}{\alpha} - \theta_{in}^t$;// For winning bids
4:   **end if**
5:   **if** $\bar{x}_{in}^t = 0$ **then**
6:     Set $p_{in}^t = 0$;   // For other bids
7:   **end if**
8: **end for**

**Insights:** The design of Algorithm 4 is based on the sufficient and necessary conditions [43] for truthfulness and individually rationality, as formally defined in Section V-C. Note that the auction mechanism is actually transformed into a randomized one by Algorithm 3, and our payment allocation method is feasible since we ensure that the expectation of the estimated allocation $\bar{x}_{in}^t$ equals its actual value $\widetilde{x}_{in}^t$.

## V. Theoretical Analysis

We define and analyze three key performances in the following: (i) Regret and fit analysis for the sub-problems $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ (Section V-A); (ii) Asymptotic competitiveness analysis for the original problem $\mathcal{P}$ (Section V-B); (iii) Economic properties analysis for the auction mechanism (Section V-C). The structure of theoretical analysis is shown in Fig. 5.

### A. Regret and Fit Analysis

We denote the online solution returned by our algorithm, i.e., $\bar{\mathrm{Z}}_1^t = \{\bar{\boldsymbol{x}}^t, \widetilde{\boldsymbol{y}}^t, \widetilde{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ for $\mathcal{P}_1$, $\bar{\mathrm{Z}}_2^t = \{\bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ for $\mathcal{P}_2$ and $\bar{\mathrm{Z}}_3^t = \{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ for $\mathcal{P}_3$. Then we denote $\mathrm{Z}_{1,2,3}^{t\star}$ as the "one-shot" offline optimal solution at time slot $t$, i.e., breaking the problem into one-shot slices and solving each slice to its optimum individually. After that, we provide the definitions of regret and fit.

**Definition 1** (Regret)**.** *The regret defines the total discrepancy between the social cost obtained from online solution and the offline optimum. We define the regrets of $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ as*

$$Reg_{1,2,3}^T \triangleq \sum_{t=1}^T (\mathbb{E}[f_{-S}^t(\bar{\mathrm{Z}}_{1,2,3}^t)] - f_{-S}^t(\mathrm{Z}_{1,2,3}^{t\star})).$$

**Definition 2** (Fit)**.** *The fit defines the total violations of long-term constraints evaluated from the generated online solutions accordingly. We define the fits of $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ as*

$$Fit_{1,2,3}^T \triangleq \sum_{t=1}^T \|[g^t(\bar{\mathrm{Z}}_{1,2,3}^t)]^+\|.$$

**Theorem 2.** *The regrets and fits can be bounded by*

$$Reg_1^T \leq \mathcal{O}(T^{\varsigma_1}), Reg_2^T \leq \mathcal{O}(T^{\varsigma_2}), Reg_3^T \leq \mathcal{O}(T^{\varsigma_3}),$$
$$Fit_1^T \leq \mathcal{O}(T^{\sigma_1}), Fit_2^T \leq \mathcal{O}(T^{\sigma_2}), Fit_3^T \leq \mathcal{O}(T^{\sigma_3}),$$

*where* $\varsigma_1, \varsigma_2, \varsigma_3 \in (0,1)$ *and* $\sigma_1, \sigma_2, \sigma_3 \in (0,1)$.

*Proof.* See Appendix B in our supplemental material. □

**Proof Sketch:** (i) Use Lemma 1 to bound the integral regret and fit by their fractional relaxations. (ii) Use Lemma 2 to show the dual variables remain bounded. (iii) Step I sums the

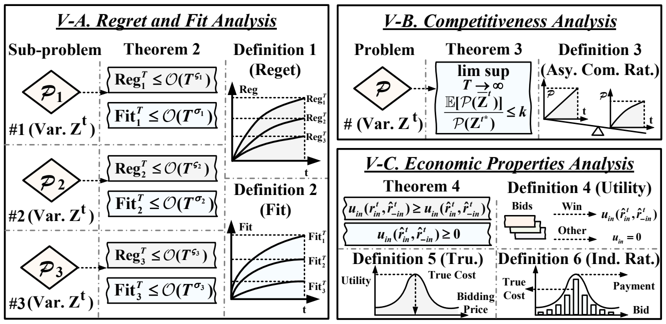



Fig. 5. Structure of theoretical analysis.

per-round inequalities to express $\widetilde{\text{Reg}}^T$ in terms of the cumulative variations of the fractional optimum and constraints. (iv) Step II bounds $\widetilde{\text{Fit}}^T$ via dual recursion. Finally, the appropriate choice of $\alpha$ and $\mu$ leads to sub-linear regret and fit.

**Insights:** (i) Sub-linear regret $\mathcal{O}(T^{\varsigma})$ means the time averaged per-slot social cost compared to offline optimum vanishes as time horizon $T$ grows. (ii) Sub-linear fit $\mathcal{O}(T^{\sigma})$ guarantees that the long-run average of the battery energy of each edge $n$ is asymptotically within its battery capacity constraint $[O_n, B_n]$. (iii) Different $\varsigma$ and $\sigma$ for $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$ reflect the varying complexity of coordinating coupled decisions, which can be verified in Fig. 8 and Fig. 9 in Section VI-B.

### B. Asymptotic Competitiveness Analysis

The asymptotic competitive ratio quantifies the limiting multiplicative difference between the expected objective value achieved by an online algorithm and the offline optimal solution as the time horizon grows. Generally, this asymptotic competitive ratio is expected to converge to a constant rather than varying with time. The formal definition is given below.

**Definition 3** (Asymptotic Competitive Ratio)**.** *The asymptotic competitive ratio $k$ of an online problem $\mathcal{P}$ satisfies $\limsup_{T\to\infty} \frac{\mathcal{P}(\bar{Z}^t)}{\mathcal{P}(Z^{t*})} \leq k$, where $\bar{Z}^t$ represents the online solution and $Z^{t*}$ denotes the offline optimal solution considering the optimal solution of $\mathcal{P}$ over the whole time horizon $T$, instead of breaking $\mathcal{P}$ into "one-shot" slices at each time slot $t$.*

**Theorem 3.** *Our online problem $\mathcal{P}$ achieves*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathcal{P}(\bar{Z}^t)]}{\mathcal{P}(Z^{t*})} \leq k,$$

*where $k$ is the constant independent of time. And we note that $\bar{Z}^t = \{\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t, \boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$. Specifically, $\bar{\boldsymbol{x}}^t$ is from $\bar{Z}_1^t$, $\{\bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t\}$ are from $\bar{Z}_2^t$ and $\{\boldsymbol{u}^t, \boldsymbol{v}^t, \boldsymbol{w}^t, \boldsymbol{q}^t\}$ are from $\bar{Z}_3^t$.*

*Proof.* See Appendix C in our supplemental material. ☐

**Proof Sketch:** (i) Step I relates the switching cost to the non-switching cost via constants $\beta$ and $\vartheta$, yielding an overall bound $k_1$ on the ratio. (ii) Step II bounds the total cost of the fractional optimal solution by a constant $k$ times the offline optimum $\mathcal{P}(Z^{t\star})$. (iii) Step III applies the sub-linear regret of Theorem 2 to obtain $\mathbb{E}[\mathcal{P}(\bar{Z}^t)] \leq k\mathcal{P}(Z^{t*}) + \mathcal{O}(T^{\varsigma_3})$. (iv) Step IV makes the additive term $\mathcal{O}(T^{\varsigma_3})$ explicit under a non-degeneracy assumption ($f_{-S}^t(\cdot) \geq c_{\min}$), giving the ratio $k$.

**Assumption:** Theorem 3 relies on the *non-degeneracy* assumption that there exists a constant $c_{\min} > 0$ such that for every time slot $t$, the per-slot non-switching cost satisfies $f_{-S}^t(\cdot) \geq c_{\min}$. It means that for the multi-edge-AI system, in each time slot there is at least one edge AI service with positive workload and energy price, incurring a basic cost, which is largely reasonable and acceptable in reality.

**Insights:** An asymptotic constant competitive ratio $k$ ensures that for sufficiently long time horizons, the expected online total social cost is at most $k$ times the offline optimum, with per-slot cost converging to the offline optimum.

### C. Economic Properties Analysis

Under the assumption that $\psi_{in}^t$ and $\theta_{in}^t$ are verifiable (or reported truthfully), we put forward three economic attributes with respect to the bidding price $r_{in}^t$. The utility of a bid reflects the gap between the payment and the true cost, which is the foundation for later definitions. The truthfulness ensures the bidders bid their true valuations without misstating, and the individual rationality means no bidder incurs negative outcome whether it wins or not. The formal definitions are as follows.

**Definition 4** (Utility)**.** *The utility of the bid $i$ for the edge $n$ at time slot $t$ is*

$$u_{in}(r_{in}^t, r_{-in}^t) = \begin{cases} p_{in}^t(\hat{r}_{in}^t, \hat{r}_{-in}^t) - r_{in}^t \mathbb{E}[\bar{x}_{in}^t(\hat{r}_{in}^t, \hat{r}_{-in}^t)], \\ \textit{if } \bar{x}_{in}^t = 1, \\ 0, \textit{otherwise}. \end{cases}$$

*where $\hat{r}_{in}^t$ is the bidding price of the bid $i$ for the edge $n$, $\hat{r}_{-in}^t$ refers to the bidding prices of all the other bids except the bid $i$ and $r_{in}^t$ is the true cost of the bid $i$.*

**Definition 5** (Truthfulness)**.** *A randomized auction is truthful in expectation if each bid maximizes its expected utility by bidding its true cost, i.e., $u_{in}(r_{in}^t, \hat{r}_{-in}^t) \geq u_{in}(\hat{r}_{in}^t, \hat{r}_{-in}^t)$.*

**Definition 6** (Individual Rationality)**.** *A randomized auction is individually rational in expectation if each bid always has a non-negative utility, i.e., $u_{in}(\hat{r}_{in}^t, \hat{r}_{-in}^t) \geq 0$.*

**Theorem 4.** *Our auction mechanism achieves both truthfulness and individual rationality in the meantime.*

*Proof.* See Appendix D in our supplemental material. ☐

**Proof Sketch:** (i) Propose the first condition for truthfulness holds from comparing the fractional optimal solution under bids $r_{in}^t \geq \dot{r}_{in}^t$, yielding $\mathbb{E}[\bar{x}_{in}^t] \leq \mathbb{E}[\bar{\dot{x}}_{in}^t]$. (ii) Propose the second condition for truthfulness holds since the expected allocation integrates to at most the maximum payment $\varkappa_{in}^t$. (iii) Propose the individual rationality follows directly from the payment rule, ensuring non-negative expected utility.

**Insights:** (i) The utility definition separates the actual payment $p_{in}^t$ from the expected true cost $r_{in}^t\mathbb{E}[\bar{x}_{in}^t]$, reflecting that a bidder's gain is measured against its private valuation rather than its declared bid $\hat{r}_{in}^t$. (ii) The truthfulness indicates that bidding true cost $r_{in}^t$ is a dominant strategy for the bidders, i.e., no bidder benefits from reporting $\hat{r}_{in}^t \neq r_{in}^t$. (iii) The individual rationality guarantees that the payment $p_{in}^t$ is always at least the expected true cost $r_{in}^t\mathbb{E}[\bar{x}_{in}^t]$ for any winning bids, ensuring no participant suffers a loss in the auctions.

## VI. Experimental Evaluations

### A. Experiment Settings

**Edge infrastructure and VPP:** The simulations are conducted based on the widely-used Australian EUA dataset [21] containing 125 edge nodes and 816 mobile users in the Melbourne central business district (CBD) area. We consider 20 edge AI service providers [10] based on the mobile user density in the CBD [22]. The battery capacity of each edge is set to support the peak power demand for 12 hours [24]. And we consider lithium-ion batteries as edge batteries, where the charging efficiency $\eta_c$ is set as 99.9% and the discharging efficiency $\eta_d$ is set as 85% [1]. The battery degradation parameter $d_n$ is set as $8.3 \times 10^{-5}$ \$/Wh [23]. Moreover, we set the VPP signal $\{D^t, s^t_+\}$ and the energy purchasing price $s^t_-$ from the power grid using the real-world VPP instance of South Australia from October 1, 2021 to October 4, 2021 [31].

**Edge AI Services:** As for edge AI service requests, We consider multiple AI models (computer vision and natural language processing) including YOLOv4 [25], DeepLabv3 [26], BERT [27], VGG16 [28], and Inceptionv3 [29] to handle different edge AI inference requests. In addition, we set the model inference error rate within 0.1∼0.3 [44] and the energy efficiency $\rho^t_{in}$ of the edge workload within $10^{-4} \sim 5 \times 10^{-2}$ Wh per inference request based on the number of users within the coverage area of each edge node [45]. For the processing procedure, we consider $|\mathcal{T}| = 300$ consecutive time slots, where each time slot equals 15 minutes based on the processing time of edge inference requests [5], [46], [47]. This time granularity is consistent with multiple VPP energy trading scenarios [48], [49], [50]. Each edge has to process 0∼8000 inference requests at each time slot $t$ [51]. Furthermore, the number of AI inference requests that arrive at each edge is set in proportion to the dynamic job arrival trace of Google clusters [30]. Then we expand these requests in different proportions to a reasonable range according to the actual situation of each type in the real world. Besides, we consider up to 1200 bidders in the system while setting the model switching cost within 10∼100s [5]. For bidding price, we set it within 0∼0.275\$ by multiplying the energy saved in the bid and the energy price purchasing from the power grid.

**Auction-based Baselines:** We retain the auction incentive mechanism and each edge AI service provider $i$ (bidder) on edge $n$ can voluntarily propose a *truthful bid* $\mathcal{V}^t_{in} \triangleq \{r^t_{in}, \psi^t_{in}, \theta^t_{in}\}$ at each time slot $t$. Then the edge infrastructure provider (auctioneer) chooses the winning bids by decision variable $x^t_{in}$. For this, we have 5 baselines: (i) `ADES`: The approach proposed in this paper. (ii) `Offline`: The offline optimal solution that solves the original problem $\mathcal{P}$ as a whole across the time horizon $T$ via the Gurobi [52] optimization solver with all inputs provided in advance. (iii) `Greedy`: The bids with the minimum bidding price and the inference error rate are received. (iv) `MPUTA` [53]: A state-of-the-art approach that maximizes the total social welfare in the system. It utilizes an interior point method to find the fractional solution and a randomized rounding technique to obtain the integral solution. (v) `IRSM` [54]: A state-of-the-art approach that minimizes the total consumption via online scheduling. It employs a Lagrange dual technique to find the fractional solution and a Poisson rounding scheme to obtain the integral solution.

TABLE IV
Types of Weight Settings

| Types | Weight Settings of Different Units | | |
|---|---|---|---|
| | Dollar (\$) | Second (s) | Percentage (%) |
| #0: non-oriented | $w_1 = 1$ | $w_2 = 1$ | $w_3 = 1$ |
| #1: \$-oriented | $w_1 \times 5$ | $w_2$ | $w_3$ |
| #2: s-oriented | $w_1$ | $w_2 \times 5$ | $w_3$ |
| #3: %-oriented | $w_1$ | $w_2$ | $w_3 \times 5$ |
| #4: \$.s-oriented | $w_1$ | $w_2$ | $w_3 \times 0.2$ |
| #5: \$.%-oriented | $w_1$ | $w_2 \times 0.2$ | $w_3$ |
| #6: %.s-oriented | $w_1 \times 0.2$ | $w_1$ | $w_3$ |

**Non-auction Baselines:** We remove the auction and each edge AI service provider $i$ on edge $n$ can voluntarily maintain a *model set* $\mathcal{M}^t_{in} \triangleq \{1, \ldots, M^t_{in}\}$ at each time slot $t$ (add or delete models at any time). For each model $M^t_{in}$, it has a parameter set $\{A^t_{in}, B^t_{in}, C^t_{in}\}$. $A^t_{in}$ is "the valuation of the model", $B^t_{in}$ is "the deployment cost of the model" and $C^t_{in}$ is "the energy consumed to serve one request". Here, $A^t_{in}$ is still set as "the energy savings achieved by the proposed model multiplied by the price of energy purchased from the grid" (which is same as $r^t_{in}$ in auction-based baseline). Then the edge infrastructure provider chooses an appropriate model for each service $i$ on edge $n$. Based on the above, we employ two baselines: (i) `DRL` [55]: Deep reinforcement learning approach that applies hybrid action algorithm to address the problem of hybrid continuous and discrete action space. We treat "model selection" and "battery charging/discharging/resting" as the discrete action space; treat "energy amounts of charging/discharging" as the continuous action space. (ii) `LYA` [56]: Lyapunov optimization algorithm that utilizes drift-plus-penalty and virtual queue technique. We regard the battery charging and discharging process as a virtual queue, and incorporate the costs related to the model and the costs related to the battery level into the drift-plus-penalty term. An intuitive comparison between auction-based and non-auction baselines is shown in Appendix E in our supplemental material.

**Weight Settings:** The objective function (1) combines terms with 3 units: model switching cost $C_s$ (measured in seconds), inference error rate $C_{\%}$ (measured in percentages) and other costs $C_{\$}$ (measured in dollars). Based on the studies [57], [58], we utilize the well-known weighted-sum approach to convert the multi-objective optimization to the single-objective optimization. That is, we normalize each original value $C_{\sim}$ ($\sim$ represents $s$ or $\%$ or \$) to the normalized value $C^{nor}_{\sim} \in [0, 1]$ using min–max normalization which is defined as $C^{nor}_{\sim} = \frac{C_{\sim} - C^{min}_{\sim}}{C^{max}_{\sim} - C^{min}_{\sim}}$. Here, $C^{min}_{\sim}$ and $C^{max}_{\sim}$ represent the minimum and the maximum value of the original value $C_{\sim}$, which can be borrowed from the numerical range of all historical input datasets. Then the objective function (1) can be written as $\mathcal{P} = w_1 C^{nor}_{\$} + w_2 C^{nor}_{s} + w_3 C^{nor}_{\%}$, where $w_1$, $w_2$ and $w_3$ are the weights, and the edge infrastructure provider can adjust these weights based on its own needs. In TABLE IV, we present seven types of weight settings, including one *non-oriented type* #0 (evaluation in Section VI-B) and six *extreme oriented types* #1∼#6 (evaluation in Section VI-C).

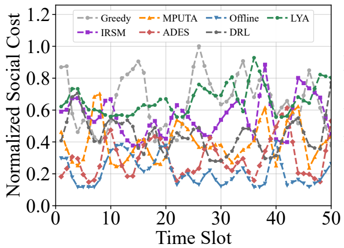


Fig. 6. Real-time social cost.

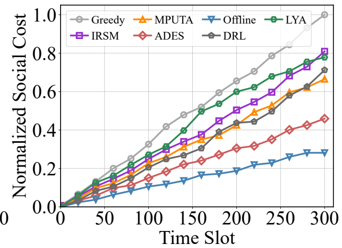


Fig. 7. Cumulative social cost.

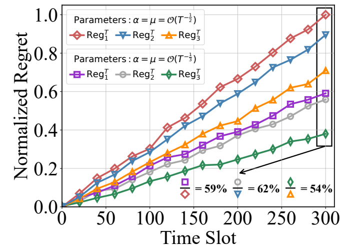


Fig. 8. Dynamic regret.

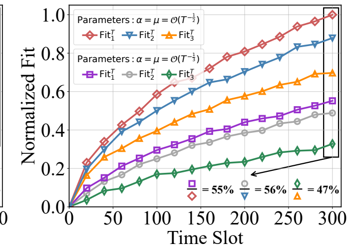


Fig. 9. Dynamic fit.

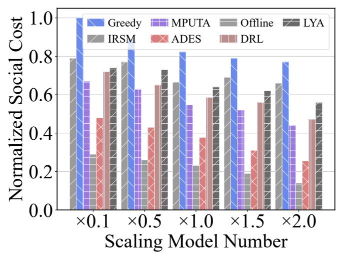


Fig. 10. Scaling model number.

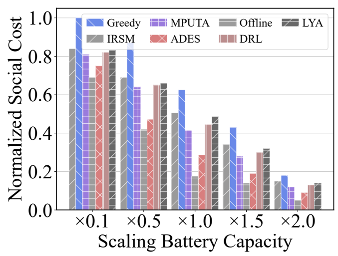


Fig. 11. Scaling battery capacity.

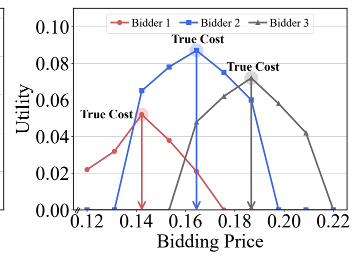


Fig. 12. Truthfulness.

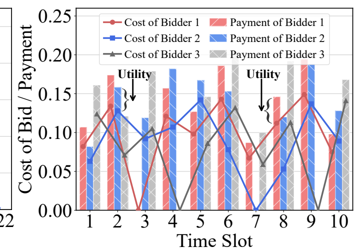


Fig. 13. Individual rationality.

For instance, `non-oriented` means no unit has an oriented weight; `$-oriented` means the unit $ has high weight, etc.

**Algorithm Parameters:** Parameter $\beta$ in Algorithm 1 is set as 0.5 based on the cost-minimizing approach [59] and the experimental verification shown in Fig. 18 and Fig. 25. The step size and constant in Algorithm 2 are set as $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{3}}) \approx 0.15$ based on the proof of Theorem 2 in Section V-A and experimental evidence shown in Fig. 8 and Fig. 9.

**Testbed Settings:** Our evaluations in Section VI contain about 12,000 lines of Python codes. We conduct it on a desktop with a 2.2-GHz Intel Core i9 CPU and 32-GB memory.

## B. Basic Evaluations

We conduct basic evaluations under *non-oriented type* (#0) of weight setting to demonstrate the performance.

**Social Cost:** Fig. 6 and Fig. 7 show the normalized real-time social cost and cumulative social cost incurred by different approaches. `ADES` outperforms others with an average cost reduction of 54% compared to `Greedy`, 43% to `MPUTA`, 31% to `IRSM`, 36% to `DRL`, and 41% to `LYA`. We can further observe that the cumulative social cost of `ADES` increases at a slower rate compared to them. This is due to these baselines used for edge computing task scheduling neglect the specific characteristics of AI inference, such as inference error rates, and also overlook the carefully designed conditions for model switching, as well as the long-term constraints of the edge batteries. Consequently, these methods may lead to higher inference error rates, model switching costs, and the expenses of energy. Compared to the `Offline`, we further calculate the competitive ratio of `ADES` and find it to be 1.63.

**Regret and Fit:** Fig. 8 and Fig. 9 visualize the regret and the fit for our proposed approach `ADES`, respectively under pre-determined step size $\mu$ and constant $\alpha$. The growth of the regret and that of the fit are both sub-linear, which is consistent with our theoretical analysis. We can also observe that the regrets and fits of $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ decrease successively. This results from we substitute $\bar{\boldsymbol{x}}^t$ as the input into $\mathcal{P}_2$ and $\{\bar{\boldsymbol{x}}^t, \bar{\boldsymbol{y}}^t, \bar{\boldsymbol{z}}^t\}$ as the inputs into $\mathcal{P}_3$, thereby reducing the computation. Compared to $\mathcal{P}_1$, the final regret of $\mathcal{P}_2$ decreases by approximately 10%, and that of $\mathcal{P}_3$ decreases by approximately 29% while $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{3}})$. Compared to $\mathcal{P}_1$, the final fit of $\mathcal{P}_2$ decreases by approximately 15%, and that of $\mathcal{P}_3$ decreases by approximately 41% while $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{3}})$. Furthermore, we can observe that the regret and the fit are greater while $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{3}})$ compared to $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{2}})$, which is also consistent with our theoretical analysis.

**Scaling Model Number:** Fig. 10 reveals the impact of model number. Specifically, as new models are added to the candidate model pool, the overall social cost continues to decline. A larger and more diverse set of models provides a wider range of options, offering greater flexibility in balancing cost and accuracy. This expanded selection allows for more refined control decisions, ultimately leading to lower optimized objective values. This increased diversity in available models enhances the ability to identify solutions that better minimize social costs, demonstrating the value of incorporating varied model candidates into the decision-making process.

**Scaling Battery Capacity:** Fig. 11 indicates the impact of battery capacity. It can be observed that increasing battery capacity can effectively reduce social cost. This is because the batteries serve as an energy buffer, capable of storing energy during low-price periods to hedge against future price surges while ensuring a stable energy supply to both the grid and edge workloads. A larger battery capacity allows for greater energy procurement from the grid when energy prices are low, thereby significantly reducing or even eliminating the need to purchase energy during subsequent high-price periods.

**Truthfulness:** Fig. 12 validates the truthfulness of our auction mechanism. We randomly select three bidders as examples. As shown, if a bidder reports its true cost, it will achieve the highest utility. If the bid is inconsistent with the true cost, the utility will drop significantly. This aligns with our theoretical analysis, indicating that no bidder can gain higher utility by misreporting the bidding price that differ from the true cost. This truthfulness encourages bidders to submit

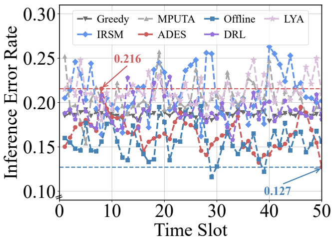



Fig. 14. Real-time inf. error rate.

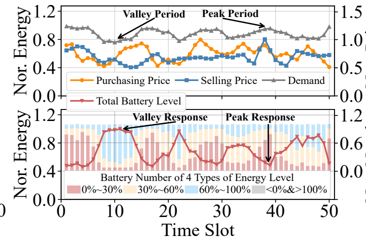



Fig. 15. Real-time VPP response.

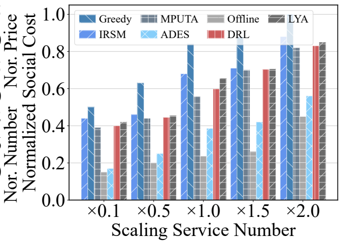



Fig. 16. Scaling service.

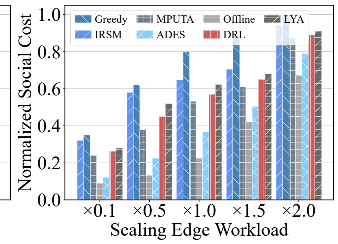



Fig. 17. Scaling edge workload.

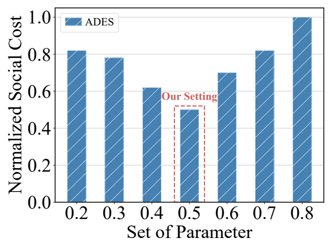



Fig. 18. Impact of parameter $\beta$.

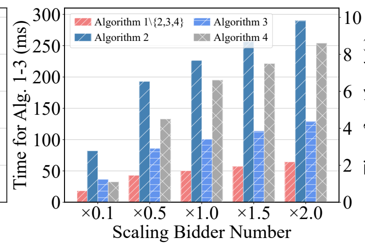



Fig. 19. Execution time.

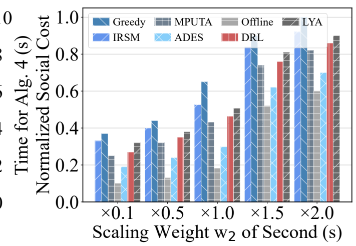



Fig. 20. Impact of second (s).

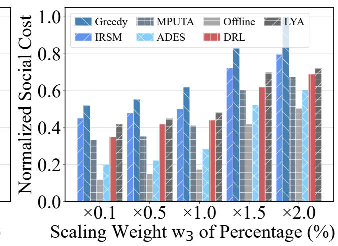



Fig. 21. Impact of percentage (%).

honest bids, thereby enhancing the credibility of the auction market and ensuring the fairness of the auction mechanism.

**Individual Rationality:** Fig. 13 reflects the individual rationality from the perspective of non-negative utility. The three bidders shown here are randomly selected from our auction. It can be observed that regardless of the auction outcome and variations in the bidding price (which equals the true cost due to truthfulness) and payment vary, each bid receives a payment no less than the corresponding bid, thereby ensuring non-negative utility. And non-winning bidders receive no payment. These results align with the definition of individual rationality and thus guarantees rational participation for all bidders.

**Real-time Inference Error Rate:** Fig. 14 displays the averaged inference error rate of all the models selected to process corresponding inference requests at the first 50 time slots. And the inference error rate is kept within an acceptable range 12.7%$\sim$21.6% since we carefully design the conditions for model switching and employ online learning algorithm. Then we further calculate the average inference error rate of `ADES` is 0.168 and find it achieves closest to the `Offline` optimum, outperforming `Greedy` by 13%, `MPUTA` by 20%, `IRSM` by 23%, `DRL` by 15%, and `LYA` by 22%.

**Real-time VPP Response:** Fig. 15 reflects the real-time VPP response of `ADES`. (i) From the first sub-figure, it can be observed that the energy prices tend to be higher during "peak period" of the energy demand and lower during "valley period". (ii) From the second sub-figure, it can be found that the total battery level changes in response to the energy demand and prices of the power grid. While the energy demand of the power grid is at "peak period", the VPP (edge infrastructure) dispatches more batteries to supply energy to the grid, causing its battery level to decrease sharply. Conversely, when the energy demand of the grid is at "valley period", the VPP allows more batteries to draw energy from the grid, leading to a rapid increase of the battery level. Furthermore, the battery number of "$e^t \leq 0\%$" and "$e^t \geq 100\%$" always keep 0. In other words, the battery level $e_n^t$ always remain within 0%$\sim$100%.

**Scaling Service:** Fig. 16 presents the social cost comparison for all the implemented algorithms as the services increase. `ADES` performs at least 31% better than other baselines and it is also the closest to the offline optimum. We can further observe that the more services the system has, the more social cost it will be. This is on account of the edge service providers needs to switch more models, in order to dispatch more AI inference services to change the energy consumption. However, the social cost does not increase significantly with the growing number of services. This can be attributed to edge infrastructure can still schedule energy to meet the demand of the power grid through battery charging and discharging. With such dual safeguards, i.e. battery buffering and model switching, the edge infrastructure can reliably work as a VPP.

**Scaling Edge Workload:** Fig. 17 exhibits the impact of edge workload on social cost. As the number of edge workloads increases, which means each AI service provider needs to handle more inference requests, the social cost not only rises but also makes the edge scheduling more complex. Furthermore, we observe that although other baselines (except `Offline`) focus on the flexibility and efficiency of the edge scheduling, `ADES` performs at least 9% superior than them.

**Impact of parameter $\beta$:** Fig. 18 demonstrates our experimental test of parameter $\beta$ in Algorithm 1. As described in the algorithm, the switching operation is triggered while $\beta$ times the cumulative non-switching cost from the last switching to the previous time slot exceeds the current switching cost. If $\beta$ is too large, the switching condition is easily satisfied, leading to frequent model switching. Conversely, if $\beta$ is too small, the switching condition is less likely to be met, causing a single model to be used for an extended period. Both scenarios increase the social cost. Therefore, we conducted experiments with different values of $\beta$ to minimize the social cost as much as possible. As shown in the figure, while $\beta$ equals 0.5, the resulting social cost is relatively low, indicating that this value effectively balances switching cost and non-switching cost.

**Execution Time:** Fig. 19 details the execution time of our approach. The results show that `ADES` finishes in hundreds of milliseconds, except the payment allocation which consumes

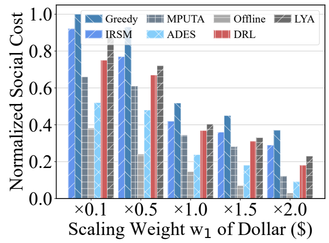



Fig. 22. Impact of dollar ($).

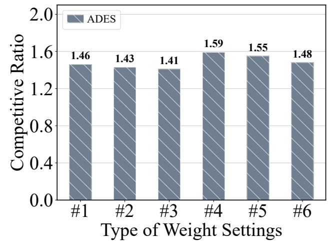



Fig. 23. Com. ratio for 6 types.

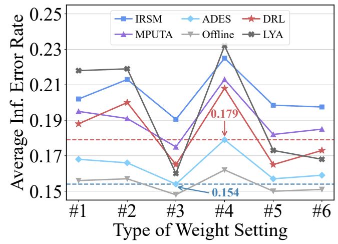



Fig. 24. Error rate for 6 types.

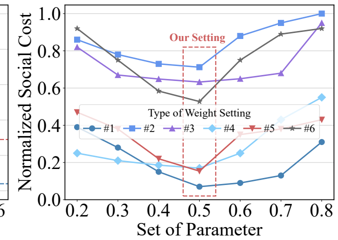
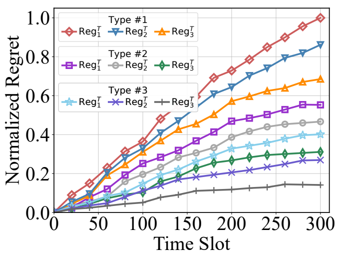
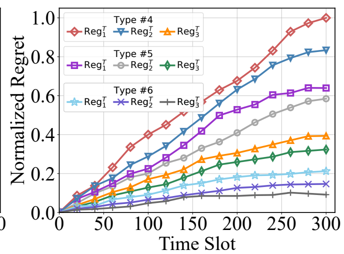


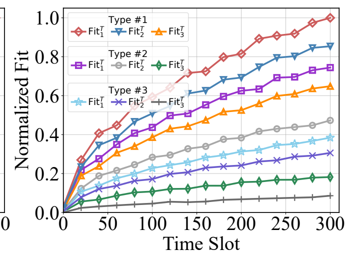

Fig. 25. Parameter $\beta$ for 6 types.



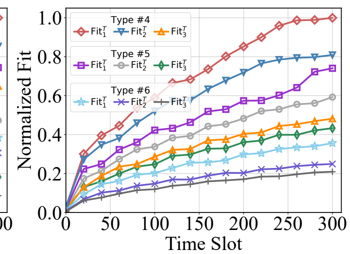

(a) Regret for types #1∼#3.




(b) Regret for types #4∼#6.




(c) Fit for types #1∼#3.




(d) Fit for types #4∼#6.

Fig. 26. Dynamic regret and dynamic fit for 6 types of weight settings.

seconds. Meanwhile, we can observe that the execution time of each algorithm experiences only a slight increase as the number of bidders grows. Overall, the execution time of `ADES` is significantly less than the length of a single time slot.

### C. Sensitivity Analyses of Weight Settings

We evaluate the robustness of main conclusions under 6 *extreme oriented types* (#1∼#6) of weight settings.

**Social Cost:** (i) Impact of second (s) and percentage (%): Fig. 20 and Fig. 21 depict the impact of the social cost incurred by the weight of model switching cost and inference error rate respectively. We can find that the social cost increases with the higher weight of the switching cost and the inference error rate. This is owing to the edge infrastructure needs to bear higher social cost while switching energy-efficient models to schedule energy so that the energy demand of the power grid could be met. (ii) Impact of dollar ($): Fig. 22 shows the social cost resulting from changing the weight of other costs. We can observe that the social cost decreases as the weight increases. This is because the increasing weight of other costs weakens the weight of model switching overhead (i.e., model switching cost and inference error rate), allowing the edge infrastructure to switch models frequently to meet the demands of the grid. (iii) Performance: `ADES` is the closest to `Offline` and consistently outperforms other baselines by at least 11%.

**Competitive Ratio:** Fig. 23 presents the competitive ratios under 6 extreme weight settings. It can be observed that these competitive ratios are smaller than the competitive ratio of 1.63 in the basic setting (Type #0) in Section VI-B. This is because when the edge infrastructure sets a certain weight too high or too low, it greatly reduces the proportion of other costs in the total social cost, thereby weakening the cost incurred by the operation of other algorithms. In other words, under extreme weight settings, it can be seen as a reduction in the number of algorithms, which lowers the competitive ratio.

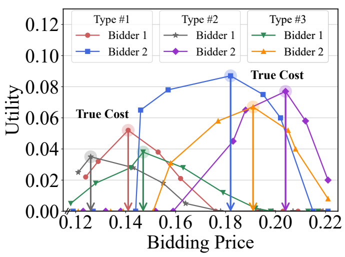
**Average Inference Error Rate:** Fig. 24 shows the average inference error rates under 6 extreme weight settings. It can be observed: (i) Type #3 achieves the lowest error rate because it excessively increases the weight of error rate, while Type #4 has the highest error rate as it excessively decreases the weight of error rate. (ii) Under each setting, the average error rate of `ADES` can reach up to 0.179 at its highest, and can be as low as 0.154 at its lowest. And `ADES` consistently remains closest to `Offline` and outperforms the other baselines.

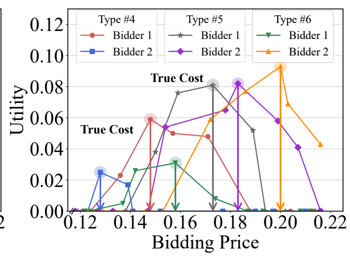
**Impact of Parameter $\beta$:** Fig. 25 illustrates the impact of parameter $\beta$ on the social cost under 6 extreme weight settings. As shown in the figure, setting $\beta = 0.5$ minimizes the social cost, thereby achieving an efficient balance between the model switching cost and the non-switching cost in Algorithm 1.

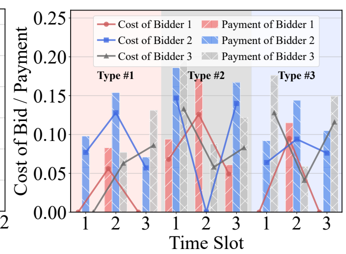
**Regret and Fit:** Fig. 26 reflects the regret and fit under 6 extreme weight settings. We have the following discoveries: (i) Although extreme weight settings affect the magnitude of regret and fit, they still remain sub-linear, which is consistent with the theoretical analysis in Section V-A. (ii) Among the 3 units of social cost, the unit corresponding to $ has the highest computational complexity. Therefore, increasing the weight of $ leads to the higher regret and fit. In comparison, the computational complexity associated with s and % decreases sequentially, and the regret and fit vary accordingly.

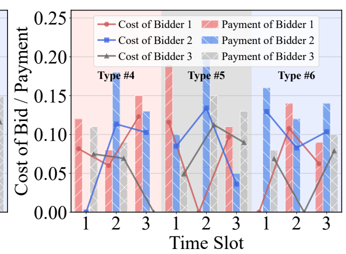
**Economic Properties:** (i) Truthfulness: Fig. 27(a) and Fig. 27(b) present the verification of truthfulness under 6 extreme weight settings for auctions. For each setting, we randomly select two bidders and assign them their true bidding costs along with a range of false costs. The results show that bidders achieve the maximum utility only when bidding truthfully, showing the auction remains robustly truthful. (ii) Individual Rationality: Fig. 27(c) and Fig. 27(d) demonstrate the verification of individual rationality under the same 6 extreme weight settings. For each setting, we randomly select three bidders and compare their bidding costs with their payments. The results indicate that our auction mechanism does not impose negative utility on bidders, thereby ensuring individual rationality.

**Scalability:** We evaluate the scalability under 6 extreme weight settings: (i) Fig. 28(a) shows the performance under



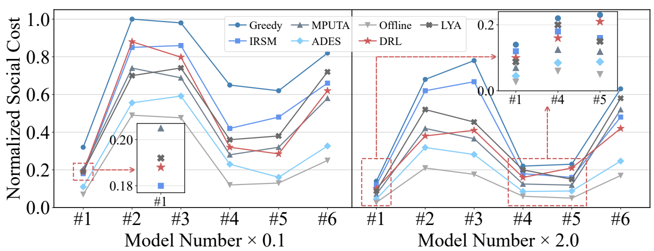



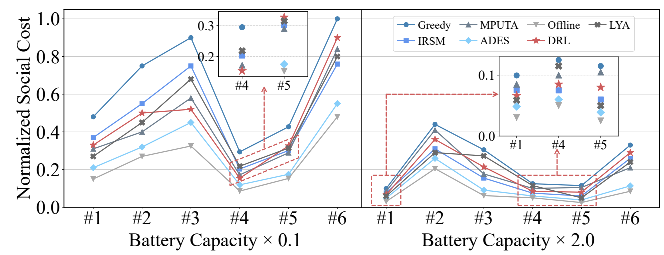




(a) Truthfulness for types #1∼#3. (b) Truthfulness for types #4∼#6. (c) Ind. rat. for types #1∼#3. (d) Ind. rat. for types #4∼#6.

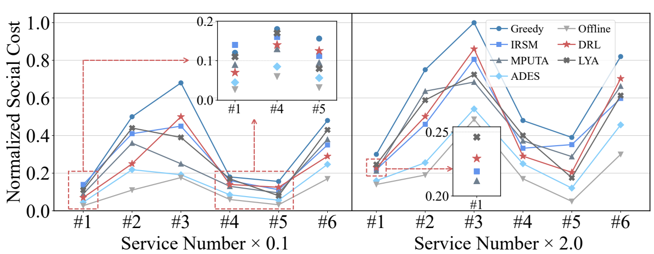
Fig. 27. Economic properties (truthfulness and individual rationality) for 6 types of weight settings.



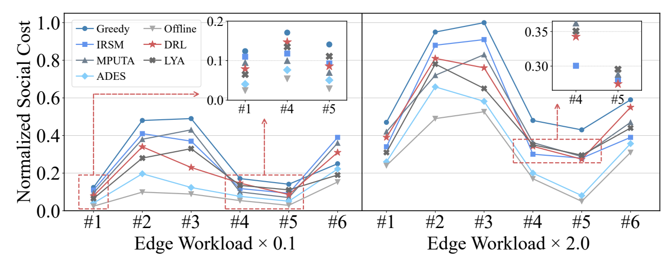

(a) Scaling model number for 6 types.




(b) Scaling battery capacity for 6 types.




(c) Scaling service number for 6 types.




(d) Scaling edge workload for 6 types.

Fig. 28. Scalability for 6 types of weight settings.

the scalability of model number; (ii) Fig. 28(b) illustrates the performance under the scalability of battery capacity; (iii) Fig. 28(c) displays the performance under the scalability of service number; (iv) Fig. 28(d) exhibits the performance under the scalability of edge workloads. We have the following observations: (i) Under different experimental settings, `ADES` consistently remains closest to `Offline` and outperforms the other baselines. (ii) Types #2, #3 and #6 yield higher social costs. This is consistent with our analysis of Figs. 20, 21 and 22: while the weights of the costs associated with model switching (switching cost and inference error rate) are larger, the social cost increases; conversely, the social cost decreases.

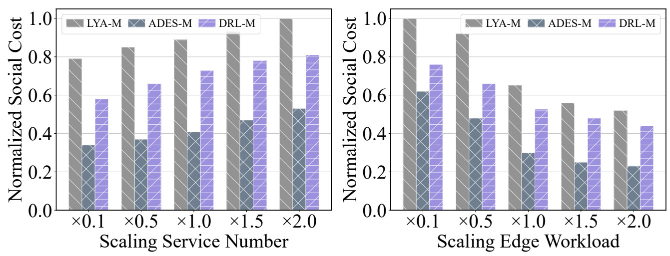
### *D. Isolation Studies*

We conduct isolation studies on the three components (incentive mechanism, model switching and battery scheduling) of the entire management process by modifying only one or two components at a time while keeping the others the same as in our approach `ADES`, and conduct multiple evaluations to demonstrate isolation performance. For instance, based on auction mechanism, `IRSM-M` is an approach for model switching in `IRSM`; based on non-auction mechanism, `DRL-B` is an approach for battery control in `DRL`, etc. See Appendix F in our supplemental material for *Detailed Isolation Setting*.

**Isolation Study for Incentive Mechanism:** We modify the incentive mechanism (coupled with model switching) while battery scheduling is resolved by `ADES-B`. `ADES-M` is based on auction mechanism, while `LYA-M` and `DRL-M` are based on non-auction mechanism. Then we have the evaluations: (i) Fig. 29(a) visualizes that under different numbers of edge AI services, `ADES-M` achieves at least a 35% lower social cost than other methods. (ii) Fig. 29(b) indicates that under different numbers of edge workloads, `ADES-M` achieves a social cost at least 18% lower than other methods.

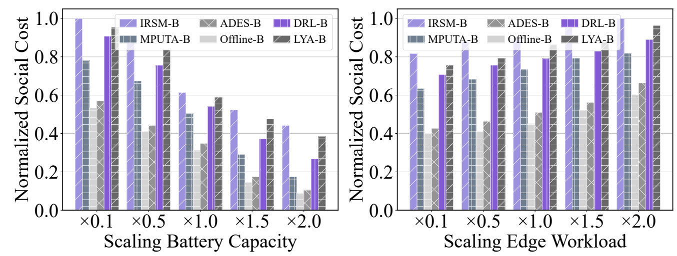
**Isolation Study for Model Switching:** We change the model switching strategy (coupled with incentive mechanism) while battery scheduling is resolved by `ADES-B`. `LYA-M` and `DRL-M` are based on non-auction mechanism while others are based on auction mechanism. Then we have the observations: (i) Fig. 30(a) shows that the average inference error rate of `ADES-M` is the closest to `Offline-M` and lower than that of other baselines. (ii) Fig. 30(b) displays that `ADES-M` is efficient while scaling model number. Compared to others (except `Offline-M`), it reduces social cost by at least 20%.

**Isolation Study for Battery Scheduling:** We only change the battery scheduling while model switching is resolved by `ADES-M` based on auction mechanism. Then we have the following assessments: (i) Fig. 31(a) demonstrates that under different battery capacities, `ADES-B` consistently achieves



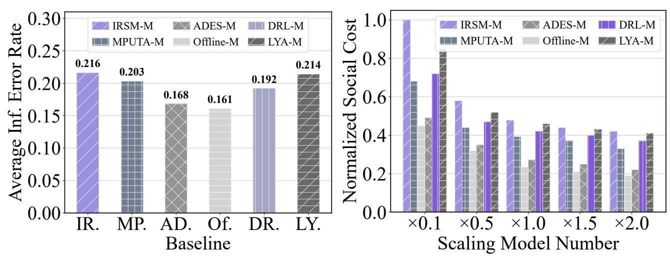

(a) Scaling service number. (b) Scaling edge workload.

Fig. 29. Isolation study for incentive mechanism.



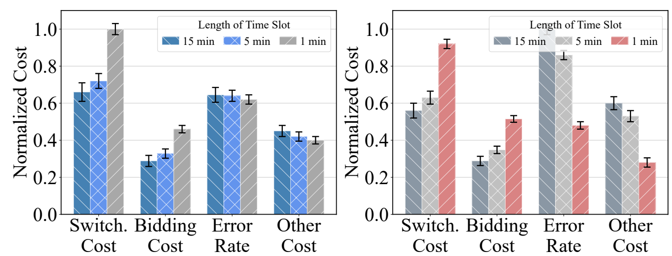

(a) Scaling battery capacity. (b) Scaling edge workload.

Fig. 31. Isolation study for battery scheduling.




(a) Average inference error rate. (b) Scaling model number.

Fig. 30. Isolation study for model switching.




(a) Low volatility. (b) High volatility.

Fig. 32. 4 tpyes of costs in social cost under low and high volatility.

lower social cost than other baselines (except `Offline-B`), with at least a 27% reduction. (ii) Fig. 31(b) reflects the performance of `ADES-B` under different scaling levels of edge workloads, it reduces the social cost by at least 19% compared to other baselines and remains the closest to `Offline-B`.

Compared with the experiments in Section VI-B (which compare `ADES` with other baselines as a whole), we find that under isolation settings, all baselines are closer to `Offline`. This indicates the effectiveness of each component in `ADES`.

### *E. Sensitivity Analysis of Intra-slot Variations*

We explain how variations within time slot are captured by increasing *"decision frequency"* through sensitivity analysis.

**Settings:** We randomly generate 1-minute data from the original 15-minute dataset under low volatility ($\pm 10\%$ across minutes) and high volatility ($\pm 50\%$ across minutes), preserving the 15-minute totals for inference requests and VPP demand while adding fluctuations to energy selling and purchasing prices. The `ADES` then runs over 75 hours at three decision granularities: (i) for 15-minute decisions, it uses the original constant 15-minute data; (ii) for 5-minute decisions, it sums the five 1-minute sub-slots for requests and demand and uses a constant 5-minute energy price; (iii) for 1-minute decisions, it directly uses the generated per-minute values.

**Results:** Averaged over 10 random seeds, the results (normalized costs with standard deviation bars) show two distinct patterns. We now explain for average values: (i) As shown in Fig. 32(a), under low volatility, finer decision granularities (5-min, 1-min) increase switching and bidding costs while only marginally reducing error rate and other cost. Thus a coarser granularity (15-min) is optimal. (ii) As shown in Fig. 32(b), under high volatility, finer granularities substantially reduce error rate and other costs, outweighing the increases in switching and bidding costs, so the finest granularity (1-min) becomes optimal. (iii) We address this intra-slot variations by increasing *"decision frequency"* while adding computational load. And the execution time in Fig. 19 remains acceptable relative to the slot length. See Appendix G in our supplemental material for insights and how to choose slot length.

## VII. Discussion

**Information Availability in Real Deployments:** For each model, the edge AI service provider can truthfully report (i) switching cost $m_{in}^t$ and $l_{in}^t$, profiled offline (deployment and loading time) [42]; (ii) inference error rate $\theta_{in}^t$, obtained online from validation [5]; (iii) energy efficiency $\rho_{in}^t$ and $\psi_{in}^t$, derived from hardware-specific power measurements [60]. Moreover, suitable lower-energy alternative models are readily available in modern edge AI [61], [62], where model compression techniques such as pruning, quantization, knowledge distillation produce a spectrum of accuracy–energy trade-offs. In practice, the edge AI service providers can maintain multiple model variants for switching purposes to meet different requirements.

**Bidder Behavior and Feasibility:** Our auction mechanism requires bidders to behave strategically through their bidding price $r_{in}^t$ while truthfully disclosing energy efficiency $\psi_{in}^t$ and inference error rate $\theta_{in}^t$ as part of the bid. This design follows typical auction practice [19], [20], [38] where non-price attributes are verifiable or contractually binding. Participation in any auction is voluntary, and bidders may skip an auction without penalty, which aligns with classical tenant behavior in multi-tenant edge systems. These requirements are grounded in practical feasibility and are commonly adopted in prior works [37], [39], [40] on auction-based edge computing environment.

**Battery Level Feasibility:** We can restore strict feasibility by adding a projection step after Line 13 of Algorithm 1: if $e_n^t \notin [O_n, B_n]$, minimally adjust $u_n^t, v_n^t, w_n^t$ to bring $e_n^t$ back into $[O_n, B_n]$. The above projection ensures that battery level is respected in each time slot, thereby no constraint violation at all. This results in even better *fit*. However, with regard to *regret*, it is *not sub-linear* in this case. The fundamental reason is that the projection-based restoration incurs the additional total cost of $\mathcal{O}(T)$. For *competitiveness*, the asymptotic competitive ratio can still be derived, as elaborated in Appendix H in our supplemental material, but is greater (i.e., worse) than our original competitive ratio when restoration is not considered.

## VIII. Conclusion

Edge infrastructure equipped with distributed energy storage operates as a novel form of VPP, enabling sustainable energy scheduling. While this has been largely ignored in prior works, our paper fills the critical gap. We study the online scheduling of edge AI inference services with fluctuating energy environments, while dynamically optimizing energy consumption and inference error rate. Moreover, we propose an auction-driven mechanism with novel online learning for long-term social cost minimization, rigorously prove the performance and further validate it through trace-driven experiments. For future work, we will continue to explore the efficient and flexible energy scheduling for energy-storage-enabled mobile edge computing systems and their interactions with AI services.

## References


[1] D. Zheng, L. Liu, G. Tang, Y. Wang, and W. Li, "Power demand reshaping using energy storage for distributed edge clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 35, no. 2, pp. 362–376, 2024.

[2] Y. Luo, L. Pu, and C.-H. Liu, "Computing power and battery charging management for solar energy powered edge computing," *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 1913–1927, 2025.

[3] Y. Qin, Y. Rao, Z. Xu, X. Lin, K. Cui, J. Du, and M. Ouyang, "Toward flexibility of user side in china: Virtual power plant (vpp) and vehicle-to-grid (v2g) interaction," *eTransportation*, vol. 18, p. 100291, 2023.

[4] C. Liu, R. J. Yang, X. Yu, C. Sun, G. Rosengarten, A. Liebman, R. Wakefield, P. S. Wong, and K. Wang, "Supporting virtual power plants decision-making in complex urban environments using reinforcement learning," *Sustainable Cities and Society*, vol. 99, p. 104915, 2023.

[5] M. Ji, H. Zhao, L. Jiao, S. Zhang, X. Li, Z. Qian, and B. Ye, "Edge ai inference as a service via dynamic resources from repeated auctions," *IEEE Transactions on Mobile Computing*, pp. 1–17, 2025.

[6] Z. Meng, H. Xu, M. Chen, Y. Xu, Y. Zhao, and C. Qiao, "Learning-driven decentralized machine learning in resource-constrained wireless edge computing," in *IEEE INFOCOM*, 2021.

[7] Q. Li, S. Wang, X. Ma, A. Zhou, Y. Wang, G. Huang, and X. Liu, "Battery-aware energy optimization for satellite edge computing," *IEEE Transactions on Services Computing*, vol. 17, no. 2, pp. 437–451, 2024.

[8] H. Liang and J. Ma, "Data-driven resource planning for virtual power plant integrating demand response customer selection and storage," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1833–1844, 2022.

[9] B. Zhou, K. Zhang, K. W. Chan, C. Li, X. Lu, S. Bu, and X. Gao, "Optimal coordination of electric vehicles for virtual power plants with dynamic communication spectrum allocation," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 450–462, 2021.

[10] H. Liu, S. Liu, S. Long, Q. Deng, and Z. Li, "Joint optimization of model deployment for freshness-sensitive task assignment in edge intelligence," in *IEEE INFOCOM*, 2024.

[11] 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*, vol. 22, no. 10, pp. 5870–5886, 2023.

[12] C. Chen, M. Herrera, G. Zheng, L. Xia, Z. Ling, and J. Wang, "Cross-edge orchestration of serverless functions with probabilistic caching," *IEEE Transactions on Services Computing*, vol. 17, no. 5, pp. 2139–2150, 2024.

[13] K. Yang, S. Long, Q. Shi, Y. Li, Z. Liu, Y. Wu, T. Yang, and Z. Jia, "Sketchint: Empowering int with towersketch for per-flow per-switch measurement," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 11, pp. 2876–2894, 2023.

[14] Q. Hu, Q. He, H. Zhong, G. Lu, X. Zhang, G. Zhai, and Y. Wang, "Varfvv: View-adaptive real-time interactive free-view video streaming with edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 7, pp. 2620–2634, 2025.

[15] X. Chen, G. Han, Y. Bi, Z. Yuan, M. K. Marina, Y. Liu, and H. Zhao, "Traffic prediction-assisted federated deep reinforcement learning for service migration in digital twins-enabled mec networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 10, pp. 3212–3229, 2023.

[16] X. Liu, M. Derakhshani, L. Mihaylova, and S. Lambotharan, "Risk-aware contextual learning for edge-assisted crowdsourced live streaming," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 3, pp. 740–754, 2023.

[17] Y. Xie, Y. Zhang, W.-J. Lee, Z. Lin, and Y. A. Shamash, "Virtual power plants for grid resilience: A concise overview of research and applications," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 329–343, 2024.

[18] R. Zhang, J. Liu, F. Liu, T. Huang, Q. Tang, S. Wang, and F. R. Yu, "Buffer-aware virtual reality video streaming with personalized and private viewport prediction," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 2, pp. 694–709, 2022.

[19] X. Chen, Y. Deng, G. Zhu, D. Wang, and Y. Fang, "From resource auction to service auction: An auction paradigm shift in wireless networks," *IEEE Wireless Communications*, vol. 29, no. 2, pp. 185–191, 2022.

[20] J. Pang, Z. Han, R. Zhou, R. Zhang, J. C. Lui, and H. Chen, "Eris: An online auction for scheduling unbiased distributed learning over edge networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 7196–7209, 2024.

[21] "Edge infrastructure dataset," https://aemo.com.au/energy-systems/energy/national-energy-market-nem/data-nem/data-dashboard-nem.

[22] G. Cui, Q. He, F. Chen, H. Jin, Y. Xiang, and Y. Yang, "Location privacy protection via delocalization in 5g mobile edge computing environment," *IEEE Transactions on Services Computing*, vol. 16, no. 1, pp. 412–423, 2023.

[23] W. Chen, J. Qiu, J. Zhao, Q. Chai, and Z. Y. Dong, "Bargaining game-based profit allocation of virtual power plant in frequency regulation market considering battery cycle life," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2913–2928, 2021.

[24] G. Tang, H. Yuan, D. Guo, K. Wu, and Y. Wang, "Reusing backup batteries as bess for power demand reshaping in 5g and beyond," in *IEEE INFOCOM*, 2021.

[25] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *ArXiv*, vol. abs/2004.10934, 2020.

[26] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[30] "Google clusters dataset," https://github.com/google/cluster-data.

[31] "Virtual power plant dataset," https://aemo.com.au/initiatives/major-programs/nem-distributed-energy-resources-der-program/.

[32] H. Chen, Y. Zhu, Z. Liang, C. Y. Chung, X. Yin, H. Yang, and J. Chen, "Energy scheduling of virtual power plants: A data-driven enclosing polyhedron method," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 2, pp. 1170–1179, 2025.

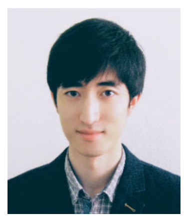
[33] M. Thomas, F. Niall, J. D. Sarah, and D. M. Malcolm, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nature Energy*, vol. 3, no. 2, p. 94–101, 2018.

[34] L. Yan, X. Zhang, Z. Ullah, H. S. Qazi, and H. M. Hasanien, "A novel solution strategy for scheduling optimization of virtual power plant considering multiple participants and peak energy market," *Renewable Energy*, vol. 250, p. 123275, 2025.

[35] X. Deng, J. Li, L. Shi, Z. Wei, X. Zhou, and J. Yuan, "Wireless powered mobile edge computing: Dynamic resource allocation and throughput

maximization," *IEEE Transactions on Mobile Computing*, vol. 21, no. 6, pp. 2271–2288, 2022.
[36] M. Abner, P. K.-Y. Wong, and J. C. Cheng, "Battery lifespan enhancement strategies for edge computing-enabled wireless bluetooth mesh sensor network for structural health monitoring," *Automation in Construction*, vol. 140, p. 104355, 2022.
[37] R. Lu, W. Zhang, Y. Wang, Q. Li, X. Zhong, H. Yang, and D. Wang, "Auction-based cluster federated learning in mobile edge computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 4, pp. 1145–1158, 2023.
[38] H. Liu, M. Wang, J. Liu, S. Yuan, J. Lou, C. Wu, and J. Li, "Cast: Cluster-driven truthful crowdfunding mechanism for shared ai service deployment," *IEEE Transactions on Services Computing*, vol. 18, no. 6, pp. 3552–3566, 2025.
[39] P. Wang, W. Sun, Y. Yang, D. Niyato, and D. O. Wu, "Mobisplit: Mobility-aware inference partitioning and offloading for efficient edge intelligence," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2025.
[40] W. Y. B. Lim, J. S. Ng, Z. Xiong, J. Jin, Y. Zhang, D. Niyato, C. Leung, and C. Miao, "Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 536–550, 2022.
[41] 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, pp. 105–122, jan 1999.
[42] Y. Zhang, L. Jiao, K. Zhu, X. Lin, and L. Zhang, "Toward market-assisted ai: Cloud inference for streamed data via model ensembles from auctions," *IEEE Transactions on Networking*, vol. 33, no. 2, pp. 793–806, 2025.
[43] A. Archer and E. Tardos, "Truthful mechanisms for one-parameter agents," in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, 2001, pp. 482–491.
[44] Y. Jin, L. Jiao, Z. Qian, S. Zhang, N. Chen, S. Lu, and X. Wang, "Provisioning edge inference as a service via online learning," in *IEEE SECON*, 2020.
[45] X. Yuan, N. Li, K. Wei, W. Xu, Q. Chen, H. Chen, and S. Guo, "Mobility and cost aware inference accelerating algorithm for edge intelligence," *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 1530–1549, 2025.
[46] 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.
[47] Y. Jin, L. Jiao, Z. Qian, S. Zhang, N. Chen, S. Lu, and X. Wang, "Provisioning edge inference as a service via online learning," in *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2020, pp. 1–9.
[48] "Vpp time granularity-sino," https://energypartnership.cn/fileadmin/china/newsroom_docs/Business_Models_of_Virtual_Power_Plants__VPPs__in_Germany_EN.pdf.
[49] "Vpp time granularity-epexspot," https://www.epexspot.com/en/15-minute-products-market-coupling.
[50] "Vpp time granularity-next kraftwerke," https://www.next-kraftwerke.com/energy-blog/day-ahead-switch-15-min.
[51] 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 *IEEE ICDCS*, 2023.
[52] "Gurobi," http://www.gurobi.com.
[53] H. Liu, S. Liu, S. Long, Q. Deng, and Z. Li, "Joint optimization of model deployment for freshness-sensitive task assignment in edge intelligence," in *IEEE INFOCOM*, 2024.
[54] G. Li, J. Cai, X. Chen, and Z. Su, "Nonlinear online incentive mechanism design in edge computing systems with energy budget," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 4086–4102, 2023.
[55] C. Huang, G. Chen, P. Xiao, Y. Xiao, Z. Han, and J. A. Chambers, "Joint offloading and resource allocation for hybrid cloud and edge computing in sagins: A decision assisted hybrid action space deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 5, pp. 1029–1043, 2024.
[56] J. Huang, J. Jiao, Y. Wang, Y. Li, and Q. Zhang, "Utility loss of information minimization with long erasure coding for task-adaptive communications in satellite-integrated internet," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 7, pp. 2604–2619, 2025.
[57] Y. Li, Z. Liu, Z. Kou, Y. Wang, G. Zhang, Y. Li, and Y. Sun, "Real-time adaptive partition and resource allocation for multi-user end-cloud inference collaboration in mobile environment," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 13 076–13 094, 2024.
[58] L. Lv, Y. Zhang, Y. Li, K. Xu, D. Wang, W. Wang, M. Li, X. Cao, and Q. Liang, "Communication-aware container placement and reassignment in large-scale internet data centers," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 540–555, 2019.
[59] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, and F. C. Lau, "Moving big data to the cloud: An online cost-minimizing approach," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 2710–2721, 2013.
[60] F. Wang, L. Jiao, K. Zhu, X. Lin, and L. Li, "Toward sustainable ai: Federated learning demand response in cloud-edge systems via auctions," in *IEEE INFOCOM*, 2023.
[61] "Edge ai modelzoo," https://ambiq.ai/articles/ambiq-modelzoo/.
[62] "Edge ai model ecosystem," https://www.thefastmode.com/technology-solutions/45592-ceva-embedur-expand-edge-ai-model-ecosystem-with-modelnova-platform.

**Hengdi Wang** is a Ph.D. student with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include online network optimization, cloud-edge collaborative computing and decentralized exchange.



**Lei Jiao** received the Ph.D. degree in computer science from the University of Göttingen, Germany. He is currently a faculty member at the University of Oregon, USA and was previously a technical staff member at Nokia Bell Labs in Ireland. His research interests include machine learning systems, edge computing, cloud computing, network optimization, and network economics.

**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 a Professor with the Beijing University of Posts and Telecommunications, Beijing, China. His research interests include Internet of Vehicles, Edge Computing and Distributed Learning.

**Lin Zhang** received the Ph.D. degrees in 2001, from the Beijing University of Posts and Telecommunications, Beijing, China. He joined Beijing University of Posts and Telecommunications in 2004 and has been a Professor since 2011. He is also the director of Beijing Big Data Center. His current research interests include mobile cloud computing.

**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.

## APPENDIX

### A. Proof of Theorem 1

*Proof.* We set $T = N = 1$, thereby dropping the superscript $t$ and the subscript $n$, and denote $\mathcal{I}_1$ as $\mathcal{I}$. Now, for each $i \in \mathcal{I}$, we set $m_i = l_i = 0$, set $s_+ = 0$, and set $s_- = d = M$, where $M$ is a huge positive constant; we also set $\lambda_i = 1$. With these settings, we have the reduction as follows. The objective is now the minimization of $\sum_{i\in\mathcal{I}}(r_i + \theta_i)x_i$. Note that due to $M$, we have $q = u = v = w = 0$. The constraints now become $\sum_{i\in\mathcal{I}} \psi_i x_i \geq \sum_{i\in\mathcal{I}} \rho_i$, which is derived from (1g). Note that all other constraints are automatically satisfied. (1a) and (1b) hold, since $e^1 = e^0$ and $O \leq e^0 \leq B$. (1c)∼(1f) can hold because $y$ and $z$ can be set to 0. That is, the problem $\mathcal{P}$ becomes the *minimum knapsack problem*. Such NP-hardness holds even in the above offline setting, not to mention that we aim to address it online, and the proof completes. □

### B. Proof of Theorem 2

*Proof.* We first define the fractional regret $\widetilde{\mathrm{Reg}}^T$ and the fractional fit $\widetilde{\mathrm{Fit}}^T$ for the problem $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$, and then propose Lemmas 1 and 2 to assist subsequent proofs.

$$\widetilde{\mathrm{Reg}}^T = \textstyle\sum_{t=1}^T f_{-S}^t(\widetilde{\mathrm{Z}}^t) - \sum_{t=1}^T f_{-S}^t(\widetilde{\mathrm{Z}}^{t\star}), \tag{1}$$

$$\widetilde{\mathrm{Fit}}^T = \|[\textstyle\sum_{t=1}^T g^t(\widetilde{\mathrm{Z}}^t)]^+\|, \tag{2}$$

where $\widetilde{\mathrm{Z}}^t$ is the online fractional solution and $\widetilde{\mathrm{Z}}^{t\star}$ is the "one-shot" fractional optimal solution, i.e., breaking the problem into one-shot slices and solving each slice individually.

**Assumptions:** To advance our following analysis, we establish the following assumptions: (i) The functions $f_{-S}^t(\widetilde{\mathrm{Z}})$ has bounded gradients on $\mathcal{Z}$, i.e., $\|\nabla f_{-S}^t(\widetilde{\mathrm{Z}})\| \leq F$, $\forall \widetilde{\mathrm{Z}} \in \mathcal{Z}$ and $g^t(\widetilde{\mathrm{Z}})$ is also bounded on $\mathcal{Z}$, i.e., $\|g^t(\widetilde{\mathrm{Z}})\| \leq G$, $\forall \widetilde{\mathrm{Z}} \in \mathcal{Z}$; (ii) The radius of the convex feasible set Z is bounded, i.e., $\|\widetilde{\mathrm{Z}}_1 - \widetilde{\mathrm{Z}}_2\| \leq R$, $\forall \widetilde{\mathrm{Z}}_1, \widetilde{\mathrm{Z}}_2 \in \mathcal{Z}$; (iii) There exists a constant $\delta > 0$ and an interior point $\ddot{\mathrm{Z}} \in \mathcal{Z}$ such that $g^t(\ddot{\mathrm{Z}}^t) \leq -\delta\mathbf{1}, \forall t \in \mathcal{T}$; (iv) The slack constant $\delta$ is larger than the point-wise maximal variation of the consecutive constraints, i.e., $\delta > V(g)$, where $V(g) = \max_{t\in\mathcal{T}} \max_{\widetilde{\mathrm{Z}}\in\mathcal{Z}} \|g^{t+1}(\widetilde{\mathrm{Z}}) - g^t(\widetilde{\mathrm{Z}})\|^+$.

**Lemma 1.** The relationship between dynamic regret and fit in the domain of integers and reals can be written as

$$\mathrm{Reg}^T \leq \widetilde{\mathrm{Reg}}^T, \mathrm{Fit}^T \leq \widetilde{\mathrm{Fit}}^T. \tag{3}$$

*Proof.* We derive the relationship between $\mathrm{Reg}^T$ and $\widetilde{\mathrm{Reg}}^T$ as

$$\begin{aligned}
&\mathrm{Reg}^T = \mathbb{E}[\textstyle\sum_{t=1}^T f_{-S}^t(\bar{\mathrm{Z}}^t)] - \sum_{t=1}^T f_{-S}^t(\mathrm{Z}^{t\star})\\
&\overset{(a)}{=} \textstyle\sum_{t=1}^T f_{-S}^t(\mathbb{E}[\bar{\mathrm{Z}}^t]) - \sum_{t=1}^T f_{-S}^t(\mathrm{Z}^{t\star})\\
&= \textstyle\sum_{t=1}^T f_{-S}^t(\widetilde{\mathrm{Z}}^t) - \sum_{t=1}^T f_{-S}^t(\mathrm{Z}^{t\star})\\
&+ \textstyle\sum_{t=1}^T f_{-S}^t(\mathbb{E}[\bar{\mathrm{Z}}^t]) - \sum_{t=1}^T f_{-S}^t(\widetilde{\mathrm{Z}}^t)\\
&\overset{(b)}{\leq} \textstyle\sum_{t=1}^T f_{-S}^t(\widetilde{\mathrm{Z}}^t) - \sum_{t=1}^T f_{-S}^t(\widetilde{\mathrm{Z}}^{t\star}) = \widetilde{\mathrm{Reg}}^T,
\end{aligned} \tag{4}$$

where (a) holds by the linearity of $f_{-S}^t(\bar{\mathrm{Z}}^t)$; (b) holds due to $\mathbb{E}[\bar{\mathrm{Z}}^t] = \widetilde{\mathrm{Z}}^t$ which ensured by randomized rounding algorithm and the fact that the objective value conducted by integer optimum is more than fractional optimum. We derive fit as

$$\begin{aligned}
&\mathrm{Fit}^T = \|[\mathbb{E}[\textstyle\sum_{t=1}^T g^t(\bar{\mathrm{Z}}^t)]]^+\| \leq \|\mathbb{E}[\sum_{t=1}^T g^t(\bar{\mathrm{Z}}^t)]\|\\
&\overset{(a)}{=} \|\textstyle\sum_{t=1}^T g^t(\mathbb{E}[\bar{\mathrm{Z}}^t])\| = \|\sum_{t=1}^T g^t(\widetilde{\mathrm{Z}}^t)\| = \widetilde{\mathrm{Fit}}^T,
\end{aligned} \tag{5}$$

where (a) follows that the value of the 2-norm will decrease due to the negative values are set to 0 by using $[\cdot]^+ = \max\{\cdot, 0\}$. The linearity of $g^t(\widetilde{\mathrm{Z}}^t)$ and the unchanged expectation property holds by the Algorithm 3 guarantee (b). □

**Lemma 2.** Under previous assumptions and the dual variable initialization of $\varphi^1 = 0$, we can then derive

$$(\|\varphi^{t+1}\|^2 - \|\varphi^t\|^2)/2 \leq \mu\, \varphi^t g^t(\widetilde{\mathrm{Z}}^t) + \mu^2\|g^t(\widetilde{\mathrm{Z}}^t)\|^2/2, \tag{6}$$

$$\|\varphi^t\| \leq \|\bar{\varphi}\| \triangleq \mu G + \frac{2FR + R^2/(2\alpha) + (nG^2)/2}{\delta - V(g)}. \tag{7}$$

*Proof.* According to the update policy of $\varphi$, we will have

$$\begin{aligned}
&\|\varphi^{t+1}\|^2 = \|[\varphi^t + \mu g^t(\widetilde{\mathrm{Z}}^t)]^+\|^2 \leq \|\varphi^t + \mu g^t(\widetilde{\mathrm{Z}}^t)\|^2\\
&= \|\varphi^t\|^2 + 2\mu\varphi^t g^t(\widetilde{\mathrm{Z}}^t) + \mu^2\|g^t(\widetilde{\mathrm{Z}}^t)\|^2.
\end{aligned} \tag{8}$$

After rearranging terms in (8), we then arrive at

$$(\|\varphi^{t+1}\|^2 - \|\varphi^t\|^2)/2 \leq \mu\varphi^t g^t(\widetilde{\mathrm{Z}}^t) + \mu^2\|g^t(\widetilde{\mathrm{Z}}^t)\|^2/2. \tag{9}$$

Since $\widetilde{\mathrm{Z}}^{t+1}$ is fractional solution in (9) of our main paper, by using the interior point $\ddot{\mathrm{Z}}^t$ in assumption (iii), we have

$$\begin{aligned}
&\nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t) + \varphi^{t+1}g^t(\widetilde{\mathrm{Z}}^{t+1}) + \|\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t\|^2/(2\alpha)\\
&\leq \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t) + \varphi^{t+1}g^t(\ddot{\mathrm{Z}}^t) + \|\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t\|^2/(2\alpha)\\
&\overset{(a)}{\leq} \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t) - \delta\varphi^{t+1}\mathbf{1} + \|\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t\|^2/(2\alpha)\\
&\overset{(b)}{\leq} \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t) - \delta\|\varphi^{t+1}\| + \|\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t\|^2/(2\alpha),
\end{aligned} \tag{10}$$

where inequality (a) holds due to assumption (ii), and inequality (b) holds because $\|\varphi^{t+1}\|$ is less or equal to $\varphi^{t+1}\mathbf{1}$ for any non-negative vector $\varphi^{t+1}$. Following this, we can obtain

$$\begin{aligned}
&\varphi^{t+1}g^t(\widetilde{\mathrm{Z}}^{t+1}) \leq (\|\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t\|^2 - \|\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t\|^2)/(2\alpha)\\
&+ \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t) - \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t)\\
&- \delta\|\varphi^{t+1}\| \overset{(a)}{\leq} \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t) - \delta\|\varphi^{t+1}\|\\
&- \nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)(\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t) + R^2/(2\alpha)\\
&\overset{(b)}{\leq} \|\nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)\|\|\ddot{\mathrm{Z}}^t - \widetilde{\mathrm{Z}}^t\| - \delta\|\varphi^{t+1}\|\\
&+ \|\nabla f_{-S}^t(\widetilde{\mathrm{Z}}^t)\|\|\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t\| + R^2/(2\alpha)\\
&\overset{(c)}{\leq} 2FR - \delta\|\varphi^{t+1}\| + R^2/(2\alpha) \triangleq \Phi^{t+1},
\end{aligned} \tag{11}$$

where inequality (a) holds by the bounded radius of the domain, and $\|\widetilde{\mathrm{Z}}^{t+1} - \widetilde{\mathrm{Z}}^t\| \geq 0$; inequality (b) holds by Cauchy-Schwartz inequality; the inequality (c) holds by using the bounded gradient in assumption (i) and bounded domain. We consider (11) with Lemma 2, and this leads to

$$\begin{aligned}
&\Delta(\varphi^{t+1}) \triangleq (\|\varphi^{t+1}\|^2 - \|\varphi^t\|^2)/2 \qquad (12)\\
&\leq \mu\varphi^{t+1}g^{t+1}(\widetilde{\mathrm{Z}}^{t+1}) + \mu^2\|g^{t+1}(\widetilde{\mathrm{Z}}^{t+1})\|^2/2
\end{aligned}$$

$$\overset{(a)}{\leq} \mu\varphi^{t+1}(g^{t+1}(\widetilde{Z}^{t+1}) - g^t(\widetilde{Z}^{t+1})) + \mu^2G^2/2 + \Phi^{t+1}$$
$$\overset{(b)}{\leq} \mu\varphi^{t+1}[g^{t+1}(\widetilde{Z}^{t+1}) - g^t(\widetilde{Z}^{t+1})]^+ + \mu^2G^2/2 + \Phi^{t+1}$$
$$\overset{(c)}{\leq} \mu V(g)\|\varphi^{t+1}\| + \mu^2G^2/2 + 2FR - \delta\|\varphi^{t+1}\| + R^2/(2\alpha),$$

where inequality (a) holds by using the upper bound of $g^t(\cdot)$; inequality (b) holds by the non-negative property of $\varphi^{t+1}$; and inequality (c) holds by assumption (iv).

Next, we show the correctness of (7) by contradiction. Without loss of generality, we suppose that $t+2$ is the first time index that breaks (7), that is:

$$\|\varphi^{t+1}\| \leq \|\bar{\varphi}\| < \|\varphi^{t+2}\|. \tag{13}$$

By using the update policy of $\varphi$, the relationship can be obtained on $\varphi$ between consecutive time slots as follows:

$$\begin{aligned}
&\|\varphi^{t+1}\| \overset{(a)}{\geq} \|\varphi^{t+2}\| - \|\varphi^{t+2} - \varphi^{t+1}\| \\
&= \|\varphi^{t+2}\| - \|[\varphi^{t+1} + \mu g^{t+1}(\widetilde{Z}^{t+1})]^+ - \varphi^{t+1}\| \\
&\geq \|\varphi^{t+2}\| - \|\varphi^{t+1} + \mu g^{t+1}(\widetilde{Z}^{t+1}) - \varphi^{t+1}\| \\
&= \|\varphi^{t+2}\| - \|\mu g^{t+1}(\widetilde{Z}^{t+1})\| \overset{(b)}{>} \|\bar{\varphi}\| - \mu G,
\end{aligned} \tag{14}$$

where inequality (a) holds by the triangle inequality, and inequality (b) holds by (13). Then we can obtain $\Delta(\varphi^{t+1}) < 0$, leading to $\|\varphi^{t+2}\| < \|\varphi^{t+1}\|$, which contradicts $\|\varphi^{t+1}\| \leq \|\bar{\varphi}\| < \|\varphi^{t+2}\|$. Thus, Lemma 2 can be proved. □

Here, we proceed with the proof of Theorem 2 in two steps. **Step I:** The objective in (9) of our main paper indicates that it is $1/\alpha$-strongly convex with respect to $\widetilde{Z}$, which we denote as $L^t(\widetilde{Z}^t)$. This can be formally stated as: for $\forall \widetilde{Z}_1, \widetilde{Z}_2 \in \mathcal{Z}$,

$$L^t(\widetilde{Z}_2) \geq L^t(\widetilde{Z}_1) + \nabla L^t(\widetilde{Z}_1)(\widetilde{Z}_2 - \widetilde{Z}_1) + \|\widetilde{Z}_2 - \widetilde{Z}_1\|^2/(2\alpha). \tag{15}$$

Since $\widetilde{Z}^{t+1}$ is the optimum for $\min_{\widetilde{Z}\in\mathcal{Z}} L^t(\widetilde{Z})$, then we have

$$\nabla L^t(\widetilde{Z}^{t+1})(\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}) \geq 0. \tag{16}$$

We set $\widetilde{Z}_1 = \widetilde{Z}^{t+1}$ and $\widetilde{Z}_2 = \widetilde{Z}^{t\star}$, and plugging (16) into (15), then we arrive at the following expression:

$$L^t(\widetilde{Z}^{t\star}) \geq L^t(\widetilde{Z}^{t+1}) + \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2/(2\alpha). \tag{17}$$

After adding $f^t_{-S}(\widetilde{Z}^t)$ on both two sides, expanding $L^t(\cdot)$ according to its definition and using the property of convex function on $f^t_{-S}(\cdot)$, i.e., $f^t_{-S}(\widetilde{Z}^{t\star}) \geq f^t_{-S}(\widetilde{Z}^t) + \nabla f^t_{-S}(\widetilde{Z}^t)(\widetilde{Z}^{t\star} - \widetilde{Z}^t)$, then we can further obtain

$$\begin{aligned}
&f^t_{-S}(\widetilde{Z}^t) + \nabla f^t_{-S}(\widetilde{Z}^t)(\widetilde{Z}^{t+1} - \widetilde{Z}^t) \\
&+ \varphi^{t+1}g^t(\widetilde{Z}^{t+1}) + \|Z^{t+1} - Z^t\|^2/(2\alpha) \\
&\leq f^t_{-S}(\widetilde{Z}^{t\star}) + \|Z^{t\star} - Z^t\|^2/(2\alpha) \\
&- \|Z^{t\star} - Z^{t+1}\|^2/(2\alpha) + \varphi^{t+1}g^t(\widetilde{Z}^{t\star}) \\
&\overset{(a)}{\leq} f^t_{-S}(\widetilde{Z}^{t\star}) + (\|Z^{t\star} - Z^t\|^2 - \|Z^{t\star} - Z^{t+1}\|^2)/(2\alpha),
\end{aligned} \tag{18}$$

where inequality (a) holds by $\varphi^{t+1} \geq 0$ and the per-slot optimal solution $\widetilde{Z}^{t\star}$ is feasible, i.e., $g^t(\widetilde{Z}^{t\star}) \leq 0$. We then proceed to analyze the gradient term, which is given by

$$\begin{aligned}
&- \nabla f^t_{-S}(\widetilde{Z}^t)(\widetilde{Z}^{t+1} - \widetilde{Z}^t) \leq \|\nabla f^t_{-S}(\widetilde{Z}^t)\|\|\widetilde{Z}^{t+1} - \widetilde{Z}^t\| \\
&\overset{(a)}{\leq} \|\nabla f^t_{-S}(\widetilde{Z}^t)\|^2/(2\zeta) + \zeta\|\widetilde{Z}^{t+1} - \widetilde{Z}^t\|^2/2 \\
&\overset{(b)}{\leq} F^2/(2\zeta) + \zeta\|\widetilde{Z}^{t+1} - \widetilde{Z}^t\|^2/2,
\end{aligned} \tag{19}$$

where $\zeta$ is an arbitrary positive constant. Inequality (a) holds by $a^2+b^2 \geq 2ab$, inequality (b) holds by the bounded gradient of $f^t_{-S}(\cdot)$. Then we plug (19) into (18) and yield

$$\begin{aligned}
&f^t_{-S}(\widetilde{Z}^t) + \varphi^{t+1}g^t(\widetilde{Z}^{t+1}) \leq f^t_{-S}(\widetilde{Z}^{t\star}) \\
&+ (\zeta/2 - 1/(2\alpha))\|\widetilde{Z}^{t+1} - \widetilde{Z}^t\|^2 + F^2/2\zeta \\
&+ (\|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 - \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2)/(2\alpha) \overset{(a)}{=} f^t_{-S}(\widetilde{Z}^{t\star}) \\
&+ (\|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 - \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2)/(2\alpha) + \alpha F^2/2,
\end{aligned} \tag{20}$$

where inequality (a) holds by setting $\zeta = 1/\alpha$. By plugging (20) into the Lemma 2, we can further derive

$$\begin{aligned}
&f^t_{-S}(\widetilde{Z}^t) + \Delta(\varphi^{t+1})/\mu \leq f^t_{-S}(\widetilde{Z}^t) + \mu\|g^{t+1}(\widetilde{Z}^{t+1})\|/2 \\
&+ \varphi^{t+1}g^{t+1}(\widetilde{Z}^{t+1}) + \varphi^{t+1}g^t(\widetilde{Z}^{t+1}) - \varphi^{t+1}g^t(\widetilde{Z}^{t+1}) \\
&= f^t_{-S}(\widetilde{Z}^t) + \varphi^{t+1}g^t(\widetilde{Z}^{t+1}) + \mu\|g^{t+1}(\widetilde{Z}^{t+1})\|^2/2 \\
&+ \varphi^{t+1}g^{t+1}(\widetilde{Z}^{t+1}) - \varphi^{t+1}g^t(\widetilde{Z}^{t+1}) \overset{(a)}{\leq} f^t_{-S}(\widetilde{Z}^{t\star}) \\
&+ (\|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 - \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2)/(2\alpha) + (\alpha F^2)/2 \\
&+ \mu\|g^{t+1}(\widetilde{Z}^{t+1})\|^2/2 + \varphi^{t+1}(g^{t+1}(\widetilde{Z}^{t+1}) - g^t(\widetilde{Z}^{t+1})) \\
&\leq f^t_{-S}(\widetilde{Z}^{t\star}) + (\|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 - \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2)/(2\alpha) \\
&+ (\alpha F^2 + \mu G^2)/2 + \varphi^{t+1}[g^{t+1}(\widetilde{Z}^{t+1}) - g^t(\widetilde{Z}^{t+1})] \\
&\overset{(b)}{\leq} f^t_{-S}(\widetilde{Z}^{t\star}) + (\|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 - \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2)/(2\alpha) \\
&+ (\alpha F^2 + \mu G^2)/2 + \|\varphi^{t+1}\|V(g^t),
\end{aligned} \tag{21}$$

where inequality (a) holds by (20), inequality (b) holds by assumption (iv). Then we consider the intermediate term as:

$$\begin{aligned}
&\|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 = \|\widetilde{Z}^{t\star} - \widetilde{Z}^t\|^2 - \|\widetilde{Z}^t - \widetilde{Z}^{t-1\star}\|^2 \\
&+ \|\widetilde{Z}^t - \widetilde{Z}^{t-1\star}\|^2 \overset{(a)}{=} \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t-1\star}\|\|\widetilde{Z}^{t\star} - 2\widetilde{Z}^t + \widetilde{Z}^{t-1\star}\| \\
&+ \|\widetilde{Z}^t - \widetilde{Z}^{t-1\star}\|^2 \overset{(b)}{\leq} 2R\|\widetilde{Z}^{t\star} - \widetilde{Z}^{t-1\star}\| + \|\widetilde{Z}^t - \widetilde{Z}^{t-1\star}\|^2,
\end{aligned} \tag{22}$$

where equation (a) holds by applying the difference of two squares on the first two terms, inequality (b) holds by the triangle inequality for vectors and the bounded radius on domain. After applying (22) to (21), we then proceed to get

$$\begin{aligned}
&\Delta(\varphi^{t+1})/\mu + f^t_{-S}(\widetilde{Z}^t) \leq f^t_{-S}(\widetilde{Z}^{t\star}) + \|\varphi^{t+1}\|V(g^t) \\
&+ (\alpha F^2 + \mu G^2)/2 + R\|\widetilde{Z}^{t\star} - \widetilde{Z}^{t-1\star}\|/\alpha \\
&+ \|\widetilde{Z}^t - \widetilde{Z}^{t-1\star}\|^2/(2\alpha) - \|\widetilde{Z}^{t\star} - \widetilde{Z}^{t+1}\|^2/(2\alpha).
\end{aligned} \tag{23}$$

Sum the values of (23) over the time slot $t$, we will have

$$\textstyle\sum_{t=1}^T \Delta(\varphi^{t+1})/\mu + \sum_{t=1}^T f^t_{-S}(\widetilde{Z}^t) \leq \sum_{t=1}^T f^t_{-S}(\widetilde{Z}^{t\star})$$

$$
\begin{aligned}
&+ R\cdot V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})/\alpha + \textstyle\sum_{t=1}^{T}(\|\varphi^{t+1}\|V(g^t))\\
&+ \textstyle\sum_{t=1}^{T}(\|\widetilde{\mathrm{Z}}^{t} - \widetilde{\mathrm{Z}}^{t-1\star}\|^2 - \|\widetilde{\mathrm{Z}}^{t\star} - \widetilde{\mathrm{Z}}^{t+1}\|^2)/(2\alpha)\\
&+ (\alpha F^2 T + \mu G^2 T)/2 \overset{(a)}{\leq} R\cdot V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})/\alpha\\
&+ \|\widetilde{\mathrm{Z}}^{1} - \widetilde{\mathrm{Z}}^{0\star}\|^2/(2\alpha) - \|\widetilde{\mathrm{Z}}^{T\star} - \widetilde{\mathrm{Z}}^{T+1}\|^2/(2\alpha)\\
&+ \textstyle\sum_{t=1}^{T} f_{-S}^{t}(\widetilde{\mathrm{Z}}^{t\star}) + (\alpha F^2 T + \mu G^2 T)/2\\
&+ \|\bar{\varphi}\| \textstyle\sum_{t=1}^{T} V(g^t) \overset{(b)}{\leq} +(\alpha F^2 T + \mu G^2 T)/2\\
&+ R\cdot V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})/\alpha + \textstyle\sum_{t=1}^{T} f_{-S}^{t}(\widetilde{\mathrm{Z}}^{t\star})\\
&+ \|\bar{\varphi}\| V(\{g^t\}_{t=1}^{T}) + \|\widetilde{\mathrm{Z}}^{1} - \widetilde{\mathrm{Z}}^{0\star}\|^2/(2\alpha),
\end{aligned}
\tag{24}
$$

where $V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T}) \triangleq \sum_{t=1}^{T}\|\widetilde{\mathrm{Z}}^{t\star} - \widetilde{\mathrm{Z}}^{t-1\star}\|$, inequality (a) holds by the definition of $\|\bar{\varphi}\|$, inequality (b) holds due to $V(\{g^t\}_{t=1}^{T}) \triangleq \sum_{t=1}^{T} \max_{\widetilde{\mathrm{Z}}\in\mathcal{Z}}[g^{t+1}(\widetilde{\mathrm{Z}}) - g^t(\widetilde{\mathrm{Z}})]^+$. Then,

$$
\begin{aligned}
&\mathrm{Reg}^T = \textstyle\sum_{t=1}^{T} f_{-S}^{t}(\widetilde{\mathrm{Z}}^{t}) - \sum_{t=1}^{T} f_{-S}^{t}(\widetilde{\mathrm{Z}}^{t\star})\\
&\leq (\alpha F^2 T + \mu G^2 T)/2 + \|\bar{\varphi}\| V(\{g^t\}_{t=1}^{T})\\
&+ R\cdot V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})/\alpha + \|\widetilde{\mathrm{Z}}^{1} - \widetilde{\mathrm{Z}}^{0\star}\|^2/(2\alpha)\\
&- \textstyle\sum_{t=1}^{T} \Delta(\varphi^{t+1})/\mu = (\alpha F^2 T + \mu G^2 T)/2\\
&+ \|\bar{\varphi}\| V(\{g^t\}_{t=1}^{T}) + \|\widetilde{\mathrm{Z}}^{1} - \widetilde{\mathrm{Z}}^{0\star}\|^2/(2\alpha) + \|\varphi^{t=2}\|^2/(2\mu)\\
&+ R\cdot V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})/\alpha - \|\varphi^{T+2}\|^2/(2\mu) \overset{(a)}{\leq} \mathcal{R}^T,
\end{aligned}
\tag{25}
$$

where we denote that $\mathcal{R}^T \triangleq R\cdot V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})/\alpha + \alpha F^2 T/2 + \mu G^2(T+1)/2 + R^2/2\alpha + \|\bar{\varphi}\| V(\{g^t\}_{t=1}^{T})$. Inequality (a) holds due to $\|\widetilde{\mathrm{Z}}^{1} - \widetilde{\mathrm{Z}}^{0\star}\|^2$ has been bounded by $R$ according to the bounded radius of domain, $\|\varphi^{T+2}\|^2 \geq 0$, and $\|\varphi^{t=2}\|^2 \leq \mu^2 G^2$ if $\varphi^1 = 0$.

**Step II:** According to the dual recursion in Algorithm 2, the following result is obtained:

$$
[\varphi^T + \mu g^T(\widetilde{\mathrm{Z}}^{T})]^+ \geq \cdots \geq \varphi^1 + \textstyle\sum_{t=1}^{T} \mu g^t(\widetilde{\mathrm{Z}}^{t}). \tag{26}
$$

Since $\varphi^1 = 0$, we can rearrange the terms, and then obtain

$$
\textstyle\sum_{t=1}^{T} g^t(\widetilde{\mathrm{Z}}^{t}) \leq \varphi^{T+1}/\mu - \varphi^1/\mu \leq \varphi^{T+1}/\mu. \tag{27}
$$

Therefore, $\widetilde{\mathrm{Fit}}^T = \|[\sum_{t=1}^{T} g^t(\widetilde{\mathrm{Z}}^{t})]^+\|$ can be treated as

$$
\widetilde{\mathrm{Fit}}^T \leq \left\|\textstyle\sum_{t=1}^{T} g^t(\widetilde{\mathrm{Z}}^{t})\right\| \leq \left\|\varphi^{T+1}/\mu\right\| \leq \|\bar{\varphi}\|/\mu. \tag{28}
$$

Then we can obtain the following results: $\mathrm{Reg}^T \leq \widetilde{\mathrm{Reg}}^T \leq \mathcal{R}^T$ and $\mathrm{Fit}^T \leq \widetilde{\mathrm{Fit}}^T \leq \varphi^{T+1}/\mu \leq \|\bar{\varphi}\|/\mu$. By controlling the parameters as $\alpha = \mu = \max\{\sqrt{\frac{V(\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})}{T}}, \sqrt{\frac{V(\{g^t\}_{t=1}^{T})}{T}}\}$, we have $\mathrm{Reg}^T = \mathcal{O}(\max\{\sqrt{V(\{\widetilde{\mathrm{z}}^{t\star}\}_{t=1}^{T})T}, \sqrt{V(\{g^t\}_{t=1}^{T})T}\})$ and $\mathrm{Fit}^T \leq \|\bar{\varphi}\|/\mu = \mathcal{O}(\max\{\frac{T}{V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})}, \frac{T}{V(\{g^t\}_{t=1}^{T})}\})$. If we further set the parameters as $\alpha = \mu = \mathcal{O}(T^{-\frac{1}{3}})$, we can get $\mathrm{Reg}^T = \mathcal{O}(\max\{V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T})T^{\frac{1}{3}}, V(\{g^t\}_{t=1}^{T})T^{\frac{1}{3}}, T^{\frac{2}{3}}\})$ and $\mathrm{Fit}^T = \mathcal{O}(T^{\frac{2}{3}})$. Then the sub-linear regret and fit of $\mathcal{O}(T^{\frac{2}{3}})$ can be achieved if we have $V(\{\widetilde{\mathrm{Z}}^{t\star}\}_{t=1}^{T}) \in \mathcal{O}(T^{\frac{2}{3}})$ and $V(\{g^t\}_{t=1}^{T}) \in \mathcal{O}(T^{\frac{2}{3}})$ in the meantime. □

### C. Proof of Theorem 3

*Proof.* **Step I:** We denote the timestamps that the model switching operations occur as $t'_l$, $\forall 1 \leq l \leq l'$, where $l'$ is the total number of the model switching operations over time horizon $T$. Then the relation between the switching cost and the accumulated non-switching cost is denoted by

$$
f_S^{t'_l}(\bar{\mathrm{Z}}^{t'_l}) \leq \beta \textstyle\sum_{t=t'_l}^{t'_{l+1}-1} f_{-S}^{t}(\mathrm{Z}_3^{t\star}), \tag{29}
$$

where $\mathrm{Z}_3^{t\star}$ is the offline optimum result of $\mathcal{P}_3$. Furthermore, when the switching operation occurs in $t'_l$, the ratio of switching cost to non-switching cost can be bounded as

$$
\frac{f_S^{t'_l}(\bar{\mathrm{Z}}^{t'_l})}{f_{-S}^{t'_l}(\bar{\mathrm{Z}}^{t'_l})} \leq \frac{\sum_{n=1}^{N}\sum_{i=1}^{I}\max_i\{m_{in}, l_{in}\}}{\sum_{n=1}^{N} s_+^{t'_l}(O_n - B_n)} \triangleq \vartheta, \tag{30}
$$

where $\vartheta$ is a constant. After that, we have the following:

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{t} f_S^{t}(\bar{\mathrm{Z}}^{t}) = \sum_{1\leq l\leq l'} f_S^{t'_l}(\bar{\mathrm{Z}}^{t'_l}) + f_S^{t'_{l'}}(\bar{\mathrm{Z}}^{t'_l})\\
&\overset{(a)}{\leq} \textstyle\sum_{1\leq l\leq l'}[\beta \sum_{t=t'_l}^{t'_{l+1}-1} f_{-S}^{t}(\mathrm{Z}_3^{t\star})]\\
&+ \vartheta \textstyle\sum_{t=t'_{l'}+1}^{t} f_{-S}^{t}(\mathrm{Z}_3^{t\star}) + \vartheta f_{-S}^{t'_{l'}}(\bar{\mathrm{Z}}^{t'_l})\\
&= \beta \textstyle\sum_{t=1}^{t'_{l'}-1} f_{-S}^{t}(\mathrm{Z}_3^{t\star}) + \vartheta \sum_{t=t'_{l'}}^{t} f_{-S}^{t}(\mathrm{Z}_3^{t\star})\\
&\leq \max\{\beta, \vartheta\} \textstyle\sum_{t=1}^{t} f_{-S}^{t}(\mathrm{Z}_3^{t\star}) \triangleq k_1 \cdot \sum_{t=1}^{t} f_{-S}^{t}(\mathrm{Z}_3^{t\star}),
\end{aligned}
\tag{31}
$$

where $k_1$ is a constant. Inequality (a) a holds since (29), (30) and the extra added term $\vartheta \sum_{t=t'_{l'}}^{t} f_{-S}^{t}(\mathrm{Z}_3^{t\star})$ is non-negative.

**Step II:** We have the definition for the non-switching cost:

$$
\xi \triangleq \max_{t\in\mathcal{T}}\{\max_{\mathrm{Z}_3^{t\star}} f_{-S}^{t}(\mathrm{Z}_3^{t\star})/\min_{\mathrm{Z}_3^{t\star}} f_{-S}^{t}(\mathrm{Z}_3^{t\star})\}, \tag{32}
$$

where $\xi$ is a constant. After taking $t = 1$ to $t = T$, we obtain

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{T} f_{-S}^{t}(\mathrm{Z}_3^{t\star}) \leq \xi \sum_{t=1}^{t} f_{-S}^{t}(\mathrm{Z}^{t*})\\
&\leq \xi \textstyle\sum_{t=1}^{T}(f_S^{t}(\mathrm{Z}^{t*}) + f_{-S}^{t}(\mathrm{Z}^{t*})) \leq \xi \mathcal{P}(\mathrm{Z}^{t*}),
\end{aligned}
\tag{33}
$$

where we link the non-switching cost and the total cost. Then the overall cost can be bounded by

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{T}(f_S^{t}(\bar{\mathrm{Z}}^{t}) + f_{-S}^{t}(\mathrm{Z}_3^{t\star})) \overset{(a)}{\leq} (1+k_1)\sum_{t=1}^{T} f_{-S}^{t}(\mathrm{Z}_3^{t\star})\\
&\overset{(b)}{\leq} \xi(1+k_1)\mathcal{P}(\mathrm{Z}^{t*}) \triangleq k \cdot \mathcal{P}(\mathrm{Z}^{t*}),
\end{aligned}
\tag{34}
$$

where $k$ is a constant. Inequality (a) holds by (31) and Inequality (b) holds by (33).

**Step III:** For the expectation $\mathbb{E}[\mathcal{P}(\bar{\mathrm{Z}}^{t})]$, we can obtain

$$
\begin{aligned}
&\mathbb{E}[\mathcal{P}(\bar{\mathrm{Z}}^{t})] = \mathbb{E}[\textstyle\sum_{t=1}^{T}(f_S^{t}(\bar{\mathrm{Z}}^{t}) + f_{-S}^{t}(\bar{\mathrm{Z}}^{t}))]\\
&\overset{(a)}{\leq} \textstyle\sum_{t=1}^{T}(f_S^{t}(\bar{\mathrm{Z}}^{t}) + f_{-S}^{t}(\bar{\mathrm{Z}}_3^{t}))\\
&\overset{(b)}{\leq} \textstyle\sum_{t=1}^{T}(f_S^{t}(\bar{\mathrm{Z}}^{t}) + f_{-S}^{t}(\mathrm{Z}_3^{t\star})) + \mathcal{O}(T^{\varsigma_3})\\
&\overset{(c)}{\leq} k \cdot \mathcal{P}(\mathrm{Z}^{t*}) + \mathcal{O}(T^{\varsigma_3}),
\end{aligned}
\tag{35}
$$

where $\bar{\mathrm{Z}}_3^{t}$ is the online result for $\mathcal{P}_3$. Inequality (a) holds since $\mathbb{E}[\bar{\boldsymbol{x}}^t] = \widetilde{\boldsymbol{x}}^t$, $\mathbb{E}[\bar{\boldsymbol{y}}^t] = \widetilde{\boldsymbol{y}}^t$ and $\mathbb{E}[\bar{\boldsymbol{z}}^t] = \widetilde{\boldsymbol{z}}^t$ in Algorithm 3; inequality (b) holds since $\mathrm{Reg}_3^T \leq \mathcal{O}(T^{\varsigma_3})$ in Theorem 2; inequality (c) holds since inequality (34).

**Step IV:** We note that the $\mathcal{O}(T^{\varsigma_3})$ notation used in (35) can be made explicit: there exists a constant $C_3 > 0$ such that the bound is at most $C_3T^{\varsigma_3}$. With this explicit constant, we start from the additive bound established in the previous step:

$$\mathbb{E}[\mathcal{P}(\bar{\mathrm{Z}}^t)] \leq k \cdot \mathcal{P}(\mathrm{Z}^{t*}) + C_3T^{\varsigma_3}, \qquad \varsigma_3 \in (0,1). \tag{36}$$

If we simply divide both sides by $\mathcal{P}(\mathrm{Z}^{t*})$, we obtain

$$\frac{\mathbb{E}[\mathcal{P}(\bar{\mathrm{Z}}^t)]}{\mathcal{P}(\mathrm{Z}^{t*})} \leq k + \frac{C_3T^{\varsigma_3}}{\mathcal{P}(\mathrm{Z}^{t*})}. \tag{37}$$

The second term on the right does not necessarily vanish as $T$ grows unless we know that $\mathcal{P}(\mathrm{Z}^{t*})$ grows at least linearly in $T$; otherwise the ratio might not be bounded by a constant.

To address this, we introduce a mild non-degeneracy assumption: there exists a constant $c_{\min} > 0$ such that for every time slot $t$ and any feasible solution, the per-slot non-switching cost satisfies $f_{-S}^t(\cdot) \geq c_{\min}$. This assumption is natural because energy prices are positive and workloads are non-zero, implying a basic cost per time slot. Consequently, the offline optimal social cost must satisfy

$$\begin{aligned}\mathcal{P}(\mathrm{Z}^{t*}) &= \textstyle\sum_{t=1}^T f_S^t(\mathrm{Z}^{t*}) + \sum_{t=1}^T f_{-S}^t(\mathrm{Z}^{t*})\\ &\geq \textstyle\sum_{t=1}^T f_{-S}^t(\cdot) \geq \sum_{t=1}^T c_{\min} = c_{\min}T,\end{aligned} \tag{38}$$

i.e., $\mathcal{P}(\mathrm{Z}^{t*}) = \Omega(T)$. Since $\varsigma_3 < 1$, the additive term $C_3T^{\varsigma_3}$ is of lower order than $T$, and therefore

$$\lim_{T\to\infty} \frac{C_3T^{\varsigma_3}}{\mathcal{P}(\mathrm{Z}^{t*})} \leq \lim_{T\to\infty} \frac{C_3T^{\varsigma_3}}{c_{\min}T} = \lim_{T\to\infty} \frac{C_3}{c_{\min}} T^{\varsigma_3-1} = 0. \tag{39}$$

Now, dividing both sides of (36) by $\mathcal{P}(\mathrm{Z}^{t*})$ and taking the limit superior yields

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathcal{P}(\bar{\mathrm{Z}}^t)]}{\mathcal{P}(\mathrm{Z}^{t*})} \leq \limsup_{T\to\infty} \left(k + \frac{C_3T^{\varsigma_3}}{\mathcal{P}(\mathrm{Z}^{t*})}\right) = k. \tag{40}$$

Thus, under the non-degeneracy assumption, the online algorithm achieves an asymptotic competitive ratio of $k$ for sufficiently large $T$. □

### *D. Proof of Theorem 4*

*Proof.* A randomized auction is *truthfulness* in expectation if and only if the randomized auction satisfies (i) $\mathbb{E}[\bar{x}_{in}^t]$ is monotonically non-increasing in $r_{in}^t$; (ii) $\int_0^\infty \mathbb{E}[\bar{x}_{in}^t]dr < \infty$.

For condition (i), we define $\widetilde{x}_{in}^t$ and $\widetilde{\dot{x}}_{in}^t$ as the optimal fractional results of the bidding price $r_{in}^t$ and $\dot{r}_{in}^t$. In the case that $r_{in}^t \geq \dot{r}_{in}^t$, we can therefore derive

$$f_{-S}^t(\widetilde{x}_{in}^t, r_{in}^t, r_{-in}^t) \leq f_{-S}^t(\widetilde{\dot{x}}_{in}^t, r_{in}^t, r_{-in}^t), \tag{41}$$

$$f_{-S}^t(\widetilde{\dot{x}}_{in}^t, \dot{r}_{in}^t, r_{-in}^t) \leq f_{-S}^t(\widetilde{x}_{in}^t, \dot{r}_{in}^t, r_{-in}^t). \tag{42}$$

By combining the inequalities (41) and (42) together, based on $\mathbb{E}[\bar{x}_{in}^t] = \widetilde{x}_{in}^t$, we consequently arrive at

$$\begin{aligned}&\widetilde{x}_{in}^t(r_{in}^t - \dot{r}_{in}^t) \leq \widetilde{\dot{x}}_{in}^t(r_{in}^t - \dot{r}_{in}^t)\\ &\Rightarrow \widetilde{x}_{in}^t \leq \widetilde{\dot{x}}_{in}^t \Rightarrow \mathbb{E}[\bar{x}_{in}^t] \leq \mathbb{E}[\bar{\dot{x}}_{in}^t],\end{aligned} \tag{43}$$

then $\mathbb{E}[\bar{x}^t]$ is monotonically non-increasing in $r_{in}^t$.

For condition (ii), as $\varkappa_{in}^t$ is the maximum payment, we get

$$\textstyle\int_0^\infty \mathbb{E}[\bar{x}_{in}^t(r, r_{-in}^t)]dr = \int_0^{\varkappa_{in}^t} \mathbb{E}[\bar{x}_{in}^t(r, r_{-in}^t)]dr \leq \varkappa_{in}^t < \infty. \tag{44}$$

As for *individual rationality*, since $\mathbb{E}[\bar{x}_{in}^t] = \widetilde{x}_{in}^t$, we have

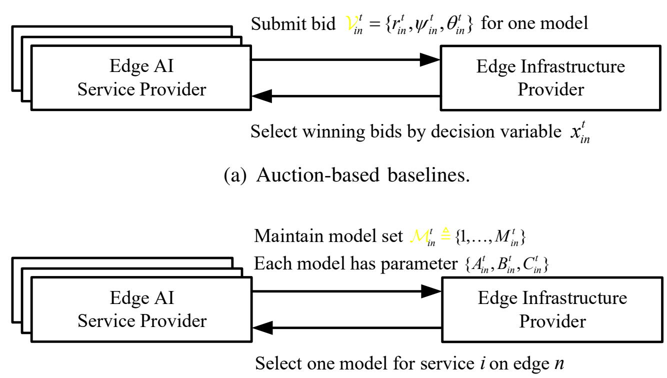
$$\begin{aligned}u_{in}^t &= p_{in}^t - r_{in}^t \mathbb{E}[\bar{x}_{in}^t(r_{in}^t, r_{-in}^t)]\\ &= \textstyle\int_{r_{in}^t}^{\varkappa_{in}^t} \mathbb{E}[\bar{x}_{in}^t(r, r_{-in}^t)]dr \geq 0.\end{aligned} \tag{45}$$

Following this, we will be able to reliably obtain it. □

### *E. An Intuitive Comparison between Auction-based and Non-auction Baselines (shown in Fig. 1 and Fig. 2)*

We have the following notations for our baselines:

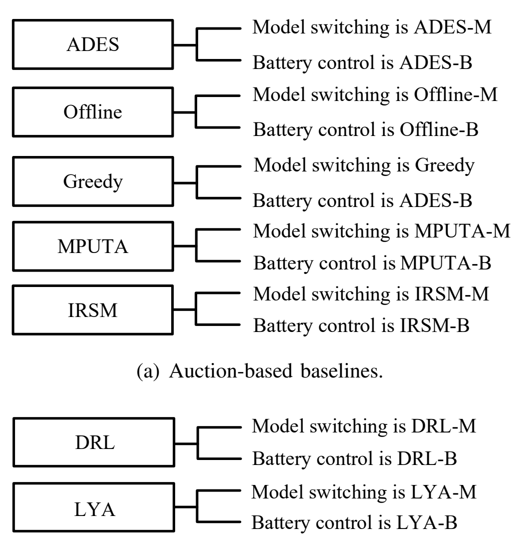
(i) For incentive mechanism, we should choose auction mechanism or non-auction mechanism. (ii) For model switching, if we choose the auction mechanism, we need to decide $x_{in}^t$, it can be solved by ADES-M, Offline-M, Greedy, MPUTA-M and IRSM-M; if we choose non-auction mechanism, we need to choose a model for model set, it can be solved by DRL-M and LYA-M. (iii) For battery control, we need to decide $\{y_n^t, z_n^t, q_n^t, u_n^t, v_n^t, w_n^t\}$, this can be solved by ADES-B, Offline-B, MPUTA-B, IRSM-B, DRL-B and LYA-B.




(a) Auction-based baselines.

(b) Non-auction baselines.

Fig. 1. Comparison between auction-based and non-auction baselines.




(a) Auction-based baselines.

(b) Non-auction baselines.

Fig. 2. Composition of auction-based and non-auction baselines.

### F. Detailed Isolation Setting

The problem consists of three components: (i) incentive mechanism: auction or non-auction; (ii) model switching; (iii) battery scheduling. In fact, there are two pairs of coupling relationships: (i) The incentive mechanism (auction or non-auction) is coupled with the model switching. For example, based on auction mechanism, the model switching can only be solved by ADES-M, Offline-M, Greedy, MPUTA-M and IRSM-M; based on non-auction mechanism, the model switching can only be solved by DRL-M or LYA-M. (ii) The model switching and battery scheduling are coupled together due to the constraint (1g) "$\sum_{i\in\mathcal{I}_n}\big(\lambda_{in}^t(\rho_{in}^t-\psi_{in}^t x_{in}^t)\big) \le q_n^t + w_n^t, \forall n, \forall t$" (i.e., model switching also affects the energy consumption thus affects the battery level). We can: (i) solve all of them all at once using one method (to evaluate coupled performance); (ii) solve them separately using different methods (to evaluate independent performance).

Thus, we design two types of experiments:

- Basic evaluation in Section VI-B: Use the baselines directly (ADES, Offline, Greedy, MPUTA, IRSM, DRL, LYA), they can all solve the problem independently.
- Isolation studies in Section VI-D: (i) For incentive mechanism, since it is coupled with model switching, we use our approach (ADES-M) and the non-auction methods (DRL-M and LYA-M) for evaluation. At the same time, we keep using ADES-B for the battery scheduling unchanged. (ii) For model switching, since it is coupled with the incentive mechanism, we evaluate our method (ADES-M), and other model switching methods. The latter include auction-based methods (IRSM-M, MPUTA-M, and Offline-M), and non-auction methods (DRL-M and LYA-M). Meanwhile, we keep using ADES-B for the battery scheduling unchanged. (iii) For battery scheduling, we evaluate our method (ADES-B) along with other methods (IRSM-B, MPUTA-B, Offline-B, DRL-M, and LYA-M). Meanwhile, we keep using the auction mechanism with ADES-M for model switching unchanged.

### G. Sensitivity Analysis of Intra-slot Variations

We propose two insights for sensitivity analysis and how to choose slot length to deal with the intra-slot variations.

**Insights:** (i) *Finer granularity incurs unavoidable switching and bidding overhead.* If the grid mandates a finer aggregation period (e.g., 5 or 1 minute), we must match that frequency. The resulting cost increase is intrinsic to more frequent adaptations, not a flaw of our algorithm. (ii) *Computational overhead remains practical:* Reducing the slot length increases the number of decision points over a fixed time horizon, which linearly increases computational load. However, our profiling shows that the most time-consuming component (payment allocation) takes at most a few seconds per slot, while other components finish in hundreds of milliseconds. Even with a 1-minute slot, this is still well below the slot length (60 seconds), leaving ample time for communication and execution. Therefore, the computational overhead is acceptable, and finer-grained control is practically feasible in VPP scheduling.

**Strategies for choosing slot length:** The choice depends on two factors: the resolution of available data (data is constant), and the magnitude of intra-slot variations (data can reflect intra-slot variations). There are 3 principles: (i) *Data resolution is the primary constraint.* The slot length cannot be smaller than the time resolution of the input data. If energy prices, VPP demand, and requests are given every 15 minutes, the finest meaningful slot is 15 minutes. Our main experiments respect this. (ii) *When variations are small, a larger slot length (e.g., 15 min) is preferable.* If energy prices, VPP demand, and requests fluctuate only mildly within a 15-minute window, a smaller slot leads to frequent model switches and the associated switching and bidding costs. Meanwhile, the inference error rate and other cost decrease very little, leading to an increased social cost. (iii) *When variations are large, a smaller slot length (e.g., 5 min or 1 min) can be beneficial.* If sharp spikes or dips occur within a 15-minute period, a finer-grained controller can adapt quickly, potentially reducing inference error rate or other cost. However, the benefit must be weighed against the increased switching and bidding overhead. In our experiments with high volatility, the reduction in costs exceeded the increase, resulting in a decrease in the social cost; for highly volatile real-world data, the trade-off may shift.

### H. Asymptotic Competitiveness after Applying Projection

The projection step modifies only the solutions $\{u_n^t, v_n^t, w_n^t\}$ of $\mathcal{P}$ to $\{\hat{u}_n^t, \hat{v}_n^t, \hat{w}_n^t\}$. Therefore, the only cost difference between the projected solution $\hat{\mathrm{Z}}^t$ and the non-projected solution $\bar{\mathrm{Z}}^t$ comes from "the part of the social cost realted to $\{u_n^t, v_n^t, w_n^t\}$". We explicitly write this part as a function:

$$\begin{aligned} f_{\text{uvw}}^t &\triangleq \textstyle\sum_{n=1}^N (s_-^t + d_n\eta_c)u_n^t \\ &\quad + \textstyle\sum_{n=1}^N(-s_+^t + d_n/\eta_d)v_n^t + \sum_{n=1}^N (d_n/\eta_d)w_n^t, \end{aligned} \tag{46}$$

where the coefficients of $u_n^t, v_n^t, w_n^t$ are

$$\alpha_n^t \triangleq s_-^t + d_n\eta_c, \quad \beta_n^t \triangleq -s_+^t + d_n/\eta_d, \quad \gamma_n \triangleq d_n/\eta_d. \tag{47}$$

These coefficients are bounded in magnitude because

- Energy prices $s_-^t, s_+^t$ are finite and bounded in practice (they come from real-world VPP markets).
- Battery degradation cost $d_n$ is finite (computed from battery acquisition cost and cycle life).
- Charging/discharging efficiencies $\eta_c, \eta_d \in (0,1)$.

Based on the above analysis, there exists a constant

$$L' \triangleq \max_{t,n} \big\{|\alpha_n^t|,\ |\beta_n^t|,\ |\gamma_n|\big\} < \infty. \tag{48}$$

Now, for any two decisions $\hat{\mathrm{Z}}^t$ and $\bar{\mathrm{Z}}^t$ that differ only in $\{u_n^t, v_n^t, w_n^t\}$, we have the following bound:

$$\begin{aligned} &\big|f_{\text{uvw}}^t(\hat{\mathrm{Z}}^t) - f_{\text{uvw}}^t(\bar{\mathrm{Z}}^t)\big| \\ &\le L' \textstyle\sum_{n=1}^N \big(|\hat{u}_n^t - u_n^t| + |\hat{v}_n^t - v_n^t| + |\hat{w}_n^t - w_n^t|\big). \end{aligned} \tag{49}$$

Since $u_n^t, v_n^t, w_n^t$ are bounded within $[O_n, B_n]$, we can get

$$\begin{aligned} &\textstyle\sum_{n=1}^N \big(|\hat{u}_n^t - u_n^t| + |\hat{v}_n^t - v_n^t| + |\hat{w}_n^t - w_n^t|\big) \\ &\le \textstyle\sum_{n=1}^N 3(B_n - O_n) \triangleq C, \end{aligned} \tag{50}$$

which leads to

$$\left|f^t_{\mathrm{uvw}}(\hat{\mathrm{Z}}^t) - f^t_{\mathrm{uvw}}(\bar{\mathrm{Z}}^t)\right| \leq L'C \triangleq L. \tag{51}$$

Then we obtain the bound:

$$f^t_{-S}(\hat{\mathrm{Z}}^t) - f^t_{-S}(\bar{\mathrm{Z}}^t) \leq L \tag{52}$$

Summing (52) over $T$ time slots gives

$$\textstyle\sum_{t=1}^{T} f^t_{-S}(\hat{\mathrm{Z}}^t) - \sum_{t=1}^{T} f^t_{-S}(\bar{\mathrm{Z}}^t) \leq LT. \tag{53}$$

Then we can obtain

$$\mathbb{E}[\mathcal{P}(\hat{\mathrm{Z}}^t)] - \mathbb{E}[\mathcal{P}(\bar{\mathrm{Z}}^t)] \leq LT. \tag{54}$$

Combining this with (36), we get

$$\mathbb{E}[\mathcal{P}(\hat{\mathrm{Z}}^t)] \leq k \cdot \mathcal{P}(\mathrm{Z}^{t*}) + C_3 T^{\varsigma_3} + LT, \quad \varsigma_3 \in (0,1). \tag{55}$$

If we simply divide both sides by $\mathcal{P}(\mathrm{Z}^{t*})$, we obtain

$$\frac{\mathbb{E}[\mathcal{P}(\hat{\mathrm{Z}}^t)]}{\mathcal{P}(\mathrm{Z}^{t*})} \leq k + \frac{C_3 T^{\varsigma_3} + LT}{\mathcal{P}(\mathrm{Z}^{t*})}. \tag{56}$$

According to (38), i.e., $\mathcal{P}(\mathrm{Z}^{t*}) \geq c_{\min} T$, we can acquire

$$\begin{aligned}
&\lim_{T\to\infty} \frac{C_3 T^{\varsigma_3} + LT}{\mathcal{P}(\mathrm{Z}^{t*})} \leq \lim_{T\to\infty} \frac{C_3 T^{\varsigma_3} + LT}{c_{\min} T} \\
&= \lim_{T\to\infty} \left(\frac{C_3}{c_{\min}} T^{\varsigma_3 - 1} + \frac{L}{c_{\min}}\right) = \frac{L}{c_{\min}} \triangleq k_2.
\end{aligned} \tag{57}$$

Now, taking the limit superior of both sides of (56) yields

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{\mathbb{E}[\mathcal{P}(\hat{\mathrm{Z}}^t)]}{\mathcal{P}(\mathrm{Z}^{t*})} \leq \limsup_{T\to\infty} \left(k + \frac{C_3 T^{\varsigma_3} + LT}{\mathcal{P}(\mathrm{Z}^{t*})}\right) \\
&= k + k_2 \triangleq k'.
\end{aligned} \tag{58}$$

Thus, after the projection step, the problem $\mathcal{P}$ achieves an asymptotic competitive ratio of $k'$ for sufficiently large $T$.