Incentive-Based Load Curtailment with Limited Information: A Bilevel Zeroth-Order Learning Approach

Zhisen Jiang, Florian Dörfler, Saverio Bolognani

ETH Zurich

Suggested Citation

Z. Jiang, F. Dörfler, and S. Bolognani, "Incentive-Based Load Curtailment with Limited Information: A Bilevel Zeroth-Order Learning Approach," in *Proceedings of the PowerUp Conference*, 2026.

BibTeX Entry

```
@inproceedings{jiang2026incentivebased,
  author = {Jiang, Zhisen and D{\"o}rfler, Florian and Bolognani, Saverio},
  title = {Incentive-Based Load Curtailment with Limited Information: A Bilevel
      Zeroth-Order Learning Approach},
  booktitle = {Proceedings of the PowerUp Conference},
  year = {2026},
}
```

Incentive-Based Load Curtailment with Limited Information: A Bilevel Zeroth-Order Learning Approach

Zhisen Jiang, Florian Dörfler, and Saverio Bolognani

***Abstract*—Incentive-based load curtailment unlocks critical demand-side flexibility but is hindered by the limited knowledge of private user parameters and the inherent nonsmoothness of responses due to physical device constraints. We address this via a constrained bilevel optimization framework and propose the Bi-ZOL (Bilevel Zeroth-Order Learning) algorithm. Unlike conventional black-box methods, Bi-ZOL exploits the bilevel structure to decompose the hypergradient, integrating the exact analytical information of the SO's objective with a zeroth-order estimate of the unknown response sensitivity. This structural decomposition-based learning method mathematically smoothes the nonsmooth response landscape and reduces hypergradient estimation error. We provide theoretical convergence guarantees to an approximate stationary point and demonstrate through simulations that Bi-ZOL achieves near-optimal performance.**

***Index Terms*—Load Curtailment, Bilevel Optimization, Zeroth-Order Estimation, Nonsmooth Optimization.**

I. Introduction

The rapid growth of renewable energy resources and the increasing volatility of demand profiles pose significant challenges for System Operators (SOs) in maintaining power balance. Conventionally, when demand exceeds available generation, SOs maintain balance by ramping up or committing reserve generation. However, relying on peaking units can be economically inefficient and is limited by ramping constraints. Alternatively, SOs can activate demand-side flexibility by temporarily curtailing load from end-users. In this context, demand-side flexibility, e.g., controllable load curtailment from end-users, has emerged as a key resource to support system-level power balance.

This paradigm is reflected in real-world power systems where verified demand-side load curtailment is treated as an operational resource. For instance, the PJM Interconnection has established that demand resources can substitute for generation by providing identifiable capacity services when dispatched [1]. Similarly, in France, RTE has implemented mechanisms enabling demand-response participants to bid verified load reductions into electricity markets and balancing arrangements [2]. In California, CAISO has introduced market participation models under which demand-response resources can offer verified load reductions into wholesale markets [3]. Furthermore, policy reforms in Japan have explicitly institutionalized "negawatt trading," validating the procurement of demand reductions as a flexibility resource [4].

Manuscript received January 30, 2026; accepted April 10, 2026. This work was supported by the Swiss Federal Office of Energy (grant SI/502734 MAESTRO) and by the Swiss National Science Foundation under the NCCR Automation (grant agreement 51NF40_225155).

The authors are with the Automatic Control Laboratory, ETH Zurich, 8092 Zurich, Switzerland. Email: {zhijiang, dorfler, bsaverio}@ethz.ch.

While these examples demonstrate that demand curtailment can function as an operational substitute for upward generation, the practical question is how the SO reliably and efficiently achieves the desired curtailment.

Broadly, approaches range from direct operator control to incentive-based procurement that leverages voluntary participant responses. While Direct Load Control (DLC) has been employed to manage loads, such centralized approaches face privacy concerns. They necessitate invasive access to end-users' private constraints and utility functions. Consequently, there is a paradigm shift towards incentive-based load curtailment. In this framework, the SO uses financial incentives to induce desired load reduction among self-interested end-users in a privacy-preserving and incentive-compatible manner. Existing literature has explored various incentive forms. Differentiated contracts that distinguish between automated and voluntary curtailment are designed based on user opportunity costs [5]. Risk-aware contract designs have been developed to ensure system reliability while respecting individual utility [6], [7]. Recently, dynamic pricing for demand response [8] and price-based flexibility contracts with competitive guarantees [9] are proposed. Moreover, applications of mechanism design provide the theoretical foundation for ensuring that these incentives remain robust against strategic behavior [10].

From the SO's perspective, the primary objective is to determine the optimal incentives that elicit sufficient response from end-users while minimizing the total payment for the curtailment [11]. However, designing optimal incentives is intrinsically a complex optimization problem characterized by limited knowledge of end-users' private information. This opacity necessitates a learning-based approach, where the SO optimizes incentives using observed feedback rather than end-users' private parameters. Nevertheless, a critical challenge arises because end-users are subject to strict physical limits (e.g., device capacities). When these limits are reached, the response naturally saturates, causing abrupt changes in how load reacts to the incentive. This saturation creates "kinks" in the response profile, which renders the underlying function mathematically nonsmooth. This nonsmoothness exposes the limitations of existing learning methods. First-order feedback methods [12], which rely on response sensitivities estimation, may suffer from oscillation because the gradient of response function is undefined at saturation points. Conversely, standard zeroth-order (ZO) methods utilize smoothing schemes to bypass these differentiability issues [12], [13]. While these methods are robust to nonsmoothness, they treat the problem as a complete black box. They fail to exploit the known structure of the SO's cost function, leading to significant suboptimality and computational inefficiency.

These challenges motivate us to explicitly formulate the

incentive-based load curtailment as a constrained bilevel optimization problem, a framework that naturally captures the hierarchical structure while explicitly modeling the user-side decision-making process. To solve this problem, we propose the Bi-ZOL (Bilevel Zeroth-Order Learning) algorithm. Distinguishing itself from standard black-box approaches that indiscriminately estimate the gradient of the entire objective function, Bi-ZOL exploits the bilevel structure to decompose the hypergradient via the chain rule. This decomposition reveals that the nonsmoothness stems solely from the response sensitivity term (the Jacobian of the lower-level solution). Consequently, Bi-ZOL focuses on learning this aggregate response sensitivity directly from end-user feedback via a zeroth-order gradient estimator, which effectively provides a smoothed approximation of the landscape. By isolating the learning target, our method incorporates the exact analytical gradient of the SO's known cost function, thereby enhancing optimality while rigorously handling the nonsmooth nature of user behaviors.

In this paper, we first formulate the incentive-based load curtailment as a bilevel optimization problem and derive the piecewise affine structure of the users' response, identifying it as the source of nonsmoothness in Section II. In Section III, we present our Bi-ZOL algorithm and provide theoretical convergence guarantees to an approximate stationary point. Then, we illustrate the algorithm's performance through numerical simulations in Section IV. Finally, we discuss the remaining open challenges in the Conclusions.

II. Problem: Incentive-Based Load Curtailment

In this section, we formulate incentive-based load curtailment as a bilevel optimization problem. The SO, acting as the leader, determines and broadcasts *nodal curtailment incentives* to procure demand reductions from self-interested end-users connected to each node. As followers, end-users react to these incentives by optimally adjusting their heterogeneous flexible loads subject to device limits.

Our mathematical formulation is similar to the load control framework in [12]. In contrast to [12], which models the end-user response as a generic black-box mapping subject to high-level behavioral properties (e.g., monotonicity), we explicitly characterize the decision-making process as a lower-level optimization problem. This structural approach allows us to mathematically derive the specific nature of the nonsmoothness arising from device constraints. For clarity, we simplify the load control formulation by focusing solely on active load reduction requirements and excluding the underlying grid power flow constraints.

A. SO's upper-level problem (leader)

Let $\mathcal{N}$ denote the set of nodes, with $|\mathcal{N}| = N$. We define the following vectors in $\mathbb{R}^N$:

- $\boldsymbol{\lambda}$ nodal curtailment incentives, where λ_i is the incentive broadcast at node i;
- $\mathbf{R}$ aggregate nodal load reduction responses, where R_i is the total reduction at node i contributed by connected end-users;
- $\mathbf{P}^{\mathrm{b}}$ baseline loads, where P_i^{b} is the baseline load of node i.

The SO aims to minimize total incentive payments while driving the total system load toward a target P^{target}. Given a price vector $\boldsymbol{\lambda}$ and response $\mathbf{R}$, we define the global load mismatch as:

$$E \coloneqq \mathbf{1}^\top\big(\mathbf{P}^{\mathrm{b}} - \mathbf{R}\big) \;-\; P^{\mathrm{target}}. \tag{1}$$

The SO solves the following optimization problem:

$$\min_{\boldsymbol{\lambda}\in\Lambda,\mathbf{R}} \; \varphi(\boldsymbol{\lambda},\mathbf{R}) \;\coloneqq\; \boldsymbol{\lambda}^\top\mathbf{R} \;+\; \rho E^2, \tag{2}$$

where $\rho > 0$ is a penalty parameter and Λ is the admissible price set.

Note that $\varphi(\boldsymbol{\lambda},\mathbf{R})$ is smooth and Lipschitz continuous on compact sets. We adopt the following standard assumption regarding the constraints:

Assumption 1. *The set Λ is nonempty, convex, compact, and bounded with diameter D, i.e., $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \le D$ for any $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda$.*

This assumption is readily satisfied in practice, typically taking the form of box constraints $\Lambda \coloneqq [\underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}]$.

B. End-users' lower-level problem (followers)

End-users are partitioned by nodes. Let $\mathcal{U}$ be the set of users and $\mathcal{U}_i \subseteq \mathcal{U}$ denote the subset of users connected to node i (where $\{\mathcal{U}_i\}_{i\in\mathcal{N}}$ are disjoint). Each user u controls a set of flexible devices indexed by $\mathcal{K}_u$. For each device $k \in \mathcal{K}_u$, let $r_{u,k} \in \mathbb{R}$ denote the load reduction, subject to the capacity constraint:

$$0 \le r_{u,k} \le C_{u,k}, \tag{3}$$

where $C_{u,k} > 0$ represents the maximum load reduction for the device, which is determined by the device capacity.

Given the nodal incentive λ_i, each end-user $u \in \mathcal{U}_i$ minimizes their net cost (discomfort minus incentive payment). The decision problem for user u is formulated as:

$$\min_{\{r_{u,k}\}_{k\in\mathcal{K}_u}} \quad \sum_{k\in\mathcal{K}_u} \Big(\frac{1}{2\alpha_{u,k}} r_{u,k}^2 - \lambda_i r_{u,k}\Big) \tag{4a}$$

$$\text{s.t.} \;\; 0 \le r_{u,k} \le C_{u,k}, \qquad \forall k \in \mathcal{K}_u, \tag{4b}$$

where $\alpha_{u,k} > 0$ denotes the device discomfort sensitivity. Unlike [12], our formulation explicitly models the aggregation of heterogeneous devices. The quadratic term captures the *increasing marginal disutility* of load reduction. For example, while dimming lights slightly may cause negligible annoyance, curtailing HVAC usage during extreme weather incurs significantly higher discomfort per unit of energy reduced.

Remark 1. *The proposed quadratic discomfort model* (4a) *implies that all devices begin reducing load as soon as $\lambda_i > 0$. This formulation can be trivially extended to include a linear discomfort term $\beta_{u,k} r_{u,k}$ (where $\beta_{u,k} > 0$), representing a minimum "activation cost" for device k. Physically, a device only participates when the incentive λ_i exceeds its specific discomfort threshold $\beta_{u,k}$. This allows the model to capture sequential activation behaviors, where users prioritize*

"cheaper" reduction before committing to "expensive" devices as incentives increase.

Let $\{r^\star_{u,k}(\lambda_i)\}_{k\in\mathcal{K}_u}$ denote the optimal solution to (4). The aggregate nodal response is then given by:

$$R_i(\lambda_i) \coloneqq \sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} r^\star_{u,k}(\lambda_i). \tag{5}$$

We now characterize the properties of $\mathbf{R}$ to identify the source of nonsmoothness in the bilevel problem.

Proposition 1. *Fix a node $i \in \mathcal{N}$. The aggregate nodal response $R_i(\lambda_i)$ satisfies the following properties:*

1) *(*Uniqueness*) For every $\lambda_i \in \mathbb{R}$, the response $R_i(\lambda_i)$ is unique.*
2) *(*Closed-form piecewise affine structure*) $R_i(\lambda_i)$ admits the closed-form expression:*

$$R_i(\lambda_i) = \sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} \min\big\{C_{u,k}, \max\{0, \alpha_{u,k}\lambda_i\}\big\}. \tag{6}$$

Consequently, $R_i(\cdot)$ is piecewise affine with breakpoints contained in the set $\{0\}\cup\{C_{u,k}/\alpha_{u,k} : k \in \mathcal{K}_u, u \in \mathcal{U}_i\}$.
3) *(*Lipschitz Continuity*) $R_i(\cdot)$ is globally Lipschitz continuous with constant:*

$$L_i \coloneqq \sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} \alpha_{u,k}.$$

Proof. We start the proof with the device optimal solution $r^\star_{u,k}(\lambda_i)$ admitted by (4). Since $\alpha_{u,k} > 0$ for all k, the Hessian matrix is diagonal with strictly positive entries $\{\frac{1}{\alpha_{u,k}}\}_{k\in\mathcal{K}_u}$, the objective is strictly convex in $\{r_{u,k}\}$. Together with the nonempty compact constraints (4b) and first order optimality condition, for any device (u,k) and any $\lambda_i \in \mathbb{R}$, the problem (4) admits a unique optimizer given by

$$r^\star_{u,k}(\lambda_i) = \min\big\{C_{u,k}, \max\{0, \alpha_{u,k}\lambda_i\}\big\}. \tag{7}$$

Summing these unique device responses over all users at node i proves statements (1) and (2). Regarding (3), $r^\star_{u,k}(\cdot)$ has slope either 0 or $\alpha_{u,k}$ (except at two breakpoints), so it is globally Lipschitz with constant $\alpha_{u,k}$. Therefore,

$$\begin{aligned}|R_i(\lambda) - R_i(\lambda')| &\le \sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} |r^*_{u,k}(\lambda) - r^*_{u,k}(\lambda')| \\ &\le \sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} \alpha_{u,k}|\lambda - \lambda'| \\ &= L_i|\lambda - \lambda'|. \end{aligned}$$

□

To illustrate the piecewise affine responses in Proposition 1, Fig. 1 depicts responses for a node with three connected end-users. The individual device responses exhibit saturation at their capacities C_1, C_2, C_3, and the aggregate R_i is the sum with breakpoints at $\{C_{u,k}/\alpha_{u,k}\}$ as in (6).

Lemma 1. *The aggregate nodal response vector $\mathbf{R}(\boldsymbol{\lambda})$ is globally Lipschitz continuous with constant $L_R \coloneqq \max_{i\in\mathcal{N}} L_i$, and is bounded by $R_{\max} \coloneqq \sqrt{\sum_{i=1}^N (\sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} C_{u,k})^2}$.*

Proof. Using the vector components:

$$\begin{aligned}\|\mathbf{R}(\boldsymbol{\lambda}) - \mathbf{R}(\boldsymbol{\lambda}')\| &= \sqrt{\sum_{i=1}^N |R_i(\lambda_i) - R_i(\lambda'_i)|^2} \\ &\le \sqrt{\sum_{i=1}^N L_i^2|\lambda_i - \lambda'_i|^2} \\ &\le \sqrt{\Big(\max_j L_j^2\Big)\sum_{i=1}^N |\lambda_i - \lambda'_i|^2} \\ &= L_R\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2.\end{aligned}$$

Boundedness follows directly from the fact that $0 \le R_i \le \sum_{u,k} C_{u,k}$. □

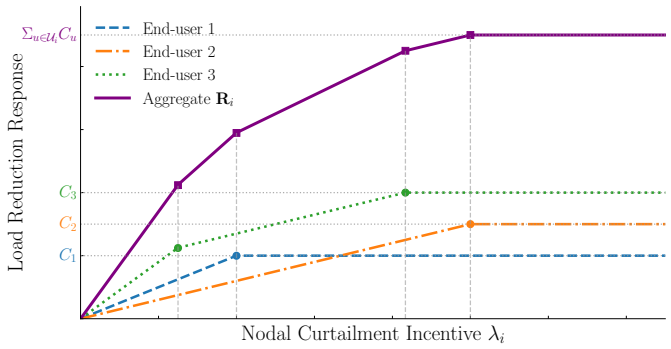

Fig. 1. Responses for a node with three connected end-users: individual load reduction responses and aggregate nodal response R_i as functions of the nodal incentive λ_i.

C. Bilevel Formulation

We now combine the SO's upper-level problem (2) and the end-users' lower-level problem (4). The resulting leader–follower interaction can be written as the following bilevel program:

$$\min_{\boldsymbol{\lambda}\in\Lambda,\mathbf{R}} \ \varphi(\boldsymbol{\lambda},\mathbf{R}) \coloneqq \boldsymbol{\lambda}^\top\mathbf{R} + \rho E^2 \tag{8a}$$

$$\text{s.t. } R_i = \sum_{u\in\mathcal{U}_i}\sum_{k\in\mathcal{K}_u} r^\star_{u,k}(\lambda_i), \quad \forall i \in \mathcal{N},$$

$$E = \mathbf{1}^\top\big(\mathbf{P}^{\mathrm{b}} - \mathbf{R}\big) - P^{\mathrm{target}},$$

$$\forall u \in \mathcal{U}:$$

$$r^\star_{u,k}(\lambda_i) = \arg\min_{k\in\mathcal{K}_u}\left\{\sum_{k\in\mathcal{K}_u}\Big(\frac{1}{2\alpha_{u,k}} r_{u,k}^2 - \lambda_i r_{u,k}\Big)\right\} \tag{8b}$$

$$\text{s.t. } 0 \le r_{u,k} \le C_{u,k}, \quad \forall k \in \mathcal{K}_u. \tag{8c}$$

The bilevel formulation (8) captures the SO's incentive-based load curtailment under limited information. The SO can observe the aggregate nodal response $\mathbf{R}$ after broadcasting the price vector $\boldsymbol{\lambda}$. However, it does not know the closed-form mapping from the incentive vector to the aggregate response because it is induced by private end-user discomfort parameters and device constraints.

The general interaction between the SO and the end-users is illustrated in Fig. 2.

III. Solution: Bilevel Zeroth-Order Learning

In this section, we propose the Bilevel Zeroth-Order Learning (Bi-ZOL) algorithm to solve the bilevel problem (8). The

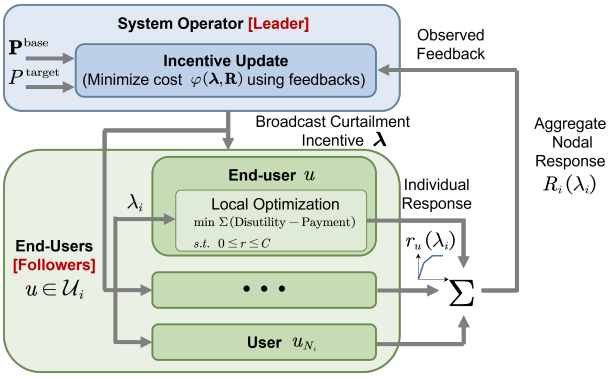

Fig. 2. Incentive-based load curtailment architecture.

core principle is to learn the sensitivity of the aggregate nodal response to incentives from observed feedback using a zeroth-order estimator. From a practical perspective, this learning process circumvents the SO's limited knowledge regarding end-user parameters. Mathematically, this corresponds to a randomized smoothing of the nonsmooth response landscape, which is essential for ensuring algorithmic convergence. Furthermore, unlike standard black-box methods, Bi-ZOL fully utilizes the SO's available analytical information of the upper-level cost function to enhance the precision of the hypergradient estimate.

A. Approximate Stationary Point

To formally introduce the algorithm, we first define the stationary point condition we aim to satisfy. We view the bilevel problem (8) as a general constrained nonsmooth nonconvex optimization problem:

$$\min_{\boldsymbol{\lambda}\in\Lambda} \ \tilde{\varphi}(\boldsymbol{\lambda}) \coloneqq \varphi(\boldsymbol{\lambda},\mathbf{R}(\boldsymbol{\lambda})) = \boldsymbol{\lambda}^\top \mathbf{R}(\boldsymbol{\lambda}) \ + \ \rho E(\boldsymbol{\lambda})^2, \quad (9)$$

where the equilibrium constraints (8b)-(8c) are implicitly captured by the single-valued mapping $\mathbf{R}(\boldsymbol{\lambda})$.

Wherever $\tilde{\varphi}(\cdot)$ is differentiable, the hypergradient is given by the chain rule:

$$\nabla\tilde{\varphi}(\boldsymbol{\lambda}) = \nabla_1\varphi(\boldsymbol{\lambda},\mathbf{R}) + \mathbf{JR}(\boldsymbol{\lambda})^\top \nabla_2\varphi(\boldsymbol{\lambda},\mathbf{R}), \quad (10)$$

where the Jacobian matrix $\mathbf{JR}(\boldsymbol{\lambda})$ captures the sensitivity of the aggregate response to incentives. However, as discussed in Section II-B, $\mathbf{R}(\boldsymbol{\lambda})$ is nonsmooth, possessing numerous points where the Jacobian is undefined. This renders standard gradient-based methods invalid.

To address this, in our algorithm, we propose to approximate the ill-defined hypergradient by replacing the $\mathbf{JR}(\boldsymbol{\lambda})$ with the approximate Jacobian $\mathbf{JR}_\delta(\boldsymbol{\lambda})$, where $\mathbf{R}_\delta(\boldsymbol{\lambda})$ is the smoothed response function by randomized smoothing defined as:

$$\mathbf{R}_\delta(\boldsymbol{\lambda}) = \mathbb{E}_{\mathbf{u}\sim\text{Unif}(\mathbb{B})}\left[\mathbf{R}(\boldsymbol{\lambda}+\delta\mathbf{u})\right], \quad (11)$$

where $\text{Unif}(\mathbb{B})$ is the uniform distribution over the closed unit ball and $\delta > 0$ is the smoothing radius. Consequently, we have our approximate hypergradient:

$$g_\delta(\boldsymbol{\lambda}) = \nabla_1\varphi(\boldsymbol{\lambda},\mathbf{R}) + \mathbf{JR}_\delta(\boldsymbol{\lambda})^\top \nabla_2\varphi(\boldsymbol{\lambda},\mathbf{R}). \quad (12)$$

Crucially, unlike fully black-box approaches, our formulation retains the exact values of the partial derivatives $\nabla_1\varphi$ and $\nabla_2\varphi$. By isolating the zeroth-order estimation solely to the unknown response sensitivity, (12) significantly reduces estimation error and represents the true descent direction more faithfully.

Note that g_δ is defined everywhere, and it serves as a proxy for the ill-defined hypergradient. Since the true objective violates the standard Lipschitz gradient assumption, we cannot rely on the smooth case first-order optimality conditions ($\nabla\tilde{\varphi}(\boldsymbol{\lambda}^\star) = 0$). Moreover, (9) is constrained. Thus, we define the (δ,ϵ)-Bi-ZOL Frank–Wolfe Stationary Point (FWSP) to characterize stationarity for the constrained, nonsmooth landscape.

Definition 1. *A point $\boldsymbol{\lambda}\in\Lambda$ is a (δ,ϵ)-Bi-ZOL Frank–Wolfe Stationary Point of problem* (9) *if:*

$$\max_{\mathbf{z}\in\Lambda} \langle \mathbf{z}-\boldsymbol{\lambda}, -g_\delta(\boldsymbol{\lambda})\rangle \le \epsilon. \quad (13)$$

This definition relaxes the standard stationarity condition in nonsmooth optimization by allowing for a smoothing error δ and a convergence tolerance ϵ, which is consistent with the literature on zeroth-order optimization for nonsmooth functions [14].

Remark 2. *The rationale for this stationarity definition is grounded in nonsmooth analysis. For a nonsmooth function, standard gradients are ill-defined, and first-order optimality is typically characterized through the Clarke subdifferential $\partial^C\tilde{\varphi}(\boldsymbol{\lambda})$. Our approximate hypergradient g_δ is inspired by the concept of the Goldstein subdifferential, a computationally tractable relaxation of $\partial^C\tilde{\varphi}(\boldsymbol{\lambda})$ constructed by sampling gradients within a δ-neighborhood. g_δ should be interpreted as a structured approximate stationarity direction: it preserves the exact upper-level partial derivatives and smooths only the unknown response-sensitivity term. As the smoothing radius $\delta \to 0$, any limit point of the sequence of (δ,ϵ)-Bi-ZOL FWSPs satisfies the necessary optimality condition defined by $\partial^C\tilde{\varphi}(\boldsymbol{\lambda})$.*

B. Bilevel Zeroth-Order Learning Algorithm

The proposed Bi-ZOL algorithm employs a Frank–Wolfe (FW) update scheme and operates by alternating between learning the response sensitivity via a two-point estimator and updating the incentives along a descent direction.

We leverage the fact that the SO has full analytical knowledge of the upper-level objective φ. At iteration k, given the current pair $(\boldsymbol{\lambda}_k,\mathbf{R}_k)$, the partial gradients $\nabla_1\varphi$ and $\nabla_2\varphi$ are computed exactly:

$$\nabla_1\varphi(\boldsymbol{\lambda},\mathbf{R}) = \mathbf{R}_k, \quad \nabla_2\varphi(\boldsymbol{\lambda},\mathbf{R}) = \boldsymbol{\lambda}_k - 2\rho E(\mathbf{R}_k)\mathbf{1}. \quad (14)$$

The remaining challenge is to estimate the smoothed Jacobian $\mathbf{JR}_\delta(\boldsymbol{\lambda}_k)$. It is a known result in zeroth-order optimization that the gradient of a function smoothed over the unit ball can be estimated using samples from the unit sphere. Accordingly, we generate a random direction $\mathbf{w}\sim\text{Unif}(\mathbb{S}^{N-1})$ and utilize a two-point estimator:

$$\widehat{\mathbf{JR}}(\boldsymbol{\lambda}_k) \ = \ \frac{N}{2\delta}\Big(\mathbf{R}(\boldsymbol{\lambda}_k+\delta\mathbf{w}) - \mathbf{R}(\boldsymbol{\lambda}_k-\delta\mathbf{w})\Big)\mathbf{w}^\top. \quad (15)$$

Algorithm 1 Bilevel Zeroth-Order Learning (Bi-ZOL)

Input: Stepsize γ; smoothing radius $\delta > 0$.
Initialize: Choose $\boldsymbol{\lambda}_0 \in \Lambda$.
for $k = 0, 1, 2, \ldots$ **do**
  Query response: Broadcast $\boldsymbol{\lambda}_k$, observe $\mathbf{R}_k = \mathbf{R}(\boldsymbol{\lambda}_k)$.
  Compute mismatch: $E_k = \mathbf{1}^\top\big(\mathbf{P}^{\mathrm{b}} - \mathbf{R}_k\big) - P^{\mathrm{target}}$.
  Compute partials: $\nabla_1\varphi = \mathbf{R}_k$; $\nabla_2\varphi = \boldsymbol{\lambda}_k - 2\rho E_k\mathbf{1}$.
  ZO Jacobian estimation:
    Sample direction $\mathbf{w} \sim \mathrm{Unif}(\mathbb{S}^{N-1})$.
    Broadcast $\boldsymbol{\lambda}_k + \delta\mathbf{w}$, observe $\mathbf{R}^+ = \mathbf{R}(\boldsymbol{\lambda}_k + \delta\mathbf{w})$.
    Broadcast $\boldsymbol{\lambda}_k - \delta\mathbf{w}$, observe $\mathbf{R}^- = \mathbf{R}(\boldsymbol{\lambda}_k - \delta\mathbf{w})$.
    Estimate $\widehat{\mathbf{J}}\mathbf{R}(\boldsymbol{\lambda}_k) \leftarrow \dfrac{N}{2\delta}\big(\mathbf{R}^+ - \mathbf{R}^-\big)\mathbf{w}^\top$.
  Hypergradient estimate:
    $\widehat{\mathbf{g}}_k \leftarrow \nabla_1\varphi + \widehat{\mathbf{J}}\mathbf{R}(\boldsymbol{\lambda}_k)^\top\nabla_2\varphi$.
  FW linear oracle: $\mathbf{z}_k \in \arg\max_{\mathbf{z}\in\Lambda}\ \langle \mathbf{z}, -\widehat{\mathbf{g}}_k\rangle$.
  FW update: $\boldsymbol{\lambda}_{k+1} \leftarrow \boldsymbol{\lambda}_k + \gamma\big(\mathbf{z}_k - \boldsymbol{\lambda}_k\big)$.
end for

This estimator serves two critical functions: (1) It allows the SO to learn the aggregate behavior of end-users without accessing their private parameters; (2) It effectively estimates the Jacobian of the *smoothed* response $\mathbf{R}_\delta$, thereby addressing the nonsmoothness issues of the original function $\mathbf{R}$.

Combining (14) and (15), Bi-ZOL constructs the hypergradient estimate:

$$\widehat{\mathbf{g}}_k \;=\; \nabla_1\varphi(\boldsymbol{\lambda}_k, \mathbf{R}_k) + \widehat{\mathbf{J}}\mathbf{R}(\boldsymbol{\lambda}_k)^\top\nabla_2\varphi(\boldsymbol{\lambda}_k, \mathbf{R}_k). \tag{16}$$

With $\widehat{\mathbf{g}}_k$, the Frank–Wolfe linear oracle computes the descent direction:

$$\mathbf{z}_k \in \arg\max_{\mathbf{z}\in\Lambda}\ \langle \mathbf{z}, -\widehat{\mathbf{g}}_k\rangle, \tag{17}$$

followed by the update step:

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \gamma\big(\mathbf{z}_k - \boldsymbol{\lambda}_k\big), \tag{18}$$

where $\gamma \in (0, 1]$ is the stepsize. Since Λ is convex, the update (18) ensures that the new incentives remain feasible without requiring projection.

The complete procedure is summarized in Algorithm 1.

Remark 3. *In a practical implementation, the iterative updates of $\boldsymbol{\lambda}$ and the probing of $\mathbf{R}(\boldsymbol{\lambda}\pm\delta\mathbf{w})$ occur within a digital negotiation phase (e.g., via Home Energy Management Systems) prior to the physical load adjustment. The SO exchanges incentives with the HEMS agents to estimate sensitivities without physically altering the load consumption. This avoids the latency and frequent physical actuation. Moreover, it is possible to adapt the algorithm to use a one-point estimator [15], [16]. A one-point approach would require only a single broadcast per iteration, simplifying the communication overhead. We leave the exploration of one-point variants for real-time operations to future work.*

C. Algorithm Convergence

In this section, we establish the convergence of the Bi-ZOL algorithm to a (δ, ϵ)-Bi-ZOL FWSP. The analysis relies on the smoothness properties of the smoothed response function $\mathbf{R}_\delta$ and the ancillary function $\bar{\varphi}_\delta(\boldsymbol{\lambda}) \coloneqq \varphi(\boldsymbol{\lambda}, \mathbf{R}_\delta(\boldsymbol{\lambda}))$. By quantifying the bias introduced by smoothing and the variance of the zeroth-order estimator, we derive the following convergence guarantee.

Theorem 1. *With Assumption 1, the Bi-ZOL algorithm (Algorithm 1) with constant stepsize $\gamma \in (0, 1]$ guarantees the following bound on the expected Frank–Wolfe gap after T iterations:*

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathcal{G}_\delta(\boldsymbol{\lambda}_k)\right] \le \frac{\Delta_0}{\gamma T} + \frac{L_{\nabla\bar{\varphi}_\delta}D^2}{2}\gamma + C_{\mathrm{bias}}D\delta + C_{\mathrm{noise}}D, \tag{19}$$

where $\mathcal{G}_\delta(\boldsymbol{\lambda}) \coloneqq \max_{\mathbf{z}\in\Lambda}\langle \mathbf{z} - \boldsymbol{\lambda}, -g_\delta(\boldsymbol{\lambda})\rangle$ is the FW gap associated with the approximate hypergradient g_δ. The constants are defined as:

- $\Delta_0 \coloneqq \bar{\varphi}_\delta(\boldsymbol{\lambda}_0) - \min_{\boldsymbol{\lambda}}\bar{\varphi}_\delta(\boldsymbol{\lambda})$,
- $L_{\nabla\bar{\varphi}_\delta} \coloneqq L_R(2 + 2\rho N L_R) + \frac{cNL_R}{\delta}(\Lambda_{\max} + 2\rho\sqrt{N}(|c_0| + \sqrt{N}R_{\max}))$,
- $C_{\mathrm{bias}} \coloneqq (1 + 2\rho N L_R)L_R$,
- $C_{\mathrm{noise}} \coloneqq 4M_{\varphi 2}(2\pi)^{1/4}NL_R$, *with* $M_{\varphi 2} \coloneqq \Lambda_{\max} + 2\rho\sqrt{N}(|c_0| + \sqrt{N}R_{\max})$.

The proof of Theorem 1, along with the supporting lemmas regarding the properties of the estimator and the ancillary function, are provided in Appendix.

Remark 4. *The expectation in the convergence measure is due to the randomized zeroth-order estimator, and the iteration average can be interpreted as the expected gap $\mathbb{E}\left[\max_{\mathbf{z}\in\Lambda}\left\langle -g_\delta\left(\boldsymbol{\lambda}_R\right), \mathbf{z} - \boldsymbol{\lambda}_R\right\rangle\right]$ where $\boldsymbol{\lambda}_R$ is uniformly sampled from $\{\boldsymbol{\lambda}_k\}_{k=0}^{T-1}$, as considered in [16].*

IV. Numerical Experiments

In this section, we evaluate the performance of the proposed Bi-ZOL algorithm to solve the incentive-based load curtailment problem on a simulated network.

A. Simulation Setup

The simulations are conducted on a 3-node network populated with 8 end-users randomly connected to the nodes. Each end-user manages 4 distinct flexible devices capable of providing load reduction.

The nodal curtailment incentives are constrained within the range $\Lambda = [0, 5]^3$ to ensure economic feasibility. The end-users are modeled as rational agents minimizing a quadratic cost function as defined in (4). To capture population heterogeneity, the discomfort sensitivity coefficient $\alpha_{u,k}$ for each device is sampled uniformly from the interval $[0.5, 3.0]$. Device capacities $C_{u,k}$ are randomly assigned to reflect diverse appliance ratings.

We implement the Bi-ZOL algorithm with a constant step size $\gamma = 10^{-3}$ and a smoothing radius $\delta = 10^{-3}$, values determined via empirical tuning. To establish a ground truth for optimality, we leverage the structural properties of the lower-level problem. Since the device response function (7) is piecewise affine, the bilevel optimization problem can

be exactly reformulated as a Mixed-Integer Linear Program (MILP) by introducing binary variables to model the saturation states of each device. The global optimal solution $\boldsymbol{\lambda}^*$ and corresponding cost $\tilde{\varphi}(\boldsymbol{\lambda}^*)$ is obtained by solving this MILP using the Gurobi optimizer.

B. Simulation Results

The convergence behavior of the algorithm and comparison with the global optimum is illustrated in Fig. 3. Fig. 3 shows the evolution of the objective function value. Bi-ZOL exhibits a steady descent and stabilizes to the (δ, ϵ)-Bi-ZOL FWSP. Despite treating the user response as a black box, Bi-ZOL achieves a final objective value of 40.51, which is remarkably close to the global optimum of 38.34. The relative optimality gap is approximately 5.66%, demonstrating that the proposed zeroth-order estimator effectively captures the descent direction even in the presence of nonsmooth response functions.

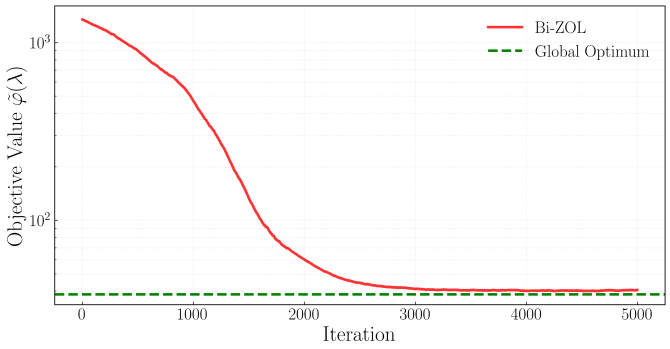

Fig. 3. Convergence performance of Bi-ZOL. The objective value approaches the global optimum .

We further analyze the incentives and the corresponding end-users' load reduction responses. Fig. 4 illustrates the convergence trajectories of the incentives $\boldsymbol{\lambda}$ across three nodes. Throughout the iterations, the algorithm encounters nonsmooth points a total of 614 times, as indicated by the markers on each component's curve. The stable convergence, despite these frequent nonsmooth encounters, demonstrates that the smoothing technique within our Bi-ZOL algorithm effectively addresses the nonsmooth issues.

In Bi-ZOL iterations, λ_1 and λ_3 almost track the optimal prices, while λ_2 is a bit higher. The reason is twofold. First, we actually deal with a nonsmooth nonconvex optimization, in which global optimality is not guaranteed. Our method only guarantees convergence to a (δ, ϵ)-Bi-ZOL FWSP, which is probably close to a local optimum. Second, even though we select a most exact approximation for the hypergradient, the approximation error for the smoothing is still non-negligible.

Fig. 5 illustrates the load reduction response of the end-users. The responses vary significantly across users, confirming the heterogeneity of the population. Notably, at the converged state, four devices belonging to end-users 1, 3, and 5 reach their load reduction limits. This implies that the algorithm ultimately converges to a nonsmooth point. Crucially, all end-users' responses are stable after iterations,

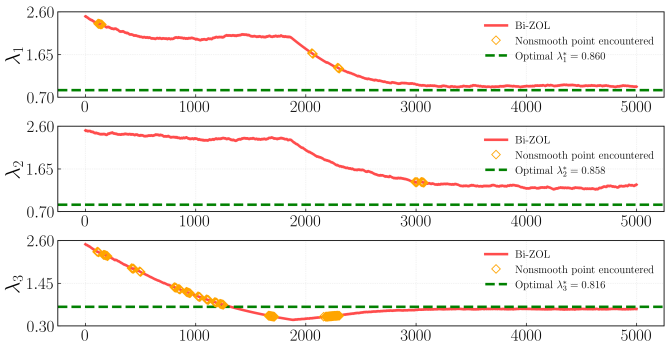

Fig. 4. Incentive Trajectories. The yellow diamonds mark instances where the algorithm encounters nonsmooth points. Two nodal incentives among three nodes converge close to the optimal solution. On the other hand, the incentive of the second node is a bit higher, which is probably due to the learning error and the local optimum.

which means our method is able to learn the end-users' responses and provide a stable incentive signal to control the loads.

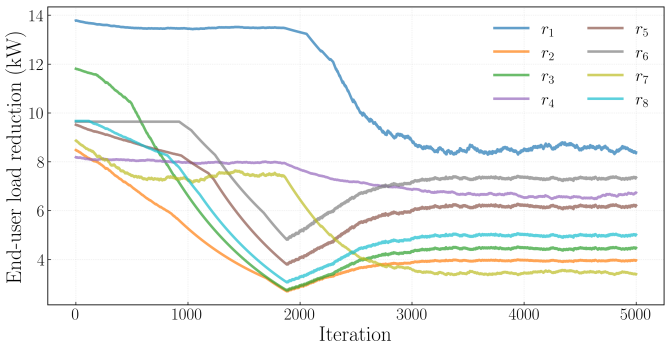

Fig. 5. End-user Response Trajectories. The response of heterogeneous users stabilizes after iterations.

V. Conclusion

This paper presents a bilevel optimization framework for incentive-based load curtailment. By explicitly modeling the decision-making process of end-users subject to device capacity constraints, we identified that standard gradient-based methods fail due to undefined sensitivities at saturation points, while standard zeroth-order methods suffer from inefficiency by ignoring known system structures. To bridge this gap, we proposed the Bi-ZOL algorithm, which synergizes analytical model information with learning response sensitivities. Utilizing a structural hypergradient decomposition and a two-point estimator for sensitivity, Bi-ZOL rigorously handles the nonsmooth points in aggregate responses. Theoretical analysis confirms that Bi-ZOL converges to a (δ, ϵ)-Bi-ZOL Frank–Wolfe stationary point under mild assumptions. Furthermore, numerical experiments validate that our approach stabilizes the incentive near the global optimum. Future work will extend this framework to incorporate grid constraints, exploring how voltage limits affect the learning of response sensitivities. Further research could also investigate methods that reduce active probing, for instance by employing one-point gradient

estimators or leveraging historical feedback to estimate sensitivities. Additionally, investigating the tracking performance of Bi-ZOL in time-varying environments remains a promising direction for enhancing real-time grid operations.

APPENDIX

In this appendix, we provide the detailed proofs for the convergence of the Bi-ZOL algorithm. We use $\|\cdot\|$ to denote the l_2 norm for vectors and the spectral norm for matrices.

A. Properties of the Zeroth-Order Estimator

We begin by characterizing the statistical properties of the zeroth-order Jacobian estimator defined in (15).

Lemma 2. *The two-point estimator* (15) *is an unbiased estimator of the smoothed Jacobian and satisfies the following variance bound:*

$$\begin{aligned} \mathbb{E}\Big[\widehat{\mathbf{J}}\mathbf{R} \mid \boldsymbol{\lambda}\Big] &= \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}), \\ \mathbb{E}\Big[\big\|\widehat{\mathbf{J}}\mathbf{R}\big\|^2 \mid \boldsymbol{\lambda}\Big] &\le 16\sqrt{2\pi}\, N^2 L_R^2. \end{aligned} \tag{20}$$

Proof. For each $i \in \{1,\dots,N\}$, the i-th row of the Jacobian matrix $\widehat{\mathbf{J}}\mathbf{R}$ corresponds to the transposed vector estimator for the gradient of the i-th component function R_i:

$$\widehat{\mathbf{J}}\mathbf{R}_i \;\coloneqq\; \frac{N}{2\delta}\Big(R_i(\boldsymbol{\lambda}+\delta\mathbf{w}) - R_i(\boldsymbol{\lambda}-\delta\mathbf{w})\Big)\mathbf{w}^\top \in \mathbb{R}^{1\times N}.$$

Consider the randomized smoothing of the scalar function R_i over the unit ball:

$$R_{i,\delta}(\boldsymbol{\lambda}) \;\coloneqq\; \mathbb{E}_{\mathbf{u}\sim\mathrm{Unif}(\mathbb{B})}\big[R_i(\boldsymbol{\lambda}+\delta\mathbf{u})\big].$$

From standard properties of zeroth-order estimators on the unit sphere in [17], the expectation of the column estimator is the gradient. Taking the transpose, we have:

$$\mathbb{E}\big[\widehat{\mathbf{J}}\mathbf{R}_i \mid \boldsymbol{\lambda}\big] = \nabla R_{i,\delta}(\boldsymbol{\lambda})^\top.$$

Let $\mathbf{R}_\delta(\boldsymbol{\lambda}) \coloneqq (R_{1,\delta}(\boldsymbol{\lambda}),\dots,R_{N,\delta}(\boldsymbol{\lambda}))^\top$. Stacking the expectations of the rows yields the Jacobian matrix:

$$\mathbb{E}\big[\widehat{\mathbf{J}}\mathbf{R}\big] = \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}).$$

To bound the second moment, we utilize the Frobenius norm decomposition:

$$\big\|\widehat{\mathbf{J}}\mathbf{R}\big\|^2 \le \big\|\widehat{\mathbf{J}}\mathbf{R}\big\|_F^2 = \sum_{i=1}^N \big\|\widehat{\mathbf{J}}\mathbf{R}_i\big\|^2.$$

Applying the bound for two-point estimators on Lipschitz functions from [17, Lemma D.1], we have for each i:

$$\mathbb{E}\big[\|\widehat{\mathbf{J}}\mathbf{R}_i\|^2 \mid \boldsymbol{\lambda}\big] \le 16\sqrt{2\pi}\, N\, L_i^2,$$

where L_i is the Lipschitz constant of $R_i(\cdot)$. Summing over all N rows yields:

$$\mathbb{E}\big[\|\widehat{\mathbf{J}}\mathbf{R}\|^2 \mid \boldsymbol{\lambda}\big] \le 16\sqrt{2\pi}\, N \sum_{i=1}^N L_i^2.$$

Using Lemma 1, we substitute $L_i \le L_R$ for all i, which implies $\sum_{i=1}^N L_i^2 \le N L_R^2$. Thus:

$$\mathbb{E}\big[\|\widehat{\mathbf{J}}\mathbf{R}\|^2 \mid \boldsymbol{\lambda}\big] \le 16\sqrt{2\pi}\, N^2 L_R^2,$$

which completes the proof. □

B. Smoothness Properties

Next, we establish the Lipschitz continuity of the smoothed response and its Jacobian.

Lemma 3. $\mathbf{R}_\delta(\boldsymbol{\lambda})$ *is Lipschitz continuous with Lipschitz constant* L_R *and bounded by* $R_{\max}$.

Proof. By definition, we have:

$$\begin{aligned} \|\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{R}_\delta(\boldsymbol{\lambda}')\| &= \big\|\mathbb{E}\left[\mathbf{R}(\boldsymbol{\lambda}+\delta\mathbf{u}) - \mathbf{R}(\boldsymbol{\lambda}'+\delta\mathbf{u})\right]\big\| \\ &\le \mathbb{E}\left[\big\|\mathbf{R}(\boldsymbol{\lambda}+\delta\mathbf{u}) - \mathbf{R}(\boldsymbol{\lambda}'+\delta\mathbf{u})\big\|\right] \\ &\le L_R\left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\right\|. \end{aligned}$$

The first inequality is due to $\|\mathbb{E}[X]\| \le \mathbb{E}[\|X\|]$, and the second inequality is due to the Lemma 1. Similarly, we can prove $\|\boldsymbol{R}_\delta(\boldsymbol{\lambda})\| \le R_{\max}$. □

Lemma 4. *The* $\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})$ *is Lipschitz continuous with Lipschitz constant* $\frac{cNL_R}{\delta}$*, where* c *is a constant.*

Proof. Fix any $i \in \{1,\dots,N\}$ and define the scalar smoothing $R_{i,\delta}(\boldsymbol{\lambda}) \coloneqq \mathbb{E}_{\mathbf{u}\sim\mathrm{Unif}(\mathbb{B})}[R_i(\boldsymbol{\lambda}+\delta\mathbf{u})]$. Since $R_i(\cdot)$ is L_i-Lipschitz by Proposition 1, $R_{i,\delta}$ is differentiable and its gradient is Lipschitz with constant $\frac{cL_i\sqrt{N}}{\delta}$, i.e.,

$$\left\|\nabla R_{i,\delta}(\boldsymbol{\lambda}) - \nabla R_{i,\delta}(\boldsymbol{\lambda}')\right\| \le \frac{cL_i\sqrt{N}}{\delta}\left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\right\|. \tag{21}$$

The inequality is due to [17, Proposition 2.2]. Recall that the i-th row of $\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})$ equals $\nabla R_{i,\delta}(\boldsymbol{\lambda})^\top$. Therefore, using the Frobenius norm,

$$\begin{aligned} \left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}')\right\|_F^2 &= \sum_{i=1}^N \left\|\nabla R_{i,\delta}(\boldsymbol{\lambda}) - \nabla R_{i,\delta}(\boldsymbol{\lambda}')\right\|^2 \\ &\le \sum_{i=1}^N \left(\frac{cL_i\sqrt{N}}{\delta}\left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\right\|\right)^2 \\ &= \left(\frac{cNL_R}{\delta}\right)^2 \left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\right\|^2. \end{aligned}$$

Taking square roots gives

$$\left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}')\right\|_F \le \frac{cNL_R}{\delta}\left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\right\|.$$

Finally, since $\|\cdot\| \le \|\cdot\|_F$, we finish the proof. □

C. Properties of the Ancillary Function

To prove the convergence of the Bi-ZOL algorithm, we utilize the ancillary function defined as:

$$\bar{\varphi}_\delta(\boldsymbol{\lambda}) \coloneqq \varphi\left(\boldsymbol{\lambda}, \mathbf{R}_\delta\right) = \boldsymbol{\lambda}^\top \mathbf{R}_\delta(\boldsymbol{\lambda}) \;+\; \rho E_\delta\left(\boldsymbol{\lambda}\right)^2, \tag{22}$$

where $E_\delta \coloneqq \mathbf{1}^\top\big(\mathbf{P}^{\mathrm{b}} - \mathbf{R}_\delta\big) \;-\; P^{\mathrm{target}}$. We now prove $\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda})$ is Lipschitz continuous.

Lemma 5. *The gradient of the ancillary function* $\nabla\bar{\varphi}_\delta(\cdot)$ *is Lipschitz continuous on* Λ *with constant:*

$$\begin{aligned} L_{\nabla\bar{\varphi}_\delta} =& L_R(2 + 2\rho N L_R) \\ &+ \frac{cNL_R}{\delta}\left(\Lambda_{\max} + 2\rho\sqrt{N}(|c_0| + \sqrt{N}R_{\max})\right), \end{aligned} \tag{23}$$

where $\Lambda_{\max} \coloneqq \sup_{\boldsymbol{\lambda}\in\Lambda}\|\boldsymbol{\lambda}\|$.

Proof. We first derive the expression for $\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda})$. Let $c_0 \coloneqq \mathbf{1}^\top\mathbf{P}^{\mathrm{b}} - P^{\mathrm{target}}$. Recall that $E_\delta(\boldsymbol{\lambda}) = c_0 - \mathbf{1}^\top\mathbf{R}_\delta(\boldsymbol{\lambda})$. Thus, $\nabla E_\delta(\boldsymbol{\lambda}) = -\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\mathbf{1}$. Applying the chain rule to $\bar{\varphi}_\delta$:

$$\begin{aligned}\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}) &= \mathbf{R}_\delta(\boldsymbol{\lambda}) + \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\boldsymbol{\lambda} - 2\rho E_\delta(\boldsymbol{\lambda})\,\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\mathbf{1}\\ &= \mathbf{R}_\delta(\boldsymbol{\lambda}) + \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\mathbf{v}(\boldsymbol{\lambda}),\end{aligned}$$

where we define the auxiliary vector $\mathbf{v}(\boldsymbol{\lambda}) \coloneqq \boldsymbol{\lambda} - 2\rho E_\delta(\boldsymbol{\lambda})\mathbf{1}$.

Fix arbitrary $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$ and let $\Delta\boldsymbol{\lambda} \coloneqq \boldsymbol{\lambda} - \boldsymbol{\lambda}'$. Using the triangle inequality:

$$\begin{aligned}\left\|\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}) - \nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}')\right\| \le{}& \left\|\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{R}_\delta(\boldsymbol{\lambda}')\right\|\\ &+ \left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}')^\top\mathbf{v}(\boldsymbol{\lambda}')\right\|.\end{aligned}$$

For the second term, we add and subtract $\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\mathbf{v}(\boldsymbol{\lambda}')$:

$$\begin{aligned}&\left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}')^\top\mathbf{v}(\boldsymbol{\lambda}')\right\|\\ &\le \left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})^\top\big(\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{v}(\boldsymbol{\lambda}')\big)\right\|\\ &\quad + \left\|\big(\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}')\big)^\top\mathbf{v}(\boldsymbol{\lambda}')\right\|\\ &\le \left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})\right\|\cdot\left\|\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{v}(\boldsymbol{\lambda}')\right\|\\ &\quad + \left\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda}')\right\|\cdot\left\|\mathbf{v}(\boldsymbol{\lambda}')\right\|.\end{aligned}$$

We now bound each term. First, for $\|\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{v}(\boldsymbol{\lambda}')\|$, substituting the definition of $\mathbf{v}$:

$$\left\|\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{v}(\boldsymbol{\lambda}')\right\| \le \|\Delta\boldsymbol{\lambda}\| + 2\rho\left|E_\delta(\boldsymbol{\lambda}) - E_\delta(\boldsymbol{\lambda}')\right|\cdot\|\mathbf{1}\|.$$

The mismatch difference is bounded by:

$$\begin{aligned}\left|E_\delta(\boldsymbol{\lambda}) - E_\delta(\boldsymbol{\lambda}')\right| &= \left|\mathbf{1}^\top\big(\mathbf{R}_\delta(\boldsymbol{\lambda}') - \mathbf{R}_\delta(\boldsymbol{\lambda})\big)\right|\\ &\le \|\mathbf{1}\|\cdot\left\|\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{R}_\delta(\boldsymbol{\lambda}')\right\|\\ &\le \sqrt{N}\,L_R\,\|\Delta\boldsymbol{\lambda}\|.\end{aligned}$$

Substituting $\|\mathbf{1}\| = \sqrt{N}$ yields:

$$\left\|\mathbf{v}(\boldsymbol{\lambda}) - \mathbf{v}(\boldsymbol{\lambda}')\right\| \le (1 + 2\rho N L_R)\,\|\Delta\boldsymbol{\lambda}\|. \tag{24}$$

Next, we bound $\|\mathbf{v}(\boldsymbol{\lambda}')\|$. Using Lemma 3:

$$\begin{aligned}|E_\delta(\boldsymbol{\lambda}')| &= |c_0 - \mathbf{1}^\top\mathbf{R}_\delta(\boldsymbol{\lambda}')|\\ &\le |c_0| + \|\mathbf{1}\|\cdot\|\mathbf{R}_\delta(\boldsymbol{\lambda}')\|\\ &\le |c_0| + \sqrt{N}R_{\max}.\end{aligned}$$

Therefore,

$$\left\|\mathbf{v}(\boldsymbol{\lambda}')\right\| \le \Lambda_{\max} + 2\rho\sqrt{N}\big(|c_0| + \sqrt{N}R_{\max}\big). \tag{25}$$

Finally, recall that $\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})\| \le L_R$ (because of Lemma 3, we have $\mathbf{R}_\delta$ is L_R-Lipschitz and differentiable). Using Lemma 4, the Jacobian difference is bounded by $\frac{cNL_R}{\delta}\,\|\Delta\boldsymbol{\lambda}\|$. Combining these into the triangle inequality:

$$\begin{aligned}&\|\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}) - \nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}')\|\\ &\le L_R\|\Delta\boldsymbol{\lambda}\| + L_R(1 + 2\rho N L_R)\|\Delta\boldsymbol{\lambda}\|\\ &\quad + \frac{cNL_R}{\delta}\Big(\Lambda_{\max} + 2\rho\sqrt{N}\big(|c_0| + \sqrt{N}R_{\max}\big)\Big)\|\Delta\boldsymbol{\lambda}\|\\ &= \Bigg[L_R(2 + 2\rho N L_R)\\ &\quad + \frac{cNL_R}{\delta}\Big(\Lambda_{\max} + 2\rho\sqrt{N}\big(|c_0| + \sqrt{N}R_{\max}\big)\Big)\Bigg]\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|.\end{aligned}$$

This completes the proof. □

D. Bias Error Bound

With the smoothness of the ancillary function established, we now quantify the difference between approximate hypergradient g_δ (which uses the observed response $\mathbf{R}$) and the gradient of the ancillary function $\nabla\bar{\varphi}_\delta$ (which uses the smoothed response $\mathbf{R}_\delta$).

Lemma 6. *The difference between the approximate hypergradient and the gradient of the ancillary function is bounded by:*

$$\|\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}) - g_\delta(\boldsymbol{\lambda})\| \le (1 + 2\rho N L_R)\,\delta L_R. \tag{26}$$

Proof. First, we show that the partial derivative $\nabla_2\varphi(\boldsymbol{\lambda}, \cdot)$ is Lipschitz continuous with respect to the response vector $\mathbf{R}$. Recall from (14) that $\nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R}) = \boldsymbol{\lambda} - 2\rho(\mathbf{1}^\top\mathbf{P}^{\mathrm{b}} - \mathbf{1}^\top\mathbf{R} - P^{\mathrm{target}})\mathbf{1}$. For any vectors $\mathbf{R}, \mathbf{R}'$:

$$\begin{aligned}\|\nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R}') - \nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R})\| &= \left\|-2\rho(-\mathbf{1}^\top\mathbf{R}' - (-\mathbf{1}^\top\mathbf{R}))\mathbf{1}\right\|\\ &= 2\rho\left\|(\mathbf{1}^\top\mathbf{R} - \mathbf{1}^\top\mathbf{R}')\mathbf{1}\right\|\\ &\le 2\rho\|\mathbf{1}\|\cdot|\mathbf{1}^\top(\mathbf{R} - \mathbf{R}')|\\ &\le 2\rho\sqrt{N}\cdot\|\mathbf{1}\|\|\mathbf{R} - \mathbf{R}'\|\\ &= 2\rho N\,\|\mathbf{R} - \mathbf{R}'\|.\end{aligned}$$

Thus, the Lipschitz constant is $L_{\varphi 2} \coloneqq 2\rho N$.

Next, we expand the difference between the gradients. Recall $g_\delta = \mathbf{R} + \mathbf{J}\mathbf{R}_\delta^\top\nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R})$ and $\nabla\bar{\varphi}_\delta = \mathbf{R}_\delta + \mathbf{J}\mathbf{R}_\delta^\top\nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R}_\delta)$.

$$\begin{aligned}&\|\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}) - g_\delta(\boldsymbol{\lambda})\|\\ &= \left\|(\mathbf{R}_\delta - \mathbf{R}) + \mathbf{J}\mathbf{R}_\delta^\top\left(\nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R}_\delta) - \nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R})\right)\right\|\\ &\le \|\mathbf{R}_\delta - \mathbf{R}\| + \|\mathbf{J}\mathbf{R}_\delta\|\,\|\nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R}_\delta) - \nabla_2\varphi(\boldsymbol{\lambda}, \mathbf{R})\|\\ &\le \|\mathbf{R}_\delta - \mathbf{R}\| + L_R(2\rho N)\,\|\mathbf{R}_\delta - \mathbf{R}\|\\ &= (1 + 2\rho N L_R)\,\|\mathbf{R}_\delta - \mathbf{R}\|.\end{aligned}$$

The second inequality uses $\|\mathbf{J}\mathbf{R}_\delta(\boldsymbol{\lambda})\| \le L_R$ (from Lemma 3).

Finally, we bound the smoothing error $\|\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{R}(\boldsymbol{\lambda})\|$. Using the definition of randomized smoothing over the unit ball $\mathbb{B}$:

$$\begin{aligned}\|\mathbf{R}_\delta(\boldsymbol{\lambda}) - \mathbf{R}(\boldsymbol{\lambda})\| &= \left\|\mathbb{E}_{\mathbf{u}\sim\mathrm{Unif}(\mathbb{B})}[\mathbf{R}(\boldsymbol{\lambda} + \delta\mathbf{u})] - \mathbf{R}(\boldsymbol{\lambda})\right\|\\ &\le \mathbb{E}_{\mathbf{u}}[\|\mathbf{R}(\boldsymbol{\lambda} + \delta\mathbf{u}) - \mathbf{R}(\boldsymbol{\lambda})\|]\\ &\le \mathbb{E}_{\mathbf{u}}[L_R\|\delta\mathbf{u}\|]\\ &= \delta L_R\mathbb{E}_{\mathbf{u}}[\|\mathbf{u}\|]\\ &\le \delta L_R.\end{aligned}$$

The first inequality is due to the Jensen's inequality. Substituting this back yields the final bound:

$$\|\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}) - g_\delta(\boldsymbol{\lambda})\| \le (1 + 2\rho N L_R)\,\delta L_R. \qquad \square$$

E. Proof of Theorem 1

Proof. Let $\hat{\mathbf{z}}_k \coloneqq \arg\max_{\mathbf{z}\in\Lambda}\langle\mathbf{z}, -g_\delta(\boldsymbol{\lambda}_k)\rangle$ be the optimal direction for the approximate hypergradient, and recall that $\mathbf{z}_k$ is the oracle output for the stochastic estimate $\widehat{\mathbf{g}}_k$.

Using the Lipschitz smoothness of the ancillary function $\bar{\varphi}_\delta$ (Lemma 5) and the update $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \gamma(\mathbf{z}_k - \boldsymbol{\lambda}_k)$:

$$\bar{\varphi}_\delta(\boldsymbol{\lambda}_{k+1}) \le \bar{\varphi}_\delta(\boldsymbol{\lambda}_k) + \gamma\langle\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}_k), \mathbf{z}_k - \boldsymbol{\lambda}_k\rangle$$

$$+\frac{L_{\nabla\bar{\varphi}_\delta}\gamma^2}{2}\|\mathbf{z}_k-\boldsymbol{\lambda}_k\|^2$$
$$\leq \bar{\varphi}_\delta(\boldsymbol{\lambda}_k)+\gamma\langle\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}_k),\mathbf{z}_k-\boldsymbol{\lambda}_k\rangle+\frac{L_{\nabla\bar{\varphi}_\delta}D^2}{2}\gamma^2.$$

We decompose the inner product term to isolate the FW gap for g_δ:

$$\langle\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}_k),\mathbf{z}_k-\boldsymbol{\lambda}_k\rangle$$
$$=\underbrace{\langle g_\delta(\boldsymbol{\lambda}_k),\hat{\mathbf{z}}_k-\boldsymbol{\lambda}_k\rangle}_{\text{True Gap }(-\mathcal{G}_\delta)}+\underbrace{\langle g_\delta(\boldsymbol{\lambda}_k),\mathbf{z}_k-\hat{\mathbf{z}}_k\rangle}_{\text{Optimization Error}}$$
$$+\underbrace{\langle\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}_k)-g_\delta(\boldsymbol{\lambda}_k),\mathbf{z}_k-\boldsymbol{\lambda}_k\rangle}_{\text{Bias Error}}.$$

Bounding the Optimization Error: Recall that $\mathbf{z}_k$ minimizes $\langle\widehat{\mathbf{g}}_k,\cdot\rangle$. Thus, $\langle\widehat{\mathbf{g}}_k,\mathbf{z}_k-\boldsymbol{\lambda}_k\rangle\leq\langle\widehat{\mathbf{g}}_k,\hat{\mathbf{z}}_k-\boldsymbol{\lambda}_k\rangle$. We can write:

$$\begin{aligned}\langle g_\delta(\boldsymbol{\lambda}_k),\mathbf{z}_k-\hat{\mathbf{z}}_k\rangle &= \langle g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}_k,\mathbf{z}_k-\hat{\mathbf{z}}_k\rangle\\ &\quad+\langle\widehat{\mathbf{g}}_k,\mathbf{z}_k-\hat{\mathbf{z}}_k\rangle\\ &\leq \langle g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}_k,\mathbf{z}_k-\hat{\mathbf{z}}_k\rangle+0\\ &\leq \|g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}_k\|\cdot\|\mathbf{z}_k-\hat{\mathbf{z}}_k\|\\ &\leq \|g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}_k\|D.\end{aligned}$$

Bounding the Bias Error: Using Lemma 6:

$$\begin{aligned}\langle\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}_k)-g_\delta(\boldsymbol{\lambda}_k),\mathbf{z}_k-\boldsymbol{\lambda}_k\rangle &\leq \|\nabla\bar{\varphi}_\delta(\boldsymbol{\lambda}_k)-g_\delta(\boldsymbol{\lambda}_k)\|D\\ &\leq (1+2\rho NL_R)L_R\delta D\\ &= C_{\text{bias}}D\delta.\end{aligned}$$

Substituting these back into the descent inequality and rearranging:

$$\gamma\mathcal{G}_\delta(\boldsymbol{\lambda}_k)\leq\bar{\varphi}_\delta(\boldsymbol{\lambda}_k)-\bar{\varphi}_\delta(\boldsymbol{\lambda}_{k+1})$$
$$+\gamma\|g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}_k\|D+\gamma C_{\text{bias}}D\delta+\frac{L_{\nabla\bar{\varphi}_\delta}D^2}{2}\gamma^2.$$

Taking the expectation and summing over $k=0,\ldots,T-1$:

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\mathcal{G}_\delta(\boldsymbol{\lambda}_k)]\leq\frac{\Delta_0}{\gamma T}+\frac{D}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}_k\|]$$
$$+C_{\text{bias}}D\delta+\frac{L_{\nabla\bar{\varphi}_\delta}D^2}{2}\gamma.$$

Bounding the Stochastic Gradient Error: Recall $g_\delta=\nabla_1\varphi+\mathbf{J}\mathbf{R}_\delta^\top\nabla_2\varphi$ and $\widehat{\mathbf{g}}_k=\nabla_1\varphi+\widehat{\mathbf{J}}\mathbf{R}^\top\nabla_2\varphi$.

For the second term in the right-hand side, we have:

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}(\boldsymbol{\lambda}_k)\|\right]$$
$$\overset{(s.1)}{\leq}\sqrt{\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|g_\delta(\boldsymbol{\lambda}_k)-\widehat{\mathbf{g}}(\boldsymbol{\lambda}_k)\right\|^2\right]}$$
$$\overset{(s.2)}{\leq}M_{\varphi 2}\sqrt{\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{J}\mathbf{R}_\delta-\widehat{\mathbf{J}}\mathbf{R}\right\|_F^2\right]}$$
$$\overset{(s.3)}{\leq}M_{\varphi 2}\sqrt{\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|\widehat{\mathbf{J}}\mathbf{R}\right\|_F^2\right]}$$
$$\overset{(s.4)}{\leq}M_{\varphi 2}\sqrt{16\sqrt{2\pi}\,N^2L_R^2}$$
$$=4M_{\varphi 2}\,(2\pi)^{1/4}\,NL_R.$$

where $M_{\varphi 2}$ is the upper bound of $\nabla_2\varphi$. The $(s.1)$ is due to the AM-QM inequality and Jensen's inequality. The $(s.2)$ follows from the upper bound of $\nabla_2\varphi$ and $\|\boldsymbol{A}^\top\boldsymbol{v}\|\leq\|\boldsymbol{A}\|_F\|\boldsymbol{v}\|$. The $(s.3)$ follows from the fact $\mathbb{E}[\|\,X-E[X]\,\|_F^2]\leq\mathbb{E}[\|\,X\,\|_F^2]$ by conditional unbiasedness in Lemma 2. The $(s.4)$ is from the proof in Lemma 2. Substituting this constant bound into the summation yields the final result. □

REFERENCES

[1] L. Monitoring Analytics, "2024 state of the market report for pjm: Volume 2: Detailed analysis," report, 2025.

[2] R. France, "Règles pour la valorisation des effacements de consommation sur les marchés de l'énergie nebef 3.1," report, 2018.

[3] C. P. U. Commission, "Decision regarding phase four direct participation: Authorization for investor-owned utilities to participate in bidding of the CAISO's proxy demand resource product," report, 2010.

[4] H. Yamada and O. Ikki, "National survey report of pv power applications in japan 2017," report, 2018.

[5] K. Daniels and R. Lobel, "Demand response in electricity markets: Voluntary and automated curtailment contracts," *Available at SSRN 2505203*, 2014.

[6] A. Mijatović, J. Moriarty, and J. Vogrinc, "Procuring load curtailment from local customers under uncertainty," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 375, no. 2100, 2017.

[7] R. Aïd, D. Possamaï, and N. Touzi, "Optimal electricity demand response contracting with responsiveness incentives," *Mathematics of Operations Research*, vol. 47, no. 3, pp. 2112–2137, 2022.

[8] S. Xu, L. Yu, X. Lin, J. Dong, and Y. Xue, "Online distributed price-based control of dr resources with competitive guarantees," in *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*, pp. 17–33, 2022.

[9] T. Zhao, M. Zhou, Y. Mo, J. M. Wang, J. Luo, X. Pan, and M. Chen, "Optimizing demand response in distribution network with grid operational constraints," in *Proceedings of the 14th ACM International Conference on Future Energy Systems*, pp. 299–313, 2023.

[10] K. Abedrabboh and L. Al-Fagih, "Applications of mechanism design in market-based demand-side management: A review," *Renewable and Sustainable Energy Reviews*, vol. 171, p. 113016, 2023.

[11] O. Abrishambaf, P. Faria, and Z. Vale, "Application of an optimization-based curtailment service provider in real-time simulation," *Energy Informatics*, vol. 1, no. 1, p. 3, 2018.

[12] A. Lechowicz, J. Comden, and A. Bernstein, "Optimizing individualized incentives from grid measurements under limited knowledge of agent behavior," in *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, pp. 544–563, 2025.

[13] C. Maheshwari, J. Cheng, S. Sastry, L. Ratliff, and E. Mazumdar, "Follower agnostic learning in stackelberg games," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 222–228, IEEE, 2024.

[14] Z. Liu, C. Chen, L. Luo, and B. K. H. Low, "Zeroth-order methods for constrained nonconvex nonsmooth stochastic optimization," in *Proceedings of the 41st International Conference on Machine Learning* (R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, eds.), vol. 235 of *Proceedings of Machine Learning Research*, pp. 30842–30872, PMLR, 21–27 Jul 2024.

[15] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "A new one-point residual-feedback oracle for black-box learning and control," *Automatica*, vol. 136, p. 110006, 2022.

[16] Z. He, S. Bolognani, J. He, F. Dörfler, and X. Guan, "Model-free nonlinear feedback optimization," January 01, 2022 2022. Published on IEEE Transactions on Automatic Control; doi:10.1109/TAC.2023.3341752.

[17] T. Lin, Z. Zheng, and M. Jordan, "Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26160–26175, 2022.

I. Review #102A

A. Paper summary

This submission presents a bi-level optimal incentive-based curtailment methodology for distributed loads. The approach considers a central leader agent (the system operator) at the upper level of the bi-level problem formulation and end-user follower agents, each controlling a set of flexible loads subject to capacity constraints. The approach is demonstrated on an ad-hoc numerical simulation example for 8 end-users managing 4 loads each randomly connected to a 3-node network.

B. Comments for authors

The focus of the paper is inherently methodological, delving into theoretical properties and convergence behavior of the proposed approach, which circumvents computational hurdles arising from non-differentiable load constraints. Network state or power flow constraints are disregarded in the formulation, which narrows the technological value of the contribution to that of a preliminary study on optimal scheduling of flexible loads via price incentives. The more interesting integration of network power flow constraints is left for future work.

C. Summarize your recommendation in one to two sentences

Incorporating network power flow constraints to the formulation would make the paper more meaningful in addressing load management in real-world power systems. Many bi-level formulations of optimal demand management exist in the academic literature. It is not clear what one learns from the paper that can inform real-world power system operation, planning and / or policies.

— *Julio Braslavsky*

II. Review #102B

A. Paper summary

In this work, the Authors consider incentive-based demand response via a bi-level optimization method. The upper level models the system operator whereas the lower one the load behaviour. The load response to an incentive is modelled as balancing a linear reward in the incentive and a quadratic discomfort with unknown scaling. A hybrid gradient-based/zeroth-order method is then proposed to solved the bi-level problem by combining full information gradient for the upper level and a two-point gradient estimate from only the observed lower-level optimal load reduction for the lower level. The Authors show that the iteration/time-averaged expected value of the Frank-Wolf stationary point gap is bounded. A Small-scale simulation example is provided.

B. Comments for authors

- The context of the work is interesting, relevant, and timely. The set-up is interesting with the system operator being white box and the loads blackboxes. The Authors derive a sound approach and motivate each of their steps.
- The deployment of the approach is unclear. At this time, the Authors assume that they can query the system and load as many times as they want to compute the incentive. This rather feels like an online optimization settings (e.g., see online convex optimization) where one submits a decision to an environment (the load in this case), then gets to observe its outcome (optimal reward), which can be used to compute the next round's decision. At this time, one can picture that the local optimization is performed and the feedback is returned. However, in the current setting, it seems like many lower-level optimizations are solved but without implementation, so how is comforting really assess? Switching the online setting seems like it would generalize the approach because it would be such that even the load does not need to know their α, which is the case I believe. At his time, the single-point gradient estimator may be advantageous as well.
- What is the intuition behind a (time/iteration-averaged) expected Frank-Wolfe stationary point gap?
- The approach claims an advantage in using some full information in the hypergradient. How can one support this claim? How is it reflected in the performance analysis?
- How does this approach compared to ZO methods for bilevel optimization? The literature includes ZO first- and second-order methods.
- Combining the two above items, would it not be preferable to use a performance metric akin to the regret? Static regret would capture what is currently proposed and dynamic regret could extend it to a setting where load response evolves through times (e.g., varying degree of flexibility through the day or following sustained curtailment). The main difference is that controlling the regret would mean that the algorithm does not lead the load into high-loss regimes.
- Otherwise, if loads can be queried like in the current setting, it raises the question: why not 'simply' building a supervised learning model for α and then solve the optimization problem to (global) optimality in the same way it is done in the example section?
- The numerical example is tiny. Demand response is intrinsically a large-scale problem. Please discuss scalability.

– Section IV-B: there seem to be confusion with global optimal/local optimal/stationary points.
– It would be interesting to the compare the difference between full ZO and the proposed approach in the numerical simulations.

C. Summarize your recommendation in one to two sentences

This paper would be improved for clarifying the deployment scheme of the method (how is feedback received from the load? Online or not? What is loads change) and largely expanding the numerical case study. Technical ideas behind the hybrid ZO approach is interesting but it would be great to further see the advantage of using the full gradient when available. A major revision would be required for acceptation.

— Anonymous reviewer

III. Review #102C

A. Paper summary

This paper tackled the problem of real-time demand response incentive design under the realistic conditions that the system operator does not know the private information of the customers. It formulated the problem as a bilevel optimization problem and designed a zero-order learning approach with theoretical guarantees.

B. Comments for authors

I thought that this paper was great! It is a strong theoretical paper considering the space limitations and addressed many of the practical considerations when formulating a problem and solution simple enough to analyze. I believe that this is important work that could be applied to many other contexts. I only wonder how the analysis would change if end-user discomfort functions were linear instead of quadratic.

C. Summarize your recommendation in one to two sentences

This paper should be presented and considered for best paper. I believe that this paper provides excellent theoretical contributions to an extremely important problem within the conference's scope.

— Joshua Comden

IV. Brief response to reviewers (optional)

Response to Reviewer #102A:

We thank the reviewer for this comment. We agree that incorporating network power flow constraints better reflects real power systems. In this conference version, we intentionally isolate the incentive-response layer to study learning difficulties stemming from private user parameters and device-induced nonsmoothness. This allows us to clearly study how a system operator can learn aggregate response sensitivities from feedback and use them for price-based load curtailment.

The distinction from existing bilevel demand-management formulations is that our work does not only formulate a leader-follower problem. Instead, we exploit the bilevel structure to decompose the hypergradient and estimate only the unknown response sensitivity through zeroth-order feedback, while keeping the known upper-level objective analytically exact. Crucially, we provide a theoretically grounded algorithm for nonsmooth bilevel optimization, yielding stationarity-certified incentives instead of relying on blind search.

Response to Reviewer #102B:

We thank the reviewer for the careful and constructive comments. Stationarity, comparisons with vanilla zeroth-order (VZO) methods, and Bi-ZOL's scalability require broader treatment than this conference format allows. We studied these issues in our extended manuscript, "Bi-ZOL: Bilevel Zeroth-Order Learning with Nonsmooth Responses," available from the ETH Research Collection at http://hdl.handle.net/20.500.11850/802376. The main distinction is that VZO smooths the whole reduced hyperobjective, while Bi-ZOL keeps the exact upper-level partial derivatives and applies zeroth-order smoothing only to the missing response-sensitivity term. This gives an approximate hypergradient with an anchored chain-rule interpretation. Furthermore, Bi-ZOL's Frank-Wolfe stationarity certificate connects to the original nonsmooth problem's Clarke Frank-Wolfe stationarity via a structural bias term. The manuscript also includes oracle complexity bounds and Bi-ZOL vs. VZO numerical comparisons to clarify empirical performance and scalability.

Regarding the online setting and regret metrics, we agree that this is an important direction. The present paper studies a static feedback-based problem, rather than a full time-varying online optimization problem. A formal regret analysis for time-varying responses belongs to online bilevel games, which is beyond the scope of this conference version and left as future work. Regarding comfort assessment, the perturbed incentives are not physically actuated during the learning phase. Instead, we assume that each household has a HEMS agent that can evaluate the hypothetical discomfort before returning aggregate feedback to the system operator.

Finally, directly learning the discomfort parameter α cannot reconstruct the constrained lower-level problem without also identifying device constraints, which requires richer data and a separate model-learning stage. Moreover, discomfort parameters may vary over time, making fixed models less suitable. Bi-ZOL avoids explicit parameter identification, estimating only the local response sensitivity needed for the current update directly from aggregate feedback.

Response to Reviewer #102C:

We thank the reviewer for this insightful question. With purely linear discomfort, lower-level problems become linear programs, causing set-valued user responses and a non-Lipschitz continuous aggregate response map. Thus, the present convergence proof would not directly apply. A standard remedy is adding a small strongly convex quadratic regularizer to the lower-level discomfort, restoring uniqueness and Lipschitz continuity while approximating the linear-discomfort model.

V. Brief summary of changes made to the final paper

We made minor revisions to the manuscript.

First, we revised the notation for the approximate hypergradient by replacing $\nabla\tilde{\varphi}_\delta(\boldsymbol{\lambda})$ with $g_\delta(\boldsymbol{\lambda})$. This avoids possible misunderstanding, since the Bi-ZOL approximate hypergradient is not necessarily the gradient of an explicitly defined smoothed reduced objective. Instead, it preserves the exact upper-level partial derivatives and smooths only the unknown response-sensitivity term. Accordingly, we renamed the (δ, ϵ)-FWSP as the (δ, ϵ)-Bi-ZOL FWSP.

Second, we revised the related stationarity remarks to clarify the meaning the expected time-averaged Frank–Wolfe gap. The new text explains that the expected time average comes from the randomized zeroth-order estimator.

Third, we added grant support and authors' names and affiliations.