Demand Response in Smart Grids: A Randomized Auction Approach

Ruiting Zhou*, Zongpeng Li*§, Chuan Wu† and Minghua Chen‡
* University of Calgary, {rzho, zongpeng}@ucalgary.ca
† The University of Hong Kong, cwu@cs.hku.hk
‡ The Chinese University of Hong Kong, minghua@ie.cuhk.edu.hk
§ School of Computer and Collaborative Innovation Center of Geospatial Technology, Wuhan University

***Abstract*—The smart grid is a modern power grid that achieves high efficiency and robustness through sophisticated information and communications technology. Demand response has great potential in helping balance demand and supply in a smart grid, cutting generation cost and carbon footprint, and improving system stability. Auctions represent a natural and efficient approach for carrying out demand response between the power grid and large electricity users, microgrids, and electricity storage devices. This work explores the modelling and design space of demand response auctions, targeting expressive power, truthful information revelation, computational efficiency, and economic efficiency. We present a randomized auction that explores the underlying problem structure of demand response, and prove that it is truthful, runs in polynomial time, and achieves $(1+\epsilon)$-optimal social cost for an arbitrarily small constant ϵ. The key technique lies in the marriage of smoothed analysis and randomized reduction, which makes its debut in this work among literature on mechanism design, and can be applied to problems where social welfare optimization is NP-hard but admits a smoothed polynomial-time algorithm.**

I. Introduction

The *smart grid*, emerging as a convergence of ICT with power system engineering, is a modern electric power grid infrastructure for enhanced efficiency and reliability through automated control, communication, sensing and metering, and the strategic optimization of demand, energy, and network availability [1], [2]. As in a traditional power grid, the quintessential problem in a smart grid is the realtime balance between supply and demand. Imbalances are to be corrected within seconds, to avoid frequency deviations that threaten grid stability [3]. Demand response facilitates cost savings by reducing and temporally shifting peak loads, arbitraging between periods of over- and under-generation. Essentially all power grids dispatch generators in a merit order, and wholesale electricity prices are in line with the highest marginal cost. As a result, significant economic benefits can be gleaned from a seemingly small reduction in peak consumption, as illustrated in Fig. 1. It was observed that in a regional power grid within the USA Eastern Interconnection, a 10% peak demand shedding translates into $28 billion annual savings in electricity cost [4].

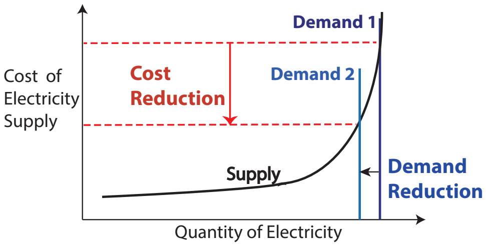

Fig. 1. Benefit of demand response: a small reduction in peak consumption (and hence peak generation rate) leads to significant cost savings and carbon footprint reduction. Electricity generation cost is non-linear, and the marginal cost per Watt increases as the generation rate increases.

In its definition of demand response, the Federal Energy Regulatory Commission (FERC) envisions both *elastic* and *emergent* versions: *"changes in electric usage by end-use customers from their normal consumption patterns to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized."* The first demand response model we study (Sec. III) assumes emergent demand response with a fixed target of demand reduction, while the richer problem of elastic demand response is considered later in Sec. VI.

A *demand response target*, measured as the reduction in net electricity consumption, can be achieved through three types of actions:

◇ *Demand curtailing and temporal shifting.* Large electricity users exemplified by data centres are ideal candidates for participating in demand response. In 2013, U.S. data centers consumed an estimated 91 billion kilowatt-hours of electricity, equivalent to the annual output of 34 large (500-megawatt) coal-fired power plants. They incur $9 billion in electricity bills and emit 97 million metric tons of carbon pollution per year [5]. At the same time, computing tasks are often elastic by nature [6], making temporal workload shifting feasible.

◇ *Onsite generation.* Another salient characteristic of a modern smart grid is the growing penetration of distributed generation [7], where microgrids and large users are in procession of their own generation units that include renewable generation with unstable output (wind, solar) and stand-by generation that can be started and tuned on-demand (diesel, fuel cell). Such quick-start generation can contribute to a demand response process

Manuscript received on March 29, 2015; revised on July 11, 2015. This work is supported by Hidaca Inc, grants from Hong Kong RGC under the contracts HKU 718513 and 17204715, National Basic Research Program of China (Project No. 2013CB336700) and the University Grants Committee of Hong Kong (Theme-based Research Scheme Project No. T23-407/13-N and General Research Fund No. 14201014).

by increasing the output level during peak demand periods.

◇ *Electricity release to the grid.* Recent technology advances are making active (*e.g.*, batteries) and passive (*e.g.*, Plug-in Electric Vehicles, or PEVs) electricity storage economically feasible. Ideally, these devices are charged during periods of low demand and low prices, and discharged at periods of high demand and high prices. A demand response auction provides the necessary catalyst that makes such electricity arbitrage practically possible.

While a vast literature from the past three decades focus on engineering challenges of demand response [8], [9], they often fail to address a fundamental question of practical importance: *why would individual loads, microgrids and storage devices voluntarily respond to grid instability at their own cost?* As admitted by the UK government [10]: *"an effective market mechanism must be created to reward installation of the (demand response) technology fairly."*

The evolution of a traditional power grid to the smart grid makes demand response auctions particularly suitable. First, the two-way information communication infrastructure enables realtime bid submission and auction result declaration [1]. Second, agents in the grid now have their own intelligence module based on software algorithms, and are capable of submitting demand response bids and executing auction algorithms [1]. Large scale realtime auctions in a network environment have now proven practically feasible. Thousands of ad impression auctions are executed on the Internet by Google per second, or billions per day [11]; *myThings*, the personalized retargeting company in Europe, handles over 50 million realtime bids per day [12]; Plethora Mobile receives up to $40,000$ bids per second for audience targeting opportunities [13].

A main alternative to auction is pre-defined electricity price offers, *e.g.*, a substantially high metering price for discouraging consumption when supply is tight [4]. While straightforward, fixed price schemes have their limitations. First, what price to offer is always a tricky question; the grid may end up resorting to heuristic guesses and trial-and-error. Second, it is hard to predict by how much the demand would decrease as a result, problematic for emergent demand response. Third, on-site generation and electricity release need to be considered and priced separately. A well designed demand response auction automatically resolves all three problems. First, when properly designed, an auction discovers the market price of a demand response bid automatically. Second, a demand response target can be explicitly set and achieved in an auction. Third, it is natural to implement a unified type of bid that models all three options of demand response: curtailing electricity consumption, on-site generation, and electricity discharges by storage devices.

This work explores the modelling and design space of demand response auctions in a smart grid, and aims to test the limits of the performance of such auction mechanisms, in terms of expressive power, truthful information revelation, computational efficiency, and economic efficiency. **First**, we consider both emergent demand response where a fixed target in net demand reduction is to be achieved, and elastic demand response where the grid has a concave utility function over a range of flexible reduction ranges. We model demand response bids in both forms of demand reduction and supply augmentation. Quick-start generation at the power grid itself, with linear or non-linear generation cost, is further included. **Second**, we require that the auction be truthful — all demand response participants achieve their respective maximum utility by biding truthfully, regardless of other bidders' strategies. **Third**, for practical feasibility, we assume that computational power at the grid is not unlimited, and the auction algorithm must execute in polynomial-time. **Fourth**, we aim at tuning the auction algorithm to elicit desirable behaviors from agents in the grid, such that a demand response goal is met with minimum possible grid-wide cost.

We present the design of demand response auctions that consider all three types of bids, execute in polynomial time, and achieve near-optimal social welfare. A key technique is the combination of smoothed analysis with randomized MIDR auction design. This is enabled by a pair of associated perturbations that facilitates the design of a smoothed polynomial time algorithm and turns it into a truthful auction. This new technique of designing smoothed polynomial-time auctions is applicable to a broad range of auction design problems, where social welfare optimization can be modelled into a linear integer program that is NP-hard in general, but admits a smoothed polynomial time algorithm.

In the rest of the paper, we review related literature in Sec. II, and define the demand response problem in Sec. III. Sec. IV designs a smoothed polynomial-time demand response algorithm, and Sec. V converts the algorithm into an FPTAS auction. Simulation studies are presented in Sec. VII, and Sec. VIII concludes the paper.

II. Related Work

Over the past decade, demand response has been extensively studied for various management objectives in power grids. Logenthiran *et al.* [2] use a day-ahead load shifting technique to help providers reshape the load profile and reduce peak demand. They formulate demand side management as minimizing the gap between objective consumption and actual consumption, with a heuristic evolutionary algorithm adopted. Qian *et al.* [14] propose a real-time pricing scheme that helps reduce the peak load and realize demand response management in smart grid systems. Shi *et al.* [15] consider residential demand response in a power distribution network with power flow and system operational constraints, and propose a distributed scheme can to compute an optimal demand schedule. Saber *et al.* [16] study two possible models to utilize PEVs: the load-leveling model and the smart grid model, and show that the latter with renewable energy sources is a promising approach. Different from the above literature, this work focuses on the auction design that provides the necessary financial catalyst for realizing demand response. More importantly, these existing demand response mechanisms sidestep the computational challenges by avoiding making win-lose decisions and assume mandatory participation of every agent, which compromises optimal social welfare.

A series of recent work start to examine the design of auction mechanisms for realizing demand response in smart

	Auction	Truthful	Voluntary	Approx. ratio	Type of DR
this work	✓	✓	✓	$1+\epsilon$	DC, OG, ER
[2], [14], [15]	✗	N/A	✗	No guarantee	DC
[16]	✗	N/A	✗	No guarantee	ER
[17]	✓	✓	✗	No guarantee	DC
[18]	✓	✓	✓	Not a constant	ER
[19]	✓	✓	✓	Not a constant	DC
[20]	✗	N/A	✓	close to 3	OG
[7]	✓	No guarantee	✓	Not a constant	OG

TABLE I
COMPARISON BETWEEN EXISTING DEMAND RESPONSE LITERATURE AND THIS WORK. DC = DEMAND CURTAILING AND TEMPORAL SHIFTING; OG = ONSITE QUICK-START GENERATION; ER = ELECTRICITY RELEASE TO THE GRID.

grids. Samadi *et al.* [17] propose a VCG mechanism that aims to maximize the social welfare of a smart grid. Their design requires users to report their energy demand, and computes each user's electricity bill payment. They verify that their mechanism guarantees economic efficiency and user truthfulness. Zhou *et al.* [18] propose an truthful online auction to incentivize the participation of storage devices in power demand response. The approximation ratio of their primal-dual approach is not a constant, but is close to 2 in typical scenarios. Another recent work [19] study datacenter demand response where geo-distributed clouds participate in demand response activities at multiple power grids. A decentralized mechanism is designed for each datacenter to elicit truthful bids and to determine the winning ones. Again, mandatory participation in the demand response is assumed in the first work [17]. Although latest studies [18], [19] model voluntary participation, most of them provide no proven guarantee for approximation ratio in social welfare. Our work is among the first that applies smoothed analysis techniques to design auction mechanisms, and guarantees $(1+\epsilon)$-optimal social cost for an arbitrarily small ϵ.

On the optic of integrating microgrids into a modern smart grid, Lu *et al.* [20] propose an online algorithm for the microgrid generation scheduling problem, which achieves a small competitive ratio below 3. Moreover, a few studies have started to investigate auction design for microgrids. An auction framework for electricity trading between a power grid and microgrids is presented by Zhang *et al.* [7]. Both grid-to-microgrid and microgrid-to-grid energy sales are studied, with truthful bidding guaranteed for the latter case only. The above literature models the voluntary participation of agents, but does not always guarantee truthful bidding, and cannot provide a guarantee of near-optimal social welfare. Table I summarizes the comparison between existing literature and this work.

A polynomial-time approximation scheme (PTAS) [21] is a type of approximation algorithm for NP-hard problems. It takes two parameters: $\epsilon > 0$ and problem size n, and produces a solution that is $(1+\epsilon)$-optimal for minimization problems, or $(1-\epsilon)$-optimal for maximization problems. The running time of a PTAS is required to be polynomial in n, but can be exponential to $\frac{1}{\epsilon}$. If we further require the complexity to be polynomial in both n and $\frac{1}{\epsilon}$, a PTAS become a fully polynomial-time approximation scheme (FPTAS).

Smoothed analysis [22] is a relatively new technique for analyzing the expected running time of an algorithm with a randomly perturbed problem instance. It originates from attempts to understand and analyze the behavior of algorithms that have a bad worst-case performance but a good performance in practice, such as the simplex algorithm for linear programming. To our knowledge, this work is the first that adopts the idea of smoothed analysis to mechanism design. Dough and Roughgarden [23] studied mechanism design where social welfare maximization has a packing structure. They show that if an FPTAS exists when truthful bids are known, then such truthful bids can be elicited through a truthful auction that retains the FPTAS property. While this work has been inspired in part by their randomization techniques, we do not require the existence of an FTPAS in the first place. We resort to the art of smoothed polynomial-time algorithm design instead.

III. SYSTEM MODEL AND PRELIMINARIES

We consider a smart grid system where a power grid is connected with agents that include microgrids, large electricity users (*e.g.*, data centers), and storage devices (*e.g.*, batteries, PEVs), as illustrated in Fig. 2. When the power grid predicts a time period in which supply may fail to meet demand, it acts as the auctioneer and initiates a demand response process by calling for bids from agents through a *reverse auction*, *a.k.a.* a *procurement auction*. Each agent's bid is a pair (e_m, b_m). Here e_m is the power it can supply to the power grid or the amount of power consumption (in W) it is willing to shed. b_m is the corresponding remuneration asked for. The power grid is in possession of its own stand-by generators (*e.g.*, diesel), which can be turned on to supply electricity as well.

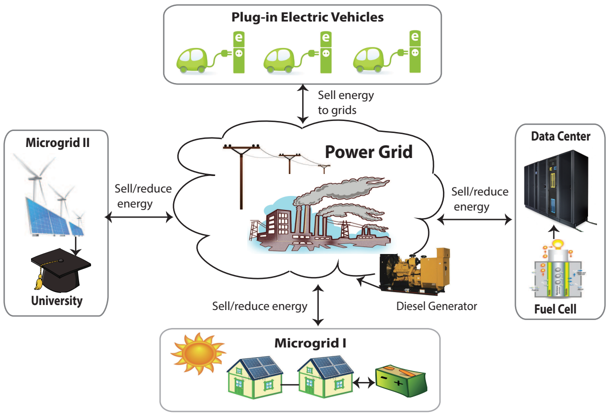

Fig. 2. An illustration of the demand response auction in a smart grid.

Let $[M]$ denote the integer set $\{1, 2, \ldots, M\}$. Assume M agents participate in the auction. D is the demand response target, *i.e.*, power shortage (in W) in the upcoming time period. We define c and z_{max} as the unit cost (in \$ per W) and maximum available power (in W) of the grid's diesel generators. We also assume that any $(M-1)$ agents' bids can cover the demand response target, *i.e.*, $\sum_{m\in[M-n]} e_m \geq D$, $\forall n \in [M]$. At the end of the auction, the auctioneer announces: (i) A binary number x_m corresponding to each agent m, where x_m is 1 if the grid accepts its bid, and 0 otherwise. (ii) A payment p_m to each winning agent m. Finally, the power grid determines the total output rate z of its stand-by generators.

Let v_m be the true cost of e_m, which is private information known to agent m itself only. Let b_{-m} be the set of all bids expect that of agent m. The utility of agent m is:

$$u_m(b_m, b_{-m}) = \begin{cases} p_m - v_m & \text{if } x_m = 1 \\ 0 & \text{otherwise} \end{cases}$$

In term of strategic behaviours, an agents is assumed to be selfish and rational, with a natural goal of maximizing its own utility. An agent may choose to misreport its cost $(b_m \neq v_m)$, if doing so leads to a higher utility. The auctioneer instead aims to maximize the social welfare of the entire grid, for which it is important to elicit truthful bids from agents.

Definition (Truthful auction): An auction is *truthful* if for any agent m, its dominant strategy is to report the true cost v_m of e_m, regardless of other agents' bids. In other words, for all $b_m \neq v_m$ and b_{-m}, the following always holds: $u_m(v_m, b_{-m}) \geq u_m(b_m, b_{-m})$.

Definition (Social welfare, social cost): The social welfare in a demand response auction is the aggregate utility of the grid $(-\sum_{m\in[M]} p_m - cz)$ and the bidding agents $(\sum_{m\in[M]}(p_m - v_m x_m))$. Payments between agents and the grid cancel themselves, and the social welfare is equal to $-\sum_{m\in[M]} v_m x_m - cz$. Maximizing the social welfare is equivalent to minimizing the social cost $\sum_{m\in[M]} v_m x_m + cz$, which in turn is equivalent to minimize $\sum_{m\in[M]} b_m x_m + cz$ under truthful bidding.

Social Cost Optimization in A Demand Response Auction. Under the assumption of truthful bidding, the social cost minimization problem for demand response can be modelled by the following mixed integer linear program (MILP):

$$\text{Minimize} \sum_{m\in[M]} b_m x_m + cz \tag{1}$$

$$\text{Subject to:} \quad \sum_{m\in[M]} e_m x_m + z \geq D \tag{1a}$$

$$x_m \in \{0,1\},\ \forall m \in [M] \tag{1b}$$

$$0 \leq z \leq z_{max} \tag{1c}$$

Constraint (1a) guarantees that the successful bids and the diesel generation are together sufficient to cover the grid's demand response target D. Constraint (1b) models binary decision making. Constraint (1c) limits the output of the diesel generators by their maximum capacity.

MILP (1) is a generalization of the NP-hard problem of minimum knapsack [24], and is hence unlikely to have optimal polynomial-time algorithms. We are interested in polynomial-time demand response auctions that are truthful and can approach the optimal social cost as closely as possible. A table of notations is provided below for ease of reference.

M	# of agents	$[M]$	integer set $\{1,\dots,M\}$
$\vec{1}$	all-one vector	$\vec{0}$	all-zero vector
P	perturbation matrix	$\hat{y}$	$P^T y^p$
α	parameter in (0,1)	ϵ	$\epsilon = \alpha M$
b_m	asking price of e_m	$\tilde{b}_m$	perturbed b_m
v_m	cost of e_m	p_m	payment to agent m
z	output rate of diesel generators		
c	unit cost of diesel generators		
D	demand response target		
e_m	power agent m can supply/reduce		
x_m	agent m's bid successful (1) or not (0)		
D_c	$D_c = \sum_{m\in[M]} e_m - D$		
(x, z_1)	a feasible solution of MILP (1), $x = [x_1\ \dots\ x_M]^T$		
(y, z_2)	a feasible solution of MILP (2), $y = [y_1\ \dots\ y_M]^T$		
$\mathcal{P}(m)$	Pareto optimal set for the first m agents		
β_m	parameter in $[0, \alpha/M]$		
(y^*, z^*)	optimal solution for MILP (2)		
(y^p, z^p)	optimal solution for MILP (5)		
$D(y,z)$	a solution (y,z)'s distribution function		
(x^f, z^f)	final solution for MILP (1)		
γ	$\max_{i,j\in[M]}\{b_i/b_j\}$		
$T(x^f, z^f)$	$b^T x^f + cz^f$, total cost with the solution (x^f, z^f)		

IV. The Smoothed Polynomial-time Algorithm

In this section, we first formulate a complementary problem to MILP (1) in Sec. IV-A, such that a feasible solution to the complementary problem can be easily converted to a feasible solution to MILP (1). Then we design an exact algorithm to solve the complementary problem in Sec. IV-B. Sec. IV-C applies the smoothed analysis technique to perturb bidding prices b_m. With a carefully designed perturbation matrix, the exact algorithm to the complementary problem can be utilized to return a $(1+\epsilon)$-optimal solution to MILP (1), with expected polynomial running time.

A. A Complementary Problem to MILP (1)

Let's define $D_c = \sum_{m\in[M]} e_m - D$. A complementary MILP to MILP (1) is:

$$\text{Maximize} \sum_{m\in[M]} b_m y_m - cz \tag{2}$$

$$\text{Subject to:} \quad \sum_{m\in[M]} e_m y_m - z \leq D_c \tag{2a}$$

$$y_m \in \{0,1\},\ \forall m \in [M] \tag{2b}$$

$$0 \leq z \leq z_{max} \tag{2c}$$

Let (x, z_1) and (y, z_2) be solutions to (1) and (2), respectively, where $x = [x_1\ x_2\ \dots\ x_M]^T$ and $y = [y_1\ y_2\ \dots\ y_M]^T$. Let $y = \vec{1} - x$ and $z_2 = z_1$. Clearly, if (x, z_1) is a feasible solution to MILP (1), then (y, z_2) is a feasible solution to MILP (2).

Consequently, (x, z_1) is an optimal solution to MILP (1) *if and only if* (y, z_2) is an optimal solution to MILP (2), where $\vec{\mathbf{1}}$ is a $M \times 1$ vector of 1's.

B. An Exact Algorithm to the Complementary Problem

We next design an algorithm to solve the complementary problem in MILP (2). The algorithm is *exact* in that it always returns the optimal solution, but it may not terminate in polynomial time. The solution consists of two sets of values: y and z_2. A naive approach is to enumerate all the possible combinations of y, and complement each with a z_2: if $\sum_{m \in [M]} e_m y_m$ is already smaller than D_c, then z_2 is set to zero; otherwise, $z_2 = \sum_{m \in [M]} e_m y_m - D_c$. The optimal solution is the one with the maximum value in $(\sum_{m \in [M]} b_m y_m - c z_2)$. However, this is inefficient since the number of possible y's grows exponentially as the size of the input increases. Let $b = [b_1\ b_2\ \ldots\ b_M]^T$, $e = [e_1\ e_2\ \ldots\ e_M]^T$. Our first idea is to only enumerate the "good" y's and ignore the "bad" ones, based on the following observation: a vector y cannot be optimal if it is dominated by another vector y', *i.e.*, if $b^T y'$ is larger than $b^T y$ and $e^T y'$ is smaller than $e^T y$. We formalize the concept of "good" vectors using *Pareto optimal* vectors:

Definition (Pareto optimal vector): A vector y is *Pareto optimal* if there does not exist a vector y' dominating y, *i.e.*, $\nexists y'$ such that $b^T y' \geq b^T y$ and $e^T y' \leq e^T y$, with at least one inequality being strict.

Lemma 1: Let $\mathcal{P}(m)$ be the set of all Pareto optimal vectors when only the first m agents are considered. If $y^{(m)} \in \mathcal{P}(m)$, then the vector obtained from $y^{(m)}$ by removing its m-th element is a Pareto optimal vector in $\mathcal{P}(m-1)$, $\forall m \in [2, 3, \ldots, M]$.
Proof: Consider a vector $y^{(m)} \in \mathcal{P}(m)$. By the definition of Pareto optimal vectors, $y^{(m)}$ is not dominated by another vector. By way of contradiction, suppose that a vector $y^{(m-1)}$ obtained by removing the last element $y_m^{(m)}$ from $y^{(m)}$ is not Pareto optimal. Then there exists a Pareto optimal vector $y^{(m-1)'}$ dominating $y^{(m-1)}$. In addition, $y^{(m-1)'} + y_m^{(m)}$ (*i.e.*, the vector obtained by appending $y_m^{(m)}$ to the end of $y^{(m-1)'}$) dominates $y^{(m)}$, which leads to a contradiction. □

Lemma 1 suggests that the Pareto optimal set $\mathcal{P}(m)$ can be computed from $\mathcal{P}(m-1)$. Furthermore, it must be contained in the set $\mathcal{P}(m-1) + 0 \cup \mathcal{P}(m-1) + 1$, where $\mathcal{P}(m-1) + 0$ is obtained by appending 0 as the m-th element to each vector in $\mathcal{P}(m-1)$, similar for $\mathcal{P}(m-1) + 1$. An exact algorithm is shown in Algorithm 1, adopting the classic dynamic programming method for constructing the Pareto optimal set. First, it initializes an empty set A and constructs the bottom set $\mathcal{P}(1)$ at line 1. By the definition of a Pareto optimal vector, both solution 1 (accept the first agent's bid) and 0 (reject the the first agent's bid) are included in the set $\mathcal{P}(1)$. Then a *for* loop in lines 2-5 computes $\mathcal{P}(2), \ldots, \mathcal{P}(M)$. At each iteration, $\mathcal{P}(m)$ is derived by eliminating all the dominated vectors (line 4) from the set $\mathcal{P}(m-1) + 0 \cup \mathcal{P}(m-1) + 1$. Another *for* loop in lines 6-12 computes the value of variable z_2. For each y in Pareto optimal set $\mathcal{P}(M)$, if the total $e^T y$ is smaller than or equal to D_c, z is set to zero to maximize the cost. Otherwise, z_2 is set to the gap between $e^T y$ and D_c. If z_2 satisfies constraint (2c), the value of y and z_2 is stored in the set A at line 10. Line 13 returns the solution with maximum objective value among all the feasible solutions in set A.

Algorithm 1 An Exact Algorithm for MILP (2)

```
Input: $b, e, D_c$
Output: optimal solution $y^*$ and $z^*$ to MILP (2)
 1: $A = \emptyset$; $\mathcal{P}(1) = \{0, 1\}$;
 2: for all $m \in [2, 3, \ldots, M]$ do
 3:     Merge $\mathcal{P}(m-1) + 0$ and $\mathcal{P}(m-1) + 1$ into $\mathcal{P}(m)'$ such
        that $\mathcal{P}(m)'$ is sorted in non-decreasing order of social cost;
 4:     Construct $\mathcal{P}(m) = \{y^{(m)} \in \mathcal{P}(m)' | \nexists y^{(m)'} \in \mathcal{P}(m)' :$
        $y^{(m)'}$ dominates $y^{(m)}\}$;
 5: end for
 6: for all $y \in \mathcal{P}(M)$ do
 7:     if $e^T y \leq D_c$ then $z_2 = 0$;
 8:     else $z_2 = e^T y - D_c$;
 9:     end if
10:     if $z_2 \leq z_{max}$ $A = A \cup (y, z_2)$ then;
11:     end if
12: end for
13: Return $y^*, z^* = \arg\max_{(y, z_2) \in A} b^T y - c z_2$
```

Lemma 2: The number of Pareto optimal vectors $|\mathcal{P}(m)|$ does not decrease when m increases, *i.e.*, $\mathcal{P}(1) \leq \cdots \leq \mathcal{P}(M)$.
Proof: In Algorithm 1, $|\mathcal{P}(m)|$ is computed from $\mathcal{P}(m)'$ by pruning the non-Pareto optimal vectors. When we merge $\mathcal{P}(m-1) + 0$ and $\mathcal{P}(m-1) + 1$ into $\mathcal{P}(m)'$, if all solutions in $\mathcal{P}(m-1) + 0$ are retained in $\mathcal{P}(m)$, then $|\mathcal{P}(m)| \geq |\mathcal{P}(m-1)|$. If some solutions in $\mathcal{P}(m-1) + 0$ are eliminated, then there are other solutions in $\mathcal{P}(m-1) + 1$ that dominate them. Therefore, we can always find a vector to replace the removed one. That finishes the proof of $|\mathcal{P}(m)| \geq |\mathcal{P}(m-1)|$. □

Lemma 3: Upon termination, Algorithm 1 returns an optimal solution (y^*, z^*) to MILP (2).
Proof: We first prove that the vector y^* in the optimal solution set (y^*, z^*) for MILP (2) must be a Pareto optimal vector in $\mathcal{P}(M)$. Otherwise, there exists a Pareto optimal vector y' that dominates y^*, *i.e.*, $b^T y' \geq b^T y^*$ and $e^T y' \leq e^T y^*$ with at least one inequality being strict. Therefore, y^* is not an optimal solution to MILP (2), which is a contradiction. Hence, the optimal y^* comes from Pareto optimal set $\mathcal{P}(M)$. In Algorithm 1, we subsequently calculate the corresponding z_2 for each Pareto optimal vector in $\mathcal{P}(M)$ to satisfy constraints (2a) and (2c), and output the one with maximum objective value as the final (optimal) solution. □

Theorem 1: The running time of Algorithm 1 is polynomial to the number of Pareto optimal vectors, and is $O(M|\mathcal{P}(M)|)$.
Proof: Lines 1 in Algorithm 1 can be executed in $O(1)$ steps. The running time of line 13 is polynomial to $|\mathcal{P}(M)|$. During each iteration of the first loop, line 3 constructs set $\mathcal{P}(m)'$ by merging the two sets $\mathcal{P}(m-1) + 0$ and $\mathcal{P}(m-1) + 1$. Both of the these sets are sorted in non-decreasing order of cost due to the assumption of $\mathcal{P}(m-1)$. Thus, we can compute $\mathcal{P}(m)'$ in $O(|\mathcal{P}(m-1)|)$ steps such that it is also sorted. Given this order of vectors in $\mathcal{P}(m)'$, the set $\mathcal{P}(m)$ can be founded in linear time

at line 4. The running time of the second loop (lines 6-12) is polynomial to the number of vectors in the Pareto optimal set $\mathcal{P}(M)$. In summary, the overall time complexity of Algorithm 1 is upper bounded by $\sum_{m=1}^{M-1} O(|\mathcal{P}(m)|) + O(|\mathcal{P}(M)|)$. By Lemma 2, $O(\sum_{m=1}^{M-1} |\mathcal{P}(m)|) + O(|\mathcal{P}(M)|) \leq O(M|\mathcal{P}(M)|) + O(|\mathcal{P}(M)|) = O(M|\mathcal{P}(M)|)$. □

C. The Smoothed Polynomial-Time Algorithm

We next design a randomized algorithm with expected polynomial running time for MILP (1), based on randomized perturbation. The basic idea is to first construct a set of feasible solutions to the complementary problem (2) through the exact Algorithm 1, then randomly output a solution from this set following a well-designed distribution, such that the gap between the expectation of the chosen solution and the optimal solution is very small. Next, we utilize the complementarity between MILP (2) and MILP (1) to compute a corresponding feasible solution to MILP (1), and prove that the expectation of our solution is away from optimum by at most an additive factor $\epsilon \max_{m \in [M]} b_m$, where ϵ is a small parameter that can be arbitrarily close to zero. In order to compute the feasible set efficiently, we use a perturbation matrix to perturb the input bids based on smoothed analysis techniques.

Given an approximation parameter $\alpha \in (0, 1)$, we draw M random variables uniformly from $[0, \alpha/M]$, forming a vector $\beta = [\beta_1 \ \beta_2 \ \ldots \ \beta_M]^T$. Define a perturbation matrix:

$$P = (1 - \alpha)I + \frac{\beta \vec{\mathbf{1}}^T}{M}, \tag{3}$$

where I is a $M \times M$ identity matrix. We utilize the perturbation matrix to perturb the cost vector b into a new vector $\tilde{b} = Pb$, such that each new cost $\tilde{b_m}$ can be expressed as:

$$\tilde{b}_m = (1 - \alpha)b_m + \frac{\beta_m \sum_{j=1}^{M} b_j}{M}, \forall m \in [M]. \tag{4}$$

The perturbed complementary problem is:

$$\text{Maximize} \sum_{m \in [M]} \tilde{b}_m y_m - cz \tag{5}$$

$$\text{Subject To: Constraints } (2a)(2b)(2c).$$

Algorithm 1 can be executed to solve the above perturbed problem (5), and it outputs an optimal solution (y^p, z^p). The value of the objective function is $POBJ = \tilde{b}^T y^p - cz^p$.

Let (y^*, z^*) and $OBJ_2 = b^T y^* - cz^*$ be the optimal solution to the complementary problem in (2) and the value of the corresponding objective function, respectively. We then have:

$$\begin{aligned} POBJ &= \tilde{b}^T y^p - cz^p = (Pb)^T y^p - cz^p \geq (Pb)^T y^* - cz^* \\ &= b^T((1 - \alpha)I + \frac{\vec{\mathbf{1}}\beta^T}{M})y^* - cz^* \geq (1 - \alpha)b^T y^* - cz^* \\ &= OBJ_2 - \alpha b^T y^* \end{aligned} \tag{6}$$

The first inequality holds because (y^p, z^p) is the optimal solution to the perturbed problem (5). We can observe that a possible solution to problem (2) is $(\hat{y} = P^T y^p, z^p)$, which has only a small loss $\alpha b^T y^*$. However, $(\hat{y}, z^p)$ may not be a feasible solution because $P^T y^p$ may have fractional entries due to the setting of P. Hence, constraint (2b) is violated, and constant (2a) may not be satisfied either. Although we can not use $(\hat{y}, z^p)$ directly as the solution for problem (2), we can design a randomized algorithm that outputs a sample following a well-designed distribution, such that *the expectation of the random sample equals* $(\hat{y}, z^p)$. As a result, our approach can solve problem (2) with a small loss $\alpha b^T y^*$ in expectation.

Let y be a M-dimensional all-zero vector $\vec{\mathbf{0}}$, then $(\vec{\mathbf{0}}, z^p)$ is a feasible solution to (5). If we assume $e_m \leq D_c, \forall m \in [M]$, let $l_1, l_2, \ldots, l_M$ denote basis vectors, *i.e.*, $l_m^m = 1$ and $l_m^{m'} = 0$, $\forall m' \neq m$. Then for any m, (l_m, z^p) lies in the feasible set of (5). Note that (y^p, z^p) is also a feasible solution to (2) because the constrains in MILP (5) are identical to those in MILP (2). The final output for MILP (2) is (y^f, z^f), where y^f is a sample randomly produced from the set $\{y^p, l_1, l_2, \ldots, l_M, \vec{\mathbf{0}}\}$ and $z^f = z^p$. The final output follows the distribution $D(y^f, z^f)$:

$$D(y^f, z^f) = \begin{cases} Pr[y^f = y^p, z^f = z^p] = 1 - \alpha \\ Pr[y^f = l_m, z^f = z^p] = \dfrac{\sum_{j=1}^{M} \beta_j y_j^p}{M}, \forall m \in [M] \\ Pr[y^f = \vec{\mathbf{0}}, z^f = z^p] \\ = 1 - Pr[y^f = y^p] - \displaystyle\sum_{m=1}^{M} Pr[y^f = l_m]. \end{cases} \tag{7}$$

We can verify that the expectation of y^f is equal to $\hat{y}$:

$$E[y^f] = (1 - \alpha)y^p + \frac{\sum_{j=1}^{M} \beta_j y_j^p}{M}(\sum_{m=1}^{M} l_m) = P^T y^p = \hat{y} \tag{8}$$

Thus, the expected value of the objective function when (y^f, z^f) follows the distribution $D(y^f, z^f)$ is

$$E[b^T y^f - cz^f] = b^T \hat{y} - cz^p \geq OBJ_2 - \alpha b^T y^* \tag{9}$$

We know that MILP (2) is a complementary problem to MILP (1). A solution (x^f, z^f) to MILP (1) can be obtained by letting $x^f = \vec{\mathbf{1}} - y^f$ and $z^f = z^p$. According to the assumption $\sum_{m \in [M-n]} e_m \geq D, \forall n \in [M]$ (Sec. III), (x^f, z^f) must be a feasible solution to MILP (1). Algorithm 2 is our randomized algorithm that utilizes this property to solve the original MILP (1). We next analyze the approximation guarantee of Algorithm 2.

Theorem 2: The expected social cost of the solution (x^f, z^f) returned by Algorithm 2 is at most an additive $\epsilon \max_{m \in [M]} b_m$ more than the optimal social cost, where $\epsilon = \alpha M$.
Proof: Define OBJ_1 as the optimal social cost of problem (1). The expected objective value returned by Algorithm 2 is

$$\begin{aligned} &E[b^T x^f + cz^f] = E[b^T(\vec{\mathbf{1}} - y^f) + cz^p] \\ &= \sum_{m \in [M]} b_m - E[b^T y^f - cz^p] \leq \sum_{m \in [M]} b_m - (1 - \alpha)b^T y^* + cz^* \\ &= (\sum_{m \in [M]} b_m - b^T y^* + cz^*) + \alpha b^T y^* \\ &\leq OBJ_1 + \epsilon/M \sum_{m \in [M]} b_m \leq OBJ_1 + \epsilon \max_{m \in [M]} b_m. \end{aligned} \tag{10}$$

□

We next show in Lemma 4 and Theorem 3 that the expected running time of Algorithm 2 is polynomial to the input size. Intuitively, the time complexity of Algorithm 2 depends on the number of Pareto optimal vectors, for which we establish an upper-bound that is a polynomial of M and $\frac{1}{\epsilon}$.

Lemma 4: For the perturbed complementary maximization problem (5) with perturbation matrix P produced from equation (3), the expected number of Pareto optimal vectors $E[|\mathcal{P}(M)|]$ is upper bounded by $1 + \frac{M^4}{\alpha}$.

Proof: Let $c(y)$ be the social cost under vector y and perturbed cost $\tilde{b}$. Let $e(y) = e^T y$. Each Pareto optimal vector has a total cost in $[0, Mb_{max}]$ because each agent's perturbed cost is at most b_{max}. Assuming that no two vectors are identical, we can partition $[0, Mb_{max}]$ into small intervals such that there is at most one Pareto optimal vector in each small interval. As a result, the expected number of Pareto optimal vectors is:

$$E[|\mathcal{P}(M)|] = 1 + \lim_{N \to \infty} \sum_{n=0}^{N-1} Pr[\exists y \in [\mathcal{P}(M)] : \\ c(y) \in (\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]].$$

Where the additional 1 corresponds to the vector $\vec{\mathbf{0}}$, which is Pareto optimal by definition. To estimate the probability in each interval, we first define some variables. Let y^{n*} be the vector that has the largest $e(y)$ and satisfies $c(y) \leq \frac{Mb_{max}n}{N}$, *i.e.*, $y^{n*} = \arg\max\{e(y)|c(y) \leq \frac{Mb_{max}n}{N}\}$. For $n \geq 0$, y^{n*} must always exist. Let $y^{\hat{n}} = \arg\min\{c(y)|e(y) > e(y^{n*}) \cap c(y) > \frac{Mb_{max}n}{N}\}$ be the vector that has the smallest cost such that $e(y) > e(y^{n*})$ and $c(y) > \frac{Mb_{max}n}{N}$.

If $c(y^{\hat{n}})$ exists, then we define a random variable $\Lambda(\frac{Mb_{max}n}{N}) = c(y^{\hat{n}}) - \frac{Mb_{max}n}{N}$, and claim that,

Claim 1: If and only if $\Lambda(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}$, there exists a Pareto optimal vector y such that $c(y) \in (\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]$.

Proof: Assume there is a Pareto optimal vector with the cost in $(\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]$, and let y^n be the Pareto optimal vector with the smallest cost in $(\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]$. Then according the definition, $y^n = y^{\hat{n}}$ and $\Lambda(\frac{Mb_{max}n}{N}) = c(y^n) - \frac{Mb_{max}n}{N} \in (0, \frac{Mb_{max}}{N}]$. Conversely, if $\Lambda(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}$, $y^{\hat{n}}$ must be a Pareto optimal vector whose cost lies in the range of $(\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]$. □

Hence, we can rewrite the expected number of Pareto optimal vectors as:

$E[|\mathcal{P}(M)|] = 1 + \lim_{N \to \infty} \sum_{n=0}^{N-1} Pr[\Lambda(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}]$.

Furthermore, we define $y^{(n,m-)}$ as the vector that rejects agent m's bid and has a cost of at most $\frac{Mb_{max}n}{N}$. Let $S^{(n,m-)} = \{y|c(y) \leq \frac{Mb_{max}n}{N} \cap y_m = 0\}$ be the set of all $y^{(n,m-)}$. Further define $y^{(n*,m+)}$ as: $y^{(n*,m-)} = \arg\max\{e(y)|c(y) \leq \frac{Mb_{max}n}{N} \cap y_m = 0\}$. We define another variable $y^{(\hat{n},m+)}$ as $y^{(\hat{n},m+)} = \arg\min\{c(y)|e(y) > e(y^{n*,m-}) \cap c(y) > \frac{Mb_{max}n}{N} \cap y_m = 1\}$. Similarly, we define a random variable $\Lambda^m(\frac{Mb_{max}n}{N}) = c(y^{\hat{n},m+}) - \frac{Mb_{max}n}{N}$ when $y^{\hat{n},m+}$ exists, then we have the following claim:

Claim 2: When $\Lambda(\frac{Mb_{max}n}{N})$ is defined, there exists an index $m \in [M]$ such that $\Lambda(\frac{Mb_{max}n}{N}) = \Lambda^m(\frac{Mb_{max}n}{N})$.

Proof: When $\Lambda(\frac{Mb_{max}n}{N})$ is defined, there exist the corresponding $y^{\hat{n}}$ and y^{n*}. Because $c(y^{n*}) < c(y^{\hat{n}})$, there must exist at least one agent's bid, indexed by $m \in [M]$, being accepted by $y^{\hat{n}}$ but rejected by y^{n*} *i.e.*, $y_m^{\hat{n}} = 1, y_m^{n*} = 0$. We claim that for this index m, $\Lambda(\frac{Mb_{max}n}{N}) = \Lambda^m(\frac{Mb_{max}n}{N})$. In order to prove it, we first observe that $y^{n*} = y^{n*,m-}$ This is due to the reason that y^{n*} is the vector with the highest $e(y)$ among all vectors with cost at most $\frac{Mb_{max}n}{N}$. It is belong to the set $S^{n,m-}$, it is in particular the vector with the highest $e(y)$ among all vectors that reject agent m's bid and have the cost at most $\frac{Mb_{max}n}{N}$. Similarly arguments can be applied to prove $y^{\hat{n}} = y^{\hat{n},m+}$. This directly implies that $\Lambda(\frac{Mb_{max}n}{N}) = \Lambda^m(\frac{Mb_{max}n}{N})$. □

Claim 3: : $\forall m \in [M]$, $Pr[\Lambda^m(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}] \leq \frac{M^3}{\alpha N}$.

Proof: In the perturbed MILP (5), the value of $\tilde{b}_m$ lies in $[(1-\alpha)b_m, (1-\alpha)b_m + \frac{\alpha \sum_{j=1}^{M} b_j}{M^2}]$ since β_m is drawn from $[0, \alpha/M]$. The length of the interval is $\frac{\alpha \sum_{j=1}^{M} b_j}{M^2}$, which is no smaller than $\frac{\alpha b_{max}}{M^2}$ as $\sum_{j=1}^{M} b_j \geq b_{max}$. The density of $\tilde{b}_m$ is upper-bounded by $\frac{M^2}{\alpha b_{max}}$ everywhere in the interval $[(1-\alpha)b_m, (1-\alpha)b_m + \frac{\alpha \sum_{j=1}^{M} b_j}{M^2}]$.

In order to prove this lemma, we exploit the randomness of cost $\tilde{b}_m$ for a given m, other agents' cost $\tilde{b}_j$, $j \neq m$, can be considered as arbitrarily fixed parameters. Then the vectors from set $S^{n,m-}$ are fixed and hence also the vector $y^{n*,m-}$ is fixed. Let $S = \{y|e(y) > e(y^{n*,m-}) \cap c(y) > \frac{Mb_{max}n}{N} \cap y_m = 1\}$. If the vector $y^{n*,m-}$ is fixed, then set S is also fixed. $y^{\hat{n},m+}$ is the vector with the minimal cost in set S. To prove Claim 3, we only need to find the probability of $c(y^{\hat{n},m+}) \in (\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]$. Because other agents' costs can be considered as fixed parameters, the value of $\tilde{b}_m$ determines whether $c(y^{\hat{n},m+})$ lies in the interval $(\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]$ or not. We can rewrite this event as $\{\frac{Mb_{max}n}{N} < \sum_{j \neq m} \tilde{b}_j + \tilde{b}_m \leq \frac{Mb_{max}(n+1)}{N}]\}$. Let $\lambda = \frac{Mb_{max}n}{N} - \sum_{j \neq m} \tilde{b}_j$, then the above event is the same as the event $\tilde{b}_m \in (\lambda, \lambda + \frac{Mb_{max}}{N}]$. Hence, the probability of this event is upper-bounded by $\frac{Mb_{max}}{N} \times \frac{M^2}{\alpha b_{max}} = \frac{M^3}{\alpha N}$. □

Combining Claim 1, Claim 2 and Claim 3, we have:

$$\begin{aligned} & Pr[\exists y \in [\mathcal{P}(M)] : c(y) \in (\frac{Mb_{max}n}{N}, \frac{Mb_{max}(n+1)}{N}]] \\ & = Pr[\Lambda(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}] \\ & \leq Pr[\exists m \in [M] : \Lambda^m(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}] \\ & = \bigcup_{m=1}^{M} Pr[\Lambda^m(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}] \\ & \leq \sum_{m=1}^{M} Pr[\Lambda^m(\frac{Mb_{max}n}{N}) \leq \frac{Mb_{max}}{N}] \leq \frac{M^4}{\alpha N}. \end{aligned}$$

Therefore, we have: $E[|\mathcal{P}(M)|] \leq 1 + \lim_{N \to \infty} \sum_{n=0}^{N-1} \frac{M^4}{\alpha N} = 1 + \frac{M^4}{\alpha}$. □

Theorem 3: The expected running time of Algorithm 2 for solving MILP (1) is polynomial.

Proof: In Algorithm 2, lines 1-2 generate the perturbation matrix P, and their running time is polynomial to M. Lines 4-5 output a random solution y^f according to the distribution $D(y^p)$, and the running time is polynomial as well. Line 6 takes one step to compute x^f and returns the result. Now, we only need to examine the running time of line 3. Combining Theorem 1 and Lemma 4, the expected running to solve the perturbed MILP (5) is

$$O(ME[|\mathcal{P}(M)|]) \leq O(M(1 + \frac{M^4}{\alpha})) \leq O(M^5 \frac{1}{\alpha}) = O(M^6 \frac{1}{\epsilon}).$$

Hence, the overall expected running time of Algorithm 2 is $O(M^6 \frac{1}{\epsilon})$. □

Note that Algorithm 2 is effectively a randomized *additive FPTAS*, since it solves MILP (1) with expected polynomial-time, and outputs a solution that is at most an additive $\epsilon \max_{m \in [M]} b_m$ more than the optimal social cost.

Further define $\gamma = \max_{i,j\in[M]}\{b_i/b_j\}$. We have that $\max_{m\in[M]} b_m/OBJ_1 \leq \gamma$, as OBJ_1 includes at least one agent's bid. Thus, the expectation of social cost achieved by Algorithm 2 is: $E[b^T x^f + cz^f] \leq (1+\gamma\epsilon)OBJ_1$. In other words, Algorithm 2 can achieve $(1+\gamma\epsilon)$-optimality where ϵ can be arbitrarily close to zero. With γ being a constant, Algorithm 2 becomes a randomized FPTAS that can output a $(1+\epsilon)$-optimal solution.

Algorithm 2 A Smoothed Polynomial-Time Algorithm for MILP (1)

```
Input: $\alpha \in (0,1), b, e, D_c$
Output: A solution $(x^f, z^f)$ for MILP (1)
Generate $\beta = [\beta_1\ \beta_2\ \ldots\ \beta_M]^T$ uniformly randomly from the interval $[0, \alpha/M]$ ;
Compute the perturbation matrix: $P = (1-\alpha)I + \frac{\beta\vec{1}^T}{M}$;
Run Algorithm 1 with the input $(Pb, e, D_c)$, and obtain output $(y^p, z^p)$;
Produce a distribution function $D(y^f, z^f)$ as shown in (7);
Choose $(y^f, z^f)$ according to the distribution function $D(y^f, z^f)$;
$x^f = 1 - y^f$; Return $(x^f, z^f)$.
```

V. An FPTAS Demand-Response Auction

We now translate the smoothed polynomial-time algorithm (Algorithm 2) into a truthful auction, adding truthfulness to the algorithm while retaining its FPTAS property. This is achieved through a recent technique of maximal-in-distributional range (MIDR) algorithms [25], the second main technique in this work besides smoothed analysis. An MIDR algorithm refers to a randomized algorithm that outputs a sample from a set of feasible solutions, according to a distribution that does not depend on agent bids, achieving the largest social welfare among all such distributions in the range. Combined with VCG-style payments following a similar distribution, an MIDR algorithm yields an auction that is truthful in expectation. We will still utilize the complementarity between MILPs (1) and (2). First, we generate a distribution range for MILP (2), then covert it to one for MILP (1). Then we prove that our randomized Algorithm 2 is an MIDR algorithm. At the end, a randomized VCG-like payment scheme that works in concert with the MIDR algorithm is designed to obtain a truthful demand response auction.

Theorem 4: Algorithm 2 is an MIDR algorithm for the original cost minimization problem in MILP (1).

Proof: An MIDR algorithm pre-commits to a distribution range (a set of probability distributions over feasible solutions) independent of agents' bids, and returns a sample based on a distribution that is from the distribution range to maximize the expected social welfare. In a procurement auction, social welfare maximization is equivalent to social cost minimization. Therefore, an MIDR algorithm for MILP (1) is to output a sample form a distribution that minimizes the expected social cost over the distribution range.

First, we construct a distribution range for MILP (2), and convert it to a distribution range for the original problem in MILP (1). Let S denote the set of all feasible solutions (y', z') of MILP (5). For each feasible solution $(y', z') \in S$, we can construct a distribution for (y, z_2) similar to the one in (7):

$$D(y, z_2) = \begin{cases} Pr[y = y', z_2 = z'] = 1 - \alpha \\ Pr[y = l_m, z_2 = z'] = \dfrac{\sum_{j=1}^{M} \beta_j y'_j}{M}, \forall m \in [M] \\ Pr[y = \vec{\mathbf{0}}, z_2 = z'] \\ = 1 - Pr[y = y'] - \displaystyle\sum_{m=1}^{M} Pr[y = l_m]. \end{cases} \tag{11}$$

Where (y, z_2) is a feasible solution for MILP (5) sampled following distribution $D(y, z_2)$. Constraints in MILP (5) is the same as those in MILP (2), hence (y, z_2) is also a feasible solution to MILP (2) and (y, z_2) is independent of agents' bids. Therefore, we can construct a compact set $R_y = \{D(y, z_2), \forall (y', z') \in S\}$ including all the distributions indexed by feasible solution (y', z'), and R_y is the distribution range for MILP (2).

Let $D(x^f, z^f)$ be the distribution that draws (y^f, z^f) according to distribution $D(y^f, z^f)$ and outputs $(x^f = \vec{\mathbf{1}} - y^f, z^f = z^f)$. Similarly, let $D(x, z_1)$ be the distribution that draws (y, z_2) according to distribution $D(y, z_2)$ and computes a solution $(x = \vec{\mathbf{1}} - y, z_1 = z_2)$. Then, the distribution range R_x for MILP (1) includes all the possible $D(x, z_1)$. We have

$$\begin{aligned} & E_{(x^f, z^f)\sim D(x^f, z^f)}[b^T x^f + cz^f] \\ & = E_{(y^f, z^f)\sim D(y^f, z^f)}[b^T(\vec{\mathbf{1}} - y^f) + cz^f] \\ & = \sum_{m\in[M]} b_m - b^T(P^T y^p) + cz^p \\ & = \sum_{m\in[M]} b_m - \max_{(y', z')\in S}(b^T(P^T y') - cz') \\ & = \sum_{m\in[M]} b_m - \max_{(y', z')\in S} E_{(y, z_2)\sim D(y, z_2)}[b^T y - cz_2] \\ & = \min_{(y', z')\in S} E_{(y, z_2)\sim D(y, z_2)}[b^T(\vec{\mathbf{1}} - y) + cz_2] \\ & = \min_{D(x, z_1)\in R_x} E_{(x, z_1)\sim D(x, z_1)}[b^T x + cz_1] \end{aligned} \tag{12}$$

The first two equalities above follow from the definition of $D(x^f, z^f)$ and y^f's expected value in equation (8). The third equality holds because (y^p, z^p) is the optimal solution to the perturbed MILP (5). The fourth equality holds since $E_{(y, z_2)\sim D(y, z_2)}[y] = P^T y'$, which can be derived according to (8). The last two equalities come from the definition of $D(x, z_1)$ and R_x. In summary, Algorithm 2 is an MIDR algorithm that achieves the smallest expected social cost among all the solutions produced following the distributions in the distribution range R_x. □

Next, towards designing a truthful-in-expectation auction, we first describe an important property of an MIDR algorithm: analogous to the VCG mechanism, there is a deterministic payment rule p_m^{vcg} that can be coupled with an MIDR algorithm to yield a truthful-in-expectation mechanism, and

$$p_m^{vcg} = E[T(x_{-m}^f, z_{-m}^f) - (T(x^f, z^f) - b_m x_m^f)], \forall m \in [M]. \tag{13}$$

Here p_m^{vcg} is the payment for each agent m. We do not need to consider the payment to the power grid's own quick-start generators. (x_{-m}^f, z_{-m}^f) is the output of Algorithm 2 by setting agent m's asking price to infinity. $T(x_{-m}^f, z_{-m}^f)$ is the total social cost with agent m excluded from the auction. $T(x^f, z^f)$ is the total social cost when agent m participates and (x^f, z^f) is the solution returned by Algorithm 2. We define x^f_m as the m-th element of x^f, then $T(x^f, z^f) - b_m x^f_m$ is the overall cost except agent m, when every agent participates in the auction.

It is not always possible to compute the expected value in (13) efficiently. Nonetheless, if the expectation of a randomized payment scheme is equal to p^{vcg}_m, then this payment also guarantees truthfulness in expectation [23]. Therefore, we compute the payments as follows:

$$p_m = T(x^f_{-m}, z^f_{-m}) - (T(x^f, z^f) - b_m x^f_m), \forall m \in [M] \quad (14)$$

Lemma 5: The payment scheme in (14) yields a truthful auction in expectation.

Proof: Intuitively, the randomized MIDR auction is truthful because both winner determination and payment computation are bid independent, and it is known that an auction is truthful if and only if it is bid independent. More specifically, It is easy to observe that $E[p_m] = E[T(x^f_{-m}, z^f_{-m}) - (T(x^f, z^f) - b_m x^f_m)] = p^{vcg}_m$. According to the properties of MIDR algorithms [25], [23], such VCG-type payment renders an MIDR algorithm truthful in expectation. □

Theorem 5: The randomized algorithm in Alg.3 combined with the randomized VCG payment (14) is a truthful-in-expectation mechanism, parametrized by ϵ, that runs in polynomial time in expectation, and outputs a solution with expected social cost at most an additive $\epsilon \max_{m \in [M]} b_m$ more than the optimal value.

Proof: The theorem follows from Theorem 2, Theorem 3 and Lemma 5. □

Algorithm 3 A Randomized Auction Mechanism

```
Input: $\alpha \in (0, 1), b, e, D_c$
Output: A solution $(x^f, z^f)$ to MILP (1) and payment $p_m$
Run Algorithm 2 with the input $(\alpha, b, e, D_c)$, the output is $(x^f, z^f)$;
Compute the payment for each winning agent, $p_m = T(x^f_{-m}, z^f_{-m}) - (T(x^f, z^f) - b_m x^f_m), \forall m \in [M]$ ;
Return a solution $(x^f, z^f)$; Return the payment $p_m$ for each winning agent $m$.
```

VI. Extensions and Discussions

A. Non-linear Cost in Electricity Generation

In the real world, the unit cost for the power grid to generate its own power is not linear to the quantity of generation z. The cost of stand-by generation is a function of the generation rate, $\Delta(z)$. It is determined by all the costs over the operation time period t, and varies for different values of z. The objective function of MILP (1) can be rewritten as:

$$\text{Minimize} \sum_{m \in [M]} b_m x_m + \Delta(z)$$

For example, the total cost $\Delta(z)$ of a diesel generator to generate zW power consists of the following components [26]: $\Delta(z) = I_t + M_t(z) + F_t(z) + W_t(z)$. During the generator's running time period t, I_t is investment cost, $M_t(z)$ is operation and maintenance cost, $F_t(z)$ is fuel cost, and $W_t(z)$ is waste disposal and emission control cost. I_t is a constant value, which is the one-time purchase cost of the generator amortized to the running period t. $M_t(z)$, $F_t(z)$ and $W_t(z)$ are functions of the generation rate z.

The same strategy (Algorithm 3) can be adapted to such a non-linear cost model for designing an FPTAS demand-response mechanism. The difference lies in the last step in Algorithm 1. It returns a solution $(y, z_2) \in A$ with maximum value of $b^T y - \Delta(z_2)$.

B. Elastic Demand Response

While the power grid may have a most preferred value for the demand response target, such a target is often not the only option, and a small variation is tolerable though less preferred. In a real-world power grid, such deviation, up to a certain threshold value, can be absorbed by the generating plants through *ex post* primary frequency control [3]. Realtime imbalances in grid-wide electricity demand and supply are reflected in measured deviations in power frequency from its nominal value (50 Hz in the majority of the world, 60 Hz in North America and small parts of Asia). Generating units are then regulated on a second-by-second basis to correct such deviation in a closed-loop control fashion. Such primary frequency control comes at its own cost [9], and we capture that through a utility function of the power grid, which associates each feasible demand response target value with a difference level of preference:

$$\text{Minimize} \sum_{m \in [M]} b_m x_m + \Delta(Z) - U(D) \quad (15)$$

$$\text{Subject To:} \quad \sum_{m \in [M]} e_m x_m + z \geq D \quad (15a)$$

$$x_m \in \{0, 1\}, \ \forall m \in [M] \quad (15b)$$

$$0 \leq z \leq z_{max} \quad (15c)$$

$$D_{min} \leq D \leq D_{max} \quad (15d)$$

$U(D)$ is the grid's utility function that depends on the target reduction D in power consumption. $U(D)$ is often a concave function in practice [17]. D_{min} and D_{max} are the lower-bound and upper-bound of the demand response target of the power grid, respectively, which bound the net reduction within an acceptable range. The power grid can accept an aggregated net reduction from agents that is smaller or larger than the actual demand response target D', but it has the highest preference level at point D'. When $D \in [D_{min}, D']$, $U(D)$ is increasing and concave; when $D \in [D', D_{max}]$, $U(D)$ is decreasing and concave. The corresponding complementary problem is formulated as:

$$\text{Maximize} \sum_{m \in [M]} b_m y_m - \Delta(Z) + U(D) \quad (16)$$

$$\text{Subject To:} \quad \sum_{m \in [M]} e_m y_m - z \leq \sum_{m \in [M]} e_m - D \quad (16a)$$

$$y_m \in \{0, 1\}, \ \forall m \in [M] \quad (16b)$$

$$0 \leq z \leq z_{max} \quad (16c)$$

$$D_{min} \leq D \leq D_{max} \quad (16d)$$

Let (x, z_1, D_1) and (y, z_2, D_2) be a feasible solution for problem (15) and problem (16), respectively. We have $y = \vec{1} - x$, $z_2 = z_1$ and $D_2 = D_1$. (x, z_1, D_1) is an optimal solution to problem (15) *if and only if* (y, z_2, D_2) is an optimal solution to problem (16).

A challenge in the auction design for such a new model lies in computing the optimal solution (y, z_2, D_2) to the complementary problem (16) (Sec. IV-B). The rest of the FPTAS auction design is similar. We perturb the bids b_m to obtain a randomized algorithm with excepted polynomial running time. The expectation of the solution is at most $\epsilon \max_{m \in [M]} b_m$ more than the optimal social cost. A similar VCG-type randomized payment scheme then completes the truthful-in-expectation auction.

In Algorithm 1, for every Pareto optimal vector $y \in \mathcal{P}(M)$, we need to compute values for z_2 and D_2 that maximize $U(D) - \Delta(Z)$ and satisfy constraints (16c) and (16d):

$$\text{Maximize} \quad U(D) - \Delta(Z) \tag{17}$$

$$\text{Subject To:} \qquad D - z \leq \sum_{m \in [M]} e_m - e^T y \tag{17a}$$

$$0 \leq z \leq z_{max} \tag{17b}$$

$$D_{min} \leq D \leq D_{max} \tag{17c}$$

After computing z_2 and D_2 for each Pareto optimal vector $y \in \mathcal{P}(M)$, Algorithm 1 selects a solution that maximizes $\sum_{m \in [M]} b_m y_m - \Delta(Z) + U(D)$.

Assume that $\Delta(z)$ is an increasing convex function, as the marginal cost of electricity generation grows with the output rate z increases, then $-\Delta(z)$ is a concave function. Recall that $U(D)$ is a concave function. The sum of concave functions is still a concave function [27]. Therefore, the objective function $U(D) - \Delta(Z)$ is concave, and problem (17) becomes a classic convex minimization problem, which can be solved in polynomial-time using standard convex optimization methods such as the interior-point algorithm [28].

VII. Performance Evaluation

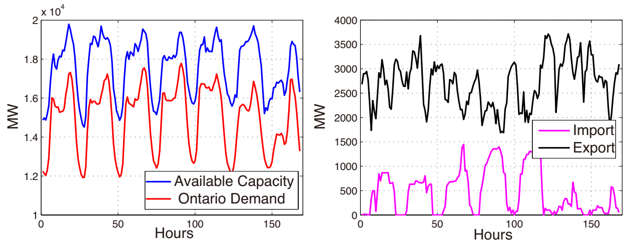

Fig. 3. Power supply and demand in Ontario, Canada, October 27, 2014 to November 2, 2014.

We evaluate our FPTAS demand response mechanism through trace-driven simulation studies, based on real-world demand data in Ontario, Canada in 2014. The left of Fig. 3 shows hourly demand and capacity of Ontario's grid from October 27, 2014 to November 2, 2014 [29]. *Available capacity* represents the capacity of Ontario's power market, including both local generation and imports. *Ontario demand* represents the actual power demand within Ontario, and is calculated by subtracting exports from the total power generation. We can obverse that the market capacity is always larger than actual demand. Furthermore, power generation and demand are imbalanced at times. Consequently, the power grid needs to import/export energy from/to other provinces. The right of Fig. 3 illustrates hourly electricity trading data. In the following simulations, the demand response target D is set to 100MW, or one fifth of the average hourly import, under the assumption that one fifth of the shortage is supplied by the auction and four-fifths is purchased from other provinces directly. Following a report about cost of electricity by source [26], the value of b_m is generated uniformly randomly from the interval $[200, 2000]$. The amount of supply/reduction e_m is a uniformly distributed number between 0MW to 10MW. The unite cost of the diesel generator is set to \$180 per MW, with a maximum output capacity of 10MW [30]. Each set of simulation is repeated ten times, and results are averaged.

Percentage of Cost Savings. Fig. 4 shows the percentage of social cost savings by our demand response auction with different demand response target D and different number of agents. Let C_o be the cost when the power grid uses its diesel generators to cover the demand-supply gap instead of resorting to demand response. Let C_n be the social cost returned by our randomized algorithm. The percentage of cost saving is computed as $\frac{C_o - C_n}{C_o}$. The cost data is taken from reports of US Energy Information Administration (EIA) of the U.S. Department of Energy [26], [30]. From Fig. 4, we can see that the demand response approach can save more than 50% of the cost when 40 agents submit demand response bids. Even when there are less agents (20 agents) submitting bids, it can still save more than 20% of the total cost. Moreover, the change of the demand response target D doesn't influence the percentage of cost saving.

Approximation Ratio. A salient feature of our randomized auction is its FPTAS property, *i.e.*, it achieves $(1 + \epsilon)$-optimal social cost. We first evaluate the approximation ratio of our Algorithm 2 under different system settings. Recall that Algorithm 2 solves the original minimization problem (1) by first solving the complementary problem (2), then converting it to a solution for problem (1). Fig. 5 and Fig. 6 compare the approximation ratio achieved by Algorithm 2 to solve MILP (1) and MILP (2), respectively. The ratios are computed by comparing the social cost achieved by Algorithm 2 to the optimal social cost.

Given $\alpha = 0.03$, Fig. 5 shows the approximation ratio with different number of agents. We can observe that the approximation ratio (red bars on the right side) for MILP (2) remains around 0.96 with the growth of the number of agents. That is because equation (9) indicates the gap between the approximation ratio of MILP (2) and optimal ratio 1 is $\frac{\alpha b^T y^*}{b^T y^* - c z^*}$, which is determined by the value of α as $\alpha c z^*$ is a very small number. The blue bars on the left side that represents the approximation ratio to solve the original problem (1) increases when the number of agents increases. This trend is in line with the theoretical analysis in Theorem 2. The difference between the objective value return by Algorithm 2 and the optimum is $\alpha b^T y^*$. Recall that y^* is the optimal solution to maximization problem (2). The value of $b^T y^*$ increases when number of agents grows, since y^* includes larger bids to maximize $b^T y^*$. We have proved that the gap between the blue bars (left side) and dotted line is

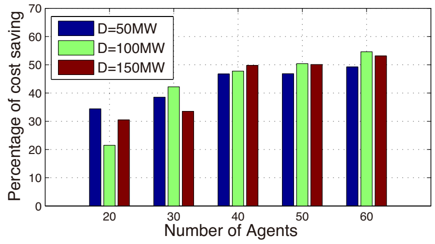

Fig. 4. Percentage of cost saving with different D and number of agents

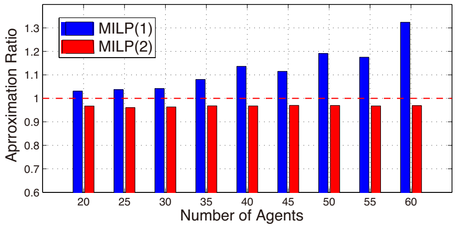

Fig. 5. Approximation ratio with different number of agents.

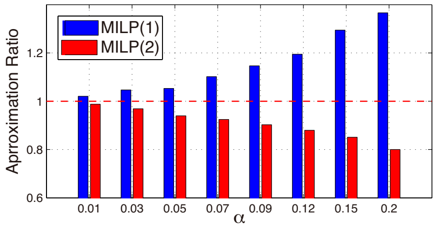

Fig. 6. Approximation ratio with different α.

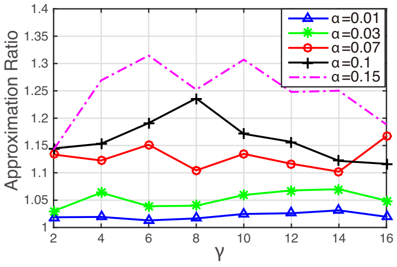

Fig. 7. Approximation ratio with different γ.

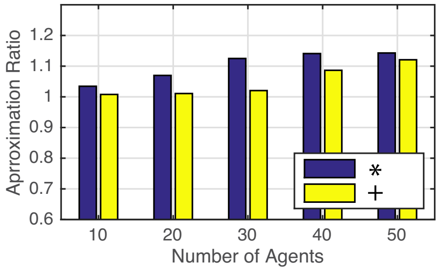

Fig. 8. Comparison between our algorithm^{+} with Algorithm 2 from Zhang *et al.*[7]*.

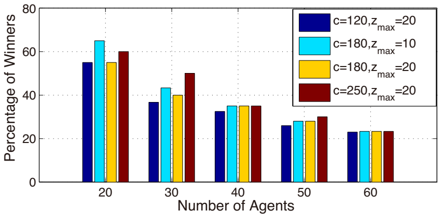

Fig. 9. Percentage of winners.

bounded by $\gamma\epsilon = \gamma M\alpha$. Our simulation results suggest that the gap is substantially smaller than the theoretical bound. Fig. 6 illustrates the approximation ratio under different α when 30 agents participate in the demand response process. With the increase of α, the approximation ratio of MILP (2) decreases and the approximation ratio of MILP (1) increases. Similar to the explanation for Fig. 5, the difference between approximation ratio of MILP (2) and optimal ratio 1 is close to α, while the gap between approximation ratio of MILP (1) and optimal ratio 1 is larger than α.

At the end of Sec. IV, we mention that the approximation ratio of our algorithm is upper bounded by $(1+\gamma\epsilon)$. Fig. 7 shows the approximation ratio achieved by Algorithm 2 to solve MILP (1) with different γ, where $\gamma = \max_{i,j\in[M]}\{b_i/b_j\}$. Although the theoretical analysis proves that $(1+\gamma\epsilon)$ is the upper bound of the approximation ratio, our simulations reveal a more rosy picture in practice. Furthermore, the value of γ doesn't affect the ratio, while α dominates the final ratio. We can observe that the approximation ratio fluctuates with the decrease of γ, but decreases monotonically when α drops. This can be intuitively explained as following: γ is used merely to bound the approximation ratio. It indicates the ratio in the worst case scenario rather than determines the ratio. Theorem 2 shows the real ratio still depends on the value of α. We also compare our algorithm with Algorithm 2 from Zhang *et al.* [7] to examine the approximation ratio when $\alpha = 0.01$, as shown in Fig. 8. It can be observed that both algorithms perform relatively well under the given input, achieving small approximation ratios that are between 1 and < 1.2. For all different number of agents tested, our algorithm outperforms that of Zhang *et al.* slightly.

Percentage of Winners. We next study the performance of our randomized algorithm in terms of winner satisfaction, as measured by the percentage of agents whose bid is accepted by the grid. Fig. 9 shows that more agents are selected by the grid when the number of participating agents is small. This is because the grid needs a large fraction of the agents to cover its shortage when there is only small number of choice. The cost and maximum capacity of diesel generators also influence the percentage of winners. Compare the percentage of winners when $c = \$120$ and $c = \$250$, we can observe that the percentage of winners with high cost is always larger. This can be explain as follows: when diesel generation is economical, our algorithm will first utilize diesel generators, and then consider demand response bids. But the grid selects all the bids from agents and avoids to use diesel generators when their costs are high.

Social Cost. Fig. 10 illustrates social cost computed by our randomized algorithm with varying number of agents and α. The smallest value occurs at the left bottom of the surface where α takes the smallest value and the number of agents is large. The social cost decreases when the number of agents increases. This is because a larger agent pool includes a larger number of agents who submit low-cost bids. With a small number of agents, the grid is forced to select expensive bids to meet the demand response target. Furthermore, given the same number of agents, a smaller α means the social cost is closer to the smallest social cost. This is the reason why we can observe a downward trend from the right side to the left side of the surface.

In Sec. VI, we extended our studies to non-constant demand response targets that are captured by the grid's utility function. Fig. 12 shows social cost achieved by 40 agents with different utility functions. We consider four quadratic utility functions [17] shown in Fig. 11. They are concave functions of the target D. The maximum and minimum target that grid can accept is 150MW and 50MW, respectively. The actual target of grid is set to 100MW. When $D = 100$MW, grid has the highest level of preference. For example, the dotted line represents the utility function $U(D) = 4(D-50)(D-150)$, the maximum level of preference is equivalent to to \$10000 when $D = 100$MW. When D is smaller or larger than

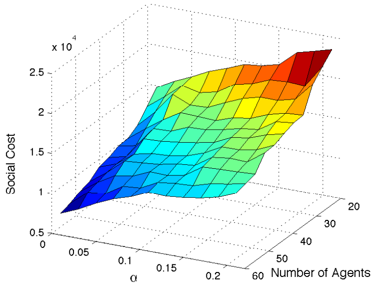

Fig. 10. Social cost with different α and number of agents.

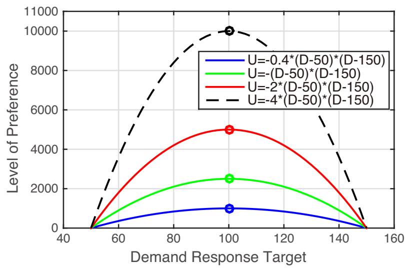

Fig. 11. Grid's utility function

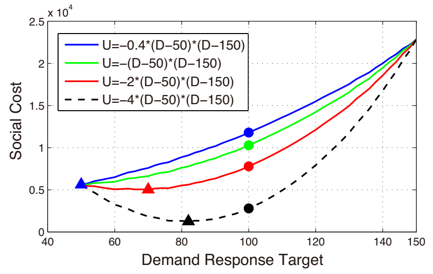

Fig. 12. Social cost with utility function

100MW, the level of preference decreases and reaches zero when $D = 50$MW or $D = 150$MW. However, the trend in Fig. 12 is not the same as in Fig. 11. We use circles to indicate the social cost when D is equal to the grid's actual demand response target, and triangles to mark the optimal social cost. Although the grid prefers 100MW the most, it is clear that the lowest cost may not occur at such a preferred point. Our algorithm can quickly compute a $(1 + \epsilon)$-optimal solution to these models in polynomial time.

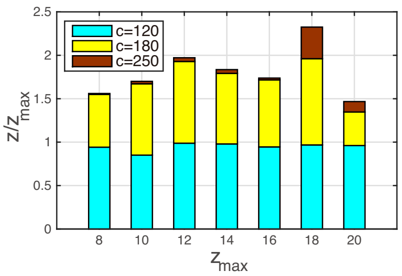

Fig. 13. Usage of diesel generators under different c and z_{max}.

Usage of Diesel Generators. Finally, we evaluate the usage of on-site quick-start generators, as exemplified by diesel generators, under different cost and generation capacity. We assumes that 40 agents participate in the auction. The Y axis in Fig. 13 represents the ratio of the output rate to the maximum power of the diesel generators, and the X axis is the maximum available power z_{max}. Here c is the unit cost, which equals \$120, or \$180 or \$250 per MW. Clearly, the higher the unit cost is, the lower the usage of diesel generators is. When the unit cost is low (light blue bars on the bottom layer), diesel generators operate at the maximum capacity. Red bars on the top layer show that the grid avoids to start its diesel generators when it is expensive to operate diesel generators. Furthermore, it is apparent that the cost, rather than the generation capacity, determines the usage of diesel generators.

VIII. Conclusions

This work formulates general and expressive models for demand response auctions. Through a new technique that combines smoothed polynomial-time algorithm design with randomized reduction, we designed demand response mechanisms that are truthful, polynomial-time computable, and can approach optimal social cost arbitrarily closely. The new technique of designing randomized auction mechanisms through smoothed polynomial-time algorithms may be applied to a broad range of problems where social welfare maximization is hard in the worst case, but the hard cases are relatively rare and isolated. However, similar to many other FPTAS types of algorithms and mechanisms in the future, our auction algorithms can be sometimes slow in practice despite its polynomial time running time, due to large exponents in the running time. Designing demand response auctions that are even more efficient for large scale practical applications remains an interesting and challenging problem.

References

[1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid?the new and improved power grid: A survey," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[2] T. Logenthiran, D. Srinivasan, and T. Shun, "Demand side management in smart grid using heuristic optimization," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1244–1252, 2012.

[3] A. Molina-García, F. Bouffard, and D. Kirschen, "Decentralized demand-side contribution to primary frequency control," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 411–419, 2011.

[4] K. Spees and L. Lave, "Impacts of responsive load in pjm: load shifting and real time pricing," *The Energy Journal*, pp. 101–121, 2008.

[5] NRDC, *America's Data Centers Consuming and Wasting Growing Amounts of Energy*, http://www.nrdc.org/energy/data-center-efficiency-assessment.asp.

[6] N. R. Herbst, S. Kounev, and R. Reussner, "Elasticity in cloud computing: What it is, and what it is not." in *Proc. of ICAC*, 2013.

[7] L. Zhang, Z. Li, and C. Wu, "Randomized auction design for electricity markets between grids and microgrids," in *Proc. of ACM SIGMETRICS*, 2014.

[8] D. P. Chassin, J. E. Dagle, and M. K. Donnelly, "Electrical power distribution control methods, electrical energy demand monitoring methods, and power management devices," 2006, uS Patent 7,149,605.

[9] K. Samarakoon, J. Ekanayake, and N. Jenkins, "Investigation of domestic load control to provide primary frequency response using smart meters," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 282–292, 2012.

[10] Wikipedia, *Dynamic demand*, http://en.wikipedia.org/wiki/Dynamic_demand_%28electric_power%29.

[11] Google, *AdWords*, http://www.google.com/adwords/?sourceid=awo&subid=ww-et-awhp_nelsontest_con&clickid.

[12] MyThings, *The global leader in customized programmatic ad solutions*, http://www.mythings.com/.

[13] P. Mobile, *The mobile media OSP*, http://www.plethoramobile.com/#!platform/c46c.

[14] L. Qian, Y. J. Zhang, J. Huang, and Y. Wu, "Demand response management via real-time electricity price control in smart grids," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1268–1280, 2013.

[15] W. Shi, N. Li, X. Xie, C. Chu, and R. Gadh, "Optimal residential demand response in distribution networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 7, pp. 1441–1450, 2014.

[16] A. Saber and G. Venayagamoorthy, "Resource scheduling under uncertainty in a smart grid with renewables and plug-in vehicles," *IEEE Systems Journal*, vol. 6, no. 1, pp. 103–109, 2012.

[17] P. Samadi, H. Mohsenian-Rad, R. Schober, and V. Wong, “Advanced demand side management for the future smart grid using mechanism design,” *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1170–1180, 2012.

[18] R. Zhou, Z. Li, and C. Wu, “An online procurement auction for power demand response in storage-assisted smart grids,” in *Proc. of IEEE INFOCOM*, 2015.

[19] Z. Zhou, F. Liu, Z. Li, and H. Jin, “When smart grid meets geo-distributed cloud: An auction approach to datacenter demand response,” in *Proc. of IEEE INFOCOM*, 2015.

[20] L. Lu, J. Tu, C. Chau, M. Chen, and X. Lin, “Online energy generation scheduling for microgrids with intermittent energy sources and co-generation,” in *Proc. of the ACM SIGMETRICS*, 2013.

[21] S. Martello and P. Toth, *Knapsack problems*. Wiley New York, 1990.

[22] D. A. Spielman and S. Teng, “Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time,” *Journal of the ACM (JACM)*, vol. 51, no. 3, pp. 385–463, 2004.

[23] S. Dughmi and T. Roughgarden, “Black-box randomized reductions in algorithmic mechanism design,” *SIAM Journal on Computing*, vol. 43, no. 1, pp. 312–336, 2014.

[24] J. Csirik, J. Frenk, M. Labbé, and S. Zhang, *Heuristics for the 0-1 min-knapsack problem*. European Institute for Advanced Studies in Management, 1990.

[25] S. Dobzinski and S. Dughmi, “On the power of randomization in algorithmic mechanism design,” in *Proc. of IEEE FOCS*, 2009.

[26] Wikipedia, *Cost of electricity by source*, http://en.wikipedia.org/wiki/Cost_of_electricity_by_source.

[27] Yahoo, *Sum of concave functions*, https://answers.yahoo.com/question/index?qid=20071021151748AAuasTT.

[28] Y. Nesterov and Y. Nemirovskii, A.and Ye, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[29] *Independent electricity system operator (ieso)*, https://ieso-public.sharepoint.com/.

[30] Wikipedia, *Diesel generator*, http://en.wikipedia.org/wiki/Diesel_generator.

Ruiting Zhou received a B.E. degree in telecommunication engineering from Nanjing University of Post and Telecommunication, China, in 2007, a M.S. degree in telecommunications from Hong Kong University of Science and Technology, Hong Kong, in 2008 and a M.S. degree in computer science from University of Calgary, Canada, in 2012. Since January, 2014, she has been a PhD student at the Department of Computer Science, University of Calgary, Canada. Her research interests include smart grids, cloud computing and mobile network optimization.

Zongpeng Li received his B.E. degree in Computer Science and Technology from Tsinghua University (Beijing) in 1999, his M.S. degree in Computer Science from University of Toronto in 2001, and his Ph.D. degree in Electrical and Computer Engineering from University of Toronto in 2005. Since August 2005, he has been with the Department of Computer Science in the University of Calgary. In 2011-2012, Zongpeng was a visitor at the Institute of Network Coding, Chinese University of Hong Kong. His research interests are in computer networks.

Chuan Wu received her B.Engr. and M.Engr. degrees in 2000 and 2002 from the Department of Computer Science and Technology, Tsinghua University, China, and her Ph.D. degree in 2008 from the Department of Electrical and Computer Engineering, University of Toronto, Canada. Since September 2008, Chuan Wu has been with the Department of Computer Science at the University of Hong Kong, where she is currently an Associate Professor. Her research is in the areas of cloud computing, online and mobile social networks. She was the co-recipient of the best paper award of HotPOST 2012.

Minghua Chen received his B.Eng. and M.S. degrees in EE from Tsinghua University in 1999 and 2001, respectively. He received his Ph.D. degree in EECS from University of California at Berkeley in 2006. He spent one year visiting Microsoft Research Redmond as a Postdoc Researcher. He joined the Dept. of Information Engineering, the Chinese University of Hong Kong in 2007, where he is currently an Associate Professor. He is also an Adjunct Associate Professor in Institute of Interdisciplinary Information Sciences, Tsinghua University. He received the Eli Jury award from UC Berkeley in 2007 and The Chinese University of Hong Kong Young Researcher Award in 2013. Minghua also received the IEEE ICME Best Paper Award in 2009, the IEEE Transactions on Multimedia Prize Paper Award in 2009, and the ACM Multimedia Best Paper Award in 2012. He is currently an Associate Editor of the IEEE/ACM Transactions on Networking. His recent research interests include energy systems, distributed optimization, multimedia networking, wireless networking, network coding, and delay-constrained communication.