Constrained Optimization with Decision-Dependent Distributions

Zifan Wang, Changxin Liu, Thomas Parisini, Michael M. Zavlanos, and Karl H. Johansson

***Abstract*—In this paper we deal with stochastic optimization problems where the data distributions change in response to the decision variables. Traditionally, the study of optimization problems with decision-dependent distributions has assumed either the absence of constraints or fixed constraints. This work considers a more general setting where the constraints can also dynamically adjust in response to changes in the decision variables. Specifically, we consider linear constraints and analyze the effect of decision-dependent distributions in both the objective function and constraints. Firstly, we establish a sufficient condition for the existence of a constrained equilibrium point, at which the distributions remain invariant under retraining. Moreover, we propose and analyze two algorithms: repeated constrained optimization and repeated dual ascent. For each algorithm, we provide sufficient conditions for convergence to the constrained equilibrium point. Furthermore, we explore the relationship between the equilibrium point and the optimal point for the constrained decision-dependent optimization problem. Notably, our results encompass previous findings as special cases when the constraints remain fixed. To show the effectiveness of our theoretical analysis, we provide numerical experiments on both a market problem and a dynamic pricing problem for parking based on real-world data.**

***Index Terms*—Constrained optimization, decision-dependent distributions, dual ascent algorithms**

I. Introduction

Machine learning algorithms typically use historical data under the assumption that these data adequately reflect future system behavior. However, in many settings, decisions made by machine learning algorithms will influence behavior of the system and lead to changes to the data distribution. For example, in dynamic pricing problems for parking management [1], [2], the users may adjust their decisions on whether and for how long to park, causing modifications in the distributions of future data. This phenomenon appears in numerous applications, such as online labor markets [3], predictive policing [4], and vehicle sharing markets [5]–[7], where strategic users react to changes of decisions and make adjustments accordingly.

To account for distribution shifts in the data in response to decisions influenced by machine learning models, new frameworks under the name of *performative prediction* or *optimization with decision-dependent distributions* have recently been developed [8]. Specifically, the term "performative prediction" was coined in [8] to illustrate the phenomenon that predictive models can trigger reactions that influence the outcome they aim to predict. This framework also incorporates a theoretical analysis that relies on defining a distribution map that relates decisions to data distributions and associating this map with a sensitivity property that employs the Wasserstein distance metric.

Many optimization problems with decision-dependent distributions also involve constraints that are themselves also influenced by the decisions. For example, in dynamic pricing for parking management problems, the total parking time, as a strategic feature in response to the change of the price (decision), makes the constraints decision-dependent if the total parking time plays a role in defining the constraints. Motivated by such scenarios, in this paper we investigate constrained optimization with decision-dependent distributions affecting both the objective function and constraints. Specifically, we consider linear constraints that depend on the decision variables and introduce two notions of solution points: the constrained equilibrium point and constrained optimal point. Assuming that the distribution maps in both the objective function and constraints are insensitive, we show that the constrained equilibrium point is unique and repeated constrained minimization (RCM) converges to this unique equilibrium point. Notably, when the distributions in the constraints are independent of the decisions, the sufficient condition for convergence of RCM reduces to that for repeated risk minimization in [8]. As RCM requires the exact solution of a constrained optimization problem, we relax this requirement and propose the repeated dual ascent (RDA) algorithm. We show that convergence of RDA to the constrained equilibrium point is guaranteed under more restrictive conditions compared to RCM. Additionally, we present convergence analysis for both RCM and RDA when only the constraints are decision-dependent and the distribution in the objective function remains fixed. Finally, we analyze the relation between the constrained equilibrium points and the constrained optimal points. Under mild conditions, we

This work was supported in part by Swedish Research Council Distinguished Professor Grant 2017-01078, Knut and Alice Wallenberg Foundation, Wallenberg Scholar Grant, the Swedish Strategic Research Foundation CLAS Grant RIT17-0046, AFOSR under award #FA9550-19-1-0169, NSF under award CNS-1932011, the Digital Futures Scholar-in-Residence Program, and the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 739551 (KIOS CoE). (Corresponding author: Changxin Liu.)

Zifan Wang and Karl H. Johansson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, and also with Digital Futures, SE-10044 Stockholm, Sweden (e-mail: zifanw@kth.se; kallej@kth.se).

Changxin Liu is with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mail: changxinl@ecust.edu.cn).

Thomas Parisini is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK, with the Department of Electronic Systems, Aalborg University, Denmark, and with the Department of Engineering and Architecture, University of Trieste, Italy (e-mail: t.parisini@imperial.ac.uk).

Michael M. Zavlanos is with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA (e-mail: michael.zavlanos@duke.edu).

show that the distance between these points can be bounded. This result also encompasses the case when the constraints are fixed, as explored in [8]. To validate our theoretical analysis, we present numerical experiments on a market problem and a dynamic pricing problem for parking management that utilizes an open-sourced dataset.

To the best of our knowledge, constrained optimization with decision-dependent constraints has not been explored in the literature. Most closely related to our study is the vast literature on optimization with decision-dependent distributions that exclusively occur in the objective function [7]–[18]. In this context, there are no constraints or the constraints remain fixed throughout the optimization process. Among these works, the authors in [8] first provide a comprehensive framework for handling decision-dependent problems. Specifically, they define the equilibrium and optimal points, and show that repeated risk minimization and repeated gradient descent converge to the equilibrium point when the distribution map is less sensitive. Moreover, under certain conditions, [8] shows that the distance between equilibrium and optimal points can be bounded. In the case that only stochastic gradient is available, the work in [9] studies stochastic optimization problems with decision-dependent distributions and proposes two variants of the stochastic gradient method that converge to the equilibrium point. The follow-up work in [10] focuses on directly optimizing the performative risk and proposes algorithms to find the optimal point. Subsequent works have extended this framework to different scenarios. For example, [7] investigates multi-agent games with decision-dependent distributions and analyzes the convergence to the Nash equilibrium in strongly monotone games. Similarly, [19] defines performative equilibrium equilibrium in multi-agent networked games and analyzes convergence to this equilibrium. However, common in all the above works is that there are no constraints or the constraints are fixed. The introduction of decision-dependent constraints adds significant complexity to the optimization process, as these constraints now adapt to the decisions made, necessitating the development of more sophisticated techniques in the dual domain.

Related to our work is also the literature on optimization with time-varying constraints [20]–[23]. The reason is that when the decisions change during the optimization process, the constraints become effectively time-varying and impact the feasible solution space and further the optimal solution. The work in [20] derives a quantitative bound on the rate of change of the optimizer when the objective function and constraints are perturbed concurrently. This bound is then used to quantify the tracking performance for continuous-time algorithms in time-varying constrained optimization. In the context of optimization with decision-dependent distributions, all existing works assume static constraints, with the exception of [24]. In [24], the authors consider online optimization with decision-dependent distributions where the objective function and constraints vary over time, leading to time-varying equilibrium points. They propose a projected gradient descent method and demonstrate its ability to track the trajectory of the evolving equilibrium points. However, all works discussed before, including [24], do not leverage the crucial information that the perturbations in the objective function and/or constraints are induced by the changes in the decisions. As a result, the techniques presented in these works cannot be used to analyze optimization problems with decision-dependent constraints as the ones considered here.

The paper is organized as follows. In Section II, we formulate our problem and provide some preliminary results. In Section III, we propose two algorithms and analyze their convergence to the constrained equilibrium point. Section IV analyzes the relation between the constrained equilibrium and optimal points. Section V provides the numerical validation of the proposed algorithms for a market problem and a dynamic pricing problem for parking management. Finally, concluding remarks are given in Section VI.

II. Problem Setup

In this section, we formulate the problem and provide some preliminary results. Throughout the paper, we let $\mathbb{R}^n$ denote n-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and 2-norm $\|\cdot\|$. We equip $\mathbb{R}^n$ with σ-algebra. For a matrix $A \in \mathbb{R}^{m\times n}$, we denote by $\|A\|_2$ the maximum singular value of the matrix A.

A. Main Definitions

Consider the following linearly constrained optimization problem:

$$\begin{aligned}\min_{x}\ & \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)}[l(x,z)]\\ \text{s.t.}\quad & Gx \le \mathop{\mathbb{E}}_{w\sim\mathcal{D}_g(x)}[w],\end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the decision variable, $G \in \mathbb{R}^{d_w\times n}$ is a matrix associated with the linear constraint, and $z \in \mathbb{R}^{d_z}$ and $w \in \mathbb{R}^{d_w}$ are the random variables in the objective function and constraint, respectively. We assume that the loss function $l : \mathbb{R}^n\times\mathbb{R}^{d_z} \to \mathbb{R}$ is twice continuously differentiable in (x, z). The random variable z follows the distribution $\mathcal{D}(x)$ that depends on the decision variable x, where $\mathcal{D}$ is the distribution map from $\mathbb{R}^n$ to the space of distributions. In addition, we assume that the constraints are decision-dependent, i.e., the random variable w depends on the decision variable through the distribution $\mathcal{D}_g(x)$, where $\mathcal{D}_g$ is the distribution map from $\mathbb{R}^n$ to the space of distributions. Let z and w take values in the metric spaces $\mathcal{Z}$ and $\mathcal{W}$, respectively. We assume that suitable Borel measurability conditions hold so that the expected value operators $\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)}[\cdot]$ and $\mathop{\mathbb{E}}_{z\sim\mathcal{D}_g(x)}[\cdot]$ are well-defined, hence implying that the minimization operations are well-defined. We define the solution points below.

Definition 1. *(Constrained Optimal Point). A vector x_o is called a constrained optimal point if it satisfies*

$$\begin{aligned}x_o = \arg\min_{x}\ & \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)}[l(x,z)]\\ \text{s.t.}\quad & Gx \le \mathop{\mathbb{E}}_{w\sim\mathcal{D}_g(x)}[w].\end{aligned}$$

A constrained optimal point solves the problem (1), but is usually hard to compute since the distribution maps $\mathcal{D}$ and

$\mathcal{D}_g$ are generally unknown to the decision maker. Therefore, we introduce an alternative solution point below, which can be regarded as a fixed point of a repeated training process.

Definition 2. *(Constrained Equilibrium Point). A vector x_s is called a constrained equilibrium point if it satisfies*

$$x_s = \arg\min_{x} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)] \quad \text{s.t.} \quad Gx \leq \mathop{\mathbb{E}}_{w\sim\mathcal{D}_g(x_s)}[w].$$

The equilibrium point defined in Definition 2 denotes a class of points, not necessarily optimal for (1), that solve the constrained optimization problem using the distributions they induce. It is also referred to as a stable point in [8].

In this work, we make the assumption that a constrained optimal point exists for (1), which is a commonly accepted assumption in decision-dependent optimization and linearly constrained optimization problems. Furthermore, our analysis of two optimization methods will demonstrate the existence of equilibrium points under certain conditions on the loss function and the distribution map.

Our goal is to solve problem (1) by first determining the existence and uniqueness of constrained equilibrium points and next developing algorithms that converge to the constrained equilibrium points. Moreover, we aim to explore the relation between the constrained equilibrium points and the optimal points.

B. Key Assumptions and Preliminary Results

The distribution maps in (1) play a crucial role in determining how the distributions respond to changes in decisions. In order to upper bound the rate of change at which the distributions change, we impose a Lipschitz condition on the distribution map, known as ϵ-sensitivity.

Definition 3. *(ϵ-sensitivity). We say that a distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive if for all x, x', we have*

$$W_1(\mathcal{D}(x), \mathcal{D}(x')) \leq \epsilon \|x - x'\|, \tag{2}$$

where W_1 denotes the earth mover's distance.

The earth mover's distance, also known as 1-Wasserstein distance, is a natural metric for quantifying the dissimilarity between probability distributions. According to Kantorovich-Rubinstein duality theorem [25], the earth mover's distance of two probability distributions P and Q can be expressed as

$$W_1(P,Q) = \sup_{f\sim \mathrm{Lip}_1} \{\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]\},$$

where Lip_1 represents the set of 1-Lipschitz continuous functions. We make the following assumptions on the cost function and the distribution maps.

Assumption 1. *The loss function $l(x,z)$ is γ-strongly convex in x for every $z \in \mathcal{Z}$, i.e.,*

$$l(x,z) \geq l(x',z) + \langle \nabla_x l(x',z), x - x' \rangle + \frac{\gamma}{2}\|x - x'\|^2,$$

for all $x, x' \in \mathbb{R}^n$.

Assumption 2. *$\nabla_x l(x,z)$ is β_z-Lipschitz continuous in z for every $x \in \mathbb{R}^n$, i.e.,*

$$\|\nabla_x l(x,z) - \nabla_x l(x,z')\| \leq \beta_z \|z - z'\|,$$

for all $z, z' \in \mathcal{Z}$, and $\nabla_x l(x,z)$ is β_x-Lipschitz continuous in x for every $z \in \mathcal{Z}$, i.e.,

$$\|\nabla_x l(x,z) - \nabla_x l(y,z)\| \leq \beta_x \|x - y\|,$$

for all $x, y \in \mathbb{R}^n$.

Assumption 3. *The distribution maps $\mathcal{D}$ and $\mathcal{D}_g$ are ϵ- and ϵ_g-sensitive, respectively, namely,*

$$W_1(\mathcal{D}(x), \mathcal{D}(y)) \leq \epsilon \|x - y\|,$$
$$W_1(\mathcal{D}_g(x), \mathcal{D}_g(y)) \leq \epsilon_g \|x - y\|,$$

for all $x, y \in \mathbb{R}^n$.

It is worth noting that Assumptions 1–3 are quite standard in the literature, e.g., [8]–[10]. Moreover, we introduce the following assumption on the constraints.

Assumption 4. *The matrix G has full row rank.*

Assumption 4 is a technical condition that ensures the strong concavity of the Lagrange dual function, and the uniqueness of the optimal dual variable [26]. These two properties facilitate the contraction analysis of the proposed dual ascent algorithm. This assumption is commonly used in the literature regarding resource allocation [27] and primal-dual optimization [28]. Relaxing the rank condition would introduce additional complexity to the theoretical analysis, making it more challenging to establish rigorous results. It is interesting to examine whether it can be relaxed to other constraint qualifications, e.g., linear independence constraint qualification (LICQ) or Slater condition; we leave it for future research.

In what follows, we present two important lemmas that will be used in the subsequent analysis.

Lemma 1. *[8] Suppose that the loss $l(x,z)$ is γ-strongly convex in x for every $z \in \mathcal{Z}$, $\nabla_x l(x,z)$ is β_z-Lipschitz continuous in z, and the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive. Define $F(x') = \arg\min_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')}[l(x,z)]$, where $\mathcal{X}$ is a convex set. For all $x, x' \in \mathcal{X}$, we have*

$$\|F(x) - F(x')\| \leq \frac{\epsilon\beta_z}{\gamma}\|x - x'\|. \tag{3}$$

Lemma 1 shows that the variation of the optimal solution across two distributions is bounded by the distance between the corresponding points that generate these distributions. Note that Lemma 1 assumes fixed constrains and, therefore, cannot be used here. The following lemma explores the sensitivity of the optimal value with respect to perturbations of the constraint.

Lemma 2. *[20] Consider the convex optimization problem:*

$$\min_{x\in\mathcal{X}} \quad f(x) \quad \text{s.t.} \quad Gx \leq v, \tag{4}$$

Algorithm 1: Repeated Constrained Minimization

1: **Input**: Initial variable x_0.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: $\quad x_{t+1} = \arg\min_x f_{x_t}(x) \quad \text{s.t.} \quad Gx \leq \xi(x_t)$;
4: **end for**
5: **Output**: Sequence $\{x_t\}$

where f is γ-strongly convex, β-smooth and twice continuously differentiable, $G \in \mathbb{R}^{d_w \times n}$ has full row rank, and $\mathcal{X}$ is a convex set. Suppose that $GG^{\mathrm{T}} \geq \underline{w}^2 I$ for some $\underline{w} > 0$. Then, there exists a continuous map $x^*(v)$ such that $x^*(v)$ is the optimum of the problem (4) *and satisfies $\|\nabla_v x^*(v)\| \leq l_{x^*} = \sqrt{\frac{\beta}{\gamma}}\frac{1}{\underline{w}}$.*

III. Algorithms that Converge to Equilibrium Points

This section focuses on the analysis of constrained equilibrium points. Specifically, we discuss the existence and uniqueness of constrained equilibrium points. Moreover, we introduce two algorithms: RCM and RDA, and present sufficient conditions required for each algorithm to converge to the equilibrium point.

A. Repeated Constrained Minimization

We begin by analyzing RCM. We establish a sufficient condition for RCM to converge, which also ensures the uniqueness of the constrained equilibrium point.

Starting with an initial decision variable, RCM repeatedly solves a constrained optimization problem with the distribution induced by the previous decision. The RCM algorithm is presented in Algorithm 1. Specifically, at iteration $t+1$, we perform the following update:

$$x_{t+1} = \arg\min_x f_{x_t}(x) \quad \text{s.t.} \quad Gx \leq \xi(x_t), \tag{5}$$

where $f_{x_t}(x) := \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_t)}[l(x,z)]$ and $\xi(x_t) := \mathop{\mathbb{E}}_{w\sim\mathcal{D}_g(x_t)}[w]$. The convergence of RCM is closely related to the existence and uniqueness of the constrained equilibrium point. The main result is presented in the following theorem. The proof can be found in Appendix A.

Theorem 1. *Suppose Assumptions 1–4 hold and let $L_{x^*} := \sqrt{\frac{\beta_x}{\gamma\lambda_{\min}(GG^{\mathrm{T}})}}$. If*

$$\frac{\epsilon\beta_z}{\gamma} + L_{x^*}\epsilon_g < 1, \tag{6}$$

then problem (1) *admits a unique equilibrium point x_s and RCM converges to x_s at a linear rate.*

Theorem 1 demonstrates that RCM converges to the unique equilibrium point at a linear rate when the parameters ϵ and ϵ_g are small enough, i.e., the sensitivity of the optimization problem is small. When the constraints are fixed ($\epsilon_g = 0$), RCM is equivalent to repeated risk minimization (RRM) discussed in [8]. In this case, the sufficient condition in (6) required for RCM reduces to $\frac{\epsilon\beta_z}{\gamma} < 1$, which aligns with the previous result. Therefore, RCM subsumes RRM as a special case.

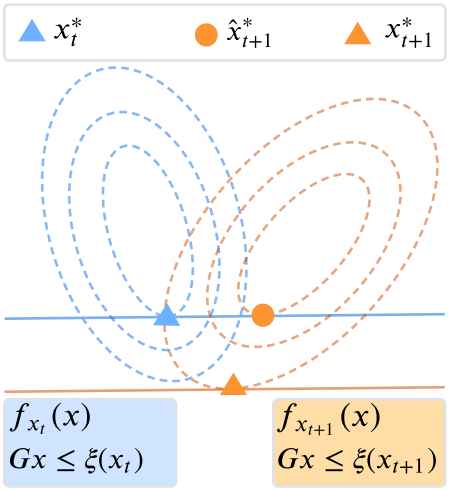

Fig. 1. Illustration of an RCM iteration. The decision-dependent constraint shifts the location of the optimum.

When the constraints change as a function of decisions, the RCM dynamics become more intricate than RRM in [8], as the solution at each iteration may be influenced by the variation in the constraints. Fig. 1 illustrates the dynamics of RCM. Under the constraint $Gx \leq \xi(x_t)$, the points x_t^* and $\hat{x}_{t+1}^*$ represent minima of the objective functions $f_{x_t}(x)$ and $f_{x_{t+1}}(x)$, respectively. However, since the constraint shifts to $Gx \leq \xi(x_{t+1})$, the minimum at iteration $t+1$ is x_{t+1}^* instead of $\hat{x}_{t+1}^*$. This constraint shift, which is usually hard to predict, complicates the analysis of RCM as shown in the proof of Theorem 1.

When the objective function is fixed ($\epsilon = 0$), implying that the distribution in the objective function is decision-independent, the convergence result is as follows. The proof follows directly from Theorem 1.

Corollary 1. *Suppose $\epsilon = 0$ and Assumptions 1–4 hold, and let $L_{x^*} := \sqrt{\frac{\beta_x}{\gamma\lambda_{\min}(GG^{\mathrm{T}})}}$. If $L_{x^*}\epsilon_g < 1$, then problem* (1) *admits a unique equilibrium point x_s and RCM converges to x_s at a linear rate.*

In the case that the distribution in the objective function remains fixed, the convergence of RCM is thus ensured under the condition $L_{x^*}\epsilon_g < 1$. The parameter L_{x^*} can be interpreted as an intrinsic property of the system that characterizes the sensitivity to perturbations in the constraints. The parameter ϵ_g quantifies the extent to which the constraints are sensitive to changes in the decisions. For convergence, it is essential that both L_{x^*} and ϵ_g are sufficiently small. In other words, it is important that the system exhibits low sensitivity to perturbations in the constraints, while simultaneously ensuring that the constraints are also minimally affected by changes in decisions. This ensures that the sequence of optimal solutions remains insensitive to the changes in the decision-dependent distributions.

When $\epsilon = 0$, the sufficient condition $L_{x^*}\epsilon_g < 1$ for RCM to converge is in fact tight. This can be illustrated in the following

simple example.

Example. Consider the minimization problem with the loss function x^2 and the constraint $x \geq \xi(x_t)$, where $x \in \mathbb{R}$ is the decision variable, and $\xi(x_t) = \theta x_t$ with $\theta > 0$. In this problem, it is easy to verify that $\epsilon = 0$, $\epsilon_g = \theta$, $\gamma = \beta_x = 2$, $G = 1$, and $L_{x^*} = 1$. If we run RCM from a positive initial point $x_0 > 0$, it is easy to verify that $x_t = \theta^t x_0$. The convergence of RCM requires the condition $\theta < 1$, which is equivalent to $L_{x^*}\epsilon_g < 1$.

Remark 1. *Although the sufficient condition* (6) *seems to be conservative, it is in fact tight in two specific scenarios: when $\epsilon = 0$ and when $\epsilon_g = 0$, as evidenced by Proposition 3.6 in [8] and the example provided above, respectively. Since this condition is applicable to a wide range of distribution maps, relaxing it when ϵ and ϵ_g are both non-zero poses a considerable challenge, which is left as our future work.*

B. Repeated Dual Ascent

Theorem 1 analyzes the convergence of RCM when the sensitivity parameters ϵ, ϵ_g are small enough. However, implementing RCM requires the exact solution of a constrained optimization problem at each iteration, which might be computationally inefficient. To address this limitation, we introduce a dual ascent algorithm RDA in this section. Sufficient conditions are proposed that guarantee the convergence of this algorithm to the constrained equilibrium point.

To facilitate the analysis of varying constraints, we define the Lagrangian function

$$L_{x'}(x, \lambda) := f_{x'}(x) + \lambda^{\mathrm{T}}(Gx - \xi(x'))$$

where $f_{x'}(x) := \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')}[l(x, z)]$, and the dual function

$$d_{x'}(\lambda) := \min_x L_{x'}(x, \lambda).$$

It is easy to verify that the loss function $f_{x'}(x)$ is γ-strongly convex and β_x-smooth in x for every $x' \in \mathbb{R}^n$ by taking the expectation of the loss function $l(x, z)$ with respect to $z \sim \mathcal{D}(x')$.

Note that in this setting the dual function can be written as

$$d_{x'}(\lambda) = -\tilde{f}_{x'}(-G^{\mathrm{T}}\lambda) - \lambda^{\mathrm{T}}\xi(x'), \tag{7}$$

where the conjugate function $\tilde{f}_{x'}$ is defined as $\tilde{f}_{x'}(y) := \max_x \langle y, x\rangle - f_{x'}(x)$. The gradient of the dual function is given by

$$\nabla d_{x'}(\lambda) = G\nabla\tilde{f}_{x'}(-G^{\mathrm{T}}\lambda) - \xi(x'). \tag{8}$$

We present the RDA as Algorithm 2. Specifically, at iteration t, the primal variable x_t is obtained by minimizing the Lagrangian function $L_{x_{t-1}}(x, \lambda_t)$, i.e.,

$$x_t = \arg\min_x L_{x_{t-1}}(x, \lambda_t). \tag{9}$$

Then, the dual variable is updated via gradient ascent:

$$\lambda_{t+1} = [\lambda_t + \eta\nabla d_{x_t}(\lambda_t)]_+ = [\lambda_t + \eta(G\bar{y}_t - \xi(x_t))]_+, \tag{10}$$

where $\bar{y}_t = \arg\min_x L_{x_t}(x, \lambda_t)$ and $[a]_+ = \max\{a, 0\}$. Note that the update direction in (10) is the positive gradient of the dual function d_{x_t} instead of $d_{x_{t-1}}$. In order to compute the gradient of the new dual function $\nabla d_{x_t}(\lambda_t)$, we need to find the minima $\bar{y}_t$ by minimizing the loss function $L_{x_t}(x, \lambda_t)$.

Algorithm 2: Repeated Dual Ascent

1: **Input**: Initial variables λ_1, x_0, step size η.
2: **for** $t = 1, 2, \ldots$ **do**
3: $\quad x_t = \arg\min_x L_{x_{t-1}}(x, \lambda_t)$;
4: $\quad \bar{y}_t = \arg\min_x\{f_{x_t}(x) + \lambda_t^{\mathrm{T}}(Gx - \xi(x_t))\}$;
5: $\quad \lambda_{t+1} = [\lambda_t + \eta(G\bar{y}_t - \xi(x_t))]_+$;
6: **end for**
7: **Output**: Sequences $\{x_t\}$, $\{\lambda_t\}$

Given that $f_{x'}(x)$ is γ-strongly convex and β_x-smooth in x for every $x' \in \mathbb{R}^n$, it can be shown that the dual function $d_{x'}(\lambda)$ is strongly concave and smooth. This result is summarized in the following lemma and the proof can be found in the literature; see Proposition 3.1 in [29] for the analysis of strong concavity and Lemma 3.2 in [30] for the analysis of smoothness.

Lemma 3. *Suppose Assumptions 1, 2, and 4 hold. There holds that, for every $x' \in \mathbb{R}^n$, $d_{x'}(\lambda)$ is γ_d-strongly concave, i.e.,*

$$d_{x'}(\lambda) \leq d_{x'}(\lambda') + \langle \lambda - \lambda', \nabla d_{x'}(\lambda')\rangle + \frac{\gamma_d}{2}\|\lambda - \lambda'\|^2,$$

with $\gamma_d = \frac{\lambda_{\min}(GG^{\mathrm{T}})}{L}$, and L_d-smooth, i.e.,

$$\|\nabla d_{x'}(\lambda) - \nabla d_{x'}(\lambda')\| \leq L_d\|\lambda - \lambda'\|,$$

with $L_d = \frac{\|G\|_2^2}{\gamma}$.

We say that the linear independence constraint qualification (LICQ) is satisfied if the gradients of all active constraints are linearly independent. The following lemma discusses LICQ and the uniqueness of the optimal dual variable.

Lemma 4. *[26] Suppose Assumptions 1 and 4 hold. Then, the LICQ holds and the optimal dual variable is unique.*

Now we are ready to present the convergence result of RDA. The proof can be found in Appendix B.

Theorem 2. *Suppose Assumptions 1–4 hold, and denote by $\{(x_t, \lambda_t)\}$ the sequence generated by RDA. If*

$$\left(\frac{\epsilon\beta_z\|G\|_2}{\gamma} + \epsilon_g\right)\left(1 + \frac{\epsilon\beta_z\|G\|_2}{\gamma^2} + \frac{\|G\|_2^2}{\gamma^2}\right) < 2\gamma_d \tag{11}$$

and

$$\frac{\epsilon\beta_z}{\gamma}\left(1 + \frac{\epsilon\beta_z\|G\|_2}{\gamma^2}\right) < 1 \tag{12}$$

hold, then there exist $s_2 > s_1 \geq 0$ such that for all step size $\eta \in (s_1, s_2)$ we have

$$\begin{aligned}&\|\lambda_{t+1} - \lambda_s^*\|^2 + \alpha\|x_t - x_s\|^2 \\ &\leq \kappa^t(\|\lambda_1 - \lambda_s^*\|^2 + \alpha\|x_0 - x_s\|^2),\end{aligned} \tag{13}$$

where λ_s^* is the optimal dual variable of the dual function $d_{x_s}(\lambda)$, and

$$\begin{aligned}
&\rho_1 = \frac{\epsilon\beta_z}{\gamma},\ \rho_2 = \rho_1 \|G\|_2 + \epsilon_g,\\
&\rho_3 = \frac{\rho_1 \|G\|_2}{\gamma} + \frac{\|G\|_2^2}{\gamma^2},\ \rho_4 = \rho_1^2 + \frac{\rho_1 \|G\|_2}{\gamma},\\
&a_1 = L_d^2 + L_d\rho_2 + \rho_2^2\rho_3 + L_d\rho_2\rho_3,\\
&a_2 = 2\gamma_d - \rho_2(1+\rho_3),\ a_3 = \alpha\rho_3,\\
&s_1 = \frac{a_2 - \sqrt{a_2^2 - 4a_1a_3}}{2a_1},\\
&b_1 = \rho_4(\rho_2^2 + L_d\rho_2),\ b_2 = \rho_2\rho_4,\ b_3 = \alpha(1-\rho_4),\\
&s_2 = \frac{-b_2 + \sqrt{b_2^2 + 4b_1b_3}}{2b_1},\ \alpha = (1-\sqrt{\bar{\alpha}})\frac{a_2^2}{4\rho_3 a_1},\\
&\sqrt{\bar{\alpha}} = \max\left\{0, 1 - \frac{L_d^2}{4\gamma_d(\rho_2 + L_d)}\right\},\\
&\kappa = \max(\kappa_1, \kappa_2),\ \kappa_1 = \eta^2 a_1 - \eta a_2 + \alpha a_3 < 1,\\
&\kappa_2 = \eta^2 b_1 + \eta b_2 + b_3 < 1.
\end{aligned}$$

Theorem 2 shows that RDA converges to the equilibrium point at a linear rate under conditions (11) and (12). Note that the sufficient condition in (11) is more restrictive than the condition in (6) that guarantees the convergence of RCM to the equilibrium point, because

$$\begin{aligned}
&\frac{1}{2\gamma_d}\left(\frac{\epsilon\beta_z}{\gamma}\|G\|_2 + \epsilon_g\right)\left(1 + \frac{\epsilon\beta_z\|G\|_2}{\gamma^2} + \frac{\|G\|_2^2}{\gamma^2}\right)\\
&\geq \frac{1}{2\gamma_d}\left(\frac{\epsilon\beta_z}{\gamma}\|G\|_2 + \epsilon_g\right)\frac{2\|G\|_2}{\gamma}\\
&= \frac{\epsilon\beta_z}{\gamma}\frac{\|G\|_2^2\beta_x}{\gamma\lambda_{\min}(GG^{\mathrm{T}})} + L_{x^*}\epsilon_g\sqrt{\frac{\|G\|_2^2\beta_x}{\gamma\lambda_{\min}(GG^{\mathrm{T}})}}\\
&\geq \frac{\epsilon\beta_z}{\gamma} + L_{x^*}\epsilon_g,
\end{aligned}$$

where in the first inequality we use $1 + \frac{\|G\|_2^2}{\gamma^2} \geq \frac{2\|G\|_2}{\gamma}$ and in the equality we use the definitions $\gamma_d = \frac{\lambda_{\min}(GG^{\mathrm{T}})}{\beta_x}$ and $L_{x^*} = \sqrt{\frac{\beta_x}{\gamma\lambda_{\min}(GG^{\mathrm{T}})}}$. The last inequality follows from the facts that the strong convexity parameter is not larger than the smooth parameter, i.e., $\gamma \leq \beta_x$ and $\lambda_{\min}(GG^{\mathrm{T}}) \leq \|G\|_2^2$. Moreover, it follows that the condition (11) together with Assumptions 1–4 ensures the uniqueness of the equilibrium point.

Remark 2. *We would like to highlight the technical innovations of RDA compared to RGD in [8]. Unlike RGD, which is analyzed in the primal space, the convergence of RDA is examined in the dual space by analyzing the dynamics of the dual variable. The proof of RDA, which involves the use of Lagrangian and conjugate functions in the dual space, is novel and has not been previously employed in the study of decision-dependent optimization.*

Remark 3. *We note that RDA deviates from the traditional dual ascent algorithm by involving the solution of two minimization problems at each iteration, as opposed to the traditional algorithm that only requires one. The additional computation of $\bar{y}_t$ is implemented to ensure that the dual variable performs an update in the direction of the updated gradient $\nabla d_{x_t}(\lambda_t)$. An alternative way to update the dual variable is to use $\nabla d_{x_{t-1}}(\lambda_t) = Gx_t - \xi(x_{t-1})$, which eliminates the need for computing $\bar{y}_t$. However, the theoretical analysis of the algorithm with this update becomes challenging and we leave it as a subject for future work.*

When the objective function is fixed ($\epsilon = 0$), implying that the distribution in the objective function is decision-independent, the convergence of RDA is presented in the following result. The proof can be found in Appendix C.

Corollary 2. *Suppose $\epsilon = 0$ and Assumptions 1–4 hold, and denote by $\{(x_t, \lambda_t)\}$ the sequence generated by RDA. If*

$$\epsilon_g\left(1 + \frac{\|G\|_2^2}{\gamma^2}\right) < 2\gamma_d \tag{14}$$

and

$$\eta < \frac{2\gamma_d - \epsilon_g(1 + \frac{\|G\|_2^2}{\gamma^2})}{L_d^2 + \epsilon_g L_d + \epsilon_g^2\frac{\|G\|_2^2}{\gamma^2} + \epsilon_g^2 L_d\frac{\|G\|_2^2}{\gamma^2}},$$

then there holds

$$\|\lambda_{t+1} - \lambda_s^*\|^2 \leq \kappa_3^t \|\lambda_1 - \lambda_s^*\|^2 \tag{15}$$

and

$$\|x_{t+1} - x_s\|^2 \leq \kappa_3^{t-1}\frac{\|G\|_2^2}{\gamma^2}\|\lambda_1 - \lambda_s^*\|^2, \tag{16}$$

where $\kappa_3 = 1 - 2\eta\gamma_d + \eta^2 L_d^2 + \epsilon_g\eta + \epsilon_g L_d\eta^2 + \frac{\|G\|_2^2}{\gamma^2}(\epsilon_g^2\eta^2 + \epsilon_g\eta + \epsilon_g L_d\eta^2)$.

Condition (14) required for RDA to converge is more stringent than the condition $L_{x^*}\epsilon_g < 1$ that ensures the convergence of RCM, because

$$\begin{aligned}
&\frac{1}{2\gamma_d}\epsilon_g\left(1 + \frac{\|G\|_2^2}{\gamma^2}\right) \geq \frac{1}{2\gamma_d}\epsilon_g\frac{2\|G\|_2}{\gamma}\\
&\geq \epsilon_g L_{x^*}\sqrt{\frac{\beta_x\|G\|_2^2}{\gamma\lambda_{\min}(GG^{\mathrm{T}})}} \geq \epsilon_g L_{x^*},
\end{aligned}$$

where the last inequality holds since $\beta_x \geq \gamma$ and $\|G\|_2^2 \geq \lambda_{\min}(GG^{\mathrm{T}})$. This aligns with the case when $\epsilon > 0$.

Remark 4. *Theorem 2 demonstrates that the step size η cannot be arbitrarily small to guarantee convergence of RDA. In contrast, the traditional dual ascent method allows for an arbitrarily small step size. This requirement arises here since a small step size may hinder the dual variable from keeping pace with changes in the optimal dual variable which are induced by changes in distributions in the objective function. From a technical perspective, a small step size cannot guarantee that the term $\|\lambda_t - \lambda_s^*\|^2$ will decrease with time. However, when the distribution in the objective function is fixed ($\epsilon = 0$), this concern is eliminated, and the step size can be selected arbitrarily small, as shown in Corollary 2.*

Remark 5. *We note that extending linear constraints to general convex constraints is not straightforward due to two main challenges. On the one hand, even if the original constraint*

function is strongly convex, the presence of distribution shifts can render the decision-dependent constraint non-convex when x and ξ in the constraint function are not separable. On the other hand, the dual function $d_x(\lambda)$ is not necessarily strongly concave and thus the optimal dual variable may not be unique. The non-uniqueness of dual variable, combined with the distribution shift effect, would significantly complicate the analysis of the general constraint case.

IV. Relating Equilibrium and Optimal Points

The results presented so far focus on the convergence to the equilibrium points. However, a natural question to consider is the relation between the constrained equilibrium and optimal points. In this section, we quantify the distance between them. To proceed, we introduce the following additional assumption.

Assumption 5. *$l(x,z)$ is L_z-Lipschitz continuous in z for every x.*

Given Assumption 5, our next result shows that all the equilibrium and optimal points are close to each other. The proof can be found in Appendix D.

Theorem 3. *Suppose Assumptions 1–3 and 5 hold. If $L_{x^*}\epsilon_g < 1$, then for every constrained equilibrium point x_s and every optimal point x_o, we have*

$$\|x_o - x_s\| \leq \frac{2(L_z\epsilon + \epsilon_g\sqrt{d_w}\,\|\lambda_s^*\|)}{\gamma(1 - L_{x^*}\epsilon_g\sqrt{d_w})^2}. \tag{17}$$

Note that Theorem 3 remains valid also for cases when constrained equilibrium points and constrained optimal points are not necessarily unique. Theorem 3 shows that all equilibrium and optimal points are close to each other if the parameters ϵ and ϵ_g are small. Consequently, constrained optimal points can be approximated by constrained equilibrium points which can be derived using RCM and RDA when ϵ and ϵ_g meet the specified conditions.

When $\epsilon_g = 0$, the bound in (17) becomes equivalent to the result in [8]. Therefore, our result subsumes the previous result as a special case. This demonstrates the broader applicability of our framework.

To end this section, we briefly discuss the challenges encountered in the analysis of Theorem 3. When analyzing the distance between the equilibrium and optimal points, a standard technique is to use the strong convexity property, which is also used in our proof. However, this property cannot be applied in a straightforward way here since the constraint sets induced by the distributions at the constrained equilibrium and optimal points are not the same. The presence of inconsistent constraints necessitates us to define a new optimal point, subject to newly constructed constraints that ensure that all these points satisfy the same constraints. By introducing this intermediate point and conducting the sensitivity analysis, we are able to address the challenges arising from the presence of inconsistent constraints. Specifically, we can establish an upper bound on the distance between the equilibrium and optimal points and provide valuable insights into the relation between the equilibrium and optimal solutions for the constrained problem considered here.

V. Numerical Experiments

In this section, we conduct experiments to illustrate the performance of RCM and RDA on market and dynamic pricing problems.

A. Market Problem

We consider a market problem with the cost function $l(x,z) = -x_1z_1 - x_2z_2$. Here, x_i is the price of the i-th good and z_i is the random demand of the i-th good, $i = 1, 2$. We assume that the demand of the first good satisfies $z_1 = \zeta_1(x_1) - a_1x_1$, where the random variable $\zeta_1(x_1)$ follows the uniform distribution $U[\zeta_{L_1}, \zeta_{R_1} + \epsilon x_1]$. For the second good, the random demand is given by $z_2 = \zeta_2 - a_2x_2$, where ζ_2 follows a fixed uniform distribution $U[\zeta_{L_2}, \zeta_{R_2}]$. Additionally, we introduce the cost of production v_i for each good. We assume that the cost of the first good, v_1, follows a uniform distribution $U[\underline{v}_1, 1.2\underline{v}_1 + \epsilon_g x_1]$, where a lower price leads to higher demands and, consequently, lower average cost. The cost of the second good, v_2, follows a uniform distribution $U[\underline{v}_2, 1.2\underline{v}_2]$. The objective is to maximize the gross sales, represented by $x_1z_1 + x_2z_2$. We consider the constraint $Gx \leq \xi(x)$, where $G = [-a_3, -a_4]$, $\xi(x) = \mathbb{E}[-v_1(x_1) - v_2 - e_1]$, which ensures that the selling price is greater than the weighted average cost plus e_1.

The decision-dependent constrained optimization problem is defined as

$$\begin{aligned} \min \quad & -x_1(\mathbb{E}[\zeta_1(x_1)] - a_1x_1) - x_2(\mathbb{E}[\zeta_2] - a_2x_2) \\ \text{s.t.} \quad & a_3x_1 + a_4x_2 \geq e_1 + \mathbb{E}[v_1(x_1)] + \mathbb{E}[v_2]. \end{aligned}$$

The distributions in both the objective function and constraints depend on the decisions. Next, we analyze the sensitivity of the distribution maps. Without loss of generality, we assume $x_1' > x_1$. By virtue of the definition of Wasserstein distance, we have

$$\begin{aligned} & W_1(v_1(x_1), v_1(x_1')) \\ & = 1 - \frac{\eta_{R_1} + \epsilon x_1 - \zeta_{L_1}}{\zeta_{R_1} + \epsilon x_1' - \zeta_{L_1}} + 1 - \frac{\zeta_{R_1} + \epsilon x}{\zeta_{R_1} + \epsilon x'} \\ & \leq \epsilon|x_1 - x_1'|\left(\frac{1}{\zeta_{R_1} + \epsilon x_1' - \zeta_{L_1}} + \frac{1}{\zeta_{R_1} + \epsilon x'}\right) \\ & \leq \epsilon\left(\frac{1}{\zeta_{R_1} - \zeta_{L_1}} + \frac{1}{\zeta_{R_1}}\right)|x_1 - x_1'|. \end{aligned}$$

Based on the last inequality, we select the sensitivity parameter of the distribution map in the objective function as $\bar{\epsilon} = \epsilon(\frac{1}{\zeta_{R_1} - \zeta_{L_1}} + \frac{1}{\zeta_{R_1}})$. Similarly, for the sensitivity parameter of the distribution map of the constraints, we set $\bar{\epsilon}_g = \epsilon_g(\frac{1}{0.2\underline{v}_1} + \frac{1}{1.2\underline{v}_1})$.

We consider the following parameter values for the market problem: $a_1 = 0.8$, $a_2 = 0.2$, $a_3 = 0.6$, $a_4 = 1$, $\zeta_{L_1} = 1$, $\zeta_{R_1} = 5.5$, $\zeta_{L_2} = 0.5$, $\zeta_{R_2} = 2.2$, $e_1 = 1.2$, $\underline{v}_1 = 1.7$, $\underline{v}_2 = 2.5$. All the minimization problems are solved using the CVXOPT toolbox. The simulation results for RCM and RDA are presented in Figs 2–5. In particular, Figs. 2 and 3 show the evolution of the decision variable x for RCM and RDA, respectively. In Figs. 4 and 5, we plot the performance of RCM and RDA for different values of the sensitivity parameters.

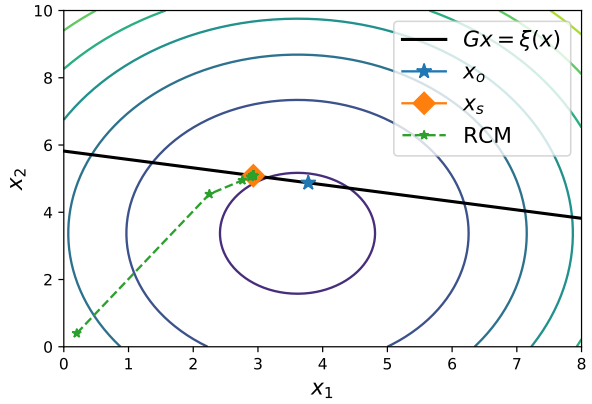

Fig. 2. Trajectory of RCM when $\epsilon = 0.7$, $\epsilon_g = 0.7$ in the market problem. As expected, it converges quickly to the equilibrium point x_s, which is close to the optimal point x_o.

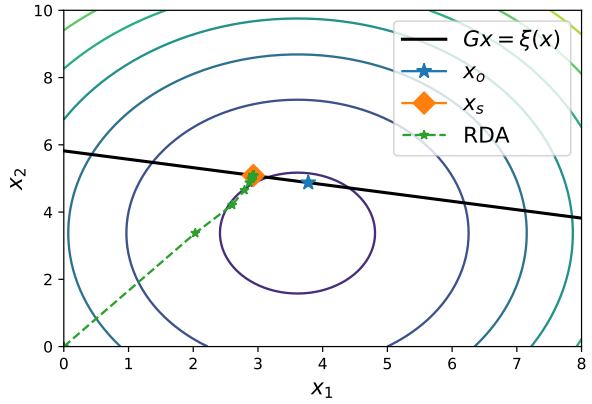

Fig. 3. Trajectory of RDA when $\epsilon = 0.7$, $\epsilon_g = 0.7$ in the market problem. Slightly slower convergence compared to RCM in Fig. 2.

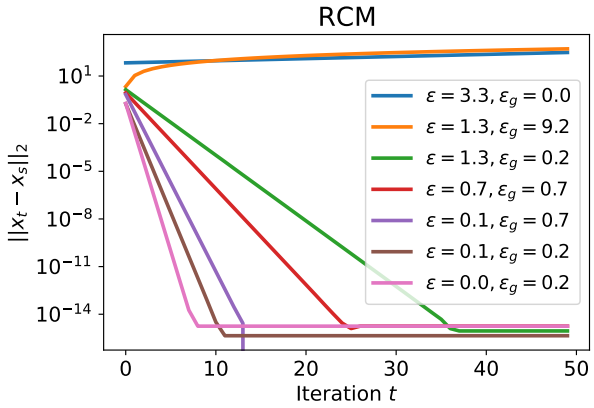

Fig. 4. Convergence of RCM for different ϵ, ϵ_g parameters in the market problem.

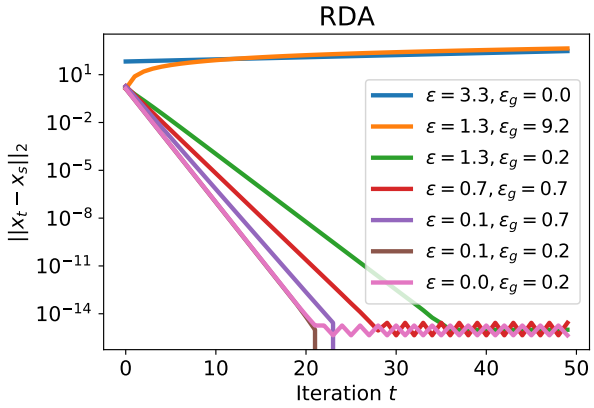

Fig. 5. Convergence of RDA for different ϵ, ϵ_g parameters in the market problem.

We observe that when ϵ or ϵ_g is large, both methods fail to converge to the equilibrium point. When ϵ and ϵ_g are small, both methods converge rapidly within a few iterations. Moreover, when the sensitivity parameter ϵ is not that small, RCM performs similarly to RDA. However, when ϵ is small, RCM outperforms RDA, demonstrating faster convergence.

B. Dynamic Pricing Problem

In this subsection, we present the application of our algorithms to a dynamic pricing experiment using real-world parking data from SFpark in San Francisco [31]. The decision to use a personal vehicle for a trip is heavily influenced by parking availability, parking location, and price. The primary goal of the SFpark pilot project was to make it easy to find a parking space. To this end, SFpark implemented the following adjustments to hourly rates based on occupancy levels: a) When the occupancy is between 80% and 100%, rates are increased by \$0.25; b) When the occupancy is between 60% and 80%, no adjustment is made to the rates; c) When the occupancy is between 30% and 60%, rates are decreased by \$0.25; d) When the occupancy is below 30%, rates are decreased by \$0.50. SFpark aims to maintain occupancy z_1 between 60% and 80% and at the same time maximize the revenue $z_2(x + \bar{x})$, where z_2 represents the total parking time and x represents the difference between the parking price and the nominal price $\bar{x} = 3$. We use a parameter $t > 0$ to quantify the trade-off between these two objectives. The loss function for this dynamic pricing problem is defined as follows:

$$l(x, z_1, z_2) = (z_1 - 0.7)^2 - tz_2(x + \bar{x}) + \frac{v}{2} \|x\|^2 .$$

Here, $v > 0$ is a regularization parameter. When analyzing the individual response to a price adjustment x, we assume that each user updates their behavior using the best response strategy: $z_2' = z_2 - \epsilon x$. Here, ϵ is a positive constant that quantifies the sensitivity of the distribution map. It is easy to verify that the distribution map for z_2 is ϵ-sensitive.

We use the data from the street of Beach ST 600, containing a total of $n = 10224$ samples. Six features are considered, with the total occupied time treated as the strategic feature. The occupancy is computed by dividing the total time by the total occupied time. As in [12], we approximate the distribution of occupancy as $z_1 = \zeta - Ax$, where $A \approx 0.157$ is estimated from data, $\zeta \sim p_0$ is a fixed distribution and p_0 is sampled from data of which the price is at the nominal price.

The constraint is selected as $\mathop{\mathbb{E}}_{z_2 \sim \mathcal{D}(x)} [z_2] \leq c_1 x + c_2$, meaning that the total occupied time should be linearly bounded. We note that also the distribution in the constraint is ϵ-sensitive.

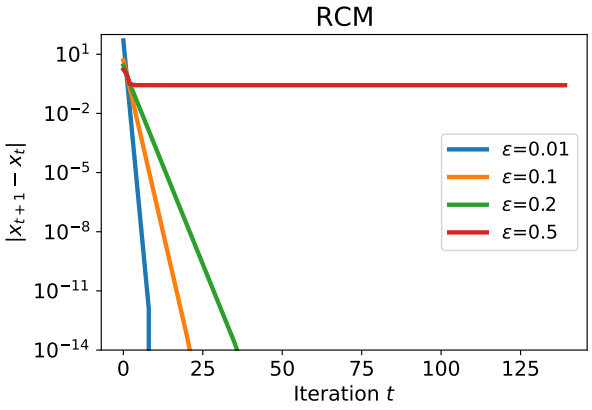

Fig. 6. Convergence of RCM for different ϵ-sensitivity parameters in the dynamic parking pricing problem.

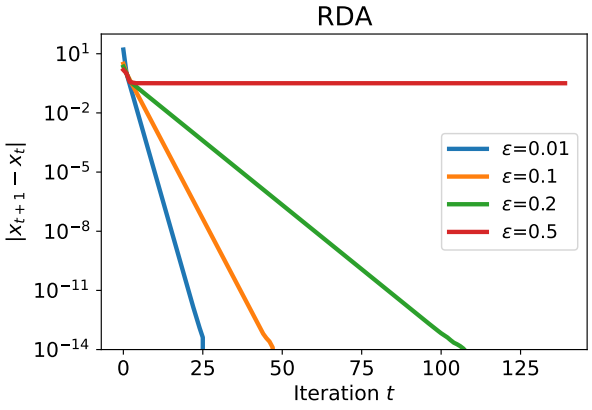

Fig. 7. Convergence of RDA for different ϵ-sensitivity parameters in the dynamic parking pricing problem.

At iteration t, the decision-dependent problem can be written as

$$\min_x \ (\frac{v}{2}+A^2)x^2 - 2A \mathop{\mathbb{E}}_{\zeta\sim p_0}[\zeta]x + 1.4Ax - t \mathop{\mathbb{E}}_{z_2\sim\mathcal{D}(x_t)}[z_2]$$
$$\text{s.t.} \ -c_1 x \le c_2 - \mathop{\mathbb{E}}_{z_2\sim\mathcal{D}(x_t)}[z_2].$$

We select $v=0.03$, $t=0.005$, $c_1=0.5$, $c_2=5$, $\eta=0.18$ and $x_0=-1$. All the minimization problems are solved using the CVXOPT toolbox. The convergence results for RCM and RDA are presented in Figs. 6 and 7, respectively. In Fig. 6, we observe that RCM does indeed converge in only a few iterations for small values of ϵ while it divergences if ϵ is too large. Fig. 7 shows that RDA also converges linearly for small values of ϵ, however, with a slower rate compared to RCM.

VI. Conclusions

In this work, we developed a framework for solving constrained optimization with decision-dependent distributions in both the objective function and linear constraints. Firstly, we established a sufficient condition that guarantees the uniqueness of the constrained equilibrium point and at the same time ensures that the RCM algorithm converges to this point. Furthermore, we presented sufficient conditions for the RDA algorithm to converge to the constrained equilibrium point. Additionally, we derived a bound for the distance between the constrained equilibrium and optimal points. Our results showed that the decision-dependent optimization in [8] can be viewed as a special case of our framework. Finally, we demonstrated the effectiveness of our algorithms using illustrative experiments on both a synthetic market problem and a dynamic pricing problem based on an open-source dataset.

The work presents numerous opportunities for future research. For instance, it would be intriguing to expand the framework to encompass broader constraints. One limitation of our work is that the designed algorithms assume access to expectation computation. In future work, it will be worth considering the use of stochastic optimization algorithms. Moreover, examining the impact of solving a single subproblem per iteration within RDA on convergence holds great interest and warrants further investigation.

Appendix

A. Proof of Theorem 1

To ease notation, for fixed vectors $a, b \in \mathbb{R}^n$, we define

$$M(a,b) = \arg\min_x \mathop{\mathbb{E}}_{z\sim\mathcal{D}(a)}[l(x,z)] \quad \text{s.t.} \quad Gx \le \xi(b).$$

We note that $M(x',x')$ and $M(x'',x')$ are minima of the objective functions $\mathbb{E}_{z\sim\mathcal{D}(x')}[l(x,z)]$ and $\mathbb{E}_{z\sim\mathcal{D}(x'')}[l(x,z)]$, respectively, under the same constrained set. Hence, we can use Lemma 1 to bound the distance of these two minima, which gives

$$\|M(x',x') - M(x'',x')\| \le \frac{\epsilon\beta_z}{\gamma}\|x'-x''\|. \tag{18}$$

The two points $M(x'',x'')$ and $M(x'',x')$ are minima of the same objective function in two different constraint sets. By virtue of Lemma 2, the distance between $M(x'',x'')$ and $M(x'',x')$ can be bounded by

$$\begin{aligned}\|M(x'',x'') - M(x'',x')\| &\le L_{x^*}\|\xi(x') - \xi(x'')\| \\ &\le L_{x^*}\epsilon_g\|x'-x''\|. \end{aligned}\tag{19}$$

Combining (18) and (19), we have

$$\begin{aligned}&\|M(x',x') - M(x'',x'')\| \\ &\le \|M(x',x') - M(x'',x')\| \\ &\quad + \|M(x'',x'') - M(x'',x')\| \\ &\le \left(\frac{\epsilon\beta_z}{\gamma} + L_{x^*}\epsilon_g\right)\|x'-x''\|. \end{aligned}\tag{20}$$

From the update equation (5), we have $x_{t+1} = M(x_t,x_t)$ and $x_s = M(x_s,x_s)$. From (20), we have

$$\begin{aligned}\|x_t - x_s\| &= \|M(x_{t-1},x_{t-1}) - M(x_s,x_s)\| \\ &\le \left(\frac{\epsilon\beta_z}{\gamma} + L_{x^*}\epsilon_g\right)\|x_{t-1}-x_s\| \\ &\le \left(\frac{\epsilon\beta_z}{\gamma} + L_{x^*}\epsilon_g\right)^t\|x_0-x_s\|, \end{aligned}$$

which shows that RCM converges at a linear rate if $\frac{\epsilon\beta_z}{\gamma} + L_{x^*}\epsilon_g < 1$. Given that the mapping $x \to M(x,x)$ is a contraction, the uniqueness of x_s directly follows from the Banach fixed-point theorem. The proof is complete.

B. Proof of Theorem 2

We first note that all the items are positive since $\rho_2(1+\rho_3) < 2\gamma_d$ and $\rho_4 < 1$.

By using the KKT conditions, it is easy to show that

$$\begin{aligned} x_s &= \arg\min_{x\in\mathbb{R}^n} L_{x_s}(x, \lambda_s^*) \\ \lambda_s^* &= [\lambda_s^* + \eta\nabla d_{x_s}(\lambda_s^*)]_+, \end{aligned} \tag{21}$$

where λ_s^* is the optimal multiplier with respect to $d_{x_s}(\lambda)$. We first analyze the dynamics of the dual variable. From the update rule (10), we have

$$\begin{aligned} &\|\lambda_{t+1} - \lambda_s^*\|^2 \\ &= \|[\lambda_t + \eta\nabla d_{x_t}(\lambda_t)]_+ - [\lambda_s^* + \eta\nabla d_{x_s}(\lambda_s^*)]_+\|^2 \\ &\le \|[\lambda_t + \eta\nabla d_{x_t}(\lambda_t)] - [\lambda_s^* + \eta\nabla d_{x_s}(\lambda_s^*)]\|^2 \\ &= \|\lambda_t - \lambda_s^*\|^2 + \eta^2\|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\|^2 \\ &\quad + 2\eta\langle\lambda_t - \lambda_s^*, \nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\rangle. \end{aligned} \tag{22}$$

Since $d_{x'}(\lambda)$ is γ_d-strongly concave for every x', we have

$$\begin{aligned} &\langle\lambda_t - \lambda_s^*, \nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\rangle \\ &= \langle\lambda_t - \lambda_s^*, \nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\rangle \\ &\quad + \langle\lambda_t - \lambda_s^*, \nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\rangle \\ &\le \langle\lambda_t - \lambda_s^*, \nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\rangle - \gamma_d\|\lambda_t - \lambda_s^*\|^2 \\ &\le \|\lambda_t - \lambda_s^*\|\|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\| - \gamma_d\|\lambda_t - \lambda_s^*\|^2. \end{aligned} \tag{23}$$

Moreover, we have

$$\begin{aligned} &\|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\| \\ &= \left\|G\nabla\tilde{f}_{x_t}(-G^{\mathrm{T}}\lambda_t) - \xi(x_t) - G\nabla\tilde{f}_{x_s}(-G^{\mathrm{T}}\lambda_t) + \xi(x_s)\right\| \\ &= \|G\bar{y}_t - G\tilde{y}_t - \xi(x_t) + \xi(x_s)\|, \end{aligned} \tag{24}$$

where $\bar{y}_t = \arg\min_{x\in\mathbb{R}^n}(f_{x_t}(x) + \lambda_t^{\mathrm{T}}Gx - \xi(x_t))$, $\tilde{y}_t = \arg\min_{x\in\mathbb{R}^n}(f_{x_s}(x) + \lambda_t^{\mathrm{T}}Gx - \xi(x_s))$. By virtue of Lemma 1, we have

$$\|\bar{y}_t - \tilde{y}_t\| \le \frac{\epsilon\beta_z}{\gamma}\|x_t - x_s\| = \rho_1\|x_t - x_s\|. \tag{25}$$

Substituting (25) into (24), we obtain

$$\begin{aligned} &\|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\| \\ &\le \left(\|G\|_2\frac{\epsilon\beta_z}{\gamma} + \epsilon_g\right)\|x_t - x_s\| = \rho_2\|x_t - x_s\|. \end{aligned} \tag{26}$$

Combining (26) with (23), we have

$$\begin{aligned} &\langle\lambda_t - \lambda_s^*, \nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\rangle \\ &\le \rho_2\|x_t - x_s\|\|\lambda_t - \lambda_s^*\| - \gamma_d\|\lambda_t - \lambda_s^*\|^2. \end{aligned} \tag{27}$$

Besides, we have

$$\begin{aligned} &\|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\|^2 \\ &= \|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t) + \nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\|^2 \\ &= \|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\|^2 + \|\nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\|^2 \\ &\quad + 2\langle\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t), \nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\rangle \\ &\le \rho_2^2\|x_t - x_s\|^2 + \|\nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\|^2 \\ &\quad + 2\|\nabla d_{x_t}(\lambda_t) - \nabla d_{x_s}(\lambda_t)\|\|\nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\| \\ &\le \rho_2^2\|x_t - x_s\|^2 + L_d^2\|\lambda_t - \lambda_s^*\|^2 \\ &\quad + 2L_d\rho_2\|\lambda_t - \lambda_s^*\|\|x_t - x_s\|, \end{aligned} \tag{28}$$

where the first inequality follows from the Cauchy–Schwarz inequality. The last inequality follows from the smoothness of $d_{x_s}(\lambda)$, which gives $\|\nabla d_{x_s}(\lambda_t) - \nabla d_{x_s}(\lambda_s^*)\| \le L_d\|\lambda_t - \lambda_s^*\|$. From (22), (27), and (28), we have

$$\begin{aligned} &\|\lambda_{t+1} - \lambda_s^*\|^2 \\ &\le \|\lambda_t - \lambda_s^*\|^2 + \eta^2\Big(\rho_2^2\|x_t - x_s\|^2 + L_d^2\|\lambda_t - \lambda_s^*\|^2 \\ &\qquad + 2L_d\rho_2\|\lambda_t - \lambda_s^*\|\|x_t - x_s\|\Big) \\ &\quad + 2\eta\Big(\rho_2\|x_t - x_s\|\|\lambda_t - \lambda_s^*\| - \gamma_d\|\lambda_t - \lambda_s^*\|^2\Big) \\ &= \Big(1 + \eta^2L_d^2 - 2\eta\gamma_d\Big)\|\lambda_t - \lambda_s^*\|^2 + \eta^2\rho_2^2\|x_t - x_s\|^2 \\ &\quad + 2\rho_2(L_d\eta^2 + \eta)\|x_t - x_s\|\|\lambda_t - \lambda_s^*\| \\ &\le \Big(1 + \eta^2L_d^2 - 2\eta\gamma_d + \rho_2\eta + \rho_2L_d\eta^2\Big)\|\lambda_t - \lambda_s^*\|^2 \\ &\quad + \Big(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2\Big)\|x_t - x_s\|^2, \end{aligned} \tag{29}$$

where the last inequality follows from the fact that $2ab \le a^2 + b^2$.

Next we analyze the dynamics of the primal variable. Recall that

$$\begin{aligned} x_t &= \arg\min_x(f_{x_{t-1}}(x) + \lambda_t^{\mathrm{T}}(Gx - \xi(x_{t-1}))) \\ &= \arg\min_x \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_{t-1})}[l(x,z) + \lambda_t^{\mathrm{T}}Gx] \\ x_s &= \arg\min_x(f_{x_s}(x) + \lambda_s^{*\mathrm{T}}(Gx - \xi(x_s))) \\ &= \arg\min_x \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z) + \lambda_s^{*\mathrm{T}}Gx], \end{aligned}$$

and denote

$$\begin{aligned} y_t :&= \arg\min_x(f_{x_s}(x) + \lambda_t^{\mathrm{T}}(Gx - \xi(x_s))) \\ &= \arg\min_x \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z) + \lambda_t^{\mathrm{T}}Gx] \end{aligned}$$

It can be verified that the function $l(x,z) + \lambda_t^{\mathrm{T}}Gx$ is γ-strongly convex in x and its gradient $\nabla_x l(x,z) + G^{\mathrm{T}}\lambda_t$ is β-Lipschitz continuous in z. By virtue of Lemma 1, the distance between x_t and y_t can be bounded by

$$\|x_t - y_t\| \le \frac{\epsilon\beta_z}{\gamma}\|x_{t-1} - x_s\| = \rho_1\|x_{t-1} - x_s\|. \tag{30}$$

Now we bound $\|y_t - x_s\|$. Since $f_{x_s}(x)$ is strongly convex in x, we have

$$\begin{aligned} &\gamma\|x_s - y_t\|^2 \le \langle x_s - y_t, \nabla f_{x_s}(x_s) - \nabla f_{x_s}(y_t)\rangle \\ &\le \|x_s - y_t\|\|\nabla f_{x_s}(x_s) - \nabla f_{x_s}(y_t)\|. \end{aligned} \tag{31}$$

Since $f_{x_s}(x) + \lambda_t^{\mathrm{T}}Gx$ is also strongly convex in x, y_t is unique and satisfies $\nabla f_{x_s}(y_t) + G^{\mathrm{T}}\lambda_t = 0$. Similarly, $f_{x_s}(x) + \lambda_s^{*\mathrm{T}}Gx$ is strongly convex in x and thus $\nabla f_{x_s}(x_s) + G^{\mathrm{T}}\lambda_s^* = 0$. Then, (31) yields

$$\|x_s - y_t\| \le \frac{1}{\gamma}\|\nabla f_{x_s}(x_s) - \nabla f_{x_s}(y_t)\|$$

$$= \frac{1}{\gamma}\left\|-G^{\mathrm{T}}\lambda_s^* + G^{\mathrm{T}}\lambda_t\right\| \leq \frac{\|G\|_2}{\gamma}\|\lambda_t - \lambda_s^*\|. \quad (32)$$

From (30) and (32), we have

$$\begin{aligned}
&\|x_t - x_s\|^2 = \|x_t - y_t + y_t - x_s\|^2 \\
&= \|x_t - y_t\|^2 + \|y_t - x_s\|^2 + 2\langle x_t - y_t, y_t - x_s\rangle \\
&\leq \rho_1^2\|x_{t-1} - x_s\|^2 + \frac{\|G\|_2^2}{\gamma}\|\lambda_t - \lambda_s^*\|^2 \\
&\quad + 2\rho\frac{\|G\|_2}{\gamma}\|x_{t-1} - x_s\|\|\lambda_t - \lambda_s^*\| \\
&\leq \left(\rho_1^2 + \frac{\rho_1\|G\|_2}{\gamma}\right)\|x_{t-1} - x_s\|^2 \\
&\quad + \left(\frac{\|G\|_2^2}{\gamma^2} + \frac{\rho_1\|G\|_2}{\gamma}\right)\|\lambda_t - \lambda_s^*\|^2 \\
&= \rho_4\|x_{t-1} - x_s\|^2 + \rho_3\|\lambda_t - \lambda_s^*\|^2. \quad (33)
\end{aligned}$$

Combining (29) and (33), we have

$$\begin{aligned}
&\|\lambda_{t+1} - \lambda_s^*\|^2 + \alpha\|x_t - x_s\|^2 \\
&\leq \left(1 + \eta^2L_d^2 - 2\eta\gamma_d + \rho_2\eta + \rho_2L_d\eta^2\right)\|\lambda_t - \lambda_s^*\|^2 \\
&\quad + \left(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2 + \alpha\right)\|x_t - x_s\|^2 \\
&\leq \left(1 + \eta^2L_d^2 - 2\eta\gamma_d + \rho_2\eta + \rho_2L_d\eta^2\right)\|\lambda_t - \lambda_s^*\|^2 \\
&\quad + \left(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2 + \alpha\right) \quad (34) \\
&\quad \times \left(\rho_4\|x_{t-1} - x_s\|^2 + \rho_3\|\lambda_t - \lambda_s^*\|^2\right) \\
&= \Big(1 + \eta^2L_d^2 - 2\eta\gamma_d + \rho_2\eta + \rho_2L_d\eta^2 \\
&\quad + \rho_3\Big(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2 + \alpha\Big)\Big)\|\lambda_t - \lambda_s^*\|^2 \\
&\quad + \frac{\rho_4}{\alpha}\Big(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2 + \alpha\Big)\alpha\|x_{t-1} - x_s\|^2 \\
&= (1 + H_\lambda(\eta))\|\lambda_t - \lambda_s^*\|^2 \\
&\quad + (1 + H_x'(\eta))\alpha\|x_{t-1} - x_s\|^2, \quad (35)
\end{aligned}$$

where we define $H_\lambda(\eta) := \eta^2L_d^2 - 2\eta\gamma_d + \rho_2\eta + \rho_2L_d\eta^2 + \rho_3\Big(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2 + \alpha\Big) = a_1\eta^2 - a_2\eta + a_3$ and $H_x'(\eta) = \frac{\rho_4}{\alpha}\Big(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2 + \alpha\Big) - 1 = \frac{1}{\alpha}\Big(b_1\eta^2 + b_2\eta - b_3\Big) := \frac{1}{\alpha}H_x(\eta)$. It suffices to show that the choice of $\eta \in (s_1, s_2)$ assures $H_\lambda(\eta) < 0$ and $H_x(\eta) < 0$.

Since $a_2^2 > 4a_1a_2$, there exist two distinct solutions to $H_\lambda(\eta) = 0$, which we denote by s_1, s_1' with $s_1 < s_1'$. When $\eta \in (s_1, s_1')$, we have $H_\lambda(\eta) < 0$. Similarly, we denote by s_2, s_2' the solutions to $H_x(\eta) < 0$ with $s_2 > s_2'$. When $\eta \in (s_2', s_2)$, we have $H_\lambda(x) < 0$. When η takes values on the intersection of these two regions, we can make sure that $H_\lambda(\eta) < 0$ and $H_x(\eta) < 0$ simultaneously. Obviously, this happens if $s_1 < s_2$. In what follows, we show that the choice of α ensures $s_1 < s_2$. Specifically, the condition $s_1 < s_2$ is equivalent to

$$\begin{aligned}
&\frac{a_2 - \sqrt{a_2^2 - 4a_1a_3}}{2a_1} < \frac{-b_2 + \sqrt{b_2^2 + 4b_1b_3}}{2b_1} \\
\Leftrightarrow\quad & a_2b_1 + a_1b_2 < b_1\sqrt{a_2^2 - 4a_1a_3} + a_1\sqrt{b_2^2 + 4b_1b_3} \\
\Leftrightarrow\quad & \rho_2\rho_4(L_d^2 + L_d\rho_2 + \rho_2^2\rho_3 + L_d\rho_2\rho_3) \\
& + \rho_4(\rho_2^2 + L_d\rho_2)(2\gamma_d - \rho_2(1 + \rho_3)) \\
& < b_1\sqrt{a_2^2 - 4a_1a_3} + a_1\sqrt{b_2^2 + 4b_1b_3} \\
\Leftrightarrow\quad & \rho_2\rho_4(2\gamma_d\rho_2 + 2\gamma_dL_d + L_d^2 - \rho_2^2) \\
& < b_1\sqrt{a_2^2 - 4a_1a_3} + a_1\sqrt{b_2^2 + 4b_1b_3}. \quad (36)
\end{aligned}$$

Since $\sqrt{b_2^2 + 4b_1b_3} > b_2$, a sufficient condition for (36) is

$$\begin{aligned}
&\rho_2\rho_4(2\gamma_d\rho_2 + 2\gamma_dL_d + L_d^2 - \rho_2^2) \\
&< b_1\sqrt{a_2^2 - 4a_1a_3} + a_1b_2 \\
\Leftrightarrow\quad & \rho_2\rho_4(2\gamma_d\rho_2 + 2\gamma_dL_d + L_d^2 - \rho_2^2) \\
& < b_1\sqrt{a_2^2 - 4a_1a_3} + \rho_2\rho_4(L_d^2 + L_d\rho_2 + \rho_2^2\rho_3 \\
& \quad + L_d\rho_2\rho_3) \\
\Leftrightarrow\quad & 2\gamma_d\rho_2 + 2\gamma_dL_d < (\rho_2 + L_d)\sqrt{a_2^2 - 4a_1a_3} \\
& + L_d^2 + L_d\rho_2 + \rho_2^2\rho_3 + L_d\rho_2^2\rho_3 + \rho_2^2. \quad (37)
\end{aligned}$$

Since $\alpha = (1 - \bar{\alpha})\frac{a_2^2}{4\rho_3a_1}$, the condition (37) is equivalent to

$$\begin{aligned}
&2\gamma_d\rho_2 + 2\gamma_dL_d < \sqrt{\bar{\alpha}}(\rho_2 + L_d)(2\gamma_d - \rho_2(1 + \rho_3)) \\
&+ L_d^2 + L_d\rho_2 + \rho_2^2\rho_3 + L_d\rho_2^2\rho_3 + \rho_2^2 \\
\Leftrightarrow\quad & 2\gamma_d\rho_2 + 2\gamma_dL_d + \sqrt{\bar{\alpha}}\rho_2(1 + \rho_3)(\rho_2 + L_d) \\
& < 2\sqrt{\bar{\alpha}}\gamma_d(\rho_2 + L_d) + L_d^2 + L_d\rho_2 + \rho_2^2\rho_3 \\
& + L_d\rho_2^2\rho_3 + \rho_2^2. \quad (38)
\end{aligned}$$

Since $\sqrt{\bar{\alpha}} \in [0, 1)$ and $\gamma_d < L_d$, we further simplify the sufficient condition (38) to

$$\begin{aligned}
&2\gamma_d\rho_2 + 2\gamma_dL_d < \sqrt{\bar{\alpha}}(2\gamma_d\rho_2 + 2\gamma_dL_d) + L_d^2 \\
\Leftrightarrow\quad & (1 - \sqrt{\bar{\alpha}})(2\gamma_d\rho_2 + 2\gamma_dL_d) < L_d^2 \\
\Leftrightarrow\quad & \sqrt{\bar{\alpha}} > 1 - \frac{L_d^2}{2\gamma_d(\rho_2 + L_d)}, \quad (39)
\end{aligned}$$

which holds since $\sqrt{\bar{\alpha}} = \max\left\{0, 1 - \frac{L_d^2}{4\gamma_d(\rho_2 + L_d)}\right\}$. Therefore, when $\eta \in (s_1, s_2)$, we have $H_x(\lambda) := \kappa_1 < 1$ and $H_x(\eta) := \kappa_2 < 1$. For $\kappa = \max\{\kappa_1, \kappa_2\}$, from (34), we have $\|\lambda_{t+1} - \lambda_s^*\|^2 + \alpha\|x_t - x_s\|^2 \leq \kappa(\|\lambda_t - \lambda_s^*\|^2 + \alpha\|x_{t-1} - x_s\|^2)$. Iteratively using this inequality completes the proof.

C. Proof of Corollary 2

Following the steps as in the proof of Theorem 2, we have that

$$\begin{aligned}
&\|\lambda_{t+1} - \lambda_s^*\|^2 \\
&\leq \left(1 + \eta^2L_d^2 - 2\eta\gamma_d + \rho_2\eta + \rho_2L_d\eta^2\right)\|\lambda_t - \lambda_s^*\|^2 \\
&\quad + \left(\eta^2\rho_2^2 + \rho_2\eta + \rho_2L_d\eta^2\right)\|x_t - x_s\|^2, \quad (40)
\end{aligned}$$

and

$$\|x_{t+1} - x_s\|^2 \leq \frac{\|G\|_2^2}{\gamma^2}\|\lambda_t - \lambda_s^*\|^2. \quad (41)$$

Substituting (41) into (40), we have

$$\begin{aligned}
&\|\lambda_{t+1}-\lambda_s^*\|^2\\
&\le \Big(1-2\eta\gamma_d+\eta^2L_d^2+\epsilon_g\eta+\epsilon_gL_d\eta^2\\
&\quad+\frac{\|G\|_2^2}{\gamma^2}(\epsilon_g\eta^2+\epsilon_g\eta+\epsilon_gL_d\eta^2)\Big)\|\lambda_t-\lambda_s^*\|^2\\
&=\Big(1-\eta(2\gamma_d-\epsilon_g-\epsilon_g\frac{\|G\|_2^2}{\gamma^2})+\eta^2(L_d^2+\epsilon_gL_d\\
&\quad+\epsilon_g\frac{\|G\|_2^2}{\gamma^2}+\epsilon_gL_d\frac{\|G\|_2^2}{\gamma^2})\Big)\|\lambda_t-\lambda_s^*\|^2\\
&=\kappa_3\|\lambda_t-\lambda_s^*\|^2. \qquad (42)
\end{aligned}$$

Since $\eta<\frac{2\gamma_d-\epsilon_g(1+\frac{\|G\|_2^2}{\gamma^2})}{L_d^2+\epsilon_gL_d+\epsilon_g^2\frac{\|G\|_2^2}{\gamma^2}+\epsilon_gL_d\frac{\|G\|_2^2}{\gamma^2}}$, we have $\kappa_3<1$. Iteratively using (42), we can obtain (15). Substituting (41) into (15), we obtain (16). The proof is complete.

D. Proof of Theorem 3

Recall the definition $f_{x'}(x)=\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')}[l(x,z)]$. By the definition of x_o and x_s, we have $Gx_o\le\xi(x_o)$ and $Gx_s\le\xi(x_s)$. Since both x_o and x_s are feasible points of the problem (1), we have $f_{x_o}(x_o)\le f_{x_s}(x_s)$. We note that the inequality $f_{x_s}(x_s)\le f_{x_s}(x_o)$ does not necessarily hold since the constraints $Gx_o\le\xi(x_s)$ do not necessarily hold. In fact, we can only guarantee that

$$\begin{aligned}
G_ix_o&\le\xi_i(x_o)-\xi_i(x_s)+\xi_i(x_s)\\
&\le\xi_i(x_s)+\epsilon_g\|x_o-x_s\|, \qquad (43)
\end{aligned}$$

for all $i\in\{1,\dots,d_w\}$, where $\xi_i(x_s)$ denotes the i-th element of the vector $\xi(x_s)$.

To bound $\|x_o-x_s\|$, we define the point

$$\begin{aligned}
\hat{x}_s=&\arg\min_x\Big\{\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)]\Big|\\
&G_ix\le\xi_i(x_s)+\epsilon_g\|x_o-x_s\|, i\in\{1,\dots,d_w\}\Big\}.
\end{aligned}$$

Since both $\hat{x}_s$ and x_o satisfy the constraint (43), for all $i\in\{1,\dots,d_w\}$, the strongly convex property of $f_{x'}(x)$ yields

$$f_{x_s}(x_o)-f_{x_s}(\hat{x}_s)\ge\frac{\gamma}{2}\|x_o-\hat{x}_s\|^2. \qquad (44)$$

Now we bound $|f_{x_s}(x_s)-f_{x_s}(\hat{x}_s)|$. Denote by λ_s^* the optimal multiplier of the problem $\max_{\lambda\ge0}g_{x_s}(\lambda)$, where

$$g_{x_s}(\lambda)=\min_x\Big\{\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)]+\lambda^{\mathrm{T}}(Gx-\xi(x_s))\Big\}.$$

Consider the convex minimization problem:

$$\begin{aligned}
\min_x\ &\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)]\\
\text{s.t.}\ \ &G_ix\le\xi_i(x_s)+u,\ i=1,\dots,d_w,
\end{aligned}$$

where u is a scalar. Define $p^*(u)=\inf_x\Big\{\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)]\Big|G_ix\le\xi_i(x_s)+u, i\in\{1,\dots,d_w\}\Big\}$. Set $u_0=\epsilon_g\|x_o-x_s\|$. For any feasible point x that satisfies the constraint $G_ix\le\xi(x_s)+u_0$, $i=1,\dots,d_w$, we have

$$\begin{aligned}
f_{x_s}(x_s)&=p^*(0)=g_{x_s}(\lambda_s^*)\\
&\le\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)]+\lambda_s^{*\mathrm{T}}(Gx-\xi(x_s))\\
&\le\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(x,z)]+\lambda_s^{*\mathrm{T}}\mathbf{1}u_0,
\end{aligned}$$

where the equality holds due to the strong duality, the first inequality follows from the definition of g_{x_s}, and the second inequality follows from the fact that $\lambda_s^*\ge0$. Here, the notation $\mathbf{1}$ denotes the vector with suitable dimension of which all the elements are 1. Since $\hat{x}_s$ is a feasible point with respect to the constraint $G_ix\le\xi_i(x_s)+u_0$, $i=1,\dots,d_w$, we further obtain

$$\begin{aligned}
f_{x_s}(x_s)&\le\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x_s)}[l(\hat{x}_s,z)]+\lambda_s^{*\mathrm{T}}\mathbf{1}u_0\\
&\le f_{x_s}(\hat{x}_s)+\lambda_s^{*\mathrm{T}}\mathbf{1}u_0\\
&\le f_{x_s}(\hat{x}_s)+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|\|x_o-x_s\|. \qquad (45)
\end{aligned}$$

From (44) and (45), we have

$$\begin{aligned}
f_{x_o}(x_o)&\le f_{x_s}(x_s)\\
&\le f_{x_s}(\hat{x}_s)+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|\|x_o-x_s\|\\
&\le f_{x_s}(x_o)-\frac{\gamma}{2}\|x_o-\hat{x}_s\|^2\\
&\quad+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|\|x_o-x_s\|. \qquad (46)
\end{aligned}$$

Since $l(x,z)$ is L_z-Lipschitz continuous in z, we have

$$f_{x_s}(x_o)\le f_{x_o}(x_o)+L_z\epsilon\|x_o-x_s\|. \qquad (47)$$

From (46) and (47), we have

$$\begin{aligned}
\frac{\gamma}{2}\|x_o-\hat{x}_s\|^2&\le L_z\epsilon\|x_o-x_s\|\\
&\quad+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|\|x_o-x_s\|. \qquad (48)
\end{aligned}$$

By virtue of Lemma 2, we have

$$\|x_s-\hat{x}_s\|\le L_{x^*}\epsilon_g\sqrt{d_w}\|x_o-x_s\|. \qquad (49)$$

Combining (48) with (49), we have

$$\begin{aligned}
&\gamma\|x_o-x_s\|^2=\gamma\|x_o-\hat{x}_s+\hat{x}_s-x_s\|^2\\
&\le\gamma(1+\frac{1}{c})\|x_o-\hat{x}_s\|^2+\gamma(1+c)\|\hat{x}_s-x_s\|^2\\
&\le2(1+\frac{1}{c})(L_z\epsilon+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|)\|x_o-x_s\|\\
&\quad+\gamma(1+c)L_{x^*}^2\epsilon_g^2d_w\|x_o-x_s\|^2, \qquad (50)
\end{aligned}$$

where the first inequality holds, since $2ab\le ca^2+\frac{1}{c}b^2$ for any $c>0$. Upon performing a transformation, (50) can be expressed as:

$$\|x_o-x_s\|\le\frac{2(1+\frac{1}{c})(L_z\epsilon+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|}{\gamma(1-(1+c)L_{x^*}^2\epsilon_g^2d_w)}.$$

By setting $c=-1+\frac{1}{L_{x^*}\epsilon_g\sqrt{d_w}}$, we get

$$\|x_o-x_s\|\le\frac{2(L_z\epsilon+\epsilon_g\sqrt{d_w}\|\lambda_s^*\|)}{\gamma(1-L_{x^*}\epsilon_g\sqrt{d_w})^2},$$

which completes the proof.

REFERENCES

[1] Gregory Pierce and Donald Shoup. Sfpark: Pricing parking by demand. In *Parking and the City*, pages 344–353. Routledge, 2018.

[2] Chase P Dowling, Lillian J Ratliff, and Baosen Zhang. Modeling curbside parking as a network of finite capacity queues. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1011–1022, 2019.

[3] John J Horton. Online labor markets. In *International Workshop on Internet and Network Economics*, pages 515–522, 2010.

[4] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[5] Siddhartha Banerjee, Carlos Riquelme, and Ramesh Johari. Pricing in ride-share platforms: A queueing-theoretic approach. In *ACM Conference on Economics and Computation*, 2015.

[6] Gianluca Bianchin, Miguel Vaquero, Jorge Cortes, and Emiliano Dall'Anese. Online stochastic optimization for unknown linear systems: Data-driven controller synthesis and analysis. *IEEE Transactions on Automatic Control*, 69(7):4411–4426, 2023.

[7] Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian Ratliff. Learning in stochastic monotone games with decision-dependent data. In *International Conference on Artificial Intelligence and Statistics*, pages 5891–5912. PMLR, 2022.

[8] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[9] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939, 2020.

[10] John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

[11] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

[12] Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *AAAI Conference on Artificial Intelligence*, pages 8081–8088, 2022.

[13] Songtao Lu. Bilevel optimization with coupled decision-dependent distributions. In *International Conference on Machine Learning*, 2023.

[14] Yulai Zhao. Optimizing the performative risk under weak convexity assumptions. *arXiv preprint arXiv:2209.00771*, 2022.

[15] Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 38449–38460, 2022.

[16] Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022.

[17] Killian Reed Wood and Emiliano Dall'Anese. Online saddle point tracking with decision-dependent data. In *Learning for Dynamics and Control Conference*, pages 1416–1428. PMLR, 2023.

[18] Killian Wood and Emiliano Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *SIAM Journal on Optimization*, 33(3):1943–1967, 2023.

[19] Xiaolu Wang, Chung-Yiu Yau, and Hoi To Wai. Network effects in performative prediction games. In *International Conference on Machine Learning*, 2023.

[20] Irina Subotić, Adrian Hauswirth, and Florian Dörfler. Quantitative sensitivity bounds for nonlinear programming and time-varying optimization. *IEEE Transactions on Automatic Control*, 67(6):2829–2842, 2021.

[21] Xuanyu Cao and KJ Ray Liu. Online convex optimization with time-varying constraints and bandit feedback. *IEEE Transactions on Automatic Control*, 64(7):2665–2680, 2018.

[22] Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl Henrik Johansson. Distributed bandit online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Automatic Control*, 66(10):4620–4635, 2020.

[23] Mahyar Fazlyab, Santiago Paternain, Victor M Preciado, and Alejandro Ribeiro. Prediction-correction interior-point method for time-varying convex optimization. *IEEE Transactions on Automatic Control*, 63(7):1973–1986, 2017.

[24] Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6:1646–1651, Oct, 2021.

[25] Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

[26] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 2013.

[27] Xuyang Wu, Sindri Magnússon, and Mikael Johansson. A new family of feasible methods for distributed resource allocation. In *2021 60th IEEE Conference on Decision and Control*, pages 3355–3360. IEEE, 2021.

[28] Guannan Qu and Na Li. On the exponential stability of primal-dual gradient dynamics. *IEEE Control Systems Letters*, 3(1):43–48, 2018.

[29] Vincent Guigues. On the strong concavity of the dual function of an optimization problem. *arXiv preprint arXiv:2006.16781*, 2020.

[30] Zhi-Quan Luo and Paul Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Mathematics of Operations Research*, 18(4):846–867, 1993.

[31] Sfpark parking sensor data hourly occupancy 2011 – 2013. *https://www.sfmta.com/getting-around/drive-park/demand-responsive-pricing/sfpark-evaluation*, 2013.

Zifan Wang (Student Member, IEEE) received the B.S. and M.S. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2019 and 2021, respectively. He is currently working toward the Ph.D. degree in electrical engineering with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden.

His current research interests lie in online optimization, risk-averse learning, game theory and control theory.

Changxin Liu (Member, IEEE) received his Ph.D. in mechanical engineering from the University of Victoria, Victoria, BC, Canada, in 2021. From 2021 to 2024, he worked as a postdoctoral researcher at the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. In October 2024, he joined the Department of Automation at East China University of Science and Technology, Shanghai, China, where he is currently a Professor. His research interests include distributed optimization and control, machine learning, and combinatorial optimization, with applications to the process industry.

Thomas Parisini (Fellow, IEEE) received the Ph.D. degree in electronic engineering and computer science from the University of Genoa, Italy, in 1993. He was an Associate Professor with Politecnico di Milano, Milano, Italy. He currently holds the Chair of industrial control and is the Head of the Control and Power Research Group, Imperial College London, London, U.K. He also holds a Distinguished Professorship at Aalborg University, Denmark. Since 2001, he has been the Danieli Endowed Chair of automation engineering with the University of Trieste, Trieste, Italy, where from 2009 to 2012, he was the Deputy Rector. In 2023, he held a "Scholar-in-Residence"visiting position with Digital Futures-KTH, Stockholm, Sweden. He has authored or coauthored a research monograph in the Communication and Control Series, Springer Nature, and more than 400 research papers in archival journals, book chapters, and international conference proceedings. Dr. Parisini was the recipient of the Knighthood of the Order of Merit of the Italian Republic for scientific achievements abroad awarded by the Italian President of the Republic in 2023. In 2018 he received the Honorary Doctorate from the University of Aalborg, Denmark and in 2024, the IEEE CSS Transition to Practice Award. Moreover, he was awarded the 2007 IEEE Distinguished Member Award, and was co-recipient of the IFAC Best Application Paper Prize of the Journal of Process Control, Elsevier, for the three-year period 2011-2013 and of the 2004 Outstanding Paper Award of IEEE TRANSACTIONS ON NEURAL NETWORKS. In 2016, he was awarded as Principal Investigator with Imperial of the H2020 European Union flagship Teaming Project KIOS Research and Innovation Centre of Excellence led by the University of Cyprus with an overall budget of over 40 million Euros. He was the 2021-2022 President of the IEEE Control Systems Society and he was the Editor-in-Chief of IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY (2009-2016). He was the Chair of the IEEE CSS Conference Editorial Board (2013-2019). Also, he was the associate editor of several journals including the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and the IEEE TRANSACTIONS ON NEURAL NETWORKS. He is currently an Editor of Automatica and the Editor-in-Chief of the European Journal of Control. He was the Program Chair of the 2008 IEEE Conference on Decision and Control and General Co-Chair of the 2013 IEEE Conference on Decision and Control. He is a Fellow of IFAC. He is a Member of IEEE TAB Periodicals Review and Advisory Committee.

Michael M. Zavlanos (Senior Member, IEEE) received the Diploma in mechanical engineering from the National Technical University of Athens (NTUA), Greece, in 2002, and the M.S.E. and Ph.D. degrees in electrical and systems engineering from the University of Pennsylvania in 2005 and 2008, respectively. He is currently a Professor in the Department of Mechanical Engineering and Materials Science at Duke University. He also holds a secondary appointment in the Department of Electrical and Computer Engineering and the Department of Computer Science. He currently serves as the Director of the Healthcare Systems Optimization program with Duke AI Health and is also an Amazon Scholar with Amazon Robotics. His research focuses on control theory, optimization, machine learning, and AI with applications in robotics and autonomous systems, cyber-physical systems, and healthcare/medicine.

Dr. Zavlanos is a recipient of various awards including the 2014 Office of Naval Research Young Investigator Program (YIP) Award and the 2011 National Science Foundation Faculty Early Career Development (CAREER) Award.

Karl Henrik Johansson (Fellow, IEEE) is Swedish Research Council Distinguished Professor in Electrical Engineering and Computer Science at KTH Royal Institute of Technology in Sweden and Founding Director of Digital Futures. He earned his MSc degree in Electrical Engineering in 1992 and PhD in Automatic Control in 1997, both from Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU and other prestigious institutions. His research interests focus on networked control systems and cyber-physical systems with applications in transportation, energy, and automation networks. For his scientific contributions, he has received numerous best paper awards and various other distinctions from IEEE, IFAC, and other organizations. He has been awarded Distinguished Professor by the Swedish Research Council, Wallenberg Scholar by the Knut and Alice Wallenberg Foundation, Future Research Leader by the Swedish Foundation for Strategic Research. He has also received the triennial IFAC Young Author Prize, IEEE CSS Distinguished Lecturer, IFAC Outstanding Service Award, and IEEE CSS Hendrik W. Bode Lecture Prize. His extensive service to the academic community includes being President of the European Control Association, IEEE CSS Vice President Diversity, Outreach & Development, and Member of IEEE CSS Board of Governors and IFAC Council. He has served on the editorial boards of Automatica, IEEE TAC, IEEE TCNS and many other journals. He has also been a member of the Swedish Scientific Council for Natural Sciences and Engineering Sciences. He is Fellow of both the IEEE and the Royal Swedish Academy of Engineering Sciences.