Advances in Stochastic Optimization with Decision-Dependent Distributions

by

Killian R. Wood

B.A., California State University Fullerton, 2019

M.S., University of Colorado Boulder, 2022

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2024

Committee Members:

Emiliano Dall'Anese, Chair

Stephen Becker

William Kleiber

Francois Meyer

Rafael Frongillo

Wood, Killian R. (Ph.D., Applied Mathematics)

Advances in Stochastic Optimization with Decision-Dependent Distributions

Thesis directed by Prof. Emiliano Dall'Anese

The success of stochastic optimization hinges on the assumption that the distribution of the data remains stationary both throughout the run of an optimization algorithm and after deployment of a solution. However, in applications where data acquisition requires feedback from humans with vested interests in optimization outcomes, this assumption often fails as humans tend to modify their attributes to achieve desired results, leading to a changing data distribution. To capture this optimization induced distributional shift, we pose the formulation of stochastic optimization problems in which the data distribution depends explicitly on optimization variables.

We characterize two distinct types of solutions that arise: optimal points that are universally best but require significant investment to find, and stable points that can be found during " standard operation" but are only optimal for the behaviors they induce. This work provides convergence guarantees for stochastic gradient algorithms that find stable points using only feedback from the system. We demonstrate online tracking for a time-varying extension in expectation, and high probability. We show that stochastic saddle point problems with decision-dependence can be solved using derivative free methods, and the resulting stable point problem can be solved using stochastic primal-dual. Furthermore, we extend this framework to continuous games, demonstrating that a approximate Nash equilibrium can be achieved when players are capable of learning a parameterized model of their distribution.

Acknowledgements

I would like express my immense gratitude to my advisor, Emiliano Dall'Anese, for his guidance throughout this incredible journey. To Stephen Becker, for sparking my interest in optimization and always taking the time to share his wisdom. To my friends for braving this path alongside me. And to my partner Ellie, whose support and insight fueled my perseverance.

Contents

Figures

Figure

Chapter 1

Introduction

Stochastic optimization plays a central role in computing, statistical science, and engineering systems in which the goal is to find an optimal decision from a limited dataset that generalizes well to the unseen data [10, 50]. In its simplest form, these problems typically appear as the optimization problem

$$x^* \in \arg\min_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z \sim D}[f(x, z)] \tag{1.1}$$

where $f : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ is such that $x \mapsto f(x, z)$ is smooth and convex for all $z \in \mathbb{R}^k$, and $\mathcal{X} \subseteq \mathbb{R}^d$ is convex. Solving problems of this form amounts to collecting $m \in \mathbb{N}$ samples $\{z_i\}_{i=1}^m \overset{i.i.d.}{\sim} D$, either prior to optimization or throughout the run of an optimization algorithm, thus giving rise to the empirical risk minimization problem

$$x_m^* \in \arg\min_{x \in \mathcal{X}} \frac{1}{m} \sum_{i=1}^m f(x, z_i) \tag{1.2}$$

The result is a solution x_m^* that will approximate x^* for a sufficient number of samples.

In practice, this decision x_m^* will be deployed into the system or population from which data has been gathered and will retain its optimality guarantees insofar as the assumption that $\{z_i\}_{i=1}^m \overset{i.i.d.}{\sim} D$ still holds. In many applications however, the distribution of data does not remain stationary after deployment of decision variables. This phenomena is typically referred to as *distributional shift* within the optimization literature and its study is primarily focused on two distinct sources of this occurrence. The first of which is temporal, wherein data from the systems evolves according to a time series even though the cost function itself may be time invariant. While this setting is not the focus of this work, we will draw on the analysis of this case for inspiration later. The second, and the primary focus of this work, is a change in the distribution due to the deployment of the optimization variables themselves. In these systems, the data is in some way dependent on the decision with which it is used to make and deploying a new decision will cause the distribution to shift shortly after deployment.

1.1 Motivating Examples

Though not exhaustive, this phenomena can be observed in learning tasks and engineering systems in which the objective used to make optimal decisions for a population of humans in the loop.

Classification. Gaming is a common behavior observed in response to classification tasks whereby a strategic or adversarial population respond to the deployment of a classifier by modifying their attributes to receive a desired outcome from the classifier [12, 19, 31]. Here, the decision variable x parameterizes a latent function h so that $h_x(a) \approx b$, where $z = (z, b)$ and a are population features and $b \in \{0, 1\}$ are labels. The optimization problem then uses training data to find optimal parameters x via some appropriate loss $f(x, z)$, and test data drawn from the same distribution to measure generalize ability to unseen data. In this instance, the assumption that the training and testing data is drawn from the same distribution is violated as the distribution changes before test data is drawn.

Markets. Dynamic price models are a common feature used in markets, where they serve as a mechanism by which a firm or service provider set prices x to incentive users to shift their demand z. Ride-hailing services use dynamic pricing to both incentives users to request rides after querying the app, and incentive drivers to accept ride requests by minimizing operational cost or the estimated time of arrival of drivers to customers [8, 24, 32, 66]. Energy markets use dynamic pricing to incentives users to disperse demand and avoid spikes, or use services when demand can be met by maximizing utility [17, 27, 41, 42, 56, 58]. In either case, optimization used to modify demand of service to accommodate or combat period of oversupply and under-supply. While the efficacy of this method is predicated on the price-demand relationship existing, the explicit dependence of demand on price is typically not modeled within these problems.

Vehicle Routing. Problems in vehicle routing seek to choose optimal routes x for users subject to traffic flow z while minimizing travel time and encountered traffic congestion [1, 4]. However, for sufficiently large platforms, choosing routes for a fleet can in turn directly impact the traffic flow by changing the number of vehicles along a specific route.

1.2 Formalizing Decision-dependence

Critical to our success is the observation that since the data distribution D will change after *any* deployment from the learner, the population and learner create a closed feedback loop. This can be explicitly expressed by representing the populations data distribution as a *distributional map* $D : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^k)$, so that for any $x \in \mathbb{R}^d$, $D(x)$ is a stationary probability distribution supported on $\mathbb{R}^k$.

An optimal decision in this setting is one that is not just optimal for the current state of the system that is observable to the learner, but is optimal over all possible states. This new problem takes the form

$$x^* \in \arg\min_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z \sim D(x)}[f(x, z)]. \tag{1.3}$$

While the power of this framework is in allowing us to express explicit dependence of the data z on x, it also introduces a new challenge. Finding x^* using standard stochastic gradient techniques requires that we estimate the gradient of $\mathbb{E}_{z \sim D(x)} f(x, z)$, which will in turn require that we have complete knowledge of the distribution $D(x)$. Indeed, we will assume that since z is data from a large system, then $D(x)$ is a continuous probability distribution with density function $p_x(z)$ so that

$$\mathop{\mathbb{E}}_{z \sim D(x)}[f(x, z)] = \int_{\mathbb{R}^k} f(x, z) p_x(z) dz, \tag{1.4}$$

and its gradient can be conveniently represented as

$$\nabla \mathop{\mathbb{E}}_{z \sim D(x)}[f(x, z)] = \mathop{\mathbb{E}}_{z \sim D(x)}[\nabla_x f(x, z)] + \mathop{\mathbb{E}}_{z \sim D(x)}[f(x, z) \nabla_x \log p_x(z)]. \tag{1.5}$$

While (1.4) can be estimated purely from samples from $D(x)$ for each evaluation of x, (1.5) requires that we be able to compute $\nabla_x p_x(z) = p_x(z)\nabla_x \log p_x(z)$. In this work, we will discuss how to overcome this challenge first by using derivative free optimization, and later by learning a model for $D(x)$ from samples. These approaches are not without their own drawbacks, however. The former uses a gradient approximation with only a single function evaluation, making it quite slow to converge. While the latter enjoys a faster convergence, leveraging statistical learning means that we must contend with the bias-variance trade-off.

For this reason, the approach taken at the inception of this framework can still find value. A common practice in the literature on distributional shift is the notion of *repeated retraining*: each time the distribution changes, solve the new optimization problem to convergence. This involves formulating a sequence $\{x_t\}_{t\geq 0}$ satisfying

$$x_{t+1} \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim D(x_t)}[f(x,z)]. \tag{1.6}$$

When the source of distributional shift is due to time alone, this approach is particularly appealing—provided that the problem can be solved within the time scale —as it gives us a sequence of decision that we can repeatedly deploy. However, since we are interested in distributional shift that is due to an explicit response and does not necessarily evolve in time, it is possible that this repeated retraining procedure converges (provided that the change due to x is small enough). The limit point of repeated retraining will be referred to as *stable points* in this work, are points $\bar{x} \in \mathcal{X}$ satisfying

$$\bar{x} \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim D(\bar{x})}[f(x,z)]. \tag{1.7}$$

Intuitively, $\bar{x}$ is a decision which is optimal for the stationary stochastic optimization problem that it induces. Namely, when the system is in a state induced by decision $\bar{x}$, the learners optimal decision is also $\bar{x}$. This approach is the discussion of preliminary works on the subject when convexity conditions for the map $x \mapsto \mathbb{E}_{z\sim D(x)} f(x,z)$ where unknown [23, 49, 64]. When stable points are known to be unique, one can find them using simple stochastic first order optimization methods, preventing the need to run repeated retraining to convergence. An advantage of this approach is that $\bar{x}$ can be found by standard operation of the system, and in systems in which estimating $D(x)$ is not possible or appropriate. Relative to optimizers x^*, which represent *global* solutions to the problem (1.3) in the sense that they are optimal for all possible states $D(x)$, stable points $\bar{x}$ are a *local* solution in that they are optimal when the system is in state $\bar{x}$.

There are caveats of course. As we will discuss in Chapter 2, $\bar{x}$ is decidedly **not** optimal for all possible states of the system; in fact stable points can be arbitrarily far from optimal points. The very nature of $\bar{x}$ and the ability to find them by interacting with the system implies that finding $\bar{x}$ amounts to controlling or steering the driving the system—driving the system to a desirable state in which the problem can be solves—rather the being agnostic to the state. Furthermore, finding stable points via standard stochastic gradient methods hinges on the assumption that feedback from the system in the form of z_t can be readily and *quickly* acquired. However, if the rate of feedback from the system is the limiting factor in time, it is possible that the optimization problem changes in time throughout the run of the algorithm.

1.3 Preliminaries

This section introduces the notational conventions and core definitions used in this work. Throughout, $\mathbb{R}^d$ denotes the d-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$, and Euclidean norm $\| \cdot \|$. For a matrix $X \in \mathbb{R}^{n \times m}$, $\|X\|$ denotes the spectral norm. For a given integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$ and $\mathcal{S}^{n-1}$ denotes the Euclidean hypersphere in n dimensions, $\{x \in \mathbb{R}^n | \; \|x\| = 1\}$. The symbol $\mathbb{1}_d$ is used to denote the d-dimensional vector of all ones, and I_d is the $d \times d$ identity matrix. Given vectors $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, we let $(x, z) \in \mathbb{R}^{n+m}$ denote their concatenation.

For a symmetric positive definite matrix $W \in \mathbb{R}^{d \times d}$, the weighted inner product is defined by $\langle x, y \rangle_W = \langle x, Wy \rangle$ and corresponding weighted norm $\|x\|_W = \sqrt{\langle x, x \rangle_W}$ for any $x, y \in \mathbb{R}^d$. The weighted projection onto a set $\mathcal{X} \subseteq \mathbb{R}^d$ with respect to the symmetric positive definite matrix

$W \in \mathbb{R}^{d\times d}$ is given by the map

$$\mathsf{proj}_{\mathcal{X},W}(x) := \arg\min_{y\in\mathcal{X}} \frac{1}{2}\|x-y\|_W^2 \tag{1.8}$$

for any $x \in \mathbb{R}^d$. When $W = I_d$, we simply write $\mathsf{proj}_{\mathcal{X}}$.

1.3.1 Probability measures

We restrict our focus to random variables drawn from continuous probability distributions supported over the Euclidean space. When random variables $X, Y \in \mathbb{R}^k$ are equal in distribution, i.e., $P(X \leq x) = P(Y \leq x)$ for all $x \in \mathbb{R}^k$, we write $X \overset{d}{=} Y$. We will denote the point mass distribution at $a \in \mathbb{R}$ as δ_a, so that $P(x = a) = 1$ and $P(x \neq a) = 0$.

To compare probability distributions, we will be interested in computing the distance between their associated probability measures—for which we need a complete metric space. We let $\mathcal{P}(\mathbb{R}^k)$ denote the set of finite first moment probability measures supported on $\mathbb{R}^k$ and write the Wasserstein-1 distance as

$$W_1(\mu,\nu) = \sup_{h\in\mathcal{L}_1}\left\{\mathbb{E}_{X\sim\mu}[h(X)] - \mathbb{E}_{Y\sim\nu}[h(Y)]\right\} \tag{1.9}$$

for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^k)$, where $\mathcal{L}_1$ is the set of all 1-Lipschitz continuous functions $h : \mathbb{R}^k \to \mathbb{R}$. We note that this representation is due to Kantorovich-Rubenstein duality, which holds under the Euclidean space setting we impose [9]. Additionally, under these conditions, the set $(\mathcal{P}(\mathbb{R}^k), W_1)$ forms a complete metric space [9].

Our analysis includes study of sub-Gaussian, sub-exponential, and sub-Weibull random variables as a tool to discuss high probability guarantees. We adopt the definition of sub-Weibull random variables from [61].

Definition 1 (Sub-Weibull Random Variable)**.** *If random variable X satisfies*

$$\mathbb{P}\left(|X| \geq x\right) \leq a \exp\left(-\left(\frac{x}{\omega}\right)^{\frac{1}{\theta}}\right) \tag{1.10}$$

for $a, \nu, \theta > 0$, then X is a sub-Weibull random variable with tail parameter θ and variance proxy ω. We denote this as $X \sim subW(\theta, \omega)$.

The sub-Weibull family of distributions offer a convenient theoretical tool; they include a tail parameter $\theta > 0$ that measure the thickness of the tail of a distribution, allowing us to capture sub-exponential with $\theta = 1$ and sub-Gaussian with $\theta = 1/2$ distributions. The parameter ω represents a proxy for the variance of X [61, 63]. The following result provides a bridge between alternative characterization that we may use in our analysis. Moreover, the closure properties will allow us to develop high probability bounds without appealing to concentration inequalities—which may loosen the resulting bound in our arguments.

Proposition 1 (Equivalent Characterizations)**.** *For any random variable X, the following characterizations are equivalent:*

(C1) $\exists\, \omega_1 > 0$ *such that* $\mathbb{P}(|X| \geq x) \leq 2\exp\left(-\left(\omega_1^{-1}x\right)^{1/\theta}\right)$ *for all* $x \geq 0$.

(C2) $\exists\, \omega_2 > 0$ *such that* $\|z\|_k \leq \omega_2 k^{\theta}$ *for all* $k \geq 1$.

(C3) $\exists\, \omega_3 > 0$ *such that* $\mathbb{E}[\exp(\lambda|X|^{1/\theta})] \leq \exp((\omega_3\lambda)^{1/\theta})$ *for all* $0 \leq \lambda \leq \omega_3^{-1}$.

(C4) $\exists\, \omega_4 > 0$ *such that* $\mathbb{E}[\exp(|\omega_4^{-1}X|^{1/\theta})] \leq 2$ *for all* $0 \leq \lambda \leq \omega_4^{-1}$.

Proposition 2 (Sub-Weibull Inclusion)**.** *If $X \sim subW(\theta,\omega)$ and $\theta',\omega' > 0$ are such that $\theta \leq \theta'$ and $\omega \leq \omega'$ then $X \sim subW(\theta',\omega')$.*

Proposition 3 (Sub-Weibull Closure)**.** *If $X_1 \sim subW(\theta_1,\omega_1)$, $X_2 \sim subW(\theta_2,\omega_2)$ are (possibly coupled) sub-Weibull random variables and $c \in \mathbb{R}$, then the following hold:*

(1) $X_1 + X_2 \sim subW(\max\{\theta_1,\theta_2\},\omega_1+\omega_2)$;

(2) $X_1X_2 \sim subW(\theta_1+\theta_2, g(\theta_1,\theta_2)\omega_1\omega_2)$, $g(\theta_1,\theta_2) := (\theta_1+\theta_2)^{\theta_1+\theta_2}/(\theta_1^{\theta_1}\theta_2^{\theta_2})$;

(3) $cX_1 \sim subW(\theta_1, |c|\omega_1)$.

The proofs of these lemmas can be found in [61, 63]. Our high probability analysis will primarily use characterizations (C1) and (C2) in Proposition 1. We note that if $X \sim subW(\theta,\omega)$, then (C2) holds with $\omega' = \left(\frac{\theta}{2e}\right)^{\theta}\omega$.

The class of sub-Weibull distributions allows one to consider variety of error models. For instance, it includes sub-Gaussian and sub-exponential as sub-cases by setting $\theta = 1/2$ and $\theta = 1$, respectively. We notice that a sub-Gaussian assumption was typically utilized in prior works on stochastic gradient descent; for example, the assumption $\mathbb{E}[\exp\left(\xi^2/\sigma^2\right)] \leq e$ in [48] corresponds to sub-Gaussian tail behavior. However, recent works suggest that stochastic gradient descent may exhibit errors with tails that are heavier than a sub-Gaussian (see, e.g., [33]). To further elaborate on the flexibility offered by a sub-Weibull model, we provide the following additional examples.

Example 1. *Suppose that each entry of ξ follows is a sub-Weibull distribution in the sense that $e_i^T \xi_t \sim subW(\theta, \omega)$ for all $i \in [d]$. Then $\|\xi_t\|$ is sub-Weibull with $\|\xi_t\| \sim subW(\theta, 2^\theta \sqrt{d}\nu)$ [5].* □

Example 2. *Suppose that each entry of ξ is Gaussian zero mean and variance ς^2; then, it it sub-Gaussian with sub-Gaussian norm $C\varsigma$, with C an absolute constant [59], and it is therefore $subW(1/2, C'\varsigma)$ with C' an absolute constant.* □

Example 3. *Suppose that ξ is a random variable with mean $\mu := \mathbb{E}\xi_t$, such that $\xi \in [a, b]$ almost surely for some $a, b \in \mathbb{R}$. Then $\xi - \mu \sim subW(1/2, (a-b)/\sqrt{2})$ [5].* □

1.3.2 Convex Analysis

In our minimization problem formulation, we will consider cost functions that are (strongly) convex and smooth so as to restrict ourselves to problems with a desirable geometry.

Definition 2 (Convexity)**.** *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex on $\mathcal{X}$ provided that, for any $x, y \in \mathcal{X}$*

$$f(\tau x + (1-\tau)y) \leq \tau f(x) + (1-\tau) f(y) \tag{1.11}$$

for all $\tau \in [0, 1]$.

Intuitively, a convex function is such that the line joining any two points lies above its graph. We are primarily interested in continuously differentiable convex functions, which satisfy the characterization:

$$f(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle \tag{1.12}$$

for any $x, y \in \mathbb{R}^d$. Convexity is typically a desired property for objectives within the optimization literature as the geometry informs the existence of minimizers. Indeed, convex functions have the unique property that every *local* minimizers is a *global* minimizer, and hence the value $f(x^*)$ is unique for all $x^* \in \arg\min_x f(x)$. An even stronger notion of convexity, strong-convexity, allows us to characterize functions with *unique* minimizers.

Definition 3 (Strong-Convexity). *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is γ-strongly convex if the function*

$$x \mapsto f(x) - \frac{\gamma}{2}\|x\|^2$$

is convex.

Intuitively, this says that f can be lower bounded by a quadratic function with modulus γ. When $\gamma = 0$, then f is merely convex. In our analysis, we will frequently make use of the characterization that γ-strongly convex function satisfy

$$f(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{\gamma}{2}\|x - y\|^2 \tag{1.13}$$

for any $x, y \in \mathbb{R}^d$.

Even when convexity fails, the notion of Lipschitz continuity is still required to achieve desirable outcomes in optimization. We will refer to a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ as L-smooth if its gradient ∇f is L-Lipschitz continuous.

Definition 4 (Smoothness). *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth provided that ∇f satisfies*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{1.14}$$

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

Smoothness of the cost limits the rate of change of the gradient, and gives us the relationship:

$$f(x) \leq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2}\|x - y\|^2 \tag{1.15}$$

for any $x, y \in \mathbb{R}^d$. If f is twice continuously differentiable, γ-strongly convex, and L-smooth then the hessian satisfies

$$\gamma I_d \leq \nabla^2 f(x) \leq L I_d \tag{1.16}$$

for all $x \in \mathbb{R}^d$.

Frequently we will restrict the scope of our problem to a subset $\mathcal{X} \subseteq \mathbb{R}^d$, where $\mathcal{X}$ typically captures a set of implicit constraints that are relevant to the problem. In this case, the notion of local (strong) convexity and smoothness simply hold by replacing $\mathbb{R}^d$ with the set $\mathcal{X}$. Then the problem can be solved provided that the set $\mathcal{X}$ is convex.

Definition 5 (Convex Set)**.** *A set $\mathcal{X} \subseteq \mathbb{R}$ is convex provided that, for all $x, y \in \mathcal{X}$*

$$\tau x + (1 - \tau) y \in \mathcal{X} \tag{1.17}$$

for all $\tau \in [0, 1]$.

Similar to convex functions, a set is convex provided that a line segment joining any two points inside $\mathcal{X}$ is completely contained in $\mathcal{X}$.

1.3.3 Games

In our continuous game formulation, we consider a game that consists of n players. Each player has a cost function F_i, distributional map D_i, and decision set $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$. Hence, each player chooses a decision, or strategy $x_i \in \mathcal{X}_i \subseteq \mathbb{R}^{d_i}$. The concatenation of the decision variables is written as $x = (x_1, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{R}^d$ where $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$ and $d = \sum_{i=1}^n d_i$. For a fixed agent i, we will decompose the decision x as $x = (x_i, x_{-i})$ where $x_{-i} \in \mathbb{R}^{d-d_i}$ is the strategy vector of all agents excluding the ith one.

The collection of costs F_i and decision sets $\mathcal{X}_i$ defines the game

$$\min_{x_i \in \mathcal{X}_i} F_i(x_i, x_{-i}), \quad i \in [n]. \tag{1.18}$$

A Nash equilibrium of this game is a point $x^* \in \mathcal{X}$ provided that

$$x_i^* \in \arg\min_{x_i \in \mathcal{X}_i} F_i(x_i, x_{-i}^*) \tag{1.19}$$

for all $i \in [n]$. Intuitively, x^* is a strategy such that no agent can be incentivized by its cost to deviate from x_i^* when all other agents play x_{-i}^*. Finding Nash equilibria is the primary focus in this setting.

Games of this form are commonly cast into a variational inequality framework. This is due, in part, to the observation that the Nash equilibria $x^* \in \mathcal{X}$ are the solutions to the variational inequality

$$\langle x - x^*, G(x^*) \rangle \geq 0, \quad \forall\, x \in \mathcal{X},$$

where the gradient map $G : \mathbb{R}^d \to \mathbb{R}^d$ is defined as

$$G(x) = (\nabla_1 F_1(x), \ldots, \nabla_n F_n(x)) \,. \tag{1.20}$$

Here, the notation ∇_i is used to represent the partial gradient ∇_{x_i}. We will denote the set of Nash equilibria of a game with gradient map G and domain $\mathcal{X}$ as $\texttt{NASH}(G, \mathcal{X})$. Existence of solutions to variational inequalities of this form is guaranteed provided that the set $\mathcal{X}$ is convex and compact and the gradient map G is monotone; uniqueness is guaranteed when G is strongly-monotone [25]. We say that G is α-strongly-monotone on $\mathcal{X}$ provided that there exists $\alpha > 0$ such that

$$\langle x - y, G(x) - G(y) \rangle \geq \alpha \|x - y\|^2, \;\; \forall\, x, y \in \mathcal{X}, \tag{1.21}$$

and monotone when $\alpha = 0$. In this work, we primarily focus on strongly-monotone games. While monotone games are tractable, methods for solving them with decision-dependent distributions require alternative gradient estimators—a topic we leave to future work.

1.4 Organization

This thesis is based on the following four papers centered around stochastic optimization problems with decision dependent distributions:

(1) 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, 2021

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

(3) Killian 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.

(4) Killian Wood, Ahmed Zamzam, and Emiliano Dall'Anese. Solving decision-dependent games by learning from feedback. IEEE Open Journal of Control Systems, 2024.

To provide context for the exposition, we will reference material from pertinent references therein. In Chapter 2, we lay the theoretical foundations for solving the convex optimization problem in (1.3) using both optimal and stable approaches by drawing on the works of (1) and (4). In Chapter 3, we build on the first Chapter by considering a Saddle Point problem with decision-dependent distributions based the works of (2) and (3). In Chapter 4, we move to non-cooperative multiplayer games, which is the subject of (4).

Chapter 2

Convex Optimization

In this chapter, we provide an overview of the stochastic optimization problem with decision-dependent distributions, with the ultimate goal of highlighting the work in [64]. Formally, this problem takes the form

$$x^* \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ F(x) := \mathop{\mathbb{E}}_{z \sim D}[f(x,z)] \right\} \tag{2.1}$$

with cost $f : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$, distributional map $D : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^k)$, and domain $\mathcal{X} \subseteq \mathbb{R}^d$. As we discussed in the introductory chapter, the power in this problem statement lies in in the fact that it allows the learner to explicitly express the dependence of a data distribution on the optimization as a means to combat optimization-induced distributional shift. In doing so however, we have constructed a problem that is significantly more difficult to solve. The minimizers x^* are appealing in that they represent a decision that is uniformly best for all possible states of the system in which we acquire data. It is precisely this expression that prevents us from formulating a basic stochastic gradient algorithm that is capable of finding x^*: computing the gradient ∇F requires the gradient of the probability density of the distributional map with respect to x, and thus estimating ∇F requires complete knowledge of the distribution.

For this reason, the work of "Performative Prediction" draws on the analogy of decision-dependent to time-varying distributional shift to formulate a repeated retraining heuristic. This posits that we solve a new optimization problem each time the data distribution changes. In this setting, the distribution changes due to deploying the previous optimizer, but the idea is the same.

This amounts to formulating the sequence

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z \sim D(x_t)}[f(x,z)].$$

Unique to our setting, however, is the fact that the degree to which the iterates x_t and x_{t+1} is entirely dependent on properties of the cost f and the distributional map D. Thus if, these properties can be characterized, and do not change *too much*, it is possible that the sequence x_t converges to a limit in $\mathcal{X}$. In the next section, we will discuss the conditions required for convergence of repeated retraining and alternative ways of finding these limit points.

2.1 The Stability Problem

In the work that follows, we denote stable points $\bar{x} \in \mathcal{X}$ as the limit points of repeated retraining. Hence, they satisfy the relation

$$\bar{x} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z \sim D(\bar{x})}[f(x,z)]. \tag{2.2}$$

Relative to the optimizers x^*, they only satisfy the local property of being optimal for the stationary problem that they induce; however, the conditions for their existence and uniqueness as well as the mechanism required for finding them is more mild than that of optimizers. This work is primarily interested in the case of uniqueness of solutions and hence we only present these conditions. For conditions on mere existence, see [49], though finding $\bar{x}$ in this case remains an open problem.

In the following, we present a main result of [49]: that in setting where the effects of "performativity" or decision-dependence are bounded by the condition number of the cost, repeated retraining terminates and the limit is a stable point.

Theorem 4 (Repeated Retraining Convergence, [49])**.** *Suppose that the following hold:*

(i) $x \mapsto f(x,z)$ *is* γ*-strongly convex for all* $z \in \mathbb{R}^k$*,*

(ii) $z \mapsto \nabla f(x,z)$ *is* L*-Lipschitz continuous for all* $x \in \mathcal{X}$*,*

(iii) $x \mapsto D(x)$ *is* ν*-Lipschitz continuous on* $(\mathcal{P}(\mathbb{R}^k), W_1)$*,*

(iv) $\mathcal{X} \subseteq \mathbb{R}^d$ *is closed and convex.*

If $\nu L/\gamma < 1$, *then sequence* $\{x_t\}_{t \geq 0}$ *given by*

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z \sim D(x_t)} [f(x, z)].$$

converges to a unique limit $\bar{x} \in \mathcal{X}$.

Proof of this result is due to [49] and amounts to showing that the fixed point iteration satisfies the Banach-Picard fixed point theorem. We note that conditions (i) and (ii) are standard conditions for convex optimization when pursuing unique solutions and hence represent a best-case scenario. The novelty of this result is condition (iii), typically referred to as ν-sensitivity in the literature, which quantifies the decision-dependent component of the problem. Formally, this condition states that

$$W_1(D(x), D(y)) \leq \nu \|x - y\| \tag{2.3}$$

for all $x, y \in \mathbb{R}^d$. Simple examples of distributional maps that satisfy this include location scale families, in which $z \sim D(x)$ takes the form

$$z \overset{d}{=} \xi + Bx \tag{2.4}$$

where ξ is some zero-mean stationary random variable, and $B \in \mathbb{R}^{k \times d}$. A simple calculation yields that

$$W_1(D(x), D(y)) \leq \|B\| \|x - y\|, \tag{2.5}$$

and hence ν-sensitivity provided that $\nu = \|B\|$ is well-defined.

Conversely, the univariate Gaussian $D(x) = \mathcal{N}(\sqrt{x}, \sigma^2)$ is not ν-sensitive for any finite $\nu > 0$. Indeed, from [16, Theorem 3.1] we have that

$$W_1(D(x), D(y)) = |\sqrt{x} - \sqrt{y}| \tag{2.6}$$

for any $x, y \geq 0$, so D is ν-sensitive if and only if $x \mapsto \sqrt{x}$ is ν-Lipschitz continuous for some ν. To see that this is not the case, we can consider a simple contradiction argument. Suppose that there does exist $\nu > 0$ such that

$$|\sqrt{x} - \sqrt{y}| \leq \nu |x - y| \tag{2.7}$$

for any $x, y \geq 0$. Then it must hold for $x = 1/c$ for $c > 0$ and $y = 0$. Substituting into the above and rearranging yields the implication that $\sqrt{c} \leq \nu$. However, choosing $c = \nu^2 + 1$ yields a clear contradiction.

While repeated retraining is a useful conceptual and theoretical tool for this analysis, using this as an algorithmic method is highly impractical. Though the rate of this fixed point iteration is linear, performing the assignment requires that we solve an expected minimization problem at each iteration. To avoid this, we leverage the observation that $\bar{x}$ satisfies the fixed point relation

$$\bar{x} = \mathsf{proj}_{\mathcal{X}} \left[\bar{x} - \eta \mathop{\mathbb{E}}_{z \sim D(\bar{x})} [\nabla_x f(\bar{x}, z)] \right], \tag{2.8}$$

for any $\eta > 0$, and hence the update

$$x_{t+1} = \mathsf{proj}_{\mathcal{X}} \left[x_t - \eta_t \mathop{\mathbb{E}}_{z \sim D(x_t)} [\nabla_x f(x_t, z)] \right], \tag{2.9}$$

should be suitable for finding stable points. Since this amounts to deterministic gradient descent, it can be shown that this algorithm converges linearly to $\bar{x}$ provided that $x \mapsto f(x, z)$ is strongly-convex and L-smooth [49, Theorem 3.8]. In practice, we will use responses $z_t \sim D_t(x_t)$ to formulate some stochastic gradient estimator g_t and do

$$x_{t+1} = \mathsf{proj}_{\mathcal{X}} [x_t - \eta_t g_t]. \tag{2.10}$$

For this to work, we must assume the existence of a mechanism by which the learner can acquire samples from $D(x_t)$ at each iteration (i.e. an oracle). Though this assumption is somewhat mild in terms of application, it can cause issue if the data acquisition step requires more time than the algorithmic update—an observation that motivates our next section. To proceed, we will further assume the that stochastic gradient estimator satisfy the following assumptions.

Assumption 1 (Stochastic Framework). *Let $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ with elements*

$$\mathcal{F}_t := \sigma(g_\tau,\ \tau \leq t) \tag{2.11}$$

be the natural filtration of the Borel σ-algebra over $\mathbb{R}^d$ with respect to g_t. Denote $\mathbb{E}_t[\ \cdot\ ] = \mathbb{E}[\cdot|\mathcal{F}_t]$ as the conditional expectation with respect $\mathcal{F}_t$ over distribution $D(x_t)$. Suppose that g_t satisfies the following:

(1) ***(Unbiased)*** $\mathbb{E}_t[g_t] = \mathbb{E}_{z\sim D_t(x_t)}[\nabla_x f(x_t, z)]$

(2) ***(Bounded Variance)*** $\mathbb{E}_t\|g_t - \mathbb{E}_{z\sim D_t(x_t)}[\nabla_x f(x_t, z)]\|^2 \leq \sigma^2$

This assumption is a common stochastic framework used to study convergence of optimization algorithms in expectation. We will revisit variations of this framework throughout the rest of the work. With these assumptions, we are able to state the following convergence result due to [22].

Theorem 5. *Suppose that the following hold with $\nu L/\gamma < 1$*

(i) $x \mapsto f(x, z)$ *is γ-strongly convex for all $z \in \mathbb{R}^k$*

(ii) $z \mapsto \nabla f(x, z)$ *is L-Lipschitz continuous for all $x \in \mathcal{X}$*

(iii) $x \mapsto D(x)$ *is ν-Lipschitz continuous on $(\mathcal{P}(\mathbb{R}^k), W_1)$.*

(iv) $\mathcal{X} \subseteq \mathbb{R}^d$ *is closed and convex.*

Then the sequence $\{x_t\}$ generated by (2.10) *satisfies the following*

(1) if $\eta_t = \eta > 0$ then

$$\mathbb{E}\|x_t - \bar{x}\|^2 \leq (1 - (\gamma - \nu L)\eta)^t \|x_0 - \bar{x}\|^2 + \frac{2\sigma^2}{1 - \eta} \tag{2.12}$$

(2) if $\eta_t = 2/(r + t)$ for $r > 0$ then

$$\mathbb{E}\|x_t - \bar{x}\|^2 \leq \frac{M}{r + t} \tag{2.13}$$

where

$$M = \max\left\{ r\|x_0 - \bar{x}\|^2, \frac{1}{(\gamma - \nu L)} \right\} \tag{2.14}$$

Since this is a standard result within the literature, we refer the reader to [22, 44] for proof. With this result in mind, we have effectively demonstrated that the stable problem enjoys the same convergence guarantees as standard stochastic optimization with stochastic gradient descent. This result does however rely heavily on the fact that samples can be readily acquired at each iteration t to form the stochastic gradient estimator g_t. The result above demonstrates that, for a nicely conditioned problem, a reasonable degree of accuracy can be achieved with a decaying step-size after 1,000 iterations. When the time required to get feedback is a mere 5 seconds or less, then we can reach an answer in less than 1.4 hours. However, in the extreme case, 20 seconds between updates can take 5.5 hours to reach an answer—a time window over which many real-world problems and systems will naturally evolve dynamically. For this reason, it is necessary to develop a time-varying formulation of this stable problem to accommodate such a case by tracking a trajectory of stable points as it evolves in time.

2.2 Time-varying Stability

This section considers the problem of developing and analyzing online algorithms to track the trajectory of solutions for time-varying stochastic optimization problems with decision-dependent distributions. The sequence of problems has the form

$$\min_{x \in \mathcal{X}_t} \mathop{\mathbb{E}}_{z \sim D_t(x)} \left[f_t(x, z) \right], \tag{2.15}$$

where $t \in \mathbb{N}_0$ is a time index, $x \in \mathbb{R}^d$ is the decision variable, $D_t : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^k)$ is a map from the set $\mathbb{R}^d$ to the space of distributions, z is a random variable supported on $\mathbb{R}^k$, $f_t : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ is the loss function, and $\mathcal{X}_t \subseteq \mathbb{R}^d$ is a closed and convex set.

At each time t, we have a new problem instance and hence a distinct solution. Here we are only interested in the trajectory of stable points $\{\bar{x}_t\}_{t \geq 0}$, which satisfies the relation

$$\bar{x}_t \in \arg\min_{x \in \mathcal{X}_t} \mathop{\mathbb{E}}_{z \sim D_t(\bar{x}_t)} \left[f_t(x, z) \right]. \tag{2.16}$$

Like the time-invariant case, the stable point $\bar{x}_t$ is optimal for the stationary stochastic optimization it induces. To proceed, we demonstrate some of the basic properties of stable points still hold in this setting.

2.2.1 Time-varying Framework

The assumptions that follow are time-varying analogs of the standard optimization framework we pose in the previous section. We adopt the following list of assumptions at each time $t \in \mathbb{N}_0$.

Assumption 2 (Strong Convexity)**.** *There exists $\gamma_t > 0$ such that $x \mapsto f_t(x,z)$ is γ_t-strongly convex for all $z \in \mathbb{R}^k$.*

Assumption 3 (Joint smoothness)**.** *There exist $L_t > 0$ such that $x \mapsto \nabla_x f_t(x,z)$ is L_t-Lipschitz continuous and $z \mapsto \nabla_x f_t(x,z)$ is L_t-Lipschitz continuous for all $x \in \mathbb{R}^d$.*

Assumption 4 (Distributional Sensitivity)**.** *There exists $\nu_t > 0$ such that*

$$W_1(D_t(x), D_t(x')) \leq \nu_t \|x - x'\| \tag{2.17}$$

for any $x, x' \in \mathbb{R}^d$. □

Assumption 5 (Convex Constraint Set)**.** *The set $\mathcal{X}_t \subseteq \mathbb{R}^d$ is closed and convex.* □

Assumption 6 (Time Variability)**.** *There exists $\Delta \in (0,\infty)$ such that the stable drift sequence $\Delta_t := \|\bar{x}_{t+1} - \bar{x}_t\|$ satisfies $\Delta_t \leq \Delta$.*

Assumptions 2, 3, and 5 are standard assumption used throughout stochastic optimization to grantee convergence (time-invariant) and tracking (time-varying) of unique solutions, and hence are rather mild. Assumption 4 is unique to the literature on decision-dependent distributions and is necessary to characterize the degree of decision-dependence within the system [22, 49]. Note that $\nu_t = 0$ here naturally implies the absence of decision-dependence and hence stable points and optimal points coincide. Lastly, Assumption 6 is standard in the literature on time-varying optimization problems as it prevents the trajectory of solutions from changing too much due to adversarially chosen functions f_t. This is perhaps the most mild assumption one could place on time varying assumption, as it merely assumes that the worst drift is finite. If this were not the case, then Δ_t would grow unbounded and tracking $\{\bar{x}_t\}_{t\geq}$ would be meaningless.

In addition to these assumption, we will suppose that there exist finite constants $\gamma, L, \nu > 0$ such that $\gamma = \inf_t \gamma_t$, $L = \sup_t L_t$, and $\nu = \sup_t \nu_t$. These serve as uniform constants for which the above assumptions hold for all t, and will allow us to develop a worst-case analysis.

Lemma 6 ([49, Theorem 3.5])**.** *Let Assumptions 2-5 hold, and suppose that* $\frac{\nu_t L_t}{\gamma_t} < 1$ *for all* $t \in \mathbb{N}_0$*. Then, a sequence of performatively stable points* $\{\bar{x}_t\}_{t\in\mathbb{N}_0}$ *exists and is unique.*

In general, performatively stable points may not coincide with the optimizers of the original problem (2.15). However, an explicit error bound can be derived, as formally stated next.

Lemma 7 ([49, Theorem 4.3])**.** *Suppose that the function* $z \mapsto f_t(x, z)$ *is* L_t*-Lipschitz continuous for all* $x \in \mathbb{R}^d$ *and* $t \in \mathbb{N}_0$*. Then, under the same assumptions of Lemma 6, it holds that*

$$\|\bar{x}_t - x_t^*\| \leq 2\nu_t L_t \gamma_t^{-1}, \text{ for all } \in \mathbb{N}_0. \tag{2.18}$$

The proof Lemma 7 follows from [49, Thm 3.5, Thm 4.3]. In the remainder of this paper, we assume that the assumptions of Lemma 6 are satisfied, so that the performatively stable point sequence is unique. We illustrate the difference between $\bar{x}_t$ and x_t^* in the following example.

Example 4. *Consider an instance of* (2.15) *where* $f_t(x, z) = x^2 + z$, $\mathcal{X}_t = \mathbb{R}$, $D_t(x) = \mathcal{N}(\mu_t x, \sigma_t^2)$, $\mu_t, \sigma_t > 0$*. In this case, the objective can be specified in closed form as:* $\underset{z\sim D_t(x)}{\mathbb{E}}\left[x^2 + z\right] = x^2 + \mu_t x$*, and thus the unique performatively optimal point is given by* $x_t^* = -\mu_t/2$*. To determine the performatively stable point, notice that* $\nabla_x f_t(x, z) = 2x$*, and thus* $\bar{x}_t$ *satisfies* $\underset{z\sim D_t(\bar{x}_t)}{\mathbb{E}}[2\bar{x}_t] = 0$*, which implies* $\bar{x}_t = 0$*. The bound in* (2.18) *thus holds by noting that* $\nu_t = \mu_t$*,* $\gamma_t = 1$*, and* $\gamma_t = 2$*.* □

2.2.2 Deterministic Tracking

For the sake of notational ease, we will denote the decoupled cost as

$$F_t(x|y) = \mathbb{E}_{z\sim D_t(y)}[f_t(x, z)], \tag{2.19}$$

and the corresponding gradient as

$$G_t(x|y) = \mathbb{E}_{z\sim D_t(y)}[\nabla_x f_t(x, z)]. \tag{2.20}$$

In order to successfully track stable points, we will employ a time varying analog of the procedure in (2.10). This takes the form

$$x_{t+1} = \mathsf{proj}_{\mathcal{X}_t}\left[x_t - \eta_t G_t(x_t|x_t)\right]. \tag{2.21}$$

This is an example of an *online* algorithm as it uses information that arrives sequentially in time to perform its update. Observe that the update above can be expressed using the algorithmic map $\mathcal{A}_t : \mathcal{X}_t \times \mathcal{X}_t \to \mathcal{X}_t$ given by

$$\mathcal{A}_t(x|y) = \mathsf{proj}_{\mathcal{X}_t}\left[x - \eta_t G_t(x|y)\right], \tag{2.22}$$

for all $x, y \in \mathcal{X}_t$.

To analyze the behavior of our deterministic update, we rely on characterizing the gradient map G_t in a way that will allow us to use classical optimization results. The first of which is the so-called *gradient deviations* property observed in [23, 49]. We refer the reader to these works for a detailed proof.

Lemma 8 (Gradient Deviations)**.** *If Assumption 2 holds, then for any $t \geq 0$, $x \mapsto G_t(x_0|x)$ is $\nu_t L_t$-Lipschitz continuous for any $x_0 \in \mathbb{R}^d$. Specifically,*

$$\|G_t(x_0|x) - G_t(x_0|y)\| \leq \nu_t L_t \|x - y\| \tag{2.23}$$

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

This result allows us to describe the impact of decision-dependent behavior on the gradient. Our analysis $\mathcal{A}_t$ functions by relating the trajectory of $\{x_t\}_{t\geq 0}$ to the trajectory of online gradient descent in the solution state $D_t(\bar{x}_t)$. Hence, we first show that an online gradient descent step is Lipschitz continuous. Proof of this result relies on classical convex analysis result.

Lemma 9 (Contractive Map)**.** *If Assumptions 2-3 then the map* $x \mapsto \mathcal{A}_t(x|x_0)$ *is* ρ_t*- Lipschitz continuous for any* $x_0 \in \mathbb{R}^d$ *where* $\kappa_t = \max\{|1-\gamma_t\eta_t|, |1-L_t\eta_t|\}$*. That is,*

$$\|\mathcal{A}_t(x|x_0) - \mathcal{A}_t(y|x_0)\| \leq \kappa_t\|x-y\|, \tag{2.24}$$

for any $x, y \in \mathbb{R}^d$*. Furthermore, if* $\kappa_t < 1$ *then* $\bar{x}_t = \mathcal{A}_t(\bar{x}_t|\bar{x}_t)$*.*

Proof. Fix $x_0, x, y \in \mathbb{R}^d$. We note that $x \mapsto x - \eta_t G_t(x|x_0)$ has Jacobian J_t defined by $J_t(x) = I_d x - \eta_t \nabla G_t(x|x_0)$ where $\nabla G_t(x|x_0) = \mathbb{E}_{z\sim D_t(x_0)}[\nabla_x^2 f_t(x,z)]$. Due to non-expansiveness, we have that

$$\|\mathcal{A}_t(x|x_0) - \mathcal{A}_t(y|x_0)\| \leq \|\,(x - \eta_t G_t(x|x_0)) - (y - \eta_t G_t(y|x_0))\,\| \leq \sup_x \|J_t(x)\|\|x-y\|,$$

so the result follows if $\sup_x \|x - \eta_t \nabla G_t(x|x_0)\| \leq \rho_t$. By assumption we have that

$$\gamma_t\|x\|^2 \leq \langle x, \nabla G_t(x|x_0)\rangle \leq L_t\|x\|^2,$$

and hence

$$(1 - L_t\eta_t)\,\|x\|^2 \leq \langle x, J_t(x))\rangle \leq (1-\gamma_t\eta_t)\,\|x\|^2.$$

From this the result follows. □

In characterizing the individual components of change in our algorithmic step, we have enabled the ability to demonstrate how the error incurred by the sequence $x_{t+1} = \mathcal{A}_t(x_t|x_t)$ propagates relative to its starting position $\|x_0 - \bar{x}_0\|$ and the stable drift Δ_t.

Lemma 10 (Tracking Error Bound)**.** *Let Assumptions 2-5 hold, and suppose that*$\frac{\nu_t L_t}{\gamma_t} < 1$ *for all* $t \geq 0$*. Then the sequence* $\{x_t\}_{t\geq 0}$ *defined by* $x_{t+1} = \mathcal{A}_t(x_t|x_t)$ *satisfies*

$$\|x_{t+1} - \bar{x}_{t+1}\| \leq \; a_t\|x_0 - \bar{x}_0\| + \sum_{i=0}^{t} b_i\Delta_i, \tag{2.25}$$

where $a_t := \prod_{i=1}^t \kappa_t + \nu_t L_t \eta_i$,

$$b_i := \begin{cases} 1 & \text{if } i = t, \\ \prod_{k=i+1}^t \kappa_i + \eta_i L_i \eta_i & \text{if } i \neq t, \end{cases}$$

Proof. Note that $x_t \in \mathcal{X}_t$ for all $t \geq 0$ directly follows by definition of Euclidean projection. By using the triangle inequality, we find that

$$\begin{aligned}\|x_{t+1} - \bar{x}_{t+1}\| &\leq \|x_{t+1} - \bar{x}_t\| + \|\bar{x}_t - \bar{x}_{t+1}\| \\ &= \|\mathcal{A}_t(x_t|x_t) - \mathcal{A}_t(\bar{x}_t|\bar{x}_t)\| + \Delta_t \\ &\leq \|\mathcal{A}_t(x_t|x_t) - \mathcal{A}_t(x_t|\bar{x}_t)\| + \|\mathcal{A}_t(x_t|\bar{x}_t) - \mathcal{A}_t(\bar{x}_t|\bar{x}_t)\| + \Delta_t,\end{aligned}$$

where the first identity follows from the definition of $x \mapsto \mathcal{A}_t(x|x)$ and the second inequality follows from telescoping. Applying (2.23) and Lemma 9 yields:

$$\begin{aligned}\|x_{t+1} - \bar{x}_{t+1}\| &\leq \eta_t\|G_t(x_t|x_t) - G_t(x_t|\bar{x}_t)\| + \|G_t(x_t|\bar{x}_t) - G_t(\bar{x}_t|\bar{x}_t))\| + \Delta_t \\ &\leq \nu_t L_t \eta_t \|x_t - \bar{x}_t\| + \kappa_t\|x_t - \bar{x}_t\| + \Delta_t \\ &= (\kappa_t + \nu_t L_t \eta_t)\,\|x_t - \bar{x}_t\| + \Delta_t. \qquad (2.26)\end{aligned}$$

Thus we obtain the following by expanding the recursion:

$$e_t \leq \left(\prod_{i=0}^{t} \lambda_i\right) e_0 + \Delta_t + \sum_{i=0}^{t-1}\left(\prod_{k=i+1}^{t} \lambda_k\right)\Delta_i,$$

where we defined $\lambda_t := \kappa_t + \nu_t L_t \eta_t$. The bound (2.25) then follows by definition of the sequences $\{a_t\}$ and $\{b_t\}$. □

We note that a key feature of this result is that the contraction coefficient $\rho_t = \kappa_t - \nu_t L_t$ is comprised of the contraction due to online gradient descent in κ_t and the decision-dependent shift in the gradient represented by $\nu_t L_t$. Hence, in the absence of decision-dependence, $\nu_t = 0$ and we recover the online gradient descent bound. Furthermore, we know that $\rho_t \geq 0$. Indeed $\rho_t = \gamma_t L_t(\gamma_t + L_t)^{-1} - \nu_t L_t \geq 0$ if and only if

$$\nu_t \leq \frac{\gamma_t}{\gamma_t + L_t},$$

which is equivalent to

$$\nu_t \leq \frac{\gamma_t/L_t}{\gamma_t/L_t + 1} < \gamma_t/L_t$$

Theorem 11 (Neighborhood Tracking)**.** *Let Assumptions 2-5 hold, and suppose that η_t satisfies*

$$\eta_t \in \left[\frac{\varepsilon}{\gamma_t - \nu_t L_t}, \frac{2}{\gamma_t + L_t}\right] \tag{2.27}$$

for some $\varepsilon \in (0,1)$ for all $t \geq 0$. Then,

$$\limsup_{t\to\infty} \|x_t - \bar{x}_t\| \leq \frac{\Delta}{\varepsilon}. \tag{2.28}$$

Proof. Fix $\varepsilon \in (0,1)$. Observe that since $\eta_t \leq 2(\gamma_t + L_t)^{-1}$, then $\kappa_t = 1 - \gamma_t\eta_t$. Hence,

$$\kappa_t + \nu_t L_t \eta_t = 1 - \gamma_t\eta_t + \nu_t L_t \eta_t \leq 1 - \varepsilon \tag{2.29}$$

if and only if $\eta_t \geq \varepsilon(\gamma_t - \nu_t L_t)^{-1}$. Note that since $\nu_t L_t \gamma_t^{-1} < 1$ by assumption, then $\gamma_t - \nu_t L_t > 0$.

It follows from the proof of Lemma 10 that

$$\|x_t - \bar{x}_t\| \leq (1 - \kappa_t\eta_t + \nu_t L_t \eta_t)\,\|x_{t-1} - \bar{x}_{t-1}\| + \Delta_t \leq (1-\varepsilon)\,\|x_{t-1} - \bar{x}_{t-1}\| + \Delta.$$

By repeatedly applying this bound, we find that

$$\|x_t - \bar{x}_t\| \leq (1-\varepsilon)^t \|x_0 - \bar{x}_0\| + \Delta \sum_{k=1}^{t-1} (1-\varepsilon)^k \leq (1-\varepsilon)^t \|x_0 - \bar{x}_0\| + \frac{\Delta}{\varepsilon}$$

where the last step follows by bounding the geometric series by its limit as $t \to \infty$. Taking the limit supremum of both sides yields the result. □

This section provides a baseline for our analysis of online gradient descent in the decision-dependent setting, but with full information. In the next section, we extend this to algorithms that use a stochastic gradient direction.

2.3 Online Stochastic Gradient Descent

In practice, the development of the previous section is not possible to achieve as we merely have access to samples from $D_t(x_t)$. In this section, we demonstrate that the online stochastic gradient update

$$x_{t+1} = \mathsf{proj}_{\mathcal{X}_t}\left[x_t - \eta_t g_t\right], \tag{2.30}$$

permits us to track stable points in expectation and high probability. Throughout our analysis, we interpret 2.30 as an inexact online gradient update whose additive error is captured by

$$\xi_t := g_t - G_t(x_t|x_t). \tag{2.31}$$

In order to develop high probability bounds, we introduce the following assumption on the tails of this gradient error, as is common for a result of this type.

Assumption 7 (Sub-Weibull Error). *For all $t \geq 0$, $\|\xi_t\| \sim subW(\theta, \nu_t)$ for some $\theta, \omega_t > 0$.*

Assumption 7 allows us to describe a variety of sub-cases, including scenarios where the error follows sub-Gaussian and sub-Exponential distributions [59], or any distribution with finite support. Further, notice that Assumption 7 does not require the random variables $\{\xi_t\}_{t\in\mathbb{N}_0}$ to be independent.

Theorem 12 (Expected and High-probability Bounds). *Let Assumptions 2-5 hold, and suppose that $\frac{\nu_t L_t}{\gamma_t} < 1$ for all $t \in \mathbb{N}_0$. Then, the following estimates hold for* (2.30)*:*

(1) For all $t \in \mathbb{N}$,

$$\mathbb{E}\|x_{t+1} - \bar{x}_{t+1}\| \leq (\kappa_t + \nu_t L_t \eta_t)\|x_0 - \bar{x}_0\| + \sum_{i=1}^{t} b_i(\Delta_i + \eta_i \mathbb{E}\|\xi_i\|]) . \tag{2.32}$$

(2) If, additionally, Assumption 7 holds and $\delta \in (0,1)$, then with probability $1-\delta$:

$$e_{t+1} \leq \left(\frac{2e}{\theta}\right)^{\theta} \log^{\theta}\left(\frac{2}{\delta}\right)\left(a_t\|x_0 - \bar{x}_0\| + \sum_{i=1}^{t} b_i(\Delta_i + \eta_i \nu_i)\right), \tag{2.33}$$

where $\{a_t\}$ and $\{b_i\}$ are as in Theorem 10.

Proof. Note that $x_t \in \mathcal{X}_t$ for all $t \in \mathbb{N}$ directly follows by definition of Euclidean projection. To prove the first result, we first find a stochastic recursion. By the triangle inequality:

$$\|x_{t+1} - \bar{x}_{t+1}\| \leq \|g_t - G_t(\bar{x}_t|\bar{x}_t)\| + \Delta_t \leq \|g_t - G_t(x_t|x_t)\| + \|G_t(x_t|x_t) - G_t(\bar{x}_t|\bar{x}_t)\| + \Delta_t,$$

where the second inequality follows by adding and subtracting $G_t(x_t|x_t)$. By iterating (2.26), we have

$$\|G_t(x_t|x_t) - G_t(\bar{x}_t|\bar{x}_t)\| \leq \lambda_t\|x_t - \bar{x}_t\| + \Delta_t,$$

where $\lambda_t := \kappa_t + \nu_t L_t \eta_t$, and thus

$$\|x_{t+1} - \bar{x}_{t+1}\| \leq \eta_t \|g_t - G_t(x_t|x_t)\| + \lambda_t \|x_t - \bar{x}_t\| + \Delta_t.$$

This yields the stochastic recursion $\|x_{t+1} - \bar{x}_{t+1}\| \leq \lambda_t \|x_t - \bar{x}_t\| + \Delta_t + \eta_t \|\xi_t\|$. Expanding the recursion yields

$$\|x_{t+1} - \bar{x}_{t+1}\| \leq \left(\prod_{i=0}^{t} \lambda_i\right) e_0 + \Delta_t + \sum_{i=0}^{t-1} \left(\prod_{k=i+1}^{t} \lambda_k\right) (\Delta_i + \eta_i \|\xi_i\|),$$

or, equivalently,

$$\|x_{t+1} - \bar{x}_{t+1}\| \leq a_t \|x_0 - \bar{x}_0\| + \sum_{i=0}^{t} b_i (\Delta_i + \eta_i \|\xi_i\|). \tag{2.34}$$

Thus, (2.32) follows by taking the expectation on both sides.

To prove (2.33), we demonstrate that the right-hand side of (2.34) is sub-Weibull distributed. Since $\xi_i \sim subW(\theta, \omega_i)$, Proposition 3 implies that $b_i(\Delta_i + \eta_i \|\xi_i\|) \sim subW\left(\theta, b_i(\Delta_i + \eta_i \omega_i)\right)$. By summing over i, we obtain:

$$\sum_{i=0}^{t} b_i (\Delta_i + \eta_i \|\xi_i\|) \sim subW\left(\theta,\ \sum_{i=0}^{t} b_i (\Delta_i + \eta_i \omega_i)\right).$$

Denoting $S_t := a_t \|x_0 - \bar{x}_0\| + \sum_{i=0}^{t} b_i (\Delta_i + \eta_i \|\xi_i\|)$, we conclude that $S_t \sim subW\left(\theta, \upsilon_t\right)$, where $\zeta_t = \ a_t \|x_0 - \bar{x}_0\| + \sum_{i=0}^{t} b_i (\Delta_i + \eta_i \omega_i)$. From our sub-Weibull definition, we have that

$$\mathbb{P}(|\omega_t| \geq \varepsilon) \leq 2 \exp\left(-\frac{\theta}{2e}\left(\frac{\varepsilon}{\zeta_t}\right)^{\frac{1}{\theta}}\right). \tag{2.35}$$

Now let $\delta \in (0,1)$ be fixed and let ε be such that $\delta = 2\exp(-\theta(2e)^{-1}\varepsilon^{1/\theta}\zeta_t^{-1/\theta})$. Solving for ε yields $\varepsilon = \log^\theta\left(\frac{2}{\delta}\right)\left(\frac{2e}{\theta}\right)^\theta \zeta_t$. It follows that $S_t \leq \left(\frac{2e}{\theta}\right)^\theta \log^\theta\left(\frac{2}{\delta}\right) \zeta t$, with probability $1 - \delta$. Finally, (2.33) follows by substitution. □

The bound (2.32) generalizes the estimate in Lemma 10 by accounting for the gradient error. It is also worth pointing out that (2.32) and (2.33) have a similar structure; indeed, (2.33) differs only by a logarithmic factor and by the introduction of the tail parameters ω_i (which replaces the expectation term).

Remark 1. *An alternative high probability bound can be obtained by using* (2.32) *and Markov's inequality. For any $\delta \in (0,1)$, then Markov's inequality guarantees that:*

$$\|x_{t+1} - \bar{x}_{t+1}\| \leq \frac{1}{\delta}\left(a_t\|x_0 - \bar{x}_0\| + \sum_{i=1}^{t} b_i(\Delta_i + \eta_i \mathbb{E}\|\xi_i\|)\right), \tag{2.36}$$

with probability at least $1-\delta$. However, if we increase the confidence of the bound by allowing $\delta \to 0$, the right-hand-side of (2.36) *grows more rapidly than* (2.33). □

Note that the bounds in Theorem 12 are valid for any $t \in \mathbb{N}$. The asymptotic behavior is noted in the next remark.

Remark 2. *If* (2.27) *holds, then* $\limsup_{t\to+\infty} \|x_t - \bar{x}_t\| \leq (1-\widetilde{\lambda})^{-1}(\widetilde{\Delta} + \widetilde{\eta}\widetilde{\xi})$ ***almost surely****, where $\widetilde{\eta}$ and $\widetilde{\xi}$ are upper bounds on the step size and $\mathbb{E}\|\xi_t\|$; the proof is omitted because of space limits, but follows arguments similar to [5].* □

2.4 Application to Electric Vehicle Charging

This section illustrates the use of the proposed algorithms in an application inspired from [58], where the operator of a fleet of electric vehicles (EVs) seeks to determine an optimal charging policy in order to minimize its charging costs. The region of interest is modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node in $\mathcal{V}$ represents a charging station (or a group thereof), and an edge (i,j) in $\mathcal{E}$ allows vehicles to transfer from node i to j. We assume that the graph is strongly connected, so that EVs can be redirected from one node to any other node. We let $x_i \in \mathbb{R}_{>0}$ denote the energy requested by the fleet at node $i \in \mathcal{V}$. We assume that the net energy available is limited, and define the set $\mathcal{X}_t := \{x \in \mathbb{R}^d : \sum_{i\in\mathcal{V}} x_i \leq \mathcal{X}_t\}$, for a given $\mathcal{X}_t \in \mathbb{R}_{>0}$. Given $\{x_i\}$, the operator of the power grid strategically chooses a price per unit of energy so as to optimize its revenue from selling energy; we let $z_i \in \mathbb{R}_{>0}$ denote the selected price in region i, and we hypothesise that $z_i \sim \mathcal{N}(\mu_t x_i, \sigma_t^2)$, $\mu_{i,t}, \sigma_t \in \mathbb{R}_{>0}$ as an example. We note that, although the grid operator can choose the price arbitrarily large to maximize its revenue, large prices may compel the fleet operator to withdraw its demand, thus motivating the use of a model where the mean grows linearly with the

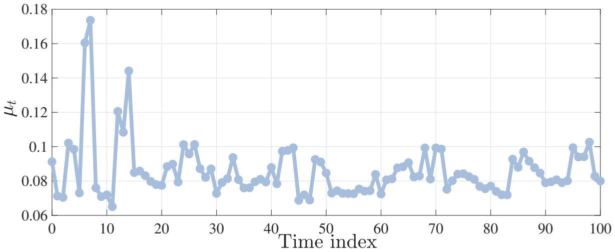


Figure 2.1: Time series data representing the price of energy in dollars per kilowatt hour (kWh). Each time step represents 5 minutes.

energy demand. Accordingly, we model the cost function of the EV operator as follows [58]:

$$f_t(x, z) = \sum_{i \in \mathcal{V}} z_i x_{i,t} - \gamma_{i,t} x_i + \kappa_{i,t} x_i^2, \tag{2.37}$$

where $\gamma_{i,t} \in \mathbb{R}_{>0}$, models the charging aggressiveness of the fleet operator, and $\kappa_{i,t} x_{i,t}^2$ quantifies the satisfaction the fleet operator achieves from consuming one unit of energy. In (2.37), the term $z_{i,t} x_{i,t}$ describes the charging cost at station i, the quantity $\gamma_{i,t} x_{i,t}$, and models the energy demand at the i-th station. Notice that, because the displacement of vehicles can change over time, we assume that the parameters $\gamma_{i,t}$ and $\xi_{i,t}$ are time dependent. We note that: (i) because of the capacity constraint $x_t \in \mathcal{X}_t$, the decision variables $x_{i,t}, i \in \mathcal{V}$, are coupled, and (ii) although the optimization could be solved in a distributed fashion since (2.37) is separable, our focus is to solve it in a centralized way since the EV operator is unique.

We apply the proposed methods to a system of 10 homogeneous charging stations over 100 time steps with fixed net energy ($\mathcal{X}_t = 10$). Namely, $\gamma_{i,t} = -1/100|t - 50| + 1$ and $\kappa_{i,t} = 2$ for $i \in \{1, \ldots, 10\}$. The charging cost distribution is informed by μ_t and σ_t; in our case, μ_t is the time series data of CAISO real-time prices deposited in Fig 2.1 (taken from `http://www.energyonline.com`) and $\sigma_t = 1$. Given these parameter values, the cost is γ_t-strongly convex and L_t-jointly smooth with $\gamma_t = L_t = 2$. Following the results in [29], the distributional maps are ν_t-sensitive with $\nu_t = \mu_t$. The sequence of performatively stable points are computed in closed form by solving the KKT equations.

For each experiment, we run online stochastic gradient descent and full-information online

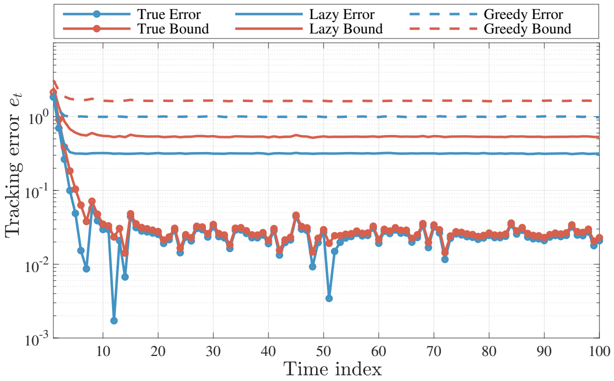


Figure 2.2: Online stochastic gradient descent with variable step-size compared to a full-information online gradient descent baseline.

gradient descent with fixed step size $\eta_t = 0.3$ by drawing initial state x_0 uniformly from a sphere of radius 5. We compute the mean tracking error for both single-sample and batch deployments. The mean tracking error for each is computed via Monte Carlo simulation using $1,000$ realizations of the initial state.

In Fig. 2.2, we demonstrate the error bound results in the previous section. Here “True” (i.e., true gradient) refers to the full-information case whereas “greedy” to the stochatic algirithm with $N_t = 1$, and “lazy” to the case where $N_t = 10$.

Chapter 3

Saddle Point Problems

We are interested in solving a stochastic saddle point problem where the data distribution shifts in response to decision variables. This feature yields the problem:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left\{ F(x,y) := \mathop{\mathbb{E}}_{z \sim D(x,y)} [f(x,y,z)] \right\}, \tag{3.1}$$

where $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ are compact constraint sets, $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is a scalar-valued function of the decision variables (x, y) parameterized by a random vector w, D is a distribution inducing map. Hereafter, we refer to F as the objective and the function f as the minimax function. We remark that the distribution of w depends on the decision variables (x, y). When solutions to the problem (3.1) exist, we will denote these solutions as (x^*, y^*).

For general distributional maps D, solving (3.1) directly is intractable. Indeed, F may be non-convex-non-concave even when f is strongly-convex-strongly-concave. Additionally, estimating the gradients $\nabla_x F$ and $\nabla_y F$ from samples requires differentiating the probability density of D. If we could do this freely, then the (3.1) amounts to a deterministic problem, which is well-studied in the literature on saddle-point problems. To proceed, we will assume that D us unknown and can only be queried to receive responses z.

A common heuristic when dealing with non-stationary data distributions is to recompute optimal decisions each time a new data distribution is revealed. For minimax problems, this corresponds to generating a sequence of decisions $\{(x_t, y_t)\}_{t\geq 0}$ such that:

$$\begin{aligned} x_{t+1} &\in \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathop{\mathbb{E}}_{z\sim D(x_t,y_t)}[f(x,y,z)] \\ y_{t+1} &\in \arg\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim D(x_t,y_t)}[f(x,y,z)]. \end{aligned} \tag{3.2}$$

We will refer to fixed points of this sequence as *stable points*. These can be seen as the generalization of the so-called performatively stable points in [23, 49, 64] in our stochastic minimax setup (3.1). A primary objective of this work is illustrating sufficient conditions for the existence and uniqueness of stable points. In particular, existence of the set stable points is shown when the minimax function is convex in x and concave in y for a given w, and under continuity of the distributional map. Building on these results, and focusing on strongly-convex-strongly-concave functions f, we then develop deterministic and stochastic projected primal-dual algorithms that can determine stable points.

However, as discussed in the paper, stable points and saddle points are qualitatively distinct. stable points are saddle points for the stationary problem that they induce, but need not be necessarily optimal. For this reason, we investigate a sufficient condition on the distributional map D that allows us to guarantee strong-convexity-strong-concavity of the objective F. We call this condition *opposing mixture dominance*, and provide a detailed example of a practical class of distributions that satisfy this assumption. Since gradient based algorithms will require us to have knowledge of the explicit dependence D has on the decision variables, we turn to zeroth order algorithms. We demonstrate that derivative-free algorithms with a single function evaluation are capable of approximating saddle points provided that F is strongly-convex-strongly-concave.

3.1 Stable Points

In our problem formulation in (3.1), we will assume that the sets $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ are convex and compact and that the data is supported on euclidean space, i.e., $z \in \mathbb{R}^k$. Let $\mathcal{P}(\mathbb{R}^k)$ be the set of finite-first moment probability measures supported on $\mathbb{R}^k$. Then the objective function can be written in integral form as $F(x,y) = \int_{\mathbb{R}^k} f(x,y,z)\mu_{(x,y)}(dz)$ where $\mu_{(x,y)} \in \mathcal{P}(\mathbb{R}^k)$ is given as the output of the distributional map D for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Classical solutions to this problem take the form of saddle points, as defined next.

Definition 6 (Saddle Points). *A pair $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a saddle point for the problem in* (3.1) *provided that $F(x^*, y) \leq F(x^*, y^*) \leq F(x, y^*)$, $\forall\, x \in \mathcal{X}, y \in \mathcal{Y}$.*

Sufficient conditions for the existence of saddle points consist of F being convex-concave while $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact [51, Ex. 11.52]. When minimax equality holds, we can equivalently characterize saddle points as a pair that satisfies:

$$x^* \in \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x, y), \qquad y^* \in \arg\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} F(x, y).$$

In practice, computing saddle points directly is computationally intractable. Namely, the dependence of the distributional map on the decision variables implies that even when f is convex-concave F may not be and hence saddle points will not even exist. Hence, we direct our attention to the fixed point of the repeated retraining heuristic in (3.2).

Definition 7 (Stable Points). *The point $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ is an stable point provided that*

$$\begin{aligned} \bar{x} &\in \arg\min_{x\in\mathcal{X}} \left\{ \max_{y\in\mathcal{Y}} \underset{z\sim D(\bar{x},\bar{y})}{\mathbb{E}} [f(x, y, z)] \right\}, \\ \bar{y} &\in \arg\max_{y\in\mathcal{Y}} \left\{ \min_{x\in\mathcal{X}} \underset{z\sim D(\bar{x},\bar{y})}{\mathbb{E}} [f(x, y, z)] \right\}. \end{aligned} \tag{3.3}$$

Intuitively, $(\bar{x}, \bar{y})$ are saddle points for the stationary saddle point problem induced by the distribution $D(\bar{x}, \bar{y})$. These are desirable as alternative solutions as they exist under mild convexity assumptions for problems with compact decision sets. Furthermore, we note that compactness here is not a limitation, as even unconstrained problems can be artificially constrained to a sufficiently large compact set without changing the solutions [36].

Our first objective in this work will be to provide conditions for the existence and uniqueness of these stable points. Later, we develop first order algorithms and demonstrate their convergence to stable points. In analysis, we will frequently refer to the the "decoupled objective", defined by

$$F(x, y|x', y') := \mathbb{E}_{z\sim D(x',y')}[f(x, y, z)] \tag{3.4}$$

for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$ as a means of separating the effects of f and D in the problem. To characterize stable points, we will consider the repeated retraining correspondence $H : \mathcal{X} \times \mathcal{Y} \to P(\mathcal{X} \times \mathcal{Y})$, defined by

$$H(x,y) := \left(\arg\min_{x' \in \mathcal{X}} \max_{y' \in \mathcal{Y}} F(x', y'|x, y), \ \ \arg\max_{y' \in \mathcal{Y}} \min_{x' \in \mathcal{X}} F(x', y'|x, y) \right) \tag{3.5}$$

which maps pairs in the product space to its power set $P(\mathcal{X} \times \mathcal{Y})$. In light of Definition 7, the stable points are fixed points of H; that is, $(\bar{x}, \bar{y}) \in H(\bar{x}, \bar{y})$.

For notional convenience, we will refer to the concatenated vector $w = (x, y) \in \mathbb{R}^{d_x + d_y}$ and its associated domain $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d_x + d_y}$ (consequently, we can identify $H(w)$ and $F(w, w')$ with the above functions whenever convenient). In the following section, we provide sufficient conditions for the existence of stable points.

3.1.1 Existence of Stable Points

Our goal is to demonstrate the existence and uniqueness of stable points. First, we demonstrate the existence of stable points by showing that the fixed point set of H, defined as $\text{Fix}(H) := \{w \in \mathcal{X} \times \mathcal{Y} |\ w \in H(w)\}$, is nonempty. The crux of our proof is showing that, under appropriate assumptions, H is an upper hemicontinous function. Next, we provide this definition as well as the notion of a topological neighborhood.

Definition 8 (Neighborhood, [3, Sec. 17.2])**.** *If A is a topological space and $x \in A$, then a neighborhood of x is a set $V \subset A$ such that there exists an open set U with $x \in U \subset V$. If the set V is open, then we say that V is an open neighborhood.*

Definition 9 (Upper Hemicontinuity,[3, Sec. 17.2])**.** *If A and B are two topological metric spaces, then a set valued function $\varphi : A \mapsto P(B)$ is upper hemicontinuous (uhc) at $x \in A$ provided that for every neighborhood U of $\varphi(x) \subset B$, the upper inverse set $\varphi^u(U) = \{x : \varphi(x) \subset U\}$ is a neighborhood of x. If φ is uhc at every x in A, then we say that φ is uhc on A.*

We next state our result for the existence of stable points.

Theorem 13 (Existence of Stable Points)**.** *Suppose that the following assumptions hold:*

i) $x \mapsto f(x, y, w)$ *is convex in* x *for all* $y \in \mathcal{Y}$ *and for all realizations of* z*;*

ii) $y \mapsto f(x, y, w)$ *is concave in* y *for all* $x \in \mathcal{X}$ *and for all realizations of* z*;*

iii) $(x, y) \mapsto f(x, y, z)$ *is continuous on* $\mathcal{X} \times \mathcal{Y}$ *for all* z*;*

iv) $\mathcal{X} \subset \mathbb{R}^{d_x}, \mathcal{Y} \subset \mathbb{R}^{d_y}$ *are convex compact subsets;*

v) the distributional map $D : \mathcal{Z} \to (\mathcal{P}(M), W_1)$ *is continuous.*

Then the fixed point set $Fix(H)$ *is nonempty and compact.*

Proof. The proof amounts to showing that H satisfies the hypotheses of Kakutani's Fixed Point Theorem [3, Corollary 17.55] for correspondences (set-valued functions). Since the domain $\mathcal{X} \times \mathcal{Y}$ is convex and compact by hypothesis, we show that H has a closed graph and non-empty convex and compact set values in $P(\mathcal{X} \times \mathcal{Y})$. Following the Closed Graph Theorem [3, Theorem 17.11], compactness of $\mathcal{X} \times \mathcal{Y}$ implies that H has closed graph if and only if it is closed valued and upper hemicontinous. Hence our proof reduces to showing that (i) H has non-empty closed values, (ii) H is upper hemicontinuous, and (iii) H has convex values.

Define the intermediate functions

$$f(x'|w) = \max_{y' \in \mathcal{Y}} F(x', y'|w) \quad \text{and} \quad g(y'|w) = \min_{x \in \mathcal{X}} F(x', y'|w) \tag{3.6}$$

as well as the realization functions

$$F(w) = \arg\min_{x' \in \mathcal{X}} f(x'|w) \quad \text{and} \quad G(w) = \arg\max_{y' \in \mathcal{Y}} g(y'|w). \tag{3.7}$$

for all $x' \in \mathcal{X}$, $y' \in \mathcal{Y}$, and $w \in \mathcal{X} \times \mathcal{Y}$. Using this convention, H can be written compactly as $H(w) = (F(w), G(w))$. It follows from continuity of f and D on $\mathcal{X} \times \mathcal{Y}$, as well as compactness of $\mathcal{X}$ and $\mathcal{Y}$ that f and g are continuous [3, Theorem 17.31]. The Maximum Theorem applied to F and G implies that F and G are upper hemicontinuous and have nonempty compact set values. Here, compactness implies closed-ness. Thus the values of H are closed since the Cartesian product of closed sets is closed. This proves (i).

To see that H is upper hemicontinuous, fix $w \in \mathcal{X} \times \mathcal{Y}$ and let U be an open set such that $H(z) \subset U$. Then H will be upper hemicontinuous provided that we can show that there exists an open neighborhood W of w such that $H(W) \subset U$. Given that $H(w)$ is a compact subset of U, [3, Theorem 2.62] guarantees the existence of open sets $V_x \subset \mathcal{X}$ and $V_y \subset \mathcal{Y}$ such that $H(z) \subset V_x \times V_y \subset U$. Since F and G are upper hemicontinuous, then the upper inverse sets $F^u(V_x) = \{w : F(w) \subset V_x\}$ and $G^u(V_y) = \{z : G(z) \subset V_y\}$ are open in $\mathcal{X} \times \mathcal{Y}$. Let $W = F^u(V_x) \cap G^u(V_y)$. Then $w \in W$ by construction, so $W, H(W) \neq \emptyset$. Furthermore, W is an open neighborhood of z and $H(W) \subset V_x \times V_y \subset U$. Thus condition (ii) holds.

Observe that since $x' \mapsto f(x'|w)$ is convex for all w and $\mathcal{X}$ is convex, then $F(w)$ is convex for all $w \in \mathcal{X} \times \mathcal{Y}$. Similarly, $G(w)$ is convex for all z. Since the Cartesian product of convex sets is convex, then condition (iii) follows. □

Recall that the intuition for the stable points is that they are the saddle points of the stationary saddle point problem that they induce. In this next results, we summarize this characterization.

Proposition 14. *Suppose that an stable point exists. Then $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ is an stable point if and only if*

$$F(\bar{x}, y|\bar{x}, \bar{y}) \leq F(\bar{x}, \bar{y}|\bar{x}, \bar{y}) \leq F(x, \bar{y}|\bar{x}, \bar{y}) \tag{3.8}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

We omit the proof as it is amounts to the same proof technique for the classical saddle point characterization result.

We will leverage the results of this section in the analysis of first-order methods that will be utilized to solve the stochastic minmax problem. In the following, we outline some working assumptions used in the algorithmic synthesis and analysis, and provide additional intermediate results.

3.1.2 Stable Points for Strongly Monotone Gradient Maps

In what follows, we outline relevant assumptions that we use in this paper for the synthesis and analysis of first-order deterministic and stochastic algorithms to identify stable points.

Assumption 8 (Strong-Convexity-Strong-Concavity)**.** *For any realization $z \in \mathbb{R}^k$, the function $w \mapsto f(w, z)$ is differentiable. The function f γ-strongly-convex-strongly-concave, for any realization of w; that is, f is γ-strongly-convex in x for all $y \in \mathbb{R}^{d_y}$ and γ-strongly-concave in y for all $x \in \mathbb{R}^{d_x}$.*

Assumption 9 (Joint Smoothness)**.** *The stochastic gradient map g given by $g(w, z) := (\nabla_x f(w, z), -\nabla_y f(w, z))$ is L-Lipschitz in w and z. Namely,*

$$\|g(w, z) - g(w', z)\| \leq L\|w - w'\|, \quad \|g(w, z) - g(w, z')\| \leq L\|z - z'\|.$$

for any $w, w' \in \mathbb{R}^{d_x+d_y}$ and z, z' supported on $\mathbb{R}^k$.

Assumption 10 (Distributional Sensitivity)**.** *The distributional map $D : \mathbb{R}^{d_x+d_y} \to \mathcal{P}(\mathbb{R}^k)$ is ν-Lipschitz. Namely,*

$$W_1(D(w), D(w')) \leq \nu\|w - w'\|$$

for any $w, w' \in \mathbb{R}^{d_x+d_y}$, where W_1 is the Wasserstein-1 distance.

Assumption 11 (Compact Convex Sets)**.** *The sets $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ are compact and convex.*

Typically, the assumption of strong-convexity-strong-concavity enables linear convergence to saddle-points in standard primal-dual methods [36]. Furthermore, strong-convexity-strong-concavity implies uniqueness of saddle point solutions; this allows to derive convergence results to the unique saddle-point in the static case, and tracking results in the context of time-varying minmax problems [18]. We also note that this assumption is useful in this paper in order to characterize the intrinsic relationship between optimal solutions to (3.1) and stable points. We also note that, for simplicity, the assumption imposes a common geometry parameter in f for both the x and y values; however, our analysis is the same for functions f being γ_1-strongly-convex in x and γ_2-strongly-concave in y (as we can take $\gamma = \min\{\gamma_1, \gamma_2\}$).

Assuming that the distributional map is ν-Lipschitz and the gradient is Lipschitz in the random variable is commonplace in the literature on decision-dependent distributions to characterize the overall effects of the distributional maps on the random variables [23, 49, 64]. Since we assume the support of our random variables w reside in a complete and separable metric space (Polish space), then a natural way to relate the resulting distributions is via the Wasserstein-1 metric. Following Kantorovich-Rubenstein Duality [9, 35], this metric can be written as

$$W_1(\mu, \nu) = \sup \left\{ \mathop{\mathbb{E}}_{z \sim \mu}[g(z)] - \mathop{\mathbb{E}}_{z \sim \nu}[g(z)] \;\middle|\; g : M \to \mathbb{R},\ \mathrm{Lip}(g) \leq 1 \right\}$$

for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^k)$. Here the supremum is taken over all Lipschitz-continuous functionals on $\mathbb{R}^k$ with Lipschitz constant less than or equal to one.

Closed and convex constraint sets are common in the literature on primal-dual methods, which are the main algorithms that will be considered shortly [30, 36]. Due to Heine-Borel, compactness of $\mathcal{X}$ and $\mathcal{Y}$ simply means closed and bounded. The addition of boundedness here is not restrictive; one can assume boundedness while the underlying sets can still be made arbitrarily large to include the saddle-points. As an illustration, consider the closed rectangles $\mathcal{X} = [-r, r]^{d_x}$ and $\mathcal{Y} = [-r, r]^{d_y}$ for some $r > 0$. Then $\mathcal{X}$ and $\mathcal{Y}$ are compact and convex for any $r > 0$, and r can be made an arbitrarily large positive number. See, e.g., [36] for an example in the context of constrained optimization problems.

To proceed, we cast the stable point problem into the variational inequality framework. We show that the stable problem is equivalent to a variational inequality over $\mathcal{W} := \mathcal{X} \times \mathcal{Y}$, where we use the concatenated variable $w = (x, y)$ when convenient. We then demonstrate uniqueness of the stable points for saddle point problems that satisfy the above assumptions.

Recall that in Assumption 9, we introduce the stochastic gradient map g given by $g(w, z) = (\nabla_x f(w, z), -\nabla_y f(w, z))$. Using this convention, we denote the decoupled gradient map as

$$G(w|w') = \mathop{\mathbb{E}}_{z \sim D(w')}[g(w, z)]. \tag{3.9}$$

This motivates the following characterization, which highlights the fact that stable points are solutions to the decoupled gradient variational inequality.

Theorem 15 (Stable Variational Inequality)**.** *A point $\bar{w} \in \mathcal{W}$ is an stable point provided that*

$$\langle w - \bar{w}, G(\bar{w}|\bar{w}) \rangle \geq 0 \,, \quad \forall w \in \mathcal{W}. \tag{3.10}$$

Proof of this fact follows steps that are similar to the ones in [51, Example 12.50]. In light of Definition 7, this result suggest that $\bar{z}$ are solutions to variational inequality induced by the stationary distribution $D(\bar{w})$. In the following, we show that when $w \mapsto G(w|w')$ is strongly monotone for all $w' \in \mathbb{R}^{d_x+d_y}$, a unique stable point exists. Furthermore, under this assumption, we can show that the distance between the saddle points for the original problem in (3.1) and the unique stable point is bounded.

Proposition 16. *Suppose that Assumption 8 holds. Then, for any $z \in \mathbb{R}^k$, $z \mapsto g(w, z)$ is γ-strongly-monotone. Furthermore, for any $w' \in \mathbb{R}^{d_x+d_y}$, $w \mapsto G(w|w')$ is γ-strongly-monotone.*

Proof of this result is immediate. Below we provide a Lemma that allows us to characterize the changes in the distributional argument of the decoupled gradient map G. This amounts to the decoupled gradient map being Lipschitz continuous in the distributional argument.

Lemma 17 (Gradient Deviations)**.** *Suppose that Assumptions 9-11 hold. Then, for any $\widehat{w} \in \mathbb{R}^{d_x+d_y}$, the map $w \mapsto G(\widehat{w}|w)$ is νL-Lipschitz. Furthermore, the restriction to $\mathcal{W}$ is bounded in the following way: for any $\widehat{w} \in \mathcal{W}$*

$$\|G(\widehat{w}|w) - G(\widehat{w}|w')\| \leq \nu L D_{\mathcal{W}} \tag{3.11}$$

for all $w, w' \in \mathcal{W}$ where $D_{\mathcal{W}} = diam(\mathcal{W}) < \infty$.

Proof. Let $v \in \mathbb{R}^{d_x+d_y}$ be an arbitrary unit vector and fix $\widehat{w}, w, w' \in \mathbb{R}^{d_x+d_y}$. It follows that

$$\langle v, G(\widehat{w}|w) - G(\widehat{w}|w') \rangle = \mathop{\mathbb{E}}_{z \sim D(w)} [\langle v, g(\widehat{w}, z) \rangle] - \mathop{\mathbb{E}}_{w \sim D(z')} [\langle v, g(\widehat{w}, z) \rangle].$$

By our assumption, we have that $z \mapsto \langle v, g(\widehat{w}, z) \rangle$ is Lipschitz with constant $L\|v\|$. From Kantorivich and Rubenstein, we have that

$$\mathop{\mathbb{E}}_{z \sim D(w)} [\langle v, g(\widehat{w}, z) \rangle] - \mathop{\mathbb{E}}_{z \sim D(w')} [\langle v, g(\widehat{w}, z) \rangle] \leq L W_1(D(w), D(w')) \leq \nu L \|w - w'\|,$$

where that last inequality follows from ν-sensitivity of D. Thus we have that for any unit vector v, $\langle v, (G(\widehat{w}|w) - G(\widehat{w}|w')\rangle \leq \nu L\|w - w'\|$, Hence, choosing

$$v = \frac{(G(\widehat{w}|w) - G(\widehat{w}|w'))}{\|(G(\widehat{w}|w) - G(\widehat{w}|w'))\|}$$

yields the result. Lastly, by Assumption 11, $\mathcal{W}$ is compact and hence $\|w - w'\| \leq D_{\mathcal{W}} < \infty$ for any $w, w' \in \mathcal{W}$. Thus, Lemma 3.11 follows. □

In what follows, we demonstrate existence and uniqueness of stable points. Similar to the statement of existence, we show that H satisfies the Banach-Picard Fixed Point Theorem by providing conditions for which H is a strict contraction.

Theorem 18 (Existence and Uniqueness of Stable Points)**.** *Suppose that Assumptions 8-11 hold. Then:*

(1) For all $w, w' \in \mathcal{W}$, $\|H(w) - H(w')\| \leq \frac{\nu L}{\gamma}\|w - w'\|$,

(2) If $\frac{\nu L}{\gamma} < 1$, then there exists a unique stable point $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$.

Proof. Let $\widehat{w}, \widetilde{w} \in \mathcal{W}$ be fixed. Then the maps $w \mapsto G(w|\widehat{w})$ and $w \mapsto G(w|\widetilde{w})$ are γ-strongly-strongly monotone. Furthermore, our strong-convexity and strong-concavity assumptions on f imply that $H(\widehat{w})$ and $H(\widetilde{w})$ and are single valued in $\mathcal{X} \times \mathcal{Y}$. Recall from our definition of H that $H(\widehat{w})$ and $H(\widetilde{w})$ are solutions to the variational inequalities induced by $\widehat{w}$ and $\widetilde{w}$ respectively. That is, for all $z \in \mathcal{X} \times \mathcal{Y}$,

$$\langle w - H(\widehat{w}), G(H(\widehat{w})|\widehat{w})\rangle \geq 0 \quad \text{and} \quad \langle w - H(\widetilde{w}), G(H(\widetilde{w})|\widetilde{w})\rangle \geq 0. \tag{3.12}$$

It follows from strong monotonicity that $\langle H(\widehat{w}) - H(\widetilde{w}), G(H(\widehat{w})|\widehat{w}) - G(H(\widetilde{w})|\widehat{w})\rangle \geq \gamma\|H(\widehat{w}) - H(\widetilde{w})\|^2$, and Theorem 3.12 imply that $\langle H(\widetilde{w}) - H(\widehat{w}), G(H(\widehat{w})|\widehat{w})\rangle \geq 0$. Hence,

$$\langle H(\widehat{w}) - H(\widetilde{w})), G(H(\widetilde{w})|\widehat{w})\rangle \leq -\gamma\|H(\widehat{w}) - H(\widetilde{w})\|^2. \tag{3.13}$$

To proceed, we provide a lower bound for the quantity on the left-hand side. By applying Cauchy-Schwartz and Lemma 17, we get that

$$\langle H(\widehat{w}) - H(\widetilde{w}), G(H(\widetilde{w})|\widetilde{w}) - G(H(\widetilde{w})|\widehat{w})\rangle \leq \nu L\|H(\widehat{w}) - H(\widetilde{w})\|\|\widetilde{w} - \widehat{w}\|.$$

Since Theorem 3.12 implies that $\langle H(\widehat{w}) - H(\widetilde{w}), G(H(\widetilde{w}|\widetilde{w})\rangle \geq 0$, then we get that

$$\langle H(\widehat{w}) - H(\widetilde{w}), G(H(\widetilde{w}|\widehat{w})\rangle \geq -\nu L \|H(\widehat{w}) - H(\widetilde{w})\| \|\widetilde{w} - \widehat{w}\|. \tag{3.14}$$

Combining inequalities Propositions 3.13 and 3.14 yields

$$-\gamma \|H(\widehat{w}) - H(\widetilde{w})\|^2 \geq -\nu L \|H(\widehat{w}) - H(\widetilde{w})\| \|\widetilde{w} - \widehat{w}\|,$$

and simplifying yields the result.

Since H is Lipschitz continuous, it is a strict contraction if $\nu L/\gamma < 1$. Uniqueness of the fixed point follows from the Banach-Picard Fixed-Point Theorem. □

We have demonstrated existence and uniqueness of stable points for some classes of problems; next, we characterize the relationship between stable points and solutions of the original problem in (3.1). First, an important observation is that when $\nu = 0$, the problem statement in (3.1) has a stationary probability distribution with respect to the decisions. Hence, saddle points coincide with stable points. When $\nu > 0$, we provide a guarantee on the distance between solutions of the two problems.

Proposition 19 (Bounded Distance)**.** *Suppose that Assumptions8-11 hold. Let w^* be the optimal solution of* (3.1)*, and let $\bar{w}$ be the stable point. Then,*

$$\|w^* - \bar{w}\| \leq \frac{\nu L}{\gamma} D_{\mathcal{W}}. \tag{3.15}$$

Proof. From the optimality conditions, we have that the decoupled gradient map satisfies $\langle \bar{w} - w^*, G(w^*|w^*)\rangle \geq 0$ and $\langle w^* - \bar{w}, G(\bar{w}|\bar{w})\rangle \geq 0$. By combining these results with results with our gradient deviation bound in Lemma 17, we obtain the following:

$$\begin{aligned}
\langle \bar{w} - w^*, G(\bar{w}|\bar{w}) - G(w^*|\bar{w})\rangle &= \langle \bar{w} - w^*, G(\bar{w}|\bar{w})\rangle - \langle \bar{w} - w^*, G(w^*|\bar{w})\rangle \\
&\leq \langle \bar{w} - w^*, G(w^*|w^*) - G(w^*|\bar{w})\rangle \\
&\leq \|\bar{w} - w^*\| \; \|G(w^*|w^*) - G(w^*|\bar{w})\| \\
&\leq \nu L D_{\mathcal{W}} \|\bar{w} - w^*\|,
\end{aligned}$$

where the second to last step follows from the Cauchy-Schwartz inequality. It follows from γ-strong-monotonicity that

$$\gamma\|\bar{w} - w^*\|^2 \leq \langle \bar{w} - w^*, G(\bar{w}|\bar{w}) - G(w^*|\bar{w})\rangle \leq \nu L D_{\mathcal{W}}\|\bar{w} - w^*\|$$

so that canceling terms and dividing by γ yields the result. □

3.1.3 Finding the Stable oint via Primal-Dual Algorithm

In this section, we discuss a primal-dual method for finding the stable points and demonstrate its linear convergence for strongly-convex-strongly-concave f. After choosing a starting point $w_0 \in \mathbb{R}^{d_x+d_y}$, we proceed by using the iterative update

$$w_{t+1} = \mathsf{proj}_{\mathcal{W}}\left(w_t - \eta G(w_t|w_t)\right) \tag{3.16}$$

where $\eta > 0$ is a positive step size. A key feature of this method is that each step projects onto the constraint sets, and hence $w_t \in \mathcal{W}$ for all $t \geq 1$ for any initial condition w_0. Observe that the update above can be expressed using the algorithmic map $\mathcal{A} : \mathcal{W} \times \mathcal{W} \to \mathcal{W}$ given by

$$\mathcal{A}(w|w') = \mathsf{proj}_{\mathcal{W}}\left(w - \eta G(w|w')\right), \tag{3.17}$$

for all $w, w' \in \mathbb{R}^{d_x+d_y}$. In the following result, we demonstrate that stable points are fixed points of $w \mapsto \mathcal{A}(w|w)$ over $\mathcal{W}$.

Proposition 20 (Fixed Point Characterization). *Let Assumptions 8-11 hold and suppose that* $\frac{\nu L}{\gamma} < 1$. *A point* $\bar{w} \in \mathcal{W}$ *is an stable point if and only if* $\bar{w} = \mathcal{A}(\bar{w}|\bar{w})$.

Proof. We want to show that $\bar{w}$ solving the variational inequality in (3.10) is equivalent to being a fixed point of the algorithmic map $\mathcal{A}$ over $\mathcal{W}$. From [26, Theorem 1.5.5], for any $\widehat{w} \in \mathbb{R}^{d_x+d_y}$, $\mathsf{proj}_{\mathcal{W}}(\widehat{w})$ is the unique element of $\mathcal{W}$ such that

$$\langle w - \mathsf{proj}_{\mathcal{W}}(\widehat{w}), \mathsf{proj}_{\mathcal{W}}(\widehat{w}) - \widehat{w}\rangle \geq 0 \tag{3.18}$$

holds for any $w \in \mathcal{W}$. As for the forward direction, if $\bar{w}$ as stable point, then (3.10) is equivalent to

$$\langle w - \bar{w}, \bar{w} - (\bar{w} - \eta G(\bar{w}|\bar{w}))\ \rangle \geq 0. \tag{3.19}$$

In setting $\widehat{w} = \bar{w} - \eta G(\bar{w}|\bar{w})$ in Proposition 3.18, we get that $\bar{w} = \mathsf{proj}_{\mathcal{W}}(\bar{w} - \eta G(\bar{w}|\bar{w}))$. Conversely, if $\bar{w}$ is such that $\bar{w} = \mathsf{proj}_{\mathcal{W}}(\bar{w} - \eta G(\bar{w}|\bar{w}))$, then by substituting $\widehat{w} = \bar{w} - \eta G(\bar{w})$ into (3.19), we have that $\bar{w}$ satisfies (3.10). □

To proceed, we will use a fixed point analysis to demonstrate convergence to the unique stable point.

Theorem 21 (Primal-Dual Convergence)**.** *Suppose that Assumptions 8-11 hold and that $\frac{\nu L}{\gamma} < 1$. Then the sequence $w_{t+1} = \mathcal{A}(w_t|w_t)$ satisfies the bound*

$$\|w_t - \bar{w}\| \leq \alpha^t \|w_0 - \bar{w}\| \tag{3.20}$$

for any initial point $w_0 \in \mathcal{W}$, and for $\alpha := \sqrt{1 - 2\eta\gamma + \eta^2 L^2} + \eta\nu L$. Furthermore, if

$$\eta \in \left(0, \frac{2(\gamma - \nu L)}{L^2(1 - \nu^2)}\right), \tag{3.21}$$

then z_t converges linearly to the unique stable point $\bar{w}$.

Proof. By applying our fixed point result in Propisition 20 and the triangle inequality, we get that

$$\|w_{t+1} - \bar{w}\| = \|\mathcal{A}(w_t|w_t) - \mathcal{A}(\bar{w}|\bar{w})\| \leq \|\mathcal{A}(w_t|w_t) - \mathcal{A}(w_t|\bar{w})\| + \|\mathcal{A}(w_t|\bar{w}) - \mathcal{A}(\bar{w}|\bar{w})\|. \tag{3.22}$$

Bounding the first quantity amounts to applying our gradient deviation result in Lemma 17. Hence,

$$\|\mathcal{A}(w_t|w_t) - \mathcal{A}(w_t|\bar{w})\| \leq \eta\|G(w_t|w_t) - G(w_t|\bar{w})\| \leq \eta\nu L\|w_t - \bar{w}\|.$$

The second quantity is the standard analysis for stationary primal-dual. Namely,

$$\begin{aligned}\|\mathcal{A}(w_t|\bar{w}) - \mathcal{A}(\bar{w}|\bar{w}))\|^2 &\leq \|(w_t - \bar{w}) - \eta(G(w_t|\bar{w}) - G(\bar{w}|\bar{w})\|^2 \\ &= \|w_t - \bar{w}\|^2 + \eta^2\|(G(w_t|\bar{w}) - G(\bar{w}|\bar{w})\|^2 - 2\eta\langle w_t - \bar{w}, G(w_t|\bar{w}) - G(\bar{w}|\bar{w})\rangle \\ &\leq (1 - 2\eta\gamma + \eta^2 L^2)\|w_t - \bar{w}\|^2.\end{aligned}$$

hence adding yields $\|w_{t+1} - \bar{w}\| \leq (1 - 2\eta\gamma + \eta^2 L^2)\|w_t - \bar{w}\|$ so that repeated application yields the result in (3.20). Convergence requires choosing step-size $\eta > 0$ such that $0 < \alpha < 1$. We observe that if $0 < \eta < 2\gamma/L^2$, then the quantity $\sqrt{1 - 2\eta\gamma + \eta^2 L^2}$ is real-valued.

Additionally, we find that $\alpha < 1$ provided that $0 < \eta\left(\eta L^2(\nu^2-1) - 2(\nu L - \gamma)\right)$ and hence we must have that $\eta L^2(\nu^2-1) - 2(\nu L - \gamma) > 0$. Finally, we note that

$$\frac{2(\gamma - \nu L)}{L^2(1-\nu^2)} \leq \frac{2\gamma}{L^2},$$

thus the result follows. □

In the next section, we focus on a stochastic algorithm. This stochastic method operates as an inexact version of the primal-dual algorithm, a fact which we highlight in our results.

3.1.4 Stochastic Primal-Dual Method

In the previous section, we showed convergence of the deterministic primal-dual algorithm for finding stable points in the full-information setting. Clearly this setting is unrealistic as it requires knowledge of the entire distributional map, for which we need not even appeal to stable points in the first place. This section will build on this analysis by investigating the stochastic primal update

$$w_{t+1} = \mathsf{proj}_{\mathcal{W}}\left(w_t - \eta g_t\right), \tag{3.23}$$

for gradient estimator g_t. A critical assumption we make is the existence of an oracle that provides unbiased estimators g_t for $G(x_t|x_t)$ at each iteration of the algorithm. In the simplest case, this merely implies that we can observe feedback $z_t \sim D_t(w_t)$ after deploying x_t into the relevant system and waiting for a response. Upon receiving z_t, we compute $g(w_t, z_t)$ and use this as the direction in our primal-dual algorithm.

3.1.4.1 Constant step-size

To build on this concept, we do not require this exact estimator be used; instead we simply assume that the estimator be unbiased and that tails of the estimator can be described by a sub-Weibull distribution.

In the following, we formally outline the stochastic framework that will be used in our analysis.

Assumption 12 (Sub-Weibull Framework)**.** *Denote* $\mathbb{E}_t = \mathbb{E}_{z\sim D(w_t)}$*. We assume the existence of an Oracle that will provide gradient estimator* g_t *for* $G(x_t|x_t)$ *at each iteration* $t \geq 0$ *such that the stochastic gradient error sequence* ξ_t *given by*

$$\xi_t = g_t - G(x_t|x_t) \tag{3.24}$$

satisfies the following properties for all $t \geq 0$*:*

(1) ***(Unbiased)*** *The gradient estimator* g_t *is an unbiased estimator for* G_t *in the sense that* $\mathbb{E}_t\xi_t = 0$*.*

(2) ***(Sub-Weibull Error)*** *The stochastic gradient error* ξ_t *is sub-Weibull in norm in the sense that* $\|\xi_t\| \sim subW(\theta, \omega_t)$ *for some* $\omega_t > 0$ *and that* $\omega = \sup_t \omega_t < \infty$*. Hence,*

$$\mathbb{P}\bigg(\|\xi_t\| \geq \varepsilon\bigg) \leq \exp\left(-\left(\frac{\varepsilon}{\omega_t}\right)^{1/\theta}\right). \tag{3.25}$$

for all $\varepsilon \geq 0$*.*

A typical assumption in the literature is that $\|\xi_t\|^2$, and hence the trace of the covariance of ξ_t, is uniformly bounded in expectation. Here we assume that the norm of the gradient error is distributed according to a heavy-tailed distribution. Note that, due to Proposition 1, $\omega = \sup_t \omega_t < \infty$ recovers this uniform boundedness property. By assuming θ is fixed for all t, we assume that all realizations of the process belong to the same sub-Weibull class.

Theorem 22. *Suppose that Assumptions 8-11 hold and* $\frac{\nu L}{\gamma} < 1$*. If Assumption 12 holds, and* η *satisfies the bound in* (3.21) *then the sequence* $\{w_t\}_{t\geq 0}$ *generated by* (3.16) *satisfies:*

(1) ***(Expectation)****. For any* $t \geq 0$*,*

$$\mathbb{E}\|w_t - \bar{w}\| \leq \rho^t\|w_0 - \bar{w}\| + \frac{\omega\eta}{1-\rho}. \tag{3.26}$$

(2) ***(High Probability)****. For any* $\delta \in (0, 1)$*, and* $t \geq 0$*,*

$$\mathbb{P}\left(\|w_t - \bar{w}\| \leq \rho^t\|w_0 - \bar{w}\| + c(\theta)\log^\theta\left(\frac{2}{\delta}\right)\frac{\omega\eta}{1-\rho}\right) \geq 1 - \delta \tag{3.27}$$

with $c(\theta) := \left(\frac{2e}{\theta}\right)^\theta$*.*

Proof. Observe that

$$\|w_{t+1} - \bar{w}\| = \|w_{t+1} - \mathcal{A}(\bar{w}|\bar{w})\| \leq \|w_{t+1} - \mathcal{A}(w_t|w_t)\| + \|\mathcal{A}(w_t|w_t) - \mathcal{A}(\bar{w}|\bar{w})\|. \tag{3.28}$$

To bound the first quantity, Applying non-expansiveness and Assumption12 yields

$$\|w_{t+1} - \mathcal{A}(w_t|w_t)\| \leq \eta\|g_t - G(w_t|w_t)\| = \eta\|\xi_t\|.$$

It follows from Theorem 21 that $\|\mathcal{A}(w_t|w_t) - \mathcal{A}(\bar{w}|\bar{w})\| \leq \rho\|w_t - \bar{w}\|$ so that combining and taking the conditional expectation yields

$$\mathbb{E}_t\|w_{t+1} - \bar{w}\| \leq \rho\|w_t - \bar{w}\| + \eta\mathbb{E}_t\|\xi_t\| \leq \rho\|w_t - \bar{w}\| + \eta\omega.$$

By bounding the finite geometric series $\sum_{i=0}^t \rho^i$ by its limit $(1-\rho)^{-1}$ and applying the law of total expectation to the recursion, we obtain $\mathbb{E}[e_t] \leq \rho^t e_0 + \eta\omega(1-\rho)^{-1}$.

For notational convenience, in what follows we will denote the error sequence as $e_t := \|w_t - \bar{w}\|$. From the above, we have that

$$e_{t+1} \leq \rho^{t+1} e_0 + \eta \sum_{i=0}^{t-1} \rho^i \|\xi_{t-i}\|. \tag{3.29}$$

From our sub-Weibull assumption, we have that $\|\xi_{t-i}\| \sim subW(\theta, \omega_{t-i})$. By applying the additive closure property in Proposition 3, we find that $S_t := \sum_{i=0}^t \rho^i \|\xi_{t-i}\| \sim subW(\theta, \sum_{i=1}^t \rho^i \omega_{t-i})$. Furthermore, closure again implies that $S_t \sim subW(\theta, \omega\eta(1-\rho)^{-1})$. Hence,

$$\mathbb{P}\left(S_t \geq \varepsilon\right) \leq 2\exp\left(-\frac{\theta}{2e}\left(\frac{(1-\rho)\varepsilon}{\omega\eta}\right)^{\frac{1}{\theta}}\right). \tag{3.30}$$

By setting the right-hand side equal to δ, we find that $\varepsilon = (\frac{2e}{\theta})^\theta \log^\theta\left(\frac{2}{\delta}\right)\omega\eta(1-\rho)^{-1}$. Now, observe that our stochastic recursion implies that for any $a > 0$, $\mathbb{P}(C_t + S_t \geq a) \geq \mathbb{P}(e_t \geq a)$. It follows that setting $a = C_t + \varepsilon$ yields

$$\mathbb{P}(e_t \leq C_t + \varepsilon) \geq \mathbb{P}(C_t + S_t \leq C_t + \varepsilon) = \mathbb{P}(S_t \leq \varepsilon) \geq 1 - \delta,$$

thus the result follows by substituting the expression for C_t and ε. □

The bounds naturally translate to convergence results by considering the limit supremum. Now we demonstrate that the algorithm converges to a neighborhood of the the stable in expectation and almost surely.

Theorem 23 (Neighborhood Convergence)**.** *Suppose that Assumptions 8-11 hold, and that g_t satisfies Assumption 12. Assume that η satisfies the condition of* (3.21)*. Then, the sequence of iterates $\{w_t\}_{t\geq 0}$ converges to a neighborhood of $\bar{w}$ in expectation and almost surely. In particular,*

$$\limsup_{t\to\infty} \mathbb{E}\|w_t - \bar{w}\| \leq \frac{\eta\omega}{1-\rho}, \quad \textit{and} \quad \mathbb{P}\left(\limsup_{t\to\infty} \|w_t - \bar{w}\| \leq \frac{\eta\omega}{1-\rho}\right) = 1.$$

Proof. The limit of the expectation follows immediately from above. As for almost sure convergence, we simply apply the Borel-Cantelli Lemma. As before we let $e_t = \|w_t - \bar{w}\|$, so that the result in Theorem 3.26 can be compactly written as $\mathbb{E}[e_t] \leq \rho^t e_0 + \eta\omega(1-\rho)^{-1}$. Denote $E_t = \max\{0, e_t\}$ so that $\mathbb{E}[E_t] \leq \rho^t e_0$.

By Markov's inequality, $P(E_t \leq \varepsilon) \leq \frac{\mathbb{E}[E_t]}{\varepsilon} \leq \frac{\rho^{t+1} e_0}{\varepsilon}$, for any $\varepsilon > 0$. Summing over t yields $\sum_{t=0}^{\infty} P(E_t \geq \varepsilon) \leq \frac{e_0}{\varepsilon(1-\rho)} < \infty$. It follows from the Borel-Cantelli Lemma that, since the sum of tail probabilities is finite, then $P(\limsup_{t\to\infty} E_t \leq \varepsilon) = 1$. Since this is true for any $\varepsilon > 0$, then the result follows. □

Notice that Theorem 22 requires only our heavy tail assumption and not a filtration that is standard with a $\|w_t - \bar{w}\|^2$ analysis. A drawback to this first-moment analysis, however, is that it only demonstrates convergence to a neighborhood whose radius is dictated by the proxy variance, and hence the quality of the estimator.

3.1.4.2 Decaying step-size

In what follows, we part with our heavy-tail assumption and demonstrate that we are able to obtain stronger convergence results at the expense of requiring our estimator to be unbiased and introducing a filtration on the probability space.

To do so, we will additionally require that our stochastic primal-dual algorithm use a decaying step-size instead of a fixed one. Hence, we will use an update of the form

$$w_{t+1} = \mathsf{proj}_{\mathcal{W}}\left[w_t - \eta_t g_t\right]. \tag{3.31}$$

In the following, we state an alternative stochastic framework to Assumption 12.

Assumption 13 (Filtration Framework)**.** *We assume the existence of an Oracle that will provide gradient estimator g_t for $G(x_t|x_t)$ at each iteration $t \geq 0$. Let $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ with elements*

$$\mathcal{F}_t := \sigma(g_\tau,\ \tau \leq t) \tag{3.32}$$

be the natural filtration of the Borel σ-algebra over $\mathbb{R}^{d_x+d_y}$ with respect to G_t. Let $\mathbb{E}_t[\ \cdot\ ] = \mathbb{E}[\cdot|\mathcal{F}_t]$ denote the conditional expectation with respect $\mathcal{F}_t$ over distribution $D(w_t)$ and $\xi_t := g_t - G_t(w_t|w_t$ denote the stochastic gradient error.

We assume that the stochastic gradient oracle returns estimators g_t satisfying the following properties:

*(1) **(Measurable)** For all $t \geq 0$, ξ_t is $\mathcal{F}_{t+1}$-measurable.*

*(2) **(Unbiased)** For all $t \geq 0$, $\mathbb{E}_t\xi_t = 0$.*

*(3) **(Bounded Variance)** There exists $\sigma > 0$ such that $\mathbb{E}_t\|\xi_t\|^2 \leq \sigma^2$, for all $t \geq 0$.*

Theorem 24 (Convergence)**.** *Suppose that Assumptions 8-11 and Assumption 13 hold. Then the sequence $\{w_t\}_{t\geq 0}$ given by $w_{t+1} = \mathsf{proj}_{\mathcal{W}_t}\left[w_t - \eta_t g_t\right]$ satisfies the following:*

*(1) **One Step Bound**. For all $t \geq 0$,*

$$\mathbb{E}_t\|w_{t+1} - \bar{w}\|^2 \leq \left(1 - 2(\gamma - \nu L)\eta_t + 2(1+\nu)^2 L^2 \eta_t^2\right)\|w_t - \bar{w}\|^2 + \eta_t^2\sigma^2.$$

*(2) **Convergence**. If $\eta_t = \ell(r+t)^{-1}$ where*

$$\ell > \frac{1}{2(\gamma - \nu L)} \quad \text{and} \quad r > \frac{(1+\nu)^2 L^2}{(\gamma - \nu L)^2} \tag{3.33}$$

then,

$$\mathbb{E}\|w_t - \bar{w}\|^2 \leq \frac{M}{r+t}, \text{ where } M := \max\left\{r\|w_0 - \bar{w}\|^2, \frac{\ell^2\sigma^2}{2(\gamma - \nu L)\ell - 1}\right\}. \tag{3.34}$$

Proof. By applying the algorithmic map, and using non-expansiveness of the projection operator we obtain the following relationship:

$$\mathbb{E}_t\|w_{t+1}-\bar{w}\|^2 \leq \|w_t-\bar{w}\|^2 - 2\eta_t\langle w_t-\bar{w}, G(w_t|w_t)-G(\bar{w}|\bar{w})\rangle + \eta_t^2\mathbb{E}_t\|g_t-G(\bar{w}|\bar{w})\|^2$$

To bound the inner product term, we use γ-strong-monotonicity and the Gradient Deviations result from Lemma 2.8:

$$\langle w_t-\bar{w}, G(w_t|w_t)-G(\bar{w}|\bar{w})\rangle \leq (\gamma-\nu L)\|w_t-\bar{w}\|^2. \tag{3.35}$$

From Young's inequality, we get that

$$\begin{aligned}\mathbb{E}_t\|g_t-G(\bar{w})\|^2 &= \mathbb{E}_t\|g_t-G(w_t|w_t)+G(w_t|w_t)-G(\bar{w}|\bar{w})\|^2\\ &\leq 2\mathbb{E}_t\|g_t-G(w_t|w_t)\|^2+2\mathbb{E}_t\|G(w_t|w_t)-G(\bar{w}|\bar{w})\|^2\\ &\leq 2\sigma^2+2(1+\nu)^2L^2\|w_t-\bar{w}\|^2,\end{aligned}$$

where the last inequality follows from the fact that $w\mapsto G(w|w)$ is $(1+\nu)L$-Lipschitz continuous. Combining yields the one step improvement bound.

To prove (b), we first the quadratic contraction parameter using convexity. Observe that $0<\eta_t\leq(\gamma-\nu L)(2(1+\nu)^2L^2)^{-1}$, implies that

$$1-2(\gamma-\nu L)\eta_t+2(1+\nu)^2L^2\eta_t^2 \leq 1-2(\gamma-\nu L)\eta_t.$$

Denoting $\rho=2(\gamma-\nu L)$, follows that

$$\mathbb{E}_t\|w_{t+1}-\bar{w}\|^2 \leq (1-\rho\eta_t)\|w_t-\bar{w}\|^2+\eta_t^2\sigma^2. \tag{3.36}$$

We proceed by induction. Clearly the bound in Theorem 3.34 holds for $t=0$. Supposing it holds for t, we have that

$$\begin{aligned}\mathbb{E}\|w_{t+1}-\bar{w}\| &\leq \left(1-\frac{\rho\ell}{r+t}\right)\frac{M}{r+t}+\frac{\sigma^2\ell^2}{(r+t)^2}\\ &\leq \frac{r+t-1}{(r+t)^2}M-\frac{\rho\ell-1}{(r+t)^2}M+\frac{\sigma^2\ell^2}{(r+t)^2}\\ &\leq \frac{r+t-1}{(r+t)^2}M\\ &\leq \frac{M}{(r+(t+1))^2},\end{aligned}$$

where the penultimate step follows from the fact that $(\rho\ell - 1)M + \sigma^2\ell^2 < 0$. □

This concludes our analysis of stable points. In the following section, we discuss how to compute saddle points.

3.2 Saddle Points and Mixture Dominance

By introducing the stable point problem, we have shifted the attention to a class of solutions that are less computationally burdensome to obtain while still serving as meaningful solutions within the context of decision-dependent stochastic problems. In this section, we demonstrate that finding saddle points is still possible for some well-behaved distributional maps. We consider a condition which we call *opposing mixture dominance*, and show that this condition is sufficient for guaranteeing existence of saddle points.

Assumption 14 (Opposing Mixture Dominance)**.** *For any $x, x', x_0 \in \mathbb{R}^{d_x}$, $y, y', y_0 \in \mathbb{R}^{d_y}$ and $\tau \in [0,1]$, the distributional map satisfies a convex shift in x*

$$\mathop{\mathbb{E}}_{z\sim D(\tau x+(1-\tau)x',y)}[f(x_0,y_0,z)] \le \mathop{\mathbb{E}}_{z\sim \tau D(x,y)+(1-\tau)D(x',y)}[f(x_0,y_0,z)],$$

and concave shift in y

$$\mathop{\mathbb{E}}_{z\sim \tau D(x,y)+(1-\tau)D(x,y')}[f(x_0,y_0,z)] \le \mathop{\mathbb{E}}_{z\sim D(x,\tau y+(1-\tau)y')}[f(x_0,y_0,z)].$$

As an example, we show that Bernoulli mixtures satisfies this assumption.

Example 5 (Bernoulli Mixtures)**.** *If the distributional map $D : \mathbb{R}^{d_x+d_y} \to \mathcal{P}(\mathbb{R}^k)$ is given by $D(x,y) =$ Bernoulli$(p(x,y))$ where $p : \mathbb{R}^{d_x+d_y} \to \mathbb{R}$ is the bilinear function*

$$p(x,y) = \langle x, Ay\rangle + \langle b, x\rangle + \langle c, y\rangle + d$$

then Assumption 14 is satisfied since $D(\tau x+(1-\tau)x', y) = \tau D(x,y)+(1-\tau)D(x',y)$ and $\tau D(x,y)+(1-\tau)D(x,y') = D(x, \tau y + (1-\tau)y')$.

Example 6. *(Location-Scale Families) A distributional map $D : \mathbb{R}^{d_x+d_y} \to \mathcal{P}(\mathbb{R}^k)$ induces a location-scale family provided that for any $z \in \mathbb{R}^{d_x+d_y}$, $z \sim D(w)$ if and only if $z \overset{d}{=} Az_0 + Bw + c$ where z_0 is some stationary zero-mean random variable. A sufficient condition for Assumption 14 to hold is that f is convex in the random variable z. A detailed proof of this fact is provided in the next section.*

In the previous section, we made the assumption that our random variables are supported on some general Polish space and are induced by a Radon probability measure parameterized by $w = (x, y) \in \mathbb{R}^{d_x+d_y}$. Here, we assume without loss of generality that the distributional map induces a probability density function $p(z|x, y)$ and write the objective as $F(x, y) = \int_{\mathbb{R}^k} f(x, y, z)p(z|x, y)dz$. The analysis that follows is identical for the case when the density $p(z|x, y)$ corresponds to discrete probability distribution parameterized by (x, y) and the proofs follow *mutatis mutandis.*

Below, we demonstrate that the opposing mixed dominance assumption is sufficient to guarantee that the objective is convex-concave in the distribution inducing arguments. The crux of this proof is observing that convex combinations of probability distributions have a density function defined by the convex combination of the underlying density functions.

Lemma 25. *Let Assumption14 hold. Then, for any $w_0 \in \mathbb{R}^{d_x+d_y}$, the function $(x, y) \mapsto \mathbb{E}_{z\sim D(x,y)}[f(w_0, z)]$ is convex-concave on $\mathbb{R}^{d_x+d_y}$.*

Proof. Fix $w_0 \in \mathcal{W}$, $x, x' \in \mathcal{X}$, and $y, y' \in \mathcal{Y}$ and let $\tau \in [0, 1]$. Observe that since the distribution $\tau D(x, y) + (1-\tau)D(x', y)$ is a convex mixture, then its probability density function is convex sum of the probability density functions for $D(x, y)$ and $D(x', y)$. That is, if p_τ is the density function for the convex mixture, and p_1 and p_2 are the density functions for $D(x, y)$ and $D(x', y)$, respectively, then $p_\tau(z) = \tau p_1(z) + (1 - \tau)p_2(z)$. From this, we conclude that

$$\mathop{\mathbb{E}}_{z\sim\tau D(x,y)+(1-\tau)D(x',y)}[f(w_0, z)] \le \tau \mathop{\mathbb{E}}_{z\sim D(x,y)}[f(w_0, z)] + (1 - \tau) \mathop{\mathbb{E}}_{z\sim\tau D(x',y)}[f(w_0, z)].$$

Combining this with Assumption 14, we get that

$$\mathop{\mathbb{E}}_{z\sim D(\tau x+(1-\tau)x',y)}[f(w_0, z)] \le \tau \mathop{\mathbb{E}}_{z\sim D(x,y)}[f(w_0, z)] + (1 - \tau) \mathop{\mathbb{E}}_{z\sim D(x',y)}[f(w_0, z)].$$

This proves convexity of $x \mapsto \mathbb{E}_{z\sim D(x,y)}[f(w_0, z)]$ for any y. The concavity in y can be shown using similar steps. □

We can then utilize this result in conjunction with our previous assumptions to get strong-convexity-strong-concavity of the objective F.

Theorem 26 (Strong-Convexity-Strong-Concavity)**.** *If Assumption 8-10 and Assumption 14 hold, then $(x, y) \mapsto F(x, y)$ is $(\gamma - 2\nu L)$-strongly-convex-strongly-concave over $\mathbb{R}^{d_x+d_y}$.*

Proof. We prove the assertion by first demonstration that strong-convexity holds in x for y fixed. Strong-concavity will follow similarly. By applying γ-strong-concavity of f in x, we get that

$$F(x', y|x'y) - F(x, y|x', y) \geq \langle x' - x, \mathop{\mathbb{E}}_{z\sim D(x',y)}[\nabla_x f(x, y, z)]\rangle + \frac{\gamma}{2}\|x - x'\|^2. \tag{3.37}$$

By the L-smoothness of the gradient, we get that

$$\langle x' - x, \mathop{\mathbb{E}}_{z\sim D(x,y)}[\nabla_x f(x, y, z)] - \mathop{\mathbb{E}}_{z\sim D(x',y)}[\nabla_x f(x, y, z)]\rangle \leq \nu L\|x - x'\|^2$$

which is equivalent to

$$0 \geq \langle x' - x, \mathop{\mathbb{E}}_{z\sim D(x,y)}[\nabla_x f(x, y, z)] - \mathop{\mathbb{E}}_{z\sim D(x',y)}[\nabla_x f(x, y, z)]\rangle - \frac{2\nu L}{2}\|x - x'\|^2. \tag{3.38}$$

Since for any $w_0 \in \mathbb{R}^{d_x+d_y}$ the function $(x, y) \mapsto \mathbb{E}_{z\sim D(x,y)}[f(w_0, z)]$ is convex-concave, we have that

$$F(x, y|x, y) - F(x, y|x', y) \geq \langle x - x', \mathop{\mathbb{E}}_{z\sim D(x',y)}[f(x, y, z)\nabla_x \log p(z|x, y)]\rangle \tag{3.39}$$

by setting $w_0 = (x, y)$. By adding inequalities (3.37)-(3.39) we obtain

$$F(x', y) - F(x, y) \geq \langle x' - x, \nabla_x F(x, y)\rangle + \frac{\gamma - 2\nu L}{2}\|x - x'\|^2,$$

which is equivalent to strong-convexity in x. Proof of strong-concavity in y follows similarly and it is omitted due to space limitations. □

3.2.1 Location-Scale Families

In this section, we are interested in solidifying the claims made in Example 6 on Location-scale families, which have seen much attention in the literature on decision-dependent distributions as it arises naturally in many common examples [45]. A formal definition is provided next.

Definition 10 (Location-Scale Family)**.** *The distributional map $D : \mathbb{R}^{d_x+d_y} \to \mathcal{P}(\mathbb{R}^k)$ forms a location-scale family provided that for every $z \in \mathbb{R}^{d_x+d_y}$ and $z \sim D(w)$, $z \overset{d}{=} Az_0 + Bw + c$ where $z_0 \sim D_0$. In this model, $D_0 \in \mathcal{P}(\mathbb{R}^k)$ is a zero-mean stationary distribution while $A_0 \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times(d_x+d_y)}$, and $c \in \mathbb{R}^k$ are model parameters.*

To demonstrate that Location-scale Families satisfy Assumption 14, we introduce the notion of convex stochastic orders. This is an ordering of random variables induced by convex functions.

Definition 11 (Convex Order, [52, Definition 7.A.1])**.** *If two k-dimensional random vectors z and u are such that $\mathbb{E}[f(u)] \leq \mathbb{E}[f(z)]$, for all convex functions $f : \mathbb{R}^k \to \mathbb{R}$, then we say that u is less than z in the convex order and write $u \leq_{cx} z$.*

Demonstrating an ordering from this definition alone proves difficult. Instead, we look to the following theorem that characterizes random variables in the convex stochastic order via couplings.

Theorem 27 ([52, Theorem 7.A.1])**.** *The random vectors $u \sim \mu$ and $z \sim \nu$ satisfy $u \leq_{cx} z$ if and only if there exists $\widehat{u} \overset{d}{=} u$ and $\widehat{z} \overset{d}{=} z$ such that $\mathbb{E}[\widehat{z}|\widehat{u}] = \widehat{u}$ a.s.*

Following this characterization, we demonstrate that location-scale families have a special relationship between the convex-combination family and the corresponding convex-mixture.

Lemma 28. *Let the distributional map $D : \mathbb{R}^{d_x+d_y} \to \mathcal{P}(\mathbb{R}^k)$ be a location scale family. Then for any $w, w' \in \mathbb{R}^{d_x+d_y}$ and $\tau \in [0, 1]$,*

$$\mathop{\mathbb{E}}_{z\sim D(\tau w+(1-\tau)w')}[f(z)] = \mathop{\mathbb{E}}_{z\sim \tau D(w)+(1-\tau)D(w')}[f(z)]$$

for any convex function $f : \mathbb{R}^k \to \mathbb{R}$.

Proof. Fix $\tau \in [0,1]$ and $w, w' \in \mathbb{R}^{d_x+d_y}$. In this proof, we use Theorem 27 to show that if $z \sim D(\tau w + (1-\tau)w')$ and $z' \sim \tau D(w) + (1-\tau)D(w')$, then we can define couplings that imply that $z \leq_{cx} z'$ and $z' \leq_{cx} z$. To this end, a key observation is that, if we denote the discrete random variable T as

$$T = \begin{cases} z \text{ w.p. } \tau, \\ \\ z' \text{ w.p } 1-\tau, \end{cases}$$

then $z' \sim \tau D(w) + (1-\tau)D(w')$ if and only if $z \overset{d}{=} Az_0 + BT + c$.

First, we suppose that $z \sim D(\tau w + (1-\tau)w')$. Then let $z' \overset{d}{=} w - B(\tau w + (1-\tau)w') + BT$. It follows that $\mathbb{E}[z'|z] = z$, and $z' \overset{d}{=} Az_0 + BT + c$. Hence $z' \sim \tau D(w) + (1-\tau)D(w')$. This proves that $z \leq_{cx} z'$.

Conversely, if we suppose that $z' \sim \tau D(w) + (1-\tau)D(w')$ and set $z \overset{d}{=} z' + B(\tau w + (1-\tau)w') - BT$ then $z' \leq_{cx} z$ follows. The statement follows from the definition of the convex order. □

Since this Lemma holds for any convex function f, it holds for stochastic payoff f provided that it is convex in w. This combined with the fact that Location-Scale Families are ν-Lipschitz with $\nu = \|B\|_2$ is sufficient for F to be strongly-convex-strongly-concave.

Theorem 29 (Strong-convexity-strong-concavity)**.** *Suppose that f satisfies Assumption 89, and the constraint sets $\mathcal{X}$ and $\mathcal{Y}$ satisfy Assumption 11. If D if a location-scale family and f is convex in w, then F is $(\gamma - 2\nu L)$- strongly-convex-strongly-concave.*

Proof. The proof amounts to demonstrating that D being a location-scale family and f being convex in z is sufficient to satisfy Assumptions 14-and 10. The result then follows by Theorem 26. We observe that Lemma 28 implies that Assumption 14 holds. As for D being ν- Lipschitz, we claim that $W_1(D(w), D(w')) \leq \|B\|_2 \|w - w'\|$. Then the Assumption holds with $\nu = \|B\|_2$. By definition,

$$W_1(D(w), D(w')) = \inf_{\Pi(D(w),D(w'))} \mathbb{E}_{(z,z') \sim \Pi(D(w),D(w'))} \|w - w'\|_2$$

where the infimum is taken over all couplings of the distributions $D(w)$ and $D(w')$. We find that if $z_0 \sim D_0$, then setting $z \overset{d}{=} Az_0 + Bw + c$ and $z' \overset{d}{=} Az_0 + Bw' + c$ implies that $z \sim D(w)$ and $z' \sim D(w')$ and $\|w - w'\| = \|B(w - w')\|$. Thus, the result follows. □

3.2.2 Derivative-Free Primal-Dual

In this section we study a derivative-free primal-dual algorithm for computing saddle points without eliciting distribution information from D. A unique feature of this algorithm is that it uses a stochastic gradient estimator with only a single cost function evaluation. This algorithm is suitable in the setting where opposing mixture dominance in Assumption 14 is known to hold, but a model for the distributional map is not available. The use of zeroth-order algorithms has been studied extensively within the context of derivative-free games in [11, 21]. For $d > 0$, we will denote $\mathbb{B}_d$ and $\mathbb{S}_d$ as the uniform distributions over the unit ball, $\mathcal{B}_d = \{x \in \mathbb{R}^d | \ \|x\| \leq 1\}$ and unit sphere, $\mathcal{S}_d = \{x \in \mathbb{R}^d | \ \|x\| = 1\}$, in $\mathbb{R}^d$ respectively. Additionally, denote $\mathbb{S}$ and $\mathbb{B}$ as joint distributions such that $v = (v_x, v_y) \sim \mathbb{B}$, $u = (u_x, u_y) \sim \mathbb{S}$ with $v_x \sim \mathbb{B}_{d_x}$, $v_y \sim \mathbb{B}_{d_y}$ and $u_x \sim \mathbb{S}_{d_x}$, $u_y \sim \mathbb{S}_{d_y}$. The derivative free algorithm map performs the update

$$w_{t+1} = \mathsf{proj}_{(1-\delta)\mathcal{W}} \left(w_t - \eta_t g_\delta(w_t, z_t) \right) \tag{3.40}$$

for $\eta_t > 0$, with zeroth-order gradient map

$$g_\delta(w, z) = \left(\frac{d_x}{\delta} f(w + \delta u_x, z) u_x, \ -\frac{d_y}{\delta} f(w + \delta u_y, z) u_y \right) \tag{3.41}$$

where $\delta > 0$, and $u = (u_x, u_y)$ with $u_x \sim \mathbb{S}_{d_x}$ and $u_y \sim \mathbb{S}_{d_y}$. Note that by projecting onto the restricted set $(1 - \delta)\mathcal{W}$ we retain feasibility throughout the iterations of the algorithm. Since we evaluating the stochastic objective at points perturbed by vectors on the unit sphere, we must introduce an additional assumption to ensure that the domain of our function is appropriate.

Assumption 15 (Boundedness)**.** *There exist positive radii $r, R > 0$ such that $\mathcal{W}$ satisfies $r\mathbb{B}_{d_x+d_y} \subseteq \mathcal{W} \subseteq R\mathbb{B}_{d_x+d_y}$.*

The gradient estimator in (3.41) naturally arises when considering the smoothed objective over the unit ball, given by

$$F_\delta(w) = \mathop{\mathbb{E}}_{v\sim\mathbb{B}}[F(w+\delta v)] = \mathop{\mathbb{E}}_{v\sim\mathbb{B}}\left[\mathop{\mathbb{E}}_{w\sim D(w+\delta v)}[f(w+\delta v, z)]\right] \tag{3.42}$$

and its associated gradient map $G_\delta(w) = (\nabla_x F_\delta(w), -\nabla_y F_\delta(w))$. These together form the perturbed saddle point problem

$$\min_{x\in(1-\delta)\mathcal{X}} \max_{y\in(1-\delta)\mathcal{Y}} F_\delta(x,y), \tag{3.43}$$

whose solutions we will we denote $w_\delta^* = (x_\delta^*, y_\delta^*)$. It follows that g_δ is an unbiased estimator of this gradient map, and hence it will allow us to find saddle points without requiring more information about the objective or distributional map. We formalize this in the following.

Lemma 30 (Unbiasedness, [11, Lemma C.1]). *If $\delta > 0$, then $\mathbb{E}_{u\sim\mathbb{S}}[\mathbb{E}_{z\sim D(w)} g_\delta(w,z)] = G_\delta(w)$, for all $w \in \mathbb{R}^{d_x+d_y}$.*

The the fact that we can estimate the gradient map using only a single function evaluation is an attractive feature of (3.40). There are alternatives multi-point estimators that use more function evaluations, but since the expectation in our problem also depends on the decision variables, they are biased. Furthermore, in the following we show that the considered perturbed gradient map retains strong-monotonicity.

Lemma 31 (Strong Monotonicity). *If the gradient of the objective F, given by $G(z) = (\nabla_x F(w), -\nabla_y F(w))$ is $(\gamma - 2\nu L)$-strongly-monotone, then G_δ is $(\gamma - 2\nu L)$-strongly-monotone for any $\delta > 0$.*

Indeed, by perturbing the objective and the constraint set by δ, the solution of the perturbed saddle point problem will may *may* be different from the solutions of the original problem. In the following, we bound the discrepancy between solutions.

Lemma 32 (Bounded Approximation). *If $\delta < r$, and G is $(\gamma - \nu L)$-strongly monotone, then*

$$\|w^* - w_\delta^*\| \le \delta\left(\left(1 + \frac{\sqrt{2L}}{(\gamma - 2\nu L)}\right)\|w^*\| + \frac{2L}{(\gamma - 2\nu L)}\right). \tag{3.44}$$

Finally, we are ready to demonstrate the performance of the algorithm. Here we impose two additional restrictions: (i) δ may not exceed the radius of the largest ball completely contained in $\mathcal{W}$, which we denoted as r; (ii) the map $(w, z) \mapsto f(w, z)$ is bounded over $\mathcal{W} \times \mathbb{R}^k$.

Theorem 33 (Convergence to the δ-Solution). *Suppose that $\delta \leq r$ and $\eta_t = \ell(r+t)^{-1}$ for $\ell > (2(\gamma - 2\nu L))^{-1}\ r > 0$ and that $B = \sup_{w \in \mathcal{W}, z \in \mathbb{R}^k} |f(w, z)| < \infty$. Then, the sequence of iterates $\{w_t\}_{t \geq 0}$ generated by the derivative free stochastic method satisfy*

$$\mathbb{E}\|w_t - w_\delta^*\|^2 \leq \frac{M}{r+t}, \quad \textit{where} \quad M := \max\left\{ r\|w_0 - w_\delta^*\|^2,\ \frac{B^2(d_x^2 + d_y^2)\ell^2}{\delta^2(2(\gamma - 2\nu L)\ell - 1)} \right\}. \tag{3.45}$$

Proof. For notational convenience, we write $\widehat{\gamma} = \gamma - 2\nu L$, and $C = B^2(d_x^2 + d_y^2)\delta^{-2}$. By applying non-expansiveness of the projection map, we get

$$\begin{aligned}
\mathbb{E}_t\|w_{t+1} - w_\delta^*\|^2 &\leq \mathbb{E}_t\|w_t - w_\delta^*\|^2 - 2\eta_t \mathbb{E}_t\langle w_t - w_\delta^*, g_\delta(w_t, z_t)\rangle + \eta_t^2 \mathbb{E}_t\|g_\delta(w_t, z_t)\|^2 \\
&\leq \|w_t - w_\delta^*\|^2 - 2\widehat{\gamma}\eta_t\|w_t - w_\delta^*\|^2 + C\eta_t^2 \\
&= (1 - 2\widehat{\gamma}\eta_t)\|w_t - w_\delta^*\|^2 + C\eta_t^2.
\end{aligned}$$

In substituting the step size $\eta_t = (\ell\gamma(r+t))^{-1}$, we find that

$$\mathbb{E}_t\|w_{t+1} - w_\delta^*\|^2 \leq \frac{r + t - 2\widehat{\gamma}\ell}{r+t}\|w_t - w_\delta^*\|^2 + \frac{C}{(r+t)^2}.$$

As in the proof of Theorem 24, the result follows by induction. □

This concludes our proof of convergence to the perturbed saddle point w_δ^*. Obtaining convergence to the saddle point w^* is a matter of applying the stochastic algorithm in stages with a geometrically decaying step size.

3.3 Numerical Experiments for Electric Vehicle Charging

To illustrate our results, we apply our algorithms to an electric vehicle charging problem in which two service providers set optimal prices for their service using demand data. We motivate this problem formulation in the following exposition.

3.3.1 Relative Cost Maximization in Competitive Markets

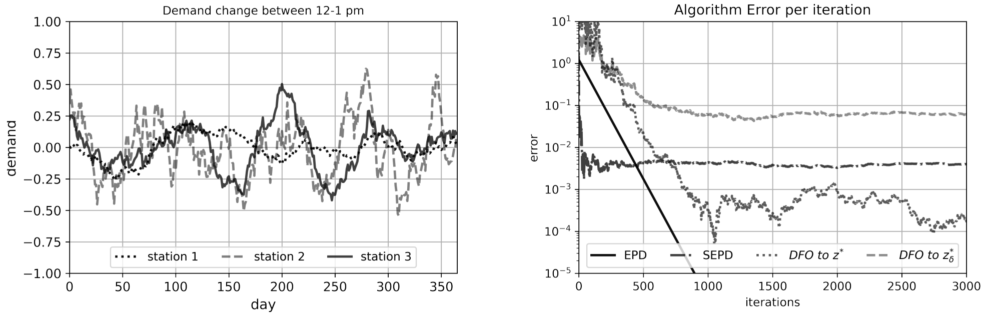
Consider a game in which two competing service providers aim to maximize their relative profits in a region partitioned in d zones. This applies to, for example, ride sharing [7] and power providers [2]. Focusing on electric vehicle charging station providers [38], at each zone $i \in [d]$ we denote the average baseline price as p_i and the price differential to charge per-minute set by provider one as x_i. The revenue of provider one is $(z_x)_i(x_i+p_i)$, based on their demand $(z_x)_i$. However, they must incorporate a zone based utility cost $\theta_i(x_i+p_i)$, as well as well as a term enforcing quality of service $\gamma_{1,i}x_i^2$ (the quadratic term balances the utility of the provider with the cost of ensuring quality of service by penalizing large deviations from the baseline price).

In total, the profit for provider one over all d zones is given by $u_1(x, z_x) = \langle z_x + \theta, x + p\rangle - \|\Gamma_1 x\|^2$, with $\Gamma_1 = \mathsf{diag}\{\gamma_{1,1}, \ldots, \gamma_{1,n}\}$. If the price and demand of service for provider two are given by y and z_y respectively, then their profit is similarly represented as $u_2(y, z_y) = \langle z_y + \theta, y + p\rangle - \|\Gamma_2 y\|^2$. Each provider has finite bounds on the prices they are willing to set in each zone, and hence their prices are constrained to the closed rectangles $\mathcal{X} = \times_{i=1}^{d}[-p_i, c_{1,i}p_i]$ and $\mathcal{Y} = \times_{i=1}^{d}[-p_i, c_{2,i}p_i]$ with multiplicative factors $c_{j,i} > 0$. The service demand vectors a and b are unknown quantities that will depend not only on the price set by their respective providers, but also their competition. One such example of a dependence is a best response model. It has been shown that best response models with linear utility and quadratic cost associated with changing features give rise to location-scale models of the form:

$$z_x \overset{d}{=} \xi_x + A_1 x + A_2 y,$$

$$z_y \overset{d}{=} \xi_y + B_1 x + B_2 y,$$

where $\xi_x \sim D_x$ and $\xi_y \sim D_y$, for which D_x and D_y represent stationary distributions for the demand associated with providers one and two respectively [49]. In order to maximize their expected profit relative to provider two, provider one will minimize the negative of their relative profit given by $u_1(x, z_x) - u_2(y, z_y)$, and hence the optimal strategies for both providers are solutions to the saddle



(a)

(b)

Figure 3.1: Data and results from numerical experiments. In (a) deviation in average demand for provider one's stations between 12 and 1 pm over 365 days.(b) the error of each algorithm depicted over 3,00 iterations. Error of the derivative-free method is depicted in both distance to the saddle point w^* as well as distance to the perturbed saddle point w^*_δ.

point problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \underset{(z_x,z_y)\sim D(x,y)}{\mathbb{E}} \|\Gamma_1 x\|^2 - \|\Gamma_2 y\|^2 - \langle z_x + \theta, x\rangle + \langle z_y + \theta, y\rangle, \tag{3.46}$$

where the dependence on baseline price p has been removed as it has no impact on the optimality criterion.

3.3.2 Numerical Simulations

In our simulation, each provider has access to three distinct regions, each of which having one station. The demand for each station is dictated by the data distributions from [28]. Each station is comprised of 50, 150, or 350 kW chargers with either 2 or 6 ports. We randomize this allocation at initialization. Data is processed by averaging the demand over each hour-long time window. After picking an hour block, we re-scale the data by subtract the mean and dividing by the variance. We choose the demand change in the 12-1pm block, and depict data for the year in Figure 3.1. Our simulations use charging utility values of $\gamma_{j,i} = 1$ for $j \in [2], i \in [3]$, elasticity values of $(A_1)_{i,j} = (-0.3)\delta_{i,j}$, $(A_2)_{i,j} = (0.3)\delta_{i,j}$, $B_1 = A_2$, and $B_2 = A_1$, and location utility values $r_i = 0$ for each station. The price deviations x and y are restricted to the interval $[-1, 2]$ for each station, representing a nominal price of \$1 and a maximum price change of twice the nominal price.

Hence $\mathcal{X} = \mathcal{Y} = [-1, 2]^3$.

We run each algorithm for 10,000 iterations, and depict the first 3,000 iterations in Figure 3.1 to provide a side-by-side comparison. The stable points and saddle points are computed via primal-dual with constant step size $\eta = 0.001$ as a means to compute the norm squared errors $\|w_t - \bar{w}\|^2$ and $\|w_t - w^*\|^2$. We run stochastic primal-dual and the zeroth order algorithm with the polynomial decay step-size schedules described in Theorems 3.34 and 33. In the latter, we choose a fixed δ value of 0.05. Relative to EPD, our results for these stochastic algorithms only guarantee sub-linear convergence at best; the step-size effectively converges to zero faster than the error resulting in the plateau of our error curves. The Python code is publicly available[1] .

3.4 Time-varying Extension

This work considers the problem of tracking the solution trajectories for problems of the form:

$$\min_{x \in \mathcal{X}_t} \max_{y \in \mathcal{Y}_t} \left\{ F_t(x, y) := \mathop{\mathbb{E}}_{z \sim D_t(x,y)} [f_t(x, y, z)] \right\} \tag{3.47}$$

where t is a time index, $\mathcal{X}_t \subseteq \mathbb{R}^n$ and $\mathcal{Y}_t \subseteq \mathbb{R}^m$ are convex and compact sets capturing time-varying constraints, $f_t : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^k \to \mathbb{R}$ is a strongly-convex-strongly-concave function revealed at time t, and $D_t : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^k)$ is a distributional map that maps decision variables to the set of finite-first moment probability distributions supported on $\mathbb{R}^k$ denoted by $\mathcal{P}(\mathbb{R}^k)$. Without loss of generality, we refer to the support of w as $\mathbb{R}^k$ (even if w is matrix valued, our analysis holds as w is isomorphic to its vectorization over $\mathbb{R}^k$).

Classical solutions to (3.47) are saddle points, which we denote $w_t^* = (x_t^*, y_t^*) \in \mathcal{X}_t \times \mathcal{Y}_t$. Under appropriate conditions, namely minimax equality, saddle points satisfy

$$x_t^* \in \arg\min_{x \in \mathcal{X}_t} \max_{y \in \mathcal{Y}_t} F_t(x, y), \quad y_t^* \in \arg\max_{y \in \mathcal{Y}_t} \min_{x \in \mathcal{X}_t} F_t(x, y). \tag{3.48}$$

In this setting, saddle points are optimal decisions that effectively anticipate the distributional shift, and hence are optimal even after the data distribution has changed in the system. While these are

[1] `https://github.com/killianrwood/charging-market`

ideal, finding them is typically computationally intractable. While sufficient conditions for their existence and uniqueness have been studied, guarantees for convergence to saddle points are only approximate or require explicit knowledge of a model for the distributional map [47, 65]. A common heuristic to overcome distributional shift in general is to repeatedly retrain the optimal decisions each time the distribution shifts. This amounts to forming a sequence $\{z_t^\ell\}_{\ell\geq 0} = \{(x_t^\ell, y_t^\ell)\}_{\ell\geq 0}$ at each time t defined by

$$\begin{aligned} x_t^{\ell+1} &\in \arg\min_{x\in\mathcal{X}_t}\max_{y\in\mathcal{Y}_t} \mathop{\mathbb{E}}_{z\sim D_t(x_t^\ell,y_t^\ell)}[f_t(x,y,z)], \\ y_t^{\ell+1} &\in \arg\max_{y\in\mathcal{Y}_t}\min_{x\in\mathcal{X}_t} \mathop{\mathbb{E}}_{w\sim D_t(x_t^\ell,y_t^\ell)}[f_t(x,y,w)]. \end{aligned} \tag{3.49}$$

The fixed points of this repeated retraining procedure have been coined *equilibrium points*, and are known to exist under mild conditions. In what follows we provide algorithms capable of tracking the equilibrium point trajectory $\{\bar{z}_t\}_{t\geq 0} = \{\bar{x}_t, \bar{y}_t\}_{t\geq 0}$ without requiring that we take the sequences in 3.49 to convergence ($\ell \to \infty$). This will be crucial for our online setting, as we assume that each time t, a new function and distributional map arrive ([6, 34, 13, 54, 18, 64]).

3.4.1 Stable Points

In this section we define the equilibrium problem, the fixed points of the repeated retraining heuristic in (3.49), and provide sufficient conditions for their existence. We start from the definition of equilibrium points.

Definition 12 (Equilibrium Points). *A pair $(\bar{x}_t, \bar{y}_t) \in \mathbb{R}^{d_x+d_y}$ is an equilibrium point if:*

$$\begin{aligned} \bar{x}_t &\in \arg\min_{x\in\mathcal{X}_t}\left\{\max_{y\in\mathcal{Y}_t} \mathop{\mathbb{E}}_{w\sim D_t(\bar{x}_t,\bar{y}_t)}[f_t(x,y,z)]\right\}, \\ \bar{y}_t &\in \arg\max_{y\in\mathcal{Y}_t}\left\{\min_{x\in\mathcal{X}_t} \mathop{\mathbb{E}}_{w\sim D_t(\bar{x}_t,\bar{y}_t)}[f_t(x,y,z)]\right\}. \end{aligned} \tag{3.50}$$

Sequences of equilibrium points are defined as $(\bar{x}_t, \bar{y}_t)_{t\in\mathbb{N}}$. □

In essence, equilibrium points are the solutions to the stationary saddle point problem that they induce. In this way, they are optimal decisions when data distribution is in state $D_t(\bar{x}_t, \bar{y}_t)$ but need not be optimal otherwise. Existence of these points is contingent on the distributional function being continuous on the set of probability distributions, and f_t being at least convex-concave.

Assumption 16 (Strong-Convexity-Strong-Concavity). *The function $(x, y) \mapsto f_t(x, y, w)$ is continuously differentiable over $\mathbb{R}^{d_x+d_y}$ for any realization of w. The function $(x, y) \mapsto f_t(x, y, w)$ is γ-strongly-convex-strongly-concave, for any realization of w; that is, f_t is γ-strongly-convex in x for all $y \in \mathbb{R}^m$ and γ-strongly-concave in y for all $x \in \mathbb{R}^n$.* □

Assumption 17 (Joint Smoothness). *The map $g_t(w, z) := (\nabla_x f_t(w, z), -\nabla_y f_t(w, z))$ is L-Lipschitz in z and w. Namely, $\|g_t(w, z) - g_t(w', z)\| \leq L\|w - w'\|$, $\|g_t(w, z) - g_t(w, z')\| \leq L\|w - w'\|$, for any $z, z' \in \mathbb{R}^{d_x+d_y}$ and z, z' supported on $\mathbb{R}^k$, for some $L \geq 0$.* □

Assumption 18 (Lipschitz-Continuous Distributional Map). *The distributional maps $D_t : \mathbb{R}^{d_x+d_y} \to \mathcal{P}(M)$ are ν-Lipschitz. Namely, $W_1(D_t(w), D_t(w')) \leq \nu\|w - w'\|$, for any $w, w' \in \mathbb{R}^{d_x+d_y}$, where W_1 is the Wasserstein-1 distance.* □

Assumption 19 (Compact Sets). *The sets $\mathcal{X}_t \subset \mathbb{R}^{d_x}$ and $\mathcal{Y}_t \subset \mathbb{R}^{d_y}$ are compact and convex.* □

Assumption 20 (Bounded Drift). *There exists a $\Delta > 0$ such that the equilibrium drift sequence defined by $\Delta_t := \|\bar{z}_{t+1} - \bar{z}_t\|$ is uniformly bounded by Δ. Namely, $\Delta_t \leq \Delta$ for all $t \geq 0$.* □

These assumptions provided are sufficient to guarantee uniqueness of the equilibrium point, and convergence of primal-dual algorithms in the offline (time-invariant) setting.

Theorem 34 (Equilibrium Point Uniqueness). *If Assumptions 16-19 are satisfied such that $\nu L < \gamma$, then a unique equilibrium point exists.*

Proof of this results amounts to showing that the repeated retraining heuristic in 3.2 is a strict contraction and hence satisfies the Banach-Picard Fixed Point Theorem. .

Given that the data distribution is shifting, it is necessary to characterize this shift and its effect on the gradient. The key to computing equilibrium points will be the gradients of f_t. We note that this is only one term required to compute the gradients of F_t, effectively ignoring the dependence of D_t on the decision variables. For now, we will denote the decoupled gradient map as the function G_t defined by

$$G_t(w|w') := \mathop{\mathbb{E}}_{z \sim D_t(w')} g_t(w, z) = \left(\mathop{\mathbb{E}}_{z \sim D_t(w')} \nabla_x f_t(w, z), \mathop{\mathbb{E}}_{z \sim D_t(w')} - \nabla_y f_t(w, z) \right) \tag{3.51}$$

for all $w, w' \in \mathbb{R}^{d_x+d_y}$. Note that we refer to this gradient map as "decoupled" as we separate the decision variable in the stochastic objective and the distributional map. This will allow us to characterize these behaviors separately.

Lemma 35 (Gradient Map Characterization)**.** *If Assumptions 16-19 hold, then:*

*(1) (**Gradient Deviation**) For any fixed $w_0 \in \mathbb{R}^{d_x+d_y}$, the map $w \mapsto G_t(w_0|w)$ is νL-Lipschitz-continuous. That is, $\|G_t(w_0|w) - G_t(w_0|w)\| \leq \nu L\|w - w'\|$, for all $w, w' \in \mathbb{R}^{d_x+d_y}$.*

*(2) (**Strong-Monotonicity**) The map $w \mapsto G_t(w|w)$ is $(\gamma - \nu L)$-strongly-monotonic.*

*(3) (**Lipschitz-Continuity**) The map $w \mapsto G_t(w|w)$ is $(L + \nu L)$-Lipschitz Continuous.*

Proof of the Gradient Deviation property follows by combining the properties allowed from joint smoothness and lipschitz continuity of the distributional map (Assumptions 17 and 18 respectively. Strong monotonicity and Lipschitz continuity of $z \mapsto G_t(w|w)$ then follow immediately. With this lemma, we can effectively deal with the decoupled gradient map by passing variables into both the D_t and g_t simultaneously. Going forward, we will simply write G_t to mean the gradient map given by $w \mapsto G_t(w|w)$.

3.5 Online Algorithms

In this section, we provide online analogs for both the conceptual and stochastic primal-dual algorithms discussed in Section 3.1.

3.5.1 A Conceptual Primal-Dual Algorithm

In this section, we demonstrate that a full-information primal-dual algorithm is capable of tracking stable points up to drift error Δ. This provides a basis of comparison for our analysis in the next section where we use a stochastic gradient estimator in place of G_t. Our online primal-dual update is given by

$$w_{t+1} = \mathsf{proj}_{\mathcal{W}_t}\left[w_t - \eta G_t(w_t|w_t),\right] \tag{3.52}$$

which can be represented via the algorithmic map $\mathcal{A}_t : \mathcal{W}_t \times \mathcal{W}_t \to \mathcal{W}_t$ given by

$$\mathcal{A}_t(w|w') = \mathsf{proj}_{\mathcal{W}_t}\left[w - \eta G_t(w|w')\right]. \tag{3.53}$$

To proceed, we observe that equilibrium points are the fixed points of the primal-dual algorithmic map.

Proposition 36 (Fixed Point Characterization)**.** *Let Assumptions 16-19 hold and suppose that* $\frac{\nu L}{\gamma} < 1$*. A point* $\bar{w}_t \in \mathcal{W}_t$ *is an equilibrium point if and only if* $\bar{w}_t = \mathcal{A}_t(\bar{w}_t|\bar{w}_t)$*.* □

This proposition will allow us to cast our analysis into a fixed point framework, using the equilibrium points as the fixed points of the distributional map.

Theorem 37 (Primal-Dual Tracking)**.** *Suppose that Assumptions 16-19 hold and that* $\frac{\nu L}{\gamma} < 1$*. Then the sequence* $w_{t+1} = \mathcal{A}_t(w_t|w_t)$ *satisfies the bound*

$$\|w_t - \bar{w}_t\| \leq \alpha^t \|w_0 - \bar{w}_0\| + \Delta(1-\alpha)^{-1} \tag{3.54}$$

for any initial point $w_0 \in \mathbb{R}^{d_x + d_y}$ *and* $\alpha := \sqrt{1 - \eta(\gamma - \nu L)}$ *provided that*

$$\eta < \min\left\{\frac{1}{\gamma - \nu L}, \frac{\gamma - \nu L}{(1+\nu)^2 L^2}\right\} \tag{3.55}$$

Furthermore, $\{w_t\}_{t\geq 0}$ *ultimately tracks the sequence of unique equilibrium points* $\{\bar{w}_t\}_{t\geq 0}$ *in the sense that* $\limsup_{t\to\infty} \|w_t - \bar{w}_t\| \leq (1-\alpha)^{-1}\Delta$*.* □

Proof. It follows from the triangle inequality that $\|w_{t+1} - \bar{w}_{t+1}\| \leq \|w_{t+1} - \bar{w}_t\| + \|\bar{w}_t - \bar{w}_{t+1}\| = \|w_{t+1} - \bar{w}_t\| + \Delta_t$, and hence we simply need to bound $\|w_{t+1} - \bar{w}_t\|$. We observe that

$$\begin{aligned}
\|w_{t+1} - \bar{w}_t\|^2 &= \|\mathsf{proj}_{\mathcal{W}_t}\left[w_t - \eta G_t(w_t|w_t)\right] - \mathsf{proj}_{\mathcal{W}_t}\left[\bar{w}_t - \eta G_t(\bar{w}_t|\bar{w}_t)\right]\|^2 \\
&\leq \|(w_t - \bar{w}_t) - \eta\left(G_t(w_t|w_t) - G_t(\bar{w}_t|\bar{w}_t)\right)\|^2 \\
&\leq \|w_t - \bar{w}_t\|^2 - 2\eta\langle w_t - \bar{w}_t, G_t(w_t|w_t) - G_t(\bar{w}_t|\bar{w}_t)\rangle + \eta^2\|G_t(w_t|w_t) - G_t(\bar{w}_t|\bar{w}_t)\|^2.
\end{aligned}$$

If we denote $\bar{\gamma} = \gamma - \nu L$ and $\bar{L} = L + \nu L$, then from Lemma 35 we have that G_t is $\bar{\gamma}$-strongly monotone and $\bar{L}$-Lipschitz continuous. Combining these facts yields

$$\langle w_t - \bar{w}_t, G_t(w_t) - G_t(\bar{w}_t)\rangle \geq \frac{\bar{\gamma}}{2}\|w_t - \bar{w}_t\|^2 + \frac{\bar{\gamma}}{2\bar{L}^2}\|G_t(w_t) - G_t(\bar{w}_t)\|^2.$$

Substituting into the above yields

$$\|w_{t+1} - \bar{w}_t\|^2 \leq (1 - \eta\bar{\gamma})\|w_t - \bar{w}_t\|^2 + \eta\left(\eta - \frac{\bar{\gamma}}{\bar{L}^2}\right)\|G_t(w_t) - G_t(\bar{w}_t)\|^2 \leq (1 - \eta\bar{\gamma})\|w_t - \bar{w}_t\|^2$$

where the last inequality follows provided that $\eta \leq \bar{\gamma}/\bar{L}^2$. It follows that if $\eta < 1/\bar{\gamma}$ as well, then $1 - \eta\bar{\gamma} < 1$ and the bound in Theorem (3.54) follows. Considering the limit supremum of the bound in (3.54) yields the result. □

We note that the noise due to the drift in (3.54) increases as we decrease the step size η. Hence it is impossible to completely remove this disturbance from the algorithm. This reflects intuition however as very small step sizes would make it difficult to ever reach the solution trajectory. Meanwhile, larger step sizes decrease this noise while simultaneously decreasing the rate at which we overcome the error $z_{t+1} - \bar{z}_t$ between successive iterates. We build on this intuition in our stochastic algorithm. This concludes our discussion of the conceptual primal-dual algorithm. In the next section, we demonstrate tracking of a stochastic primal-dual algorithm.

3.5.2 A Stochastic Primal-Dual Algorithm

In previous section, we demonstrated that a conceptual first-order algorithm is capable of tracking the trajectory of equilibrium point. In this section, we extend this result to a practical implementation based on a stochastic gradient estimator. To remain consistent with the rest of this work, we will abuse notion and denote g_t as the stochastic gradient estimator of G_t, which we expect will be some evaluation of the map $(w, z) \mapsto g_t(w, z)$.

Given a starting point w_0, the stochastic primal-dual algorithm performs the update

$$w_{t+1} = \mathsf{proj}_{\mathcal{W}_t}\left[w_t - \eta_t g_t\right] \tag{3.56}$$

Crucial to our analysis will be providing reasonable assumptions regarding the quality of the gradient estimator g_t. The case where $g_t = g_t(w_t, z_t)$ for $z_t \sim D_t(w_t)$ is particularly appealing in applications such as competitive markets, strategic classification, etc. since is does not require coordinating the algorithm to allow many sources feedback from the population.

Assumption 21 (Sub-Weibull Framework)**.** *Denote the gradient error incurred throughout the stochastic algorithm as $\xi_t = g_t - G_t(w_t|w_t)$. Then there exists constants $\theta, \nu > 0$ and a sequence $\{\omega_t\}_{t\geq 0} \subseteq \mathbb{R}_+$ such that the following hold:*

(1) ***Sub-Weibull Gradient Error****. For each $t \geq 0$, $\|\xi_t\|$ is a sub-Weibull random variable such that $\|\xi_t\| \sim subW(\theta, \omega_t)$.*

(2) ***Bounded Variance Proxies****. The sequence of variance proxies $\{\omega_t\}_{t\geq 0}$ is bounded by ω.*

With this assumption, the main convergence result is stated next.

Theorem 38. *Suppose that Assumptions 16-21 hold and $\frac{\nu L}{\gamma} < 1$. If η satisfies the bound in 3.55 then the following hold:*

(1) ***Expectation****. The sequence $\{z_t\}_{t\geq 0}$ satisfies the bound in expectation*

$$\mathbb{E}\|w_t - \bar{w}_t\| \leq \rho^t \|w_0 - \bar{w}_0\| + \frac{\Delta + \eta\omega}{1 - rho}. \tag{3.57}$$

for all $t \geq 0$, for any initial point $w_0 \in \mathbb{R}^d$, and $\rho := \sqrt{1 - \eta(\gamma - \nu L)}$.

(2) ***High Probability****. For any $\delta \in (0,1)$, and $t \geq 0$,*

$$\mathbb{P}\left(\|w_t - \bar{w}_t\| \leq \rho^t \|w_0 - \bar{w}_0\| + \frac{\Delta}{1-\rho} + c(\theta) \log^\theta \left(\frac{2}{\delta}\right) \frac{\eta\nu}{1-\rho} \right) \geq 1 - \delta \,. \tag{3.58}$$

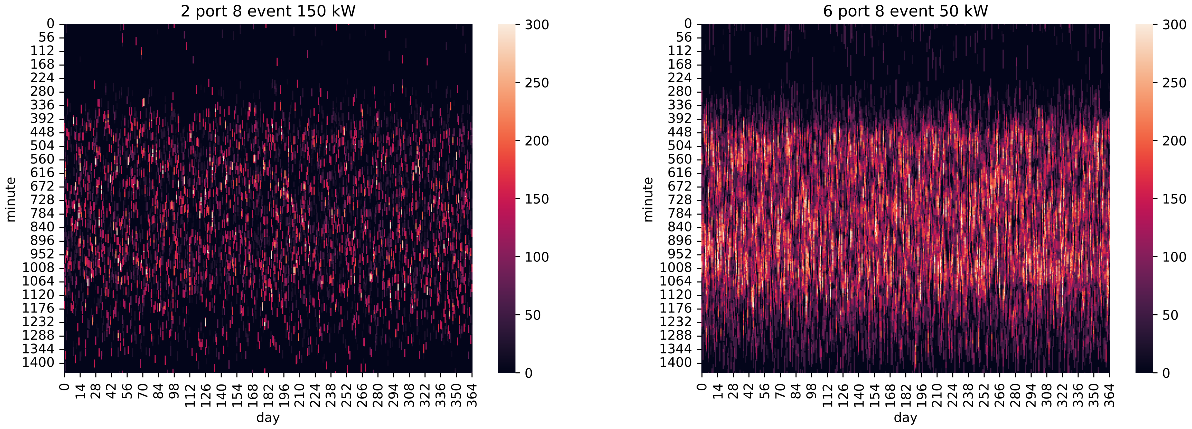
with $c(\theta) := \left(\frac{2e}{\theta}\right)^\theta$, for any initial point $w_0 \in \mathcal{W}_t$.

Proof. As before, we have that $\|w_{t+1} - \bar{w}_{t+1}\| \leq \|w_{t+1} - \bar{w}_t\| + \Delta_t$ where

$$\begin{aligned} \|w_{t+1} - \bar{w}_t\| &\leq \|(w_t - \bar{w}_t) - \eta(g_t - G_t(\bar{w}_t|\bar{w}_t)\| \\ &= \|(w_t - \bar{w}_t) - \eta(G_t(w_t|w_t) - G_t(\bar{w}_t|\bar{w}_t) - \eta\xi_t\| \\ &\leq \rho\|w_t - \bar{w}_t\| + \eta\|\xi_t\|. \end{aligned}$$

This yields that stochastic recursion $\|w_t - \bar{w}_t\| \leq \rho^t\|w_0 - \bar{w}_0\| + \Delta \sum_{i=0}^t \rho^i + \eta \sum_{i=0}^t \rho^i \|\xi_{t-i}\|$. Recall that when η satisfies the condition in (3.55), $\rho < 1$. Hence assuming this fact and taking the expectation of both sides yields

$$\mathbb{E}\|w_t - \bar{w}_t\| \leq \rho^t\|w_0 - \bar{w}_0\| + \frac{\Delta}{1-\rho} + \eta \sum_{i=0}^t \rho^i \mathbb{E}\|\xi_{t-i}\|$$



(a) (b)

Figure 3.2: Demand time series visualization: horizontal axis is time of day, vertical axis is the day of the year between 1 and 365. Brightness indicates intensity of the demand.

so that the result in (3.57) follows. To prove the result in (3.27), we denote $e_t = \|w_t - \bar{w}_t\|$, $E_t = \rho^t\|w_0 - \bar{w}_0\| + \Delta(1-\rho)^{-1}$, and $S_t = \eta\sum_{i=0}^t \rho^i\|\xi_{t-i}\|$. Observe that, due to our closure properties,

$$\|S_t\|_p \leq \sum_{i=0}^{t} \rho^i \mathbb{E}\|\xi_t\|^p]^{1/p} \leq \frac{\eta\omega}{1-\rho}p^\theta$$

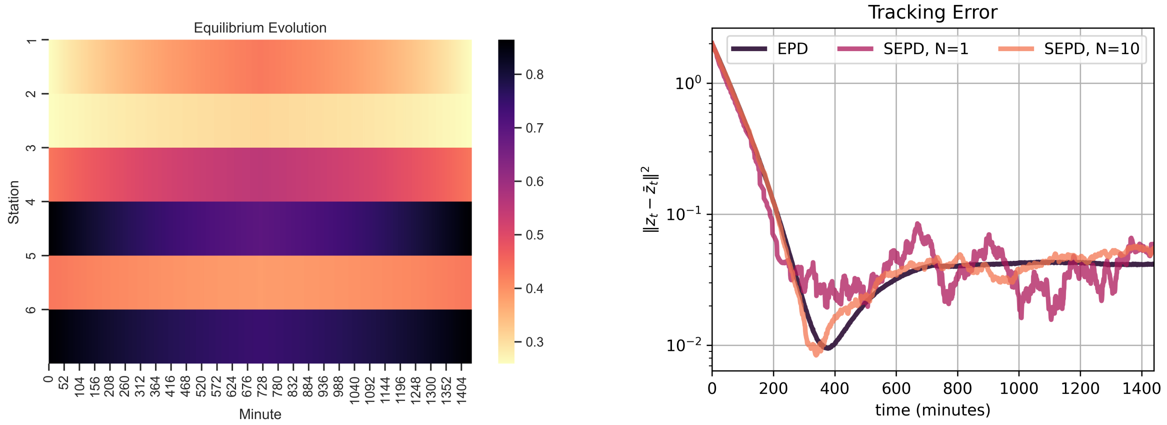
for any $p \geq 1$ and hence $S_t \sim subW(\theta, \eta\nu(1-\rho)^{-1})$. It follows that

$$\mathbb{P}\left(S_t \geq \varepsilon\right) \leq 2\exp\left(-\frac{\theta}{2e}\left(\frac{(1-\rho)\varepsilon}{\eta\omega}\right)^{\frac{1}{\theta}}\right), \tag{3.59}$$

and setting the right hand side above equal to $\delta > 0$ yields $\varepsilon = c(\theta)\log^\theta\left(\frac{2}{\delta}\right)\eta\omega(1-\rho)^{-1}$. Now, observe that our stochastic recursion implies that for any $a > 0$, $\mathbb{P}(E_t + S_t \geq a) \geq \mathbb{P}(e_t \geq a)$. It follows that setting $a = E_t + \varepsilon$ yields $\mathbb{P}(e_t \leq E_t + \varepsilon) \geq \mathbb{P}(E_t + S_t \leq E_t + \varepsilon) = \mathbb{P}(S_t \leq \varepsilon) \geq 1 - \delta$, thus the result follows. □

3.5.3 Numerical Simulations on Electric Vehicle Charging

Examples of problems of the form (3.47) emerge in profit maximization in competitive markets, where the (stochastic) demand shifts in response to prices (see, e.g., [40, 57]), and in applications in adversarial strategic classification, finance, energy systems, transportation networks, and ride-sharing—just to mention a few. Focusing on the first example, we consider a competition



(a) (b)

Figure 3.3: Results: in (a) we depict the evolution of the equilibrium points over the time horizon plotted in absolute value. In (b), we depict the tracking error for both algorithms.

between two service providers in an area with d distinct regions for which each provider seeks to maximize their relative revenue, and when the demand for each provider's service changes in response to the price variation set by both providers. This problem can be written as the saddle point problem

$$\min_{x \in \mathcal{X}_t} \max_{y \in \mathcal{Y}_t} \left\{ F_t(x, y) = \mathbb{E}_{(z_x, z_y) \sim D_t(x,y)} \|\Gamma_t^1 x\|^2 - \|\Gamma_t^2 y\|^2 - \langle z_x + \beta_t^1, x \rangle + \langle z_y + \beta_t^2, y \rangle \right\} , \tag{3.60}$$

where $x = (x_i)_{i=1}^d$ and $y = (y_i)_{i=1}^d$ are vectors of price deviations from a nominal value for providers one and two respectively (components x_i and y_i are the prices in region $i \in [d]$); raw profit for each provider is represented by terms $\langle z_x, x \rangle$ and $\langle z_y, y \rangle$; the terms $\langle -\beta_t^1, x \rangle + \|\Gamma_t^1 x\|^2$ and $-\langle b, y \rangle + \|\Gamma_t^2 y\|^2$ capture the risk-aversion associated with setting large price deviations; and $z_x, z_y \in \mathbb{R}^d$ are changes in demand in each region (in response to price changes) with distributions $z_x \overset{d}{=} \xi_x^t + A_1^t x + B_1^t y$, and $z_y \overset{d}{=} \xi_y^t + A_2^t x + B_2^t y$. Here $\{\xi_x^t\}_{t \geq 0}$ and $\{\xi_y^t\}_{t \geq 0}$ are time series such that each ξ_x^t and ξ_y^t are zero-mean random vectors representing the (nominal) demand of users in the market in the absence of decision-dependence.

Our experiment use nominal demand data from [28]. This dataset consists of a years of worth of electricity demand for up to 18 charging stations with entries for each minute of the year. Each file represents a different type of charging station positioned near commercial uses with varying number of ports (2 or 6) and port power output (50, 150, or 350 kW), and demand profile (low, medium, or high). We randomly allocate each provider with three medium demand stations. At each time, demand data is standardized across all stations simultaneously.

A representative example of the raw demand data is provided in Figure 1, with time in minutes along the horizontal axis, day of the year along the vertical, and color intensity representing demand value. The price elasticity is dictated by the function $h_t(p) = (-0.8/m|t - m| + 0.8)$ and $m = 720$ is the midpoint of the time horizon. The elasticity matrices are then given by $(A_1^t)_{ij} = -h_t(p_i)\delta_{i,j}$, $(B_1^t)_{ij} = -h_t(p_i)\delta_{i,j}$ for $i \in [3]$ where p_i is the power of each port at the ith station belonging to the provider and $B_2^t = -B_1^t$ and $A_2^t = -A_1^t$. We set $\Gamma_t^1 = \Gamma_t^2 = I$ and choose β_t^1, β_t^2 dependent on the port power of each station: $\beta_{i,t}^j = c(p_i^j)$ where p_i^j is power of port j for provider j such that $c(50) = 1.0$, $c(150) = 0.5$, and $c(350) = 0.3$. From this we conclude that for all t, G_t is 1-strongly monotone and 1-Lipschitz. Hence our results apply provided that $\eta < (1 - \nu)/(1 + \nu)^2$ where $\nu = 0.8$.

We compute the equilibrium points by executing a batch primal-dual algorithm for 2000 iterations with a step size of $\eta = 10^{-3}$. We then run the online primal-dual and stochastic primal-dual algorithms over the time horizon and plot the distance to the solutions in Figure 2. We observe that the primal-dual algorithm is capable of reasonably tracking the trajectory. Furthermore the stochastic primal-dual algorithm captures the trajectory with noise decreasing as the number samples increases.

Chapter 4

Monotone Games

We focus on solving the Nash equilibrium problem of a game, which is to find a decision from which no agent is incentivized by their own cost to deviate when played. Formally, the stochastic Nash equilibrium problem with decision-dependent distributions considered in this paper is to find a point $x^* = (x_1^*, \dots, x_n^*) \in \mathbb{R}^d$ such that

$$x_i^* \in \operatorname*{arg\,min}_{x_i \in \mathcal{X}_i} F_i(x_i, x_{-i}^*), \quad \forall\, i \in \{1, \dots, n\} \tag{4.1}$$

with $F_i(x_i, x_{-i}^*)$ defined as:

$$F_i(x_i, x_{-i}^*) := \mathop{\mathbb{E}}_{z_i \sim D_i(x_i, x_{-i}^*)} f_i(x_i, x_{-i}^*, z_i) \tag{4.2}$$

where: z_i denotes a random variable supported on $\mathbb{R}^{k_i}$, $f_i : \mathbb{R}^d \times \mathbb{R}^{k_i} \to \mathbb{R}$ is a scalar valued function that is convex and continuously differentiable in x_i, $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ is a compact convex set, and $D_i : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^{k_i})$ is a *distributional map* whose output is a probability distribution supported on $\mathbb{R}^{k_i}$.

Standard stochastic first-order methods are insufficient for solving problems of this form. As we will demonstrate later in the paper, even estimating the expected gradient from samples requires knowledge of the probability density function associated with D_i—which is not possible in a majority of practical applications.

Hereafter, we use the term "system" to refer to a population or a collection of automated controllers producing a response $z_i \in \mathbb{R}^{k_i}$ upon observing x. To illustrate our setup, consider again the example where each agent represents an EV charging provider. Here, $x_i \in \mathbb{R}^{d_i}$ represents the charging price at a station managed by provider i, expressed in \$/kWh. Correspondingly, z_i indicates demand for the service at that price, while f_i is the service cost (or the negative of the total profit) for provider i. This is an example of a competitive market in which the demand for service is a function of the price of all providers; see, for example, the game-theoretic approaches presented in [43, 27] and the Stackelberg game presented in [58]. However, compared to existing game-theoretic models for EV markets, the framework proposed in this paper allows for an uncertain response of EV owners to price variations; this randomness is difficult to model, as it it related to the drivers' preferences and other externalities such as the locations of the charging stations, etc., as explained in, e.g., [43, 37, 17].

Challenges in solving problems of this form typically stem from the that fact that the distributional maps D_i are often unknown [55, 20, 15, 14]. To overcome this challenge, we propose a learning-based optimization procedure – in the spirit of the methods proposed for convex optimization in [39, 45] – to tackle the multi-player decision-dependent stochastic game. The key idea behind this framework is that we first propose a parameterization for the distributional map in the system and estimate it from responses. Then, we use the estimated distributional map throughout the game without requiring further interaction with the system.

4.0.1 Monotonicity in Decision-Dependent Games

In this work, we introduce the additional complexity to the formulation in (1.18) that the F_i's are the expected cost over a distributional map $D_i : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^{k_i})$. In particular, we write the cost as

$$F_i(x_i, x_{-i}) := \mathop{\mathbb{E}}_{z_i \sim D_i(x_i, x_{-i})} f_i(x_i, x_{-i}, z_i). \tag{4.3}$$

This can be written alternatively as the integral

$$F_i(x) = \int_{\mathbb{R}^{k_i}} f_i(x, z_i) p_i(z_i, x) \mathrm{d}z_i \tag{4.4}$$

where p_i is the probability density function for the distribution $D_i(x)$. When the integral satisfies the Dominated Convergence Theorem, computing the gradient amounts to differentiating under the integral and using the product rule. We then obtain

$$\nabla_i F_i(x) = \mathop{\mathbb{E}}_{z_i \sim D_i(x)} \left[\nabla_{x_i} f_i(x, z_i) + f_i(x, z_i) \nabla_i \log p_i(x; z_i)\right], \tag{4.5}$$

where we recall that $G(x) = (\nabla_1 F_1(x), \ldots, \nabla_n F_n(x))$. In short, characterizing the gradient of this decision-dependent game requires assumptions not only on f_i, but also on the properties of the distributional map D_i. Sufficient conditions for strong monotonicity of the game in (1.18) are due to [47] and are stated in terms of the *decoupled* costs, given by

$$F_i(x, y) = \mathop{\mathbb{E}}_{z_i \sim D_i(y)} f_i(x, z_i) \tag{4.6}$$

for all $x, y \in \mathbb{R}^d$, and their associated decoupled partial gradients

$$G_i(x, y) = \mathop{\mathbb{E}}_{z_i \sim D_i(y)} \nabla_i f_i(x, z_i), \tag{4.7}$$

for all $x, y \in \mathbb{R}^d$ and

$$H_i(x, y) = \nabla_{y_i} \mathop{\mathbb{E}}_{z_i \sim D_i(y)} f_i(x, z_i) \tag{4.8}$$

for all $x, y \in \mathbb{R}^d$. A key observation used in the proof is that $G_i(x) = \nabla_i F_i(x) = G_i(x, x) + H_i(x, x)$.

Theorem 39 (Strong Monotonicity, [47]). *Suppose that,*

(i) For all $y \in \mathcal{X}$, $x \mapsto G(x, y)$ is λ-strongly monotone,

(ii) For all $x \in \mathcal{X}$, $y \mapsto H(x, y)$ is monotone,

and that for all $i \in [n]$,

(iii) For all $x \in \mathcal{X}, z_i \mapsto \nabla_i f_i(x, z_i)$ is L_i-Lipschitz continuous,

(iv) $y \mapsto D_i(y)$ is ν_i-lipschitz continuous on $(\mathcal{P}(\mathbb{R}^{k_i}), W_1)$.

Set $\kappa = \sqrt{\sum_{i=1}^{n} (\frac{\nu_i L_i}{\lambda})^2}$. Then if $\kappa < 1/2$, $x \mapsto G(x)$ is $\gamma = (1 - 2\kappa)\lambda$-strongly monotone. □

Algorithm 1: Multi-phase Optimization

```
Input: $m, \{D_{x_i}\}_{i=1}^n$
for $j \in [m]$ do
    for $i \in [n]$ do
        Draw $x_i^{(j)} \sim D_{x_i}$ ;
    end
    Deploy $x^{(j)}$ ;
    Observe $z_i^{(j)} \sim D_i(x^{(j)})$ ;
end
for $i \in [n]$ do
    Fit $\widehat{\beta}_i \in \arg\min_{\beta_i \in \mathcal{B}_i} \frac{1}{m} \sum_{j=1}^m R_i(x^{(j)}, z_i^{(j)}, \beta_i)$ ;
end
Compute $\widehat{x} \in \texttt{Nash}(G_{\widehat{\beta}}, \mathcal{X})$ ;
```

4.1 Learning-based Decision-Dependent Games

In this work, we aim to solve the stochastic Nash equilibrium problem with decision-dependent data distributions as formulated in (4.1). Methods for finding Nash equilibrium for games with decision dependent data distributions either use derivative free optimization, at the expense of an extremely slow rate, or use derivative information in conjunction with a learned model of the distributional map [47].

In [39], it is shown that a "plug-in" optimization approach, whereby a model for the distributional map is learned from samples prior to optimization, yields a bounded excess risk for the convex optimization problems with decision-dependent data. In this work, we leverage the properties of the system to simplify the communication structure of our approach. We assume that realizations of z_i can be directly observed from the system, and the decisions x_{-i} can be obtained from a server or are made public (for example, the prices of EV charging of different providers can be observed at the various stations).

To accommodate this setting, our algorithm proposes a multistage approach consisting of the following phases: (i) sampling; (ii) learning; (iii) optimization. It is important to note that following the learning phase players only need to participate in gradient play without receiving any additional feedback from the system in the form of $z_i \sim D_i(x)$. This is distinct from existing approaches in which performatively stable points can only be reached after several (even thousands of) rounds of feedback [49, 47, 65], and performatively optimal points can only be reached for models known to be location scale families a priori [45, 47].

Sampling. In the sampling phase we require that players collaborate by each deploying a set of decisions $\{x_i^{(j)}\}_{j=1}^m \overset{i.i.d}{\sim} D_{x_i}$ so that they can collectively receive feedback $z_i^{(j)} \sim D_i(x^{(j)})$ from the system (in response to their deployed decisions $\{x_i^{(j)}\}_{j=1}^m$). The result is that each agent has access to a dateset $\{x^{(j)}, z_i^{(j)}\}_{j=1}^m$ which they can use to learn their distributional map D_i

Learning. In this procedure, each player will choose a hypothesis class of parameterized functions

$$\mathcal{H}_{\mathcal{B}_i} = \left\{ D_{\beta_i} \middle| \, \beta_i \in \mathcal{B}_i \subseteq \mathbb{R}^{\ell_i} \right\}, \tag{4.9}$$

as well as a suitable criterion or risk function R_i, to formulate their own expected risk minimization problem

$$\beta_i^* \in \operatorname*{arg\,min}_{\beta_i \in \mathcal{B}_i} \mathop{\mathbb{E}}_{x \sim D_x, z_i \sim D_i(x)} R_i(x, z_i, \beta_i) \tag{4.10}$$

over the random variable (x, z_i) drawn from the coupled distribution $(D_x, D_i(x))$. Then, using the set of samples from the previous sampling phase, they can formulate the corresponding empirical risk minimization problem

$$\widehat{\beta}_i \in \operatorname*{arg\,min}_{\beta_i \in \mathcal{B}_i} \frac{1}{m} \sum_{j=1}^m R_i(x^{(j)}, z_i^{(j)}, \beta_i). \tag{4.11}$$

The result is a learned distributional map $D_{\widehat{\beta}_i}$ approximating D_i, which we can now use to solve the approximate Nash equilibrium problem.

Optimization. Following the approximation phase, each player now has an learned model of their distributional map $D_{\widehat{\beta}_i}$, which can be used to formulate an approximation of the ground-truth cost F_i and hence an approximate Nash equilibrium problem:

$$\widehat{x}_i \in \operatorname*{arg\,min}_{x_i \in \mathcal{X}_i} F_{\widehat{\beta}_i}(x_i, \widehat{x}_{-i}) \tag{4.12}$$

for all $i \in [n]$, where

$$F_{\widehat{\beta}_i}(x_i, \widehat{x}_{-i}) := \mathop{\mathbb{E}}_{z_i \sim D_{\widehat{\beta}_i}(x_i, \widehat{x}_{-i})} f_i(x_i, \widehat{x}_{-i}, z_i)\,. \tag{4.13}$$

Hereafter, we denote the Nash equilibrium of the approximate game as $\widehat{x}$ to distinguish it from the ground truth x^*. In Algorithm 1, we write the set of Nash equilibria for the operator $G_{\widehat{\beta}}$ with domain $\mathcal{X}$ as $\texttt{Nash}(G_{\widehat{\beta}}, \mathcal{X})$. In practice, we will assume the appropriate assumptions to guarantee uniqueness of this assignment; in which case the set inclusion is simply an equality.

By solving (4.12) instead of (4.1) we have introduced two errors: (i) the approximation error of the distributional map D_i by elements of the hypothesis class $\mathcal{H}_{\mathcal{B}_i}$, and (ii) the estimation or statistical error by solving the ERM problem instead of the expected risk minimization problem. In [39], the main result demonstrates that these two sources of error propagate through the optimization problem, and that the resulting excess risk can be bounded in terms of the sample complexity m. Our goal is to expand this result and provide additional analysis to our setting.

4.1.1 Parameter Estimation for Regular Problems

A critical component of our analysis is the estimation or learning of the distributional map and the subsequent characterization of the estimation error. In this section, we outline a class of expected risk minimization problems, which we call *regular problems*, for which we can characterize the distance between expected risk minimization solutions and empirical risk minimization solutions. Throughout, we write $R_i(\beta_i) = \mathbb{E}_{(x,z)}[R(x, z_i, \beta_i)]$ and $\widehat{R}_i(\beta_i) = (1/m)\sum_{j=1}^{m} R_i(x^{(j)}, z_i^{(j)}, \beta_i)$ for $\beta_i \in \mathbb{R}^{\ell_i}$ to denote the expected and empirical risk, respectively.

Definition 13 (Map Learning Regularity)**.** *A map learning problem, consisting of the optimization problems with costs R_i and $\widehat{R}_i$ over $\mathcal{B}_i$, is regular provided that:*

(a) ***Convexity:*** *The expected risk $\beta_i \mapsto R_i(\beta_i)$ is μ_i-strongly convex, and the empirical risk $\beta_i \mapsto \widehat{R}_i(\beta_i)$ is convex.*

(b) ***Smoothness:*** *For all realizations of $x \in \mathcal{X}$ and $z_i \in \mathbb{R}^{k_i}$, $\beta_i \mapsto \nabla_{\beta_i} R_i(x, z_i, \beta_i)$ is L_{β_i}-Lipschitz continuous.*

(c) ***Boundedness:*** *The set $\mathcal{B}_i \subseteq \mathbb{R}^{\ell_i}$ is convex and compact.*

(d) ***Sub-Exponential gradient:*** *For all $\beta_i \in \mathcal{B}_i$, $\nabla_{\beta_i} R_i(x, z_i, \beta_i)$ is a sub-exponential vector with parameter $\theta_i > 0$.* □

Items (a) and (c), taken together, guarantee existence of $\widehat{\beta}$ and uniqueness of β^* as defined in (4.11) and (4.10), respectively. Furthermore, the inclusion of item (b) is necessary to guarantee that first-order stochastic gradient methods will converge at least sub-linearly to $\widehat{\beta}$. Lastly, the heavy-tail assumption [60] will allow us to describe the concentration of the gradient estimates. Together, they allow us to relate the solutions to the sample complexity in the following lemma.

Lemma 40 (Uniform Gradient Bound)**.** *If the smoothness and sub-exponential gradient assumptions in Definition 13 hold for player $i \in [n]$, then for any $\delta \in (0, 1/2)$ and any m such that $m/\log(m) \geq 2(\ell_i + \log(1/\delta))$, we have that:*

$$\sup_{\beta \in \mathcal{B}} \|\nabla \widehat{R}_i(\beta) - \nabla R_i(\beta)\| \leq 4 \max\{L_{\beta_i}/15r_i, \theta_i\} \sqrt{\frac{\log(m)(\ell_i + \log(1/\delta))}{m}} \tag{4.14}$$

with probability at least $1 - \delta$.

Proof. For the sake of notation convenience, and visual clarity, we will suppress the i index throughout the proof. We denote the gradient error by $J(\beta) = \nabla \widehat{R}(\beta) - \nabla R(\beta)$ for all $\beta \in \mathbb{R}^\ell$.

To begin, we will generate coverings for the unit sphere in $\mathbb{R}^\ell$ and $\mathcal{B} \subseteq \mathbb{R}^\ell$ and use a discretization argument to create bounds over these finite sets. Fix $\beta \in \mathcal{B}$ and $u \in \mathcal{S}^{\ell-1}$. Let $\{u_j\}_{j=1}^N$ be an arbitrary 1/2-covering of the sphere $\mathbb{S}^{d_{\ell_i}}$ with respect to the Euclidean norm. From [62, Lemma 5.7], we know that $N \leq 5^\ell$. From our covering, we have that there exists u_j in the covering such that $\|u - u_j\| \leq 1/2$. Hence,

$$\begin{aligned}
\langle u, J(\beta)\rangle &= \langle u_j + (u - u_j), J(\beta)\rangle \\
&= \langle u_j, J(\beta)\rangle + \langle u - u_j, J(\beta)\rangle \\
&\leq \langle u_j, J(\beta)\rangle + \|u - u_j\|\|J(\beta)\| \\
&\leq \langle u_j, J(\beta)\rangle + \frac{1}{2}\|J(\beta)\| \\
&\leq \max_{j \in [N]} \langle u_j, J(\beta)\rangle + \frac{1}{2}\|J(\beta)\| \,.
\end{aligned}$$

Since this is true for any $u \in \mathcal{S}^{d-1}$, then it holds for $u = J(\beta)/\|J(\beta)\|$. Thus the above becomes

$$\|J(\beta)\| \leq 2\langle u_j, J(\beta)\rangle \leq 2 \max_{j\in[N]} \langle u_j, J(\beta)\rangle. \tag{4.15}$$

Now we fix $\nu \in (0,1]$, and choose and ε-covering for the set $\mathcal{B}$, which we will write as $\{\beta_k\}_{k=1}^M$. Recall that $\mathcal{B}$ is bounded, so there exists a constant $r > 0$ such that for all $\beta \in \mathcal{B}$, $\|\beta\| \leq r$. Hence $\mathcal{B} \subseteq B(r)$. From [60, Proposition 4.2.12], we have that

$$M \leq \frac{\text{vol}\left(B(r) + \frac{\varepsilon}{2}B(1)\right)}{\text{vol}\left(\frac{\varepsilon}{2}B(1)\right)} = \frac{\text{vol}\left(\frac{3}{2}B(r)\right)}{\text{vol}\left(\frac{\varepsilon}{2}B(1)\right)} = \left(\frac{3r}{\varepsilon}\right)^{\ell}. \tag{4.16}$$

Thus, we conclude that $M \leq (3r/\varepsilon)^{\ell}$.

Now by our discretization argument, there exists $k \in [M]$ such that $\|\beta - \beta_k\| \leq \varepsilon$ and hence

$$\begin{aligned}
\max_{j\in[N]}\langle u_j, J(\beta)\rangle &= \max_{j\in[N]}\langle u_j, J(\beta_k) + (J(\beta) - J(\beta_k)\rangle \\
&= \max_{j\in[N]}\langle u_j, J(\beta_k)\rangle + \langle u_j, J(\beta) - J(\beta_k)\rangle \\
&\leq \max_{j\in[N]}\langle u_j, J(\beta_k)\rangle + \max_{j\in[N]}\langle u_j, J(\beta) - J(\beta_k)\rangle \\
&\leq \max_{k\in[M]}\max_{j\in[N]}\langle u_j, J(\beta_k)\rangle + \sup_{\|\alpha-\alpha'\|\leq\varepsilon}\max_{j\in[N]}\langle u_j, J(\alpha) - J(\alpha')\rangle.
\end{aligned}$$

We observe that if $\alpha, \alpha' \in \mathcal{B}$ are such that $\|\alpha - \alpha'\| \leq \varepsilon$, then applying our smoothness assumption yields

$$\begin{aligned}
\langle u_j, J(\alpha) - J(\alpha')\rangle &= \langle u_j, (\nabla\widehat{R}(\alpha) - \nabla R(\alpha)) - (\nabla\widehat{R}(\alpha') - \nabla R(\alpha'))\rangle \\
&= \langle u_j, \nabla\widehat{R}(\alpha) - \nabla\widehat{R}(\alpha')\rangle + \langle u_j, \nabla R(\alpha') - \nabla R(\alpha)\rangle \\
&\leq \|u_j\|\|\nabla\widehat{R}(\alpha) - \nabla\widehat{R}(\alpha')\| + \|u_j\|\|\nabla R(\alpha) - \nabla R(\alpha')\| \\
&\leq L_{\beta_i}\|\alpha - \alpha'\| + L_{\beta_i}\|\alpha - \alpha'\| \\
&\leq 2L_{\beta}\varepsilon,
\end{aligned}$$

where the second-to-last inequality uses $\|u_j\| = 1$.

To bound the remaining term, we use the concentration of sub-exponential random variables, due to Bernstein's Inequality combined with the Union Bound. We have that

$$\mathbb{P}\left(\langle u_j, J(\beta_k)\rangle \geq t\right) \leq 2\exp\left(-\frac{mt^2}{2\theta^2}\right)$$

for all $t \leq \theta$, and hence

$$\begin{aligned}
\mathbb{P}\left(\max_{k\in[M]}\max_{j\in[N]}\langle u_j, J(\beta_k\rangle \geq t\right) &= \mathbb{P}\left(\bigcup_{k\in[M]}\bigcup_{j\in[N]}\{\langle u_j, J(\beta_k\rangle \geq t\}\right) \\
&\leq \sum_{k\in[M]}\sum_{j\in[N]}\mathbb{P}\left(\{\langle u_j, J(\beta_k\rangle \geq t\}\right) \\
&\leq \sum_{k\in[M]}\sum_{j\in[N]} 2\exp\left(-\frac{mt^2}{2\theta^2}\right) \\
&= M\cdot N\cdot 2\exp\left(-\frac{mt^2}{2\theta^2}\right) \\
&\leq 2\left(\frac{15r}{\varepsilon}\right)^{\ell}\exp\left(-\frac{mt^2}{2\theta^2}\right)
\end{aligned}$$

for all $t \leq \theta$, where we used the fact that $M \leq (3r/\varepsilon)^{\ell}$ and $N \leq 5^{\ell}$. Setting the right hand side equal to 2δ yields

$$t = \sqrt{2}\theta\sqrt{\frac{\ell\log(15r/\varepsilon)+\log(1/\delta)}{m}}. \tag{4.17}$$

Next we choose $\varepsilon = \frac{1}{15r}\sqrt{\frac{\ell+\log(1/\delta)}{m}}$ so that

$$\begin{aligned}
t &= \sqrt{2}\theta\sqrt{\frac{\ell\log(15r/\varepsilon)+\log(1/\delta)}{m}} \\
&= \sqrt{2}\theta\sqrt{\frac{\frac{\ell}{2}\log(m)-\frac{\ell}{2}\log(\ell+\log(1/\delta))+\log(1/\delta)}{m}} \\
&\leq \sqrt{2}\theta\sqrt{\frac{\ell\log(m)+\log(1/\delta)}{m}} \\
&\leq \sqrt{2}\theta\sqrt{\frac{\log(m)(\ell+\log(1/\delta))}{m}}.
\end{aligned}$$

By requiring that m satisfy $m/\log(m) \geq 2(\ell+\log(1/\delta))$, we enforce that $t \leq \theta$. In combining, we observe that

$$\begin{aligned}
t + 2\varepsilon L &\leq \sqrt{2}\theta\sqrt{\frac{\log(m)(\ell+\log(1/\delta))}{m}} + \frac{2L}{15r}\sqrt{\frac{\ell+\log(1/\delta)}{m}} \\
&\leq 2\left(\theta+\frac{L}{15r}\right)\sqrt{\frac{\log(m)(\ell+\log(1/\delta))}{m}} \\
&\leq 4\max\left\{\frac{L}{15r},\theta\right\}\sqrt{\frac{\log(m)(\ell+\log(1/\delta))}{m}},
\end{aligned}$$

and the result follows. □

This result offers a broad generalization of [46, Equation (19b)] to any risk with Lipschitz-continuous sub-exponential gradients over any convex and compact set. Our result is comparable to the $\mathcal{O}(\sqrt{\ell_i}m)$ rate that can be found for specific problem instances such as linear least squares regression and logistic regression, but with the addition of a $\sqrt{\log m}$ factor. Indeed, the generality of the risk function requires that we enforce compactness of the domain, thus giving rise to this extra logarithmic factor. This gradient estimation result will now allow us to reach our desired bounded distance result, which we present in the following theorem.

Theorem 41 (ERM Approximation)**.** *If the map learning problem is regular for player $i \in [n]$ (i.e., it satisfies the assumptions in Definition 13), then for any $\delta \in (0, 1/2)$ and any m such that $m/\log(m) \geq 2(\ell_i + \log(1/\delta))$ we have that:*

$$\|\widehat{\beta}_i - \beta_i^*\| \leq C_i \sqrt{\frac{\log(m)(\ell_i + \log(1/\delta))}{m}} \tag{4.18}$$

with probability at least $1 - \delta$, where $C_i = (4/\mu_i)\max\{L_{\beta_i}/15r_i, \theta_i\}$. □

Proof. We suppress the subscript i for notational simplicity. We recall that that the μ-strong convexity of the map $\beta \mapsto R(x, z; \beta)$ implies μ-strong monotonicity of $\nabla R(\beta)$, and $\nabla \widehat{R}(\beta)$. It follows that

$$\begin{aligned}
\mu\|\widehat{\beta} - \beta^*\|^2 &\leq \langle \widehat{\beta} - \beta^*, \nabla R(\widehat{\beta}) - \nabla R(\beta^*) \rangle \\
&= \langle \widehat{\beta} - \beta^*, \nabla R(\widehat{\beta}) \rangle - \langle \widehat{\beta} - \beta^*, \nabla R(\beta^*) \rangle \\
&\leq \langle \widehat{\beta} - \beta^*, \nabla R(\widehat{\beta}) \rangle \\
&\leq \langle \widehat{\beta} - \beta^*, \nabla R(\widehat{\beta}) \rangle + \langle \beta^* - \widehat{\beta}, \nabla \widehat{R}(\widehat{\beta}) \rangle \\
&= \langle \widehat{\beta} - \beta^*, \nabla R(\widehat{\beta}) - \nabla \widehat{R}(\widehat{\beta}) \rangle \\
&\leq \|\widehat{\beta} - \beta^*\| \sup_{\beta \in \mathcal{B}} \|\nabla R(\beta) - \nabla \widehat{R}(\beta)\|
\end{aligned}$$

and hence

$$\|\widehat{\beta} - \beta^*\| \leq \frac{1}{\mu} \sup_{\beta \in \mathcal{B}} \|\nabla R(\beta) - \nabla \widehat{R}(\beta)\|. \tag{4.19}$$

The result now follows by applying Lemma 40. □

The power in this characterization lies in the fact that it holds for any statistical learning problem satisfying the assumptions listed in Definition 13, and is not specific to the setting of learning distributional maps. We note that our Definition 13, which is a property used in the Theorem 41, is different from the one in [39] and it involves conditions that are easier to check.

As an example, we provide conditions for which a linear least squares problem satisfies the regularity conditions and hence is subject to the above ERM approximation result.

Proposition 42 (Linear Least Squares Regularity)**.** *Consider the linear least squares problem with expected risk problem*

$$B_i^* \in \operatorname*{arg\,min}_{B\in\mathcal{B}_i} \frac{1}{2}\mathbb{E}_{(x,z)}\|Bx - z\|^2,$$

and empirical risk minimization problem

$$\widehat{B}_i \in \operatorname*{arg\,min}_{B\in\mathcal{B}_i} \frac{1}{2m}\sum_{j=1}^{m}\|Bx_i^{(j)} - z_i^{(j)}\|^2.$$

Let $x_i \sim D_i$ with zero mean and covariance matrix Σ_i. If

(i) There exist $\gamma_i, L_i > 0$ such that $\gamma_i I \leq \Sigma_i \leq L_i I$,

(ii) The entries of xx^T and zx^T are sub-exponential,

(ii) The constraint set $\mathcal{B}_i$ is convex and compact.

Then, the map learning problem is regular. □

Proof. We suppress the i index throughout. The associated risk function is $R(x, z, B) = \frac{1}{2}\|Bx-z\|^2$, so that $\nabla R(x, z, B) = (Bx - z)x^T = Bxx^T - zx^T$ and $\nabla^2 R(x, z, B) = xx^T$ are the corresponding gradient and hessian. We observe that enforcing $\gamma I \leq \mathbb{E}[xx^T] \leq LI$ for some $\gamma, L > 0$ ensures γ-strong convexity and L-smoothness of the expected risk. Similarly, the empirical risk has gradient $\nabla R_m(B) = 1/m(BXX^T - ZX^T)$, and hessian $\nabla^2 R_m(B) = (1/m)XX^T$. Thus R_m is convex the hessian is symmetric, then it is positive semi-definite and thus R_m is convex. Furthermore, smoothness of R_m follows with constant $\max\{L, \|XX^T\|_2\}$. Lastly, since zx^T and xx^T have sub-exponential entries, the gradient is sub-exponential and the result follows. □

Deriving conditions for the more general case of non-linear regression is attainable but outside the scope of this work.

4.1.2 Bounding the Approximation Error

Finding a relationship between $\widehat{x}$ and x^* will require that we first characterize an appropriate hypothesis class of distributions for learning. Here, we formalize the notion of misspecification and sensitivity for a hypothesis class $\mathcal{H}_{\mathcal{B}_i}$.

Definition 14 (Misspecification, [39])**.** *A hypothesis class $\mathcal{H}_{\mathcal{B}_i}$ is ζ_i-misspecified provided that there exists a $\zeta_i > 0$ such that*

$$W_1(D_{\beta_i^*}(x), D_i(x)) \leq \zeta_i \tag{4.20}$$

for all $x \in \mathcal{X}$. □

We note that, although η_i is not known to agents in practice, it is a useful conceptual quantity that can be used to represent the expressiveness of the parameterization relative to the ground truth; it also captures the ability of the chosen risk function to fit a parameterization. This is similar to the notion of approximation error used in classical statistical learning methods [53]. However, unlike this setting, we note that $\eta_i = 0$ implies that $D_{\beta_i^*}(x) = D_i(x)$ for all $x \in \mathcal{X}$; hence, $z \sim D(x)$ and $z' \sim D_{\beta_i^*}(x)$ yields $z \overset{d}{=} z'$ but not necessarily $z = z'$ almost everywhere as we might like.

Definition 15 (Sensitivity, [39])**.** *The hypothesis class $\mathcal{H}_{\mathcal{B}_i}$ is ν_i-sensitive if, for any $\beta_i, \beta_i' \in \mathcal{B}_i$,*

$$W_1(D_{\beta_i}(x), D_{\beta_i'}(x)) \leq \upsilon_i \|\beta_i - \beta_i'\| \tag{4.21}$$

for all $x \in \mathcal{X}$. □

Sensitivity of $\mathcal{H}_{\mathcal{B}_i}$ is merely a convenient name for the condition that $\beta \mapsto D_{\beta_i}(x)$ be υ_i-Lipschitz continuous for all realizations of $x \in \mathcal{X}$. In the result that follows, we demonstrate that an appropriately misspecified and sensitive hypothesis class induces a cost that has bounded distance to the ground truth cost in (4.1).

Theorem 43 (Bounded Approximation)**.** *Suppose that the following conditions hold for all $i \in [n]$:*

(i) The hypothesis class $\mathcal{H}_{\mathcal{B}_i}$ is η_i-misspecified, and ν_i-sensitive.

(ii) The map learning problem is regular.

(iii) For all $x \in \mathcal{X}_i$, $z_i \mapsto f_i(x, z_i)$ is L_{z_i}-Lipschitz continuous.

Then, the bound

$$|F_{\widehat{\beta}_i}(x) - F_i(x)| \leq \zeta_i L_{z_i} + L_{z_i} \upsilon_i C_i \sqrt{\frac{\log(m)(\ell_i + \log(1/\delta))}{m}}, \tag{4.22}$$

holds with probability $1 - \delta$ for any $x \in \mathcal{X}$.

Proof. We observe that for any fixed $x \in \mathcal{X}$, we have that

$$|F_{\widehat{\beta}_i}(x) - F_i(x)| \leq |F_{\widehat{\beta}_i}(x) - F_{\beta_i^*}(x)| + |F_{\beta_i^*}(x) - F_i(x)|.$$

The first term describes our statistical error at x. We denote $\Pi(D_{\widehat{\beta}_i}, D_{\beta_i^*})$ as a coupling on $\mathcal{P}(\mathbb{R}^{k_i})$ so that

$$\begin{aligned}
|F_{\widehat{\beta}_i}(x) - F_{\beta_i^*}(x)| &= \left| \inf_{\Pi(D_{\widehat{\beta}_i}(x), D_{\beta_i^*}(x))} \mathbb{E}_{(z,z')\sim\Pi(D_{\widehat{\beta}_i}(x), D_{\beta_i^*}(x))} \left(f(x,z) - f(x,z')\right) \right| \\
&\leq \inf_{\Pi(D_{\widehat{\beta}_i}(x), D_{\beta_i^*}(x))} \mathbb{E}_{(z,z')\sim\Pi(D_{\widehat{\beta}_i}(x), D_{\beta_i^*}(x))} \left|f(x,z) - f(x,z')\right| \\
&\leq L_{z_i} \left(\inf_{\Pi(D_{\widehat{\beta}_i}(x), D_{\beta_i^*}(x))} \mathbb{E}_{(z,z')\sim\Pi(D_{\widehat{\beta_i}}(x), D_{\beta_i^*}(x))} \|z_i - z_i^{'}\| \right) \\
&= L_{z_i} W_1(D_{\widehat{\beta}_i}(x), D_{\beta_i^*}(x)) \\
&\leq L_{z_i} \varepsilon_i \|\widehat{\beta}_i - \beta_i^*\|\,.
\end{aligned}$$

By similar argument, we find that $|F_{\beta_i^*}(x) - F_i(x)| \leq L_{z_i} W_1(D_{\beta_i^*}(x), D_i(x)) \leq L_{z_i}\zeta_i$. In combining, we get $|F_{\widehat{\beta}_i}(x) - F_i(x)| \leq L_{z_i} \upsilon i \|\widehat{\beta}_i - \beta_i^*\| + L_{z_i}\zeta_i$. Lastly, $\|\widehat{\beta}_i - \beta_i^*\|$ can be bounded as in Theorem 41.

Regarding the second bound, we have that

$$\begin{aligned}
F_i(\widehat{x}) - F_i(x^*) = &\left[F_i(\widehat{x}) - F_{\beta_i^*}(\widehat{x})\right] + \left[F_{\beta_i^*}(\widehat{x}) - F_{\widehat{\beta}_i}(\widehat{x})\right] + \left[F_{\widehat{\beta}_i}(\widehat{x}) - F_{\widehat{\beta}_i}(x^{**})\right] \\
&+ \left[F_{\widehat{\beta}_i}(x^{**}) - F_{\beta_i^*}(x^{**})\right] + \left[F_{\beta_i^*}(x^{**}) - F_{\beta_i^*}(x^*)\right] + \left[F_{\beta_i^*}(x^*) - F_i(x^*)\right] \\
\leq &\, 2\|F_i - F_{\beta_i^*}\|_\infty + 2\|F_{\beta_i^*} - F_{\widehat{\beta}_i}\|_\infty + \left[F_{\widehat{\beta}_i}(\widehat{x}) - F_{\widehat{\beta}_i}(x^{**})\right] + \left[F_{\beta_i^*}(x^{**}) - F_{\beta_i^*}(x^*)\right]
\end{aligned}$$

where $x_i^{**} \in \mathcal{X}^{**}$, where $\mathcal{X}^{**}$ is the set of equilibria of

$$x_i^{**} \in \operatorname*{arg\,min}_{x_i \in \mathcal{X}_i} F_{\beta_i^*}(x_i, x_{-i}^{**}), \quad i \in [n] \tag{4.23}$$

where

$$F_{\beta_i^*}(x_i, x_{-i}^{**}) := \mathop{\mathbb{E}}_{z_i \sim D_{\beta_i^*}(x_i, x_{-i}^{**})} f_i(x_i, x_{-i}^{**}, z_i)\,.$$

Then,

$$\begin{aligned} F_{\widehat{\beta}_i}(\widehat{x}) - F_{\widehat{\beta}_i}(x^{**}) &\le \Big[F_{\widehat{\beta}_i}(\widehat{x}_i, \widehat{x}_{-i}) - F_{\widehat{\beta}_i}(x_i^{**}, \widehat{x}_{-i})\Big] + \Big[F_{\widehat{\beta}_i}(x_i^{**}, \widehat{x}_{-i}) - F_{\widehat{\beta}_i}(x_i^{**}, x_{-i}^{**})\Big] \\ &\le L_i^{\widehat{\beta}} \|\widehat{x}_i - x_i^{**}\| + L_{-i}^{\widehat{\beta}} \|\widehat{x}_{-i} - x_{-i}^{**}\| \\ &\le \sqrt{2} \max\{L_i^{\widehat{\beta}}, L_{-i}^{\widehat{\beta}}\} \|\widehat{x} - x^{**}\| \end{aligned}$$

where we have used the inequality $\sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}$ for some $a, b \ge 0$.

Next, consider

$$\begin{aligned} F_{\beta_i^*}(x^{**}) - F_{\beta_i^*}(x^*) &= [F_{\beta_i^*}(x_i^{**}, x_{-i}^{**}) - F_{\beta_i^*}(x_i^{**}, x_{-i}^*)] + [F_{\beta_i^*}(x_i^{**}, x_{-i}^*) - F_{\beta_i^*}(x_i^*, x_{-i}^*)] \\ &\le L_{-i}^{\beta^*} \|x_{-i}^{**} - x_{-i}^*\| + L_i^{\beta^*} \|x_i^{**} - x_i^*\| \\ &\le \sqrt{2} \max\{L_i^{\beta^*}, L_{-i}^{\beta^*}\} \|x^{**} - x^*\|\,. \end{aligned}$$

Combining the bounds yields

$$\begin{aligned} F_i(\widehat{x}) - F_i(x^*) &\le 2\|F_i - F_{\beta_i^*}\|_\infty + 2\|F_{\beta_i^*} - F_{\widehat{\beta}_i}\|_\infty + \sqrt{2}\max\{L_i^{\widehat{\beta}}, L_{-i}^{\widehat{\beta}}\}\|\widehat{x} - x^{**}\| \\ &\quad + \sqrt{2}\max\{L_i^{\beta^*}, L_{-i}^{\beta^*}\}\|x^{**} - x^*\| \\ &\le 2\zeta_i L_{z_i} + 2\upsilon L_{z_i}\|\widehat{\beta}_i - \beta_i^*\| + \sqrt{2}(\max\{L_i^{\widehat{\beta}}, L_{-i}^{\widehat{\beta}}\} + \max\{L_i^{\beta^*}, L_{-i}^{\beta^*}\})\mathsf{diam}(\mathcal{X}) \\ &\le 2\zeta_i L_{z_i} + 2\upsilon L_{z_i}\|\widehat{\beta}_i - \beta_i^*\| + 2\sqrt{2}\bar{L}_i \mathsf{diam}(\mathcal{X}) \end{aligned}$$

Then, (4.25) follows using the bound on $\|\widehat{\beta}_i - \beta_i^*\|$ from Theorem 41. □

Note that since each F_{β_i} is assumed to be continuously differentiable and $\mathcal{X} \subseteq \mathbb{R}^d$ is compact, then $x \mapsto F_{\beta_i}(x)$ is L_{β_i}-Lipschitz continuous on $\mathcal{X}$ with

$$L_{\beta_i} = \max_{x \in \mathcal{X}} \|\nabla F_{\beta_i}(x)\|. \tag{4.24}$$

Leveraging this fact allows us to demonstrate that the excess cost can be bounded—an analog of the main result in [39].

Corollary 44. *Suppose that the hypothesis of Theorem 43 holds. Then,*

$$|F_i(\widehat{x}) - F_i(x^*)| \leq 2\zeta_i L_{z_i} + 2\upsilon_i C_i \sqrt{\frac{\log(m)(\ell_i + \log(1/\delta))}{m}} + 2\max\{L_{\widehat{\beta}_i}, L_{\beta_i^*}\}\mathsf{diam}(\mathcal{X}_{-i}) \tag{4.25}$$

hold with probability $1-\delta$ *for any* $\widehat{x} \in \texttt{NASH}(G_{\widehat{\beta}}, \mathcal{X})$ *and* $x^* \in \texttt{NASH}(G, \mathcal{X})$*, where* $\mathcal{X}_{-i} = \prod_{j \neq i} \mathcal{X}_j$.

Proof. Observe that

$$\begin{aligned} F_i(\widehat{x}) - F_i(x^*) &\leq \left[F_i(\widehat{x}) - F_{\beta_i^*}(\widehat{x})\right] + \left[F_{\beta_i^*}(\widehat{x}) - F_{\widehat{\beta}_i}(\widehat{x})\right] + \left[F_{\widehat{\beta}_i}(\widehat{x}) - F_{\widehat{\beta}_i}(x^{**})\right] \\ &\quad + \left[F_{\widehat{\beta}_i}(x^{**}) - F_{\beta_i^*}(x^{**})\right] + \left[F_{\beta_i^*}(x^{**}) - F_{\beta_i^*}(x^*)\right] + \left[F_{\beta_i^*}(x^*) - F_i(x^*)\right] \\ &\leq 2\|F_i - F_{\beta_i^*}\|_\infty + 2\|F_{\beta_i^*} - F_{\widehat{\beta}_i}\|_\infty + \left[F_{\widehat{\beta}_i}(\widehat{x}) - F_{\widehat{\beta}_i}(x^{**})\right] + \left[F_{\beta_i^*}(x^{**}) - F_{\beta_i^*}(x^*)\right] \end{aligned}$$

where $x^{**} \in \mathcal{X}$ is the Nash equilibrium satisfying

$$x_i^{**} \in \operatorname*{arg\,min}_{x_i \in \mathcal{X}_i} F_{\beta_i^*}(x_i, x_{-i}^{**}), \quad i \in [n]. \tag{4.26}$$

It follows from (4.24) that

$$\begin{aligned} F_{\widehat{\beta}_i}(\widehat{x}) - F_{\widehat{\beta}_i}(x^{**}) &\leq \left[F_{\widehat{\beta}_i}(\widehat{x}_i, \widehat{x}_{-i}) - F_{\widehat{\beta}_i}(x_i^{**}, \widehat{x}_{-i})\right] + \left[F_{\widehat{\beta}_i}(x_i^{**}, \widehat{x}_{-i}) - F_{\widehat{\beta}_i}(x_i^{**}, x_{-i}^{**})\right] \\ &\leq F_{\widehat{\beta}_i}(x_i^{**}, \widehat{x}_{-i}) - F_{\widehat{\beta}_i}(x_i^{**}, x_{-i}^{**}) \\ &\leq L_{\widehat{\beta}_i}\|\widehat{x}_{-i} - x_{-i}^{**}\|. \end{aligned}$$

Similarly,

$$\begin{aligned} F_{\beta_i^*}(x^{**}) - F_{\beta_i^*}(x^*) &= \left[F_{\beta_i^*}(x_i^{**}, x_{-i}^{**}) - F_{\beta_i^*}(x_i^{**}, x_{-i}^*)\right] + \left[F_{\beta_i^*}(x_i^{**}, x_{-i}^*) - F_{\beta_i^*}(x_i^*, x_{-i}^*)\right] \\ &\leq F_{\beta_i^*}(x_i^{**}, x_{-i}^*) - F_{\beta_i^*}(x_i^*, x_{-i}^*) \\ &\leq L_{\beta^*}\|x_{-i}^{**} - x_{-i}^*\|. \end{aligned}$$

Combining the bounds yields

$$
\begin{aligned}
&F_i(\widehat{x}) - F_i(x^*) \\
&\leq 2\|F_i - F_{\beta_i^*}\|_\infty + 2\|F_{\beta_i^*} - F_{\widehat{\beta}_i}\|_\infty + L_{\widehat{\beta}}\|\widehat{x}_{-i} - x_{-i}^{**}\| + L_{\beta^*}\|x_{-i}^* - x_{-i}^{**}\| \\
&\leq 2\zeta_i L_{z_i} + 2\upsilon_i L_{z_i}\|\widehat{\beta}_i - \beta_i^*\| + L_{\widehat{\beta}}\mathsf{diam}(\mathcal{X}_{-i}) + L_{\beta^*}\mathsf{diam}(\mathcal{X}_{-i}) \\
&\leq 2\zeta_i L_{z_i} + 2\upsilon_i L_{z_i}\|\widehat{\beta}_i - \beta_i^*\| + 2\max\{L_{\widehat{\beta}_i}, L_{\beta_i^*}\}\mathsf{diam}(\mathcal{X}_{-i})
\end{aligned}
$$

Then, (4.25) follows using the bound on $\|\widehat{\beta}_i - \beta_i^*\|$ from Theorem 41. □

The analysis in this section demonstrates that the estimation procedure in Algorithm 1 yields a cost function that approximates the original cost in (4.1) with an error the decreases as the number of samples increases. Furthermore, this bound exists independent of the conditioning of the Nash equilibrium problem we solve in the optimization phase. We note that (4.22) is similar to the result in [39], but it is based on a different definition of regular problem (see Definition 13); the bound (4.25) is unique to this paper.

In the section that follows, we examine a family of hypothesis classes that allows the approximated game to be monotone, and provide suitable algorithms for solving them with convergence guarantees.

4.2 Solving Strongly-monotone Decision-dependent Games

Since the agents lack full knowledge of the system and hence the ground truth distributional map D_i in (4.1), we cannot hope to enforce that D_i satisfy any assumptions to encourage tractability of our optimization problem. We can however impose conditions on the hypothesis class $\mathcal{H}_{\mathcal{B}_i}$, which is chosen by the agents. To successfully find a Nash equilibrium of the approximate problem in (4.12), it will be crucial that agents choose a class that balances expressiveness of the system (thereby making η_i small) with tractability of the optimization.

Perhaps the simplest model capable of achieving this goal is the location-scale family [45, 47, 65]. In our setting, a location scale family parameterization for agent i is a distributional map D_{B_i}

having matrix parameter $B_i \in \mathbb{R}^{k_i \times d}$ where $z_i \sim D_{B_i}$ if and only if

$$z_i \overset{d}{=} \xi_i + B_i x \tag{4.27}$$

for stationary random variable $\xi_i \sim D_{\xi_i}$. We note that this parameterization can be written alternatively as $z_i \overset{d}{=} \xi_i + B_i^i x_i + B_{-i}^i x_{-i}$, where $B_i^i \in \mathbb{R}^{k_i \times d_i}$ and $B_{-i}^i \in \mathbb{R}^{k_i \times (d-d_i)}$ are block matrices such that $B_i x = B_i^i x_i + B_{-i}^i x_{-i}$ due to linearity. The resulting partial gradient has the form

$$\nabla_i F_i(x) = \mathbb{E}_{z_i \sim D(x)} \left[\nabla_i f_i(x, z_i) + (B_i^i)^T \nabla_{z_i} f_i(x, z_i) \right],$$

which is typically much simpler to analyze than alternative models. Intuitively, this model allows us to express z_i as the sum of a stationary random variable from a base distribution with a linear factor depending on x, where the matrix parameter B_i weights the responsiveness of the population to the agents decisions.

This model is particularly appealing as guarantees for learning B_i are known and established in Proposition 42. Moreover, the matter of expressiveness is due to the fact that location scale families are a particular instance of strategic regression [49, 39], in which member of the population interact with agents by modifying their stationary data (such as features in a learning task) ξ_i in an optimal way upon observing x:

$$z_i \overset{d}{=} \arg\min_{y} \left[-u_{\beta_i}(x, y) + \frac{1}{2}\|y - \xi_i\|^2 \right],$$

where u_{β_i} is a utility function parameterized by $\beta_i \in \mathcal{B}_i$ corresponding to the utility that members of the population derive from changing their data in response to the decisions in x; and the quadratic term $1/2\|y - \xi_i\|^2$ is the cost of changing their data from ξ_i to y. Indeed when $u_{\beta_i}(x,) = \langle y, B_i x \rangle$ for $\beta_i = B_i \in \mathbb{R}^{k_i \times d}$, we recover the form above.

Furthermore, location scale families immediately satisfy several of the assumption required for further analysis. In particular, it is known that Sensitivity (Definition 15) holds with $\upsilon_i = \max_{x \in \mathcal{X}} \|x\|^2$, Lipschitz continuity of $x \mapsto D_{B_i}$ holds with $\nu_i = \|B_i\|^2$, and Lipschitz continuity of G_{B_i} holds due to the following result.

Lemma 45. *(Lipschitz Gradient, [47]) Suppose that D_{β_i} is such that $z \overset{d}{=} B_i x + \xi_i$ with $\beta_i = B_i$, and that for each $i \in [n]$ there exists $\varphi_i \geq 0$ such that $(x, z_i) \mapsto \nabla_{i,z_i} f_i(x, z_i)$ is φ_i-Lipschitz continuous. Then G_{β_i} is L-Lipschitz continuous with*

$$L := \sqrt{\sum_{i=1}^{n} \varphi_i^2 \max\{1, \|B_i^i\|^2\}(1 + \|B_i\|^2)} \,. \tag{4.28}$$

□

Strong monotonicity will follow from Theorem 39 provided that G_{β_i} satisfy the remaining hypothesis on the f_{β_i}—which tends to be on a case-by-case basis. We will not require that G_{β_i} use this parameterization in our analysis, however we can proceed with knowledge a model class satisfying our hypothesis does exist.

4.2.1 Distributed Gradient-based Method

In our optimization phase, we seek to use a gradient-based algorithm that respects the agent's communication structure with the system. For the sake of readability, we will suppress the β_i subscript and instead refer to quantities G_i keeping in mind that they will correspond to the approximate Nash equilibrium problem in (4.12) with solution $\widehat{x}$.

We will assume that each agent has access to an estimator of the gradient $\nabla_i F_i$ and is capable of projecting onto their decision set $\mathcal{X}_i$. In the constant step-size setup, each agent chooses a rate $\omega_i > 0$ and performs the update

$$x_i^{t+1} = \mathsf{proj}_{\mathcal{X}_i}\left[x_i^t - \omega_i^{-1} g_i^t\right],$$

where g_i^t is a stochastic gradient estimator for $\nabla_i F_i$ used at iteration t, which is then reported to the system and made available to all agents. For the sake of analysis, we will assume without loss of generality that the steps-sizes satisfy the ordering

$$\omega_1 \geq \omega_2 \geq \ \dots \ \geq \omega_n$$

and hence $\omega_1 = \max_{i\in[n]} \omega_i$ and $\omega_n = \min_{i\in[n]} \omega_i$. The collective update can be written compactly as

$$x^{t+1} = \mathsf{proj}_{\mathcal{X},W}\left[x^t - W^{-1}g^t\right], \tag{4.29}$$

where $W = \mathsf{diag}(\omega_1 \mathbb{1}_{d_1}, \ldots, \omega_n \mathbb{1}_{d_n})$ and g^t is an estimator for $G(x^t)$ at iteration t. Convergence of this procedure hinges on the following assumptions.

Assumption 22 (Monotone and Lipschitz Gradient). *The gradient function $G : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ is γ-strongly monotone and L-Lipschitz continuous.*

Assumption 23 (Stochastic Framework). *Let $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ with elements*

$$\mathcal{F}_t = \sigma(g^\tau, \tau \leq t) \tag{4.30}$$

be the natural filtration of the Borel σ-algebra over $\mathbb{R}^d$ with respect to g^t, and use the short-hand notation $\mathbb{E}_t[\cdot] := \mathbb{E}_{z\sim D(x^t)}[\cdot|\mathcal{F}_t]$ as the conditional expectation over the product distribution $D(x^t) = \prod_{i=1}^n D_i(x^t)$. There exist bounded sequences $\{\rho^t\}_{t\geq 0}, \{\sigma^t\}_{t\geq 0} \subseteq \mathbb{R}_+$ such that

$$\textbf{\textit{(Bias)}} \qquad \|\mathbb{E}_t g^t - G(x^t)\| \leq \rho^t$$

$$\textbf{\textit{(Variance)}} \qquad \mathbb{E}_t\|g^t - \mathbb{E}_t g^t\|^2 \leq (\sigma^t)^2$$

where $\rho^t \leq \rho$ and $\sigma^2 \leq \sigma$ for all $t \geq 0$.

Assumption 22 is standard for guaranteeing convergence of gradient play [25], and the uniformly bounded variance component of Assumption 23 is standard for convergence for stochastic algorithms. As we will show shortly, convergence with bias is possible and the result reduces to the unbiased case when $\rho^t = 0$. The next result will quantify the one-step improvement of (4.2.1).

Lemma 46 (One-Step Improvement). *Let Assumptions 22 and 23 hold. Then, the sequence generated by iteration (4.29) satisfies:*

$$\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 \leq \frac{\omega_1}{\omega_n + \gamma}\|x^t - \widehat{x}\|_W^2 \; + \frac{2\omega_1\left(\omega_1\rho^2 + \gamma\sigma^2\right)}{\gamma\omega_n\left(\omega_n + \gamma\right)}$$

for all $t \geq 0$, provided that $\omega_1/\omega_n^2 \leq \gamma/(4L^2)$.

Proof. Consider the function $\varphi : \mathbb{R}^d \to \mathbb{R}$ defined by $\varphi(y) = \frac{1}{2}\|x^t - W_i^{-1} g_i^t - y\|_W^2$ for all $y \in \mathcal{X}$. Then, φ is ω_n-strongly convex over $\mathcal{X}$ and has a unique minimizer $x^{t+1} \in \mathcal{X}$. This implies that:

$$\varphi(x^*) \geq \varphi(x^{t+1}) + \langle x^* - x^{t+1}, \nabla\varphi(x^{t+1})\rangle + \frac{\omega_n}{2}\|x^{t+1} - x^*\|^2 .$$

Since $\langle x - x^{t+1}, \nabla\varphi(x^{t+1})\rangle \geq 0$ for all $x \in \mathcal{X}$, we obtain

$$\omega_n \|x_i^{t+1} - x_i^*\| \leq \|x_i^t - \eta_i g_i^t - x_i^*\|_W^2 - \|x_i^t - \eta_i g_i^t - x_i^{k+1}\|_W^2 .$$

It follows that

$$\frac{\omega_n}{\omega_1}\|x^{t+1} - \widehat{x}\|_W^2 \leq \|x^t - \widehat{x}\|_W^2 - \|x^t - x_i^{t+1}\|_W^2 - 2\langle x^t - \widehat{x}, g^t\rangle + 2\eta_i\langle x^t - x^{t+1}, g^t\rangle .$$

We now consider the above in the conditional expectation $\mathbb{E}_t \cdot := \mathbb{E}_{z_i \sim D(x_t)}[\,\cdot\,|\mathcal{F}_t]$ with $\mathcal{F}_t = \sigma(g^t, \tau \geq t)$. We find that

$$\begin{aligned}
&\frac{\omega_n}{\omega_1}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 \\
&\leq \mathbb{E}_t\|x^t - \widehat{x}\|_W^2 - \mathbb{E}_t\|x^t - x^{t+1}\|_W^2 - 2\mathbb{E}_t\langle x^t - \widehat{x}, g_i^t\rangle - 2\mathbb{E}_t\langle x^{t+1} - x^t, g^t\rangle \\
&= \|x^t - \widehat{x}\|_W^2 - \mathbb{E}_t\|x^t - x^{t+1}\|_W^2 - 2\langle x^t - \widehat{x}, \mu^t\rangle - 2\mathbb{E}_t\langle x^{t+1} - x^t, g^t\rangle \\
&= \|x^t - \widehat{x}\|_W^2 - \mathbb{E}_t\|x^t - x^{t+1}\|_W^2 + 2\mathbb{E}_t\langle x^t - x^{t+1}, g^t - \mu^t\rangle + 2\mathbb{E}_t\langle \widehat{x} - x^{t+1}, \mu^t\rangle \\
&= \|x^t - \widehat{x}\|_W^2 - \mathbb{E}_t\|x^t - x^{t+1}\|_W^2 - 2\langle x^{t+1} - \widehat{x}, G(x^{t+1})\rangle + 2\mathbb{E}_t\langle \widehat{x} - x^{t+1}, \mu^t - G(x^{t+1})\rangle \\
&\quad + 2\mathbb{E}_t\langle x^t - x^{t+1}, g^t - \mu^t\rangle .
\end{aligned}$$

To proceed, we bound the inner product terms. Using strong monotonicity, we have that

$$\mathbb{E}_t\langle \widehat{x} - x^{t+1}, G(x^{t+1})\rangle \geq \gamma\mathbb{E}_t\|x^{t+1} - \widehat{x}\|^2 \geq \frac{\gamma}{\omega_1}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 .$$

Furthermore, we observe that

$$\mathbb{E}_t\langle \widehat{x} - x_i^{t+1}, \mu^t - G(x^{t+1})\rangle = \mathbb{E}_t\langle \widehat{x} - x^{t+1}, \mu^t - G(x^t)\rangle + \mathbb{E}_t\langle \widehat{x} - x^{t+1}, G(x^t) - G(x^{t+1})\rangle .$$

To bound the remaining terms, we use arguments based on a weighted Young's inequality. Let

$\Delta_1, \Delta_2, \Delta_3 > 0$ be fixed constants. It follows that

$$\begin{aligned}
2\mathbb{E}_t\langle x^t - x^{t+1}, g^t - \mu^t\rangle &\leq \Delta_1 \mathbb{E}_t\|x^{t+1} - x^t\|^2 + \frac{1}{\Delta_1}\mathbb{E}_t\|g^t - \mu^t\|^2 \\
&\leq \frac{\Delta_1}{\omega_n}\mathbb{E}_t\|x^{t+1} - x^t\|_W^2 + \frac{1}{\Delta_1}\sum_{i=1}^n \mathbb{E}_t\|g^t - \mu^t\|^2 \\
&\leq \frac{\Delta_1}{\omega_n}\mathbb{E}_t\|x^{t+1} - x^t\|_W^2 + \frac{1}{\Delta_1}\sum_{i=1}^n \sigma_i^2 \\
&\leq \frac{\Delta_1}{\omega_n}\mathbb{E}_t\|x^{t+1} - x^t\|_W^2 + \frac{\sigma^2}{\Delta_1},
\end{aligned}$$

and

$$\begin{aligned}
2\mathbb{E}_t\langle \widehat{x} - x^{t+1}, \mu^t - G(x^t)\rangle &\leq \Delta_2 \mathbb{E}_t\|x^{t+1} - \widehat{x}\|^2 + \frac{1}{\Delta_2}\mathbb{E}_t\|\mu^t - G(x^t)\|^2 \\
&\leq \frac{\Delta_2}{\omega_n}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 + \frac{1}{\Delta_2}\sum_{i=1}^n \mathbb{E}_t\|\mu^t - G(x^t)\|^2 \\
&\leq \frac{\Delta_2}{\omega_n}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 + \frac{1}{\Delta_2}\sum_{i=1}^n \rho_i^2 \\
&\leq \frac{\Delta_2}{\omega_n}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 + \frac{\rho^2}{\Delta_2}.
\end{aligned}$$

Additionally, we have that

$$\begin{aligned}
2\mathbb{E}_t\langle \widehat{x} - x^{t+1}, G(x^t) - G(x^{t+1})\rangle &\leq \Delta_3 \mathbb{E}_t\|x^{t+1} - \widehat{x}\|^2 + \frac{1}{\Delta_3}\mathbb{E}_t\|G(x^t) - G(x^{t+1})\|^2 \\
&\leq \frac{\Delta_3}{\omega_n}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 + \frac{L^2}{\Delta_3}\mathbb{E}_t\|x^{t+1} - x^t\|^2 \\
&\leq \frac{\Delta_3}{\omega_n}\mathbb{E}_t\|x^{t+1} - \widehat{x}\|_W^2 + \frac{L^2}{\omega_n\Delta_3}\mathbb{E}_t\|x^{t+1} - x^t\|_W^2.
\end{aligned}$$

Combining these estimates yields

$$\begin{aligned}\frac{\omega_n}{\omega_1}\mathbb{E}_t\|x^{t+1}-\widehat{x}\|_W^2 &\le \|x^t-\widehat{x}\|_W^2 - \mathbb{E}_t\|x^{t+1}-x^t\|_W^2 - \frac{2\gamma}{\omega_1}\mathbb{E}_t\|x^{t+1}-\widehat{x}\|_W^2\\ &\quad+\left(\frac{\Delta_1}{\omega_n}\mathbb{E}_t\|x^{t+1}-x^t\|_W^2+\frac{\sigma^2}{\Delta_1}\right)+\left(\frac{\Delta_2}{\omega_n}\mathbb{E}_t\|x^{t+1}-\widehat{x}\|_W^2+\frac{\rho^2}{\Delta_2}\right)\\ &\quad+\left(\frac{\Delta_3}{\omega_n}\mathbb{E}_t\|x^{t+1}-\widehat{x}\|_W^2+\frac{L^2}{\omega_n\Delta_3}\mathbb{E}_t\|x^{t+1}-x^t\|_W^2\right)\\ &=\|x^t-\widehat{x}\|_W^2+\left(\frac{\Delta_1}{\omega_n}+\frac{L^2}{\omega_n\Delta_3}-1\right)\mathbb{E}_t\|x^{t+1}-x^t\|_W^2\\ &\quad+\left(\frac{\Delta_2}{\omega_n}+\frac{\Delta_3}{\omega_n}-\frac{2\gamma}{\omega_1}\right)\mathbb{E}_t\|x^{t+1}-\widehat{x}\|_W^2+\left(\frac{\sigma^2}{\Delta_1}+\frac{\rho^2}{\Delta_2}\right)\end{aligned}$$

and simplifying gives

$$\begin{aligned}\left(\frac{\omega_n}{\omega_1}+\frac{2\gamma}{\omega_1}-\frac{\Delta_2}{\omega_n}-\frac{\Delta_3}{\omega_n}\right)\mathbb{E}_t\|x^{t+1}-\widehat{x}\|_W^2 &\le \|x^t-\widehat{x}\|_W^2+\left(\frac{\sigma^2}{\Delta_1}+\frac{\rho^2}{\Delta_2}\right)\\ &\quad+\left(\frac{\Delta_1}{\omega_n}+\frac{L^2}{\omega_n\Delta_3}-1\right)\mathbb{E}_t\|x^{t+1}-x^t\|_W^2.\end{aligned}$$

To proceed, we choose $\Delta_2=\Delta_3=\frac{\gamma\omega_n}{2\omega_1}$ and $\Delta_1=\omega_n-2\omega_1L^2/(\gamma\omega_n)$ to ensure that the coefficient on the $\mathbb{E}_t\|x^{t+1}-x^t\|_W^2$ term is zero. Furthermore, enforcing that $\frac{\omega_1}{\omega_n^2}\le\frac{\gamma}{4L^2}$ guarantees that $\Delta_1^{-1}\le 2\omega_n^{-1}$. Hence the variance term is finite. Substituting these values and simplifying yields the result. □

We note that setting $\omega_i=\omega$ for some $\omega>0$ recovers the result in [47, Theorem 15]. Following this one-step analysis, we can show convergence to a neighborhood of the Nash equilibrium.

Theorem 47 (Neighborhood Convergence)**.** *Let Assumptions 22 and 23 hold, and suppose that* $(\omega_1-\omega_n)<\gamma$. *Then,*

$$\limsup_{t\to\infty}\mathbb{E}\|x^t-\widehat{x}\|^2\le\frac{2\omega_1\left(\omega_1\rho^2+\gamma\sigma^2\right)}{\gamma\omega_n\left(\omega_1-\omega_n+\gamma\right)}. \tag{4.31}$$

Proof. For notational convenience, we will use the short-hand notation $e^t:=\|x^t-\widehat{x}\|_W^2$, $c=\omega_1(\gamma+\omega_n)^{-1}$, and

$$A=2\frac{\gamma\sigma^2+\omega_1\rho^2}{\gamma\omega_n}.$$

Hence, the result in Lemma 46 can be written compactly as

$$\mathbb{E}_{t-1}e^t\le ce^{t-1}+cA.$$

By recursively applying this result and applying the law of total expectation, we find that

$$\mathbb{E}e^t \leq c^t e^0 + cA \sum_{j=1}^{t-1} c^j \leq c^t e^0 + cA \frac{1-c^t}{1-c}.$$

□

The result shows that the algorithm converges linearly to a neighborhood of the Nash equilibrium $\widehat{x}$, where the radius of the neighborhood is dictated by the step-size, variance, and bias bounds. When $\rho = \sigma = 0$, we retrieve linear convergence. In order to converge to $\widehat{x}$ directly, we will require a decaying step-size policy. For example, we consider the following policy:

$$\omega^t = \frac{\gamma(r+t-2)}{2} \tag{4.32}$$

for fixed constant $r > 2$, which we assumed to be shared by all agents. Hence, the decaying step-size update is given by

$$x^{t+1} = \mathsf{proj}_{\mathcal{X}} \left[x^t - (\omega^t)^{-1} g^t \right]. \tag{4.33}$$

In the theorem that follows, we show that this sequence converges to $\widehat{x}$ provided that the bias shares an asymptotic rate with $(\omega^t)^{-1}$.

Theorem 48 (Convergence)**.** *Suppose that Assumptions 22 and 23 hold and that there exists* $\bar{\rho}, s > 0$ *such that*

$$\|\mathbb{E}_t g^t - G(x^t)\| \leq \frac{\bar{\rho}}{s+t} \tag{4.34}$$

for all $t \geq 0$*. Then,*

$$\mathbb{E}\|x^t - \widehat{x}\|^2 \leq \frac{M}{\gamma^2(r+t)} \tag{4.35}$$

where

$$M = \max\left\{ \gamma^2 r \|x^0 - \widehat{x}\|^2, 4\bar{\rho}^2 \max\left\{\frac{r}{s}, 1\right\} + \frac{8r\sigma^2}{r-2} \right\}.$$

Proof. Fix $t \geq 0$. For notational convenience, we will denote $e^t = \|x^t - \widehat{x}\|^2$. Replacing the step-size matrix in Lemma 46 with $W = \omega^t I_d$ yields

$$\mathbb{E}_t e^{t+1} \leq \frac{\omega^t}{\omega^t + \gamma} e^t + \frac{2\sigma^2}{\omega^t(\omega^t + \gamma)} + \frac{2(\rho^t)^2}{\gamma(\omega^t + \gamma)}. \tag{4.36}$$

To proceed, we will use the observation that

$$\frac{1}{(s+t)(r+t)} = \frac{r+t}{(s+t)(r+t)^2} \leq \frac{\max\{\frac{r}{s}, 1\}}{(r+t)^2} \tag{4.37}$$

and

$$\frac{1}{(r+t)(r+t-2)} \leq \frac{\frac{r}{r-2}}{(r+t)^2}. \tag{4.38}$$

By substituting our expression for ω^t, ρ^t, and e^t into (4.36) we obtain

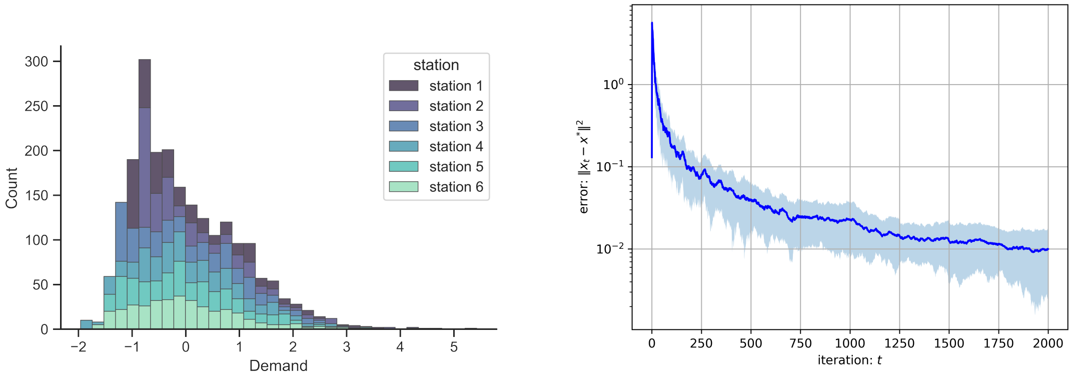
$$\begin{aligned}
\mathbb{E}_t e^{t+1} &\leq \frac{r+t-2}{\gamma^2(r+t)^2} M + \frac{8\sigma^2}{\gamma^2(r+t-2)(r+t)} + \frac{4\bar{\rho}}{\gamma^2(s+t)(r+t)} \\
&\leq \frac{r+t-2}{\gamma^2(r+t)^2} M + \frac{8\sigma^2\left(\frac{r}{r-2}\right)}{\gamma^2(r+t)^2} + \frac{4\bar{\rho}\max\left\{\frac{r}{s}, 1\right\}}{\gamma^2(r+t)^2} \\
&= \frac{r+t-1}{\gamma^2(r+t)^2} M + \frac{-M + 8\sigma^2\left(\frac{r}{r-2}\right) + 4\bar{\rho}\max\left\{\frac{r}{s}, 1\right\}}{\gamma^2(r+t)^2} \\
&\leq \frac{r+t-1}{\gamma^2(r+t)^2} M \\
&\leq \frac{M}{\gamma^2(r+t+1)}.
\end{aligned}$$

Here, the last steps follow from construction of M, and the fact that $(r+t+1)(r+t-1) \leq (r+t)^2$. □

4.3 Numerical Experiments on Electric Vehicle Charging

In this section, we consider a competitive game between n distinct electric vehicle charging station operators, where stations are equipped with renewable power sources. The goal of each player is to set prices to maximize their own profit in a system where demand for their station will change in response to the prices set by other competing stations as well. The cost function (negative profit) takes the form

$$f_i(x, z_i) = \underbrace{-z_i x_i + \frac{\lambda_i}{2} x_i^2}_{\text{service profit}} - \underbrace{p_w \phi(w_i - z_i)}_{\text{renewable profit}} + \underbrace{p_r \phi(z_i - w_i)}_{\text{operational cost}}$$



(a)

(b)

Figure 4.1: Data and results from numerical experiments: (a) Standardized demand data for six medium demand EVCS's consisting of either 2 or 6 ports and port power values of 50, 150, and 350 kWh. Standardization maps raw demand instances to instances of demand that are deviations from the average at each station; (b) Expected error curve and confidence interval for regularized stochastic gradient descent with decaying step size for a location-scale model.

where $\phi(y) = \log(1 + \exp(y))$ for all $y \in \mathbb{R}$. The renewable profit and operational cost terms allow us to describe the trade-off between profit from renewable power generation sold to the grid at rate p_w, and surplus power required from the grid to meet demand at rate p_r. To set prices, we can formulate a Nash equilibrium problem over the expected costs $F_i(x) = \mathbb{E}_{z_i \sim D_i(x)}[f_i(x, z_i)]$ for $i \in [n]$ and $x \in \mathcal{X} = \Pi_{i=1}^n \mathcal{X}_i$, where $\mathcal{X}_i = [p_w, p_r]$ is the interval of price values between the wholesale and retail price.

Since the set of reasonable prices will be quite small, we hypothesize that the the price and demand have a linear relationship of the form $z_i \overset{d}{=} \xi_i + \langle b_i, x_i \rangle$ where $b_i \in \mathbb{R}^n$ with $\xi_i \sim D_{\xi_i}$ corresponding to the base demand. Since we have a simple model, the first and second derivatives can be computed in closed form, and the relevant constants can be computed directly. Indeed, we find that the hypothesis of Theorem 39 are satisfied with $\lambda = \min_i \lambda_i$ which we set to 1, $L_i = 1$,and $\gamma_i = \|b_i\|^2$. We conclude that $G : \mathbb{R}^n \to \mathbb{R}^n$ is $\alpha = (1 - 2\|B\|_F)$-strongly monotone with where B is the parameter matrix whose columns are b_i.

Our data depicts the demand of electricity across an hour-long period for 6 ports of varying power profiles for each day in year. We standardize the data to be zero mean and unit variance across each station. Solutions are calculated by performing expected gradient play with constant step size; the expected mean is estimated via the empirical mean over the data set.

We set $b_{ii} = -1/18 + \nu$ and $b_{ij} = 1/18 + \nu$, where we use $\nu \sim \mathcal{N}(0, 10^{-5})$ to simulate learning B from samples. Hence demand for agent i decreases as their own price increases, and increases as the price of other agents decreases. We run the stochastic gradient play algorithm initialized at $x^0 = p_r \mathbb{1}_n$ with a single sample at each round and a decaying step size policy $\omega_t = \alpha(r + t - 2)/2$ for $r = 3$. In Figure 4.1b we plot the mean error trajectory an confidence interval over 50 trials of 2000 iterations.

Bibliography

[1] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Informational braess' paradox: The effect of information on traffic congestion. Operations Research, 66(4):893–917, 2018.

[2] Joydeep Acharya and Roy D. Yates. Service provider competition and pricing for dynamic spectrum allocation. In 2009 International Conference on Game Theory for Networks, pages 190–198, 2009.

[3] Charalambos D Aliprantis and Kim Border. Infinite dimensional analysis: A Hitchhiker's Guide. Springer, 2006.

[4] Babak Ghaffarzadeh Bakhshayesh and Hamed Kebriaei. Decentralized equilibrium seeking of joint routing and destination planning of electric vehicles: A constrained aggregative game approach. IEEE Transactions on Intelligent Transportation Systems, 23(8):13265–13274, 2021.

[5] Nicola Bastianello, Liam Madden, Ruggero Carli, and Emiliano Dall'Anese. A stochastic operator framework for inexact static and online optimization. arXiv preprint arXiv:2105.09884, 2021.

[6] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. Operations research, 63(5):1227–1244, 2015.

[7] Gianluca Bianchin, Miguel Vaquero, Jorge Cortes, and Emiliano Dall'Anese. Online stochastic optimization for unknown linear systems: Data-driven synthesis and controller analysis. arXiv preprint arXiv:2108.13040, 2021.

[8] Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial pricing in ride-sharing networks. Operations Research, 67(3):744–769, 2019.

[9] Vladimir I Bogachev and Aleksandr V Kolesnikov. The monge-kantorovich problem: achievements, connections, and perspectives. Russian Mathematical Surveys, 67(5):785, 2012.

[10] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. SIAM review, 60(2):223–311, 2018.

[11] Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. Advances in Neural Information Processing Systems, 31, 2018.

[12] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. The Journal of Machine Learning Research, 13(1):2617–2654, 2012.

[13] Xuanyu Cao, Junshan Zhang, and H. Vincent Poor. Online stochastic optimization with time-varying distributions. IEEE Tran. on Automatic Control, 66(4):1840–1847, 2021.

[14] Boxiao Chen, Xiuli Chao, and Cong Shi. Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. Mathematics of Operations Research, 46(2):726–756, 2021.

[15] Wang Chi Cheung, David Simchi-Levi, and He Wang. Dynamic pricing and demand learning with limited price experimentation. Operations Research, 65(6):1722–1731, 2017.

[16] Saurab Chhachhi and Fei Teng. On the 1-wasserstein distance between location-scale distributions and the effect of differential privacy. arXiv preprint arXiv:2304.14869, 2023.

[17] Nicolò Daina, Aruna Sivakumar, and John W Polak. Electric vehicle charging choices: Modelling and implications for smart charging services. Transportation Research Part C: Emerging Technologies, 81:36–56, 2017.

[18] Emiliano Dall'Anese, Andrea Simonetto, Stephen Becker, and Liam Madden. Optimization and learning with information streams: Time-varying algorithms and applications. IEEE Signal Processing Magazine, 37(3):71–83, 2020.

[19] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 99–108, 2004.

[20] Arnoud V Den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. Surveys in operations research and management science, 20(1):1–18, 2015.

[21] Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Improved rates for derivative free gradient play in strongly monotone games. In 2022 IEEE 61st Conference on Decision and Control (CDC), pages 3403–3408. IEEE, 2022.

[22] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. arXiv:2011.11173, 2020.

[23] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. Mathematics of Operations Research, 2022.

[24] Filippo Fabiani and Barbara Franci. A stochastic generalized nash equilibrium model for platforms competition in the ride-hail market. In 2022 IEEE 61st Conference on Decision and Control (CDC), pages 4455–4460. IEEE, 2022.

[25] Francisco Facchinei and Jong-Shi Pang. Finite-dimensional variational inequalities and complementarity problems. Springer, 2003.

[26] Francisco Facchinei and Jong-Shi Pang. Finite-dimensional variational inequalities and complementarity problems. Springer Science & Business Media, 2007.

[27] Filiberto Fele and Kostas Margellos. Scenario-based robust scheduling for electric vehicle charging games. In 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), pages 1–6. IEEE, 2019.

[28] Madeline Gilleran, Eric Bonnema, Jason Woods, Partha Mishra, Ian Doebber, Chad Hunter, Matt Mitchell, and Margaret Mann. Impact of electric vehicle charging on the power demand of retail buildings. Advances in Applied Energy, 4:100062, 2021.

[29] Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. Michigan Mathematical Journal, 31(2):231 – 240, 1984.

[30] Matthew T Hale, Angelia Nedić, and Magnus Egerstedt. Asynchronous multiagent primal-dual optimization. IEEE Transactions on Automatic Control, 62(9):4421–4435, 2017.

[31] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In Proceedings of the 2016 ACM conference on innovations in theoretical computer science, pages 111–122, 2016.

[32] Eryn Juan He, Sergei Savin, Joel Goh, and Chung-Piaw Teo. Off-platform threats in on-demand services. Manufacturing & Service Operations Management, 25(2):775–791, 2023.

[33] Liam Hodgkinson and Michael W Mahoney. Multiplicative noise and heavy tails in stochastic optimization. arXiv:2006.06293, 2020.

[34] Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In Artificial Intelligence and Statistics, pages 398–406. PMLR, 2015.

[35] 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.

[36] Jayash Koshal, Angelia Nedić, and Uday V Shanbhag. Multiuser optimization: Distributed algorithms and error analysis. SIAM Journal on Optimization, 21(3):1046–1081, 2011.

[37] Charilaos Latinopoulos, Aruna Sivakumar, and JW Polak. Response of electric vehicle drivers to dynamic pricing of parking and charging services: Risky choice in early reservations. Transportation Research Part C: Emerging Technologies, 80:175–189, 2017.

[38] Chaojie Li, Zhaoyang Dong, Guo Chen, Bo Zhou, Jingqi Zhang, and Xinghuo Yu. Data-driven planning of electric vehicle charging infrastructure: a case study of sydney, australia. IEEE Transactions on Smart Grid, 12(4):3289–3304, 2021.

[39] Licong Lin and Tijana Zrnic. Plug-in performative optimization. arXiv preprint arXiv:2305.18728v1, 2023.

[40] Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, S Shankar Sastry, and Lillian J Ratliff. Zeroth-order methods for convex-concave minmax problems: Applications to decision-dependent risk minimization. arXiv preprint arXiv:2106.09082, 2021.

[41] Johanna L Mathieu, Duncan S Callaway, and Sila Kiliccote. Examining uncertainty in demand response baseline models and variability in automated responses to dynamic pricing. In 2011 50th IEEE Conference on Decision and Control and European Control Conference, pages 4332–4339. IEEE, 2011.

[42] Johanna L Mathieu, Marina González Vayá, and Göran Andersson. Uncertainty in the flexibility of aggregations of demand response resources. In IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society, pages 8052–8057. IEEE, 2013.

[43] Chathurika P Mediwaththe and David B Smith. Game-theoretic electric vehicle charging management resilient to non-ideal user behavior. IEEE Transactions on Intelligent Transportation Systems, 19(11):3486–3495, 2018.

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

[45] 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.

[46] Wenlong Mou, Nhat Ho, Martin J Wainwright, Peter Bartlett, and Michael I Jordan. A diffusion process perspective on posterior contraction rates for parameters. arXiv preprint arXiv:1909.00966, 2019.

[47] 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.

[48] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. SIAM Journal on optimization, 19(4):1574–1609, 2009.

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

[50] Herbert Robbins and Sutton Monro. A stochastic approximation method. The annals of mathematical statistics, pages 400–407, 1951.

[51] R Tyrrell Rockafellar and Roger J-B Wets. Variational analysis, volume 317. Springer Science & Business Media, 2009.

[52] Moshe Shaked and J George Shanthikumar. Stochastic orders. Springer Science & Business Media, 2007.

[53] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms, 2014.

[54] Iman Shames and Farhad Farokhi. Online stochastic convex optimization: Wasserstein distance variation. arXiv:2006.01397, 2020.

[55] Zhengwei Sun, Andrea C. Hupman, Heather I. Ritchey, and Ali E. Abbas. Bayesian updating of the price elasticity of uncertain demand. IEEE Systems Journal, 10(1):136–146, 2016.

[56] Joshua A Taylor and Johanna L Mathieu. Uncertainty in demand response—identification, estimation, and learning. In The Operations research revolution, pages 56–70. Informs, 2015.

[57] Berkay Turan and Mahnoosh Alizadeh. Competition in electric autonomous mobility on demand systems. IEEE Transactions on Control of Network Systems, 2021.

[58] Wayes Tushar, Walid Saad, H Vincent Poor, and David B Smith. Economics of electric vehicle charging: A game theoretic approach. IEEE Transactions on Smart Grid, 3(4):1767–1778, 2012.

[59] Roman Vershynin. High-dimensional probability: An introduction with applications in data science. Cambridge University press, 2018.

[60] Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

[61] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. Stat, 9(1):e318, 2020.

[62] Martin J. Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[63] Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. The Annals of Statistics, 48(2):1124 – 1142, 2020.

[64] 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, 2021.

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

[66] Yuanguang Zhong, Tong Yang, Bin Cao, and TCE Cheng. On-demand ride-hailing platforms in competition with the taxi industry: Pricing strategies and government supervision. International Journal of Production Economics, 243:108301, 2022.