STOCHASTIC SADDLE POINT PROBLEMS WITH DECISION-DEPENDENT DISTRIBUTIONS[*]

KILLIAN WOOD[†] AND EMILIANO DALL'ANESE[‡]

Abstract. This paper focuses on stochastic saddle point problems with decision-dependent distributions. These are problems whose objective is the expected value of a stochastic payoff function and whose data distribution drifts in response to decision variables—a phenomenon represented by a distributional map. A common approach to accommodating distributional shift is to retrain optimal decisions once a new distribution is revealed, or repeated retraining. We introduce the notion of equilibrium points, which are the fixed points of this repeated retraining procedure, and provide sufficient conditions for their existence and uniqueness. To find equilibrium points, we develop deterministic and stochastic primal-dual algorithms and demonstrate their convergence with constant step size in the former and polynomial decay step-size schedule in the latter. By modeling errors emerging from a stochastic gradient estimator as sub-Weibull random variables, we provide error bounds in expectation and in high probability that hold for each iteration. Without additional knowledge of the distributional map, computing saddle points is intractable. Thus we propose a condition on the distributional map—which we call opposing mixture dominance—that ensures that the objective is strongly-convex–strongly-concave. Finally, we demonstrate that derivative-free algorithms with a single function evaluation are capable of approximating saddle points.

Key words. saddle point problems, minimax problems, stochastic gradient, decision-dependent distributions

MSC codes. 90C25, 90C15, 90C47

DOI. 10.1137/22M1488077

1. Introduction. The broad goal of stochastic optimization is to find an optimal decision for an objective with uncertainty in some parameters [30, 36, 42]. For example, in statistical learning, parameters may be taken to be data-label pairs in large data sets [8, 17]; in the context of optimization of physical and dynamical systems, they may model externalities and random exogenous inputs, or system parameters that are predicted from data and are accompanied by given error statistics [4]. A key assumption that is typically leveraged for providing theoretical guarantees for stochastic optimization algorithms is that the distributions of random parameters are stationary [5]. However, in modern machine learning and cyber-physical systems applications, data may be subject to decision-dependent shift, whereby the distribution is inextricably tied to the decision variables.

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

$$(1.1) \qquad \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left\{ \Phi(x,y) := \mathop{\mathbb{E}}_{w \sim D(x,y)} [\phi(x,y,w)] \right\},$$

[†]Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO USA (killian.wood@colorado.edu).
[‡]Department of Electrical, Computer, and Energy Engineering, and Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO USA (emiliano.dallanese@colorado.edu).

where $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^n$ are compact constraint sets, $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a scalar-valued function of the decision variables (x, y) parameterized by a random vector w, D is a distribution inducing map, and w is supported on a complete and separable metric space M with metric d. Hereafter, we refer to Φ as the objective and the function ϕ as the minimax function. We remark that the distribution of w depends on the decision variables (x, y). When solutions to the problem (1.1) exist, we will denote these solutions as (x^*, y^*). Additional notational conventions may be found below.[1]

For general distributional maps D, solving (1.1) directly is intractable. Indeed, Φ may be nonconvex-nonconcave even when ϕ is strongly-convex–strongly-concave. A common heuristic when dealing with nonstationary 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

$$
\begin{aligned}
x_{t+1} &\in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathbb{E}_{w \sim D(x_t, y_t)}[\phi(x, y, w)], \\
y_{t+1} &\in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \mathbb{E}_{w \sim D(x_t, y_t)}[\phi(x, y, w)].
\end{aligned}
$$
(1.2)

We will refer to fixed points of this sequence as *equilibrium points*. These can be seen as the counterparts of the so-called performatively stable points in [11, 31, 41] in our stochastic minimax setup (1.1). A primary objective of this work is illustrating sufficient conditions for the existence and uniqueness of equilibrium points. In particular, existence of the set equilibrium 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 ϕ, we then develop deterministic and stochastic projected primal-dual algorithms that can determine equilibrium points.

However, as discussed in the paper, equilibrium points and saddle points are qualitatively distinct. Equilibrium 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 Φ. 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 Φ is strongly-convex–strongly-concave.

---

[1]*Notation.* We let $\mathbb{R}$ denote the set of real numbers, and $[n] = \{1, 2, \ldots, n\}$. For given column vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we let (x, y) denote their concatenation; that is, $[x^T, y^T]^T$ with ^{T} denoting transposition; for $x, y \in \mathbb{R}^n$, we let $\langle x, y \rangle$ denote the inner product. For a given column vector $x \in \mathbb{R}^n$, $\|x\|$ is the Euclidean norm; for a matrix $X \in \mathbb{R}^{n \times m}$, $\|X\|_F$ denotes the Frobenius norm and $\|X\|_\star$ the nuclear norm. Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $\nabla f(x)$ denotes the gradient of f at x (taken to be a column vector). For a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $\nabla_x f(x, x)$ denotes the partial derivatives of f with respect to x. Given a closed convex set $C \subseteq \mathbb{R}^n$, $\Pi_C : \mathbb{R}^n \to \mathbb{R}^n$ denotes the Euclidean projection of y onto C, namely, $\Pi_C(y) := \arg\min_{v \in C} \|y - v\|$. Given a set $C \subseteq \mathbb{R}^n$, $P(C)$ denotes its power set. For a given random variable $w \in \mathbb{R}$, we write $w \sim \mu$ to mean that w is a random variable with law μ, a probability measure supported on $\mathbb{R}$. Hence $\mu(A) = \mathbb{P}(w \in A)$ for all $A \subseteq \mathbb{R}$. Furthermore, $\mathbb{E}[w]$ denotes the expected value of w, and $\mathbb{P}(w \leq \epsilon)$ denotes the probability of w taking values smaller than or equal to ϵ; $\|w\|_p := \mathbb{E}[|w|^p]^{1/p}$ for any $p \geq 1$.

1.1. Motivation. Saddle point problems arise in a variety of areas, including stochastic constrained optimization problems [22, 33], strategic classification [31], and games [13], and are relevant in several applications that span finance, energy systems, transportation networks, and ride sharing, just to mention a few. Our work naturally contributes modeling and algorithmic approaches in these areas and applications whenever the uncertainty can be considered to be decision dependent. For example, in the energy systems context, the problem of finding an optimal charging policy for a fleet of electric vehicles involves uncertainty in the price of energy; indeed, real-time prices are subject to change due to the demand itself as well as on external factors such as spot market behavior [14]. As representative examples, in the following we provide a brief description of the problem formulations in competitive markets and strategic classification.

1.1.1. 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 into n zones. This applies to, for example, ride sharing [4] and power providers [1]. Focusing on electric vehicle charging station providers [23], at each zone $i \in [n]$ 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 $a_i(x_i + p_i)$, based on their demand a_i. However, they must incorporate a zone based utility cost $\theta_i(x_i + p_i)$, 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 n zones is given by $u_1(x,a) = \langle a + \theta, x + p \rangle - \|\Gamma_1 x\|^2$ with $\Gamma_1 = \text{diag}\{\gamma_{1,1}, \ldots, \gamma_{1,n}\}$. If the price and demand of service for provider two are given by y and b, respectively, then their profit is similarly represented as $u_2(y,b) = \langle b + \theta, y + p \rangle - \|\Gamma_2 y\|^2$. 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 $\mathcal{X} = \bigtimes_{i=1}^{n}[-p_i, c_{1,i}p_i]$ and $\mathcal{Y} = \bigtimes_{i=1}^{n}[-p_i, c_{2,i}p_i]$ 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 $a \overset{d}{=} a_0 + A_1 x + A_2 y$, $b \overset{d}{=} b_0 + B_1 x + B_2 y$, where $a_0 \sim D_1$ and $b_0 \sim D_2$, for which D_1 and D_2 represent stationary prior distributions for the demand associated with providers one and two, respectively [31]. In order to maximize their expected profits relative to provider two, provider one will minimize the negative of their relative profit given by $u_1(x,a) - u_2(y,b)$ and, hence, the optimal strategies for both providers are solutions to the saddle point problem

$$(1.3) \qquad \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathbb{E}_{(a,b) \sim D(x,y)} \|\Gamma_1 x\|^2 - \|\Gamma_2 y\|^2 - \langle a + \theta, x \rangle + \langle b + \theta, y \rangle,$$

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

1.1.2. Multitask strategic classification. A second application is a multitask strategic classification problem [31]. Consider the problem of learning a collection of N tasks for a strategic population simultaneously. Such a framework is especially helpful in problems such as spam filtering; in this case, a classifier is learned for each user, however, it is expected that spam email designers will adapt to spam filters in

order to be misclassified [24, 37, 39]. Each user has a dataset $\mathcal{D}_i = \{(a_{i,j}, b_{i,j})\}_{j=1}^{n_i}$ with features from their emails and labels of either spam or not spam. The goal is to learn classifiers $\{f_i\}_{i=1}^N$ that will predict spam based not only on the available data, but on the anticipated changes spammers will make to their features. To learn $\{f_i\}_{i=1}^N$, we pose the nuclear norm regularized minimization problem

$$(1.4) \qquad \min_{x_1,\ldots,x_N \in \mathcal{X}} \quad \sum_{i=1}^N \mathop{\mathbb{E}}_{(a,b) \sim D_i(x_i)} \ell(b, f_i(a; x_i)) + \|[x_1, \ldots, x_N]\|_*.$$

A challenge in solving (1.4) is that the nuclear norm is not differentiable. To avoid using proximal methods, we can introduce dual variables by observing that the nuclear norm is the dual of the operator norm and hence $\|X\|_\star = \max_{Y^T Y \preceq I} -\langle X, Y \rangle_F$, where $\langle \cdot, \cdot \rangle_F : \mathbb{R}^{d \times m} \times \mathbb{R}^{d \times m} \mapsto \mathbb{R}$ is the Frobenius inner product defined by $\langle X, Y \rangle_F = \text{trace}(X^T Y)$ [32]. Thus, the problem in (1.4) can be rewritten as

$$(1.5) \qquad \min_{X \in \mathcal{X}^T} \max_{Y \in \mathcal{Y}} \quad \sum_{t=1}^T \mathop{\mathbb{E}}_{(a,b) \sim D_t(x_t)} \ell(b, f_t(a; x_t)) - \langle X, Y \rangle_F,$$

where $\mathcal{X}^N = \bigtimes_{i=1}^N \mathcal{X} \subseteq \mathbb{R}^{d \times N}$ and $\mathcal{Y} = \{Y | Y^T Y \preceq I\}$.

1.2. Related works. In this subsection, we review the literature on saddle point problems and stochastic optimization that is most relevant to our work.

Saddle point problems. Saddle point problems have been well investigated, with theoretical guarantees in distance metrics (squared Euclidean metric) and dual-gap metrics. Common approaches for these problems are either proximal methods or primal-dual methods, with version existing for both deterministic and stochastic problems. Proximal methods include mirror-prox [29] and extragradient [27] for deterministic objectives and stochastic mirror-prox [30] for stochastic objectives; primal-dual methods include primal-dual or gradient descent-ascent [20] and optimistic gradient descent-ascent [27]. Stochastic primal-dual methods and their accelerated varieties are studied in [43]. Convergence guarantees for stochastic optimization algorithms typically come in the form of convergence with a diminishing step-size policy [28, 29, 30], or convergence to a neighborhood with fixed step size [20, 43].

Decision-dependent distributions. Our work is most closely related to the literature on stochastic optimization with decision-dependent distributions—also referred to as performative prediction in the machine learning community. These are two related paradigms for stochastic programs where the data distribution is a function of the decision variables. In this context, equilibrium problems are presented and solved via conceptual deterministic algorithms in [31]; second moment analysis for stochastic algorithms with access to a sampling oracle is provided in [25]; and proximal first-order algorithms for regularized objectives are studied in [11]. In [41], first moment and high probability tracking analysis are provided under a sub-Weibull gradient error for a time-varying optimization problem.

Sub-Weibull error models. Sub-Gaussian and subexponential gradient error models are common in the literature on stochastic gradient methods. Empirical and theoretical results have demonstrated that heavier tailed distributions arise naturally in deep learning. The class of sub-Weibull distributions subsumes the sub-Guassian and subexponential classes of distributions, while also including heavier tailed distributions as well as distributions with bounded support [21, 38, 40]. Given their broad application, this model has been receiving increased use in the literature. See, for example, [3, 21, 41].

1.3. Contributions. In this paper, we offer the following main contributions.
(1) *The minimax equilibrium problem.* We propose a notion of equilibrium points
for the saddle point problem in (1.1). We then provide conditions to guarantee their
existence and uniqueness, and we provide bounds for the distance between the unique
equilibrium point and saddle points of (1.1).
(2) *Algorithms.* We demonstrate that primal-dual algorithms, using the gradients
of ϕ, are effective at finding equilibrium points when the stochastic objective ϕ is
strongly-convex–strongly-concave. First, we demonstrate convergence of a concep-
tual algorithm using full gradient information. We then demonstrate that stochastic
algorithms with fixed step size converge to a noise-dominated neighborhood of the
equilibrium point, as well as provide expectation bounds and high probability bounds
that hold for each iteration. Additionally, we show convergence of the stochastic
algorithm for decaying step size.
(3) *Saddle points.* We propose a sufficient condition for distributional maps that pre-
serve the strong-convexity–strong-concavity of ϕ. Hence when ϕ is strongly-convex–
strongly-concave, so too is Φ. We then discuss a zeroth-order algorithm capable of
finding an approximate saddle point using only a single function evaluation.
(4) *Experiments.* We illustrate our results on the electric vehicle charging problem
in (1.3) by incorporating synthetic demand data from [16] for a location-scale family
based distributional map.

Connection to related works. Relative to the referenced work on stochastic saddle
point problems, we consider the case where the function $\phi(x, y, w)$ is strongly-convex–
strongly-concave in the decisions (x, y) and the randomness captured by w is governed
by a family of distributions $(x, y) \mapsto D(x, y)$. To measure performance of our algo-
rithms we use the Euclidean distance to the solution. Contrary to most of the work
on the stochastic saddle point problems that use the distance metric, we analyze using
both the first moment and the second moment. We note that, however, bounds on
the first moment do not need stochastic filtrations as an underlying working assump-
tion. We also provide bounds in high probability that hold for each iteration, under
a sub-Weibull model.

The prior body of work on optimization with decision dependent-distribution has
studied both equilibrium points and minimizers, as well as stochastic algorithms for
finding them. Hence, our results extend those found in the setting of [11, 25, 26, 31]
by casting the problem into a more general variational inequality framework. The
study of saddle point problems is a unique contribution of this paper. Relative to [26],
we continue the study of stochastic orders and location-scale families by proposing a
condition on the distributional map that is suitable for convex-concave objectives. To
approximate saddle points, we analyze a zeroth-order algorithm adapted from works
such as [7, 10, 15, 26] to suit our setting.

Organization. The paper is organized as follows. In section 2 we define and solve
the equilibrium point problem in the static setting. Section 3 investigates the time-
varying equilibrium point problem via online optimization methods. Section 4 offers
sufficient conditions for which the saddle point problem can be solved. In section 5,
we illustrate our results on a bilinear matrix game.

2. The equilibrium problem. Recall the decision-dependent stochastic saddle
point problem provided in (1.1),

$$(2.1) \qquad \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left\{ \Phi(x, y) := \mathop{\mathbb{E}}_{w \sim D(x,y)} [\phi(x, y, w)] \right\},$$

where the sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ are convex and compact. Let $\mathcal{P}(M)$ be the set of Radon probability measures on a complete and separable metric space M with finite first moment, and observe that the objective function can be written in integral form as $\Phi(x,y) = \int_M \phi(x,y,w)\mu_{(x,y)}(dw)$, where $\mu_{(x,y)} \in \mathcal{P}(M)$ 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 2.1 (saddle points). *A pair $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a saddle point for the problem in* (1.1) *provided that $\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*) \; \forall x \in \mathcal{X}, y \in \mathcal{Y}$.*

Sufficient conditions for the existence of saddle points consist of Φ being convex-concave while $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact [34, Example 11.52]. When minimax equality holds, we can equivalently characterize saddle points as a pair that satisfies

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

In practice, computing saddle points directly is computationally intractable. Namely, the dependence of the distributional map on the decision variables implies that even when ϕ is convex-concave Φ 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 (1.2).

DEFINITION 2.2 (equilibrium points). *A pair $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ is an equilibrium point if*

$$(2.2) \qquad \begin{aligned} \bar{x} &\in \arg\min_{x \in \mathcal{X}} \left\{ \max_{y \in \mathcal{Y}} \; \mathbb{E}_{w \sim D(\bar{x}, \bar{y})} [\phi(x,y,w)] \right\}, \\ \bar{y} &\in \arg\max_{y \in \mathcal{Y}} \left\{ \min_{x \in \mathcal{X}} \; \mathbb{E}_{w \sim D(\bar{x}, \bar{y})} [\phi(x,y,w)] \right\}. \end{aligned}$$

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 [20].

Our first objective in this work will be to provide conditions for the existence and uniqueness of these equilibrium points. Later, we develop first-order algorithms and demonstrate their convergence to equilibrium points. Crucial to our analysis will be the properties of the "decoupled objective," which is defined as $\Phi(x,y;x',y') = \mathbb{E}_{w \sim D(x',y')}[\phi(x,y,w)]$ for $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. Here, the distribution is fixed for given points (x', y'). With these definitions, we consider a correspondence $H : \mathcal{X} \times \mathcal{Y} \to P(\mathcal{X} \times \mathcal{Y})$, defined by

$$(2.3) \qquad H(x,y) = \left(\arg\min_{x' \in \mathcal{X}} \max_{y' \in \mathcal{Y}} \Phi(x', y'; x, y), \; \arg\max_{y' \in \mathcal{Y}} \min_{x' \in \mathcal{X}} \Phi(x', y'; x, y) \right)$$

which maps pairs in the product space to its power set $P(\mathcal{X} \times \mathcal{Y})$. In light of Definition 2.2, the equilibrium points are fixed points of the map H; that is, $(\bar{x}, \bar{y}) \in H(\bar{x}, \bar{y})$.

For notional convenience, we will introduce the stacked vector in the product space $z = (x,y) \in \mathcal{X} \times \mathcal{Y}$ (consequently, we can identify $H(z)$ and $\Phi(z';z)$ with the above functions whenever convenient). In the following section, we provide sufficient conditions for the existence of equilibrium points.

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

DEFINITION 2.3 (neighborhood [2, sect. 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 2.4 (upper hemicontinuity [2, sect. 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 equilibrium points.

THEOREM 2.5 (existence of equilibrium points). *Suppose that the following assumptions hold:*
  (i) *$x \mapsto \phi(x, y, w)$ is convex in x for all $y \in \mathcal{Y}$ and for all realizations of w;*
  (ii) *$y \mapsto \phi(x, y, w)$ is concave in y for all $x \in \mathcal{X}$ and for all realizations of w;*
  (iii) *ϕ is continuous on $\mathcal{X} \times \mathcal{Y}$ for all w;*
  (iv) *$\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} \subset \mathbb{R}^n$ are convex compact subsets;*
  (iv) *the distributional map $D : \mathcal{Z} \to (\mathcal{P}(M), W_1)$ is continuous.*
Then the fixed point set $\mathrm{Fix}(H)$ is nonempty and compact.

Proof. The proof amounts to showing that H satisfies the hypotheses of Kakutani's fixed point theorem [2, 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 nonempty convex and compact set values in $P(\mathcal{X} \times \mathcal{Y})$. Following the closed graph theorem [2, Theorem 17.11], compactness of $\mathcal{X} \times \mathcal{Y}$ implies that H has a closed graph if and only if it is closed valued and upper hemicontinuous. Hence our proof reduces to showing that (i) H has nonempty closed values, (ii) H is upper hemicontinuous, and (iii) H has convex values.

Define the intermediate functions

$$(2.4) \qquad f(x'; z) = \max_{y' \in \mathcal{Y}} \Phi(x', y'; z) \quad \text{and} \quad g(y'; z) = \min_{x \in \mathcal{X}} \Phi(x', y'; z)$$

as well as the realization functions

$$(2.5) \qquad F(z) = \arg\min_{x' \in \mathcal{X}} f(x'; z) \quad \text{and} \quad G(z) = \arg\max_{y' \in \mathcal{Y}} g(y'; z)$$

for all $x' \in \mathcal{X}$, $y' \in \mathcal{Y}$, and $z \in \mathcal{X} \times \mathcal{Y}$. Using this convention, H can be written compactly as $H(z) = (F(z), G(z))$. It follows from continuity of ϕ and D on $\mathcal{X} \times \mathcal{Y}$, as well as compactness of $\mathcal{X}$ and $\mathcal{Y}$, that f and g are continuous [2, 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 closedness. 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 $z \in \mathcal{X} \times \mathcal{Y}$ 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 z such that $H(W) \subset U$. Given that $H(z)$ is a compact subset of U, [2, 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 $F^u(V_x) = \{z : F(z) \subset V_x\}$ 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 $z \in W$ 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 $x' \mapsto f(x'; z)$ is convex for all z and $\mathcal{X}$ is convex, then $F(z)$ is convex for all $z \in \mathcal{X} \times \mathcal{Y}$. Similarly, $G(z)$ is convex for all z. Since the Cartesian product of convex sets is convex, then condition (iii) follows. $\square$

Recall that the intuition for the equilibrium 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 2.6 (saddle point and equilibrium equivalence). *Suppose that an equilibrium point exists. Then* $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ *is an equilibrium point if and only if*

$$(2.6) \qquad \Phi(\bar{x}, y; \bar{x}, \bar{y}) \leq \Phi(\bar{x}, \bar{y}; \bar{x}, \bar{y}) \leq \Phi(x, \bar{y}; \bar{x}, \bar{y})$$

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

We omit the proof as it 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.

2.2. Equilibrium 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 equilibrium points.

Assumption 1 (strong-convexity–strong-concavity). The function ϕ is is continuously differentiable over for any realization of w. The function ϕ is γ-strongly-convex–strongly-concave, for any realization of w; that is, ϕ is γ-strongly-convex in x for all $y \in \mathbb{R}^n$ and γ-strongly-concave in y for all $x \in \mathbb{R}^d$.

Assumption 2 (joint smoothness). The stochastic gradient map ψ given by $\psi(z, w) := (\nabla_x \phi(z, w), -\nabla_y \phi(z, w))$ is L-Lipschitz in z and w. Namely,

$$\|\psi(z, w) - \psi(z', w)\| \leq L\|z - z'\|, \quad \|\psi(z, w) - \psi(z, w')\| \leq L \mathrm{d}(w, w')$$

for any $z, z' \in \mathbb{R}^n \times \mathbb{R}^m$ and w, w' supported on M. Here $\mathrm{d} : M \times M \to \mathbb{R}$ denotes the metric on M.

Assumption 3 (distributional sensitivity). The distributional map $D : \mathbb{R}^d \times \mathbb{R}^n \to \mathcal{P}(M)$ is ε-Lipschitz. Namely,

$$W_1(D(z), D(z')) \leq \varepsilon\|z - z'\|$$

for any $z, z' \in \mathbb{R}^n \times \mathbb{R}^m$, where W_1 is the Wasserstein-1 distance.

Assumption 4 (compact convex sets). The sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ are compact and convex.

Typically, the assumption of strong-convexity–strong-concavity enables linear convergence to saddle points in standard primal-dual methods [20]. Furthermore, strong-convexity–strong-concavity implies uniqueness of saddle point solutions; this allows one to derive convergence results to the unique saddle point in the static case, and tracking results in the context of time-varying minmax problems [9]. We also note that this assumption is useful in this paper in order to characterize the intrinsic relationship between optimal solutions to (1.1) and equilibrium points. We also note that, for simplicity, the assumption imposes a common geometry parameter in ϕ for both the x and y values; however, our analysis is the same for functions ϕ 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 ε-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 [11, 31, 41]. 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 [6, 19], this metric can be written as

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

for all $\mu, \nu \in \mathcal{P}(M)$. Here the supremum is taken over all Lipschitz-continuous functionals on M with a 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 [18, 20]. 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 $\mathcal{X} = [-r, r]^n$ and $\mathcal{Y} = [-r, r]^m$ 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., [20] for an example in the context of constrained optimization problems.

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

Recall that in Assumption 2, we introduce the stochastic gradient map $\psi(z, w) = (\nabla_x \phi(z, w), -\nabla_y \phi(z, w))$. Using this convention, we denote the decoupled gradient map as

$$\Psi(z; z') = \mathbb{E}_{w \sim D(z')}[\psi(z, w)].$$

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

THEOREM 2.7 (equilibrium variational inequality). *A point $\bar{z} \in \mathcal{Z}$ is an equilibrium point provided that*

(2.7) $$\langle z - \bar{z}, \Psi(\bar{z}; \bar{z}) \rangle \geq 0 \quad \forall z \in \mathcal{Z}.$$

Proof of this fact follows steps that are similar to the ones in [34, Example 12.50]. In light of Definition 2.2, this result suggests that the $\bar{z}$ are solutions to variational

inequality induced by the stationary distribution $D(\bar{z})$. In the following, we show that when $z \mapsto \Psi(z; z')$ is strongly monotone for all $z' \in \mathbb{R}^n \times \mathbb{R}^m$, a unique equilibrium point exists. Furthermore, under this assumption, we can show that the distance between the saddle points for the original problem in (1.1) and the unique equilibrium point is bounded.

PROPOSITION 2.8 *Suppose that Assumption* 1 *holds. Then, for any* $w \in M$, $z \mapsto \psi(z, w)$ *is* γ-*strongly-monotone. Furthermore, for any* $z' \in \mathbb{R}^n \times \mathbb{R}^m$, $z \mapsto \Psi(z; z')$ *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 Ψ. This amounts to the decoupled gradient map being Lipschitz continuous in the distributional argument.

LEMMA 2.9 (gradient deviations). *Suppose that Assumptions* 2 *to* 4 *hold. Then, for any* $\hat{z} \in \mathbb{R}^n \times \mathbb{R}^m$, *the map* $z \mapsto \Psi(\hat{z}, z)$ *is* εL-*Lipschitz. Furthermore, the restriction to* $\mathcal{Z}$ *is bounded in the following way: for any* $\hat{z} \in \mathcal{Z}$

$$(2.8) \qquad \|\Psi(\hat{z}; z) - \Psi(\hat{z}; z')\| \leq \varepsilon L D_{\mathcal{Z}}$$

for all $z, z' \in \mathcal{Z}$, *where* $D_{\mathcal{Z}} = diam(\mathcal{Z}) < \infty$.

Proof. Let $v \in \mathbb{R}^n \times \mathbb{R}^m$ be an arbitrary unit vector and fix $\hat{z}, z, z' \in \mathbb{R}^n \times \mathbb{R}^m$. It follows that

$$\langle v, \Psi(\hat{z}, z) - \Psi(\hat{z}; z') \rangle = \mathop{\mathbb{E}}_{w \sim D(z)}[\langle v, \psi(\hat{z}, w) \rangle] - \mathop{\mathbb{E}}_{w \sim D(z')}[\langle v, \psi(\hat{z}, w) \rangle].$$

By our assumption, we have that $w \mapsto \langle v, \psi(\hat{z}, w) \rangle$ is Lipchitz with constant $L\|v\| = L$. Thus, from Kantorivich and Rubenstein, we have that

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

where that last inequality follows from ε-sensitivity of D. Thus we have that for any unit vector v, $\langle v, (\Psi(\hat{z}, z) - \Psi(\hat{z}; z') \rangle \leq \varepsilon L\|z - z'\|$. Hence, choosing $v = (\Psi(\hat{z}, z) - \Psi(\hat{z}; z'))/\|(\Psi(\hat{z}, z) - \Psi(\hat{z}; z'))\|$ yields the result. Last, by Assumption 4, $\mathcal{Z}$ is compact and hence $\|z - z'\| \leq D_{\mathcal{Z}} < \infty$ for any $z, z' \in \mathcal{Z}$. Thus, (2.8) follows. □

In what follows, we demonstrate existence and uniqueness of equilibrium points. Similarly 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 2.10 (existence and uniqueness of equilibrium points). *Suppose that Assumptions* 1 *to* 4 *hold. Then*
 1. *for all* $z, z' \in \mathcal{Z}$, $\|H(z) - H(z')\| \leq \frac{\varepsilon L}{\gamma}\|z - z'\|$;
 2. *if* $\frac{\varepsilon L}{\gamma} < 1$, *then there exists a unique equilibrium point* $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$.

Proof. Let $\hat{z}, \tilde{z} \in \mathcal{Z}$ be fixed. Then the maps $z \to \Psi(z; \hat{z})$ and $z \to \Psi(z; \tilde{z})$ are γ-strongly-monotone. Furthermore, our strong-convexity and strong-concavity assumptions on ϕ imply that $H(\hat{z})$ and $H(\tilde{z})$ are single valued in $\mathcal{X} \times \mathcal{Y}$. Recall from our definition of H that $H(\hat{z})$ and $H(\tilde{z})$ are solutions to the variational inequalities induced by $\hat{z}$ and $\tilde{z}$, respectively. That is, for all $z \in \mathcal{X} \times \mathcal{Y}$,

$$(2.9) \qquad \langle z - H(\hat{z}), \Psi(H(\hat{z}); \hat{z}) \rangle \geq 0 \quad \text{and} \quad \langle z - H(\tilde{z}), \Psi(H(\tilde{z}); \tilde{z}) \rangle \geq 0.$$

It follows from strong monotonicity that $\langle H(\hat{z}) - H(\tilde{z}), \Psi(H(\hat{z}); \hat{z}) - \Psi(H(\tilde{z}); \hat{z}) \rangle \geq \gamma \|H(\hat{z}) - H(\tilde{z})\|^2$ and (2.9) imply that $\langle H(\tilde{z}) - H(\hat{z}), \Psi(H(\hat{z}); \hat{z}) \geq 0$. Hence,

$$(2.10) \qquad \langle H(\hat{z}) - H(\tilde{z})), \Psi(H(\tilde{z}); \hat{z}) \leq -\gamma \|H(\hat{z}) - H(\tilde{z})\|^2.$$

To proceed, we provide a lower bound for the quantity on the left-hand side. By applying Cauchy–Schwarz and Lemma 2.9, we get that

$$\langle H(\hat{z}) - H(\tilde{z}), \Psi(H(\tilde{z}); \tilde{z}) - \Psi(H(\tilde{z}); \hat{z}) \rangle \leq \varepsilon L \|H(\hat{z}) - H(\tilde{z})\| \|\tilde{z} - \hat{z}\|.$$

Since (2.9) implies that $\langle H(\hat{z}) - H(\tilde{z}), \Psi(H(\tilde{z}); \tilde{z}) \geq 0$, then we get that

$$(2.11) \qquad \langle H(\hat{z}) - H(\tilde{z}), \Psi(H(\tilde{z}); \hat{z}) \geq -\varepsilon L \|H(\hat{z}) - H(\tilde{z})\| \|\tilde{z} - \hat{z}\|.$$

Combining inequalities (2.10) and (2.11) yields

$$-\gamma \|H(\hat{z}) - H(\tilde{z})\|^2 \geq -\varepsilon L \|H(\hat{z}) - H(\tilde{z})\| \|\tilde{z} - \hat{z}\|,$$

and simplifying yields the result.

Since H is Lipschitz continuous, it is a strict contraction if $\varepsilon L/\gamma < 1$. Uniqueness of the fixed point follows from the Banach–Picard fixed-point theorem. $\square$

We have demonstrated existence and uniqueness of equilibrium points for some classes of problems; next, we characterize the relationship between equilibrium points and solutions of the original problem in (1.1). First, an important observation is that when $\varepsilon = 0$, the problem statement in (2.1) has a stationary probability distribution with respect to the decisions. Hence, saddle points coincide with equilibrium points. When $\varepsilon > 0$, we provide a guarantee on the distance between solutions of the two problems.

PROPOSITION 2.11 (bounded distance). *Suppose that Assumptions* 1 *to* 4 *hold. Let* z^* *be the optimal solution of* (1.1)*, and let* $\bar{z}$ *be the equilibrium point. Then,*

$$(2.12) \qquad \|z^* - \bar{z}\| \leq \frac{\varepsilon L}{\gamma} D_{\mathcal{Z}}.$$

Proof. From the optimality conditions, we have that the decoupled gradient map satisfies $\langle \bar{z} - z^*, \Psi(z^*; z^*) \rangle \geq 0$ and $\langle z^* - \bar{z}, \Psi(\bar{z}; \bar{z}) \rangle \geq 0$. By combining these results with results with our gradient deviation bound in Lemma 2.9, we obtain the following,

$$\begin{aligned}
\langle \bar{z} - z^*, \Psi(\bar{z}; \bar{z}) - \Psi(z^*; \bar{z}) \rangle &= \langle \bar{z} - z^*, \Psi(\bar{z}; \bar{z}) \rangle - \langle \bar{z} - z^*, \Psi(z^*; \bar{z}) \rangle \\
&\leq \langle \bar{z} - z^*, \Psi(z^*; z^*) - \Psi(z^*; \bar{z}) \rangle \\
&\leq \|\bar{z} - z^*\| \, \|\Psi(z^*; z^*) - \Psi(z^*; \bar{z})\| \\
&\leq \varepsilon L D_{\mathcal{Z}} \|\bar{z} - z^*\|,
\end{aligned}$$

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

$$\gamma \|\bar{z} - z^*\|^2 \leq \langle \bar{z} - z^*, \Psi(\bar{z}; \bar{z}) - \Psi(z^*; \bar{z}) \rangle \leq \varepsilon L D_{\mathcal{Z}} \|\bar{z} - z^*\|$$

so that canceling terms and dividing by γ yields the result. $\square$

2.3. Finding the equilibrium point via a primal-dual algorithm. In this section, we focus on a primal-dual method for finding the equilibrium point, and we provide results in terms of convergence at a linear rate. In particular, we focus on the equilibrium primal-dual (EPD) algorithm, which is based on the algorithmic map,

$$(2.13) \qquad \mathcal{G}(z; z') := \Pi_{\mathcal{Z}} \left(z - \eta \Psi(z; z') \right)$$

for all $z, z' \in \mathbb{R}^n \times \mathbb{R}^m$, where we recall that $\Pi_{\mathcal{Z}}(z) = \arg\min_{z' \in \mathcal{Z}} \|z - z'\|^2$ is the projection map and $\eta > 0$ is a positive step size. Given an initial point $z_0 \in \mathbb{R}^n \times \mathbb{R}^m$, the algorithm then generates a sequence via the Banach–Picard iteration

$$(2.14) \qquad z_{t+1} = G(z_t; z_t) = \Pi_{\mathcal{Z}} \left(z_t - \eta \Psi(z_t; z_t) \right), \ \ t = 1, 2, \ldots,$$

where we recall that $\Psi(z_t; z_t) = \mathbb{E}_{w \sim D(z_t)}[\psi(z_t, w)]$ with $D(z_t)$ the distribution induced by z_t. A key feature of this method is that each step projects onto the constraint sets and, hence, $z_t \in \mathcal{Z}$ for all $t \geq 1$ for any initial condition z_0. In the following we demonstrate that the equilibrium points are fixed points of the algorithmic map. We then provide linear convergence results for the case when our assumptions hold on just the constraint sets $\mathcal{Z}$, and then later globally.

PROPOSITION 2.12 (fixed point characterization). *Let Assumptions 1 to 4 hold and suppose that $\frac{\varepsilon L}{\gamma} < 1$. A point $\bar{z} \in \mathcal{Z}$ is an equilibrium point if and only if $\bar{z} = \mathcal{G}(\bar{z}; \bar{z})$.*

Proof. We want to show that $\bar{z}$ solving the variational inequality in (2.7) is equivalent to $\bar{z} = \Pi_{\mathcal{Z}}(\bar{z} - \eta \Psi(\bar{z}; \bar{z}))$ for $\eta > 0$. From [12, Theorem 1.5.5], for any $\hat{z} \in \mathbb{R}^n \times \mathbb{R}^m$, $\Pi_{\mathcal{Z}}(\hat{z})$ is the unique element of $\mathcal{Z}$ such that

$$(2.15) \qquad \langle z - \Pi_{\mathcal{Z}}(\hat{z}), \Pi_{\mathcal{Z}}(\hat{z}) - \hat{z} \rangle \geq 0$$

holds for any $z \in \mathcal{Z}$. As for the forward direction, if $\bar{z}$ is an equilibrium point, then (2.7) is equivalent to

$$(2.16) \qquad \langle z - \bar{z}, \bar{z} - (\bar{z} - \eta \Psi(\bar{z}; \bar{z})) \rangle \geq 0.$$

In setting $\hat{z} = \bar{z} - \eta \Psi(\bar{z}; \bar{z})$ in (2.15), we get that $\bar{z} = \Pi_{\mathcal{Z}}(\bar{z} - \eta \Psi(\bar{z}; \bar{z}))$. Conversely, if $\bar{z}$ is such that $\bar{z} = \Pi_{\mathcal{Z}}(\bar{z} - \eta \Psi(\bar{z}; \bar{z}))$, then by substituting $\hat{z} = \bar{z} - \eta \Psi(\bar{z})$ into (2.16), we have that $\bar{z}$ satisfies (2.7). $\qquad\square$

Now that we can equivalently characterize equilibrium points of the EPD map, we can leverage fixed point theory to analyze the EPD algorithm. In the following, we demonstrate convergence to the unique equilibrium point.

THEOREM 2.13 (EPD convergence). *Suppose that Assumptions 1 to 4 hold and that $\frac{\varepsilon L}{\gamma} < 1$. Then the sequence $z_{t+1} = \mathcal{G}(z_t; z_t)$ satisfies the bound*

$$(2.17) \qquad \|z_t - \bar{z}\| \leq \alpha^t \|z_0 - \bar{z}\|$$

for any initial point $z_0 \in \mathcal{Z}$ and for $\alpha := \sqrt{1 - 2\eta\gamma + \eta^2 L^2} + \eta \varepsilon L$. Furthermore, if

$$(2.18) \qquad \eta \in \left(0, \frac{2(\gamma - \varepsilon L)}{L^2(1 - \varepsilon^2)} \right),$$

then z_t converges linearly to the unique equilibrium point $\bar{z}$.

Proof. By applying our fixed point result in Proposition 2.12 and using the triangle inequality, we get that

(2.19)
$$\|z_{t+1} - \bar{z}\| = \|\mathcal{G}(z_t; z_t) - \mathcal{G}(\bar{z}; \bar{z})\| \leq \|\mathcal{G}(z_t; z_t) - \mathcal{G}(z_t; \bar{z})\| + \|\mathcal{G}(z_t; \bar{z}) - \mathcal{G}(\bar{z}; \bar{z})\|.$$

Bounding the first quantity amounts to applying our gradient deviation result in Lemma 2.9. Hence, $\|\mathcal{G}(z_t; z_t) - \mathcal{G}(z_t; \bar{z})\| \leq \eta \|\Psi(z_t; z_t) - \Psi(z_t; \bar{z})\| \leq \eta \varepsilon L \|z_t - \bar{z}\|$.

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

$$\begin{aligned}
\|\mathcal{G}(z_t; \bar{z}) - \mathcal{G}(\bar{z}'; \bar{z})\|^2 &\leq \|(z_t - \bar{z}) - \eta(\Psi(z_t; \bar{z}) - \Psi(\bar{z}; \bar{z}))\|^2 \\
&= \|z_t - \bar{z}\|^2 + \eta^2 \|(\Psi(z_t; \bar{z}) - \Psi(\bar{z}; \bar{z}))\|^2 \\
&\quad - 2\eta \langle z_t - \bar{z}, \Psi(z_t; \bar{z}) - \Psi(\bar{z}; \bar{z}) \rangle \\
&\leq (1 - 2\eta\gamma + \eta^2 L^2) \|z_t - \bar{z}\|^2.
\end{aligned}$$

Hence, adding yields $\|z_t - \bar{z}\| \leq (1 - 2\eta\gamma + \eta^2 L^2)\|z_{t-1} - \bar{z}\|$ so that repeated application yields the result in (2.17). 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 $0 < \eta(\eta L^2(\varepsilon^2 - 1) - 2(\varepsilon L - \gamma))$ and hence we must have that $\eta L^2(\varepsilon^2 - 1) - 2(\varepsilon L - \gamma) > 0$. Finally, we note that

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

thus the result follows. $\square$

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.

2.4. Stochastic projected primal-dual method. In the previous section, we showed convergence of a primal-dual algorithm. However, this algorithm requires the computation of the gradient; that is, one can either compute the function $\mathbb{E}_{w \sim D(z')}[\psi(z, w)]$ in closed form, or we have access to a black box evaluator that can compute this exactly for for any $z' \in \mathbb{R}^n \times \mathbb{R}^m$. In the first case, this means that we have explicit knowledge of the distributional map D. The latter assumes that the integral expression can be computed exactly or approximated to machine precision. However, this integral approximation problem is known to suffer from the curse of dimensionality.

When the computation of the gradient is not possible, a common approach is to utilize a stochastic gradient estimate. Accordingly, in this work we assume access to an oracle $\Omega : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ that returns a gradient estimator at a given point in the domain. Conceptually, Ω is a stochastic process indexed over $\mathcal{Z}$ and may take the form of common estimators used in the stochastic optimization literature,

$$\Omega(z) = \begin{cases} \psi(z, w_1), & w_1 \sim D(z), \\ \frac{1}{N} \sum_{i=1}^{N} \psi(z, w_i), & w_1, \ldots, w_N \overset{i.i.d.}{\sim} D(z), \end{cases}$$

where $\{w_i\}_{i=1}^{N}$ are random samples drawn from $D(z)$. These estimators represent the stochastic gradient and minibatch gradient estimator, respectively. Assuming access to Ω, we define the stochastic EPD (SEPD) method via the algorithmic map

$$\hat{\mathcal{G}}(z) = \Pi_{\mathcal{Z}} (z - \eta\Omega(z))$$

so that the SEPD algorithm generates the sequence $z_{t+1} = \hat{\mathcal{G}}(z_t)$, $t \geq 0$, starting from a point $z_0 \in \mathbb{R}^n \times \mathbb{R}^m$. In our analysis, we associate the gradient error with a class of potentially heavy tailed probability distributions. We consider the class of sub-Weibull random variables as defined next.

DEFINITION 2.14 (sub-Weibull random variable [38]). *The distribution of a random variable ξ is sub-Weibull, denoted $\xi \sim \mathrm{subW}(\theta, \nu)$, if there exists $\theta > 0, \nu > 0$ such that $\|z\|_k \leq \nu_2 k^\theta$ for all $k \geq 1$.*

In this definition, θ measures the heaviness of the tail (higher values of θ correspond to heavier tails); for example, $\theta = 1$ and $\theta = 1/2$ correspond to subexponential and sub-Gaussian random variables, respectively. The parameter ν represents a proxy for the variance of ξ [38, 40].

In addition to the fact that heavy tailed distributions are well-motivated in the literature on stochastic gradient methods [8, 17], they allow us to derive results that are also applicable to subexponential and sub-Gaussian random variables, as well as random variables whose distribution has a finite support.

In the following we provide results for the first moment and second moment. These results will require some of the next assumptions.

Assumption 5 (stochastic framework). Let Ω be a gradient estimator with random variables w defined over the probability space $(\mathcal{S}, \mathcal{F}, \mathbf{P})$, where $\mathcal{S}$ is a sample space with σ-algebra $\mathcal{F}$, and probability measure $\mathbf{P}$. Suppose that the following hold.
1. *Filtration*. Let $(\mathcal{S}, \mathcal{F}, \mathbf{F}, \mathbf{P})$ be a filtered probability space with filtration $\mathbf{F} = \{\mathcal{F}_t\}_{t \geq 0}$, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and Ω is $\mathcal{F}_t$ measurable for all $t \geq 0$.
2. *Sub-Weibull error*. Let Ω be such that there exist a tail parameter $\theta > 0$ and bounded variance proxy function $\nu : \mathbb{R}^d \times \mathbb{R}^n \mapsto \mathbb{R}_{\geq 0}$ such that

$$(2.20) \qquad \xi(z) = \|\Omega(z) - \mathop{\mathbb{E}}_{w \sim D(z)}[\psi(z, w)]\| \sim \mathrm{subW}(\theta, \nu(z))$$

for all $z \in \mathcal{Z}$. Denote $\tilde{\nu} \in (0, \infty)$ such that $\nu(z) \leq \tilde{\nu}$ for all $z \in \mathcal{Z}$.
3. *Unbiased Gradient Estimator*. The gradient estimator Ω is unbiased in that $\mathbb{E}_{w \sim D(z)} \Omega(z) = \Psi(z; z)$ for all $z \in \mathcal{Z}$.

While we assume that the gradient estimator may be defined over the whole space $\mathbb{R}^n \times \mathbb{R}^m$, notice that we only place the sub-Weibull assumption on the restriction of Ω to the constraint set $\mathcal{Z}$. A typical assumption in the literature is that the norm of ξ is uniformly bounded in expectation. Here, we assume that the norm of the gradient error is distributed according to a heavy tailed distribution. If the variance proxy function ν is continuous over the compact set $\mathcal{Z}$, we retrieve the uniform boundedness property. By assuming θ is fixed for any $z \in \mathcal{Z}$, we assume that all realizations of the process belong to the same sub-Weibull class.

THEOREM 2.15 (first moment). *Suppose that Assumptions 1 to 4 hold and $\frac{\varepsilon L}{\gamma} < 1$. If Ω satisfies Assumption 5.2, and η satisfies the bound in (2.18) then the following hold:*
1. *Expectation. The sequence $\{z_t\}_{t \geq 0}$ satisfies the bound in expectation,*

$$(2.21) \qquad \mathbb{E}\|z_t - \bar{z}\| \leq \alpha^t \|z_0 - \bar{z}\| + \frac{\bar{\nu}\eta}{1 - \alpha}$$

for all $t \geq 0$.

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

$$(2.22) \qquad \mathbb{P}\left(\|z_t - \bar{z}\| \leq \alpha^t \|z_0 - \bar{z}\| + c(\theta)\log^\theta\left(\frac{2}{\delta}\right)\frac{\bar{\nu}\eta}{1-\alpha} \right) \geq 1 - \delta$$

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

Before proving the result of the theorem, we provide the following supporting lemmas that will be utilized in the proof.

LEMMA 2.16 (equivalent characterizations). *If ξ is a sub-Weibull random variable with tail parameter $\theta > 0$, then the following characterizations are equivalent (we recall that $\|z\|_k = \mathbb{E}[|z|^k]^{1/k}$):*
  (c1) *Tail probability:* $\exists\, \nu_1 > 0$ *such that* $\mathbb{P}(|z| \geq \epsilon) \leq 2\exp(-(\epsilon/\nu_1))^{1/\theta}$ *for all $\epsilon > 0$.*
  (c2) *Moment:* $\exists\, \nu_2 > 0$ *such that* $\|z\|_k \leq \nu_2 k^\theta$ *for all $k \geq 1$.*
Moreover, if (c2) *holds for a given $\nu_2 > 0$, then* (c1) *holds with $\nu_1 = \left(\frac{2e}{\theta}\right)^\theta \nu_2$.*

LEMMA 2.17 (sub-Weibull inclusion). *If $\xi \sim \mathrm{subW}(\theta, \nu)$ based on* (c2) *and $\theta', \nu' > 0$ such that $\theta \leq \theta'$, and $\nu \leq \nu'$, then $\xi \sim \mathrm{subW}(\theta', \nu')$.*

LEMMA 2.18 (sub-Weibull closure). *If $\xi_1 \sim \mathrm{subW}(\theta_1, \nu_1)$, $\xi_2 \sim \mathrm{subW}(\theta_2, \nu_2)$ are (possibly coupled) sub-Weibull random variables based on* (c2) *and $c \in \mathbb{R}$, then the following hold:*
  1. $\xi_1 + \xi_2 \sim \mathrm{subW}(\max\{\theta_1, \theta_2\}, \nu_1 + \nu_2)$;
  2. $\xi_1\xi_2 \sim \mathrm{subW}(\theta_1 + \theta_2, \psi(\theta_1, \theta_2)\nu_1\nu_2)$, $\psi(\theta_1, \theta_2) := (\theta_1 + \theta_2)^{\theta_1 + \theta_2}/(\theta_1^{\theta_1}\theta_2^{\theta_2})$;
  3. $c\xi_1 \sim \mathrm{subW}(\theta_1, |c|\nu_1)$.

The proofs of these lemmas can be found in [38, 40].

Proof. *Proof of Theorem* 2.15. For notational convenience, we denote the equilibrium error as $e_t = \|z_t - \bar{z}\|$ throughout. We proceed by treating the stochastic gradient step as an inexact step using the triangle inequality:

$$(2.23) \qquad e_{t+1} = \|\hat{\mathcal{G}}(z_t) - \mathcal{G}(\bar{z}; \bar{z})\| \leq \|\hat{\mathcal{G}}(z_t) - \mathcal{G}(z_t; z_t)\| + \|\mathcal{G}(z_t; z_t) - \mathcal{G}(\bar{z}; \bar{z})\|.$$

To bound the first quantity, applying nonexpansiveness and Assumption 5.2 yield

$$\|\hat{\mathcal{G}}(z_t) - \mathcal{G}(z_t; z_t)\| \leq \eta\|\Omega(z_t) - \Psi(z_t; z_t)\| = \eta\xi(z_t).$$

It follows from Theorem 2.13 that $\|\mathcal{G}(z_t; z_t) - \mathcal{G}(\bar{z}; \bar{z})\| \leq \alpha\|z_t - \bar{z}\|$.

Combining these results yields $e_{t+1} \leq \alpha^{t+1}e_0 + \eta\sum_{i=0}^{t}\alpha^i\xi_{t-i}$. Taking the expectation yields $\mathbb{E}[e_{t+1}] \leq \alpha^{t+1}e_0 + \sum_{i=0}^{t}\alpha^i\mathbb{E}[\xi_{t-i}]$, and applying the results of Lemmas 2.17 and 2.18 gives us the expected result in (2.21).

Now, denote $\omega_{t+1} = \alpha^{t+1}e_0$ and $\sigma_t = \eta\sum_{i=1}^{t}\alpha^i\xi_{t-i}$ so that $e_{t+1} \leq \omega_{t+1} + \sigma_t$. From our sub-Weibull assumption, we have that $\xi_{t-i} \sim \mathrm{subW}(\theta, \nu_{t-i})$, where $\nu_{t-i} = \nu(z_{t-i})$. It follows from the closure under product and addition in Lemma 2.18 that $\sigma_t \sim \mathrm{subW}(\theta, \Delta)$ with $\Delta = \bar{\nu}\eta(1-\alpha)^{-1}$. Hence,

$$(2.24) \qquad \mathbb{P}(\sigma_t \geq \epsilon) \leq 2\exp\left(-\frac{\theta}{2e}\left(\frac{\epsilon}{\Delta}\right)^{\frac{1}{\theta}}\right).$$

By setting the right-hand side equal to δ, we find that $\epsilon = c(\theta)\log^\theta\left(\frac{2}{\delta}\right)\Delta$ where $c(\theta) = \left(\frac{2e}{\theta}\right)^\theta$. Now, observe that our stochastic recursion implies that for any $a > 0$, $\mathbb{P}(\omega_{t+1} + \sigma_t \geq a) \geq \mathbb{P}(e_{t+1} \geq a)$. It follows that setting $a = \omega_{t+1} + \epsilon$ yields

$$\mathbb{P}(e_{t+1} \leq \omega_{t+1} + \epsilon) \geq \mathbb{P}(\omega_{t+1} + \sigma_t \leq \omega_{t+1} + \epsilon) = \mathbb{P}(\sigma_t \leq \epsilon) \geq 1 - \delta;$$

thus the result follows by substituting the expression for ω_{t+1} and ϵ. $\square$

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

THEOREM 2.19 (neighborhood convergence). *Suppose that Assumptions 1–4 hold, and Ω satisfies Assumption 5.2. Assume that η satisfies condition* (2.18). *Then, the sequence of iterates $\{z_t\}_{t\geq0}$ converges to a neighborhood of $\bar{z}$ in expectation and almost surely. In particular,*

$$\limsup_{t\to\infty} \mathbb{E}\|z_t - \bar{z}\| \leq \frac{\eta\bar{\nu}}{1-\alpha} \quad and \quad \mathbb{P}\left(\limsup_{t\to\infty}\|z_t - \bar{z}\| \leq \frac{\eta\bar{\nu}}{1-\alpha}\right) = 1.$$

Proof. The limit of the expectation follows immediately from the above. As for almost sure convergence, we simply apply the Borel–Cantelli lemma. As before we let $e_t = \|z_t - \bar{z}\|$, so that the result in (2.21) can be compactly written as $\mathbb{E}[e_t] \leq \alpha^t e_0 + \eta\bar{\nu}(1-\alpha)^{-1}$. Denote $E_t = \max\{0, e_t\}$ so that $\mathbb{E}[E_t] \leq \alpha^t e_0$.

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

Notice that Theorem 2.15 requires only Assumption 5.2 (and it does not require the filtration in Assumption 5.1). However, a drawback to this first-moment analysis is that it only demonstrates convergence to a neighborhood whose radius is dictated by the proxy variance and, hence, the quality of the estimator. In what follows, we 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.

THEOREM 2.20 (second moment convergence). *Suppose Assumptions 1 to 5 are satisfied and we denote $\mathbb{E}_t = \mathbb{E}_{w\sim D(z_t)}[\ \cdot\ |\mathcal{F}_t]$. Then the following inequalities hold:*
  1. *One step bound. The sequence $\{z_t\}_{t\geq0}$ generated by SEPD satisfies*

$$\mathbb{E}_t\|z_{t+1} - \bar{z}\|^2 \leq \left(1 - 2(\gamma - \varepsilon L)\eta_t + 2(1+\varepsilon)^2 L^2\eta_t^2\right)\|z_t - \bar{z}\|^2 + \eta_t^2\bar{\nu}^2 2^{1+2\theta}.$$

  2. *Convergence. If the step size is $\eta_t = \ell(\kappa + t)^{-1}$, where*

$$(2.25) \qquad \ell > \frac{1}{2(\gamma - \varepsilon L)} \quad and \quad \kappa > \frac{(1+\varepsilon)^2 L^2}{(\gamma - \varepsilon L)^2},$$

  then the sequence $\{z_t\}_{t\geq0}$ generated by SEPD satisfies

$$(2.26) \quad \mathbb{E}\|z_t - \bar{z}\|^2 \leq \frac{\zeta}{\kappa + t}, \quad where \quad \zeta := \max\left\{\kappa\|z_0 - \bar{z}\|^2, \frac{\ell^2\bar{\nu}^2 2^{1+2\theta}}{2(\gamma - \varepsilon L)\ell - 1}\right\}.$$

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

$$\mathbb{E}_t\|z_{t+1} - \bar{z}\| \leq \|z_t - \bar{z}\|^2 - 2\eta_t\langle z_t - \bar{z}, \Psi(z_t; z_t) - \Psi(\bar{z}; \bar{z})\rangle + \eta_t^2\mathbb{E}_t\|\Omega(z_t) - \Psi(\bar{z}; \bar{z})\|^2.$$

To bound the inner product term, we use γ-strong-monotonicity and the gradient deviations result from Lemma 2.8: $\langle z_t - \bar{z}, \Psi(z_t; z_t) - \Psi(\bar{z}; \bar{z})\rangle \leq (\gamma - \varepsilon L)\|z_t - \bar{z}\|^2$. From the properties of the sub-Weibull random variables, we have that $\mathbb{E}_t\|\Omega(z_t) - \Psi(z_t; z_t)\|^2 \leq$

$\bar{\nu}^2 2^{2\theta}$, where $\bar{\nu}$ is an upper bound on the variance proxy function, and θ is a uniform tail parameter. By applying this result, as well as Young's inequality, we get that

$$\begin{aligned}
\mathbb{E}_t \|\Omega(z_t) - \Psi(\bar{z})\|^2 &= \mathbb{E}_t \|\Omega(z_t) - \Psi(z_t; z_t) + \Psi(z_t; z_t) - \Psi(\bar{z}; \bar{z})\|^2 \\
&\leq 2\mathbb{E}_t \|\Omega(z_t) - \Psi(z_t; z_t)\|^2 + 2\mathbb{E}_t \|\Psi(z_t; z_t) - \Psi(\bar{z}; \bar{z})\|^2 \\
&\leq \bar{\nu} 2^{1+\theta} + 2(1+\varepsilon)^2 L^2 \|z_t - \bar{z}\|^2,
\end{aligned}$$

where the last inequality follows from the fact that $z \mapsto \Psi(z; z)$ is $(1+\varepsilon)L$-Lipschitz continuous. Combining yields the one step improvement bound.

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

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

Denoting $C = 2(\gamma - \varepsilon L)$, and $\Delta = \bar{\nu}^2 \eta^2 2^{1+2\theta}$ it follows that

$$(2.27) \qquad \mathbb{E}_t \|z_{t+1} - \bar{z}\|^2 \leq (1 - C\eta_t)\|z_t - \bar{z}\|^2 + \eta_t^2 \Delta.$$

We proceed by induction. Clearly the bound in (2.26) holds for $t = 0$. Supposing it holds for t, we have that

$$\begin{aligned}
\mathbb{E}\|z_{t+1} - \bar{z}\| &\leq \left(1 - \frac{C\ell}{\kappa + t}\right) \frac{\zeta}{\kappa + t} + \frac{\Delta \ell^2}{(\kappa + t)^2} \\
&\leq \frac{\kappa + t - 1}{(\kappa + t)^2}\zeta - \frac{C\ell - 1}{(\kappa + t)^2}\zeta + \frac{\Delta \ell^2}{(\kappa + t)^2} \\
&\leq \frac{\kappa + t - 1}{(\kappa + t)^2}\zeta \\
&\leq \frac{\zeta}{(\kappa + (t+1))^2},
\end{aligned}$$

where the penultimate step follows from the fact that $(C\ell - 1)\zeta + \Delta \ell^2 < 0$. ☐

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

3. Saddle points and mixture dominance. By introducing the equilibrium point problem, we have shifted 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. In particular, we consider a condition which we call *opposing mixture dominance*. To outline the main arguments, we focus on the saddle point problem (1.1). In the following, we define the notion of opposing mixture dominance.

Assumption 6 (opposing mixture dominance). For any $x, x', x_0 \in \mathbb{R}^d$, $y, y', y_0 \in \mathbb{R}^n$, and $\tau \in [0, 1]$, the distributional map satisfies a *convex shift* in x,

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

and *concave shift* in y

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

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

Example 3.1 (Bernoulli mixtures). If the distributional map $D : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(M)$ is given by $D(x,y) = \text{Bernoulli}(p(x,y))$, where $p : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the bilinear function

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

then Assumption 6 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 3.2 (location-scale families). A distributional map $D : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^m)$ induces a location-scale family provided that for any $z \in \mathbb{R}^n \times \mathbb{R}^m$, $w \sim D(z)$ if and only if $w \stackrel{d}{=} Aw_0 + Bz + c$, where w_0 is some stationary zero-mean random variable. A sufficient condition for Assumption 6 to hold is that ϕ is convex in the random variable w. 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 $z = (x,y) \in \mathbb{R}^n \times \mathbb{R}^m$. Here, we assume without loss of generality that the distributional map induces a probability density function $p(w; x, y)$ and write the objective as $\Phi(x,y) = \int_M \phi(x,y,w)p(w; x,y)dw$. The analysis that follows is identical for the case when the density $p(w,; x, y)$ corresponds to a 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 3.3. *Let Assumption* 6 *hold. Then, for any* $z_0 \in \mathbb{R}^n \times \mathbb{R}^m$, *the function* $(x,y) \mapsto \mathbb{E}_{w \sim D(x,y)}[\phi(z_0, w)]$ *is convex-concave on* $\mathbb{R}^n \times \mathbb{R}^m$.

Proof. Fix $z_0 \in \mathcal{Z}$, $x, x' \in \mathcal{X}$, 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 a 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 $p_\tau(w) = \tau p_1(w) + (1-\tau)p_2(w)$. From this, we conclude that

$$\mathbb{E}_{w \sim \tau D(x,y)+(1-\tau)D(x',y)}[\phi(z_0, w)] \leq \tau \mathbb{E}_{w \sim D(x,y)}[\phi(z_0, w)] + (1-\tau)\mathbb{E}_{w \sim \tau D(x',y)}[\phi(z_0, w)].$$

Combining this with Assumption 6, we get that

$$\mathbb{E}_{w \sim D(\tau x+(1-\tau)x',y)}[\phi(z_0, w)] \leq \tau \mathbb{E}_{w \sim D(x,y)}[\phi(z_0, w)] + (1-\tau)\mathbb{E}_{w \sim \tau D(x',y)}[\phi(z_0, w)].$$

This proves the convexity of $x \mapsto \mathbb{E}_{w \sim D(x,y)}[\phi(z_0, w)]$ 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 Φ.

THEOREM 3.4 (strong-convexity–strong-concavity). *If Assumptions* 1 *to* 3 *and Assumption* 6 *hold, then* $(x,y) \mapsto \Phi(x,y)$ *is* $(\gamma - 2\varepsilon L)$-*strongly-convex–strongly-concave over* $\mathbb{R}^n \times \mathbb{R}^m$.

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

$$(3.1) \quad \Phi(x',y;x'y) - \Phi(x,y;x',y) \geq \left\langle x'-x, \underset{w\sim D(x',y)}{\mathbb{E}}[\nabla_x\phi(x,y,w)] \right\rangle + \frac{\gamma}{2}\|x-x'\|^2.$$

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

$$\left\langle x'-x, \underset{w\sim D(x,y)}{\mathbb{E}}[\nabla_x\phi(x,y,w)] - \underset{w\sim D(x',y)}{\mathbb{E}}[\nabla_x\phi(x,y,w)] \right\rangle \leq \varepsilon L\|x-x'\|^2$$

which is equivalent to

$$(3.2)$$
$$0 \geq \left\langle x'-x, \underset{w\sim D(x,y)}{\mathbb{E}}[\nabla_x\phi(x,y,w)] - \underset{w\sim D(x',y)}{\mathbb{E}}[\nabla_x\phi(x,y,w)] \right\rangle - \frac{2\varepsilon L}{2}\|x-x'\|^2.$$

Since for any $z_0 \in \mathbb{R}^n \times \mathbb{R}^m$ the function $(x,y) \mapsto \mathbb{E}_{w\sim D(x,y)}[\phi(z_0,w)]$ is convex-concave, we have that

$$(3.3) \quad \Phi(x,y;x,y) - \Phi(x,y;x',y) \geq \left\langle x-x', \underset{w\sim D(x',y)}{\mathbb{E}}[\phi(x,y,w)\nabla_x\log p(w;x,y)] \right\rangle$$

by setting $z_0 = (x,y)$. By adding inequalities (3.1)–(3.3) we obtain

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

which is equivalent to strong-convexity in x. Proof of strong-concavity in y follows similarly and is omitted due to space limitations. $\qquad\square$

3.1. Location-scale families. In this section, we are interested in solidifying the claims made in Example 3.2 on location-scale families, which have seen much attention in the literature on decision-dependent distributions as it arises naturally in many common examples [26]. A formal definition is provided next.

DEFINITION 3.5 (location-scale family). *The distributional map $D : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^m)$ forms a location-scale family provided that for every $z \in \mathbb{R}^n \times \mathbb{R}^m$ and $w \sim D(z)$, $w \overset{d}{=} Aw_0 + Bz + c$, where $w_0 \sim D_0$. In this model, $D_0 \in \mathcal{P}(\mathbb{R}^m)$ is a zero-mean stationary distribution while $A_0 \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times(d+n)}$, and $c \in \mathbb{R}^m$ are model parameters.*

To demonstrate that location-scale families satisfy Assumption 6, we introduce the notion of convex stochastic orders. This is an ordering of random variables induced by convex functions.

DEFINITION 3.6 (convex order [35, Definition 7.A.1]). *If two m-dimensional random vectors u and w are such that $\mathbb{E}[f(u)] \leq \mathbb{E}[f(w)]$, for all convex functions $f : \mathbb{R}^m \to \mathbb{R}$, then we say that u is less than w in the convex order and write $u \leq_{cx} w$.*

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 3.7 ([35, Theorem 7.A.1]). *The random vectors $u \sim \mu$ and $w \sim \nu$ satisfy $u \leq_{cx} w$ if and only if there exist $\hat{u} \overset{d}{=} u$ and $\hat{w} \overset{d}{=} w$ such that $\mathbb{E}[\hat{w}|\hat{u}] = \hat{u}$ 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 3.8. *Let the distributional map* $D : \mathbb{R}^d \times \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^m)$ *be a location scale family. Then for any* $z, z' \in \mathbb{R}^d \times \mathbb{R}^n$ *and* $\tau \in [0, 1]$,

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

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

Proof. Fix $\tau \in [0,1]$ and $z, z' \in \mathbb{R}^d \times \mathbb{R}^n$. In this proof, we use Theorem 3.7 to show that if $w \sim D(\tau z + (1-\tau)z')$ and $w' \sim \tau D(z) + (1-\tau)D(z')$, then we can define couplings that imply that $w \leq_{cx} w'$ and $w' \leq_{cx} w$. 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 $w' \sim \tau D(z) + (1-\tau)D(z')$ if and only if $w \overset{d}{=} Aw_0 + BT + c$.

First, we suppose that $w \sim D(\tau z + (1-\tau)z')$. Then let $w' \overset{d}{=} w - B(\tau z + (1-\tau)z') + BT$. It follows that $\mathbb{E}[w'|w] = w$ and $w' \overset{d}{=} Aw_0 + BT + c$. Hence $w' \sim \tau D(z) + (1-\tau)D(z')$. This proves that $w \leq_{cx} w'$.

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

Since this lemma holds for any convex function f, it holds for stochastic payoff ϕ provided that it is convex in w. This combined with the fact that location-scale families are ε-Lipschitz with $\varepsilon = \|B\|_2$ is sufficient for Φ to be strongly-convex–strongly-concave.

THEOREM 3.9. *Suppose that* ϕ *satisfies Assumptions* 1 *and* 2, *and the constraint sets* $\mathcal{X}$ *and* $\mathcal{Y}$ *satisfy Assumption* 4. *If* D *if a location-scale family and* ϕ *is convex in* w, *then* Φ *is* $(\gamma - 2\varepsilon L)$- *strongly-convex–strongly-concave.*

Proof. The proof amounts to demonstrating that D being a location-scale family and ϕ being convex in w is sufficient to satisfy Assumptions 6 and 3. The result then follows by Theorem 3.4. We observe that Lemma 3.8 implies that Assumption 6 holds. As for D being ε-Lipschitz, Assumption 3, we claim that $W_1(D(z), D(z')) \leq \|B\|_2 \|z - z'\|$. Then the assumption holds with $\varepsilon = \|B\|_2$. By definition,

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

where the infimum is taken over all couplings of the distributions $D(z)$ and $D(z')$. We find that if $w_0 \sim D_0$, then setting $w \overset{d}{=} Aw_0 + Bz + c$ and $w' \overset{d}{=} Aw_0 + Bz' + c$ implies that $w \sim D(z)$ and $w' \sim D(z')$ and $\|w - w'\| = \|B(z - z')\|$. Thus, the result follows. $\square$

3.2. A zeroth-order algorithm. In this section we consider the use of a zeroth-order algorithm, which we refer to as DFO, where a stochastic gradient estimator is built only using function evaluations. This algorithm is suitable in the setting where opposing mixture dominance in Assumption 6 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 [7, 10]. Denote $\mathbb{B}_k$ and $\mathbb{S}_k$ as the uniform distributions over the unit ball, $\mathcal{B}_k = \{x \in \mathbb{R}^k | \|x\| \leq 1\}$, and

unit sphere, $\mathcal{S}_k = \{x \in \mathbb{R}^k | \; \|x\| = 1\}$, in $\mathbb{R}^k$, respectively. Additionally, denote $\mathbb{S}$ and $\mathbb{B}$ as joint distributions such that $v = (v_1, v_2) \sim \mathbb{B}$, $u = (u_1, u_2) \sim \mathbb{S}$ with $v_1 \sim \mathbb{B}_d$, $v_2 \sim \mathbb{B}_n$, and $u_1 \sim \mathbb{S}_d$, $u_2 \sim \mathbb{S}_n$. The algorithmic map is then given by

$$(3.4) \qquad \mathcal{F}_t^\delta(z) = \Pi_{(1-\delta)\mathcal{Z}} \left(z - \eta_t \Omega_\delta(z) \right)$$

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

$$(3.5) \qquad \Omega_\delta(z) = \left(\frac{d}{\delta} \phi(z + \delta u, w) u_1, \; -\frac{n}{\delta} \phi(z + \delta u, w) u_2 \right),$$

where $\delta > 0$, and $u = (u_1, u_2)$ with $u_1 \sim \mathbb{S}_d$ and $u_2 \sim \mathbb{S}_n$. Note that by projecting onto the restricted set $(1-\delta)\mathcal{Z}$ we retain feasibility throughout the iterations of the algorithm. Since we are 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 7. There exist positive radii $r, R > 0$ such that $\mathcal{Z}$ satisfies $r\mathbb{B}_{d+n} \subseteq \mathcal{Z} \subseteq R\mathbb{B}_{d+n}$.

The gradient estimator in (3.5) naturally arises when considering the smoothed objective over the unit ball, given by

$$(3.6) \qquad \Phi_\delta(z) = \mathop{\mathbb{E}}_{v \sim \mathbb{B}}[\Phi(z + \delta v)] = \mathop{\mathbb{E}}_{v \sim \mathbb{B}} \left[\mathop{\mathbb{E}}_{w \sim D(z + \delta v)} [\phi(z + \delta v, w)] \right]$$

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

$$(3.7) \qquad \min_{x \in (1-\delta)\mathcal{X}} \max_{y \in (1-\delta)\mathcal{Y}} \Phi_\delta(x, y),$$

whose solutions we will we denote $z_\delta^* = (x_\delta^*, y_\delta^*)$. It follows that Ω_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 3.10 (gradient estimator). *If* $\delta > 0$, *then* $\mathbb{E}_{u \sim \mathbb{S}}[\mathbb{E}_{w \sim D(z)} \Omega_\delta(z)] = \Psi_\delta(z)$ *for all* $z \in \mathbb{R}^n \times \mathbb{R}^m$.

Proof of this result follows from [7, Lemma C.1]. The the fact that we can estimate the gradient map using only a single function evaluation is an attractive feature of (3.4). There are alternative multipoint 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 3.11 (strong monotonicity). *If the gradient of the objective* Φ, *given by* $\Psi(z) = (\nabla_x \Phi(z), -\nabla_y \Phi(z))$ *is* $(\gamma - 2\varepsilon L)$-*strongly-monotone, then* Ψ_δ *is* $(\gamma - 2\varepsilon L)$-*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 *may* be different from the solutions of the original problem. In the following, we bound the discrepancy between solutions.

LEMMA 3.12 (bounded approximation). *If $\delta < r$ and Ψ is $(\gamma - \varepsilon L)$-strongly monotone, then*

$$(3.8) \qquad \|z^* - z_\delta^*\| \le \delta \left(\left(1 + \frac{\sqrt{2L}}{(\gamma - 2\varepsilon L)}\right) \|z^*\| + \frac{2L}{(\gamma - 2\varepsilon L)} \right).$$

Finally, we are ready to demonstrate the performance of the algorithm. Here we impose the additional restriction that δ may not exceed the radius of the largest ball completely contained in $\mathcal{Z}$, which we denoted as r.

THEOREM 3.13 (convergence to the perturbed solution). *Suppose that $\delta \le r$ and $\eta_t = \ell(\kappa + t)^{-1}$ for $\ell > (2(\gamma - 2\varepsilon L))^{-1}$ $\kappa > 0$. Then, the sequence of iterates $\{z_t\}_{t \ge 0}$ generated by the derivative-free stochastic method satisfies*

$$(3.9) \quad \mathbb{E}\|z_t - z_\delta^*\|^2 \le \frac{\zeta}{\kappa + t}, \quad where \quad \zeta := \max\left\{ \kappa\|z_0 - z_\delta^*\|^2, \ \frac{B^2(n^2 + m^2)\ell^2}{\delta^2(2(\gamma - 2\varepsilon L)\ell - 1)} \right\},$$

where $B = \max_{z \in \mathcal{Z}, w \in M} |\phi(z, w)|$.

Proof. For notational convenience, we write $\hat\gamma = \gamma - 2\varepsilon L$ and $C = B^2(n^2 + m^2)\delta^{-2}$. By applying nonexpansiveness of the projection map, we get

$$\begin{aligned} \mathbb{E}_t\|z_{t+1} - z_\delta^*\|^2 &\le \mathbb{E}_t\|z_t - z_\delta^*\|^2 - 2\eta_t\mathbb{E}_t\langle z_t - z_\delta^*, \Omega_\delta(z_t)\rangle + \mathbb{E}_t\|\Omega_\delta(z_t)\|^2 \\ &\le \|z_t - z_\delta^*\|^2 - 2\hat\gamma\eta_t\|z_t - z_\delta^*\|^2 + C\eta_t^2 \\ &= (1 - 2\hat\gamma\eta_t)\|z_t - z_\delta^*\|^2 + C\eta_t^2. \end{aligned}$$

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

$$\mathbb{E}_t\|z_{t+1} - z_\delta^*\|^2 \le \frac{\kappa + t - 2\hat\gamma\ell}{\kappa + t}\|z_t - z_\delta^*\|^2 + \frac{C}{(\kappa + t)^2}.$$

As in the proof of Theorem 2.20, the result follows by induction. □

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

4. Numerical experiments for electric vehicle charging. To illustrate our results, we apply our algorithms to the electric vehicle charging problem outlined in subsection 1.1.1. This is a competition between two providers in which each provider seeks to maximize their profit.

In our simulation, each provider has access to three distinct regions, each of which has one station. The demand for each station is dictated by the data distributions from [16]. Each station is comprised of 50, 150, or 350 kW chargers with either 2 or 6 ports. We randomize this allocation at initialization. Data are processed by averaging the demand over each hour-long time window. After picking an hour block, we rescale the data by subtracting the mean and dividing by the variance. We choose the demand change in the 12–1 pm block, and depict data for the year in Figure 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 $\mathcal{X} = \mathcal{Y} = [-1, 2]^3$.
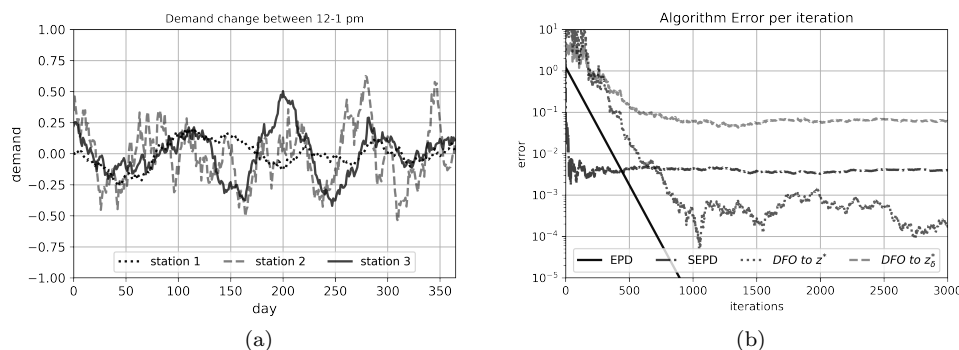
FIG. 1. *Data and results from numerical experiments. In* (a) *deviation in average demand for provider one's stations between* 12 *and* 1 *pm over* 365 *days.* (b) *The error of each algorithm depicted over* 300 *iterations. Error of the derivative-free method is depicted in both distance to the saddle point* z^* *as well as distance to the perturbed saddle point* z^*_δ.

We run each algorithm for 10,000 iterations, and depict the first 3,000 iterations in Figure 1 to provide a side-by-side comparison. The equilibrium 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 $\|z_t - \bar{z}\|^2$ and $\|z_t - z^*\|^2$. We run SEPD and the zeroth-order algorithm with the polynomial decay step-size schedules described in (2.26) and Theorem 3.13. In the latter, we choose a fixed δ value of 0.05. Relative to EPD, our results for these stochastic algorithms only guarantee sublinear 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.[2]

5. Concluding remarks. This paper focused on stochastic saddle point problems with decision-dependent distributions. We introduced the notion of equilibrium points and provided conditions for their existence and uniqueness. We showed that the distance between the two classes of solutions is bounded provided that the objective has a strongly-convex–strongly-concave payoff and Lipschitz continuous distributional map. We developed and analyzed deterministic and stochastic primal-dual algorithms. In particular, using a sub-Weibull model for the errors emerging in the gradient computation, we provided error bounds in expectation and in high probability that hold for each iteration; we also showed convergence to a neighborhood in expectation and a.s. Finally, we investigate an opposing mixture dominance condition that ensures the objective is strongly-convex-strongly–concave, and we focused on a zeroth-order algorithm.

Future directions will explore proximal-based methods for problems with a more general geometry. We will also consider alternative derivative-free methods and approaches that incorporate gradient information by estimating the distributional map.

REFERENCES

[1] J. ACHARYA AND R. D. YATES, *Service provider competition and pricing for dynamic spectrum allocation*, in 2009 International Conference on Game Theory for Networks, IEEE, Piscataway, NJ, 2009, pp. 190–198.

---

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

[2] C. D. ALIPRANTIS AND K. BORDER, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, Berlin, 2006.

[3] N. BASTIANELLO, L. MADDEN, R. CARLI, AND E. DALL'ANESE, *A Stochastic Operator Framework for Inexact Static and Online Optimization*, preprint, arXiv:2105.09884, 2021.

[4] G. BIANCHIN, M. VAQUERO, J. CORTES, AND E. DALL'ANESE, *Data driven synthesis of optimization-based controllers for regulation of unkown linear systems*, in 2021 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, pp. 5783–5788.

[5] J. R. BIRGE AND F. LOUVEAUX, *Introduction to Stochastic Programming*, Springer, New York, 2011.

[6] V. I. BOGACHEV AND A. V. KOLESNIKOV, *The Monge-Kantorovich problem: Achievements, connections, and perspectives*, Russian Math. Surveys, 67 (2012), pp. 785–890.

[7] M. BRAVO, D. LESLIE, AND P. MERTIKOPOULOS, *Bandit learning in concave n-person games*, in Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Red Hook, NY, 2018, pp. 5666–5676.

[8] U. ŞIMŞEKLI, L. SAGUN, AND M. GÜRBÜZBALABAN, *A tail-index analysis of stochastic gradient noise in deep neural networks*, in International Conference on Machine Learning, Vol. 97, Curran Associates, Red Hook, NY, 2019, pp. 5827–5837.

[9] E. DALL'ANESE, A. SIMONETTO, S. BECKER, AND L. MADDEN, *Optimization and learning with information streams: Time-varying algorithms and applications*, IEEE Signal Process. Mag., 37 (2020), pp. 71–83.

[10] D. DRUSVYATSKIY, M. FAZEL, AND L. J. RATLIFF, *Improved rates for derivative free gradient play in strongly monotone games*, in 2022 IEEE 61st Conference on Decision and Control (CDC), IEEE, Piscataway, NJ, 2022, pp. 3403–3408.

[11] D. DRUSVYATSKIY AND L. XIAO, *Stochastic optimization with decision-dependent distributions*, Math. Oper. Res., 48 (2023), pp. 954–998.

[12] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2003.

[13] F. FACCHINEI, J.-S. PANG, AND G. SCUTARI, *Non-cooperative games with minmax objectives*, Comput. Optim. Appl., 59 (2014), pp. 85–112.

[14] F. FELE AND K. 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), IEEE, Piscataway, NJ, 2019, pp. 1–6.

[15] A. D. FLAXMAN, A. T. KALAI, AND H. B. MCMAHAN, *Online convex optimization in the bandit setting: Gradient descent without a gradient*, in 16th Annual ACM-SIAM Symposium on Discrete Algorithms, 2005, pp. 385–394.

[16] M. GILLERAN, E. BONNEMA, J. WOODS, P. MISHRA, I. DOEBBER, C. HUNTER, M. MITCHELL, AND M. MANN, *Impact of electric vehicle charging on the power demand of retail buildings*, Adv. Appl. Energy, 4 (2021), 100062.

[17] M. GÜRBÜZBALABAN, U. ŞIMŞEKLI, AND L. ZHU, *The heavy-tail phenomenon in SGD*, in International Conference on Machine Learning, Vol. 139, Curran Associates, Red Hook, NY, 2021, pp. 3964–3975.

[18] M. T. HALE, A. NEDIĆ, AND M. EGERSTEDT, *Asynchronous multiagent primal-dual optimization*, IEEE Trans. Automat. Control, 62 (2017), pp. 4421–4435.

[19] L. V. KANTOROVICH AND S. RUBINSHTEIN, *On a space of totally additive functions*, Vestnik Leningrad University, 13 (1958), pp. 52–59.

[20] J. KOSHAL, A. NEDIĆ, AND U. V. SHANBHAG, *Multiuser optimization: Distributed algorithms and error analysis*, SIAM J. Optim., 21 (2011), pp. 1046–1081.

[21] A. K. KUCHIBHOTLA AND A. CHAKRABORTTY, *Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression*, Inform. Inference, 11 (2022), pp. 1389–1456.

[22] J. LE NY, A. RIBEIRO, AND G. J. PAPPAS, *Adaptive communication-constrained deployment of unmanned vehicle systems*, IEEE J. Sel. Areas Commun., 30 (2012), pp. 923–934.

[23] C. LI, Z. DONG, G. CHEN, B. ZHOU, J. ZHANG, AND X. YU, *Data-driven planning of electric vehicle charging infrastructure: A case study of Sydney, Australia*, IEEE Trans. Smart Grid, 12 (2021), pp. 3289–3304.

[24] G. LI, S. C. H. HOI, K. CHANG, W. LIU, AND R. JAIN, *Collaborative online multitask learning*, IEEE Trans. Knowl. Data Eng., 26 (2014), pp. 1866–1876.

[25] C. MENDLER-DÜNNER, J. PERDOMO, T. ZRNIC, AND M. HARDT, *Stochastic optimization for performative prediction*, Adv. Neural Inf. Process. Syst., 33 (2020), pp. 4929–4939.

[26] J. P. MILLER, J. C. PERDOMO, AND T. ZRNIC, *Outside the echo chamber: Optimizing the performative risk*, Proc. Mach. Learn. Res. (PMLR), 139 (2021), pp. 7710–7720.

[27] A. MOKHTARI, A. OZDAGLAR, AND S. PATTATHIL, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach*, Proc. Mach. Learn. Res. (PMLR), 108 (2020), pp. 1497–1507.

[28] M. NATOLE, Y. YING, AND S. LYU, *Stochastic proximal algorithms for auc maximization*, Proc. Mach. Learn. Res. (PMLR), 80 (2018), pp. 3710–3719.

[29] A. NEMIROVSKI, *Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251.

[30] A. NEMIROVSKI, A. JUDITSKY, G. LAN, AND A. SHAPIRO, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[31] J. PERDOMO, T. ZRNIC, C. MENDLER-DÜNNER, AND M. HARDT, *Performative prediction*, Proc. Mach. Learn. Res. (PMLR), 119 (2020), pp. 7599–7609.

[32] T. K. PONG, P. TSENG, S. JI, AND J. YE, *Trace norm regularization: Reformulations, algorithms, and multi-task learning*, SIAM J. Optim., 20 (2010), pp. 3465–3489.

[33] A. RIBEIRO, *Ergodic stochastic optimization algorithms for wireless communication and networking*, IEEE Trans. Signal Process., 58 (2010), pp. 6369–6386.

[34] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 2009.

[35] M. SHAKED AND J. G. SHANTHIKUMAR, *Stochastic Orders*, Springer, New York, 2007.

[36] A. SHAPIRO AND A. NEMIROVSKI, *On complexity of stochastic programming problems*, in Continuous Optimization, Springer, New York, 2005, pp. 111–146.

[37] R. K. SOLANKI, K. VERMA, AND R. KUMAR, *Spam filtering using hybrid local-global naive Bayes classifier*, in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, Piscataway, NJ, 2015, pp. 829–833.

[38] M. VLADIMIROVA, S. GIRARD, H. NGUYEN, AND J. ARBEL, *Sub-Weibull distributions: Generalizing sub-Gaussian and subexponential properties to heavier tailed distributions*, Stat, 9 (2020), e318.

[39] K. WEINBERGER, A. DASGUPTA, J. LANGFORD, A. SMOLA, AND J. ATTENBERG, *Feature hashing for large scale multitask learning*, in Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, 2009, pp. 1113–1120.

[40] K. C. WONG, Z. LI, AND A. TEWARI, *Lasso guarantees for β-mixing heavy-tailed time series*, Ann. Statist., 48 (2020), pp. 1124–1142.

[41] K. WOOD, G. BIANCHIN, AND E. DALL'ANESE, *Online projected gradient descent for stochastic optimization with decision-dependent distributions*, IEEE Control Syst. Lett., 6 (2021), pp. 1646–1651.

[42] J. ZHANG, M. HONG, M. WANG, AND S. ZHANG, *Generalization bounds for stochastic saddle point problems*, Proc. Mach. Learn. Res. (PMLR), 130 (2021), pp. 568–576.

[43] X. ZHANG, N. S. AYBAT, AND M. GÜRBÜZBALABAN, *Robust Accelerated Primal-Dual Methods for Computing Saddle Points*, preprint, arXiv:2111.12743, 2021.