Proceedings of Machine Learning Research vol 211:1–13, 2023 5th Annual Conference on Learning for Dynamics and Control

Online Saddle Point Tracking with Decision-Dependent Data

Killian Wood KILLIAN.WOOD@COLORADO.EDU
Department of Applied Mathematics, University of Colorado Boulder

Emiliano Dall'Anese EMILIANO.DALLANESE@COLORADO.EDU
Department of Electrical, Computer, and Energy Engineering and Department of Applied Mathematics, University of Colorado Boulder

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

In this work, we consider a time-varying stochastic saddle point problem in which the objective is revealed sequentially, and the data distribution depends on the decision variables. Problems of this type express the distributional dependence via a distributional map, and are known to have two distinct types of solutions—saddle points and equilibrium points. We demonstrate that, under suitable conditions, online primal-dual type algorithms are capable of tracking equilibrium points. In contrast, since computing closed-form gradient of the objective requires knowledge of the distributional map, we offer an online stochastic primal-dual algorithm for tracking equilibrium trajectories. We provide bounds in expectation and in high probability, with the latter leveraging a sub-Weibull model for the gradient error. We illustrate our results on an electric vehicle charging problem where responsiveness to prices follows a location-scale family based distributional map.

Keywords: Online optimization, minimax problems, decision-dependent data.

1. Introduction

The general goal of stochastic optimization is to find optimal decisions in systems with parameters dictated by data Nemirovski et al. (2009); Shapiro and Nemirovski (2005); Zhang et al. (2021a). In statistical learning, optimal decisions represent model parameters that best fit a mapping between feature and label data (see, e.g., Şimşekli et al. (2019); Gürbüzbalaban et al. (2021)). In the context of optimization of physical and dynamical systems, they may model externalities or system parameters that are predicted from data and are accompanied by given error statistics (see, e.g., Berberich et al. (2020); Li et al. (2021); Bianchin et al. (2021)). To analyze these problems, works posit that the data distributions are stationary Birge and Louveaux (2011); assumption may be violated when population data shifts in response to previously deployed decisions, thus making said decisions sub-optimal. Hence the distribution is inextricably tied to the decision variables.

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

$$\min_{x\in\mathcal{X}_t}\max_{y\in\mathcal{Y}_t}\left\{F_t(x,y) := \mathop{\mathbb{E}}_{w\sim D_t(x,y)}[f_t(x,y,w)]\right\} \tag{1}$$

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, $f_t : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$ 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 $\mathbb{R}^k$).

Examples of problems of the form (1) emerge in cost maximization in competitive markets, where the (stochastic) demand shifts in response to prices (see, e.g., Turan and Alizadeh (2021); Maheshwari et al. (2021)), 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, consider a competition between two service providers in an area with n 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

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

where $x=(x_i)_{i=1}^n$ and $y=(y_i)_{i=1}^n$ 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 $i\in[n]$); $\Gamma_t^1,\Gamma_t^2\in\mathbb{R}^{n\times n}$ are the charging rate utility matrices; $c_t\in\mathbb{R}^n$ is the location-based utility vector (i.e., cost of operation); and $a,b\in\mathbb{R}^n$ are changes in demand in each region (in response to price changes) with distributions $a\overset{d}{=}a_0^t+A_1^t x+B_1^t y$, and $b\overset{d}{=}b_0^t+A_2^t x+B_2^t y$. Here a_0^t and b_0^t are random variables drawn from zero-mean stationary distributions.

Classical solutions to (1) are saddle points, which we denote $z_t^*=(x_t^*,y_t^*)\in\mathcal{X}_t\times\mathcal{Y}_t$. Under appropriate conditions, namely minimax equality, saddle points satisfy

$$x_t^*\in\arg\min_{x\in\mathcal{X}_t}\max_{y\in\mathcal{Y}_t}F_t(x,y),\quad y_t^*\in\arg\max_{y\in\mathcal{Y}_t}\min_{x\in\mathcal{X}_t}F_t(x,y). \tag{3}$$

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 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 Narang et al. (2022); Wood and Dall'Anese (2022). 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

$$\begin{aligned}
x_t^{\ell+1}&\in\arg\min_{x\in\mathcal{X}_t}\max_{y\in\mathcal{Y}_t}\mathop{\mathbb{E}}_{w\sim D_t(x_t^\ell,y_t^\ell)}[f_t(x,y,w)],\\
y_t^{\ell+1}&\in\arg\max_{y\in\mathcal{Y}_t}\min_{x\in\mathcal{X}_t}\mathop{\mathbb{E}}_{w\sim D_t(x_t^\ell,y_t^\ell)}[f_t(x,y,w)].
\end{aligned} \tag{4}$$

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 $\{\bar{z}_t\}=\{\bar{x}_t,\bar{y}_t\}$ without requiring that we take the sequences in 4 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 (Besbes et al. (2015); Jadbabaie et al. (2015); Cao et al. (2020); Shames and Farokhi (2020); Dall'Anese et al. (2020); Wood et al. (2021)).

1.1. Related Work

Stochastic Saddle Point Problems. Algorithms for computing saddle points can be loosely catagorized as primal-dual based or proximal based Nemirovski (2004); Mokhtari et al. (2020); Nemirovski et al. (2009); Koshal et al. (2011); Mokhtari et al. (2020); Zhang et al. (2021b). Some works seek to find approximate saddle points by analyzing a saddle point gap. We conduct our analysis in a setting in which solutions are known to be unique, so we simply track them. Our analysis is primarily conducted through variational analysis Rockafellar and Wets (2009). Hence we define the appropriate gradient maps and demonstrate that solutions to our problems are the solution to the variational inequalities induced by said gradient maps.

Decision Dependent Distributions. This work is most closely related to the literature on stochastic optimization with decision dependent distributions, or its counterpart in learning, performative prediction. The problem of finding optimal decisions that are robust to decision-dependent data has been studied extensively, and in many distinct settings: minimization problems Drusvyatskiy and Xiao (2022); Perdomo et al. (2020), saddle-point problems Wood and Dall'Anese (2022), games Narang et al. (2022), online Wood et al. (2021), time-varying decay Ray et al. (2022). Relative to the existing work on saddle point problems in the literature, this work considers problems for which the objective, constraints and distribution are time-varying and revealed sequentially in time. The work on games is related, as specific instances of games such as two-player zero-sum may be cast into a saddle point problem. Saddle point problems however are however not a strict subset of games as they exist in their own right; arising from constrained minimization problems, etc. The most obvious inspiration for this work is that of Wood and Dall'Anese (2022), as the setting in this work is precisely stochastic saddle point problems with decision-dependent distributions for time-invariant problems. Relative to this work, the results we present here are extensions to the online setting where analysis requires handling of additional noise due to solution drift.

Online Convex Optimization. Relevant to works on online optimization that are concerned with tracking trajectories (see the representative works Popkov (2005); Selvaratnam et al. (2018); Mokhtari et al. (2016); Madden et al. (2021); Cutler et al. (2021)) or given comparator sequences Jadbabaie et al. (2015); another line of work is concerned with finding a sequence of decision that minimize a suitable dynamic regret metric Hazan (2019). Our metric is the distance to the solution of (1) at the current iteration, where we incorporate the drift of the solution trajectory. We account for the time-variability of the solution by incorporating the solution drift into our guarantees.

1.2. Contributions

Our contributions are as follows.
(c1) The Online Equilibrium Problem. We propose a notion of equilibrium points for the time-varying saddle-point problem in 1, provide conditions to guarantee existence and uniqueness, and provide bounds for the distance between the unique equilibrium points (1).
(c2) Online Algorithms. We demonstrate that primal-dual algorithms, using the gradients of f_t, are effective at finding equilibrium points when the stochastic objective f_t is strongly-convex-strongly-concave for any realization of w. First, we demonstrate effective tracking of a conceptual algorithm using full gradient information. We then demonstrate that a stochastic algorithm tracks equilibrium points with additional noise due to estimation. Furthermore, we provide provide expectation bounds and high probability bounds that hold for each iteration.

(c3) Experiments. We illustrate our results on the electric vehicle charging problem in (2) by incorporating synthetic demand data from Gilleran et al. (2021). Here, the demand changes in response to prices with a location-scale family based distributional map.

2. Equilibrium Points

In this section we define the equilibrium problem, the fixed points of the repeated retraining heuristic in 4, and provide sufficient conditions for their existence. We start from the definition of equilibrium points.

Definition 1 ***(Equilibrium Points)*** *A pair $(\bar{x}_t, \bar{y}_t) \in \mathcal{X}_t \times \mathcal{Y}_t$ is an equilibrium point if:*

$$\begin{aligned} \bar{x}_t &\in \arg\min_{x \in \mathcal{X}_t} \left\{ \max_{y \in \mathcal{Y}_t} \mathop{\mathbb{E}}_{w \sim D_t(\bar{x}_t, \bar{y}_t)} [f_t(x, y, w)] \right\}, \\ \bar{y}_t &\in \arg\max_{y \in \mathcal{Y}_t} \left\{ \min_{x \in \mathcal{X}_t} \mathop{\mathbb{E}}_{w \sim D_t(\bar{x}_t, \bar{y}_t)} [f_t(x, y, w)] \right\}. \end{aligned} \tag{5}$$

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.

Theorem 2 ***(Existence of Equilibrium Points)*** *Suppose that the following assumptions hold at time $t \geq 0$:*
i) $x \mapsto f_t(x, y, w)$ is convex in x for all $y \in \mathcal{Y}_t$ and for any realization of w;
ii) $y \mapsto f_t(x, y, w)$ is concave in y for all $x \in \mathcal{X}_t$ and for any realization of w;
iii) $(x, y) \mapsto f_t(x, y, w)$ is continuous on $\mathcal{X}_t \times \mathcal{Y}_t$ for any given w;
iv) the sets $\mathcal{X}_t \subset \mathbb{R}^n, \mathcal{Y}_t \subset \mathbb{R}^n$ are convex compact subsets;
v) the distributional map $D_t : \mathcal{Z}_t \to (\mathcal{P}(M), W_1)$ is continuous.
Then the set of equilibrium points is nonempty and compact. □

The proof follows from the fact that equilibrium points exist for each problem at time t due to (Wood and Dall'Anese, 2022, Theorem 2.10). The proof strategy amounts to demonstrating that the repeated retraining map satisfies Kakutani's Fixed Point Theorem; see, e.g., (Aliprantis and Border, 2006, Corollary 17.55); it is not provided due to space limitations.

2.1. Theoretical Framework

In light of our discussion on existence of equilibrium points, we outline the assumptions and some results that will be necessary later in our analysis. For notational convenience, we will refer to the stacked variable $z = (x, y)$ and the Cartesian product set of constraints $\mathcal{Z}_t = \mathcal{X}_t \times \mathcal{Y}_t$. We will rely on the following assumptions to hold at each time t throughout this work.

Assumption 1 ***(Strong-Convexity-Strong-Concavity)*** *The function $(x, y) \mapsto f_t(x, y, w)$ is continuously differentiable over $\mathbb{R}^n \times \mathbb{R}^m$ 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 2 *(**Joint Smoothness**) The map $g_t(z,w) := (\nabla_x f_t(z,w), -\nabla_y f_t(z,w))$ is L-Lipschitz in z and w. Namely, $\|g_t(z,w) - g_t(z',w)\| \le L\|z - z'\|$, $\|g_t(z,w) - g_t(z,w')\| \le L\,\mathrm{d}(w,w')$, for any $z, z' \in \mathbb{R}^n \times \mathbb{R}^m$ and w, w' supported on $\mathbb{R}^k$, for some $L \ge 0$, where $\mathrm{d} : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is some chosen metric on $\mathbb{R}^k$.* □

Assumption 3 *(**Lipschitz-Continuous Distributional Map**) The distributional maps $D_t : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(M)$ are ε-Lipschitz. Namely, $W_1(D_t(z), D_t(z')) \le \varepsilon\|z - z'\|$, for any $z, z' \in \mathbb{R}^n \times \mathbb{R}^m$, where W_1 is the Wasserstein-1 distance.* □

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

Assumption 5 *(**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, $\Delta_t \le \Delta$ for all $t \ge 0$.* □

These assumptions provided are sufficient to guarantee uniqueness of the equilibrium point, and convergence of primal-dual algorithms in the batch setting; see Wood and Dall'Anese (2022).

Theorem 3 *(**Equilibrium Point Uniqueness**) If Assumptions 1-4 are satisfied such that $\varepsilon L < \gamma$, then a unique equilibrium point exists.*

Proof of this results amounts to showing that the repeated retraining heuristic in 4 is a strict contraction and hence satisfies the Banach-Picard Fixed Point Theorem. For a detailed proof, see Wood and Dall'Anese (2022).

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

$$G_t(z;z') := \mathop{\mathbb{E}}_{w\sim D_t(z')} g_t(z,w) = \left(\mathop{\mathbb{E}}_{w\sim D_t(z')} \nabla_x f_t(z,w), \mathop{\mathbb{E}}_{w\sim D_t(z')} -\nabla_y f_t(z,w) \right) \tag{6}$$

for all $z, z' \in \mathbb{R}^n \times \mathbb{R}^m$. 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 4 *(**Gradient Map Characterization**) If Assumptions 1-4 hold, then:*

1. *(**Gradient Deviation**) For any fixed $\hat{z} \in \mathcal{Z}_t$, the map $z \mapsto G_t(\hat{z}, z)$ is εL-Lipschitz-continuous. That is, $\|G_t(\hat{z};z) - G_t(\hat{z};z')\| \le \varepsilon L\|z - z'\|$, for all $z, z' \in \mathcal{Z}_t$.*
2. *(**Strong-Monotonicity**) The map $z \mapsto G_t(z,z)$ is $(\gamma - \varepsilon L)$-strongly-monotonic.*
3. *(**Lipschitz-Continuity**) The map $z \mapsto G_t(z,z)$ is $(L + \varepsilon L)$-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 2 and 3 respectively. For a detailed proof, we refer the reader to Wood and Dall'Anese (2022). Strong monotonicity and Lipschitz continuity of $z \mapsto G_t(z;z)$ 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 $z \mapsto G_t(z;z)$.

3. Online Algorithms

3.1. A Conceptual Primal-Dual Algorithm

In this section, we show that if the decoupled gradient map G_t is available, then tracking the equilibrium points is possible using a projected primal-dual algorithm. 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. We denote the projection map as $\Pi_{\mathcal{Z}_t}(z) = \arg\min_{z' \in \mathcal{Z}_t} \frac{1}{2}\|z - z'\|^2$. Then, the equilibrium primal-dual algorithmic map is given by

$$\mathcal{G}_t(z) = \Pi_{\mathcal{Z}_t}\left(z - \eta G_t(z)\right), \tag{7}$$

so that the algorithm generates the sequence $\{z_t\}_{t\geq 0}$ defined by $z_{t+1} = \mathcal{G}_t(z_t)$, $t \in \mathbb{N}$. To proceed, we observe that equilibrium points are the fixed points of the primal-dual algorithmic map.

Proposition 5 ***(Fixed Point Characterization)*** *Let Assumptions 1-4 hold and suppose that $\frac{\varepsilon L}{\gamma} < 1$. A point $\bar{z}_t \in \mathcal{Z}_t$ is an equilibrium point if and only if $\bar{z}_t = \mathcal{G}_t(\bar{z}_t)$.* □

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 6 ***(Primal-Dual Tracking)*** *Suppose that Assumptions 1-4 hold and that $\frac{\varepsilon L}{\gamma} < 1$. Then the sequence $z_{t+1} = \mathcal{G}_t(z_t)$ satisfies the bound*

$$\|z_t - \bar{z}_t\| \leq \alpha^t \|z_0 - \bar{z}_0\| + (1-\alpha)^{-1}\Delta \tag{8}$$

for any initial point $z_0 \in \mathcal{Z}_t$, and $\alpha := \sqrt{1 - \eta(\gamma - \varepsilon L)}$ provided that

$$\eta < \min\left\{\frac{1}{\gamma - \varepsilon L}, \frac{\gamma - \varepsilon L}{(1+\varepsilon)^2 L^2}\right\} \tag{9}$$

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

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

$$\begin{aligned}
\|z_{t+1} - \bar{z}_t\|^2 &= \|\Pi_{\mathcal{Z}_t}\left(z_t - \eta G_t(z_t)\right) - \Pi_{\mathcal{Z}_t}\left(\bar{z}_t - \eta G_t(\bar{z}_t)\right)\|^2 \\
&\leq \|(z_t - \bar{z}_t) - \eta\left(G_t(z_t) - G_t(\bar{z}_t)\right)\|^2 \\
&\leq \|z_t - \bar{z}_t\|^2 - 2\eta\langle z_t - \bar{z}_t, G_t(z_t) - G_t(\bar{z}_t)\rangle + \eta^2\|G_t(z_t) - G_t(\bar{z}_t)\|^2.
\end{aligned}$$

If we denote $\hat{\gamma} = \gamma - \varepsilon L$ and $\hat{L} = L + \varepsilon L$, then from Lemma 4 we have that G_t is $\hat{\gamma}$-strongly monotone and $\hat{L}$-Lipschitz continuous. Combining these facts yields

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

Substituting into the above yields

$$\|z_{t+1} - \bar{z}_t\|^2 \leq (1 - \eta\hat{\gamma})\|z_t - \bar{z}_t\|^2 + \eta\left(\eta - \frac{\hat{\gamma}}{\hat{L}^2}\right)\|G_t(z_t) - G_t(\bar{z}_t)\|^2 \leq (1 - \eta\hat{\gamma})\|z_t - \bar{z}_t\|^2$$

where the last inequality follows provided that $\eta \leq \hat{\gamma}/\hat{L}^2$. It follows that if $\eta < 1/\hat{\gamma}$ as well, then $1 - \eta\hat{\gamma} < 1$ and the bound in Theorem (8) follows. Considering the limit supremum of the bound in (8) yields the result. ■

We note that the noise due to the drift in (8) 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.

3.2. 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. We say conceptual because having access to full information in G_t requires the ability to compute the expectation with respect to the distributional map D_t at each algorithmic step—which is of course impractical. Hence, we are concerned with a more pragmatic setting in which we merely have access to a stochastic gradient oracle, which we will denote H_t. We make the implicit assumption throughout that H_t is a function of the stochastic gradient function g_t defined in (6). Such functions are typically of the form

$$H_t(z) = \begin{cases} g_t(z, w_1), & w_1 \sim D_t(z), \\ \frac{1}{N}\sum_{i=1}^{N} g_t(z, w_i), & w_1, \ldots, w_N \overset{i.i.d.}{\sim} D_t(z). \end{cases} \tag{10}$$

Then, given a starting point z_0, the stochastic primal-dual algorithm performs the update

$$z_{t+1} = \hat{\mathcal{G}}_t(z_t), \text{ where } \hat{\mathcal{G}}_t(z_t) = \Pi_{\mathcal{Z}_t}\left(z_t - \eta H_t(z_t)\right) \tag{11}$$

Crucial to our analysis will be providing reasonable assumptions regarding the quality of the gradient estimator H_t. The case where $H_t(z) = g_t(z, w_1)$ is particularly appealing in applications such as competitive markets, strategic classification, etc., where $g_t(z, w_1)$ can be computed using an observation of w (in our example in competitive markets, we would observe the demands a and b).

We are interested in providing results in expectation as well as high-probability. A common assumption throughout the literature is to use a sub-Gaussian error model on this gradient error quantity—an observation supported by the central limit theorem when using a sufficiently large batch size N in (10). While this may hold in some cases, it has been observed that this requires a prohibitively large set of data while also assuming the data is of sufficiently good quality Vladimirova et al. (2020); this has also been observed in works on stochastic gradient methods such as in Şimşekli et al. (2019); Gürbüzbalaban et al. (2021). A more mild assumption then is to assume a larger class of heavy-tailed distributions known as sub-Weibull distributions, which we formalize in the following.

Definition 7 *(**Sub-Weibull Random Variable** Vladimirova et al. (2020)) 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 $\|\xi\|_p \leq \nu p^{\theta}$, for all $p \geq 1$.* □

Assumption 6 ***(Stochastic Framework)*** *Denote the gradient error incurred throughout the stochastic algorithm as* $\xi_t = H_t(z_t) - G_t(z_t)$*. Then there exists constant* $\theta, \nu > 0$ *and a sequence* $\{\nu_t\}_{t\geq 0} \subseteq \mathbb{R}_+$ *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* $\|\xi_t\| \sim \mathrm{subW}(\theta, \nu_t)$*.*

2. ***Bounded Variance Proxies****. The sequence of variance proxies* $\{\nu_t\}_{t\geq 0}$ *is bounded by* ν*.* □

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

Theorem 8 *Suppose that Assumptions 1-6 hold and* $\frac{\varepsilon L}{\gamma} < 1$*. If* η *satisfies the bound in 9 then the following hold:*

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

$$\mathbb{E}[\|z_t - \bar{z}_t\|] \leq \alpha^t \|z_0 - \bar{z}_0\| + (1-\alpha)^{-1}\Delta + (1-\alpha)^{-1}\eta\nu. \tag{12}$$

for all $t \geq 0$*, for any initial point* $z_0 \in \mathcal{Z}_t$*, and* $\alpha := \sqrt{1 - \eta(\gamma - \varepsilon L)}$*.*

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

$$\mathbb{P}\left(\|z_t - \bar{z}_t\| \leq \alpha^t \|z_0 - \bar{z}_0\| + \frac{\Delta}{1-\alpha} + c(\theta)\log^\theta\left(\frac{2}{\delta}\right)\frac{\eta\nu}{1-\alpha}\right) \geq 1 - \delta\,. \tag{13}$$

with $c(\theta) := \left(\frac{2e}{\theta}\right)^\theta$*, for any initial point* $z_0 \in \mathcal{Z}_t$*.* □

We note that the noise terms above are diametrically opposed functions of the step-size. While the drift term grows larger for small step-size, the gradient noise decreases for smaller step-size values. This relationship makes removing the contribution of any one source of perturbation impossible. We also note that the high-probability bound scales as $\log(\delta^{-1})$, as opposed to classical bounds derived using Markov's bound that scale as δ^{-1}.

Before proving the theorem, we provide supporting lemmas that will be used in the proof.

Lemma 9 ***(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 10 ***(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 11 ***(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 Vladimirova et al. (2020); Wong et al. (2020).
Proof *of* 8. As before, we have that $\|z_{t+1} - \bar{z}_{t+1}\| \leq \|z_{t+1} - \bar{z}_t\| + \Delta_t$ where

$$\begin{aligned} \|z_{t+1} - \bar{z}_t\| &\leq \|(z_t - \bar{z}_t) - \eta(H_t(z_t) - G_t(\bar{z}_t))\| \\ &= \|(z_t - \bar{z}_t) - \eta(G_t(z_t) - G_t(\bar{z}_t) - \eta\xi_t\| \leq \alpha\|z_t - \bar{z}_t\| + \eta\|\xi_t\|. \end{aligned}$$

This yields that stochastic recursion $\|z_t - \bar{z}_t\| \leq \alpha^t\|z_0 - \bar{z}_0\| + \Delta\sum_{i=0}^t \alpha^i + \eta\sum_{i=0}^t \alpha^i\|\xi_{t-i}\|$. Recall that when η satisfies the condition in (9), $\alpha < 1$. Hence assuming this fact and taking the expectation of both sides yields

$$\mathbb{E}\,\|z_t - \bar{z}_t\| \leq \alpha^t\|z_0 - \bar{z}_0\| + \frac{\Delta}{1-\alpha} + \eta\sum_{i=0}^{t}\alpha^i\,\mathbb{E}\,\|\xi_{t-i}\|$$

so that the result in (12) follows. To prove the result in (13), we denote $e_t = \|z_t - \bar{z}_t\|$, $\omega_t = \alpha_t\|z_0 - \bar{z}_0\| + \Delta(1-\alpha)^{-1}$, and $\sigma_t = \eta\sum_{i=0}^t \alpha^i\xi_{t-i}$. Observe that, due to our closure properties,

$$\|\sigma_t\|_p \leq \sum_{i=0}^{t}\alpha^i\,\mathbb{E}[\|\xi_t\|]^p]^{1/p} \leq \frac{\eta\nu}{1-\alpha}p^\theta$$

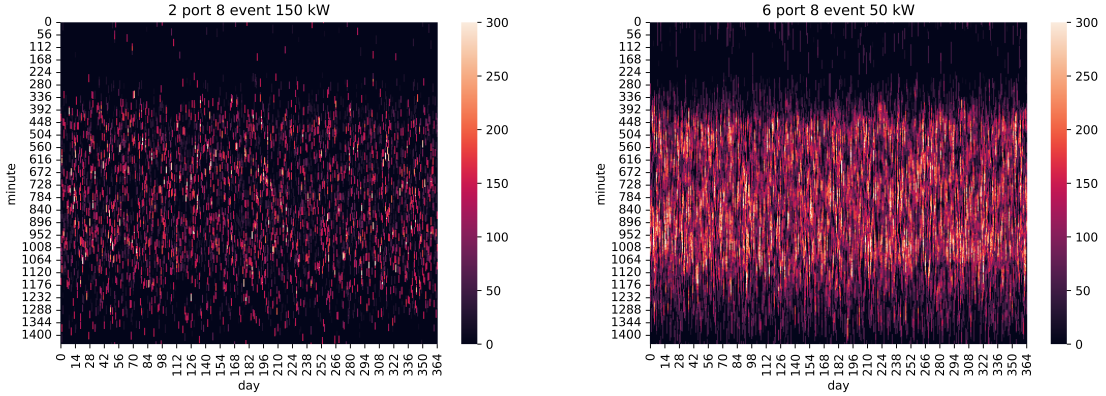
for any $p \geq 1$ and hence $\sigma_t \sim \mathrm{subW}(\theta, \eta\nu(1-\alpha)^{-1})$. It follows from Lemma 9 (c1) that

$$\mathbb{P}\left(\sigma_t \geq \epsilon\right) \leq 2\exp\left(-\frac{\theta}{2e}\left(\frac{(1-\alpha)\epsilon}{\eta\nu}\right)^{\frac{1}{\theta}}\right), \tag{14}$$

and setting the right hand side equal to $\delta > 0$ yields $\epsilon = c(\theta)\log^\theta\left(\frac{2}{\delta}\right)\eta\nu(1-\alpha)^{-1}$. Observe that our stochastic recursion implies that for any $a > 0$, $\mathbb{P}(\omega_t + \sigma_t \geq a) \geq \mathbb{P}(e_t \geq a)$. It follows that setting $a = \omega_t + \epsilon$ yields $\mathbb{P}(e_t \leq \omega_t + \epsilon) \geq \mathbb{P}(\omega_t + \sigma_t \leq \omega_t + \epsilon) = \mathbb{P}(\sigma_t \leq \epsilon) \geq 1 - \delta$, thus the result follows. ■

4. Numerical Simulations on Electric Vehicle Charging

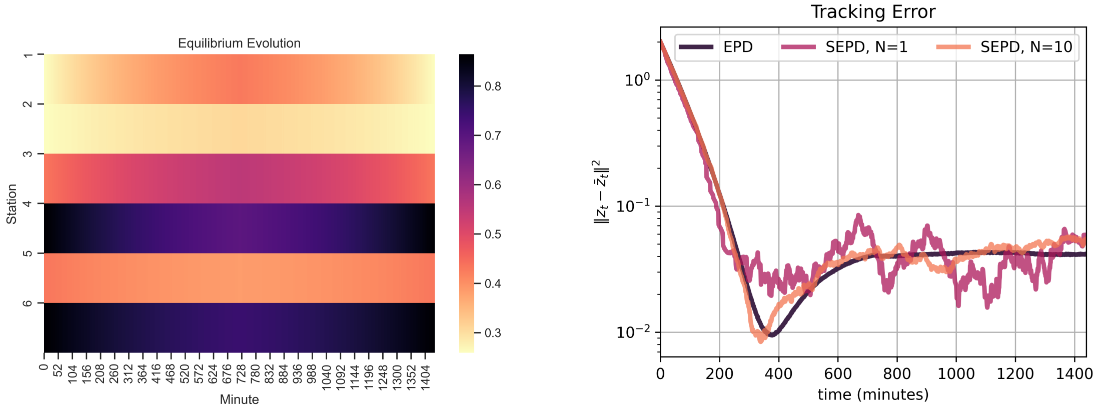
In this section, we provide a demonstration of an online electric vehicle charging market, as described in (2), with time series demand data from Gilleran et al. (2021). The data describes a years of worth of electricity demand with entries for each minute of the year. Each file represents a different type of charging station positioned near commercial uses with varying number of ports (2 or 6), frequency of use (2, 8, or 16 event), and port power output (50, 150, or 350 kW). We randomly allocate each provider with three 8-event stations and draw samples from each day of the year. The demand data is normalized by first subtracting the mean across each minute and dividing by the variance. A representative example of the raw 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 $h_t(p) = (-c(p)/m|t - m| + c(p))$ where p



(a)

(b)

Figure 1: 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.



(a)

(b)

Figure 2: 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.

is the station's port power and $c(p)$ is given by $c(p) = 0.3$ for $p \in \{50, 150\}$ and $c(p) = 0.5$ for $p = 350$. 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 $A_2^t = -A_1^t$. For the sake of simplicity, we consider service providers with unit charging speed utility rates and zero location-based utility. From this we conclude that for all t, G_t is 1-strongly monotone and 1-Lipschitz. Hence our results apply provided that $\eta < 1/3$.

We compute the equilibrium points by executing a batch primal-dual algorithm for 2000 iterations with a step size of $\eta = 0.01$. We then run the online primal-dual and stochastic primal-dual algorithms over each minute of the time series data and plot the distance to the solutions in Figure 2. We observe that the primal-dual algorithm is capable of reasonably tracking the trajectory. The noise incurred by the stochastic algorithm clearly prevents it from having identical performance, however the trajectory does decrease to an acceptable level after overcoming transient behavior for approximately 200 time steps.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) through the award CAREER 1941896.

References

Charalambos D Aliprantis and Kim Border. *Infinite dimensional analysis: A Hitchhiker's Guide*. Springer, 2006.

Julian Berberich, Johannes Köhler, Matthias A Müller, and Frank Allgöwer. Data-driven model predictive control with stability and robustness guarantees. *IEEE Transactions on Automatic Control*, 66(4):1702–1717, 2020.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

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.

John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

Xuanyu Cao, Junshan Zhang, and H Vincent Poor. Online stochastic optimization with time-varying distributions. *IEEE Transactions on Automatic Control*, 66(4):1840–1847, 2020.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 97, pages 5827–5837, 2019.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

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.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, volume 139, pages 3964–3975, 2021.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

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.

Jayash Koshal, Angelia Nedić, and Uday V Shanbhag. Multiuser optimization: Distributed algorithms and error analysis. *SIAM Journal on Optimization*, 21(3):1046–1081, 2011.

Dan Li, Dariush Fooladivanda, and Sonia Martinez. Online optimization and learning in uncertain dynamical environments with performance guarantees. *arXiv preprint arXiv:2102.09111*, 2021.

Liam Madden, Stephen Becker, and Emiliano Dall'Anese. Bounds for the tracking error of first-order online optimization methods. *Journal of Optimization Theory and Applications*, 189(2): 437–457, 2021.

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.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *IEEE Conference on Decision and Control*, pages 7195–7201, 2016.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

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.

Arkadi 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 Journal on Optimization*, 15(1):229–251, 2004.

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.

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

A Yu Popkov. Gradient methods for nonstationary unconstrained optimization problems. *Automation and Remote Control*, 66(6):883–891, 2005.

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

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Daniel D. Selvaratnam, Iman Shames, Jonathan H. Manton, and Mohammad Zamani. Numerical optimisation of time-varying strongly convex functions subject to time-varying constraints. In *IEEE Conference on Decision and Control*, pages 849–854, 2018.

Iman Shames and Farhad Farokhi. Online stochastic convex optimization: Wasserstein distance variation. *arXiv preprint arXiv:2006.01397*, 2020.

Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.

Berkay Turan and Mahnoosh Alizadeh. Competition in electric autonomous mobility on demand systems. *IEEE Transactions on Control of Network Systems*, 2021.

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.

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.

Killian Wood and Emiliano Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *arXiv preprint arXiv:2201.02313*, 2022.

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.

Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR, 2021a.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points. *arXiv preprint arXiv:2111.12743*, 2021b.