Online Projected Gradient Descent for Stochastic Optimization With Decision-Dependent Distributions

Killian Wood, Gianluca Bianchin, *Member, IEEE*, and Emiliano Dall'Anese, *Member, IEEE*

***Abstract*—This letter investigates the problem of tracking solutions of stochastic optimization problems with time-varying costs that depend on random variables with decision-dependent distributions. In this context, we propose the use of an online stochastic gradient descent method to solve the optimization, and we provide explicit bounds in expectation and in high probability for the distance between the optimizers and the points generated by the algorithm. In particular, we show that when the gradient error due to sampling is modeled as a sub-Weibull random variable, then the tracking error is ultimately bounded in expectation and in high probability. The theoretical findings are validated via numerical simulations in the context of charging optimization of a fleet of electric vehicles.**

***Index Terms*—Optimization, optimization algorithms.**

I. Introduction

THIS letter considers the problem of developing and analyzing online algorithms to track the solutions of time-varying stochastic optimization problems, where the distribution of the underlying random variables is decision-dependent. Formally, we consider problems of the form[1]:

$$x_t^* \in \arg\min_{x \in C_t} \mathbb{E}_{z \sim D_t(x)} [\ell_t(x, z)], \tag{1}$$

where $t \in \mathbb{N}_0$ is a time index, $x \in \mathbb{R}^d$ is the decision variable, D_t is a map from the set $\mathbb{R}^d$ to the space of distributions, $z \in \mathcal{Z}_t$ is a random variable (with $\mathcal{Z}_t$ the union of the support of $D_t(x)$ for all $x \in C_t$), ℓ_t:$\mathbb{R}^d \times \mathcal{Z}_t \to \mathbb{R}$ is the loss function, and $C_t \subseteq \mathbb{R}^d$ is a closed and convex set. Problems of this form arise in sequential learning and strategic classification [1], and in applications in power and energy systems [2], [3] to model uncertainty in pricing and human behavior. Moreover, the framework (1) can be used to solve control problems for dynamical systems whose dynamics are unknown, where the variable z is used to account for the lack of knowledge of the underlying system dynamics (similarly to problems in feedback-based optimization [4], [5]).

Since the distribution of z in (1) depends on the decision variable x, the problem of finding x_t^* is computationally burdensome for general cases, and intractable when D_t in unknown – even when the loss function is convex in x [6], [7]. For this reason, we focus on finding decisions that are optimal with respect to the distribution that they induce; we refer to these points as *performatively stable* [6], while we refer to solutions x_t^* to the original problem (1) as *performatively optimal*. We obtain explicit error bounds between performatively optimal and performatively stable points by leveraging tools from [6], [7]. The main focus of this letter is to propose and analyze online algorithms that can determine performatively stable points, in contexts where the loss function and constraint set are revealed sequentially. Since the distributional map D_t may be unknown in practice, we then extend our techniques to stochastic methods that only require samples of z.

Prior Work: Online (projected) gradient descent methods have been well-investigated by using tools from the controls community, we refer to the representative works [8]–[12] as well as to pertinent references therein. Convergence guarantees for online stochastic gradient methods where drift and noise terms satisfy sub-Gaussian assumptions were recently provided in [13]. Online stochastic optimization problems with time-varying distributions are studied in, e.g., [1], [14], [15]. On the other hand, time-varying costs are considered in [16], along with sampling strategies to satisfy regret guarantees. For static optimization problems, the notion of performatively stable points is introduced in [6], where error bounds for

Manuscript received July 20, 2021; revised September 20, 2021; accepted October 14, 2021. Date of publication October 29, 2021; date of current version December 6, 2021. This work was supported in part by the National Science Foundation under Award 1941896, and in part by the National Renewable Energy Laboratory through the subcontract under Grant UGA-0-41026-148. Recommended by Senior Editor F. Dabbene. *(Corresponding author: Killian Wood.)*

Killian Wood is with the Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO 80027 USA (e-mail: killian.wood@colorado.edu).

Gianluca Bianchin and Emiliano Dall'Anese are with the Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder, Boulder, CO 80033 USA.

Digital Object Identifier 10.1109/LCSYS.2021.3124187

[1]*Notation.* We let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, where $\mathbb{N}$ denotes the set of natural numbers. For a given column vector $x \in \mathbb{R}^n$, $\|x\|$ is the Euclidean 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). Given a closed convex set $C \subseteq \mathbb{R}^n$, $\text{proj}_C : \mathbb{R}^n \to \mathbb{R}^n$ denotes the Euclidean projection of y onto C, namely $\text{proj}_C(y) := \arg\min_{v \in C} \|y - v\|$. For a given random variable $z \in \mathbb{R}$, $\mathbb{E}[z]$ denotes the expected value of z, and $\mathbb{P}(z \leq \epsilon)$ denotes the probability of z taking values smaller than or equal to ϵ; $\|z\|_p := \mathbb{E}[|z|^p]^{1/p}$, for any $p \geq 1$. Finally, e denotes Euler's number.

risk minimization and gradient descent methods applied to stochastic problems with decision-dependent distributions are provided. Stochastic gradient methods to identify performatively stable points for decision-dependent distributions are studied in [7], [17] – the latter also providing results for an online setting in expectation. A stochastic gradient method for time-invariant distributional maps is presented in [18].

Contributions: We offer the following main contributions. C1) First, we propose an online projected gradient descent (OPGD) method to solve (1), and we show that the tracking error (relative to the performatively stable points) is ultimately bounded by terms that account for the temporal drift of the optimizers. C2) Second, we propose an online stochastic projected gradient descent (OSPGD) and we provide error bounds in *expectation* and in *high probability*. Our bounds in high probability are derived by modeling the gradient error as a sub-Weibull random variable [19]: this allows us to capture a variety of sub-cases, including scenarios where the error follows sub-Gaussian and sub-exponential distributions [20], or any distribution with finite support.

Relative to [1], [14]–[16] our distributions are decision dependent; relative to [6], [7], [17], [18], our cost and distributional maps are time varying. Moreover, our results do not rely on bias or variance assumptions regarding the gradient estimator. In the absence of distributional shift and without a sub-Weibull error, our upper bounds reduce to the results of [9]. Relative to [18], we seek performatively stable points rather than the performative optima. In doing so, we incur the error characterized in [6]; however, we do not restrict to distributional maps that induce continuous distributions or require finite difference approximations. With respect to the available literature on stochastic optimization, we provide for the first time explicit bounds in expectation and in high probability to solve stochastic optimization with decision dependent distributions in the presence of time-dependent distributional maps.

The remainder of this letter is organized as follows. Section II introduces some preliminaries; Section III studies the OPGD, and Section IV studies the OSPGD. Section V illustrates simulation results, and Section VI concludes this letter.

II. Preliminaries

We first introduce preliminary definitions and results. We consider random variables z that take values on a metric space (M, d), where the set M is equipped with the Borel σ-algebra induced by metric d. We assume that M is a complete and separable metric space (hence M is a Polish space). We let $\mathcal{P}(M)$ denote the set of Radon probability measures on M with finite first moment. Given $\nu \in \mathcal{P}(M)$, $z \sim \nu$ denotes that the random variable z is distributed according to ν. Due to Kantorovich-Rubenstein duality, the Wasserstein-1 distance between $\mu, \nu \in \mathcal{P}(M)$ can be defined as [21]:

$$W_1(\mu, \nu) = \sup_{g \in \mathrm{Lip}_1} \left\{ \mathop{\mathbb{E}}_{z\sim\mu}\big[g(z)\big] - \mathop{\mathbb{E}}_{z\sim\nu}\big[g(z)\big] \right\}, \tag{2}$$

where Lip_1 is the set of 1-Lipschitz functions over M. We note that the pair $(\mathcal{P}(M), W_1)$ describes a metric space of probability measures.

Heavy-Tailed Distributions: In this letter, we will utilize the sub-Weibull model [19], introduced next.

Definition 1 (Sub-Weibull Random Variable): z is a sub-Weibull random variable, denoted by $z \sim \mathrm{subW}(\theta, \nu)$, if there exists $\theta, \nu > 0$ such that $\|z\|_k \leq \nu k^\theta$ for all $k \geq 1$.

The parameter θ measures the heaviness of the tail (higher values correspond to heavier tails) and the parameter ν measures the proxy-variance [19]. In what follows, we will also use the following equivalent characterization of a sub-Weibull random variable: $z \sim \mathrm{subW}(\theta, \nu)$ if and only if $\exists\ \theta, \nu' > 0, \mathbb{P}(|z| \geq \epsilon) \leq 2\exp(-(\epsilon/\nu')^{1/\theta}$. As shown in [22], the two characterizations are equivalent by choosing $\nu = (\frac{\theta}{2e})^\theta \nu'$. The class of sub-Weibull random variables enjoys the following properties.

Proposition 1 (Closure of Sub-Weibull): Let $z \sim \mathrm{subW}(\theta_1, \nu_1)$ and $y \sim \mathrm{subW}(\theta_2, \nu_2)$ be (possibly coupled) sub-Weibull random variables and let $c \in \mathbb{R}$. Then, the following holds:

1) $z + y \sim \mathrm{subW}(\max\{\theta_1, \theta_2\}, \nu_1 + \nu_2)$;
2) $zy \sim \mathrm{subW}(\theta_1 + \theta_2, \psi(\theta_1, \theta_2)\nu_1\nu_2)$, $\psi(\theta_1, \theta_2) \coloneqq (\theta_1 + \theta_2)^{\theta_1+\theta_2}/(\theta_1^{\theta_1}\theta_2^{\theta_2})$;
3) $z + c \sim \mathrm{subW}(\theta_1, |c| + \nu_1)$;
4) $cz \sim \mathrm{subW}(\theta_1, |c|\nu_1)$.

Proof: Properties 1) and 4) are proved in [19]; property 2) is proved in [23]. To show 3), since $c \in \mathbb{R}$, then for any $k \geq 1$ $\|c\|_k = |c| \leq |c|k^\theta$. It follows that $\|z + c\|_k \leq \|z\|_k + \|c\|_k \leq \nu k^\theta + |c|k^\theta \leq (\nu + |c|)k^\theta$. ∎

III. Online Projected Gradient Descent

In this section, we propose and study an OPGD method to solve (1). In Section IV, we will leverage the results derived in this section to analyze a the stochastic version OSPGD.

We begin by outlining our main assumptions.

Assumption 1 (Strong Convexity): For a fixed $z \in \mathcal{Z}_t$, the map $x \mapsto \ell_t(x, z)$ is α_t-strongly convex, where $\alpha_t > 0$, for all $t \in \mathbb{N}_0$.

Assumption 2 (Joint Smoothness): For all $t \in \mathbb{N}_0$, $x \mapsto \nabla_x \ell_t(x, z)$ is β_t-Lipschitz continuous for all $z \in \mathcal{Z}_t$, and $z \mapsto \nabla_x \ell_t(x, z)$ is β_t-Lipschitz continuous for all $x \in \mathbb{R}^d$.

Assumption 3 (Distributional Sensitivity): For all $t \in \mathbb{N}_0$, there exists $\varepsilon_t > 0$ such that

$$W_1(D_t(x), D_t(x')) \leq \varepsilon_t \|x - x'\|_2 \tag{3}$$

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

Assumption 4 (Convex Constraint Set): For all $t \in \mathbb{N}_0$, the set C_t is closed and convex.

A. Performatively Stable Points

Since the objective function and the distribution in (1) both depend on the decision variable x, the problem (1) is intractable in general, even when the loss is convex. For this reason, we follow the approach of [6], [7] and seek optimization algorithms that can determine the performatively stable point, defined as follows:

$$\bar{x}_t \in \arg\min_{x \in C_t} \mathop{\mathbb{E}}_{z\sim D_t(\bar{x}_t)} [\ell_t(x, z)]. \tag{4}$$

Convergence to a performatively stable point is desirable because it guarantees that $\bar{x}_t$ is optimal for the distribution that it induces on z. The following result, adapted from [7, Prop. 3.3], establishes existence and uniqueness of a performatively stable point.

Lemma 1 (Existence of Performatively Stable Points) [7, Prop. 3.3]: Let Assumptions 1-4 hold, and suppose that $\frac{\varepsilon_t \beta_t}{\alpha_t} < 1$ for all $t \in \mathbb{N}_0$. Then, a sequence of performatively stable points $\{\bar{x}_t\}_{t\in\mathbb{N}_0}$ exists and is unique.

In general, performatively stable points may not coincide with the optimizers of the original problem (1). However, an explicit error bound can be derived, as formally stated next.

Lemma 2 (Error of Performatively Stable Points) [6]: Suppose that the function $z \mapsto \ell_t(x, z)$ is γ_t-Lipschitz continuous for all $x \in \mathbb{R}^d$ and $t \in \mathbb{N}_0$. Then, under the same assumptions of Lemma 1, it holds that

$$\|\bar{x}_t - x_t^*\| \leq 2\varepsilon_t \gamma_t \alpha_t^{-1}, \text{ for all } \in \mathbb{N}_0. \tag{5}$$

The proof Lemma 2 follows from [6, Ths. 3.5 and 4.3]. In the remainder of this letter, we assume that the assumptions of Lemma 1 are satisfied, so that the performatively stable point sequence is unique. We illustrate the difference between $\bar{x}_t$ and x_t^* in the following example.

Example 1: Consider an instance of (1) where $\ell(x, z) = x^2 + z$, $C_t = \mathbb{R}$, $D_t(x) = \mathcal{N}(\mu_t x, \sigma_t^2)$, $\mu_t, \sigma_t > 0$. In this case, the objective can be specified in closed form as: $\mathbb{E}_{z\sim D_t(x)}[x^2 + z] = x^2 + \mu_t x$, and thus the unique performatively optimal point is given by $x_t^* = -\mu_t/2$. To determine the performatively stable point, notice that $\nabla_x \ell(x, z) = 2x$, and thus $\bar{x}_t$ satisfies $\mathbb{E}_{z\sim D_t(\bar{x}_t)}[2\bar{x}_t] = 0$, which implies $\bar{x}_t = 0$. The bound in (5) thus holds by noting that $\varepsilon_t = \mu_t$, $\gamma_t = 1$, and $\alpha_t = 2$.

B. Online Projected Gradient Descent

We now propose an OPGD that seeks to track the trajectory of the performatively stable optimizer $\{\bar{x}_t\}_{t\in\mathbb{N}_0}$. To this end, in what follows we adopt the following notation:

$$f_t(x, \nu) \coloneqq \mathbb{E}_{z\sim\nu}[\ell_t(x, z)], \tag{6}$$

for any $x \in \mathbb{R}^d$, $\nu \in \mathcal{P}(M)$, and $t \in \mathbb{N}_0$. Notice that when ν is a distribution induced by the decision variable y, namely $\nu = D_t(y)$, we will use the notation $f_t(x, D_t(y))$. Moreover, we denote by $\nabla f_t(x, \nu)$ the gradient of $f_t(x, \nu)$ (we also note that, according to the dominated convergence theorem, the expectation and gradient operators can be interchanged).

The OPGD amounts to the following step at each $t \in \mathbb{N}_0$:

$$x_{t+1} = G_t(x_t, D_t(x_t)), \tag{7}$$

where $G_t(x_t, \nu) \coloneqq \text{proj}_{C_t}(x_t - \eta_t \nabla f_t(x_t, \nu))$, with $\eta_t > 0$ denoting a stepsize.

First, we note that a performatively stable point is a fixed point of the algorithmic map (7), namely, $\bar{x}_t = G_t(\bar{x}_t, D_t(\bar{x}_t))$. Next, we focus on characterizing the error between the updates (7) and the performatively stable points $\{\bar{x}_t\}_{t\in\mathbb{N}_0}$. To this aim, we denote the temporal drift in the performatively stable points as $\varphi_t \coloneqq \|\bar{x}_{t+1} - \bar{x}_t\|$, and the tracking error relative to the performatively stable points as $e_t \coloneqq \|x_t - \bar{x}_t\|$. Our error bound for OPGD is presented next.

Theorem 1 (Tracking Error of OPGD): Let Assumptions 1-4 hold, suppose that $\frac{\varepsilon_t \beta_t}{\alpha_t} < 1$ for all $t \in \mathbb{N}_0$, and let $\{x_t\}_{t\in\mathbb{N}_0}$ denote a sequence generated by (7). Then, for all $t \in \mathbb{N}_0$, the error $e_t = \|x_t - \bar{x}_t\|$ satisfies:

$$e_{t+1} \leq a_t e_0 + \sum_{i=0}^{t} b_i \varphi_i, \tag{8}$$

where $a_t \coloneqq \prod_{i=1}^{t} \rho_i + \eta_i \beta_i \varepsilon_i$,

$$b_i \coloneqq \begin{cases} 1 & \text{if } i = t, \\ \prod_{k=i+1}^{t} \rho_k + \eta_k \beta_k \varepsilon_k & \text{if } i \neq t, \end{cases}$$

and $\rho_t \coloneqq \max\{|1 - \eta_t \alpha_t|, |1 - \eta_t \beta_t|\}$. Moreover, if

$$\eta_t \in \left[\frac{1 - r}{\alpha_t + \beta_t \varepsilon_t}, \frac{1 + r}{\beta_t(1 + \varepsilon_t)}\right] \text{ for all } t \in \mathbb{N}_0, \tag{9}$$

for some $r \in (0, 1)$, then $\tilde{\lambda} \coloneqq \sup_{t\geq 0}\{\rho_t + \eta_t \beta_t \varepsilon_t\} \leq r$ and

$$\limsup_{t\to+\infty} e_t \leq (1 - \tilde{\lambda})^{-1} \sup_{t\geq 0}\{\varphi_t\}, \tag{10}$$

where $\tilde{\varphi} \coloneqq \sup_{t\geq 0}\{\varphi_t\}$.

Before presenting the proof, some remarks are in order.

Remark 1: By application of Lemma 2, OPGD guarantees that the error between the algorithmic updates and the performatively optimal points is bounded at all times. Precisely, the following estimate holds: $\limsup_{t\to+\infty} \|x_t - x_t^*\| \leq (1 - \tilde{\lambda})^{-1}\tilde{\varphi} + 2\sup_{t\geq 0}\{\varepsilon_t \gamma_t \alpha_t^{-1}\}$.

Remark 2: When (9) holds, one can write the bound $e_{t+1} \leq a_t e_0 + (1 - \tilde{\lambda})^{-1} \sup_i\{\varphi_i\}$; this is an exponential input-to-state-stability (E-ISS) result [24], where $\{\bar{x}_t\}$ are the equilibria of (7) and φ_i is treated as a disturbance. ISS implies that e_t is ultimately bounded as in (10).

Next, we present the proof of Theorem 1. The following lemmas are instrumental.

Lemma 3 (Gradient Deviations): Under Assumption 2, for any $t \in \mathbb{N}_0$, $x \in \mathbb{R}^d$, and measures $\mu, \nu \in \mathcal{P}(M)$, the following bound holds:

$$\|\nabla f_t(x, \mu) - \nabla f_t(x, \nu)\| \leq \beta_t W_1(\mu, \nu). \tag{11}$$

Lemma 4 (Contractive Map): Let Assumptions 1-2 and 4 hold. For any $\nu \in \mathcal{P}(M)$, the map $x \mapsto G_t(x, \nu)$ is Lipschitz continuous, namely, for any $x, y \in \mathbb{R}^d$:

$$\|G_t(x, \nu) - G_t(y, \nu)\| \leq \rho_t \|x - y\|, \tag{12}$$

where $\rho_t = \max\{|1 - \eta_t \alpha_t|, |1 - \eta_t \beta_t|\}$. Moreover, if $\rho_t < 1$ for all $t \in \mathbb{N}_0$, then $\bar{x}_t$ is the unique fixed point of (7).

The proof of Lemma 3 follows by iterating the reasoning in [7, Lemma 2.1] for all $t \in \mathbb{N}_0$; the proof of lemma 4 is standard and is omitted due to space limitations.

Proof of Theorem 1: Note that $x_t \in C_t$ for all $t \in \mathbb{N}_0$ directly follows by definition of Euclidean projection. By using the triangle inequality, we find that

$$\begin{aligned} e_{t+1} &\leq \|x_{t+1} - \bar{x}_t\| + \|\bar{x}_t - \bar{x}_{t+1}\| \\ &= \|G_t(x_t, D_t(x_t)) - G_t(\bar{x}_t, D_t(\bar{x}_t))\| + \varphi_t \end{aligned}$$

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

where the first identity follows from the definition of $G_t(\cdot,\cdot)$ and the second inequality follows by adding and subtracting $G_t(x_t, D_t(\bar{x}_t))$. Applying (11) and Lemma 4 yields:

$$\begin{aligned} e_{t+1} &\leq \eta_t \|\nabla f_t(x_t, x_t) - \nabla f_t(x_t, \bar{x}_t)\| \\ &\quad + \|G_t(x_t, D_t(\bar{x}_t)) - G_t(\bar{x}_t, D_t(\bar{x}_t))\| + \varphi_t \\ &\leq \eta_t \beta_t W_1(D_t(x_t), D_t(\bar{x}_t)) + \rho_t e_t + \varphi_t \\ &\leq \eta_t \beta_t \varepsilon_t e_t + \rho_t e_t + \varphi_t \\ &= (\rho_t + \eta_t \beta_t \varepsilon_t) e_t + \varphi_t. \end{aligned} \tag{13}$$

Thus we obtain the following by expanding the recursion:

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

where we defined $\lambda_t := \rho_t + \eta_t \beta_t \varepsilon_t$. The bound (8) then follows by definition of the sequences $\{a_t\}$ and $\{b_t\}$.

To prove (10), we show that $\sup_t \lambda_t < 1$ for appropriate η_t. Fix $r \in (0, 1)$. Then, by the definition of ρ_t, $\lambda_t \leq r$ holds if the following two conditions are satisfied simultaneously:

$$|1 - \eta_t \alpha_t| + \eta_t \beta_t \varepsilon_t \leq r, \text{ and } |1 - \eta_t \beta_t| + \eta_t \beta_t \varepsilon_t \leq r. \tag{14}$$

The first inequality holds if and only if $-r + \eta_t \beta_t \varepsilon_t < 1 - \eta_t \alpha_t < r - \eta_t \beta_t \varepsilon_t$ or, equivalently, $1 - r \leq \eta_t(\alpha_t + \beta_t \varepsilon_t) \leq 1 + r$. The second inequality holds if and only if $-r + \eta_t \beta_t \varepsilon_t < 1 - \eta_t \beta_t < r - \eta_t \beta_t \varepsilon_t$. By using $\alpha_t \leq \beta_t$, both inequalities are satisfied when

$$\frac{1-r}{\beta_t(1+\varepsilon_t)} \leq \frac{1-r}{\alpha_t + \beta_t \varepsilon_t} \leq \eta_t \leq \frac{1+r}{\beta_t(1+\varepsilon_t)} \leq \frac{1+r}{\alpha_t + \beta_t \varepsilon_t}.$$

Thus, to satisfy the maximum, its sufficient to enforce that $\eta_t \in [\frac{1-r}{\alpha_t + \beta_t \varepsilon_t}, \frac{1+r}{\beta_t(1+\varepsilon_t)}]$. The result (10) follows by utilizing the geometric series. ∎

Finally, we observe that when the objective and constraints are time-invariant, we recover the result of [7, Sec. 5] as formalized next.

Corollary 1 (Tracking Error of OPGD for Time-Invariant Problems): If the problem (1) is time independent and the assumptions in Theorem 1 hold, then OPGD with fixed step size $\eta \in (0, 2/\beta(1+\varepsilon))$ converges linearly to the performatively stable point.

Proof: When (1) is time independent, then for all $t \in \mathbb{N}_0$, $\alpha_t = \alpha$, $\beta_t = \beta$, $\varepsilon_t = \varepsilon$, $\rho_t = \rho$, $\varphi_t = 0$. Accordingly, the recursion (8) yields: $e_{t+1} \leq \lambda e_t$ with $\lambda = \rho + \eta\beta\varepsilon$. By replacing strict inequality and $r = 1$ in (14), we conclude that $\eta < 2/\beta(1+\varepsilon)$ implies $\lambda < 1$. Hence $e_{t+1}/e_t \leq \lambda < 1$. ∎

IV. Online Stochastic Gradient Descent

An exact expression for the distributional map $x_t \mapsto D_t(x_t)$ may not be available in general and, even if available, evaluating the gradient may be computational burdensome. We consider the case where we have access to a finite number of samples of z_t at each time step t to estimate the gradient $\nabla f_t(x_t, D_t(x_t))$. For example, given a mini-batch of samples $\{\hat{z}_t^i\}_{i=1}^{N_t}$ of z_t, the approximate gradient is computed as $g_t(x_t) = (1/N_t)\sum_{i=1}^{N_t} \nabla \ell_t(x_t, \hat{z}_i)$; when $N_t = 1$ we have a "greedy" estimate and when $N_t > 1$ we have a "lazy" estimate [17]. Accordingly, we consider an OSPGD described by:

$$x_{t+1} = \hat{G}_t(x_t), \;\; \hat{G}_t(x) := \text{proj}_{C_t}(x - \eta_t g_t(x)), \tag{15}$$

where $\eta_t > 0$ is a stepsize. In the remainder, we focus on finding error bounds in the spirit of Theorem 1 for OSPGD.

A. Bounds in Expectation and High-Probability

Throughout our analysis, we interpret OSPGD as an inexact OPGD with gradient error given by the random variable:

$$\xi_t := \|g_t(x_t) - \nabla f_t(x_t, D_t(x_t))\|. \tag{16}$$

We make the following assumption.

Assumption 5 (Sub-Weibull Error): The gradient error ξ_t is sub-Weibull; i.e., $\xi_t \sim \text{subW}(\theta, \nu_t)$ for some $\theta, \nu_t > 0$.

Assumption 5 allows us to describe a variety of sub-cases, including scenarios where the error follows sub-Gaussian and sub-Exponential distributions [20], or any distribution with finite support. Further, notice that Assumption 5 does not require the random variables $\{\xi_t\}_{t\in\mathbb{N}_0}$ to be independent. Examples of random variables that satisfy Assumption 5 are described in Section IV-B. Our error bounds for OSPGD are presented next.

Theorem 2 (Expectation and High-Probability Bounds for OSPGD): Let Assumptions 1-4 hold, and suppose that $\frac{\varepsilon_t \beta_t}{\alpha_t} < 1$ for all $t \in \mathbb{N}_0$. Recall that $e_t = \|x_t - \bar{x}_t\|$. Then, the following estimates hold for (15):

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

$$\mathbb{E}\big[e_{t+1}\big] \leq a_t e_0 + \sum_{i=1}^{t} b_i(\varphi_i + \eta_i \,\mathbb{E}\,[\xi_i]). \tag{17}$$

2) If, additionally, Assumption 5 holds and $\delta \in (0, 1)$, then with probability $1 - \delta$:

$$e_{t+1} \leq \left(\frac{2e}{\theta}\right)^{\theta} \log^{\theta}\left(\frac{2}{\delta}\right)\left(a_t e_0 + \sum_{i=1}^{t} b_i(\varphi_i + \eta_i \nu_i)\right), \tag{18}$$

where $\{a_t\}$ and $\{b_i\}$ are as in Theorem 1.

Proof: Note that $x_t \in C_t$ for all $t \in \mathbb{N}$ directly follows by definition of Euclidean projection. To show (17), we first find a stochastic recursion. By the triangle inequality:

$$\begin{aligned} e_{t+1} &\leq \|\hat{G}_t(x_t) - G_t(\bar{x}_t, D_t(\bar{x}_t))\| + \varphi_t \\ &\leq \|\hat{G}_t(x_t) - G_t(x_t, D_t(x_t))\| \\ &\quad + \|G_t(x_t, D_t(x_t)) - G_t(\bar{x}_t, D_t(\bar{x}_t))\| + \varphi_t, \end{aligned}$$

where the second inequality follows by adding and subtracting $G_t(x_t, D_t(x_t))$. By iterating (13), we have $\|G_t(x_t, D_t(x_t)) - G_t(\bar{x}_t, D_t(\bar{x}_t))\| \leq \lambda_t e_t + \varphi_t$, where $\lambda_t := \rho_t + \eta_t \beta_t \varepsilon_t$, and thus $e_{t+1} \leq \eta_t \|g_t(x_t) - \nabla f_t(x_t, D_t(x_t))\| + \lambda_t e_t + \varphi_t$. This yields the stochastic recursion $e_{t+1} \leq \lambda_t e_t + \varphi_t + \eta_t \xi_t$. Expanding the recursion yields

$$e_{t+1} \leq \left(\prod_{i=0}^{t} \lambda_i\right) e_0 + \varphi_t + \sum_{i=0}^{t-1} \left(\prod_{k=i+1}^{t} \lambda_k\right)(\varphi_i + \eta_i \xi_i),$$

or, equivalently,

$$e_{t+1} \leq a_t e_0 + \sum_{i=0}^{t} b_i(\varphi_i + \eta_i \xi_i). \tag{19}$$

Thus, (17) follows by taking the expectation on both sides.

To prove (18), we demonstrate that the righthand side of (19) is sub-Weibull distributed. Since $\xi_i \sim \text{subW}(\theta, \nu_i)$, Proposition 1 implies that $b_i(\varphi_i + \eta_i \xi_i) \sim \text{subW}(\theta, b_i(\varphi_i + \eta_i \nu_i))$. By summing over i, we obtain:

$$\sum_{i=0}^{t} b_i(\varphi_i + \eta_i \xi_i) \sim \text{subW}\left(\theta, \sum_{i=0}^{t} b_i(\varphi_i + \eta_i \nu_i)\right).$$

Thus, by letting $\omega_t \coloneqq a_t e_0 + \sum_{i=0}^{t} b_i(\varphi_i + \eta_i \xi_i)$, we conclude that $\omega_t \sim \text{subW}(\theta, \upsilon_t)$, where $\upsilon_t = a_t e_0 + \sum_{i=0}^{t} b_i(\varphi_i + \eta_i \nu_i)$. From Definition 1 we have

$$\mathbb{P}(|\omega_t| \geq \epsilon) \leq 2 \exp\left(-\frac{\theta}{2e}\left(\frac{\epsilon}{\upsilon_t}\right)^{\frac{1}{\theta}}\right). \tag{20}$$

Now let $\delta \in (0, 1)$ be fixed and set it equal to the right-hand side of the above inequality. Solving for ϵ yields $\epsilon = \log^{\theta}(\frac{2}{\delta})(\frac{2e}{\theta})^{\theta} \upsilon_t$. Then, we have that $\omega_t \leq (\frac{2e}{\theta})^{\theta} \log^{\theta}(\frac{2}{\delta}) \upsilon_t$, with probability $1-\delta$. Finally, (18) follows by substitution. ∎

The bound (17) generalizes the estimate in Theorem 1 by accounting for the gradient error. It is also worth pointing out that (17) and (18) have a similar structure; indeed, (18) differs only by a logarithmic factor and by the introduction of the tail parameters ν_i (which replaces the expectation term).

Remark 3: An alternative high probability bound can be obtained by using (17) and Markov's inequality. For any $\delta \in (0, 1)$, then Markov's inequality guarantees that:

$$e_{t+1} \leq \frac{1}{\delta}\left(a_t e_0 + \sum_{i=1}^{t} b_i(\varphi_i + \eta_i \mathbb{E}[e_i])\right), \tag{21}$$

with probability at least $1 - \delta$. However, if we increase the confidence of the bound by allowing $\delta \to 0$, the right-hand-side of (21) grows more rapidly than (18).

Note that the bounds in Theorem 2 are valid for any $t \in \mathbb{N}$. The asymptotic behavior is noted in the next remark.

Remark 4: If (9) holds, then $\limsup_{t \to +\infty} e_t \leq (1 - \tilde{\lambda})^{-1}(\tilde{\varphi} + \tilde{\eta}\tilde{\xi})$ *almost surely*, where $\tilde{\eta}$ and $\tilde{\xi}$ are upper bounds on the step size and $\mathbb{E}[\xi_t]$; the proof is omitted because of space limits, but follows arguments similar to [23].

B. Remarks on the Error Model

The class of sub-Weibull distributions allows one to consider variety of error models. For instance, it includes sub-Gaussian and sub-exponential as sub-cases by setting $\theta = 1/2$ and $\theta = 1$, respectively. We notice that a sub-Gaussian assumption was typically utilized in prior works on stochastic gradient descent; for example, the assumption $\mathbb{E}[\exp(\xi^2/\sigma^2)] \leq e$ in [25] corresponds to sub-Gaussian tail behavior. However, recent works suggest that stochastic gradient descent may exhibit errors with tails that are heavier than a sub-Gaussian (see, e.g., [26]). To further elaborate on the flexibility offered by a sub-Weibull model, we provide the following additional examples.

Example 2: Suppose that each entry of the gradient error $g_t(x_t) - \nabla f_t(x_t, x_t)$ follows a distribution $\text{subW}(\theta, \nu)$, $i = 1, \ldots, d$ for given $\theta, \nu > 0$. Then $\|\xi_t\|$ is sub-Weibull with $\|\xi_t\| \sim \text{subW}(\theta, 2^{\theta}\sqrt{d}\nu)$ [23].

Example 3: Suppose that an entry of the gradient error $g_t(x_t) - \nabla f_t(x_t, x_t)$ is Gaussian is zero mean and variance ς^2; then, it sub-Gaussian with sub-Gaussian norm $C\varsigma$, with C an absolute constant [20], and it is therefore a sub-Weibull $\text{subW}(1/2, C'\varsigma)$ with C' an absolute constant.

Example 4: Suppose that ξ_t is a random variable with mean $\mu_t \coloneqq \mathbb{E}[\xi_t]$, such that $\xi_t \in [\bar{\xi}, \underline{\xi}]$ almost surely. Then $\xi_t - \mu_t \sim \text{subW}(1/2, (\underline{\xi} - \bar{\xi})/\sqrt{2})$ [23].

V. Application to Electric Vehicle Charging

This section illustrates the use of the proposed algorithms in an application inspired from [3], where the operator of a fleet of electric vehicles (EVs) seeks to determine an optimal charging policy in order to minimize its charging costs. The region of interest is modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node in $\mathcal{V}$ represents a charging station (or a group thereof), and an edge (i, j) in $\mathcal{E}$ allows vehicles to transfer from node i to j. We assume that the graph is strongly connected, so that EVs can be redirected from one node to any other node. We let $x_i \in \mathbb{R}_{\geq 0}$ denote the energy requested by the fleet at node $i \in \mathcal{V}$. We assume that the net energy available is limited, and define the set $C_t \coloneqq \{x \in \mathbb{R}^d : \sum_{i \in \mathcal{V}} x_i \leq c_t\}$, for a given $c_t \in \mathbb{R}_{>0}$. Given $\{x_i\}$, the operator of the power grid strategically chooses a price per unit of energy so as to optimize its revenue from selling energy; we let $z_i \in \mathbb{R}_{\geq 0}$ denote the selected price in region i, and we hypothesise that $z_i \sim \mathcal{N}(\mu_t x_i, \sigma_t^2)$, $\mu_t, \sigma_t \in \mathbb{R}_{\geq 0}$ as an example. We note that, although the grid operator can choose the price arbitrarily large to maximize its revenue, large prices may compel the fleet operator to withdraw its demand, thus motivating the use of a model where the mean grows linearly with the energy demand. Accordingly, we model the cost function of the EV operator as follows [3]:

$$\ell_t(x, z) = \sum_{i \in \mathcal{V}} z_i x_{i,t} - \gamma_{i,t} x_i + \kappa_{i,t} x_i^2, \tag{22}$$

where $\gamma_{i,t} \in \mathbb{R}_{\geq 0}$, models the charging aggressiveness of the fleet operator, and $\kappa_{i,t} x_{i,t}^2$ quantifies the satisfaction the fleet operator achieves from consuming one unit of energy. In (22), the term $z_{i,t} x_{i,t}$ describes the charging cost at station i, the quantity $\gamma_{i,t} x_{i,t}$, and models the energy demand at the i-th station. Notice that, because the displacement of vehicles can change over time, we assume that the parameters $\gamma_{i,t}$ and $\xi_{i,t}$ are time dependent. We note that: (i) because of the capacity constraint $x_t \in C_t$, the decision variables $x_{i,t}$, $i \in \mathcal{V}$, are coupled, and (ii) although the optimization could be solved in a distributed fashion since (22) is separable, our focus is to solve it in a centralized way since the EV operator is unique.

We apply the proposed methods to a system of 10 homogeneous charging stations over 100 time steps with fixed net energy ($c_t = 10$). Namely, $\gamma_{i,t} = -1/100|t - 50| + 1$ and $\kappa_{i,t} = 2$ for $i \in \{1, \ldots, 10\}$. The charging cost distribution is informed by μ_t and σ_t; in our case, μ_t is the time series data of CAISO real-time prices deposited in Fig. 1 (taken from http://www.energyonline.com) and $\sigma_t = 1$. Given these

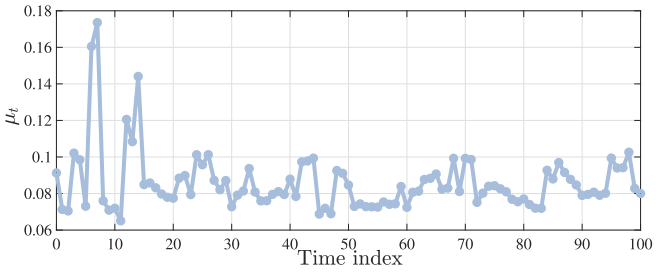

Fig. 1. Time series data representing the price of energy in dollars per kilowatt hour (kWh). Each time step represents 5 minutes.

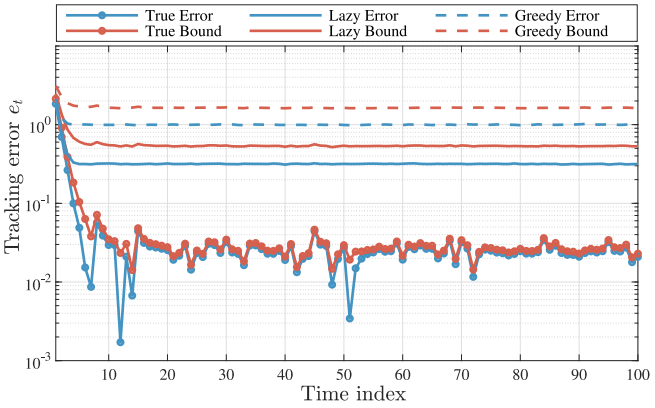

Fig. 2. Performance comparison of OPGD and OSPGD.

parameter values, the cost is α_t-strongly convex and β_t-jointly smooth with $\alpha_t = \beta_t = 2$. Following the results in [27], the distributional maps are ε_t-sensitive with $\varepsilon_t = \mu_t$. The sequence of performatively stable points are computed in closed form by solving the KKT equations.

For each experiment, we run OPGD and OSPGD with fixed step size $\eta_t = 0.3$ by drawing initial state x_0 uniformly from a sphere of radius 5. For OSPGD, we compute the mean tracking error for both greedy and lazy deployments. The mean tracking error for each is computed via Monte Carlo simulation using $1,000$ realizations of the initial state.

In Fig. 2, we illustrate the tracking errors and corresponding upper bounds presented in Theorems 1 and 2. "True" (i.e., true gradient) refers to the OPGD, "greedy" to the OSPGD with $N_t = 1$, and "lazy" to the OSPGD with $N_t = 10$. We notice that the upper bound curve mimics the evolution of the tracking error; yet, in the instance of OSPGD the relationship is looser relative to the OPGD curves.

VI. Conclusion

This letter considered online gradient and stochastic gradient methods for tracking solutions of time-varying stochastic optimization problems with decision-dependent distributions. Under a distributional sensitivity assumption, we derived explicit error bounds for the two methods. In particular, we derived convergence in expectation and in high probability for the OSPGD by assuming that the error in the gradient follows a sub-Weibull distribution. To the best of our knowledge, our convergence results for online gradient methods are the first in the literature for time-varying stochastic optimization problems with decision-dependent distributions.

References

[1] C. Wilson, Y. Bu, and V. V. Veeravalli, "Adaptive sequential machine learning," *Sequential Anal.*, vol. 38, no. 4, pp. 545–568, 2019.
[2] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Examining uncertainty in demand response baseline models and variability in automated responses to dynamic pricing," in *Proc. IEEE Conf. Decis. Control*, 2011, pp. 4332–4339.
[3] W. Tushar, W. Saad, H. V. Poor, and D. B. Smith, "Economics of electric vehicle charging: A game theoretic approach," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 1767–1778, Dec. 2012.
[4] A. Hauswirth, S. Bolognani, G. Hug, and F. Dörfler, "Optimization algorithms as robust feedback controllers," 2021, *arXiv:2103.11329*.
[5] G. Bianchin, M. Vaquero, J. Cortes, and E. Dall'Anese, "Online stochastic optimization for unknown linear systems: Data-driven synthesis and controller analysis," 2021, *arXiv:2108.13040*.
[6] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7599–7609.
[7] D. Drusvyatskiy and L. Xiao, "Stochastic optimization with decision-dependent distributions," 2020, *arXiv:2011.11173*.
[8] A. Y. Popkov, "Gradient methods for nonstationary unconstrained optimization problems," *Autom. Remote Control*, vol. 66, no. 6, pp. 883–891, 2005.
[9] D. D. Selvaratnam, I. Shames, J. H. Manton, and M. Zamani, "Numerical optimisation of time-varying strongly convex functions subject to time-varying constraints," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 849–854.
[10] A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro, "Online optimization in dynamic environments: Improved regret rates for strongly convex problems," in *Proc. IEEE Conf. Decis. Control*, 2016, pp. 7195–7201.
[11] A. Simonetto, "Time-varying convex optimization via time-varying averaged operators," 2017, *arXiv:1704.07338*.
[12] L. Madden, S. Becker, and E. Dall'Anese, "Bounds for the tracking error of first-order online optimization methods," *J. Optim. Theory Appl.*, vol. 189, no. 2, pp. 437–457, 2021.
[13] J. Cutler, D. Drusvyatskiy, and Z. Harchaoui, "Stochastic optimization under time drift: Iterate averaging, step decay, and high probability guarantees," 2021, *arXiv:2108.07356*.
[14] X. Cao, J. Zhang, and H. V. Poor, "Online stochastic optimization with time-varying distributions," *IEEE Trans. Autom. Control*, vol. 66, no. 4, pp. 1840–1847, Apr. 2021.
[15] I. Shames and F. Farokhi, "Online stochastic convex optimization: Wasserstein distance variation," 2020, *arXiv:2006.01397*.
[16] C. Wilson, V. V. Veeravalli, and A. Nedić, "Adaptive sequential stochastic optimization," *IEEE Tran. Autom. Control*, vol. 64, no. 2, pp. 496–509, Feb. 2019.
[17] C. Mendler-Dünner, J. C. Perdomo, T. Zrnic, and M. Hardt, "Stochastic optimization for performative prediction," 2020, *arXiv:2006.06887*.
[18] Z. Izzo, L. Ying, and J. Zou, "How to learn when data reacts to your model: Performative gradient descent," 2021, *arXiv:2102.07698*.
[19] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel, "Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions," *Stat*, vol. 9, no. 1, p. e318, 2020.
[20] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge, U.K.: Cambridge Univ. Press, 2018.
[21] L. V. Kantorovich and S. Rubinshtein, "On a space of totally additive functions," *Vestnik Saint Petersburg Univ. Math.*, vol. 13, no. 7, pp. 52–59, 1958.
[22] K. C. Wong, Z. Li, and A. Tewari, "Lasso guarantees for β-mixing heavy-tailed time series," *Ann. Stat.*, vol. 48, no. 2, pp. 1124–1142, 2020.
[23] N. Bastianello, L. Madden, R. Carli, and E. Dall'Anese, "A stochastic operator framework for inexact static and online optimization," 2021, *arXiv:2105.09884*.
[24] Z.-P. Jiang and Y. Wang, "Input-to-state stability for discrete-time nonlinear systems," *Automatica*, vol. 37, no. 6, pp. 857–869, 2001.
[25] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.
[26] L. Hodgkinson and M. W. Mahoney, "Multiplicative noise and heavy tails in stochastic optimization," 2020, *arXiv:2006.06293*.
[27] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions" *Michigan Math. J.*, vol. 31, no. 2, pp. 231–240, 1984.