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

Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese

***Abstract*— This paper investigates the problem of tracking solutions of stochastic optimization problems with time-varying costs and decision-dependent distributions. In this context, the paper focuses on the online stochastic gradient descent method, and establishes its convergence to the sequence of optimizers (within a bounded error) in expectation and in high probability. In particular, high-probability convergence results are derived by modeling the gradient error as a sub-Weibull random variable. The theoretical findings are validated via numerical simulations in the context of charging optimization of a fleet of electric vehicles.**

I. Introduction

This paper 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]

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

where $t \in \mathbb{N}_0$ is the 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 $x \in C_t$), $\ell_t : \mathbb{R}^d \times \mathcal{Z}_t \to \mathbb{R}$ is the loss function, and $C_t \subseteq \mathbb{R}^d$ is a closed and convex set. Problems of this form arise in several areas including sequential learning [1] and strategic classification [2], as well as in applications such as demand response in power systems [3] and electric vehicle charging [4] to model the uncertainty in the users' behavior and time-varying pricing.

Since the objective in (1) is comprised of a cost and distribution that depend on the decision variable x, the problem of finding a direct solution is intractable in general—even when the losses are convex in x [5], [6]. For this reason, we focus our attention on determining the sequence of decision variables that are optimal with respect to the distribution that they induce; we refer to these points as *performatively stable* [5]. On the other hand, the solutions x_t^* to the original problem (1) are referred to as *performatively optimal points*. We seek online algorithms that provably track performatively stable points of (1), based on functions and sets that are revealed sequentially. Tracking of the sequence of optimal solutions x_t^* is then obtained by leveraging the results of [5], [6], where the difference between performatively stable points and x_t^* is bounded under suitable assumptions that will be discussed in the paper.

Prior Work: Asymptotic tracking of online (projected) gradient descent has been extensively studied; we refer the reader to the representative works [7], [8], [9], [10] as well as pertinent references therein. Related to our setting, online stochastic optimization problems with time-varying distributions are studied in, e.g., [1], [11], [12]. On the other hand, time-varying costs are considered in [13] along with sampling strategies to satisfy regret guarantees. For static optimization problems, the notion of performatively stable points is introduced in [5], where error bounds for 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 [6], [14]—the latter also providing results for an online setting in expectation. A stochastic gradient method that converges to performatively optimal points for time-invariant distributional maps is presented in [15].

Contributions: We consider the problem (1), where the loss, the distributional map, and the constraint set may be time varying and the distribution are decision dependent. We offer the following contributions: C1) We first consider an online projected gradient descent (OPGD) method, and show that the tracking error incurred relative to the performatively stable points is bounded by terms involving the drift of the performatively stable points. C2) We then consider an online stochastic projected gradient descent (OSPGD) and derive tracking results in *expectation* and in *high probability*. The high-probability convergence results are derived by modeling the gradient error as a sub-Weibull random variable [16]; this allows us to capture a variety of models as sub-cases, including sub-Gaussian and sub-Exponential errors [17], and errors whose distribution has a finite support.

Our contributions unify the aforementioned themes of online optimization with time-varying distributions and stochastic optimization with decision-dependent distributions. Relative to [1], [11], [12], [13] our distributions are decision dependent; relative to [5], [6], [14], [15], our cost and distri-

K. Wood is with the Dept. of Applied Mathematics, University of Colorado Boulder. G. Bianchin and E. Dall'Anese are with the Dept. of Electrical, Computer, and Energy Eng., University of Colorado Boulder. This work was supported in part by the National Science Foundation ERC ASPIRE and by the National Renewable Energy Laboratory through the contract UGA-0-41026-148.

[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 $C \subseteq \mathbb{R}^n$, $\mathrm{proj}_C : \mathbb{R}^n \to \mathbb{R}^n$ denotes the Euclidean projection of y onto C, namely $\mathrm{proj}_C(y) := \arg\min_{v \in C} \|y - v\|$. 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.

butional 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 [8]. Relative to [15], we seek performatively stable points rather than the performative optima. In doing so, we incur the error characterized in [5]; however, we do not restrict to distributional maps that are continuous distributions or require finite difference approximations. Similarly to [18], we model errors as sub-Weibull random variables, but our distributions are decision dependent. To the best of our knowledge, this is the first work on stochastic optimization with decision dependent distributions that studies time dependent distributional maps and offers high probability bounds via the sub-Weibull error model.

The remainder of this paper is organized as follows. Section II introduces preliminary results; Section III considers the OPGD, while Section IV presents the main convergence results for the OSPGD. Section V provides representative results in the context of electric vehicle charging, and Section VI concludes the paper.

II. Preliminaries

We first introduce preliminary definitions and results that will be used in the paper.

Space of Probability Measures. 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 compact (and hence complete and separable [19, Lemma 1.6.2]). 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. Under our compactness assumptions, the Wasserstein-1 distance between two measures $\mu, \nu \in \mathcal{P}(M)$ can be defined as [20]:

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

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; see [21].

Heavy-Tailed Distributions. In this paper, we will utilize the sub-Weibull model [16], introduced next.

Definition 1: (***Sub-Weibull Random Variable***) The random variable z is sub-Weibull if, for some $\theta > 0$, one of the following conditions is satisfied:

a. $\exists\, \nu_1 > 0$ such that $\mathbb{P}(|z| \geq \epsilon) \leq 2e^{-\left(\frac{\epsilon}{\nu_1}\right)^{\frac{1}{\theta}}}$ for all $\epsilon > 0$.
b. $\exists\, \nu_2 > 0$ such that $\|z\|_k \leq \nu_2 k^\theta$ for all $k \geq 1$.

The above statements 1.a and 1.b are equivalent in that if condition 1.b holds for some $\nu_2 > 0$, then condition 1.a holds with $\nu_1 = \left(\frac{2e}{\theta}\right)^\theta \nu_2$ [22, Lemma 5]. The parameter θ measures the heaviness of the tail (higher values correspond to heavier tails) and the parameter ν_1 (or ν_2) measures the proxy-variance [16]. In the remainder of the paper, we will use the definition 1.b, and we use the notation $z \sim \mathrm{subW}(\theta,\nu)$ if $\|z\|_k \leq \nu k^\theta$. The class of sub-Weibull random variables enjoys the following properties.

Proposition 2.1: (***Closure of Sub-Weibull***) Suppose that $z \sim \mathrm{subW}(\theta_1,\nu_1)$ and $y \sim \mathrm{subW}(\theta_2,\nu_2)$ are possibly dependent sub-Weibull random variables according to Definition 1.b, 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) $zy \sim \mathrm{subW}(\theta_1+\theta_2, \psi(\theta_1,\theta_2)\nu_1\nu_2)$, $\psi(\theta_1,\theta_2) := (\theta_1+\theta_2)^{\theta_1+\theta_2}/(\theta_1^{\theta_1}\theta_2^{\theta_2})$;
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 [16]; property 2) is proved in [18]. 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 lay some groundwork by analyzing the performance of the OPGD method for solving (1). The results of this section will then be leveraged in Section IV to analyze the convergence of the OSPGD. Before proceeding, we first outline relevant 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, with $\alpha_t > 0$, for all $t \in \mathbb{N}_0$.

Assumption 2: (***Joint smoothness***) For all $t \in \mathbb{N}_0$, the function $(x,z) \mapsto \ell_t(x,z)$ is β_t-smooth; namely, the functions $x \mapsto \nabla_x \ell_t(x,z)$ and $z \mapsto \nabla_x \ell_t(x,z)$ are Lipschitz continuous with constant $\beta_t > 0$. □

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***) The constraint set C_t is closed and convex, uniformly in time. □

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 [5], [6] and seek optimization algorithms that converge to a performatively stable point, defined as follows:

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

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, shown in [6, Prop. 3.3], establishes the existence of performatively stable points in our setting (4).

Lemma 3.1: ***(Existence of Performatively Stable Points) [6, 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 the sequence of performatively stable points $\{\bar{x}_t\}_{t\geq 0}$ 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 3.2: ***(Error of Performatively Stable Points [5])*** 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 3.1, it holds that

$$\|\bar{x}_t - x_t^*\| \leq 2\varepsilon_t \gamma_t \alpha_t^{-1}. \tag{5}$$

The proof Lemma 3.2 follows from [5, Thm 3.5, Thm 4.3]. Finally, to better convey ideas, we illustrate the difference between $\bar{x}_t$ and x_t^* with one example.

Example 1: Consider the optimization problem

$$\min_{x \in \mathbb{R}} \quad \mathop{\mathbb{E}}_{z \sim D_t(x)} \left[x^2 + z\right], \tag{6}$$

where $D_t(x) = \mathcal{N}(\mu_t x, \sigma_t^2)$, $\mu_t, \sigma_t > 0$. We can rewrite (6) as $\min_{x \in \mathbb{R}^d} \; x^2 + \mu_t x$, and thus the unique performatively optimal point is $x_t^* = -\mu_t/2$. To determine the performatively stable points, we notice that $\nabla_x \ell(x, z) = 2x$, and thus $\bar{x}$ must satisfy $\mathop{\mathbb{E}}_{z \sim D_t(\bar{x}_t)} [2\bar{x}_t] = 0$, which implies $\bar{x}_t = 0$. □

B. Online projected gradient descent

We now consider an OPGD that seeks to track the trajectory of performatively stable optimizers $\{\bar{x}_t\}_{t \geq 0}$ of problem (4). To this end, we adopt the following notation:

$$f_t(x, \nu) := \mathop{\mathbb{E}}_{z \sim \nu} \left[\ell_t(x, z)\right], \tag{7}$$

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 $t \in \mathbb{N}_0$:

$$x_{t+1} = G_t(x_t, D_t(x_t)), \tag{8}$$

where $G_t(x_t, \nu) := \text{proj}_{C_t}(x_t - \eta_t \nabla f_t(x_t, \nu))$ and $\eta_t > 0$ denotes the stepsize. Performatively stable points are fixed points for the algorithmic map that they induce; i.e., $\bar{x}_t = G_t(\bar{x}_t, D_t(\bar{x}_t))$.

We denote the temporal drift in the performatively stable points as:

$$\varphi_t := \|\bar{x}_{t+1} - \bar{x}_t\|, \tag{9}$$

for all $t \in \mathbb{N}_0$. We also denote the tracking error relative to the sequence $\{\bar{x}_t\}_{t \geq 0}$ as $e_t := \|x_t - \bar{x}_t\|$. The convergence result for the OPGD is then presented next.

Theorem 3.3: ***(Tracking Error of OPGD)*** Let Assumptions 1-4 hold, and suppose that $\frac{\varepsilon_t \beta_t}{\alpha_t} < 1$ for all $t \in \mathbb{N}_0$. Let $\{x_t\}$ be a sequence generated by (8); then, the following holds for the errors $\{e_t = \|x_t - \bar{x}_t\|\}$:

$$e_{t+1} \leq \; a_t e_0 + \sum_{i=0}^{t} b_i \varphi_i, \text{ for all } t \in \mathbb{N}_0, \tag{10}$$

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

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

and $\rho_t := \max\{|1 - \eta_t \alpha_t|, |1 - \eta_t \beta_t|\}$. Moreover, if the sequence of step sizes satisfies

$$\eta_t < \frac{2}{\beta_t(1 + \varepsilon_t)}, \text{ for all } t \in \mathbb{N}_0, \tag{11}$$

then $\rho_t + \eta_t \beta_t \varepsilon_t < 1$ for all $t \in \mathbb{N}_0$ and

$$\limsup_{t \to +\infty} e_t \leq (1 - \tilde{\lambda})^{-1} \tilde{\varphi}, \tag{12}$$

where $\tilde{\varphi} := \sup_{t \geq 0}\{\varphi_t\}$ and $\tilde{\lambda} := \sup_{t \geq 0}\{\rho_t + \eta_t \beta_t \varepsilon_t\}$.

Before presenting the proof, some remarks are in order.

Remark 1: When the distribution of z is *independent* of x, we have $\varepsilon_t = 0$; thus, we recover the classical results for OPGD [7], [8]. Similarly, when the objective and constraints are time-invariant, i.e. $\ell_t = \ell$ and $C_t = C$ for all $t \geq 0$, then we recover the linear convergence result of [6, Section 5] with regularizer given by the indicator over the set C_t. □

Remark 2: By virtue of Lemma 3.2, the OPGM achieves tracking (within an error bound) of the optimizers $\{x_t^*\}_{t \geq 0}$ of (1). In particular, $\limsup_{t \to +\infty} \|x_t - x_t^*\| \leq (1 - \tilde{\lambda})^{-1} \tilde{\varphi} + 2\sup_{t \geq 0}\{\varepsilon_t \gamma_t \alpha_t^{-1}\}$. □

Remark 3: When (11) 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, where $\{\bar{x}_t\}$ are the equilibria of (8) and φ_i is treated as a disturbance. ISS implies that e_t is ultimately bounded as in (12). □

The following lemmas are instrumental for the proof of Theorem 3.3.

Lemma 3.4: (***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:

$$\|\nabla f_t(x, \mu) - \nabla f_t(x, \nu)\| \leq \beta_t W_1(\mu, \nu). \tag{13}$$

Lemma 3.5: (***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 $x, y \in \mathbb{R}^d$:

$$\|G_t(x, \nu) - G_t(y, \nu)\| \leq \rho_t \|x - y\|, \tag{14}$$

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 (8).

The proof of Lemma 3.4 follows by iterating the reasoning in [6, Lemma 2.1] for all $t \in \mathbb{N}_0$; the proof of lemma 3.5 is standard and is omitted due to space limitations.

Proof of Theorem 3.3: By telescoping appropriate terms and applying the triangle inequality, it follows that

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

Now, applying (13) and Lemma 3.5, we get

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

We can now repeatedly apply the recursion to get:

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

where $\lambda_t := \rho_t + \eta_t \beta_t \varepsilon_t$ for brevity. The upper bound follows by defining sequences $\{a_t\}$ and $\{b_t\}$.

For the second part, we note that $\lambda_t < 1$ is equivalent to $|1 - \eta_t \alpha_t| + \eta_t \beta_t \varepsilon_t < 1$ and $|1 - \eta_t \beta_t| + \eta_t \beta_t \varepsilon_t < 1$. The first inequality holds if and only if $-1 + \eta_t \beta_t \varepsilon_t < 1 - \eta_t \alpha_t < 2 - \eta_t \beta_t \varepsilon_t$, so that solving simultaneously yields $\eta_t \beta_t \varepsilon < \eta_t \alpha_t < 2 - \eta_t \beta_t \varepsilon$. The condition is met provided that $\frac{\varepsilon_t \beta_t}{\alpha_t} < 1$ and $\eta < \frac{2}{\alpha + \beta_t \varepsilon_t}$. Similarly, enforcing that $|1 - \eta_t \beta_t| + \eta_t \beta_t \varepsilon_t < 1$ yields $\eta_t \beta_t \varepsilon_t < \eta_t \beta_t < 1 - \eta_t \beta_t \varepsilon_t$. This requires that $\varepsilon_t < 1$ and $\eta_t < \frac{2}{\beta_t(1+\varepsilon_t)}$. Note that since $\alpha_t \leq \beta_t$, then these conditions amount to

$$\varepsilon_t < \frac{\varepsilon_t \beta_t}{\alpha_t} < 1 \quad \text{and} \quad \eta_t < \frac{2}{\beta_t(1+\varepsilon_t)} < \frac{2}{\alpha_t + \beta_t \varepsilon_t},$$

so to satisfy the maximum, its sufficient to enforce that $\frac{\varepsilon_t \beta_t}{\alpha_t} < 1$ and $\eta_t < \frac{2}{\beta_t(1+\varepsilon_t)}$. The result (12) follows by utilizing the geometric series. ■

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 [14].

The OSPGD is then described by:

$$x_{t+1} = \hat{G}_t(x_t), \quad \hat{G}_t(x) := \mathrm{proj}_{C_t}\left(x - \eta_t g_t(x)\right), \qquad (15)$$

where η_t denotes the stepsize. We are now interested in finding convergence results in the spirit of Theorem 3.3 for the OSPGD (15).

A. Convergence in expectation and high-probability

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

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

We make the following assumption.

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

We also note that the random variables $\{\xi_t\}_{t\geq 0}$ may be dependent. Assumption 5 allows us to derive general convergence results, as briefly discussed in Section IV-B. The following result characterizes the convergence in mean and in high-probability of (15).

Theorem 4.1: (***Expected and High-probability Convergence of 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 $t \in \mathbb{N}_0$:

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

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

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

where $\{a_t\}$ and $\{b_i\}$ are as in Theorem 3.3.

Proof: Apply the algorithmic map to obtain

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

From our previous analysis, we find that $\|G_t(x_t, D_t(x_t)) - G_t(\bar{x}_t, D_t(\bar{x}_t))\| \leq \lambda_t e_t + \varphi_t$, where $\lambda_t := \rho_t + \eta_t \beta_t \varepsilon_t$, and hence $e_{t+1} \leq \eta_t \|g_t(x_t) - \nabla f_t(x_t, D_t(x_t))\| + \lambda_t e_t + \varphi_t$. This yields the stochastic recursion $e_{t+1} \leq \lambda_t e_t + \varphi_t + \eta_t \xi_t$. We can now repeatedly apply the recursion to get

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

We conclude that

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

so taking the expected value of both sides yields (17).

To prove the bound in high probability, we demonstrate that the right side of (19) is sub-Weibull distributed. Since $\xi_i \sim \mathrm{subW}(\theta, \nu_i)$, then following from Prop. 2.1 we have that $b_i(\varphi_i + \eta_i \xi_i) \sim \mathrm{subW}\left(\theta, b_i(\varphi_i + \eta_i \nu_i)\right)$ as well. Summing over i, we find that

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

and hence

$$a_t e_0 + \sum_{i=0}^{t} b_i (\varphi_i + \eta_i \xi_i) \sim \mathrm{subW}\left(\theta, \upsilon_t\right).$$

with $\upsilon_t = a_t e_0 + \sum_{i=0}^{t} b_i (\varphi_i + \eta_i \nu_i)$. If we denote this upper bound as a random variable ω_t, and the scale parameter as υ_t, then Definition 1.a implies that

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

Now we let $\delta \in (0,1)$ be fixed and set it equal to the right hand side of the above inequality. Solving for ϵ yields $\epsilon = \log^\theta\left(\frac{2}{\delta}\right)\left(\frac{2e}{\theta}\right)^\theta \upsilon_t$. Then, we have that

$$\omega_t \le \left(\frac{2e}{\theta}\right)^\theta \log^\theta\left(\frac{2}{\delta}\right)\upsilon_t, \tag{21}$$

with probability $1-\delta$. Substituting our definitions yields the desired results. ■

The bound (17) generalizes the estimate in Theorem 3.3 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 4: 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:

$$e_{t+1} \le \frac{1}{\delta}\left(a_t e_0 + \sum_{i=1}^{t} b_i(\varphi_i + \eta_i \mathbb{E}[e_i])\right), \tag{22}$$

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 (22) grows more rapidly than (18). □

Note that the bounds in Theorem 4.1 are valid for any $t \in \mathbb{N}$. The asymptotic behavior is noted in the next remark.

Remark 5: If (11) holds, then $\limsup_{t\to+\infty} e_t \le (1-\tilde{\lambda})^{-1}(\tilde{\varphi}+\tilde{\eta}\tilde{\xi})$ *almost surely*, where $\tilde{\eta}$ and $\tilde{\xi}$ are upper bounds on the step size and $\mathbb{E}[\xi_t]$, respectively; the proof is omitted because of space limits, but follows arguments similar to [18, Corollary 4.8]. We also note that, since e_t is a Lyapunov function for the stochastic system (15), alternative input-to-state stability results in probability may be investigated by extending the arguments of [23, Thm. 2]. □

B. Remarks on the error model

The class of sub-Weibull distributions allows one to consider variety of error models. For instance, it includes a sub-Gaussian and a sub-exponential as sub-cases by setting $\theta = 1/2$ and $\theta = 1$, respectively. We notice that a sub-Gaussian assumption was typically utilized in prior works on stochastic gradient descent; for example, the assumption $\mathbb{E}[\exp\left(\xi^2/\sigma^2\right)] \le e$ in [24] 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, especially for small mini-batch (see, e.g., [25], [26]). To further elaborate on the flexibility offered by a sub-Weibull model, we also 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 subW(θ,ν), $i = 1,\dots,d$ for given $\theta, \nu > 0$. Then $\|\xi_t\|$ is sub-Weibull with $\|\xi_t\| \sim$ subW$(\theta, 2^\theta\sqrt{d}\nu)$ [18]. □

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 it sub-Gaussian with sub-Gaussian norm $C\varsigma$, with C an absolute constant [17], and it is therefore a sub-Weibull subW$(1/2, C'\varsigma)$ with C' an absolute constant. □

Example 4: Suppose that ξ_t is a random variable with mean $\mu := \mathbb{E}[x]$, such that $\xi_t \in [\bar{\xi}, \underline{\xi}]$ almost sure. Then $\xi_t - \mu \sim$ subW$(1/2, (\underline{\xi}-\bar{\xi})/\sqrt{2})$ [18]. □

V. Application to Electric Vehicle Charging

This section illustrates the use of the proposed algorithms in an application inspired from [4], 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 $x_i \in \mathbb{R}_{\ge 0}$ 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 $C_t := \{x \in \mathbb{R}^d : \sum_{i\in\mathcal{V}} x_i \le c_t\}$, 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 $z_i \in \mathbb{R}_{\ge 0}$ denote the selected price in region i, and we hypothesise that $z_i \sim \mathcal{N}(\mu_t x_i, \sigma_t^2)$, $\mu_{i,t}, \sigma_t \in \mathbb{R}_{\ge 0}$ 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 [4]:

$$\ell_t(x,z) = \sum_{i\in\mathcal{V}} z_i x_{i,t} - \gamma_{i,t} x_i + \kappa_{i,t} x_i^2, \tag{23}$$

where $\gamma_{i,t} \in \mathbb{R}_{\ge 0}$, 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 (23), 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 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 $i \in \{1,..,10\}$. 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 parameter values, the cost is α_t-strongly convex and β_t-jointly smooth with $\alpha_t = \beta_t = 2$. 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 the saddle point pair.

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.

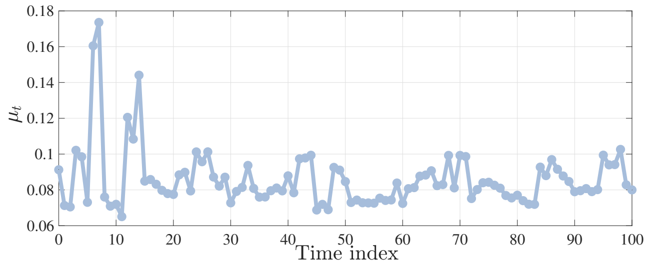

Fig. 1. Time series data representing the price of energy in dollars per kilowatt hour (kWh). Each time step represents 5 minutes.

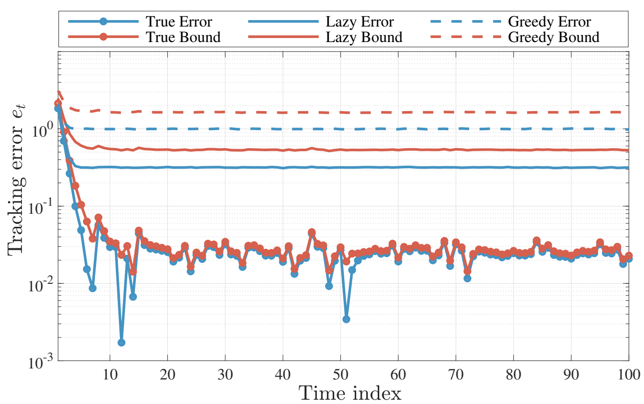

Fig. 2. Performance comparison of OPGD and OSPGD.

In Figure 2, we illustrate the the tracking errors and the corresponding time varying upper bounds presented in Theorems 3.3 and 4.1. "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 behavior of the tracking error for each algorithm; yet, in the instance of OSPGD the relationship is looser relative to the OPGD curves. Furthermore, the greedy and lazy OSPGD perform as expected in that both the mean tracking error and upper bound of the lazy deployment are smaller throughout the simulation.

VI. Conclusions

This paper 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 linear convergence results (within a bounded error) 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 Analysis*, vol. 38, no. 4, pp. 545–568, 2019.

[2] M. Brückner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2617–2654, 2012.

[3] 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 *IEEE Conference on Decision and Control*, 2011, pp. 4332–4339.

[4] W. Tushar, W. Saad, H. V. Poor, and D. B. Smith, "Economics of electric vehicle charging: A game theoretic approach," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1767–1778, 2012.

[5] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7599–7609.

[6] D. Drusvyatskiy and L. Xiao, "Stochastic optimization with decision-dependent distributions," *arXiv preprint arXiv:2011.11173*, 2020.

[7] A. Y. Popkov, "Gradient methods for nonstationary unconstrained optimization problems," *Automation and Remote Control*, vol. 66, no. 6, pp. 883–891, 2005.

[8] 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 *IEEE Conference on Decision and Control*, 2018, pp. 849–854.

[9] A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro, "Online optimization in dynamic environments: Improved regret rates for strongly convex problems," in *IEEE Conference on Decision and Control*, 2016, pp. 7195–7201.

[10] A. Simonetto, "Time-varying convex optimization via time-varying averaged operators," *arXiv preprint arXiv:1704.07338*, 2017.

[11] X. Cao, J. Zhang, and H. V. Poor, "Online stochastic optimization with time-varying distributions," *IEEE Tran. on Automatic Control*, vol. 66, no. 4, pp. 1840–1847, 2021.

[12] I. Shames and F. Farokhi, "Online stochastic convex optimization: Wasserstein distance variation," *arXiv preprint arXiv:2006.01397*, 2020.

[13] C. Wilson, V. V. Veeravalli, and A. Nedić, "Adaptive sequential stochastic optimization," *IEEE Tran. on Automatic Control*, vol. 64, no. 2, pp. 496–509, 2018.

[14] C. Mendler-Dünner, J. C. Perdomo, T. Zrnic, and M. Hardt, "Stochastic optimization for performative prediction," *arXiv preprint arXiv:2006.06887*, 2020.

[15] Z. Izzo, L. Ying, and J. Zou, "How to learn when data reacts to your model: performative gradient descent," *arXiv preprint arXiv:2102.07698*, 2021.

[16] 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.

[17] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University press, 2018.

[18] N. Bastianello, L. Madden, R. Carli, and E. Dall'Anese, "A stochastic operator framework for inexact static and online optimization," *arXiv preprint arXiv:2105.09884*, 2021.

[19] J. K. Hunter and B. Nachtergaele, *Applied analysis*. World Scientific Publishing Company, 2001.

[20] L. V. Kantorovich and S. Rubinshtein, "On a space of totally additive functions," *Vestnik of the St. Petersburg University: Mathematics*, vol. 13, no. 7, pp. 52–59, 1958.

[21] V. I. Bogachev and A. V. Kolesnikov, "The Monge-Kantorovich problem: achievements, connections, and perspectives," *Russian Mathematical Surveys*, vol. 67, no. 5, pp. 785–890, 2012.

[22] K. C. Wong, Z. Li, and A. Tewari, "Lasso guarantees for β-mixing heavy-tailed time series," *The Annals of Statistics*, vol. 48, no. 2, pp. 1124 – 1142, 2020.

[23] Y. Qin, M. Cao, and B. D. Anderson, "Lyapunov criterion for stochastic systems and its applications in distributed computation," *IEEE Tran. on Automatic Control*, vol. 65, no. 2, pp. 546–560, 2020.

[24] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[25] M. Gurbuzbalaban, U. Simsekli, and L. Zhu, "The heavy-tail phenomenon in SGD," *arXiv preprint arXiv:2006.04740*, 2020.

[26] L. Hodgkinson and M. W. Mahoney, "Multiplicative noise and heavy tails in stochastic optimization," *arXiv preprint arXiv:2006.06293*, 2020.