Convergence of the Inexact Online Gradient and Proximal-Gradient Under the Polyak-Łojasiewicz Condition

Seunghyun Kim, Liam Madden, and Emiliano Dall’Anese

Abstract—This paper focuses on the online gradient and proximal-gradient methods for optimization and learning problems with data streams. The performance of the online gradient descent method is first examined in a setting where the cost satisfies the Polyak-Łojasiewicz (PL) inequality and inexact gradient information is available. Convergence results show that the instantaneous regret converges linearly up to an error that depends on the variability of the problem and the statistics of the gradient error; in particular, we provide bounds in expectation and in high probability (that hold iteration-wise), with the latter derived by leveraging a sub-Weibull model for the errors affecting the gradient. Similar convergence results are then provided for the online proximal-gradient method, under the assumption that the composite cost satisfies the proximal-PL condition. The convergence results are applicable to a number of data processing, learning, and feedback-optimization tasks, where the cost functions may not be strongly convex, but satisfies the PL inequality. In the case of static costs, the bounds provide a new characterization of the convergence of gradient and proximal-gradient methods with a sub-Weibull gradient error. Illustrative simulations are provided for a real-time demand response problem in the context of power systems.

I. INTRODUCTION

This paper focuses on the online gradient descent and the online proximal-gradient descent methods for dynamic optimization and learning [1]–[19]. Because of their computational tractability and tracking performance, these are attractive first-order methods for solving key machine learning, data processing, and data-driven optimization tasks under streams of data, where data points and functions are processed on-the-fly and without storage; specific tasks include online classification [3], topology identification [11], logistic regression, as well as ℓ1-regularized, kernel-based, and robust linear regression [12] just to mention a few. Online gradient and proximal-gradient descent are powerful methods also in the context of online stochastic optimization [13], [14] and stochastic learning [15], [16] (along with the stochastic gradient descent counterparts), adaptive sequential learning [17], [18], and feedback-based optimization [19]–[21].

In this context, we examine the performance of online gradient and proximal-gradient descent in the presence of inexact gradient information, and when the cost to be minimized satisfies the Polyak-Łojasiewicz (PL) condition [22]–[24]. Formally, we consider time-varying optimization problems [1], [2] of the form

\[
\min_{x \in \mathbb{R}^n} F_t(x) := f_t(x) + g_t(x)
\]

where \( t \in \mathbb{N} \) is the time index, \( f_t : D \to \mathbb{R} \) is a continuously differentiable function with a Lipschitz-continuous gradient at each time \( t, D \subseteq \mathbb{R}^n \), and \( g_t : D \to \mathbb{R} \cup \{+\infty\} \) is a closed, convex and proper function uniformly in time, possibly not differentiable. Accordingly, we consider two main cases: c1) \( g_t(x) \equiv 0 \), \( x \mapsto f_t(x) \) satisfies the PL inequality uniformly in time, and an inexact gradient is available; and, c2) \( x \mapsto F_t(x) \) satisfies the proximal-PL inequality uniformly in time, and an inexact gradient is available.

Either way, the analysis is performed in terms of the instantaneous regret \( r_t := F_t(x_t) - F_t^* \), where \( F_t(x_t) \) is the cost achieved at time \( t \) by the point \( x_t \) produced by the algorithm and \( F_t^* \) is the optimal value function (that one would have achieved if the problem (1) was solved to convergence at time \( t \)). The dynamic regret \( R_t := \sum_{i=1}^t r_t \) can be analyzed as a byproduct.

Motivating examples for considering inexact gradient information are drawn from a variety of applications in learning and data-driven control; for example: i) settings where bandit and zeroth-order methods are utilized to estimate the gradient from (one or a few) functional evaluations [25]–[27]; ii) feedback-based optimization of networked systems, where errors in the gradient are due to measurement errors and asynchronous measurements [19]–[21]; and, iii) online stochastic optimization settings [13], [14], [17], i.e., when \( f_t(x) = \mathbb{E}[f_t(x,z)] \) for a given loss \( f_t : D \times \mathbb{R}^d \to \mathbb{R} \) and a random variable \( z \), and an approximate gradient may be computed using a single realization of \( z \) or a mini-batch.

As discussed in the seminal work [24], the PL inequality is satisfied by costs such as the least-squares (even in the under-determined case) and logistic regression; this is a property that has also been observed locally in the context of training of neural networks; see, e.g., [28], [29]. On the other hand, a prominent example of a problem that satisfies the proximal-PL is the sparse linear regression problem. The (proximal-)PL allows one to establish linear convergence of batch algorithms without strong convexity, as demonstrated in [24]; it is also weaker than other conditions such as essential strong convexity, weak strong convexity, and the restricted secant inequality, which have also been used to show linear...
convergence rates. Here, we intend to show that this attractive property is preserved in an online setting, and allows one to prove linear convergence to the sequence of optimal value functions \( \{ F^t \} \) within a given error bound that depends on the temporal variability of the optimization problem and the statistics of the gradient errors.

**Prior works.** Linear convergence of online gradient descent to the sequence of optimizers when the cost is strongly convex and smooth uniformly in time was shown in, e.g., [1], [2], [30] for the case of exact gradient information; similar results were derived in [5] for the online projected gradient method. A regret analysis was performed in, e.g., [7]–[9] (see also pertinent references therein), and the excess-risk was analyzed in [3]. Linear convergence of the proximal-gradient method can be readily shown by leveraging the non-expansiveness of the proximal operator; see, e.g., [4]. Inexact gradient information was considered in, e.g., [7], [10], where bounds in expectation on the regret incurred by the inexact online gradient descent were derived, and in [5] where the distance from the unique trajectory of optimizers was bounded in expectation. Convergence results in expectation were provided in the context of online stochastic optimization in, e.g., [13], [14]. Similar results in expectation can be found in, e.g., [4], [31] for the inexact proximal-gradient method.

We also acknowledge representative prior works on inexact gradient and proximal-gradient methods for batch optimization in, e.g., [32]–[34], and for the stochastic gradient descent in [35]–[38] (see also references therein). In particular, almost sure convergence to a first-order stationary point is proved assuming only strong smoothness and a weak assumption on the noise in [35]; mean convergence under the PL inequality is shown in, e.g., [37]. High-probability convergence results assuming strong smoothness and norm sub-Gaussian noise were provided in, e.g., [38], and in [39] for strongly convex functions in the non-smooth setting.

**Contributions.** We consider the cases c1) and c2) described above, and offer the following main contributions.

(i) We provide new bounds for the instantaneous regret \( r_t \) in expectation and in high probability for the inexact online gradient descent; the bounds include terms that quantify the temporal variability of the cost function, as well as the statistics of the gradient error. The high-probability convergence results are derived by adopting a sub-Weibull [40] model for the gradient error; our bounds scale more favorably compared to bounds obtained via Markov’s inequality, and hold iteration-wise. Finally, we provide an almost sure result for the asymptotic behavior of the regret \( r_t \).

(ii) Similarly, we provide new bounds for the instantaneous regret \( r_t \) in expectation and in high probability for the inexact online proximal-gradient descent; here, we consider an additional boundedness assumption on the gradient, as in prior literature (e.g., [10]).

(iii) For the case of exact gradient information, our bounds naturally extend the results of [24] to an online setting.

(iv) We also mention that, in the static optimization case (i.e., when the cost is not time-varying), we provide contributions relative to [32]–[38] by deriving high-probability convergence results under a sub-Weibull error assumption, and when the step-size is constant.

Illustrative simulations are provided for a real-time demand response problem in the context of a power systems. We conclude this section by mentioning that the sub-Weibull distribution allows one to consider a variety of error models in a unified manner; in fact, the sub-Weibull distribution includes sub-Gaussian distributions and sub-exponential distributions, and random variables whose probability density function has finite support as sub-cases [41], [42]. Furthermore, while the majority of the high-probability bounds in the context of stochastic optimization are derived using a sub-Gaussian assumption, recent works suggest that stochastic gradient descent may exhibit errors with tails that are heavier than a sub-Gaussian, especially for small mini-batch sizes (see, e.g., [43], [44]).

The rest of the paper is organized as follows. Section II introduces relevant definitions and assumptions, and Section III presents the main results for online gradient descent. Section IV focuses on the online proximal-gradient method, and Section V provides numerical results. Section VI concludes the paper.

### II. Preliminaries

We start by introducing relevant definitions and assumptions that will be utilized throughout the paper.

**Notation.** Upper-case (lower-case) boldface letters will be used for matrices (column vectors); \((\cdot)^\top\) denotes transposition. For given column vectors \( x, y \in \mathbb{R}^n \), \( \langle x, y \rangle \) denotes the inner product and \( \| x \| := \sqrt{x^\top x} \). Given a differentiable function \( f : \mathcal{D} \to \mathbb{R} \), defined over a domain \( \mathcal{D} \subseteq \mathbb{R}^n \) that is nonempty, \( \nabla f(x) \) denotes the gradient of \( f \) at \( x \) (taken to be a column vector). \( O(\cdot) \) refers to the big-O notation, whereas \( o(\cdot) \) refers to the little-o notation. For a given random variable \( \xi \in \mathbb{R} \), \( \mathbb{E}[\xi] \) denotes the expected value of \( \xi \), and \( \mathbb{P}[\xi \leq c] \) denotes the probability of \( \xi \) taking values smaller than or equal to \( c \); furthermore, \( \| \langle \xi \rangle \|_p := \mathbb{E}[|\xi|^p]^{1/p} \), for any \( p \geq 1 \). Finally, \( e \) will denote Euler’s number.

#### A. Modeling Assumptions and Definitions

We consider functions \( \{ f_t \}_{t \in \mathbb{N}} \) and \( \{ g_t \}_{t \in \mathbb{N}} \), defined over a (nonempty) domain \( \mathcal{D} \subseteq \mathbb{R}^n \), that satisfy the following assumptions.

**Assumption 1:** The function \( x \mapsto f_t(x) \) is continuously differentiable and has a Lipschitz-continuous gradient over \( \mathcal{D} \), uniformly in \( t \); i.e., \( \exists L > 0 \) such that \( \| \nabla f_t(x) - \nabla f_t(y) \| \leq L \| x - y \| \) for any \( x, y \in \mathcal{D} \), for all \( t \).

**Assumption 2:** For every \( t \in \mathbb{N} \), the function \( x \mapsto g_t(x) \) is convex, proper, and lower semi-continuous, possibly non-differentiable over \( \mathcal{D} \).

Recall that the following inequality follows from the Lipschitz-continuity of the gradient of \( f_t \):

\[
 f_t(y) \leq f_t(x) + \langle \nabla f_t(x), y - x \rangle + \frac{L}{2} \| y - x \|^2 
\]  

(2)

for any \( x, y \in \mathcal{D} \); this inequality will be utilized throughout the paper to derive a number of technical results. Let
As previously discussed, classes of functions satisfying the PL inequality and the proximal-PL inequality are described in, e.g., \cite{24}; a few examples will be provided shortly. We also refer the reader to recent discussions on the PL inequality in the context of training of neural networks in, e.g., \cite{28, 29} (and pertinent references therein).

\section{Sub-Weibull random variables}

In this section, we introduce the definition of sub-Weibull random variable \cite{40, 45}, which will be utilized to model the errors incurred by the inexact online gradient methods. We will also explain how to obtain sub-Gaussian and sub-exponential random variables \cite{41} as sub-cases.

\textbf{Definition 3 (Sub-Weibull random variable):} A random variable $\xi$ is sub-Weibull if $\exists \theta > 0, K > 0$ such that

$$E \left[ \exp \left( \left( |\xi| / K \right)^{\frac{1}{\theta}} \right) \right] \leq 2.$$  \hspace{1cm} (9)

The coefficient $\theta$ is related to the rate of decay of the tails; in particular, the tails become heavier as the parameter $\theta$ grows larger. By setting $\theta = 1/2$, one can recover the class of sub-Gaussian random variables; on the other hand, $\theta = 1$ yields the class of sub-exponential random variables (see, e.g., \cite{41}). The parameter $K > 0$ can be considered as a proxy for the variance. In the following, we will use the notation $\xi \sim \text{subW}(\theta, K)$ to refer to a sub-Weibull random variable $\xi$ with parameters $\theta$ and $K$.

We note that a sub-Weibull random variable can be equivalently characterized using the following properties, for a given $\theta > 0$ \cite[Theorem 1]{40}:

\begin{enumerate}[label=(\alph*)]
  \item $\exists K_1 > 0$ s.t. $P \left[ |\xi| \geq \varepsilon \right] \leq 2 \exp \left( - (\varepsilon / K_1)^{\frac{1}{\theta}} \right)$, $\forall \varepsilon > 0$;
  \item $\exists K_2 > 0$ s.t. $\|\xi\|_k \leq K_2^k \theta, \forall k \geq 1$;
  \item $\exists K_3 > 0$ s.t. $E \left[ \exp \left( (\lambda |\xi|)^{\frac{1}{\theta}} \right) \right] \leq \exp \left( (\lambda K_3)^{\frac{1}{\theta}} \right)$, $\forall \lambda \in (0, 1 / K_3]$;
  \item $\exists K_4 > 0$ s.t. $E \left[ \exp \left( (|\xi| / K_4)^{\frac{1}{\theta}} \right) \right] \leq 2$;
\end{enumerate}

where the parameters $K_1, K_2, K_3, K_4$ differ each by a constant that depends on $\theta$. In particular, (d) implies (a) with $K_1 = K_4$ \cite[Theorem 1]{40}.

The following lemmas will be utilized throughout the paper to derive high-probability bounds (see also \cite[Corollary 1]{40} for random variables that are identically distributed).

\textbf{Lemma 2.1:} Let $\xi \sim \text{subW}(\theta, K)$, for a given $\theta \geq 0$ and $K > 0$. Let $\alpha > 0$. Then, $\alpha \xi^2 \sim \text{subW}(2\theta, \alpha K^2)$.

\textbf{Proof:} The result follows from definition (d).

\textbf{Lemma 2.2:} Consider the random variables $\xi_1, \ldots, \xi_n$, with $\xi_i \sim \text{subW}(\theta, K_i)$ for $i = 1, \ldots, n$ where $\theta \geq 1$. Then, for all $t \geq 0$,

$$P \left[ \sum_{i=1}^{n} \xi_i \geq t \right] \leq 2 \exp \left( - \left( \frac{t}{2(2\theta)^{\frac{1}{\theta}} \sum_{i=1}^{n} K_i^{\frac{1}{\theta}}} \right)^{\frac{1}{\theta}} \right).$$  \hspace{1cm} (10)
Proof: The proof follows by combining the results of [40] Theorem 1, [45] Lemma 5, and the triangle inequality for $L^p$ spaces. For the first part of the proof, drop the subscript $i$, and consider the following inequalities for any $\xi = \xi_i$, $K = K_i$.

First, for all $t \geq 0$, one has, by Markov’s inequality:

$$
P[|\xi| \geq t] = P \left[ \exp \left( (|\xi|/K)^{1/\theta} \right) \geq \exp \left( (t/K)^{1/\theta} \right) \right] \leq 2 \exp \left( - (t/K)^{1/\theta} \right).
$$

(11)

Thus, the constant incurred by going from definition (a) to (b) is 1. Secondly, for all $p \geq 1$, using the formula of the cumulative distribution function for the expected value,

$$
\mathbb{E}[|\xi|^{p}] = \int_{0}^{\infty} P[|\xi|^{p} \geq x] dx
$$

(12a)

$$
\leq 2 \int_{0}^{\infty} \exp \left( - \left( x^{1/p}/K \right)^{1/\theta} \right) dx
$$

(12b)

$$
= 2\theta p K^p \int_{0}^{\infty} e^{-u u^{\theta p-1}} du, \quad u = \left( \frac{x^p}{K} \right)^{\frac{1}{\theta}}
$$

(12c)

$$
= 2 \theta p \Gamma(\theta p) K^p, \quad \text{as } \Gamma(z) := \int_{0}^{\infty} x^{z-1} e^{-x} dx
$$

(12d)

$$
= 2 \Gamma(\theta p + 1) K^p
$$

(12e)

$$
\leq 2 K^p \theta p, \quad \text{since } \Gamma(z+1) \leq z^z \text{ for } z \geq 1
$$

(12f)

$$
= \left( 2 \theta p K^p \theta^p \right)^p \theta^p
$$

(12g)

$$
\leq (2\theta^p K^p \theta^p)^p.
$$

(12h)

Thus, the constant incurred by going from definition (a) to (b) is $2\theta^p$. Recall that $\|\xi\|_p := \mathbb{E}[|\xi|^{p}]^{1/p}$, for any $p \geq 1$. Then, one has that $\|\xi\|_p \leq 2\theta^p K^p \theta^p \forall p \geq 1$. Using the triangle inequality, for all $p \geq 1$,

$$
\left\| \sum_{i=1}^{n} \xi_i \right\|_p \leq \sum_{i=1}^{n} \|\xi_i\|_p \leq 2\theta^p \left( \sum_{i=1}^{n} K_i \right)^{\theta^p}
$$

(13)

Next, let $\xi = \sum_{i=1}^{n} \xi_i$, and let $K = 2\theta^p \sum_{i=1}^{n} K_i$ so that $\|\xi\|_p \leq K^p \theta^p$ for all $p \geq 1$. Then, for $1/\theta \leq p < 1$, by Jensen’s inequality,

$$
\mathbb{E}[|\xi|^{1/p}] \leq \mathbb{E}[|\xi|] \leq K \leq \theta^p K^p.
$$

(14)

Now let $K = 2\theta^2 \sum_{i=1}^{n} K_i$. Then, for any $\lambda \in \left( 0, \frac{1}{(2\theta)/\theta^p} \right)$,

$$
\mathbb{E} \left[ \exp \left( (\lambda |\xi|)^{1/\theta} \right) \right] = \mathbb{E} \left[ 1 + \sum_{p=1}^{\infty} \frac{\mathbb{E}[|\lambda|^{p/\theta}]}{p!} \right]
$$

$$
\leq \sum_{p=1}^{\infty} \frac{\left( (\lambda K)^{1/p} \theta^p \right)^p}{(p/e)^p}
$$

$$
= \sum_{p=1}^{\infty} \left( (\lambda K)^{1/\theta} e/\theta \right)^p
$$

$$
= \frac{1}{1 - (\lambda K)^{1/\theta} e/\theta}
$$

$$
\leq 2.
$$

In particular, let $\lambda = \frac{1}{(2\theta)/\theta^p}$, it follows that:

$$
\mathbb{E} \left[ \exp((\lambda |\xi|)^{1/\theta}) \right] \leq \frac{1}{1 - (2\theta/e)\theta^p} = 2.
$$

Thus, the constant incurred by going from definition (b) to (d) is $\theta^p (2\theta/e)^p = (2e)^p$. Finally, let $K = 2(2e\theta)^p$. We just showed that $\xi$ satisfies definition (d) with constant $K$, so going from definition (d) to (a) again, incurring a constant of 1, completes the proof.

We conclude this section by noting that, if $\xi \sim \text{subW}(\theta, K)$, then the expected value $\mathbb{E}[|\xi|]$ can be bounded as follows for $\theta \geq 1$:

$$
\mathbb{E}[|\xi|] \leq 2\theta^p K.
$$

(15)

### III. INEXACT ONLINE GRADIENT DESCENT

We start by considering the case where $g_t \equiv 0$ for all $t$; accordingly, the time-varying problem (1) reduces here to:

$$
\min_{x} f_t(x)
$$

(16)

where we recall that $t \in \mathbb{N}$ is the time index. We consider the following inexact online gradient descent (OGD):

$$
x_{t+1} = x_t - \eta v_t
$$

(17)

where $\eta > 0$ is a given step-size and $v_t := \nabla f_t(x_t) + e_t$

(18)

is an inexact gradient of $f_t$ at $x_t$, with the vector $e_t \in \mathbb{R}^n$ modeling the errors in the gradient. We are interested in studying the performance of (17) when the function $f_t$ satisfies the PL inequality (6), and the error $\|e_t\|$ follows a sub-Weibull distribution. A discussion on the sub-Weibull model as well as the PL inequality in the context of problems in data processing, learning, and feedback-based optimization is provided in Section III-B. The main convergence results is presented next.

#### A. Convergence results

Since $g_t \equiv 0$, the instantaneous regret at time $t$ boils down here to $r_t = f_t(x_t) - f^*_t$. Throughout this section, we assume that the gradient error has a sub-Weibull distribution, as formalized next [cf. Definition 3].

**Assumption 4:** The error $\|e_t\|$ is subW($\theta, K_t$), for given $\theta \geq 1/2$ and $K_t > 0$.

In particular, Assumption 4 implies that $\mathbb{E}[\|e_t\|^2] \leq 2(2\theta)^{2\theta} K_t^2$ by Lemma 2.2 and Eq. (15). In the following, we state the main results concerning the convergence of the inexact OGD (17).

**Theorem 3.1 (Convergence of the inexact OGD):** Let Assumptions 1, 3, and 4 hold. Assume further that the function $x \mapsto f_t(x)$ satisfies the PL inequality for some $\mu > 0$ uniformly in time. Let $\{x_t\}_{t=0}^n$ be a sequence generated by the inexact OGD (17) with $\eta = 1/L$. Then, the following bounds hold for (17):

...
1) For all \( t \in \mathbb{N} \):
\[
\mathbb{E}[r_t] \leq \zeta^t r_0 + \sum_{\tau=1}^{t} \zeta^{t-\tau} \left( \frac{1}{2L^2} \mathbb{E}[\|e_{\tau-1}\|^2] + \psi_{\tau} \right) 
\]  
(19)
where \( \zeta := (1 - \frac{\mu}{L}) \).

2) If \( \delta \in (0, 1) \), then with probability \( 1 - \delta \):
\[
r_t \leq \zeta^t r_0 + \sum_{\tau=1}^{t} \zeta^{t-\tau} \left( \frac{c(\theta, \delta)}{2L} K_{\tau-1}^2 + \psi_{\tau} \right) 
\]  
(20)
where \( c(\theta, \delta) := 2(4e\theta)^{2\theta} \log^{2\theta}(2\delta^{-1}) \).

**Corollary 3.2 (Asymptotic convergence of the OGD):** Let Assumptions [1] [2] and [4] hold, and assume that \( x \mapsto f_t(x) \) satisfies the \( \mu \)-PL inequality uniformly in time. Let \( \{x_t\}_{t=0}^{\infty} \) be a sequence generated by the inexact OGD (17) with \( \eta = 1/L \). Then,
\[
\limsup_{t \to \infty} r_t \leq \frac{1}{2\mu} \bar{e} + \frac{L}{\mu} \bar{\psi} \quad \text{a.s.} 
\]  
(21)
where \( \bar{e} = \sup_{\tau} \{\mathbb{E}[\|e_{\tau}\|^2]\} \).

**Corollary 3.3 (Online optimization without gradient errors):** Let Assumptions [1] and [8] hold, and assume that \( x \mapsto f_t(x) \) satisfies the \( \mu \)-PL inequality uniformly in time. Let \( \{x_t\}_{t=0}^{\infty} \) be a sequence generated by the OGD (17) with perfect gradient information and \( \eta = 1/L \). Define the cumulative drift up to time \( t \in \mathbb{N} \) as \( \Psi_t := \sum_{\tau=1}^{t} \sigma_{\tau} + \tilde{\sigma}_{\tau}, \tilde{\sigma}_{\tau} := |F_{\tau}(x_\tau) - F_{\tau-1}(x_\tau)| \); then, the dynamic regret \( R_t := \sum_{\tau=1}^{t} r_{\tau} \) can be bounded as
\[
R_t \leq \frac{L}{\mu} r_0 + \frac{L}{\mu} \Psi_t. 
\]  
(22)

Regarding the last result, it is clear that the asymptotic behavior of \( \frac{1}{T} R_t \) now depends on \( \Psi_t \). If \( \Psi_t \) grows as \( O(t) \) then \( \frac{1}{T} R_t = O(1) \). On the other hand, if \( \Psi_t \) grows sublinearly, i.e., as \( o(T) \), then \( \frac{1}{T} R_t \) tends to 0.

Before providing the proof of the results, a few remarks are in order.

**Remark 1 (Static optimization with gradient errors):** Consider the case where the optimization problem (16) is time-invariant; i.e., \( f_t(x) = f(x) \) for all \( t \in \mathbb{N} \). Then, (19) is similar to [24, Thm. 4] (where a different step-size was used). However, relative to [24], we provide the following bound in high probability
\[
r_t \leq \zeta^t r_0 + \sum_{\tau=1}^{t} \zeta^{t-\tau} \left( \frac{c(\theta, \delta)}{2L} K_{\tau-1}^2 \right) 
\]  
(23)
which holds with probability \( 1 - \delta \) for any \( \delta \in (0, 1) \); this bound is straightforwardly derived from (20) by setting \( \psi_{\tau} = 0 \) for all \( \tau = 1, \ldots, t \). We also have the asymptotic behavior
\[
\limsup_{t \to \infty} r_t \leq \frac{1}{2\mu} \bar{e} \quad \text{a.s.} 
\]  
(24)
by Corollary [5,2].

**Remark 2 (Alternative bound in expectation):** An alternative bound in expectation can be expressed as
\[
\mathbb{E}[r_t] \leq \zeta^t r_0 + \sum_{\tau=1}^{t} \zeta^{t-\tau} \left( \frac{1}{2L^2} \mathbb{E}[\|e_{\tau-1}\|^2] + \mathbb{E}[\tilde{\psi}_{\tau}] \right) 
\]  
(25)
where \( \tilde{\psi}_{\tau} = \sigma_{\tau} + \tilde{\phi}_{\tau}, \tilde{\phi}_{\tau} := |F_{\tau}(x_\tau) - F_{\tau-1}(x_\tau)| \) (where the expectation \( \mathbb{E}[\tilde{\psi}_{\tau}] \) is taken with respect to the error \( e_{\tau-1} \), conditioned on a filtration). This leads to a tighter bound relative to (19).

**Remark 3 (Static optimization without gradient errors):** When the optimization problem (16) is time-invariant and one has access to perfect gradient information, then the result of Theorem 3.1 boils down to \( r_t \leq \zeta^t r_0 \), which coincides with [24, Thm. 1].

**Remark 4 (Markov’s inequality):** An alternative high probability bound can be obtained by using (19) and Markov’s inequality. However, the resulting bound would have a dependence \( \delta^{-1} \); on the other hand, our bound has a \( \log(\delta^{-1}) \) dependence on \( \delta \).
Next, add $-f_t(x_{t+1})$ and $f_{t+1}(x_{t+1})$ on both sides. Then, applying the definitions of $\zeta$ and $\psi_t$ to obtain:

$$f_t(x_{t+1}) + f_{t+1}(x_{t+1}) - f_{t+1}^{\star} - f_t^{\star} \leq \zeta(f_t(x_t) - f_t^{\star}) + \frac{1}{2L}||e_t||^2 + \frac{1}{2L}||e_{t+1}(x_{t+1}) - f_{t+1}^{\star}||^2 \
\implies f_{t+1}(x_{t+1}) - f_{t+1}^{\star} \leq \zeta(f_t(x_t) - f_t^{\star}) + \frac{1}{2L}||e_t||^2 + \psi_{t+1}.$$  

(29)

The result follows by taking the expectation on both sides. ■

**Proof of Theorem 3.1** From the proof of Lemma 2.2, we have the stochastic inequality $r_t \leq \zeta r_{t-1} + \frac{1}{2L}||e_{t-1}||^2 + \psi_t$, which holds almost surely; unraveling, we get

$$r_t \leq \zeta^t r_0 + \frac{1}{2L} \sum_{i=0}^{t-1} \zeta^{t-i-1}||e_i||^2 + \sum_{i=1}^t \zeta^{t-i}\psi_i.$$  

(30)

Taking the expectation on both sides, we get (19).

For the high-probability bound (20), recall that $||e_i|| \sim \text{subW}^{\star}(\theta, K_t)$; by Lemma 2.2, $||e_i||^2 \sim \text{subW}(2\theta, K_t^2)$. Applying Lemma 2.2 with probability $1 - \delta$ we have that

$$\sum_{i=0}^{t-1} \zeta^{t-i-1}||e_i||^2 \leq 2(4e\theta)^{2\delta} \log((2\delta)^{-1}) \sum_{i=0}^{t-1} \zeta^{t-i-1}K_t^2.$$  

(31)

Let $\kappa_t := \zeta^t r_0 + \sum_{i=0}^{t-1} \zeta^{t-i-1}||e_i||^2$, and rewrite (30) as $r_t \leq \kappa_t + \frac{1}{2L} \sum_{i=0}^{t-1} \zeta^{t-i-1}||e_i||^2$, which holds almost surely. Since $\kappa_t$ is a constant, for a given $\epsilon \in (0, 1)$, one has that

$$\mathbb{P}(r_t \geq \epsilon + \kappa_t) \leq \mathbb{P}(\frac{1}{2L} \sum_{i=0}^{t-1} \zeta^{t-i-1}||e_i||^2 + \kappa_t \geq \epsilon + \kappa_t) = \mathbb{P}(\frac{1}{2L} \sum_{i=0}^{t-1} \zeta^{t-i-1}||e_i||^2 \geq \epsilon).$$

Using (31), the bound (20) follows. ■

The proof of Corollary 3.2 follows similar steps as in [42, Corollary 4.8], and is omitted. The result of Corollary 3.3 can be derived from Theorem 3.1 by setting the error to zero, and by applying the definition of cumulative regret $R_t$.

**B. Remarks on applications and error model**

In this section, we provide a few examples of applications that are relevant to our setting. We also discuss the sub-Weibull error model, to highlight the merits of this modeling approach.

**Example 1 (Online least-squares):** A function of the form $f_t(x) = h_t(A_t x)$, with $h_t : \mathbb{R}^d \rightarrow \mathbb{R}$ a $\nu$-strongly convex function and $A_t \in \mathbb{R}^{d \times n}$ a given matrix, satisfies the PL inequality [24]. This class includes the least-squares (LS) by setting $f_t(x) = \frac{1}{2}||A_t x - b_t||^2$. Note that, when the matrix $A_t$ is not full-column rank, one can utilize the results of this paper to establish linear convergence of OGD for the under-determined LS problem.

**Example 2 (Online Logistic regression):** The logistic regression cost $f_t(x) = \sum_{i=1}^n \log(1 + \exp(b_i^t a_i^T x))$, with $b_{i,t} \in \mathbb{R}$ and $a_i \in \mathbb{R}^n$, satisfies the PL inequality [24].

**Example 3 (Online optimization of LTI systems):** Consider the algebraic representation of a stable linear time-invariant system $y_t = Gu + Hw_t$, where $u$ is the vector of controllable inputs and $w_t$ are unknown exogenous disturbances. Suppose that $f_t(x) = \frac{1}{2}||Gu + Hw_t - \tilde{y}_t||^2$ with $\tilde{y}_t$ a time-varying reference. Since $w_t$ is unknown, one way to compute $\nabla f_t(u_t)$ is $v_t = G^T(\tilde{y}_t - \bar{y}_t)$, where $\bar{y}_t$ is a (noisy) measurement of the output $y_t$ [19], [20].

**Example 4 (Training of neural networks):** We refer the reader to recent discussions on the PL inequality in the context of training of neural networks in, e.g., [28], [29]. The proposed framework may capture the case where stochastic gradient methods are utilized to train a neural network in an online fashion.

In terms of inexactness of the gradient, the error $e_t$ may arise in the following (application-specific) scenarios:

(i) A subset of the data points available at time $t$ are utilized to compute the gradient; for instance, in the Examples 1-2, one may utilize the data points $\{a_{i,t}, b_{i,t}\}_{i \in S_t}$, with $|S_t| < d$.

(ii) Bandit and zeroth-order methods are utilized to estimate the gradient [25]-[27].

(iii) In an online stochastic optimization setting [13], i.e. when $f_t(x) = E[f_t(x,z)]$ for a given loss $f_t : \mathbb{R}^n \rightarrow \mathbb{R}$ and a random variable $z$, the approximate gradient $v_t$ may be computed using a mini-batch.

(iv) In measurement-based algorithms as in Example 3 measurement errors and asynchronous measurements render the computation of the gradient inexact.

See also, e.g., [4], [16], [32], [33], [35], [46] for additional motivations (the proximal-gradient method will be explained in Section IV). We also point out that a classical setting for the inexact gradient descent to be well-behaved is to assume that $-v_t$ is, on average, a direction of sufficient descent; see, e.g., [46]; a more stringent assumption is the one where the mean of gradient error is zero.

Regarding the error $||e_t||$, the sub-Weibull distribution allows one to consider a variety of cases in a unified manner; in fact, the sub-Weibull distribution includes sub-Gaussian distributions and sub-exponential distributions, and random variables whose probability density function has finite support in fact, the sub-Weibull distribution includes sub-Gaussian distributions and sub-exponential distributions, and random variables whose probability density function has finite support.
Consider then the inexact online proximal-gradient method (OPGM), which involves the following step at each $t \in \mathbb{N}$:

$$
x_{t+1} = \text{prox}_{\mu g_t} \left\{ x_t - \frac{1}{L} v_t \right\}
$$  \hspace{1cm} (32)

where $v_t$ is again an estimate of $\nabla f_t(x_t)$, $\text{prox}_{\mu g_t} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ denotes the proximal operator, and the step-size is taken to be $1/L$. We are now interested in analyzing the behavior of (32) in terms of regret $r_t = F_t(x_t) - F^*$, where we recall that $F^*$ is the optimal value function, when the function $F_t$ satisfies proximal-PL inequality and the error $\|e_t\|$ follows a sub-Weibull distribution.

To enable the analysis, we introduce the following additional regularity assumption.

Assumption 5: $\exists B < \infty$ such that $\|\nabla f_t(x_t)\| \leq B$ for all $x_t \in \mathcal{D}$ generated by (32).

This is a boundedness assumption that is common in the context of stochastic gradient descent for convex losses or when the PL inequality is satisfied over the domain of interest $\mathcal{D}$. When $g_t = g$ is static, we refer to the indiscriminator function for a polytope. Assumption 5 is automatically satisfied since the gradient if $f_t$ is Lipschitz continuous. In other cases, Assumption 5 is imposed over the domain of the function, which can be taken to be the largest forward-invariant set of the algorithm (under the current assumptions, divergence of the inexact OPGM is an event with zero probability); see also the discussion in, e.g., [10, 28, 29, 33, 46].

The main convergence result for (32) is then stated next.

**Theorem 4.1 (Convergence of the inexact OPGM):** Let Assumptions 1-5 hold. Assume further that the function $x \mapsto F_t(x)$ satisfies the proximal-PL inequality for some $\mu > 0$ uniformly in time. Let $\{x_t\}_{t=0}^t$ be a sequence generated by the inexact OPGM (32). Then, the following hold:

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

$$
\mathbb{E}[r_t] \leq \zeta^t r_0 + \sum_{\tau=1}^t \zeta^{t-\tau} \left( \frac{2}{L} \epsilon_{t-\tau} + \psi_{t-\tau} \right)
$$  \hspace{1cm} (33)

where $\epsilon_{\tau} := \mathbb{E}[\|e_{\tau}\|^2] + B\mathbb{E}[\|e_\tau\|]$ and $\zeta = (1 - \frac{\mu}{L})$.

2) If $\delta \in (0,1)$, then with probability $1 - \delta$:

$$
r_t \leq \zeta^t r_0 + \sum_{\tau=1}^t \zeta^{t-\tau} \left( \frac{2c(\theta,\delta)}{L} \kappa_{t-\tau} + \psi_{t-\tau} \right)
$$  \hspace{1cm} (34)

where $c(\theta,\delta) = 2(4e\theta)^{2\theta} \log^{2\theta}(2\delta^{-1})$ and $\kappa_{t} = K_t^2 + BK_t$.

A result for the asymptotic convergence of the OPGM similar to Corollary 2.2 can be derived too, but it is omitted to avoid repetitive arguments. Similar considerations as in Remark 1 can also be drawn; in particular, in the case of an OPGM without gradient errors, a bound for the dynamic regret reduces to $R_t \leq \frac{\mu}{L} r_0 + \frac{L}{2} y_t$.

To outline the proof of the theorem, we first note that step (32) is equivalent to

$$
x_{t+1} = \arg\min_{x} \{v_t, x - x_t\} + \frac{L}{2} \|x - x_t\|^2 + g_t(x) - g_t(x_t).
$$  \hspace{1cm} (35)

We also recall the definition of $\tilde{A}_{g_t}(x_t, 1/L)$ in (38), and define $\tilde{A}_{g_t}(x_t, 1/L)$ as:

$$
\tilde{A}_{g_t}(x_t, 1/L) := -2L \min_y \{v_t, y - x_t\} + \frac{L}{2} \|y - x_t\|^2 + g_t(y) - g_t(x_t).
$$  \hspace{1cm} (36)

Lastly, for any $x \in \mathcal{D}$, we recall that $e_t \in \mathbb{R}^n$ is the gradient error, i.e., $v_t = \nabla f_t(x) + e_t$, and $\|e_t\| \sim \text{sub-Weibull}$.  

**Proof of Theorem 4.1:** We start by recalling that $F_{t+1}(x_{t+1}) = F_{t+1}(x_{t+1}) + g_{t+1}(x_{t+1})$; adding and subtracting $F_t(x_{t+1})$ on the right-hand-side, and using the definition (36), we get

$$
F_{t+1}(x_{t+1}) \leq F_t(x_t) + \langle v_t, x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 + g_{t+1}(x_{t+1}) + g_t(x_{t+1}) - g_t(x_t)
$$

where we used (2) in the last step. Next, we add and subtract $e_t$ in the inner product and use the definition $v_t = \nabla f_t(x) + e_t$ to obtain:

$$
F_{t+1}(x_{t+1}) \leq F_t(x_t) + \langle v_t, x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 + g_{t+1}(x_{t+1}) + g_t(x_{t+1}) - g_t(x_t)\tag{37}
$$

where $\epsilon_t := |\tilde{A}_{g_t}(x_t, 1/L) - \tilde{A}_{g_t}(x_t, 1/L)|$ for brevity; using the definition of $\tilde{A}_{g_t}(x_t, 1/L)$ in (38) and subtracting $\tilde{F}_{t+1}$ on both sides, we get:

$$
F_{t+1}(x_{t+1}) - F^* \leq F_t(x_t) - F^* + \frac{\mu}{L} (F_t(x_t) - F^*) + \frac{1}{2L} \epsilon_t - \langle e_t, x_{t+1} - x_t \rangle + \phi_{t+1}
$$

where $\epsilon_t := |\tilde{A}_{g_t}(x_t, 1/L) - \tilde{A}_{g_t}(x_t, 1/L)|$ for brevity. Using the fact that the proximal oper-
for any $\tau = 0$

$$z\equiv \text{minimizer); thus, substituting}$$

$$\langle W e now bound$$

$$\text{inequality, we get:}$$

$$\text{ator is quasi-non-expansive, and using the Cauchy–Schwarz}$$

$$\text{is not}\text{modify the minimizer); thus, substituting} z\text{with}\ x_{t+1}\text{we get}$$

$$\text{We now bound} \frac{1}{2L} \varepsilon_t = \min_{y} \left\{ \langle \nabla f_t(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 + g_t(y) \right\}$$

$$\text{From (33), one can notice that the minimizer of} \langle v_t, z - x_t \rangle + \frac{L}{2} \|z - x_t\|^2 + g_t(z) \text{is} x_{t+1} \text{(the constant term} g_t(x_t) \text{does not modify the minimizer); thus, substituting} z\text{with}\ x_{t+1}\text{we get}$$

$$\frac{1}{2L} \varepsilon_t \leq \min_{y} \left\{ \langle \nabla f_t(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 + g_t(y) \right\}$$

$$\text{Next, one has that} \min_y \left\{ \langle \nabla f_t(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 + g_t(y) \right\} \leq \langle \nabla f_t(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 + g_t(y) \text{for any}$$

$$\text{for any} y \in D. \text{Pick} y = x_{t+1}; \text{then, we have that:}$$

$$\frac{1}{2L} \varepsilon_t \leq \langle \nabla f_t(x_t), x_{t+1} - x_t \rangle - \langle v_t, x_{t+1} - x_t \rangle$$

$$\text{From (42a)–(42b), it then follows that}$$

$$\frac{1}{2L} \varepsilon_t \leq \frac{1}{L} \|e_t\| \|\nabla f_t(x_t)\| + \frac{1}{L} \|e_t\|^2.$$
Consider the following time-varying optimization problem for the power flow equations in case of resistive lines [19]. Negligible or they are derived based on a linearized model for the inexact OGD and, since the empirical 3-standard deviation confidence interval, and the theoretical bound [25].

We next consider the inexact OGD; in particular, we artificially corrupt the gradient with a random vector $e_t$, which is modelled as a Gaussian vector $\mathcal{N}(0, 10^{-5} \mathbf{I})$; we note that, is $e_t$ is a Gaussian vector, then $\|e_t\|^2$ is a sub-Weibull random variable. The regret is computed using a Monte Carlo approach, with 100 tests. Accordingly, Figure 2 illustrates the evolution of the expected regret obtained by averaging the trajectories of the instantaneous regret $r_t$ over the various runs, the empirical $3 - \sigma$ confidence interval, and the theoretical bound [25]. The figure validates the convergence results for the inexact OGD and, since $b_t$ continuously changes, the average $r_t$ exhibits a plateau.

**B. Real-time demand response**

We next consider an example in the context of a power distribution grid serving residential houses or commercial facilities. We consider $n$ controllable distributed energy resources (DERs) providing services to the main grid; precisely, consider the setting where the vector $x$ collects the active power outputs of the DERs, and assume the algebraic relationship $p_{0,t} = a_x x + a_w w_t$ for the net active power at the point of common coupling, where $a_x \in \mathbb{R}^n$ and $a_w \in \mathbb{R}^w$ are sensitivity coefficients, and $w_t \in \mathbb{R}^w$ is a vector collecting active powers of uncontrollable devices; in particular, $a_x$ and $a_w$ can be set to the vector of all ones when line losses are negligible [47], or they are derived based on a linearized model for the power flow equations in case of resistive lines [19].

Consider the following time-varying optimization problem for real-time management of DERs:

$$\min_x \frac{1}{2} (a_x^\top x + a_w^\top w_t - p_{0,t}^\text{ref})^2 + I_{\{|Bx| \leq c\}} \tag{52}$$

where $p_{0,t}^\text{ref}$ is a time-varying reference point for the net active power at the point of common coupling $p_{0,t}$, and $I_{\{|Bx| \leq c\}}$ is the set indicator function for the set $\{x \in \mathbb{R}^n : Bx \leq c\}$ modeling box constraints for the active powers. For example, $p_{0,t}$ may be an automatic control generation (ACG) signal, a flexible ramping signal, or a demand response setpoint. We note that the cost (52) satisfies the proximal-PL inequality [24]. The main challenge behind applying a proximal-gradient descent to (52) is that the vector $w_t$ is unknown; we therefore consider the approach of, e.g., [19], [20], where measurements of $p_{0,t}$ are utilized to estimate the gradient in lieu of the model $a_x^\top x + a_w^\top w_t$. Precisely, we compute the approximate gradient as

$$v_t = a_x (\hat{p}_{0,t} - p_{0,t}^\text{ref}) \tag{53}$$

where $\hat{p}_{0,t}$ is a measurement of $p_{0,t}$ collected at time $t$. Since measurements of $p_{0,t}$ may be affected by errors or by outliers, $v_t$ does not in general coincide with the true gradient $a_x (a_x^\top x + a_w^\top w_t - p_{0,t}^\text{ref})$.

As an example, we consider the case where $N = 500$ DERs are controlled; the limits for the active power of each device are $[-50, 50]$ kW for energy storage resources and $[0, 50]$ kW for solar inverters. We consider the case where $p_{0,t}^\text{ref}$ follows the trajectory shown in Figure 3 real data with a granularity of one second is taken from [48] to generate the non-controllable powers $w_t$, with the net power $a_w^\top w_t$ plotted in Figure 3 as...
well. The sensitivity vector \( a_w \) is computed as in [43]. A Gaussian random variable with zero mean and variance 10 kW is utilized to generate the measurement error affecting \( p_0,t \).

Figure 4 illustrates the evolution of the regret \( \tau_t \), averaged over 50 experiments, in logarithmic scale. One can notice a linear decrease of the average regret during the first iteration of the algorithms; the regret then exhibits variations that are due to the considerable time-variability of the cost function (due to the large swings in the non-controllable powers \( w_j \)). The plot also provides a zoomed version (in linear scale), where the 3-standard deviation confidence interval is also reported.

VI. CONCLUSIONS

In this paper, we showed that the online (proximal-)gradient method exhibits a linear convergence to the optimal value functions within an error, for functions satisfying the (proximal-)PL inequality, and when inexact gradient information is available. We derived bounds in expectation and in high probability, where for the latter we utilized a sub-Weibull model for the gradient errors. The convergence results are applicable to a number of data processing, learning, and feedback-optimization tasks, where the cost functions may not be strongly convex, but satisfies the PL inequality. Our results also provide new insights on the convergence of the (proximal-)gradient method for time-varying functions and with exact gradient information, and for the case of static optimization with inexact gradient information. The gradient error model is general, and it allows one to consider various sources of inaccuracy and gradient estimation techniques.

REFERENCES

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

[2] A. Simonetto, “Time-varying convex optimization via time-varying averaged operators,” 2017, [Online] Available at:https://arxiv.org/abs/1704.07338.

[3] Z. J. Towfic, J. Chen, and A. H. Sayed, “On distributed online classification in the midst of concept drifts,” Neurocomputing, vol. 112, pp. 138–152, 2013.

[4] R. Dixit, A. S. Bedi, R. Tripathi, and K. Rajawat, “Online learning with inexact proximal online gradient descent algorithms,” IEEE Transactions on Signal Processing, vol. 67, no. 5, pp. 1338–1352, 2019.

[5] A. S. Bedi, P. Sarma, and K. Rajawat, “Tracking moving agents via inexact online gradient descent algorithm,” IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 1, pp. 202–217, 2018.

[6] 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.

[7] T. Yang, L. Zhang, R. Jin, and J. Yi, “Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient,” in International Conference on Machine Learning, 2016.

[8] 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.

[9] T.-J. Chang and S. Shahrampour, “On online optimization: Dynamic regret analysis of strongly convex and smooth problems,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, 2021, pp. 6966–6973.

[10] O. Besbes, Y. Gur, and A. Zeevi, “Non-stationary stochastic optimization,” Operations research, vol. 63, no. 5, pp. 1227–1244, 2015.

[11] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, “Connecting the dots: Identifying network structure via graph signal processing,” IEEE Signal Processing Magazine, vol. 36, no. 3, pp. 16–43, 2019.
[38] X. Li and F. Orabona, “A high probability analysis of adaptive SGD with momentum,” arXiv preprint arXiv:2007.14294, 2020.
[39] N. J. Harvey, C. Liaw, and S. Randhawa, “Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent,” arXiv preprint arXiv:1909.00843, 2019.
[40] 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.
[41] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, 1st ed. Cambridge University Press, Sep. 2018.
[42] 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.
[43] M. Gurbuzbalaban, U. Simsekli, and L. Zhu, “The heavy-tail phenomenon in SGD,” arXiv preprint arXiv:2006.04740, 2020.
[44] L. Hodgkinson and M. W. Mahoney, “Multiplicative noise and heavy tails in stochastic optimization,” arXiv preprint arXiv:2006.06293, 2020.
[45] K. C. Wong, Z. Li, A. Tewari et al., “Lasso guarantees for β-mixing heavy-tailed time series,” Annals of Statistics, vol. 48, no. 2, pp. 1124–1142, 2020.
[46] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” SIAM Review, vol. 60, no. 2, pp. 223–311, 2018.
[47] S. Koch, J. L. Mathieu, D. S. Callaway et al., “Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services,” in Proc. PSCC. Citeseer, 2011, pp. 1–7.
[48] E. Dall’Anese and A. Simonetto, “Optimal power flow pursuit,” IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 942–952, March 2018.