Proceedings of Machine Learning Research vol 134:1–42, 2021 34rd Annual Conference on Learning Theory

Parameter-free Stochastic Optimization of Variationally Coherent Functions

author names withheld

Editor: Under Review for COLT 2021

Abstract

We design and analyze an algorithm for first-order stochastic optimization of a large class of functions on $\mathbb{R}^d$. In particular, we consider the *variationally coherent* functions which can be convex or non-convex. The iterates of our algorithm on variationally coherent functions converge almost surely to the global minimizer $\boldsymbol{x}^*$. Additionally, the very same algorithm with the same hyperparameters, after T iterations guarantees on convex functions that the expected suboptimality gap is bounded by $\widetilde{O}(\|\boldsymbol{x}^* - \boldsymbol{x}_0\| T^{-1/2+\epsilon})$ for any $\epsilon > 0$. It is the first algorithm to achieve both these properties at the same time. Also, the rate for convex functions essentially matches the performance of parameter-free algorithms. Our algorithm is an instance of the Follow The Regularized Leader algorithm with the added twist of using *rescaled gradients* and time-varying linearithmic regularizers.

Keywords: parameter-free, stochastic first-order optimization, FTRL, variationally coherent functions

1. Introduction

We consider the problem of finding the minimizer of a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ using access only to noisy gradients of the function. This is a fundamental problem in stochastic optimization and machine learning. Indeed, a plethora of algorithms have been proposed to solve this problem, some of them being optimal with respect to some measure (see, e.g., Bottou et al., 2018). However, the choice of an algorithm crucially depends on the assumptions on the function F.

In an effort to go beyond convex functions, we focus on *variationally coherent* functions. Variationally coherent functions (Zhou et al., 2017, 2020) are defined by the property that, at any point $\boldsymbol{x}$, the vector pointing towards the optimal solution $\boldsymbol{x}^*$ and the negative gradient form an angle of at most 90 degrees. Zhou et al. (2020) proved that this class contains convex, quasi-convex, τ-star-convex (Joulani et al., 2017), and pseudo-convex functions.

Smooth variationally coherent functions can be asymptotically minimized by Stochastic Gradient Descent (SGD) with learning rates proportional to $t^{-\alpha}$ where t is the iteration number and $\alpha > \frac{1}{2}$ (Zhou et al., 2020). If the function happens to be also convex, SGD with the same learning rate guarantees a convergence rate of $O\left(\frac{1+\|\boldsymbol{x}^*-\boldsymbol{x}_0\|^2}{T^{1-\alpha}}\right)$, where $\boldsymbol{x}_0$ is the initial point.

However, for convex functions a significantly better convergence rate $O\left(\frac{1+\|\boldsymbol{x}^*-\boldsymbol{x}_0\|\sqrt{\ln(1+\|\boldsymbol{x}^*-\boldsymbol{x}_0\|)}}{\sqrt{T}}\right)$ can be achieved by using the so-called *parameter-free* algorithms (e.g., Orabona and Pál, 2016; McMahan and Orabona, 2014; Cutkosky and Orabona, 2018) for online convex optimization and averaging their iterates. Specifically, parameter-free algorithms have better dependency on $\|\boldsymbol{x}^* - \boldsymbol{x}_0\|$, that can be arbitrarily large. Unfortunately, parameter-free algorithms are not known to work for non-convex functions.

In this paper, we design a new parameter-free algorithm for convex functions with bounded stochastic gradients that achieves convergence rate $O\left(\frac{1+\|\boldsymbol{x}^*-\boldsymbol{x}_0\|\ln(1+\|\boldsymbol{x}^*-\boldsymbol{x}_0\|)}{T^{1-\alpha}}\right)$ for $\alpha > \frac{1}{2}$ and *at the same time* guarantees asymptotic almost sure convergence to $\boldsymbol{x}^*$ for variationally coherent functions. No averaging of the iterates is required, we can guarantee convergence directly for the last iterate. As far as we know, our algorithm is the first of this kind.

Our algorithm is based on Follow The Regularized Leader (FTRL) algorithm with time-varying linearithmic regularizer and *rescaled gradients*. The regularizer we use is similar to those used in other parameter-free algorithms. Both FTRL and the regularizer are essential to guarantee the better dependency of the convergence rate on $\|\boldsymbol{x}^* - \boldsymbol{x}_0\|$ for convex functions. On the other hand, rescaling of the gradients is needed to guarantee convergence for variationally coherent functions and it is reminiscent of SGD. So our algorithm can be viewed as a novel combination of FTRL and SGD and it might be of independent interest.

The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we formally define the problem and the class of variationally coherent functions. The algorithm and the main results are stated in Section 4. In Section 5, we prove basic properties of FTRL with rescaled gradients. In Section 6, we present the time-varying regularizer and its basic properties. Section 7 contains the proofs of the main results. However, due to space limitations many supporting lemmas and their proofs are deferred to appendices. Finally, in Section 8 we conclude the paper with discussion on limitations and future work.

2. Related Work

Follow The Regularized Leader (FTRL) was introduced as an algorithm for online convex optimization (OCO) on linearized losses by Shalev-Shwartz and Singer (2006, 2007); Shalev-Shwartz (2007). The name of the algorithm comes from Abernethy et al. (2008). For offline optimization, FTRL with linearized losses was introduced under the name Dual Averaging (DA) by[1] Nesterov (2009), with the main motivation of using non-decreasing weights for the (sub)gradients, contrary to Mirror Descent (MD) (Nemirovsky and Yudin, 1983). Hence, even if the general scheme in Nesterov (2009) would support generic weights, DA in the stochastic setting is used with uniform weights. Juditsky et al. (2020) propose another way to merge aspects of MD and DA. Also, they do not allow for generic time-varying regularizers that are essential here.

Parameter-free algorithms for OCO were introduced by Orabona (2013, 2014). However, the seed of these ideas was already present in Streeter and McMahan (2012); McMahan and Abernethy (2013). The same ideas were developed in parallel for the problem of the learning with expert advice (Chaudhuri et al., 2009; Chernov and Vovk, 2010; Luo and Schapire, 2015; Koolen and van Erven, 2015). In fact, the name *parameter-free* originated in Chaudhuri et al. (2009). It is now clear that these two approaches are fundamentally the same (Orabona and Pál, 2016). Our regularizers are inspired to the ones in Koolen and van Erven (2015), but with a max rather than a prior over β (see (14)) that gives simple closed forms.

Parameter-free algorithm can easily be used in the stochastic setting through online-to-batch conversion (Cesa-Bianchi et al., 2002). In the OCO setting, regret of parameter-free algorithms have optimal dependency on $\|\boldsymbol{x}^* - \boldsymbol{x}_0\|$, while regret of online gradient descent has provably suboptimal dependency on $\|\boldsymbol{x}^* - \boldsymbol{x}_0\|$ (Streeter and McMahan, 2012; Cutkosky and Boahen, 2017). It is not known if parameter-free algorithms are optimal for stochastic optimization of convex Lipschitz functions, but it is reasonable to assume that lower bound from Streeter and McMahan (2012) can be extended to the stochastic optimization setting as well. As far as we know, there are no stochastic optimization algorithms that achieves the convergence rate of parameter-free algorithms for convex functions and guarantee convergence for variationally coherent functions.

Zhang (2004); Shamir and Zhang (2013) proved convergence of the last iterate of SGD for convex Lipschitz functions. However, the analysis critically relies on the assumption of bounded domain. Orabona

1. Note that this paper by Nesterov is actually from 2005: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=912637

(2020) proved the convergence of the last iterate of SGD on unbounded domains. We are not aware of other proofs of convergence of last iterate of FTRL-based parameter-free algorithms without changing the update rule (e.g., Cutkosky, 2019a).

Variationally coherent functions were introduced by Zhou et al. (2017, 2020). Zhou et al. (2020) also points out the connection between variationally coherent functions and variational inequalities. For this class of functions, Zhou et al. (2017) proved almost sure convergence for DA, without assuming a unique minimizer, but assuming Lipschitz gradients. The classic analysis of Bottou (1998) used essentially the same definition and proved almost sure convergence of SGD for smooth variationally coherent functions. Both definitions can be traced back to the concept of *pseudogradients* introduced by Polyak and Tsypkin (1973). They defined a pseudogradient of a function F at a point $\boldsymbol{x}$ as any vector $\boldsymbol{g}$ such that $\langle \nabla F(\boldsymbol{x}), \mathbf{E}[\boldsymbol{g}] \rangle \geq 0$. They also introduced the idea of having a pseudogradient of a surrogate objective function. In particular, they considered the surrogate function $\widetilde{F}(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2$, where $\boldsymbol{x}^*$ is the minimizer of F, so that the pseudogradient condition becomes $\langle \mathbf{E}[\boldsymbol{g}], \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$.

3. Problem Setup and Notation

Problem Setup We consider a model in which an algorithm has access to a stochastic first-order oracle for a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$. In each round $t = 1, 2, \ldots$, the algorithm computes an iterate $\boldsymbol{x}_t \in \mathbb{R}^d$. The oracle produces a stochastic gradient $\boldsymbol{g}_t \in \mathbb{R}^d$ such that

$$\mathbf{E}\left[\boldsymbol{g}_t \,\middle|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_{t-1}\right] = \nabla F(\boldsymbol{x}_t) \,. \tag{1}$$

The iterate $\boldsymbol{x}_t$ produced by the algorithm depends on the past gradients $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_{t-1}$ and past iterates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{t-1}$. Thus, even if the algorithm is deterministic, both $\boldsymbol{g}_t$ and $\boldsymbol{x}_t$ are random variables. We denote by $\mathcal{F}_t$ the σ-algebra generated by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_{t-1}$. Using this notation the condition (1) can be written as

$$\mathbf{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_t\right] = \nabla F(\boldsymbol{x}_t) \,.$$

The goal of the algorithm is to approach the minimizer $\boldsymbol{x}^*$ of F, that is,

$$\lim_{t \to \infty} \boldsymbol{x}_t = \boldsymbol{x}^* \qquad \text{almost surely} \,.$$

We make the additional assumption that

$$\|\boldsymbol{g}_t\| \leq G \qquad \text{almost surely} \tag{2}$$

where G is a positive number. The assumption (2) implies that the function F is G-Lipschitz.

We design and analyze an algorithm for two classes of functions. The first class is the class of differentiable convex functions with a minimizer $\boldsymbol{x}^*$ (possibly not unique). The second class consists of variationally coherent functions. Essentially the same class was studied already by Zhou et al. (2017) and Bottou (1998). The class contains non-convex functions, see Zhou et al. (2017) for examples.

Definition 1 (Variatonally coherent function) *A function $F : \mathbb{R}^d \to \mathbb{R}$ is called* variationally coherent *if it is continuously differentiable, has a unique minimizer $\boldsymbol{x}^*$, satisfies*

$$\langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0 \qquad \textit{for all } \boldsymbol{x} \in \mathbb{R}^d \tag{3}$$

and the equality in (3) holds if only if $\boldsymbol{x} = \boldsymbol{x}^*$.

We remark that (3) is satisfied if F is a convex differentiable function with a minimizer $\boldsymbol{x}^*$. Furthermore, any function that is continuously differentiable and strictly convex with a unique minimizer is necessarily variationally coherent.

Notation We denote by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ the standard inner product on $\mathbb{R}^d$ and $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ is the Euclidean norm. *Fenchel conjugate* of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is the function $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ defined as $f^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle - f(\boldsymbol{x})$. *Bregman divergence* associated with a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is the function $B_f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as $B_f(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$. Bregman divergence associated with a convex differentiable function is non-negative. We will use this property throughout the paper.

4. Main Results

We propose the following algorithm for our setting. The algorithm is a Follow The Regularized Leader (FTRL) algorithm with a particular sequence of regularizers operating on the sequence of *rescaled gradients* $\boldsymbol{\ell}_t = \eta_t \boldsymbol{g}_t, t = 1, 2, \ldots$. Section 5 gives a detailed explanation of FTRL. We call the scale factors $\eta_1, \eta_2, \ldots$ *learning rates*.

Algorithm 1 FTRL WITH RESCALED GRADIENTS AND LINEARITHMIC REGULARIZER

Require: Initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, sequence of learning rates $\eta_1, \eta_2, \ldots$

Initialize $S_0^2 = 4, Q_0 = 0, \boldsymbol{\theta}_0 = \mathbf{0}$
for $t = 1, 2, \ldots$ **do**
  Output

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_0 + \begin{cases} \dfrac{\|\boldsymbol{\theta}_{t-1}\|}{2S_{t-1}^2} \exp\left(\dfrac{\|\boldsymbol{\theta}_{t-1}\|^2}{4S_{t-1}^2} - Q_{t-1} \right) & \text{if } \|\boldsymbol{\theta}\| \leq S_{t-1}^2, \\ \dfrac{\boldsymbol{\theta}_{t-1}}{2\|\boldsymbol{\theta}_{t-1}\|} \exp\left(\dfrac{\|\boldsymbol{\theta}_{t-1}\|}{2} - \frac{1}{4}S_{t-1}^2 - Q_{t-1} \right) & \text{if } \|\boldsymbol{\theta}\| > S_{t-1}^2. \end{cases}$$

  Receive stochastic gradient $\boldsymbol{g}_t \in \mathbb{R}^d$ such that $\mathbf{E}[\boldsymbol{g}_t \mid \mathcal{F}_t] = \nabla F(\boldsymbol{x}_t)$ and $\|\boldsymbol{g}_t\| \leq G$
  Compute rescaled gradient $\boldsymbol{\ell}_t \leftarrow \eta_t \boldsymbol{g}_t$
  Update $S_t^2 \leftarrow S_{t-1}^2 + \|\boldsymbol{\ell}_t\|^2$
  Update $Q_t \leftarrow Q_{t-1} + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t}$
  Update sum of negative rescaled gradients $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t$
end for

We prove three results about the algorithm. Their proofs can be found in Section 7 and the proofs of supporting lemmas in Appendices C, D, and E. Theorem 2 states that, under general assumptions on the sequence of learning rates, the iterates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ converge to the minimizer $\boldsymbol{x}^*$ provided that F is variationally coherent. Theorems 3 and 4 are $\widetilde{O}(GT^{-1/2+\epsilon} \|\boldsymbol{x}^* - \boldsymbol{x}_0\|)$ bounds on the speed of convergence of the function values to the optimal value $F(\boldsymbol{x}^*)$ provided that F is convex. Theorem 3 applies to the running average of the iterates. Theorem 4 applies to iterates directly. The last two theorems require a particular sequence of learning rates $\eta_t = \frac{1}{Gt^\alpha}$, where $\alpha \in (\frac{1}{2}, 1)$.

Theorem 2 ($\boldsymbol{x}_t$ converges to $\boldsymbol{x}^*$) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a variationally coherent function with minimizer $\boldsymbol{x}^*$. Assume the stochastic gradients satisfy* (1) *and* (2)*. Assume that learning rate η_t is a non-negative*

$\mathcal{F}_t$-measurable random variable, $t = 1, 2, \dots$, and there exists a real number $\gamma > 0$ such that

$$\sum_{t=1}^{\infty} \eta_t^2 \|\boldsymbol{g}_t\|^2 < \gamma \qquad \textit{almost surely,} \tag{4}$$

$$\sum_{t=1}^{\infty} \eta_t = +\infty \qquad \textit{almost surely,} \tag{5}$$

$$\eta_t \le \frac{1}{G} \qquad \textit{for all } t = 1, 2, \dots \textit{ almost surely.} \tag{6}$$

Then, the sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ generated by Algorithm 1 satisfies

$$\lim_{t \to \infty} \boldsymbol{x}_t = \boldsymbol{x}^* \qquad \textit{almost surely.}$$

The assumption that η_t is $\mathcal{F}_t$-measurable means that η_t is an arbitrary function of $\boldsymbol{g}_1, \boldsymbol{g}_2, \dots \boldsymbol{g}_{t-1}$ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_t$. This way, $\mathbf{E}[\eta_t \boldsymbol{g}_t \mid \mathcal{F}_t] = \eta_t \mathbf{E}[\boldsymbol{g}_t \mid \mathcal{F}_t] = \eta_t \nabla F(\boldsymbol{x}_t)$. Importantly, η_t cannot depend on $\boldsymbol{g}_t$, see also discussion in Li and Orabona (2019). Assumptions (4) and (5) are essentially the same as the assumptions used in the convergence results for stochastic gradient descent algorithm (Robbins and Monro, 1951). Assumption (4) means that $\sum_{t=1}^{\infty} \eta_t^2 \|\boldsymbol{g}_t\|^2$ as a random variable is bounded. Assumption (6) ensures that $\|\boldsymbol{\ell}_t\| \le 1$ which is important for the underlying FTRL algorithm. Note that learning rate $\eta_t = \frac{1}{Gt^\alpha}$, where $\alpha \in (\frac{1}{2}, 1)$, satisfies all these assumptions.

Theorem 3 (Convergence rate of running average for convex functions) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function with a (possibly non-unique) minimizer $\boldsymbol{x}^*$. Let $\alpha \in (\frac{1}{2}, 1)$. Suppose the stochastic gradients satisfy (1) and (2). Algorithm 1 with learning rate $\eta_t = \frac{1}{Gt^\alpha}$ satisfies for all $T \ge 1$*

$$\mathbf{E}\left[F(\overline{\boldsymbol{x}}_T)\right] \le F(\boldsymbol{x}^*) + \frac{G}{T^{1-\alpha}} \left(1 + \sqrt{5 + \frac{1}{2\alpha - 1}} \|\boldsymbol{x}^* - \boldsymbol{x}_0\| \left[2\ln(1 + 2\|\boldsymbol{x}^* - \boldsymbol{x}_0\|) + 9\sqrt{5 + \frac{1}{2\alpha - 1}}\right]\right),$$

where $\overline{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$ is the running average of the iterates.

Theorem 4 (Convergence rate of last iterate for convex functions) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function with a (possibly non-unique) minimizer $\boldsymbol{x}^*$. Let $\alpha \in (\frac{1}{2}, 1)$. Suppose the stochastic gradients satisfy (1) and (2). Algorithm 1 with learning rate $\eta_t = \frac{1}{Gt^\alpha}$ satisfies for all $T \ge 1$*

$$\begin{aligned}\mathbf{E}[F(\boldsymbol{x}_T)] - F(\boldsymbol{x}^*) \le \frac{G}{T^{1-\alpha}} \left(2 + \frac{1}{e(2\alpha - 1)}\right) (&\exp(S) + 3\|\boldsymbol{x}^* - \boldsymbol{x}_0\| \\ &+ 6(S+2)\left(2 + S\|\boldsymbol{x}^* - \boldsymbol{x}_0\|\left[2\ln(1 + 2\|\boldsymbol{x}^* - \boldsymbol{x}_0\|) + 9S\right]\right)) ,\end{aligned}$$

where $S = \sqrt{5 + \frac{1}{2\alpha-1}}$ and $e = 2.718\dots$ is Euler's constant.

Theorem 3 and 4 have the same assumptions. Theorem 4 is obviously stronger, however, its proof is based on the one of Theorem 3. For that reason we include both theorems.

5. FTRL with Rescaled Gradients

As we said, Algorithm 1 is a special case of the FTRL[2] algorithm with *rescaled gradients* stated as Algorithm 2 below. The algorithm differs from the standard FTRL algorithm in that gradients $\boldsymbol{g}_t$ are rescaled by the learning rate η_t. In other words, Algorithm 2 can be viewed as the standard FTRL algorithm operating on the sequence $\boldsymbol{\ell}_t = \eta_t \boldsymbol{g}_t$, $t = 1, 2, \ldots$. As usual, the algorithm is specified by a sequence of functions $\phi_1, \phi_2, \ldots$ called *regularizers*. The regularizer ϕ_t and the learning rate η_t can depend on the previous gradients $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_{t-1}$. This way η_t and ϕ_t are $\mathcal{F}_t$-measurable random elements.

Algorithm 2 FTRL WITH RESCALED GRADIENTS

Require: Initial point $\boldsymbol{x}_0 \in R^d$, regularizers $\phi_1, \phi_2, \ldots$, and learning rates $\eta_1, \eta_2, \ldots$
1: Initialize $\boldsymbol{\theta}_0 = \mathbf{0}$
2: **for** $t = 1, 2, \ldots$ **do**
3: Predict $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_0 + \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} \; \phi_t(\boldsymbol{x}) - \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{x} \rangle$
4: Receive gradient $\boldsymbol{g}_t \in \mathbb{R}^d$
5: Compute rescaled gradient $\boldsymbol{\ell}_t \leftarrow \eta_t \boldsymbol{g}_t$
6: Update sum of negative rescaled gradients $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t$
7: **end for**

In Section 6 we present the sequence of regularizers that gives rise to Algorithm 1 and we make the derivation of the formulas used in algorithm.

Other choices of regularizers and learning rates are also possible. Two simple special cases are worth mentioning. The first special case is the choice $\eta_t = 1$ for all t which recovers standard FTRL algorithm. The second special case is the choice $\phi_t(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{x}\|^2$ for all t which recovers the standard stochastic/online gradient descent algorithm. However, in general, the algorithm is neither FTRL, nor stochastic gradient descent, not even online mirror descent algorithm.

Remark In the following, for simplicity of notation, we assume $\boldsymbol{x}_0 = \mathbf{0} \in \mathbb{R}^d$. It is easy to obtain the results for any other choice of $\boldsymbol{x}_0$ with a simple translation of the coordinate system.

The analysis of both algorithms relies on Lemma 6. The proof of Lemma 6 uses Lemma 5 (Orabona, 2019), we report its proof in Appendix A for completeness. Both lemmas are expressed in terms of the objective function that FTRL minimizes in step t,

$$H_t(\boldsymbol{x}) = \phi_t(\boldsymbol{x}) - \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{x} \rangle = \phi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \langle \boldsymbol{\ell}_i, \boldsymbol{x} \rangle \qquad \text{for } t = 1, 2, \ldots . \tag{7}$$

Lemma 5 (FTRL regret equality) *Let $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_T \in \mathbb{R}^d$ and $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} \; H_t(\boldsymbol{x})$ where $H_t(\boldsymbol{x})$ is defined in (7). Then, for any $\boldsymbol{u} \in \mathbb{R}^d$,*

$$\sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle = \phi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=1}^{T} [H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle]. \tag{8}$$

Lemma 6 (FTRL for stochastic optimization) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be any differentiable function. Suppose that for all $t \geq 1$, the gradient $\boldsymbol{g}_t$ satisfies (1) and (2) and the learning rate η_t is a non-negative $\mathcal{F}_t$-*

2. We prefer to use the name FTRL over DA because FTRL is more general: DA is a special case of FTRL when the losses are linear.

measurable random variable. Assume $\phi_1, \phi_2, \dots$ are convex differentiable and for all $t \geq 1$ satisfy

$$H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle \leq 0 \,. \tag{9}$$

Then, Algorithm 2 satisfies for all $T \geq 0$ and all $\boldsymbol{u} \in \mathbb{R}^d$,

$$\mathbf{E}[B_{\phi_{T+1}}(\boldsymbol{u}, \boldsymbol{x}_t)] + \sum_{t=1}^{T} \mathbf{E}\left[\eta_t \left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right] \leq \mathbf{E}\left[\phi_{T+1}(\boldsymbol{u})\right] - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) \,. \tag{10}$$

In particular,

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t \left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right] \leq \mathbf{E}\left[\phi_{T+1}(\boldsymbol{u})\right] - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) \,. \tag{11}$$

Additionally, if there exists $\boldsymbol{x}^* \in \mathbb{R}^d$ such that $\langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$ then

$$\mathbf{E}[B_{\phi_{T+1}}(\boldsymbol{x}^*, \boldsymbol{x}_t)] \leq \mathbf{E}\left[\phi_{T+1}(\boldsymbol{x}^*)\right] - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) \,. \tag{12}$$

Proof Lemma 5 and the assumption (9) imply that

$$\sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle \leq \phi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) \,. \tag{13}$$

Since $\boldsymbol{x}_{t+1}$ is a minimizer of H_{T+1}, $B_{H_{T+1}}(\boldsymbol{u}, \boldsymbol{x}_t) = H_{T+1}(\boldsymbol{u}) - H_{T+1}(\boldsymbol{x}_{T+1})$. Furthermore, since H_{T+1} and ϕ_{T+1} differ by a linear function, $B_{H_{T+1}}(\boldsymbol{u}, \boldsymbol{x}_t) = B_{\phi_{T+1}}(\boldsymbol{u}, \boldsymbol{x}_{T+1})$. Therefore, (13) is equivalent to

$$B_{\phi_{T+1}}(\boldsymbol{u}, \boldsymbol{x}_{T+1}) + \sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle \leq \phi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) \,.$$

Substituting $\boldsymbol{\ell}_t = \eta_t \boldsymbol{g}_t$ and taking expectation of both sides yields

$$\mathbf{E}[B_{\phi_{T+1}}(\boldsymbol{u}, \boldsymbol{x}_{T+1})] + \sum_{t=1}^{T} \mathbf{E}\left[\eta_t \left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right] \leq \mathbf{E}\left[\phi_{T+1}(\boldsymbol{u})\right] - \min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) \,.$$

We compute $\mathbf{E}\left[\eta_t \left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right]$ as

$$\mathbf{E}\left[\eta_t \left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right] = \mathbf{E}\left[\mathbf{E}[\eta_t \left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \right\rangle \mid \mathcal{F}_t]\right] = \mathbf{E}\left[\eta_t \left\langle \mathbf{E}[\boldsymbol{g}_t \mid \mathcal{F}_t], \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right] = \mathbf{E}[\eta_t \left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \right\rangle]$$

and inequality (10) follows. Inequality (11) follows from (10) and the fact that $B_{\phi_{T+1}}(\cdot, \cdot)$ is non-negative. Inequality (12) follows from (10) and the inequality $\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \geq 0$ which holds by assumption. ■

The assumption (9) might seem strange at first. Generally speaking, the analysis of FTRL with an arbitrary sequence of regularizers boils down to proving an upper bound on $H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle$. By adding a suitable constant to the regularizer ϕ_t, one can ensure that the upper bound is zero.

The surprising fact is that the regularizer ϕ_t we use to construct Algorithm 1 satisfies (9) under the assumption $\|\boldsymbol{\ell}_t\| \leq 1$ (see Lemma 20 in Appendix B) and $\min_{\boldsymbol{x} \in \mathbb{R}^d} \phi_1(\boldsymbol{x}) = \phi_1(\mathbf{0}) = -1$. These kind of regularizers and this proof technique were introduced by Orabona (2013).

The inequality (12) is essential for the proof of the asymptotic convergence for variationally coherent functions. The idea of using this Bregman divergence to guarantee convergence for non-strongly convex functions was pioneered by Dekel et al. (2010).

6. Linearithmic Regularizer

In order to define the sequence of (non-strongly convex) linearithmic regularizers, we define the functions ψ^* and ψ. The function $\psi^* : \mathbb{R} \times (0, \infty) \times [0, \infty) \to \mathbb{R}$ is defined by

$$\psi^*(\theta, S, Q) = \exp\left(\max_{\beta \in \left[-\frac{1}{2}, \frac{1}{2}\right]} \theta\beta - \beta^2 S^2 - Q \right) . \tag{14}$$

The function $\psi : \mathbb{R} \times (0, \infty) \times [0, \infty) \to \mathbb{R}$ is defined as the Fenchel conjugate of ψ^* with respect to the first argument,

$$\psi(x, S, Q) = \sup_{\theta \in \mathbb{R}} \theta x - \psi^*(\theta, S, Q) . \tag{15}$$

The maximum over β in (14) can be removed and replaced with an explicit formula

$$\psi^*(\theta, S, Q) = \begin{cases} \exp\left(\dfrac{\theta^2}{4S^2} - Q \right) & \text{if } |\theta| \le S^2, \\ \exp\left(\dfrac{|\theta|}{2} - \frac{1}{4}S^2 - Q \right) & \text{if } |\theta| > S^2. \end{cases} \tag{16}$$

Lemma 16 in Appendix B lists many properties of ψ^* and ψ, including an explicit formula for ψ of the order of $O(|x|S(\ln(|x| + 1) + Q + S))$. For now, it suffices to say that both $\theta \mapsto \psi^*(\theta, S, Q)$ and $x \mapsto \psi(x, S, Q)$ are even, strictly convex, continuously differentiable, and increasing on $[0, +\infty)$. Furthermore, $\theta \mapsto \psi^*(\theta, S, Q)$ and $x \mapsto \psi(x, S, Q)$ are Fenchel conjugates of each other. Their partial derivatives $\theta \mapsto \frac{\partial \psi^*(\theta,S,Q)}{\partial \theta}$ and $x \mapsto \frac{\partial \psi(x,S,Q)}{\partial x}$ are continuous bijections from $\mathbb{R}$ to $\mathbb{R}$ that are inverses of each other.

Definition of ϕ_t and its properties The regularizer $\phi_t : \mathbb{R}^d \to \mathbb{R}$ is defined in terms ψ as

$$\phi_t(\boldsymbol{x}) = \psi(\|\boldsymbol{x}\|, S_{t-1}, Q_{t-1}) \qquad \text{for } t = 1, 2, \ldots , \tag{17}$$

where S_t and Q_t are defined in Algorithm 1. We also define $\phi_t^* : \mathbb{R}^d \to \mathbb{R}$,

$$\phi_t^*(\boldsymbol{\theta}) = \psi^*(\|\boldsymbol{\theta}\|, S_{t-1}, Q_{t-1}) \qquad \text{for } t = 1, 2, \ldots . \tag{18}$$

Lemma 14 in Appendix B implies that ϕ_t^* and ϕ_t are Fenchel conjugates of each other. Using the properties of ψ and ψ^*, it is easy to verify that both ϕ_t^* and ϕ_t^* are strictly convex and continuously differentiable. The gradient maps $\nabla\phi_t : \mathbb{R}^d \to \mathbb{R}^d$ and $\nabla\phi_t^* : \mathbb{R}^d \to \mathbb{R}^d$ are continuous bijections and inverses of each other.

The sequences $\{S_t\}_{t=1}^\infty$, $\{Q_t\}_{t=1}^\infty$ and $\{\phi_t(\boldsymbol{x})\}_{t=1}^\infty$ are non-decreasing. Under assumption (4) the sequences are bounded and have finite limits S_∞, Q_∞ and $\phi_\infty(\boldsymbol{x}) = \psi(\|\boldsymbol{x}\|, S_\infty, Q_\infty)$ and these limits are bounded random variables; see Lemma 19 in Appendix B.

Explicit formulas We derive Algorithm 1 as a special case of Algorithm 2 with sequence of regularizers defined in (17). According to the the definitions of the algorithms, $\boldsymbol{x}_t = \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} \phi_t(\boldsymbol{x}) - \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{x} \rangle$. Since $\boldsymbol{x}_t$ is a minimizer of $\phi_t(\boldsymbol{x}) - \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{x} \rangle$, it satisfies the first order stationarity condition $\nabla\phi_t(\boldsymbol{x}_t) = \boldsymbol{\theta}_{t-1}$. Since $\nabla\phi_t^*$ and $\nabla\phi_t$ are inverses of each other, $\boldsymbol{x}_t = \nabla\phi_t^*(\boldsymbol{\theta}_{t-1})$. Formulas (16) and (18) give an explicit formula

$$\phi_t^*(\boldsymbol{\theta}) = \begin{cases} \exp\left(\dfrac{\|\boldsymbol{\theta}\|^2}{4S_{t-1}^2} - Q_{t-1} \right) & \text{if } \|\boldsymbol{\theta}\| \le S_{t-1}^2, \\ \exp\left(\dfrac{\|\boldsymbol{\theta}\|}{2} - \frac{1}{4}S_{t-1}^2 - Q_{t-1} \right) & \text{if } \|\boldsymbol{\theta}\| > S_{t-1}^2, \end{cases} \tag{19}$$

from which can compute $\boldsymbol{x}_t = \nabla\phi_t^*(\boldsymbol{\theta}_{t-1})$ and derive the formula on Line 3 of Algorithm 1.

7. Proofs of the Main Results

In this section, we present the proofs of our main results. For a matter of readability, we only present the main and most interesting steps here, leaving the proofs of the technical lemmas to the Appendix. As in Section 5, for simplicity of notation, we assume $\boldsymbol{x}_0 = \mathbf{0} \in \mathbb{R}^d$ and obtain the general results with a simple translation of the coordinate system.

7.1. Proof of Theorem 2

In the proof, we first show that $B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t)$ converges to a finite limit almost surely. Then, we show that this limit is 0. In turn, this will prove the convergence of $\boldsymbol{x}_t$ to $\boldsymbol{x}^*$, even if ϕ_t is not strongly convex. We will need the following two lemmas, the proofs are in Appendix C.

Lemma 7 (Convergence of Bregman divergences) *If $F : \mathbb{R}^d \to \mathbb{R}$ is variationally coherent then there exists a random variable B_∞ such that $\lim_{t\to\infty} B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) = B_\infty < \infty$ almost surely.*

Lemma 8 ($\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$ is squeezed) *There exists two random variables C_1 and C_2 such that with probability one, $0 < C_1 < C_2 < \infty$ and $C_1 \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \le B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) \le C_2 \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$.*

Proof [Proof of Theorem 2] Lemma 6 and $\min_{\boldsymbol{x}\in\mathbb{R}^d} \phi_1(\boldsymbol{x}) = -1$ imply that for any $T \ge 0$,

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle\right] \le 1 + \mathbf{E}\left[\phi_\infty(\boldsymbol{x}^*)\right] \ .$$

Since F is variationally coherent, $\eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle \ge 0$. Monotone convergence theorem implies

$$\mathbf{E}\left[\sum_{t=1}^{\infty} \eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle\right] \le 1 + \mathbf{E}\left[\phi_\infty(\boldsymbol{x}^*)\right] < \infty \,.$$

Therefore,

$$\sum_{t=1}^{\infty} \eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle < \infty \qquad \text{almost surely.} \tag{20}$$

Moreover, Lemma 7 implies that $\lim_{t\to\infty} B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) = B_\infty < \infty$ almost surely.

Now, we claim that $B_\infty = 0$. Clearly, $B_\infty \ge 0$. Suppose by contradiction that B_∞ is strictly positive. Then, there exists a random variable T_0 such that $T_0 < \infty$ almost surely and $\frac{1}{2} B_\infty \le B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) \le 2B_\infty$ for all $t \ge T_0$. So, Lemma 8 implies that $\sqrt{\frac{B_\infty}{2C_2}} \le \|\boldsymbol{x}_t - \boldsymbol{x}^*\| \le \sqrt{\frac{2B_\infty}{C_1}}$ for all $t \ge T_0$. Now, let

$$\delta = \inf\left\{\langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^*\rangle \ : \ \boldsymbol{x} \in \mathbb{R}^d, \sqrt{\frac{B_\infty}{2C_2}} \le \|\boldsymbol{x} - \boldsymbol{x}^*\| \le \sqrt{\frac{2B_\infty}{C_1}}\right\} \,.$$

Since F is continuously differentiable, the function $\boldsymbol{x} \mapsto \langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^*\rangle$ is continuous. The infimum is taken over a compact set $\{\boldsymbol{x} \in \mathbb{R}^d \ : \ \sqrt{\frac{B_\infty}{2C_2}} \le \|\boldsymbol{x}_t - \boldsymbol{x}^*\| \le \sqrt{\frac{2B_\infty}{C_1}}\}$. Therefore, the infimum is attained at some point $\widetilde{\boldsymbol{x}}$ in this set. That is, $\delta = \langle \nabla F(\widetilde{\boldsymbol{x}}), \widetilde{\boldsymbol{x}} - \boldsymbol{x}^*\rangle$. Since $B_\infty > 0$, $\widetilde{\boldsymbol{x}} \ne \boldsymbol{x}^*$ and therefore $\delta > 0$. Thus,

$$\sum_{t=1}^{\infty} \eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle \ge \sum_{t=T_0}^{\infty} \eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\rangle \ge \delta \sum_{t=T_0}^{\infty} \eta_t = \infty \qquad \text{almost surely ,}$$

which contradicts (20). Thus, $\lim_{t\to\infty} B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) = 0$ almost surely. Finally, Lemma 8 implies that $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|$ converges to 0 almost surely as well. ■

7.2. Proof of Theorem 3

Given the results in Lemma 6, the proof of Theorem 3 follows from standard arguments from online convex optimization and online-to-batch conversion (Cesa-Bianchi et al., 2002). We only need a technical lemma to upper bound the values of $\phi_T(\boldsymbol{u})$. Its proof is in Appendix D.

Lemma 9 (Bound on $S_T, Q_T, \phi_T(\boldsymbol{u})$) *Let $\alpha > 1/2$. If $\eta_t = \frac{1}{Gt^\alpha}$, then, for any $T \geq 0$ and any $\boldsymbol{u} \in \mathbb{R}^d$, we have*

$$S_T \leq \sqrt{5 + \frac{1}{2\alpha - 1}}, \qquad Q_T \leq \ln\left(5 + \frac{1}{2\alpha - 1}\right),$$

$$\phi_T(\boldsymbol{u}) \leq \sqrt{5 + \frac{1}{2\alpha - 1}}\, \|\boldsymbol{u}\| \left[2\ln(1 + 2\|\boldsymbol{u}\|) + 9\sqrt{5 + \frac{1}{2\alpha - 1}}\right].$$

Proof [Proof of Theorem 3] Lemma 6 and $\min_{\boldsymbol{x}\in\mathbb{R}^d} \phi_1(\boldsymbol{x}) = -1$ imply that for any $\boldsymbol{u} \in \mathbb{R}^d$,

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u}\rangle\right] \leq 1 + \mathbf{E}\left[\phi_{T+1}(\boldsymbol{u})\right] .$$

Since F is convex, $F(\boldsymbol{x}_t) - F(\boldsymbol{u}) \leq \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u}\rangle$. Substituting $\boldsymbol{x}^*$ for $\boldsymbol{u}$, we have

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))\right] \leq 1 + \mathbf{E}\left[\phi_{T+1}(\boldsymbol{x}^*)\right] .$$

Since $\{\eta_t\}_{t=1}^\infty$ is non-negative decreasing and $F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)$ is non-negative,

$$\eta_T \sum_{t=1}^{T} \mathbf{E}\left[F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)\right] \leq 1 + \mathbf{E}\left[\phi_{T+1}(\boldsymbol{x}^*)\right] .$$

By Jensen's inequality, $F(\overline{\boldsymbol{x}}_T) \leq \frac{1}{T}\sum_{t=1}^T F(\boldsymbol{x}_t)$. Thus, substituting for η_T and using Lemma 9 to upper bound $\phi_{T+1}(\boldsymbol{x}^*)$, we get the stated bound. ■

7.3. Proof of Theorem 4

Here, we prove the convergence of the last iterate, extending the approach of Orabona (2020) to FTRL with rescaled gradients. We need the following Lemmas that are proved in Appendix E.

Lemma 10 *(Orabona, 2020) Let $\eta_1, \eta_2, \ldots, \eta_T$ be a non-increasing sequence of non-negative numbers. Let $q_1, q_2, \ldots, q_T$ be non-negative. Then*

$$\eta_T q_T \leq \frac{1}{T}\sum_{t=1}^{T} \eta_t q_t + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} \eta_t (q_t - q_{T-k}) .$$

Lemma 11 (Difference of regularizers) *Let A, T be integers such that $1 \le A \le T+1$. Then,*

$$\mathbf{E}[\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A)] \le K \sum_{t=A}^{T} t^{-2\alpha}, \tag{21}$$

where $K = \frac{1}{2}\exp\left(\frac{1}{4}S\right) + 3\|\boldsymbol{x}^*\| + 5(S+2)\,\mathbf{E}[2 + \phi_\infty(\boldsymbol{x}^*)]$.

Lemma 12 (Interesting sum) *Let $\alpha > \frac{1}{2}$. Then, $\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-2\alpha} \le \frac{1}{T} + T^{-2\alpha} + \frac{1}{e(2\alpha-1)T}$.*

Lemma 13 (FTRL partial regret bound) *Let $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots, \boldsymbol{\ell}_T \in \mathbb{R}^d$ and $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} H_t(\boldsymbol{x})$ where $H_t(\boldsymbol{x})$ is defined in (7). Assume that for all $t \ge 1$, $H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle \le 0$. Then, for any $A \le T$, we have*

$$\sum_{t=A}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}_A \rangle \le \phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A)\,.$$

Proof [Proof of Theorem 4] Starting from Lemma 13, we substitute $\boldsymbol{\ell}_t = \eta_t \boldsymbol{g}_t$, take expectation of both sides, and use that $\mathbf{E}\left[\eta_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_A \rangle\right] = \mathbf{E}\left[\eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_A \rangle\right]$ by assumption (1). We get

$$\sum_{t=A}^{T} \mathbf{E}\left[\eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_A \rangle\right] \le \mathbf{E}\left[\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A)\right]\,.$$

From Lemma 11 and convexity of F, we obtain

$$\sum_{t=A}^{T} \eta_t \,\mathbf{E}\left[F(\boldsymbol{x}_t) - F(\boldsymbol{x}_A)\right] \le \sum_{t=A}^{T} \mathbf{E}\left[\eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_A \rangle\right] \le \sum_{t=A}^{T} \mathbf{E}\left[\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A)\right] \le K \sum_{t=A}^{T} t^{-2\alpha}. \tag{22}$$

We now apply Lemma 10 with $q_t = \mathbf{E}[F(\boldsymbol{x}_t)] - F(\boldsymbol{x}^*)$ and get

$$\begin{aligned}
\eta_T \,\mathbf{E}[F(\boldsymbol{x}_T) - F(\boldsymbol{x}^*)] &\le \frac{1}{T}\sum_{t=1}^{T} \eta_t \,\mathbf{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)] + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} \eta_t \,\mathbf{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{T-k})] \\
&\le \frac{1}{T}\sum_{t=1}^{T} \eta_t \,\mathbf{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)] + K \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-2\alpha}
\end{aligned}$$

where in second step we used (22) with $A = T-k$. We upper bound the first sum using convexity of F and Lemma 6 as follows

$$\sum_{t=1}^{T} \eta_t \,\mathbf{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)] \le \sum_{t=1}^{T} \eta_t \,\mathbf{E}[\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \rangle] \le 1 + \mathbf{E}[\phi_{T+1}(\boldsymbol{x}^*)]\,.$$

Finally, using Lemma 12, we have

$$\eta_T \,\mathbf{E}[F(\boldsymbol{x}_T) - F(\boldsymbol{x}^*)] \le \frac{1 + \mathbf{E}[\phi_{T+1}(\boldsymbol{x}^*)]}{T} + K\left(\frac{1}{T} + T^{-2\alpha} + \frac{1}{e(2\alpha-1)T}\right)\,.$$

We multiply both sides by $1/\eta_T = GT^\alpha$ and get and use $1/T^{1-\alpha} \ge 1/T^\alpha$ and $\phi_{T+1}(\boldsymbol{x}^*) \le \phi_\infty(\boldsymbol{x}^*)$, to get

$$\mathbf{E}[F(\boldsymbol{x}_T) - F(\boldsymbol{x}^*)] \le G\frac{1 + \mathbf{E}[\phi_\infty(\boldsymbol{x}^*)]}{T^{1-\alpha}} + GK\left(\frac{2}{T^{1-\alpha}} + \frac{1}{e(2\alpha-1)T^{\alpha-1}}\right)\,.$$

Substituting the definition of K, using Lemma 9 to upper bound $\phi_\infty(\boldsymbol{x}^*)$ and over-approximating, we obtain the stated result. ■

8. Discussions on Limitations and Future Work

We have presented the first algorithm that simultaneously achieve the best known convergence rate on convex function with bounded stochastic gradients and also guarantees asymptotic convergence with probability one on variationally coherent functions. In the following, we want to discuss some limitations and possible future directions.

Alternative assumptions Bottou (1998) uses a slightly different set of conditions in the definition of variationally coherent functions. He assumes that for every $\epsilon > 0$ there exists $\delta > 0$ such that $\langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle > \delta$ whenever $\|\boldsymbol{x} - \boldsymbol{x}^*\| > \epsilon$ and drops the condition of continuous differentiability. His assumption is incomparable with ours. Nevertheless, our Theorem 2 would still hold true as is, with a minor modification of its proof. The advantage of Bottou's condition is that Theorem 2 generalizes to (infinite-dimensional) Hilbert spaces, while our argument is based on the compactness of balls in $\mathbb{R}^d$.

Additional adaptivity It is very natural to ask if further adaptivity is possible. For example, one could think to use data-dependent learning rates that depends on the sum of the squared norms of the previous gradients. Indeed, we have an additional result in Appendix F that shows that the function value evaluate on the average iterate would convergence at a rate of $O(1/T)$ if the gradients are deterministic and it would match the convergence of Theorem 3 in the stochastic case. Moreover, the convergence result on variationally coherent functions would still hold! However, we were unable to prove the convergence of the last iterate for these learning rates and we leave it as a future direction of work.

Further applications of FTRL with rescaled gradients We firmly believe that FTRL with Rescaled Gradients might have many more applications that the one presented here. The common knowledge in OCO and optimization literature is that the degree of freedom of choosing the learning rates in OMD corresponds to the degree of freedom to choosing time-varying regularizers in FTRL. However, we have shown here that sometimes *both* degrees of freedom are necessary. Another example is the general form of the recently proposed dual-stabilized OMD (Fang et al., 2020), that with Legendre regularizer can be verified being an instantiation of FTRL with rescaled gradients *and* time-varying regularizers (Proposition H.5, Fang et al., 2020).

The need for bounded stochastic gradients Parameter-free algorithms have a fundamental limitation in the fact that the (stochastic) gradient must be bounded and the bound must be known to the algorithm, due to the lower bound in Cutkosky and Boahen (2017). In the deterministic case, it is enough to use normalized gradients to avoid the knowledge of the bound on the gradients, as explained in Nesterov (2004, Section 3.2.3). Another approach that would work also in the stochastic setting has been proposed by Cutkosky (2019b), that showed that it is possible to avoid the knowledge of the maximum gradient norm, paying an additional $O(\frac{\|\boldsymbol{x}^*\|^3}{T})$ term in the convergence guarantee. Yet, we do not know how to extend parameter-free algorithm to non-Lipschitz function, for example, to smooth functions. Note that in our theorems we proved that $\boldsymbol{x}_t$ is bounded, that would imply a bounded gradient even with smooth functions. Yet, it is unclear how to modify to the current proof to argue that $\boldsymbol{x}_t$ are bounded even in the smooth case. On the other hand, it is important to remember that assuming smoothness is not a weaker assumption than bounded gradients.

Optimality of the results As explained in Section 2, it is unclear if these results are optimal even in the stochastic convex case. We would need a lower bound for stochastic convex optimization with bounded gradients for unbounded domains, that is currently missing. Indeed, all the lower bounds we know assume a bounded domain. We conjecture that a similar lower bound to the one Streeter and McMahan (2012) could be proven for stochastic convex optimization.

References

J. D. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Rocco A. Servedio and Tong Zhang, editors, *Proc. of Conference on Learning Theory (COLT)*, pages 263–274. Omnipress, 2008.

L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. revised, May 2018.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems*, pages 359–366, 2002.

K. Chaudhuri, Y. Freund, and D. J. Hsu. A parameter-free hedging algorithm. In *Advances in neural information processing systems*, pages 297–305, 2009.

A. Chernov and V. Vovk. Prediction with advice of unknown number of experts. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

A. Cutkosky. Anytime online-to-batch, optimism and acceleration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proc. of the 36th International Conference on Machine Learning*, volume 97 of *Proc. of Machine Learning Research*, pages 1446–1454, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

A. Cutkosky. Artificial constraints and hints for unbounded online learning. In A. Beygelzimer and D. Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 874–894. PMLR, 2019b.

A. Cutkosky and K. Boahen. Online learning without prior information. In *Proc. of the 2017 Conference on Learning Theory*, volume 65 of *Proc. of Machine Learning Research*, pages 643–677, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proc. of the Conference on Learning Theory (COLT)*, 2018.

O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *Proc. of the 23rd International Conference on Learning Theory*. MIT Press, 2010.

H. Fang, N. Harvey, V. Portella, and M. Friedlander. Online mirror descent and dual averaging: keeping pace in the dynamic case. In Hal Daumé III and Aarti Singh, editors, *Proc. of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3008–3017. PMLR, 13–18 Jul 2020.

A. Hoorfar and M. Hassani. Inequalities on the Lambert W function and hyperpower function. *J. Inequal. Pure and Appl. Math*, 9(2):5–9, 2008.

P. Joulani, A. György, and C. Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In *Proc. of the International Conference on Algorithmic Learning Theory (ALT)*, volume 76, pages 681–720, 2017.

A. Juditsky, J. Kwon, and É. Moulines. Unifying mirror descent and dual averaging. *arXiv preprint arXiv:1910.13742*, 2020.

W. M. Koolen and T. van Erven. Second-order quantile methods for experts and combinatorial games. In *Proc. of the Conference On Learning Theory (COLT)*, pages 1155–1175, 2015.

X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proc. of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, 2019.

H. Luo and R. E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proc. of the Conference on Learning Theory (COLT)*, pages 1286–1304, 2015.

B. McMahan and J. Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2724–2732. Curran Associates, Inc., 2013.

H. B. McMahan and F. Orabona. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Proc of the Annual Conference on Learning Theory, COLT*, 2014.

A. S. Nemirovsky and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, New York, NY, USA, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1): 221–259, 2009.

F. Orabona. Dimension-free exponentiated gradient. In *Advances in Neural Information Processing Systems 26*, pages 1806–1814. Curran Associates, Inc., 2013.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems 27*, 2014.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

F. Orabona. Last iterate of SGD converges (even in unbounded domains), 2020. Blogpost on `http://parameterfree.com`, available at `https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/`.

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 577–585. Curran Associates, Inc., 2016.

B. Polyak and Y.Z. Tsypkin. Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, 34:377–397, 01 1973.

S. I. Resnick. *A probability path*. Birkhäuser, 1999.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22: 400–407, 1951.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

S. Shalev-Shwartz and Y. Singer. Online learning meets optimization in the dual. In *International Conference on Computational Learning Theory*, pages 423–437. Springer, 2006.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. In *Advances in neural information processing systems*, pages 1265–1272, 2007.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proc. of the International Conference on Machine Learning (ICML)*, volume 28, pages 71–79. PMLR, 2013.

M. Streeter and B. McMahan. No-regret algorithms for unconstrained online convex optimization. In *Advances in Neural Information Processing Systems 25*, pages 2402–2410. Curran Associates, Inc., 2012.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proc. of International Conference on Machine learning*, pages 919–926, New York, NY, USA, 2004. ACM.

Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. W. Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pages 7043–7052, 2017.

Z. Zhou, P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.

Appendix A. Proofs and Lemmas for FTRL with Rescaled Gradients

Proof [Proof of Lemma 5] We cancel $\sum_{t=1}^T \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle$ on both sides of (8). We get an equivalent equation

$$-\sum_{t=1}^T \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle = \phi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in\mathbb{R}^d} \phi_1(\boldsymbol{x}) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=1}^T [H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1})] \,.$$

The sum $\sum_{t=1}^T [H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1})]$ is a telescopic sum equal to $H_1(\boldsymbol{x}_1) - H_{T+1}(\boldsymbol{x}_{T+1})$. Therefore, (8) is equivalent to

$$-\sum_{t=1}^T \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle = \phi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in\mathbb{R}^d} \phi_1(\boldsymbol{x}) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + H_1(\boldsymbol{x}_1) - H_{T+1}(\boldsymbol{x}_{T+1}) \,.$$

We cancel common terms and use the fact $\min_{\boldsymbol{x}\in\mathbb{R}^d} \phi_1(\boldsymbol{x}) = H_1(\boldsymbol{x}_1)$ and we get

$$-\sum_{t=1}^T \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle = \phi_{T+1}(\boldsymbol{u}) - H_{T+1}(\boldsymbol{u}) \,,$$

which holds true by definition of $H_{T+1}(\boldsymbol{u})$. ■

Appendix B. Properties of the Regularizer

Lemma 14 (Fenchel conjugate of a function of $\|\cdot\|$) *Let $f : \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ be even and let f^* be its Fenchel conjugate. Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ be defined as $g(\boldsymbol{x}) = f(\|\boldsymbol{x}\|)$. The Fenchel conjugate of g satisfies $g^*(\boldsymbol{\theta}) = f^*(\|\boldsymbol{\theta}\|)$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$.*

Proof For any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\begin{aligned} g^*(\boldsymbol{\theta}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^d} \langle \boldsymbol{\theta}, \boldsymbol{x}\rangle - g(\boldsymbol{x}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \langle \boldsymbol{\theta}, \boldsymbol{x}\rangle - f(\|\boldsymbol{x}\|) = \sup_{\rho\in[0,\infty)} \sup_{\substack{\boldsymbol{z}\in\mathbb{R}^d \\ \|\boldsymbol{z}\|=1}} \langle \boldsymbol{\theta}, \rho\boldsymbol{z}\rangle - f(\|\rho\boldsymbol{z}\|) \\ &= \sup_{\rho\in[0,\infty)} \sup_{\substack{\boldsymbol{z}\in\mathbb{R}^d \\ \|\boldsymbol{z}\|=1}} \rho \langle \boldsymbol{\theta}, \boldsymbol{z}\rangle - f(\rho) = \sup_{\rho\in[0,\infty)} \rho \|\boldsymbol{\theta}\| - f(\rho) = \sup_{\rho\in\mathbb{R}} |\rho| \, \|\boldsymbol{\theta}\| - f(|\rho|) = \sup_{\rho\in\mathbb{R}} \rho \|\boldsymbol{\theta}\| - f(|\rho|) \\ &= \sup_{\rho\in\mathbb{R}} \rho \|\boldsymbol{\theta}\| - f(\rho) = f^*(\|\boldsymbol{\theta}\|), \end{aligned}$$

where in the second to last equality we used the fact that f is even. ■

Lemma 15 (Properties of Lambert W function) *Let $W : [0, \infty) \to [0, \infty)$ be the inverse of the function $f : [0, \infty) \to [0, \infty)$ for $f(x) = xe^x$. Then, W is a continuous increasing bijection and satisfies*

$$W(0) = 0 \,, \qquad W(xe^x) = x \,, \qquad W(x)e^{W(x)} = x \,.$$

Furthermore, for any $x \in [0, \infty)$,

$$\frac{1}{2} \ln(1 + x) \leq W(x) \leq \ln(1 + x) \,. \tag{23}$$

Proof The function $f : [0, \infty) \to [0, \infty)$, $f(x) = xe^x$, is an increasing continuous bijection. Thus its inverse is an increasing continuous bijection. The properties $W(xe^x) = x$ and $W(x)e^{W(x)} = x$ follow from the definition of inverse. The property $W(0) = 0$ is a special case of $W(xe^x) = 0$ for $x = 0$.

Both inequalities in (23) holds for $x = 0$. It thus suffices to prove them for $x > 0$. The second inequality in (23) is a special case of a more general bound proved by Hoorfar and Hassani (2008),

$$W(x) \leq \ln\left(\frac{x + y}{1 + \log y}\right) \qquad \text{for all } x > -\frac{1}{e} \text{ and all } y > -\frac{1}{e} \,,$$

for $y = 1$.

To prove the first inequality in (23) we start from $W(x)e^{W(x)} = x$. We take logarithm of both sides and we get $W(x) = \ln(x/W(x))$. Using the second inequality in (23), we have

$$W(x) = \ln\left(\frac{x}{W(x)}\right) \geq \ln\left(\frac{x}{\ln(1 + x)}\right) \,.$$

It remains to prove that for $x > 0$

$$\ln\left(\frac{x}{\ln(1 + x)}\right) \geq \frac{1}{2} \ln(1 + x) \,,$$

which is equivalent to

$$x \geq \ln(1 + x)\sqrt{1 + x} \,.$$

The last inequality holds for $x = 0$ with equality. We take derivatives of both sides. It remains to prove

$$1 \ge \frac{1 + \frac{1}{2}\ln(1+x)}{\sqrt{1+x}} \qquad \text{for } x \ge 0.$$

The last inequality is equivalent to

$$\sqrt{1+x} \ge 1 + \ln(\sqrt{1+x}) \qquad \text{for } x \ge 0.$$

Substituting $\sqrt{1+x} = 1 + z$, we need to prove

$$\ln(1+z) \le z \qquad \text{for } z \ge 0.$$

The last inequality is holds for $z = 0$. We take derivative of both sides. It remains to prove

$$\frac{1}{1+z} \le 1 \, ,$$

which clearly holds for all $z \ge 0$. ■

Lemma 16 (Properties of ψ and ψ^*) *The functions $\psi^* : \mathbb{R} \times (0,\infty) \times [0,\infty) \to \mathbb{R}$ and $\psi : \mathbb{R} \times (0,\infty) \times [0,\infty) \to \mathbb{R}$ have the following properties.*

1. *For any $S > 0$ and any $Q \ge 0$, the function $\theta \mapsto \psi^*(\theta, S, Q)$ is positive, even, continuously differentiable, strictly convex, decreasing on $(-\infty, 0]$ and increasing on $[0, +\infty)$.*
2. *For any $S > 0$ and any $Q \ge 0$, the function $x \mapsto \psi(x, S, Q)$ is even, continuously differentiable, strictly convex, decreasing on $(-\infty, 0]$ and increasing on $[0, +\infty)$.*
3. *$\psi(x, S, Q)$ is non-decreasing in its second and third argument.*
4. *If $x \ge 0$, $\frac{\partial \psi(x,S,Q)}{\partial x}$ is non-decreasing in S and Q.*
5. *For any $S' \ge S > 0$, $Q' \ge Q \ge 0$, $x' \ge x \ge 0$,*
$$\psi(x, S', Q') - \psi(x, S, Q) \le \psi(x', S', Q') - \psi(x', S, Q) \, .$$
6. *For any $S \ge 1$, $Q \ge 0$ and any $x > 0$,*
$$\sqrt{\ln(1 + 2x^2)} \le \frac{\partial \psi(x, S, Q)}{\partial x} \le x \max\left\{ 2S^2 \exp(Q),\ 4 + \frac{S^2 + 4Q - 4}{\exp\left(\frac{1}{4}S^2 - Q\right)} \right\} \tag{24}$$
7. *For any $S \ge 1$, $Q \ge 0$ and any $x \in \mathbb{R} \setminus \{0, \pm\frac{1}{2}\exp(\frac{1}{4}S^2 - Q)\}$,*
$$\frac{\min\{2, \sqrt{\ln(1+2x^2)}\}}{|x|\left(\frac{1}{2}S + 1\right)} \le \frac{\partial^2 \psi(x, S, Q)}{\partial x^2} \le \max\left\{ 2S^2 \exp(Q),\ 4\exp\left(Q - \frac{1}{4}S^2\right)\right\} \, . \tag{25}$$

8. *For any $S > 0$, $Q \geq 0$, any $x \in \mathbb{R}$,*

$$\psi(x, S, Q) = \begin{cases} -\exp(-Q) & \text{if } x = 0\,, \\ S\,|x|\,\sqrt{2}\dfrac{W(2\exp(2Q)S^2x^2) - 1}{\sqrt{W(2\exp(2Q)S^2x^2)}} & \text{if } 0 < |x| \leq \dfrac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)\,, \\ 2\,|x|\ln(2\,|x|) + |x|\left(\frac{1}{2}S^2 + Q - 2\right) & \text{if } |x| > \dfrac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)\,, \end{cases} \tag{26}$$

where $W : [0, \infty) \to [0, \infty)$ is the Lambert W-function, i.e., W is the inverse function of $x \mapsto xe^x$.

9. *For any $S \geq 1$, $Q \geq 0$, any $x \in \mathbb{R}$,*

$$\psi(x, S, Q) < S\,|x|\left[2\ln(1 + 2x^2) + 3Q + 3S\right]\,.$$

Proof

1. Equation (16) implies that for any $S > 0$, $Q \geq 0$, the function $\theta \mapsto \psi^*(\theta, S, Q)$ is positive and even. Its derivative is

$$\frac{\partial \psi^*(\theta, S, Q)}{\partial \theta} = \begin{cases} \frac{\theta}{2S^2}\exp\left(\dfrac{\theta^2}{4S^2} - Q\right) & \text{if } |\theta| \leq S^2, \\ \frac{1}{2}\text{sign}(\theta)\exp\left(\dfrac{|\theta|}{2} - \frac{1}{4}S^2 - Q\right) & \text{if } |\theta| > S^2. \end{cases}$$

The function $\theta \mapsto \frac{\partial \psi^*(\theta, S, Q)}{\partial \theta}$ is a continuous odd increasing bijection from $\mathbb{R}$ to $\mathbb{R}$ that is negative on $(-\infty, 0)$ and positive on $(0, +\infty)$. Therefore, $\theta \mapsto \psi^*(\theta, S, Q)$ is strictly convex, decreasing on $(-\infty, 0]$ and increasing on $[0, +\infty)$.

2. Since $\theta \mapsto \psi^*(\theta, S, Q)$ is even,

$$\psi(x, S, Q) = \sup_{\theta \in \mathbb{R}} \theta x - \psi^*(\theta, S, Q) = \sup_{\theta \in \mathbb{R}} \theta x - \psi^*(|\theta|, S, Q) = \sup_{\theta \in \mathbb{R}} |\theta x| - \psi^*(|\theta|, S, Q)\,,$$

and thus, the function $x \mapsto \psi(x, S, Q)$ is even as well.

Since ψ and ψ^* are Fenchel conjugates, the functions $x \mapsto \frac{\partial \psi(x, S, Q)}{\partial x}$ and $\theta \mapsto \frac{\partial \psi^*(\theta, S, Q)}{\partial \theta}$ are functional inverses of one another. We can express $\frac{\partial \psi(x, S, Q)}{\partial x}$ as

$$\frac{\partial \psi(x, S, Q)}{\partial x} = \begin{cases} \text{sign}(x) S\sqrt{2W(2\exp(2Q)S^2x^2)} & \text{if } |x| \leq \dfrac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right), \\ \text{sign}(x)\left(2\ln(2\,|x|) + \frac{1}{2}S^2 + 2Q\right) & \text{if } |x| > \dfrac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right), \end{cases} \tag{27}$$

where $W : [0, \infty) \to [0, \infty)$ is the Lambert function that is the inverse of the function $x \mapsto xe^x$. Using the properties of the Lambert function (Lemma 15), it easy to verify that the function $x \mapsto \frac{\partial \psi(x, S, Q)}{\partial x}$ is a continuous odd increasing bijection from $\mathbb{R}$ to $\mathbb{R}$ that is negative on $(-\infty, 0)$ and positive on $(0, +\infty)$. Therefore $x \mapsto \psi(x, S, Q)$ is strictly convex, decreasing on $(-\infty, 0]$ and increasing on $[0, +\infty)$.

3. Equation (16) implies that $\psi^*(\theta, S, Q)$ is non-increasing both as a function of S and as a function of Q. From the definition (15), we see that $\psi(x, S, Q)$ is non-decreasing both as a function of S and as a function of Q.

4. The equation (27) implies that for $x \geq 0$, the function $S \mapsto \frac{\partial\psi(x,S,Q)}{\partial x}$ is non-decreasing on the interval $(0,\infty)$. Likewise, for $x \geq 0$, the function $Q \mapsto \frac{\partial\psi(x,S,Q)}{\partial x}$ is non-decreasing on the interval $[0,\infty)$.

5. Using the previous property,

$$\begin{aligned}\psi(x,S',Q') - \psi(x,S,Q) &= \psi(0,S',Q') - \psi(0,S,Q) + \int_0^x \frac{\partial\psi(y,S',Q')}{\partial y} - \frac{\partial\psi(y,S,Q)}{\partial y} dy \\ &\leq \psi(0,S',Q') - \psi(0,S,Q) + \int_0^{x'} \frac{\partial\psi(y,S',Q')}{\partial y} - \frac{\partial\psi(y,S,Q)}{\partial y} dy \\ &= \psi(x',S',Q') - \psi(x',S,Q)\,.\end{aligned}$$

6. First, we prove the lower bound. If $0 < x \leq \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$ then

$$\begin{aligned}\frac{\partial\psi(x,S,Q)}{\partial x} &= \sqrt{2S^2 W(2\exp(2Q)S^2x^2)} \\ &\geq \sqrt{2W(2x^2)} && \text{(since } S \geq 1 \text{ and } Q \geq 0) \\ &\geq \sqrt{\ln(1+2x^2)} && \text{(by Lemma 15)}\,.\end{aligned}$$

If $x > \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$ then $\frac{\partial\psi(x,S,Q)}{\partial x} \geq 1$ and therefore

$$\begin{aligned}\frac{\partial\psi(x,S,Q)}{\partial x} &\geq \sqrt{\frac{\partial\psi(x,S,Q)}{\partial x}} \\ &= \sqrt{2\ln(2x) + \frac{1}{2}S^2 + 2Q} \\ &\geq \sqrt{2\ln(2x) + \frac{1}{2}} && \text{(since } S \geq 1 \text{ and } Q \geq 0) \\ &= \sqrt{\ln(4\sqrt{e}x^2)} \\ &\geq \sqrt{\ln(6x^2)} \\ &\geq \sqrt{\ln(1+2x^2)} && \text{(since } x \geq 1/2)\,.\end{aligned}$$

For the upper bound, we study the function $\frac{1}{x}\frac{\partial\psi(x,S,Q)}{\partial x}$. For $0 \leq x \leq \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$, we have that

$$\frac{1}{x}\frac{\partial\psi(x,S,Q)}{\partial x} = \frac{\sqrt{2S^2W(2\exp(2Q)S^2x^2)}}{x} = \sqrt{2S^2 \cdot 2\exp(2Q)S^2}\frac{\sqrt{W(y)}}{\sqrt{y}} \leq 2S^2\exp(Q)$$

where $y = 2\exp(2Q)S^2x^2 \geq 0$ and since $\frac{\sqrt{W(y)}}{\sqrt{y}} \leq 1$ for all $y \geq 0$. In the same way, for $x \geq \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$

$$\frac{1}{x}\frac{\partial\psi(x,S,Q)}{\partial x} = \frac{2\ln(2x) + \frac{1}{2}S^2 + 2Q}{x} \leq \frac{4x - 2 + \frac{1}{2}S^2 + 2Q}{x} \leq 4 + \frac{S^2 + 4Q - 4}{\exp\left(\frac{1}{4}S^2 - Q\right)}\,.$$

The final upper bound is the sum of the upper bounds.

7. Taking derivative of (27), we obtain the second partial derivative,

$$\frac{\partial^2 \psi(x,S,Q)}{\partial x^2} = \begin{cases} \dfrac{\sqrt{2S^2 W(2\exp(2Q)S^2x^2)}}{|x|\left(W(2\exp(2Q)S^2x^2)+1\right)} & \text{if } 0 < |x| \le \dfrac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right), \\ \dfrac{2}{|x|} & \text{if } |x| > \dfrac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right). \end{cases}$$

The lower bound holds, if $|x| > \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$. If $0 < |x| \le \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$, we have

$$\begin{aligned} \frac{\partial^2 \psi(x,S,Q)}{\partial x^2} &= \frac{\sqrt{2S^2 W(2\exp(2Q)S^2x^2)}}{|x|\left(W(2\exp(2Q)S^2x^2)+1\right)} \\ &\ge \frac{\sqrt{2S^2 W(2\exp(2Q)S^2x^2)}}{|x|\left(W(2\exp(2Q)S^2\frac{1}{4}\exp(\frac{1}{2}S^2-2Q))+1\right)} && \text{(since } W(\cdot) \text{ is increasing)} \\ &= \frac{\sqrt{2S^2 W(2\exp(2Q)S^2x^2)}}{|x|\left(W(\frac{1}{2}S^2\exp(\frac{1}{2}S^2))+1\right)} \\ &= \frac{\sqrt{2S^2 W(2\exp(2Q)S^2x^2)}}{|x|\left(\frac{1}{2}S^2+1\right)} \\ &\ge \frac{\sqrt{S^2 \ln(1+2\exp(2Q)S^2x^2)}}{|x|\left(\frac{1}{2}S^2+1\right)} && \text{(by Lemma 15)} \\ &\ge \frac{S\sqrt{\ln(1+2x^2)}}{|x|\left(\frac{1}{2}S^2+1\right)} && \text{(since } S\ge 1 \text{ and } Q \ge 0) \\ &\ge \frac{\sqrt{\ln(1+2x^2)}}{|x|\left(\frac{1}{2}S+1\right)} && \text{(since } S \ge 1)\,. \end{aligned}$$

For the upper bound, if $0 < x \le \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$, we have

$$\begin{aligned} \frac{\partial^2 \psi(x,S,Q)}{\partial x^2} &= \frac{\sqrt{2S^2 W(2\exp(2Q)S^2x^2)}}{|x|\left(W(2\exp(2Q)S^2x^2)+1\right)} = \sqrt{2S^2\cdot 2\exp(2Q)S^2}\frac{\sqrt{W(y)}}{\sqrt{y}(W(y)+1)} \\ &= 2S\exp(2Q)\frac{\sqrt{W(y)}}{\sqrt{y}(W(y)+1)}\,, \end{aligned}$$

where $y = 2\exp(2Q)S^2x^2$. It is possible to verify that $\frac{\sqrt{W(y)}}{\sqrt{y}(W(y)+1)} \le 1$ for $y \ge 0$. Hence, the first expression of the max follows. The second expression is immediate when $x \ge \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$.

8. We compute $\psi(0,S,Q)$ as

$$\psi(0,S,Q) = \sup_{\theta\in\mathbb{R}} -\psi^*(\theta,S,Q) = -\inf_{\theta\in\mathbb{R}} \psi^*(\theta,S,Q) = -\psi^*(0,S,Q) = -\exp(-Q)\,.$$

If $0 < |x| \le \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$, from (27) and the fact $x \mapsto \psi(x, S, Q)$ is even and continuous, we have that

$$\begin{aligned}
\psi(x,S,Q) &= \psi(0,S,Q) + \int_0^{|x|} \frac{\partial \psi(y,S,Q)}{\partial y} dy = \psi(0,S,Q) + \int_0^{|x|} \sqrt{2S^2 W(2\exp(2Q)S^2y^2)}dy \\
&= \psi(0,S,Q) + S\,|x|\,\sqrt{2}\frac{W(2\exp(2Q)S^2x^2) - 1}{\sqrt{W(2\exp(2Q)S^2x^2)}} + \frac{\sqrt{2S^2}}{\sqrt{2\exp(2Q)S^2}} \\
&= S\,|x|\,\sqrt{2}\frac{W(2\exp(2Q)S^2x^2) - 1}{\sqrt{W(2\exp(2Q)S^2x^2)}} .
\end{aligned}$$

If $|x| > \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$, let $x_0 = \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$. From (27) and the fact $x \mapsto \psi(x, S, Q)$ is even and continuous, we have that

$$\begin{aligned}
\psi(x,S,Q) &= \psi(x_0,S,Q) + \int_{x_0}^{|x|} \frac{\partial \psi(y,S,Q)}{\partial y} dy = \psi(x_0,S,Q) + \int_{x_0}^{|x|} 2\ln(2y) + \frac{1}{2}S^2 + 2Q dy \\
&= \psi(x_0,S,Q) + 2|x|\ln(2\,|x|) - 2x_0\ln(2x_0) + \left(\frac{1}{2}S^2 + 2Q - 2\right)(|x| - x_0) \\
&= \psi(x_0,S,Q) + 2|x|\ln(2\,|x|) - 2x_0\left(\frac{1}{4}S^2 - Q\right) + \left(\frac{1}{2}S^2 + 2Q - 2\right)(|x| - x_0) .
\end{aligned}$$

We express $\psi(x_0, S, Q)$ as

$$\begin{aligned}
\psi(x_0,S,Q) &= Sx_0\sqrt{2}\frac{W(2\exp(2Q)S^2x_0^2) - 1}{\sqrt{W(2\exp(2Q)S^2x_0^2)}} = Sx_0\sqrt{2}\frac{W(2\exp(2Q)S^2\frac{1}{4}\exp(\frac{1}{2}S^2 - 2Q)) - 1}{\sqrt{W(2\exp(2Q)S^2\frac{1}{4}\exp(\frac{1}{2}S^2 - 2Q))}} \\
&= Sx_0\sqrt{2}\frac{W(\frac{1}{2}S^2\exp(\frac{1}{2}S^2)) - 1}{\sqrt{W(\frac{1}{2}S^2\exp(\frac{1}{2}S^2))}} = Sx_0\sqrt{2}\frac{\frac{1}{2}S^2 - 1}{\sqrt{\frac{1}{2}S^2}} = 2x_0\left(\frac{1}{2}S^2 - 1\right) .
\end{aligned}$$

Hence, if $|x| > \frac{1}{2}\exp\left(\frac{1}{4}S^2 - Q\right)$, we have

$$\begin{aligned}
\psi(x,S,Q) &= 2x_0\left(\frac{1}{2}S^2 - 1\right) + 2|x|\ln(2\,|x|) - 2x_0\left(\frac{1}{4}S^2 - Q\right) + \left(\frac{1}{2}S^2 + 2Q - 2\right)(|x| - x_0) \\
&= 2|x|\ln(2\,|x|) + |x|\left(\frac{1}{2}S^2 + 2Q - 2\right) .
\end{aligned}$$

9. If $|x| \le \frac{1}{2}\exp(\frac{1}{4}S^2 - Q)$, then

$$\begin{aligned}
\psi(x,S,Q) &= S\,|x|\,\sqrt{2}\frac{W(2\exp(2Q)S^2x^2) - 1}{\sqrt{W(2\exp(2Q)S^2x^2)}} \\
&\le S\,|x|\,\sqrt{2}W(2\exp(2Q)S^2x^2) && \left(\text{since } \tfrac{W-1}{\sqrt{W}} \le W \text{ for } W \ge 0\right) \\
&\le S\,|x|\,\sqrt{2}\ln(1 + 2\exp(2Q)S^2x^2) && (\text{Lemma 15}) \\
&\le S\,|x|\,\sqrt{2}\ln(\exp(2Q)S^2 + 2\exp(2Q)S^2x^2) && (\text{since } S \ge 1 \text{ and } Q \ge 0) \\
&\le S\,|x|\,\sqrt{2}\left[\ln(1 + 2x^2) + \ln(\exp(2Q)S^2)\right] \\
&= S\,|x|\,\sqrt{2}\left[\ln(1 + 2x^2) + 2Q + 2\ln S\right] \\
&\le S\,|x|\,\sqrt{2}\left[\ln(1 + 2x^2) + 2Q + 2S\right] \\
&< S\,|x|\left[2\ln(1 + 2x^2) + 3Q + 3S\right] .
\end{aligned}$$

If $|x| > \frac{1}{2}\exp(\frac{1}{4}S^2 - Q)$, then

$$\begin{aligned}
\psi(x, S, Q) &= 2\,|x| \ln(2\,|x|) + |x| \left(\frac{1}{2}S^2 + Q - 2\right) \le 2S\,|x| \ln(2\,|x|) + |x| \left(\frac{1}{2}S^2 + Q - 2\right) \\
&\le 2S\,|x| \ln(1 + 2x^2) + |x| \left(\frac{1}{2}S^2 + Q - 2\right) \le 2S\,|x| \ln(1 + 2x^2) + |x| \left(\frac{1}{2}S^2 + SQ\right) \\
&= S\,|x| \left[2\ln(1 + 2x^2) + \frac{1}{2}S + Q\right] < S\,|x| \left[2\ln(1 + 2x^2) + 3S + 3Q\right] .
\end{aligned}$$

■

Lemma 17 (Useful inequality) *Let $a_0 \in \mathbb{R}$ and let $a_1, a_2, \ldots, a_T \in [0, \infty)$. Let $f : [a_0, \sum_{t=0}^T a_t] \to \mathbb{R}$ be a non-increasing function. Then,*

$$\sum_{t=1}^T a_t f\left(a_0 + \sum_{i=1}^t a_i\right) \le \int_{a_0}^{\sum_{t=0}^T a_t} f(x) dx .$$

Proof Denote by $s_t = \sum_{i=0}^t a_i$ for $t = 0, 1, 2, \ldots, T$ and note that $s_0 \le s_1 \le \cdots \le s_T$. Then, for any $t = 1, 2, \ldots, T$,

$$a_t f\left(a_0 + \sum_{i=1}^t a_i\right) = a_t f(s_t) = \int_{s_{t-1}}^{s_t} f(s_t) dx \le \int_{s_{t-1}}^{s_t} f(x) dx .$$

Summing over $t = 1, 2, \ldots, T$, we have the stated bound. ■

Lemma 18 (Bound on Q_T) *For any $T \ge 0$, $Q_T \le 2\ln S_T$.*

Proof We use Lemma 17 with $f(x) = 1/x$ and $a_0 = 4$ and $a_t = \|\boldsymbol{\ell}_t\|^2$. We have

$$Q_T = \sum_{t=1}^T a_t f\left(a_0 + \sum_{i=1}^t a_i\right) \le \int_{a_0}^{\sum_{t=0}^T a_t} f(x) dx = \int_4^{S_T^2} f(x) dx \le \int_1^{S_T^2} f(x) dx = 2\ln S_T .$$

■

Lemma 19 (Limits $S_\infty, Q_\infty, \psi_\infty$) *For any $\boldsymbol{x} \in \mathbb{R}^d$, the sequences $\{S_t\}_{t=0}^\infty$, $\{Q_t\}_{t=0}^\infty$, $\{\phi_t(\boldsymbol{x})\}_{t=1}^\infty$ are non-decreasing. Furthermore, the assumption (4) implies that the sequences have finite limits*

$$\begin{aligned}
S_\infty &= \lim_{t\to\infty} S_t && \textit{almost surely} , \\
Q_\infty &= \lim_{t\to\infty} Q_t && \textit{almost surely} , \\
\phi_\infty(\boldsymbol{x}) &= \lim_{t\to\infty} \phi_t(\boldsymbol{x}) = \psi\left(\|\boldsymbol{x}\|, S_\infty, Q_\infty\right) && \textit{almost surely}
\end{aligned}$$

and $S_\infty, Q_\infty, \phi_\infty(\boldsymbol{x})$ are bounded random variables.

Proof According to the definition of Algorithm 1,

$$S_t = \sqrt{4 + \sum_{i=1}^{t} \|\boldsymbol{\ell}_i\|^2} = \sqrt{4 + \sum_{i=1}^{t} \eta_i^2 \|\boldsymbol{g}_i\|^2}\,, \tag{28}$$

$$Q_t = \sum_{i=1}^{t} \frac{\|\boldsymbol{\ell}_i\|^2}{S_i^2} = \sum_{i=1}^{t} \frac{\eta_i^2 \|\boldsymbol{g}_i\|^2}{S_i^2}\,. \tag{29}$$

Clearly, the sequences $\{S_t\}_{t=0}^{\infty}$, $\{Q_t\}_{t=0}^{\infty}$ are non-decreasing and satisfy $S_t \ge 2$ and $Q_t \ge 0$. Assumption (4) implies that $S_t < \sqrt{4+\gamma}$. Therefore, the limit S_∞ exists, is finite and $2 \le S_\infty \le \sqrt{4+\gamma}$. Thus, the random variable S_∞ is bounded. By Lemma 18, $Q_t \le 2 \ln S_t < \ln(4+\gamma)$. Therefore, the limit Q_∞ exists, finite, and $0 \le Q_\infty \le \ln(4+\gamma)$. Thus, the random variable Q_∞ is bounded.

By Lemma 16, $\psi(\|\boldsymbol{x}\|, S_{t-1}, Q_{t-1}) \le \psi(\|\boldsymbol{x}\|, S_t, Q_t)$. In other words, $\{\phi_t(\boldsymbol{x})\}_{t=1}^{\infty}$ is a non-decreasing sequence. Lemma 16 also implies that $\psi(x, S_\infty, Q_\infty)$ is continuous as function on $\mathbb{R} \times (0,\infty) \times [0,\infty)$. Therefore,

$$\lim_{t\to\infty} \phi_t(\boldsymbol{x}) = \lim_{t\to\infty} \psi(\|\boldsymbol{x}\|, S_{t-1}, Q_{t-1}) = \psi\left(\|\boldsymbol{x}\|, \lim_{t\to\infty} S_{t-1}, \lim_{t\to\infty} Q_{t-1}\right) = \psi(\|\boldsymbol{x}\|, S_\infty, Q_\infty)\,.$$

Lemma 16 also implies that

$$\psi(\|\boldsymbol{x}\|, 2, 0) \le \psi(\|\boldsymbol{x}\|, S_\infty, Q_\infty) \le \psi\left(\|\boldsymbol{x}\|, \sqrt{4+\gamma}, \ln(4+\gamma)\right)\,.$$

Therefore, the random variable $\phi_\infty(\boldsymbol{x}) = \psi(\|\boldsymbol{x}\|, S_\infty, Q_\infty)$ is bounded. ■

For simplicity, we prove the next lemma using first principles, but it is also possible to observe that β_t itself is the output of a certain FTRL algorithm over a constrained set with strongly convex losses.

Lemma 20 (Key inequality) *Let H_t be defined by (7) and ψ_t be defined by (17). If $\|\boldsymbol{\ell}_t\| \le 1$ then*

$$H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle \le 0\,. \tag{30}$$

Proof Since ϕ_t^* is the Fenchel conjugate of ϕ_t, $H_t(\boldsymbol{x}_t) = -\phi_t^*(\boldsymbol{\theta}_{t-1})$, $\phi_{t+1}^*(\boldsymbol{\theta}_t) = -\phi_t^*(\boldsymbol{\theta}_{t-1})$ and $\boldsymbol{x}_t = \nabla\phi_t^*(\boldsymbol{\theta}_{t-1})$. Therefore, the left-hand side of (30) equals to

$$\phi_{t+1}^*(\boldsymbol{\theta}_t) - \phi_t^*(\boldsymbol{\theta}_{t-1}) + \langle \boldsymbol{\ell}_t, \nabla\phi_t^*(\boldsymbol{\theta}_{t-1}) \rangle\,.$$

Since

$$\phi_t^*(\boldsymbol{\theta}_{t-1}) = \begin{cases} \exp\left(\frac{\|\boldsymbol{\theta}\|^2}{4S_{t-1}^2} - Q_{t-1}\right) & \text{if } \|\boldsymbol{\theta}\| \le S_{t-1}^2, \\ \exp\left(\frac{\|\boldsymbol{\theta}\|}{2} - \frac{1}{4}S_{t-1}^2 - Q_{t-1}\right) & \text{if } \|\boldsymbol{\theta}\| > S_{t-1}^2, \end{cases},$$

the gradient $\nabla\phi_t^*(\boldsymbol{\theta}_{t-1})$ can be expressed as

$$\nabla\phi_t^*(\boldsymbol{\theta}_{t-1}) = \boldsymbol{\beta}_t \phi_t^*(\boldsymbol{\theta}_{t-1})$$

where

$$\boldsymbol{\beta}_t = \begin{cases} \dfrac{\boldsymbol{\theta}_{t-1}}{2S_{t-1}^2} & \text{if } \|\boldsymbol{\theta}_{t-1}\| \le S_{t-1}^2, \\ \dfrac{\boldsymbol{\theta}_{t-1}}{2\|\boldsymbol{\theta}_{t-1}\|} & \text{if } \|\boldsymbol{\theta}_{t-1}\| > S_{t-1}^2. \end{cases} \tag{31}$$

Therefore, the left-hand side of (30) equals to

$$\phi_{t+1}^*(\boldsymbol{\theta}_t) - \phi_t^*(\boldsymbol{\theta}_{t-1})(1 - \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle) .$$

Since $\|\boldsymbol{\beta}_t\| \le 1/2$ and $\|\boldsymbol{\ell}_t\| \le 1$ and therefore $\langle \boldsymbol{\beta}_t, \boldsymbol{\ell}_t \rangle \in [-\frac{1}{2}, \frac{1}{2}]$. Since $1 - x \ge \exp(-x - x^2)$ for any $x \in [-\frac{1}{2}, \frac{1}{2}]$ and $\phi_t^*(\cdot)$ is non-negative, the last expression can be upper bounded as

$$\begin{aligned} \phi_{t+1}^*(\boldsymbol{\theta}_t) - \phi_t^*(\boldsymbol{\theta}_{t-1})(1 - \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle) &\le \phi_{t+1}^*(\boldsymbol{\theta}_t) - \phi_t^*(\boldsymbol{\theta}_{t-1}) \exp\left(- \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle - (\langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle)^2\right) \\ &\le \phi_{t+1}^*(\boldsymbol{\theta}_t) - \phi_t^*(\boldsymbol{\theta}_{t-1}) \exp\left(- \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\ell}_t\|^2 \|\boldsymbol{\beta}_t\|^2\right) . \end{aligned}$$

It remains to show that

$$\phi_{t+1}^*(\boldsymbol{\theta}_t) - \phi_t^*(\boldsymbol{\theta}_{t-1}) \exp\left(- \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\ell}_t\|^2 \|\boldsymbol{\beta}_t\|^2\right) \le 0 ,$$

which is equivalent to

$$\phi_{t+1}^*(\boldsymbol{\theta}_t) \le \phi_t^*(\boldsymbol{\theta}_{t-1}) \exp\left(- \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 \|\boldsymbol{\ell}_t\|^2\right) .$$

We express $\phi_{t+1}^*(\boldsymbol{\theta}_t)$ and $\phi_t^*(\boldsymbol{\theta}_{t-1})$ using an explicit formula

$$\phi_t^*(\boldsymbol{\theta}_{t-1}) = \exp\left(\langle \boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 S_{t-1}^2 - Q_{t-1}\right) ,$$

where $\boldsymbol{\beta}_t$ is defined by (31). We take logarithm of both sides and get an equivalent inequality

$$\left\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} \right\rangle - \left\|\boldsymbol{\beta}_{t+1}\right\|^2 S_t^2 - Q_t \le \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 S_{t-1}^2 - Q_{t-1} - \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\ell}_t\|^2 \|\boldsymbol{\beta}_t\|^2 .$$

Using the definition Q_t and Q_{t-1}, this is equivalent to

$$\left\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} \right\rangle - \left\|\boldsymbol{\beta}_{t+1}\right\|^2 S_t^2 \le \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 S_{t-1}^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} - \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\ell}_t\|^2 \|\boldsymbol{\beta}_t\|^2 .$$

Using the definition of S_t and S_{t-1}, this is equivalent to

$$\left\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} \right\rangle - \left\|\boldsymbol{\beta}_{t+1}\right\|^2 S_t^2 \le \langle \boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 S_t^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} - \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle .$$

Using the definition of $\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\theta}_t$, this is equivalent to

$$\left\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} \right\rangle - \left\|\boldsymbol{\beta}_{t+1}\right\|^2 S_t^2 \le \langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 S_t^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} . \tag{32}$$

We prove (32) by considering several cases.

Case $\|\boldsymbol{\theta}_{t-1}\| > S_{t-1}^2$, $\|\boldsymbol{\theta}_t\| > S_t^2$. In this case, $\boldsymbol{\beta}_t = \frac{\boldsymbol{\theta}_{t-1}}{2\|\boldsymbol{\theta}_{t-1}\|}$, $\boldsymbol{\beta}_{t+1} = \frac{\boldsymbol{\theta}_t}{2\|\boldsymbol{\theta}_t\|}$ and $\|\boldsymbol{\beta}_t\| = \|\boldsymbol{\beta}_{t+1}\| = 1/2$. The inequality (32) becomes

$$\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} \rangle - \frac{1}{4} S_t^2 \leq \langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_t \rangle - \frac{1}{4} S_t^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} .$$

Multiplying by 2, cancelling common terms and rearranging terms, we get an equivalent inequality

$$\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t \rangle \leq \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} .$$

The last inequality follows since

$$\begin{aligned}
\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t \rangle &= \frac{1}{2} \left\langle \boldsymbol{\theta}_t, \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|} - \frac{\boldsymbol{\theta}_{t-1}}{\|\boldsymbol{\theta}_{t-1}\|} \right\rangle \\
&= \frac{1}{2} \left\langle \boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t, \frac{\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t}{\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|} - \frac{\boldsymbol{\theta}_{t-1}}{\|\boldsymbol{\theta}_{t-1}\|} \right\rangle \\
&= \frac{1}{2} \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\| - \frac{1}{2} \|\boldsymbol{\theta}_{t-1}\| + \frac{1}{2} \left\langle \boldsymbol{\ell}_t, \frac{\boldsymbol{\theta}_{t-1}}{\|\boldsymbol{\theta}_{t-1}\|} \right\rangle \\
&= \frac{1}{2} \frac{\langle \boldsymbol{\ell}_t, \boldsymbol{\theta}_{t-1} \rangle}{\|\boldsymbol{\theta}_{t-1}\|} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2 - 2 \langle \boldsymbol{\ell}_t, \boldsymbol{\theta}_{t-1} \rangle}{\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|} \\
&= \frac{1}{2} \langle \boldsymbol{\ell}_t, \boldsymbol{\theta}_{t-1} \rangle \frac{\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\| - \|\boldsymbol{\theta}_t\|}{\|\boldsymbol{\theta}_{t-1}\| (\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|)} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|} \\
&\leq \frac{1}{2} |\langle \boldsymbol{\ell}_t, \boldsymbol{\theta}_{t-1} \rangle| \frac{|\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\| - \|\boldsymbol{\theta}_t\||}{\|\boldsymbol{\theta}_{t-1}\| (\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|)} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|} \\
&\leq \frac{1}{2} \|\boldsymbol{\ell}_t\| \|\boldsymbol{\theta}_{t-1}\| \frac{\|\boldsymbol{\ell}_t\|}{\|\boldsymbol{\theta}_{t-1}\| (\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|)} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|} \\
&= \frac{\|\boldsymbol{\ell}_t\|^2}{\|\boldsymbol{\theta}_{t-1}\| + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t\|} \\
&\leq \frac{\|\boldsymbol{\ell}_t\|^2}{S_{t-1}^2 + S_t^2} \\
&\leq \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} ,
\end{aligned}$$

where we substituted for $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t+1}$, used that, by definition, $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t$, made some algebraic manipulation, used triangle inequality in the form $|\|\boldsymbol{a}\| - \|\boldsymbol{b}\|| \leq \|\boldsymbol{a} - \boldsymbol{b}\|$, Cauchy-Schwarz inequality and the assumptions $\|\boldsymbol{\theta}_t\| > S_t^2$, $\|\boldsymbol{\theta}_{t-1}\| > S_{t-1}^2$ and, finally, used that $S_{t-1}^2 > 0$.

Case $\|\boldsymbol{\theta}_{t-1}\| \leq S_{t-1}^2$. In this case, we have $\boldsymbol{\beta}_t = \frac{\boldsymbol{\theta}_{t-1}}{2S_{t-1}^2}$. Let

$$\boldsymbol{\beta}'_{t+1} = \frac{\boldsymbol{\theta}_t}{2S_t^2} = \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^d} \langle \boldsymbol{\theta}_t, \boldsymbol{\beta} \rangle - \|\boldsymbol{\beta}\|^2 S_t^2 .$$

The left-hand side of (32) is upper bounded as

$$\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_{t+1} \rangle - \|\boldsymbol{\beta}_{t+1}\|^2 S_t^2 \leq \langle \boldsymbol{\theta}_t, \boldsymbol{\beta}'_{t+1} \rangle - \|\boldsymbol{\beta}'_{t+1}\|^2 S_t^2 ,$$

since $\boldsymbol{\beta}_{t+1}$ and $\boldsymbol{\beta}'_{t+1}$ are the constrained and unconstrained maximizers of $\boldsymbol{\beta} \mapsto \langle \boldsymbol{\theta}_t, \boldsymbol{\beta} \rangle - \|\boldsymbol{\beta}\|^2 S_t^2$ respectively. Thus it suffices to prove

$$\langle \boldsymbol{\theta}_t, \boldsymbol{\beta}'_{t+1} \rangle - \left\|\boldsymbol{\beta}'_{t+1}\right\|^2 S_t^2 \le \langle \boldsymbol{\theta}_t, \boldsymbol{\beta}_t \rangle - \|\boldsymbol{\beta}_t\|^2 S_t^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \, .$$

Substituting $\boldsymbol{\theta}_t = 2S_t^2 \boldsymbol{\beta}'_{t+1}$, we get

$$2S_t^2 \left\|\boldsymbol{\beta}'_{t+1}\right\|^2 - \left\|\boldsymbol{\beta}'_{t+1}\right\|^2 S_t^2 \le 2S_t^2 \left\langle \boldsymbol{\beta}'_{t+1}, \boldsymbol{\beta}_t \right\rangle - \|\boldsymbol{\beta}_t\|^2 S_t^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \, ,$$

which is equivalent to

$$2S_t^2 \left\|\boldsymbol{\beta}'_{t+1} - \boldsymbol{\beta}_t\right\|^2 \le \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \, .$$

The last inequality follows from

$$\begin{aligned}
\left\|\boldsymbol{\beta}'_{t+1} - \boldsymbol{\beta}_t\right\|^2 &= \left\| \frac{\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t}{2S_t^2} - \frac{\boldsymbol{\theta}_{t-1}}{2S_{t-1}^2} \right\|^2 \\
&= \frac{1}{4} \left\| \frac{\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t}{S_t^2} - \frac{\boldsymbol{\theta}_{t-1}}{S_{t-1}^2} \right\|^2 \\
&= \frac{1}{4} \left\| \boldsymbol{\theta}_{t-1} \left(\frac{1}{S_t^2} - \frac{1}{S_{t-1}^2} \right) - \frac{\boldsymbol{\ell}_t}{S_t^2} \right\|^2 \\
&\le \frac{1}{4} \left(\left\| \boldsymbol{\theta}_{t-1} \left(\frac{1}{S_t^2} - \frac{1}{S_{t-1}^2} \right) \right\| + \frac{\|\boldsymbol{\ell}_t\|}{S_t^2} \right)^2 \\
&= \frac{1}{4} \left(\|\boldsymbol{\theta}_{t-1}\| \left(\frac{1}{S_{t-1}^2} - \frac{1}{S_t^2} \right) + \frac{\|\boldsymbol{\ell}_t\|}{S_t^2} \right)^2 \\
&\le \frac{1}{2} \|\boldsymbol{\theta}_{t-1}\|^2 \left(\frac{1}{S_{t-1}^2} - \frac{1}{S_t^2} \right)^2 + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&= \frac{1}{2} \|\boldsymbol{\theta}_{t-1}\|^2 \frac{(S_t^2 - S_{t-1}^2)^2}{S_{t-1}^4 S_t^4} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&= \frac{1}{2} \|\boldsymbol{\theta}_{t-1}\|^2 \frac{\|\boldsymbol{\ell}_t\|^4}{S_{t-1}^4 S_t^4} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&\le \frac{1}{2} S_{t-1}^4 \frac{\|\boldsymbol{\ell}_t\|^4}{S_{t-1}^4 S_t^4} + \frac{1}{2} \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&= \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \, ,
\end{aligned}$$

where we substituted for $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}'_{t+1}$, used that $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t$, used triangle inequality, the inequality $0 < S_{t-1} \le S_t$, the inequality $(a+b)^2 \le 2a^2 + 2b^2$, the fact $S_t = S_{t-1} + \|\boldsymbol{\ell}_t\|^2$ and, finally, the inequality $\|\boldsymbol{\theta}_{t-1}\| \le S_{t-1}^2$.

Case $\|\boldsymbol{\theta}_{t-1}\| > S_{t-1}^2$ and $\|\boldsymbol{\theta}_t\| \le S_t^2$. In this case, we have $\boldsymbol{\beta}_t = \frac{\boldsymbol{\theta}_{t-1}}{2\|\boldsymbol{\theta}_{t-1}\|}$ and $\boldsymbol{\beta}_{t+1} = \frac{\boldsymbol{\theta}_t}{2S_t^2}$. Since $\boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}_t + \boldsymbol{\ell}_t$, we can upper bound $\|\boldsymbol{\theta}_{t-1}\|$ as

$$\|\boldsymbol{\theta}_{t-1}\| = \|\boldsymbol{\theta}_t + \boldsymbol{\ell}_t\| \le \|\boldsymbol{\theta}_t\| + \|\boldsymbol{\ell}_t\| \le S_t^2 + \|\boldsymbol{\ell}_t\| \, .$$

Similarly, since $S_t^2 = S_{t-1}^2 + \|\boldsymbol{\ell}_t\|^2$ and $\|\boldsymbol{\ell}_2\| \leq 1$, we can lower bound $\|\boldsymbol{\theta}_{t-1}\|$ as

$$\|\boldsymbol{\theta}_{t-1}\| > S_{t-1}^2 = S_t^2 - \|\boldsymbol{\ell}_t\|^2 \geq S_t^2 - \|\boldsymbol{\ell}_t\| \ .$$

Therefore,

$$\left|\|\boldsymbol{\theta}_{t-1}\| - S_t^2\right| \leq \|\boldsymbol{\ell}_t\| \ . \tag{33}$$

In order to prove (32), we substitute $\boldsymbol{\theta}_t = 2S_t^2 \boldsymbol{\beta}_{t+1}$ in it and get

$$2S_t^2 \left\|\boldsymbol{\beta}_{t+1}\right\|^2 - \left\|\boldsymbol{\beta}_{t+1}\right\|^2 S_t^2 \leq 2S_t^2 \left\langle \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t \right\rangle - \|\boldsymbol{\beta}_t\|^2 S_t^2 + \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \ ,$$

which is equivalent to

$$2S_t^2 \left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\right\|^2 \leq \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \ .$$

The last inequality follows from

$$\begin{aligned}
\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\right\|^2 &= \left\|\frac{\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t}{2S_t^2} - \frac{\boldsymbol{\theta}_{t-1}}{2\left\|\boldsymbol{\theta}_{t-1}\right\|}\right\|^2 \\
&= \frac{1}{4}\left\|\frac{\boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t}{S_t^2} - \frac{\boldsymbol{\theta}_{t-1}}{\|\boldsymbol{\theta}_{t-1}\|}\right\|^2 \\
&= \frac{1}{4}\left\|\boldsymbol{\theta}_{t-1}\left(\frac{1}{S_t^2} - \frac{1}{\|\boldsymbol{\theta}_{t-1}\|}\right) - \frac{\boldsymbol{\ell}_t}{S_t^2}\right\|^2 \\
&\leq \frac{1}{4}\left(\left\|\boldsymbol{\theta}_{t-1}\left(\frac{1}{S_t^2} - \frac{1}{\|\boldsymbol{\theta}_{t-1}\|}\right)\right\| + \frac{\|\boldsymbol{\ell}_t\|}{S_t^2}\right)^2 \\
&\leq \frac{1}{2}\|\boldsymbol{\theta}_{t-1}\|^2\left(\frac{1}{S_t^2} - \frac{1}{\|\boldsymbol{\theta}_{t-1}\|}\right)^2 + \frac{1}{2}\frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&= \frac{1}{2}\|\boldsymbol{\theta}_{t-1}\|^2 \frac{(\|\boldsymbol{\theta}_{t-1}\| - S_t^2)^2}{S_t^4\|\boldsymbol{\theta}_{t-1}\|^2} + \frac{1}{2}\frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&\leq \frac{1}{2}\|\boldsymbol{\theta}_{t-1}\|^2 \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4\|\boldsymbol{\theta}_{t-1}\|^2} + \frac{1}{2}\frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \\
&= \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^4} \ ,
\end{aligned}$$

where we substituted for $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t+1}$, used that $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \boldsymbol{\ell}_t$, used triangle inequality, the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, and, finally, the inequality (33). ■

Appendix C. Proofs for Section 7.1

In this section, we prove Lemma 8 and Lemma 7. For its proof, we first need the following technical lemma.

Lemma 21 (Convergence of non-negative supermartingales) *Let $Y_1, Y_2, \dots$ be a non-negative supermartingale with respect to a filtration $\mathcal{F}_1, \mathcal{F}_2, \dots$, that is, $Y_t \geq 0$, $\mathbf{E}[Y_t] < \infty$ and $\mathbf{E}[Y_{t+1} \mid \mathcal{F}_t] \leq Y_t$. Then, there exists a random variable Y such that*

$$\lim_{t\to\infty} Y_t = Y < \infty \qquad \textit{almost surely.}$$

The proof of Lemma 21 can be found e.g. in Resnick (1999, Theorem 10.8.5, page 373).

We can now prove Lemma 7.

Proof [Proof of Lemma 7] Lemma 20 implies that

$$-H_{t+1}(\boldsymbol{x}_{t+1}) \leq -H_t(\boldsymbol{x}_t) - \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle \leq 0\,. \tag{34}$$

By definition of $H_t(\cdot)$ and $H_{t+1}(\cdot)$,

$$H_{t+1}(\boldsymbol{x}^*) - \phi_{t+1}(\boldsymbol{x}^*) = H_t(\boldsymbol{x}^*) - \phi_t(\boldsymbol{x}^*) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}^* \rangle \ . \tag{35}$$

We sum (34) and (35) and get

$$H_{t+1}(\boldsymbol{x}^*) - H_{t+1}(\boldsymbol{x}_{t+1}) - \phi_{t+1}(\boldsymbol{x}^*) \leq H_t(\boldsymbol{x}^*) - H_t(\boldsymbol{x}_t) - \phi_t(\boldsymbol{x}^*) - \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \ .$$

Since $\boldsymbol{x}_{T+1}$ is the minimizer of H_{T+1}, $H_{t+1}(\boldsymbol{x}^*) - H_{t+1}(\boldsymbol{x}_{t+1}) = B_{H_{t+1}}(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) = B_{\phi_{t+1}}(\boldsymbol{x}^*, \boldsymbol{x}_{t+1})$. Similarly, $H_t(\boldsymbol{x}^*) - H_t(\boldsymbol{x}_t) = B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t)$. Therefore,

$$B_{\phi_{t+1}}(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \phi_{t+1}(\boldsymbol{x}^*) \leq B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) - \phi_t(\boldsymbol{x}^*) - \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \ .$$

We add $\phi_\infty(\boldsymbol{x}^*)$ to both sides and we get

$$B_{\phi_{t+1}}(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) + \phi_\infty(\boldsymbol{x}^*) - \phi_{t+1}(\boldsymbol{x}^*) \leq B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) + \phi_\infty(\boldsymbol{x}^*) - \phi_t(\boldsymbol{x}^*) - \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \ .$$

Let $Y_t = B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) + \phi_\infty(\boldsymbol{x}^*) - \phi_t(\boldsymbol{x}^*)$. We have

$$Y_{t+1} \leq Y_t - \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \ .$$

Equivalently,

$$Y_{t+1} \leq Y_t - \eta_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \ .$$

Taking conditional expectation of both sides,

$$\mathbf{E}[Y_{t+1} \mid \mathcal{F}_t] \leq Y_t - \eta_t \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \rangle \ .$$

Since F is variationally coherent,

$$\mathbf{E}[Y_{t+1} \mid \mathcal{F}_t] \leq Y_t\,.$$

Note that Y_t is non-negative, since $B_{\phi_t}(\cdot,\cdot)$ is non-negative and $\phi_t(\boldsymbol{x}^*) \leq \phi_\infty(\boldsymbol{x}^*)$. Thus $Y_1, Y_2, \dots$ is a non-negative supermartingale. By Lemma 21, Y_t converges almost surely to a finite limit. Since $\phi_\infty(\boldsymbol{x}^*) - \phi_t(\boldsymbol{x}^*) \geq 0$ and $\{\phi_\infty(\boldsymbol{x}^*) - \phi_t(\boldsymbol{x}^*)\}_{t=1}^\infty$ is non-increasing, $\lim_{t\to\infty}(\phi_\infty(\boldsymbol{x}^*) - \phi_t(\boldsymbol{x}^*))$ exists almost surely. Therefore, $B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) = Y_t - \phi_\infty(\boldsymbol{x}^*) + \phi_t(\boldsymbol{x}^*)$ has a limit almost surely. ■

To prove Lemma 8, we need the following lemmas.

Lemma 22 (Second derivative bounds) *Let $I \subseteq \mathbb{R}$ be an open interval. Let $f : I \to \mathbb{R}$ be a function such that $f''(x)$ exists almost everywhere and $f''(x)$ is continuous almost everywhere. Let $g, h : I \to \mathbb{R}$ be twice continuously differentiable. Suppose, for almost all $x \in I$,*

$$g''(x) \leq f''(x) \leq h''(x) \, . \tag{36}$$

Then, for any $u, v \in I$, there exists z_1, z_2 between u and v such that

$$\frac{1}{2} g''(z_1)(u-v)^2 \leq f(u) - f(v) - (u-v)f'(v) \leq \frac{1}{2} h''(z_2)(u-v)^2 \, .$$

Proof The functions $g''(\cdot)$, $h''(\cdot)$ are continuous and therefore bounded on any closed interval $J \subseteq I$. Similarly, $f''(\cdot)$ is continuous almost everywhere and by assumption (36) bounded on any closed interval. Therefore, $g''(\cdot)$, $h''(\cdot)$, $f''(\cdot)$ are Riemann integrable on any closed interval $J \subseteq I$. Integrating (36), we get

$$\int_v^s g''(x)dx \leq \int_v^s f''(x)dx \leq \int_v^s h''(x)dx \qquad \text{for any } s \in I.$$

Fundamental theorem of calculus implies that

$$g'(s) - g'(v) \leq f'(s) - f'(v) \leq h'(s) - h'(v) \qquad \text{for any } s \in I.$$

Integrating one more time, we get

$$\int_v^u g'(s) - g'(v)ds \leq \int_v^u f'(s) - f'(v)ds \leq \int_v^u h'(s) - h'(v)ds$$

All integrals exists as Riemann integrals, since $f'(\cdot)$, $g'(\cdot)$, $h'(\cdot)$ are necessarily continuous. Fundamental theorem of calculus implies that

$$g(u) - g(v) - g'(v)(u-v) \leq f(u) - f(v) - (u-v)f'(v) \leq h(u) - h(v) - h'(v)(u-v) \, .$$

The lemma follows by applying Taylor's theorem to g and h. ■

Lemma 23 (Hessian of radially symmetric functions) *Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ be defined as $g(\boldsymbol{x}) = f(\|\boldsymbol{x}\|)$. If f is twice differentiable at $\|\boldsymbol{x}\|$ and $\|\boldsymbol{x}\| > 0$ then*

$$\min\left\{ f''(\|\boldsymbol{x}\|), \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|} \right\} I \preceq \nabla^2 g(\boldsymbol{x}) \preceq \max\left\{ f''(\|\boldsymbol{x}\|), \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|} \right\} I \, .$$

Proof We need to prove that

$$\forall \boldsymbol{u} \in \mathbb{R}^d \qquad \|\boldsymbol{u}\|^2 \min\left(f''(\|\boldsymbol{x}\|), \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|} \right) \leq \boldsymbol{u}^\top \nabla^2 g(\boldsymbol{x}) \boldsymbol{u} \leq \|\boldsymbol{u}\|^2 \max\left(f''(\|\boldsymbol{x}\|), \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|} \right) \, .$$

The gradient of $g(\boldsymbol{x})$ is $\nabla g(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} f'(\|\boldsymbol{x}\|)$. The Hessian of $g(\boldsymbol{x})$ is

$$\nabla^2 g(\boldsymbol{x}) = f''(\|\boldsymbol{x}\|) \frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2} + f'(\|\boldsymbol{x}\|) \left(\frac{I}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^3} \right) \, .$$

In order to upper and lower bound $\boldsymbol{u}^\top \nabla^2 g(\boldsymbol{x})\boldsymbol{u}$, we decompose the vector $\boldsymbol{u} \in \mathbb{R}^d$ as $\boldsymbol{u} = \alpha\boldsymbol{x}+\boldsymbol{v}$ where $\boldsymbol{v}$ is orthogonal to $\boldsymbol{x}$. For convenience, let $\beta = \frac{f''(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|^2} - \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|^3}$ and $\gamma = \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}$. We can express $\boldsymbol{u}^\top \nabla^2 g(\boldsymbol{x})\boldsymbol{u}$ as

$$\begin{aligned}
\boldsymbol{u}^\top \nabla^2 g(\boldsymbol{x})\boldsymbol{u} &= \boldsymbol{u}^\top \nabla^2 (\beta \boldsymbol{x}\boldsymbol{x}^\top + \gamma I)\boldsymbol{u} \\
&= (\alpha\boldsymbol{x}+\boldsymbol{v})^\top \nabla^2 (\beta \boldsymbol{x}\boldsymbol{x}^\top + \gamma I)(\alpha\boldsymbol{x}+\boldsymbol{v}) \\
&= \alpha^2 \beta \|\boldsymbol{x}\|^4 + \alpha^2\gamma \|\boldsymbol{x}\|^2 + \gamma \|\boldsymbol{v}\|^2 \\
&= \alpha^2 \left(\frac{f''(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|^2} - \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|^3} \right) \|\boldsymbol{x}\|^4 + \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}(\alpha^2 \|\boldsymbol{x}\|^2 + \|\boldsymbol{v}\|^2) \\
&= \alpha^2 f''(\|\boldsymbol{x}\|) \|\boldsymbol{x}\|^2 + \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|} \|\boldsymbol{v}\|^2 \ .
\end{aligned} \tag{37}$$

Let $A = \min\left(f''(\|\boldsymbol{x}\|), \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}\right)$ and $B = \max\left(f''(\|\boldsymbol{x}\|), \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}\right)$. We can upper and lower bound

$$A(\alpha^2 \|\boldsymbol{x}\|^2 + \|\boldsymbol{v}\|^2) \le \alpha^2 f''(\|\boldsymbol{x}\|) \|\boldsymbol{x}\|^2 + \frac{f'(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|} \|\boldsymbol{v}\|^2 \le B(\alpha^2 \|\boldsymbol{x}\|^2 + \|\boldsymbol{v}\|^2) \ .$$

Since $\boldsymbol{x}$ and $\boldsymbol{v}$ are orthogonal, $\alpha^2 \|\boldsymbol{x}\|^2 + \|\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2$. Thus,

$$A \|\boldsymbol{u}\|^2 \le \boldsymbol{u}^\top \nabla^2 g(\boldsymbol{x})\boldsymbol{u} \le B \|\boldsymbol{u}\|^2 \ .$$

■

Lemma 24 (Bounds on Bregman divergence) *For any $t \ge 1$, there exists $\widetilde{x}_t$ between $\|\boldsymbol{x}^*\|$ and $\|\boldsymbol{x}_t\|$ such that*

$$\frac{\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2}{S_{t-1}+2} \min\left\{1, \frac{1}{\widetilde{x}_t}\right\} \le B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) \le 2\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2 (S_{t-1}^2 + Q_{t-1}) \exp(Q_{t-1}) \ .$$

Proof Let us define $f : \mathbb{R} \to \mathbb{R}$, $\psi_t : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{z}_t : \mathbb{R} \to \mathbb{R}^d$ as

$$\begin{aligned}
f_t(\alpha) &= \phi_t(\alpha \boldsymbol{x}^* + (1-\alpha)\boldsymbol{x}_t) \ , \\
\psi_t(x) &= \psi(x, S_{t-1}, Q_{t-1}) \ , \\
\boldsymbol{z}_t(\alpha) &= \alpha \boldsymbol{x}^* + (1-\alpha)\boldsymbol{x}_t \ .
\end{aligned}$$

By definition of Bregman divergence,

$$B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) = f_t(1) - f_t(0) - f_t'(0) \ .$$

We we will use Lemma 22 to lower and upper bound the right-hand side. In order to apply the lemma we need upper and lower bounds on $f_t''(\alpha)$. Since $\phi_t(\boldsymbol{x}) = \psi_t(\|\boldsymbol{x}\|)$,

$$f_t''(\alpha) = (\boldsymbol{x}^* - \boldsymbol{x}_t)^\top \nabla^2 \phi_t(\alpha\boldsymbol{x}^* + (1-\alpha)\boldsymbol{x}_t)(\boldsymbol{x}^* - \boldsymbol{x}_t) = (\boldsymbol{x}^* - \boldsymbol{x}_t)^\top \nabla^2 \phi_t(\boldsymbol{z}_t(\alpha))(\boldsymbol{x}^* - \boldsymbol{x}_t)$$

Therefore, by Lemma 23,

$$\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2 \min\left\{\psi_t''(\|\boldsymbol{z}_t(\alpha)\|), \frac{\psi_t'(\|\boldsymbol{z}_t(\alpha)\|)}{\|\boldsymbol{z}_t(\alpha)\|}\right\} \le f(\alpha) \le \|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2 \max\left\{\psi_t''(\|\boldsymbol{z}_t(\alpha)\|), \frac{\psi_t'(\|\boldsymbol{z}_t(\alpha)\|)}{\|\boldsymbol{z}_t(\alpha)\|}\right\} .$$

It remains to lower bound $\min\left\{\psi_t''(\|\boldsymbol{z}_t(\alpha)\|), \frac{\psi_t'(\|\boldsymbol{z}_t(\alpha)\|)}{\|\boldsymbol{z}_t(\alpha)\|}\right\}$ and upper bound $\max\left\{\psi_t''(\|\boldsymbol{z}_t(\alpha)\|), \frac{\psi_t'(\|\boldsymbol{z}_t(\alpha)\|)}{\|\boldsymbol{z}_t(\alpha)\|}\right\}$. Note that $\psi_t'(x) = \frac{\partial \psi(x,S,Q)}{\partial x}$ and $\psi_t''(x) = \frac{\partial^2 \psi(x,S,Q)}{\partial x^2}$. We use Lemma 16 to bound these quantities. A lower bound follows from

$$
\begin{aligned}
&\min\left\{\psi_t''(\|\boldsymbol{z}_t(\alpha)\|), \frac{\psi_t'(\|\boldsymbol{z}_t(\alpha)\|)}{\|\boldsymbol{z}_t(\alpha)\|}\right\} \\
&\geq \min\left\{\frac{2}{\|\boldsymbol{z}_t(\alpha)\|\,(\frac{1}{2}S_{t-1}+1)}, \frac{\sqrt{\ln(1+2\|\boldsymbol{z}_t(\alpha)\|^2)}}{\|\boldsymbol{z}_t(\alpha)\|\,(\frac{1}{2}S_{t-1}+1)}, \frac{\sqrt{\ln(1+2\|\boldsymbol{z}_t(\alpha)\|^2)}}{\|\boldsymbol{z}_t(\alpha)\|}\right\} \\
&= \frac{\min\left\{2, \sqrt{\ln(1+2\|\boldsymbol{z}_t(\alpha)\|^2)}\right\}}{\|\boldsymbol{z}_t(\alpha)\|\,(\frac{1}{2}S_{t-1}+1)} && \text{(since } S_{t-1} \geq 1\text{)} \\
&\geq \frac{1}{\frac{1}{2}S_{t-1}+1}\min\left\{1, \frac{1}{\|\boldsymbol{z}_t(\alpha)\|}\right\} && \text{(by case analysis)}\ .
\end{aligned}
$$

An upper bound follows from

$$
\begin{aligned}
&\max\left\{\psi_t''(\|\boldsymbol{z}_t(\alpha)\|), \frac{\psi_t'(\|\boldsymbol{z}_t(\alpha)\|)}{\|\boldsymbol{z}_t(\alpha)\|}\right\} \\
&\leq \max\left\{2S_{t-1}^2\exp(Q_{t-1}), 4\exp\left(Q_{t-1} - \frac{1}{4}S_{t-1}^2\right), 4 + \frac{S_{t-1}^2 + 4Q_{t-1} - 4}{\exp\left(\frac{1}{4}S_{t-1}^2 - Q_{t-1}\right)}\right\} \\
&\leq \max\left\{4S_{t-1}^2\exp(Q_{t-1}), 4 + \frac{S_{t-1}^2 + 4Q_{t-1} - 4}{\exp\left(\frac{1}{4}S_{t-1}^2 - Q_{t-1}\right)}\right\} \\
&\leq \max\left\{4S_{t-1}^2\exp(Q_{t-1}), 4 + (S_{t-1}^2 + 4Q_{t-1} - 4)\exp(Q_{t-1})\right\} \\
&\leq \max\left\{4S_{t-1}^2\exp(Q_{t-1}), (S_{t-1}^2 + 4Q_{t-1})\exp(Q_{t-1})\right\} \\
&\leq 4(S_{t-1}^2 + Q_{t-1})\exp(Q_{t-1})\ .
\end{aligned}
$$

The second derivative $f_t''(\alpha)$ exists almost everywhere and $f_t''(\alpha)$ is continuous almost everywhere. Lemma 22 implies that there exists $\alpha^* \in [0,1]$ such that

$$
\frac{1}{2}\frac{\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2}{\frac{1}{2}S_{t-1}+1}\min\left\{1, \frac{1}{\|\boldsymbol{z}_t(\alpha^*)\|}\right\} \leq f_t(1) - f_t(0) - f_t'(0) \leq \frac{1}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2\, 4(S_{t-1}^2 + Q_{t-1})\exp(Q_{t-1})\ .
$$

Let $\widetilde{x}_t = \|\boldsymbol{z}_t(\alpha^*)\|$. The number $\widetilde{x}_t$ lies between $\|\boldsymbol{x}^*\|$ and $\|\boldsymbol{x}_t\|$, since $\boldsymbol{z}(\alpha^*)$ is a convex combination of $\alpha\boldsymbol{x}^* + (1-\alpha)\boldsymbol{x}_t$. Thus,

$$
\frac{\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2}{S_{t-1}+2}\min\left\{1, \frac{1}{\widetilde{x}_t}\right\} \leq f_t(1) - f_t(0) - f_t'(0) \leq 2\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2 (S_{t-1}^2 + Q_{t-1})\exp(Q_{t-1})\ .
$$

■

Lemma 25 (Bounds on iterates) *For any $t \geq 1$,*

$$
\|\boldsymbol{x}_t\| \leq \max\left\{\|\boldsymbol{x}^*\| + \sqrt{(S_{t-1}+2)B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t)},\ 2\|\boldsymbol{x}^*\|,\ 4(S_{t-1}+2)B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t)\right\}\ . \tag{38}
$$

In particular,

$$\sup_{t=1,2,\ldots} \|\boldsymbol{x}_t\| < \infty \qquad \textit{almost surely.}$$

Proof Lemma 24 implies

$$\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2 \le (S_{t-1}+2)\max\{1, \widetilde{x}_t\}\, B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t)\,. \tag{39}$$

for some $\widetilde{x}_t$ between $\|\boldsymbol{x}^*\|$ and $\|\boldsymbol{x}_t\|$.

In order to prove (38), we consider three cases. If $\|\boldsymbol{x}_t\| \le 2\|\boldsymbol{x}^*\|$ there is nothing to prove. If $\widetilde{x}_t \le 1$ then

$$\|\boldsymbol{x}_t\| \le \|\boldsymbol{x}^*\| + \|\boldsymbol{x}^* - \boldsymbol{x}_t\| \le \|\boldsymbol{x}^*\| + \sqrt{(S_\infty + 2)B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t)},$$

where in the last step we use (39). It remains to verify (38) when $\|\boldsymbol{x}_t\| > 2\|\boldsymbol{x}^*\|$ and $\widetilde{x}_t > 1$. We have

$$\|\boldsymbol{x}_t\| \le \frac{\|\boldsymbol{x}_t\|^2}{\widetilde{x}_t} \le \frac{4(\|\boldsymbol{x}_t\| - \|\boldsymbol{x}^*\|)^2}{\widetilde{x}_t} \le \frac{4\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2}{\widetilde{x}_t} \le 4(S_\infty + 2)B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t),$$

where in the last step we use (39).

Lemma 7 implies that $\sup_{t=1,2,\ldots} B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) = B < \infty$. Lemma 19 implies that $S_{t-1} \le S_\infty < \infty$. Therefore, inequality (38) implies that

$$\sup_{t=1,2,\ldots} \|\boldsymbol{x}_t\| \le \max\left\{\|\boldsymbol{x}^*\| + \sqrt{(S_\infty+2)B},\ 2\|\boldsymbol{x}^*\|,\ 4(S_\infty+2)B\right\}\,.$$

■

We are now ready to prove Lemma 8.

Proof [Proof of Lemma 8] Let $\sup_{t=1,2,\ldots} \|\boldsymbol{x}_t\| = X$. Lemma 25 implies $X < \infty$ almost surely. Lemma 24 implies that

$$\frac{\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2}{S_\infty + 2} \min\left\{1, \frac{1}{X}\right\} \le B_{\phi_t}(\boldsymbol{x}^*, \boldsymbol{x}_t) \le 2\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2 (S_\infty^2 + Q_\infty)\exp(Q_\infty)\,.$$

The lemma follows by defining

$$C_1 = \frac{1}{S_\infty + 2}\min\left\{1, \frac{1}{X}\right\}\,,$$
$$C_2 = 2(S_\infty^2 + Q_\infty)\exp(Q_\infty)\,.$$

■

Appendix D. Proofs for Section 7.2

Proof [Proof of Lemma 9] The bound on S_T follows from

$$\begin{aligned} S_T^2 &= 4 + \sum_{t=1}^T \|\boldsymbol{\ell}_t\|^2 = 4 + \sum_{t=1}^T \eta_t^2 \|\boldsymbol{g}_t\|^2 = 4 + \sum_{t=1}^T \frac{\|\boldsymbol{g}_t\|^2}{G^2 t^{2\alpha}} \leq 4 + \sum_{t=1}^T t^{-2\alpha} \\ &= 5 + \sum_{t=2}^T t^{-2\alpha} \leq 5 + \int_1^T x^{-2\alpha} dx = 5 + \frac{1 - T^{1-2\alpha}}{2\alpha - 1} \leq 5 + \frac{1}{2\alpha - 1}, \end{aligned}$$

where in the first inequality we used $\|\boldsymbol{g}_t\| \leq G$. The bound on $Q_T \leq 2 \ln S_T$ follows from Lemma 18. The bound on $\phi_{T+1}(\boldsymbol{u})$ follows from

$$\begin{aligned} \phi_T(\boldsymbol{u}) &= \psi(\|\boldsymbol{u}\|, S_{T-1}, Q_{T-1}) \\ &\leq S_{T-1} \|\boldsymbol{u}\| \left[2 \ln(1 + 2\|\boldsymbol{u}\|) + 3Q_{T-1} + 3S_{T-1}\right] && \text{(Lemma 16)} \\ &\leq S_{T-1} \|\boldsymbol{u}\| \left[2 \ln(1 + 2\|\boldsymbol{u}\|) + 6 \ln S_{T-1} + 3S_{T-1}\right] \\ &\leq S_{T-1} \|\boldsymbol{u}\| \left[2 \ln(1 + 2\|\boldsymbol{u}\|) + 9S_{T-1}\right] \\ &\leq \sqrt{5 + \frac{1}{2\alpha - 1}} \|\boldsymbol{u}\| \left[2 \ln(1 + 2\|\boldsymbol{u}\|) + 9\sqrt{5 + \frac{1}{2\alpha - 1}}\right] . \end{aligned}$$

■

Appendix E. Proofs for Section 7.3

Proof [Proof of Lemma 10] Let $S_k = \frac{1}{k}\sum_{t=T-k+1}^{T} \eta_t q_t$. We have

$$\sum_{t=T-k}^{T} \eta_t(q_t - q_{T-k}) \geq \sum_{t=T-k}^{T} (\eta_t q_t - \eta_{T-k} q_{T-k}) = (k+1)S_{k+1} - \eta_{T-k}(k+1)q_{T-k} .$$

That implies

$$S_{k+1} - \eta_{T-k} q_{T-k} \leq \frac{1}{k+1} \sum_{t=T-k}^{T} \eta_t(q_t - q_{T-k}) .$$

From the definition of S_k and the above inequality, we have

$$kS_k = (k+1)S_{k+1} - \eta_{T-k}q_{T-k} = kS_{k+1} + S_{k+1} - \eta_{T-k}q_{T-k} \leq kS_{k+1} + \frac{1}{k+1}\sum_{t=T-k}^{T} \eta_t(q_t - q_{T-k}) .$$

Therefore,

$$S_k \leq S_{k+1} + \frac{1}{k(k+1)} \sum_{t=T-k}^{T} \eta_t(q_t - q_{T-k}) .$$

Unrolling the inequality we get

$$\eta_T q_T = S_1 \leq S_T + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} \eta_t(q_t - q_{T-k}) .$$

Using the definition of S_T the lemma follows. ■

Proof [Proof of Lemma 11] Let $C_t = \frac{1}{2}\exp(\frac{1}{4}S_{t-1}^2 - Q_{t-1})$. We will first show that

$$\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) \leq \max\left\{C_{T+1}, \|\boldsymbol{x}_A\|\right\}(S_T^2 - S_{A-1}^2) . \tag{40}$$

Now, we claim that $C_1, C_2, \dots$ is a non-decreasing sequence. Indeed, $C_t \leq C_{t+1}$ is equivalent to

$$\frac{1}{4}S_{t-1}^2 - Q_{t-1} \leq \frac{1}{4}S_t^2 - Q_t ,$$

which is the same as

$$\frac{1}{4}\|\boldsymbol{\ell}_t\|^2 - \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \geq 0 ,$$

which trivially holds since $S_t^2 \geq 4$.

Second, observe that

$$Q_T - Q_{A-1} = \sum_{t=A}^{T} \frac{\|\boldsymbol{\ell}_t\|^2}{S_t^2} \leq \frac{1}{4}\sum_{t=A}^{T}\|\boldsymbol{\ell}_t\|^2 = \frac{1}{4}(S_T^2 - S_{A-1}^2) . \tag{41}$$

We now prove (40) by considering three cases.

Case $\|\boldsymbol{x}_A\| \le C_A$:

$$\begin{aligned}
\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) &= \psi(\|\boldsymbol{x}_A\|, S_T, Q_T) - \psi(\|\boldsymbol{x}_A\|, S_{A-1}, Q_{A-1}) \\
&\le \psi(C_A, S_T, Q_T) - \psi(C_A, S_{A-1}, Q_{A-1}) \\
&= \psi(C_A, S_T, Q_T) - C_A(S_{A-1}^2 - 2) \\
&= C_A \frac{\sqrt{2S_T^2}(W(2\exp(2Q_T)S_T^2C_A^2) - 1)}{\sqrt{W(2\exp(2Q_T)S_T^2C_A^2)}} - C_A(S_{A-1}^2 - 2) \\
&\le C_A \frac{\sqrt{2S_T^2}(W(2\exp(2Q_T)S_T^2C_{T+1}^2) - 1)}{\sqrt{W(2\exp(2Q_T)S_T^2C_{T+1}^2)}} - C_A(S_{A-1}^2 - 2) \\
&= C_A \frac{\sqrt{2S_T^2}(\frac{1}{2}S_T^2 - 1)}{\sqrt{\frac{1}{2}S_T^2}} - C_A(S_{A-1}^2 - 2) \\
&= C_A(S_T^2 - 2) - C_A(S_{A-1}^2 - 2) \\
&= C_A(S_T^2 - S_{A-1}^2) \\
&\le C_{T+1}(S_T^2 - S_{A-1}^2)\,.
\end{aligned}$$

Case $C_A \le \|\boldsymbol{x}_A\| \le C_{T+1}$:

$$\begin{aligned}
\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) &= \psi(\|\boldsymbol{x}_A\|, S_T, Q_T) - \psi(\|\boldsymbol{x}_A\|, S_{A-1}, Q_{A-1}) \\
&\le \psi(C_{T+1}, S_T, Q_T) - \psi(C_{T+1}, S_{A-1}, Q_{A-1}) \\
&= C_{T+1}\left(\frac{1}{2}S_T^2 + Q_T - \frac{1}{2}S_{A-1}^2 - Q_{A-1}\right) \\
&\le C_{T+1}(S_T^2 - S_{A-1}^2) \qquad \text{(using (41))}\,.
\end{aligned}$$

Case 3: $\|\boldsymbol{x}_A\| \ge C_{T+1}$:

$$\begin{aligned}
\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) &= \psi(\|\boldsymbol{x}_A\|, S_T, Q_T) - \psi(\|\boldsymbol{x}_A\|, S_{A-1}, Q_{A-1}) \\
&= \|\boldsymbol{x}_A\|\left(\frac{1}{2}S_T^2 + Q_T - \frac{1}{2}S_{A-1}^2 - Q_{A-1}\right) \\
&\le \|\boldsymbol{x}_A\|(S_T^2 - S_{A-1}^2) \qquad \text{(using (41))}\,.
\end{aligned}$$

Using the fact that $S_T^2 - S_{A-1}^2 = \sum_{t=A}^T \|\boldsymbol{\ell}_t\|^2$, (40) gives

$$\begin{aligned}
\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) &\le \max\left\{\frac{1}{2}\exp\left(\frac{1}{4}S_T^2 - Q_T\right), \|\boldsymbol{x}_A\|\right\}\sum_{t=A}^T \|\boldsymbol{\ell}_t\|^2 \\
&\le \max\left\{\frac{1}{2}\exp\left(\frac{1}{4}S_T^2 - Q_T\right), \|\boldsymbol{x}_A\|\right\}\sum_{t=A}^T t^{-2\alpha},
\end{aligned}$$

where we used $\|\boldsymbol{\ell}_t\| = \eta_t\|\boldsymbol{g}_t\| \le t^{-\alpha}$. We take expectation to both sides of the above equation and define

$$\kappa = \sup_{\substack{T=0,1,2,\dots\\A=1,2,\dots}} \mathbf{E}\left[\max\left\{\frac{1}{2}\exp\left(\frac{1}{4}S_T^2 - Q_T\right), \|\boldsymbol{x}_A\|\right\}\right], \tag{42}$$

to obtain

$$\mathbf{E}\left[\phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A)\right] \le \kappa \sum_{t=A}^{T} t^{-2\alpha} \,.$$

We upper bound κ as follows

$$\begin{aligned}
\kappa &= \sup_{\substack{T\ge 0\\ A\ge 1}} \mathbf{E}\left[\max\left\{\frac{1}{2}\exp\left(\frac{1}{4}S_T^2 - Q_T\right), \|\boldsymbol{x}_A\|\right\}\right] \\
&\le \sup_{A\ge 1} \mathbf{E}\left[\max\left\{\frac{1}{2}\exp\left(\frac{1}{4}S\right), \|\boldsymbol{x}_A\|\right\}\right] \\
&\le \sup_{A\ge 1} \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \mathbf{E}\left[\|\boldsymbol{x}_A\|\right] \\
&\le \sup_{A\ge 1} \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \mathbf{E}\left[\max\left\{\|\boldsymbol{x}^*\| + \sqrt{(S_{A-1}+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)},\ 2\|\boldsymbol{x}^*\|,\ 4(S_{A-1}+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)\right\}\right] \\
&\le \sup_{A\ge 1} \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \mathbf{E}\left[3\|\boldsymbol{x}^*\| + \sqrt{(S_{A-1}+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)} + 4(S_{A-1}+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)\right] \\
&\le \sup_{A\ge 1} \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \mathbf{E}\left[3\|\boldsymbol{x}^*\| + \sqrt{(S+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)} + 4(S+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)\right] \\
&\le \sup_{A\ge 1} \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \sqrt{\mathbf{E}\left[(S+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)\right]} + 3\|\boldsymbol{x}^*\| + \mathbf{E}\left[4(S+2)B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)\right] \\
&\le \sup_{A\ge 1} \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \sqrt{(S+2)\,\mathbf{E}[1+\phi_A(\boldsymbol{x}^*)]} + 3\|\boldsymbol{x}^*\| + 4(S+2)\,\mathbf{E}[1+\phi_A(\boldsymbol{x}^*)] \\
&\le \frac{1}{2}\exp\left(\frac{1}{4}S\right) + \sqrt{(S+2)\,\mathbf{E}[1+\phi_\infty(\boldsymbol{x}^*)]} + 3\|\boldsymbol{x}^*\| + 4(S+2)\,\mathbf{E}[1+\phi_\infty(\boldsymbol{x}^*)] \\
&\le \frac{1}{2}\exp\left(\frac{1}{4}S\right) + (S+2)\sqrt{\mathbf{E}[1+\phi_\infty(\boldsymbol{x}^*)]} + 3\|\boldsymbol{x}^*\| + 4(S+2)\,\mathbf{E}[1+\phi_\infty(\boldsymbol{x}^*)] \\
&\le \frac{1}{2}\exp\left(\frac{1}{4}S\right) + (S+2)\,\mathbf{E}[2+\phi_\infty(\boldsymbol{x}^*)] + 3\|\boldsymbol{x}^*\| + 4(S+2)\,\mathbf{E}[2+\phi_\infty(\boldsymbol{x}^*)] \\
&= \frac{1}{2}\exp\left(\frac{1}{4}S\right) + 3\|\boldsymbol{x}^*\| + 5(S+2)\,\mathbf{E}[2+\phi_\infty(\boldsymbol{x}^*)],
\end{aligned}$$

where we used inequality $Q_T \ge 0$, Lemma 9 that states $S_t \le S$ for all $t \ge 0$, Lemma 25 to upper bound $\|\boldsymbol{x}_A\|$, inequality $\mathbf{E}[\sqrt{\cdot}] \le \sqrt{\mathbf{E}[\cdot]}$, Lemma 6 to upper bound $\mathbf{E}[B_{\phi_A}(\boldsymbol{x}^*, \boldsymbol{x}_A)]$, and inequality $\sqrt{x} \le 1 + x$ valid for any $x \ge 0$. ■

Proof [Proof of Lemma 12] We upper bound the sum $\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-2\alpha}$ as follows.

$$\begin{aligned}
\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-2\alpha} &= \sum_{k=1}^{T-1} \sum_{t=T-k}^{T} t^{-2\alpha} \frac{1}{k(k+1)} \\
&= \sum_{k=1}^{T-1} T^{-2\alpha} \frac{1}{k(k+1)} + \sum_{t=1}^{T-1} \sum_{k=T-t}^{T-1} t^{-2\alpha} \frac{1}{k(k+1)} \\
&= T^{-2\alpha} \sum_{k=1}^{T-1} \left(\frac{1}{k} - \frac{1}{k+1} \right) + \sum_{t=1}^{T-1} t^{-2\alpha} \sum_{k=T-t}^{T-1} \left(\frac{1}{k} - \frac{1}{k+1} \right) \\
&= T^{-2\alpha} \left(1 - \frac{1}{T} \right) + \sum_{t=1}^{T-1} t^{-2\alpha} \left(\frac{1}{T-t} - \frac{1}{T} \right) \\
&= T^{-2\alpha} \left(1 - \frac{1}{T} \right) + \sum_{t=1}^{T-1} t^{-2\alpha} \frac{t}{T(T-t)} \\
&\leq T^{-2\alpha} + \sum_{t=1}^{T-1} \frac{t^{1-2\alpha}}{T(T-t)} \,.
\end{aligned}$$

Since $\alpha > \frac{1}{2}$, the function $t \mapsto t^{1-2\alpha}$ is convex on the interval $(0, \infty)$. We can upper bound $t^{1-2\alpha}$ on the interval $[1, T]$ with a linear function. That is,

$$t^{1-2\alpha} \leq \frac{T-t}{T-1} + \frac{t-1}{T-1} T^{1-2\alpha} \qquad \text{for } t \in [1, T] \,.$$

Hence,

$$\begin{aligned}
\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-2\alpha} &\leq T^{-2\alpha} + \sum_{t=1}^{T-1} \frac{t^{1-2\alpha}}{T(T-t)} \leq T^{-2\alpha} + \sum_{t=1}^{T-1} \left(\frac{1}{T(T-1)} + \frac{(t-1)T^{-2\alpha}}{(T-1)(T-t)} \right) \\
&= T^{-2\alpha} + \frac{1}{T} + \frac{T^{-2\alpha}}{(T-1)} \sum_{t=1}^{T-1} \frac{t-1}{T-t} = T^{-2\alpha} + \frac{1}{T} + \frac{T^{-2\alpha}}{(T-1)} \sum_{t=1}^{T-1} \left(\frac{T-1}{T-t} - 1 \right) \\
&= \frac{1}{T} + \frac{T^{-2\alpha}}{(T-1)} \sum_{t=1}^{T-1} \frac{T-1}{T-t} = \frac{1}{T} + T^{-2\alpha} \sum_{t=1}^{T-1} \frac{1}{T-t} = \frac{1}{T} + T^{-2\alpha} \sum_{t=1}^{T-1} \frac{1}{t} \\
&\leq \frac{1}{T} + \frac{1 + \ln(T-1)}{T^{2\alpha}} \leq \frac{1}{T} + T^{-2\alpha} + \frac{\ln T}{T^{2\alpha}} \leq \frac{1}{T} + T^{-2\alpha} + \frac{1}{e(2\alpha - 1)T},
\end{aligned}$$

where in the last step we used that $\ln x \leq \frac{x^p}{ep}$ for all $x > 0$ and all $p > 0$ with $p = 2\alpha - 1$ and $x = T$. ■

Proof [Proof of Lemma 13] For any $\boldsymbol{u} \in \mathbb{R}^d$ and any $A \leq T$,

$$-\sum_{t=A}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle = \phi_{T+1}(\boldsymbol{u}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=1}^{A-1} \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle$$

$$= \phi_{T+1}(\boldsymbol{u}) - H_A(\boldsymbol{x}_A) + H_A(\boldsymbol{x}_A) - H_{T+1}(\boldsymbol{x}_{T+1}) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=1}^{A-1} \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle$$

$$= \phi_{T+1}(\boldsymbol{u}) - H_A(\boldsymbol{x}_A) + \sum_{t=A}^{T} [H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1})] + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=1}^{A-1} \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle \ .$$

Adding $\sum_{t=A}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle$ to both sides, we get

$$\sum_{t=A}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle \leq \phi_{T+1}(\boldsymbol{u}) - H_A(\boldsymbol{x}_A) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=A}^{T} [H_t(\boldsymbol{x}_t) - H_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t \rangle] + \sum_{t=1}^{A-1} \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle \ .$$

Lemma 20 implies that

$$\sum_{t=A}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle \leq \phi_{T+1}(\boldsymbol{u}) - H_A(\boldsymbol{x}_A) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{u}) + \sum_{t=1}^{A-1} \langle \boldsymbol{\ell}_t, \boldsymbol{u} \rangle \ .$$

Set $\boldsymbol{u} = \boldsymbol{x}_A$, to obtain

$$\begin{aligned}
\sum_{t=A}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}_A \rangle &\leq \phi_{T+1}(\boldsymbol{x}_A) - H_A(\boldsymbol{x}_A) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{x}_A) + \sum_{t=1}^{A-1} \langle \boldsymbol{\ell}_t, \boldsymbol{x}_A \rangle \\
&= \phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) + H_{T+1}(\boldsymbol{x}_{T+1}) - H_{T+1}(\boldsymbol{x}_A) \\
&\leq \phi_{T+1}(\boldsymbol{x}_A) - \phi_A(\boldsymbol{x}_A) \ .
\end{aligned}$$

■

Appendix F. Adaptive Learning Rate

Lemma 26 (Bound on S_∞ for adaptive learning rate) *Let $\alpha \in (\frac{1}{2}, 1)$. If $\eta_t = \frac{G^{2\alpha-1}}{\left(2G^2+\sum_{i=1}^{t-1}\|\boldsymbol{g}_i\|^2\right)^\alpha}$ then*

$$S_\infty \leq \sqrt{4 + \frac{1}{2\alpha - 1}} \,.$$

Proof For any $T \geq 1$,

$$\begin{aligned}
S_T^2 &= 4 + \sum_{t=1}^{T} \|\boldsymbol{\ell}_t\|^2 \\
&= 4 + \sum_{t=1}^{T} \eta_t^2 \|\boldsymbol{g}_t\|^2 \\
&= 4 + G^{4\alpha-2} \sum_{t=1}^{T} \frac{\|\boldsymbol{g}_t\|^2}{(2G^2 + \sum_{i=1}^{t-1} \|\boldsymbol{g}_i\|^2)^{2\alpha}} \\
&\leq 4 + G^{4\alpha-2} \sum_{t=1}^{T} \frac{\|\boldsymbol{g}_t\|^2}{(G^2 + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2)^{2\alpha}} && \text{(since } \|\boldsymbol{g}_t\| \leq G\text{)} \\
&\leq 4 + G^{4\alpha-2} \int_{G^2}^{G^2+\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2} x^{-2\alpha} dx && \text{(Lemma 17 with } f(x) = x^{-2\alpha}, a_0 = G^2, a_t = \|\boldsymbol{g}_t\|^2\text{)} \\
&= 4 + G^{4\alpha-2} \frac{G^{2-4\alpha} - (G^2 + \sum_{t=j}^{T} \|\boldsymbol{g}_t\|^2)^{1-2\alpha}}{2\alpha - 1} \\
&\leq 4 + \frac{1}{2\alpha - 1} \,.
\end{aligned}$$

The lemma follows by taking limit $T \to \infty$. ■

Theorem 27 (Convergence rate for adaptive learning rate) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth convex function with a minimizer $\boldsymbol{x}^*$. Suppose the stochastic gradients satisfy (1), (2) and $\mathbf{E}[\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|^2 \mid \mathcal{F}_t] \leq \sigma^2$. Algorithm 1 with learning rate*

$$\eta_t = \frac{G^{2\alpha-1}}{\left(2G^2 + \sum_{i=1}^{t-1} \|\boldsymbol{g}_i\|^2\right)^\alpha} \,,$$

where $\alpha \in (\frac{1}{2}, 1)$ satisfies for all $T \geq 1$,

$$\begin{aligned}
\mathbf{E}\left[(F(\overline{\boldsymbol{x}}_T) - F(\boldsymbol{x}^*))^{1-\alpha}\right] \leq \frac{1}{T^{1-\alpha}} \max\Big\{&2^\alpha G^{(1-2\alpha)(1-\alpha)} \left(1 + \phi_\infty(\boldsymbol{x}^* - \boldsymbol{x}_0)\right)^{1-\alpha} \left(2G^2 + 2(T-1)\sigma^2\right)^{\alpha(1-\alpha)}, \\
&G^{1-2\alpha} 2^{\frac{\alpha}{1-\alpha}} \left(1 + \phi_\infty(\boldsymbol{x}^* - \boldsymbol{x}_0)\right) (4L)^\alpha \Big\} \,.
\end{aligned}$$

where

$$\overline{\boldsymbol{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \,.$$

Proof Lemma 6 states that for any $\boldsymbol{u} \in \mathbb{R}^d$,

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t \left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \right\rangle\right] \leq \mathbf{E}\left[1 + \phi_\infty(\boldsymbol{u})\right] \ .$$

Since F is convex, $F(\boldsymbol{x}_t) - F(\boldsymbol{u}) \leq \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \rangle$ and therefore

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t (F(\boldsymbol{x}_t) - F(\boldsymbol{u}))\right] \leq \mathbf{E}\left[1 + \phi_\infty(\boldsymbol{u})\right] \ .$$

Substituting $\boldsymbol{x}^*$ for $\boldsymbol{u}$, we have

$$\sum_{t=1}^{T} \mathbf{E}\left[\eta_t (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))\right] \leq \mathbf{E}\left[1 + \phi_\infty(\boldsymbol{x}^*)\right] \ .$$

Now observe that Hölder's inequality implies that $\mathbf{E}[B^p] \geq \frac{\mathbf{E}[AB]^p}{\mathbf{E}[A^q]^{p/q}}$ for all A, B non-negative random variables, $p, q \geq 1$, and $\frac{1}{p} + \frac{1}{q} = 1$. Using it with $B = \left(\eta_T \left(\sum_{t=1}^{T} (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))\right)\right)^{1-\alpha}$ and $A = \eta_T^{\alpha-1}$ and using the fact that the learning rates are decreasing and $F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)$ are non-negative, we have

$$\mathbf{E}\left[1 + \phi_\infty(\boldsymbol{x}^*)\right] \geq \mathbf{E}\left[\eta_T \sum_{t=1}^{T} (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))\right] \geq \frac{\mathbf{E}\left[\left(\sum_{t=1}^{T} (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))\right)^{1-\alpha}\right]^{\frac{1}{1-\alpha}}}{\mathbf{E}\left[\left(\frac{1}{\eta_T}\right)^{\frac{1-\alpha}{\alpha}}\right]^{\frac{\alpha}{1-\alpha}}} \ .$$

Now, observe that

$$\begin{aligned}
\mathbf{E}\left[\left(\frac{1}{\eta_T}\right)^{\frac{1-\alpha}{\alpha}}\right] &= G^{(1-2\alpha)(1-\alpha)/\alpha}\, \mathbf{E}\left[\left(2G^2 + \sum_{t=1}^{T-1} \|\boldsymbol{g}_t\|^2\right)^{1-\alpha}\right] \\
&\leq G^{(1-2\alpha)(1-\alpha)/\alpha}\, \mathbf{E}\left[\left(2G^2 + 2\sum_{t=1}^{T-1} (\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|^2 + \|\nabla F(\boldsymbol{x}_t)\|^2)\right)^{1-\alpha}\right] \\
&\leq G^{(1-2\alpha)(1-\alpha)/\alpha} \left(2G^2 + 2(T-1)\sigma^2\right)^{1-\alpha} + G^{(1-2\alpha)(1-\alpha)/\alpha}\, \mathbf{E}\left[\left(2\sum_{t=1}^{T-1} \|\nabla F(\boldsymbol{x}_t)\|^2\right)^{1-\alpha}\right] \\
&\leq G^{(1-2\alpha)(1-\alpha)/\alpha} \left(2G^2 + 2(T-1)\sigma^2\right)^{1-\alpha} + G^{(1-2\alpha)(1-\alpha)/\alpha}\, \mathbf{E}\left[\left(2\sum_{t=1}^{T} \|\nabla F(\boldsymbol{x}_t)\|^2\right)^{1-\alpha}\right] \\
&\leq G^{(2\alpha-1)(1-\alpha)/\alpha} \left(2G^2 + 2(T-1)\sigma^2\right)^{1-\alpha} + G^{(1-2\alpha)(1-\alpha)/\alpha}\, \mathbf{E}\left[\left(4L\sum_{t=1}^{T} (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))\right)^{1-\alpha}\right] .
\end{aligned}$$

Putting all together and denoting by $\Delta = \sum_{t=1}^{T} (F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*))$, we have

$$\mathbf{E}\left[\Delta^{1-\alpha}\right]^{\frac{1}{\alpha}} \leq G^{(1-2\alpha)(1-\alpha)/\alpha} \left(\mathbf{E}\left[1 + \phi_\infty(\boldsymbol{u})\right]\right)^{\frac{1-\alpha}{\alpha}} \left((2G^2 + 2(T-1)\sigma^2)^{1-\alpha} + (4L)^{1-\alpha}\, \mathbf{E}\left[\Delta^{1-\alpha}\right]\right) \ .$$

With a case analysis, we have

$$\mathbf{E}\left[\Delta^{1-\alpha}\right] \leq \max\Big(2^{\alpha} G^{(1-2\alpha)(1-\alpha)} \left(\mathbf{E}\left[1+\phi_{\infty}(\boldsymbol{u})\right]\right)^{1-\alpha} (2G^2+2(T-1)\sigma^2)^{\alpha(1-\alpha)}, \\ G^{1-2\alpha} 2^{\frac{\alpha}{1-\alpha}}\, \mathbf{E}\left[1+\phi_{\infty}(\boldsymbol{u})\right] (4L)^{\alpha}\Big) \ .$$

Jensen's inequality implies that $F(\overline{\boldsymbol{x}}_T) \leq \frac{1}{T}\sum_{t=1}^{T} F(\boldsymbol{x}_t)$, that gives the final bound. ■