Proceedings of Machine Learning Research vol 65:1–35, 2017

Online Learning Without Prior Information

Ashok Cutkosky ASHOKC@CS.STANFORD.EDU

Kwabena Boahen BOAHEN@STANFORD.EDU

Stanford University

Abstract

The vast majority of optimization and online learning algorithms today require some prior information about the data (often in the form of bounds on gradients or on the optimal parameter value). When this information is not available, these algorithms require laborious manual tuning of various hyperparameters, motivating the search for algorithms that can adapt to the data with no prior information. We describe a frontier of new lower bounds on the performance of such algorithms, reflecting a tradeoff between a term that depends on the optimal parameter value and a term that depends on the gradients' rate of growth. Further, we construct a family of algorithms whose performance matches any desired point on this frontier, which no previous algorithm reaches.

1. Problem Definition and Prior Work

Data streams, large datasets, and adversarial environments require online optimization algorithms, which continually adapt model parameters to the data. At iteration t, these algorithms pick a point $w_t \in W$, are presented with a loss function $\ell_t : W \to \mathbb{R}$, and suffer loss $\ell_t(w_t)$. The algorithm's performance is measured by *regret*, which is defined as the loss relative to some comparison point u:

$$R_T(u) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u)$$

When W is a convex set and the ℓ_t are guaranteed to be convex, the regret can be minimized using only information about the gradients of ℓ_t at w_t, leading to simple and efficient algorithms.

All online convex optimization algorithms require either a bound B on the diameter of W or a bound $L_{\max}$ on the gradients of ℓ_t, or suffer a penalty that is exponential in gradients' rate of growth when no information is given (Cutkosky and Boahen, 2016). When B is known but $L_{\max}$ is unknown, there are algorithms that can obtain regret $O(BL_{\max}\sqrt{T})$ (Duchi et al., 2010; McMahan and Streeter, 2010). Conversely, when B is infinite (e.g. W is an entire vector space) but $L_{\max}$ is known, there are algorithms that obtain $O(\|u\|L_{\max}\sqrt{T\log(\|u\|T)})$ or $O(\|u\|L_{\max}\sqrt{T}\log(\|u\|T))$ regret (McMahan and Streeter, 2012; Orabona, 2013; McMahan and Abernethy, 2013; Orabona, 2014; Orabona and Pál, 2016a). The situation does not improve when both B and $L_{\max}$ are known. In this case it is impossible to do better than $O(BL_{\max}\sqrt{T})$, so knowing just one of these parameters is essentially as good as knowing both (Abernethy et al., 2008). In the case where no prior information is given, it was recently proved by Cutkosky and Boahen (2016) that the regret must contain an additional exponential penalty $\exp(\max_t \sqrt{L_t/L_{t-1}})$, where L_t is the maximum gradient observed by iteration t.

The case in which we have no bound on either B or $L_{\max}$ is common in practice. A standard pragmatic approach to this lack of information is to simply make a guess for these parameters and then apply an algorithm that uses the guess as input, but this approach is theoretically unsound in online learning, and rather laborious and inelegant in general. We explore lower bounds and algorithms that adapt to the unknown quantities in a principled way in this paper.

Where no information is given, we prove that there is a frontier of matching lower and upper bounds on $R_T(u)$ that trades-off a $\|u\|L_{\max}\sqrt{T}\log(\|u\|T)$ term with a $\exp(\max_t L_t/L_{t-1})$ term along two dimensions, which we parametrize by k and γ.[1] Along the first dimension, the exponential penalty is reduced to $\exp((L_t/L_{t-1})/k^2)$ for any $k > 0$ at the expense of rescaling the regret's $\sqrt{T}$ term to $k\|u\|L_{\max}\sqrt{T}\log(\|u\|T)$. Along the second dimension, the logarithm's power in the $\sqrt{T}$ term is reduced to $\|u\|L_{\max}\sqrt{T}\log^{\gamma}(\|u\|T)$ for any $\gamma \in (1/2, 1]$ at the expense of increasing the exponential penalty to $\exp((L_t/L_{t-1})^{1/(2\gamma-1)})$. We prove the lower bounds by constructing a specific adversarial loss sequence, and we prove the upper bounds by providing a family of algorithms whose regret matches the lower bound frontier for any k and γ.

2. Notation and Setup

Before proceeding further, we provide a few definitions that will be useful throughout this paper. A set W is a *convex set* if W is a subset of some real vector space and $tx + (1-t)y \in W$ for all $x, y \in W$ and $t \in [0, 1]$. Throughout this paper we will assume that W is closed. A function f is a *convex function* if $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$ for all x, y and $t \in [0, 1]$. If $f : V \to \mathbb{R}$ for some vector space V, then a vector $g \in V^*$ is a *subgradient* of f at x, denoted $g \in \partial f(x)$, if $f(y) \geq f(x) + g \cdot (y - x)$ for all y. Here we use the dot product to indicate application of linear functionals in the dual space since this should cause no confusion. A *norm* $\|\cdot\|$ is a function such that $\|x\| = 0$ if and only if $x = 0$, $\|cx\| = |c|\|x\|$ for any scalar c, and $\|x + y\| \leq \|x\| + \|y\|$ for all x and y. The *dual norm* is a norm $\|\cdot\|_\star$ defined by $\|x\|_\star = \sup_{\|y\|=1} x \cdot y$. As a special case, when $\|x\| = \sqrt{x \cdot x}$ (the L_2 norm), then $\|\cdot\|_\star = \|\cdot\|$.

Online convex optimization problems can be reduced to online *linear* optimization problems in which the loss functions are constrained to be linear functions. The reduction follows by replacing the loss function $\ell_t(w)$ with the linear function $g_t \cdot w$, where g_t is a subgradient of ℓ_t at w_t. Then, by definition, $g_t \cdot w_t - g_t \cdot u \geq \ell_t(w_t) - \ell_t(u)$. Therefore the regret of our algorithm with respect to the linear loss functions $g_t \cdot w$ is an upper-bound on the regret with respect to the real loss functions ℓ_t. Because of this reduction, many online convex optimization algorithms (including ours) are *first order* algorithms, meaning they access the loss functions only through their subgradients. For the rest of this paper we will therefore assume that the losses are linear, $\ell_t(w) = g_t \cdot w$.

We will focus all of our lower bounds in Section 3 and algorithms in Section 5 on the case in which the domain W is an entire Hilbert space, so that W has infinite diameter and no boundary. This case is very common in practical optimization optimization problems encountered in machine learning, in which any constraints are often only implicitly enforced via regularization. Our objective is to design lower bounds and algorithms such that $R_T(u)$ depends on $\|u\|$, T, and $L_{\max}$ without prior knowledge of these parameters.

1. The square root is missing from the exponential term because we improved the lower bound given in Cutkosky and Boahen (2016) (see Section 3).

In the following sections we use a compressed-sum notation where subscripts with colons indicate summations: $\sum_{t=1}^{T} g_t = g_{1:T}$, $\sum_{t=1}^{T} \|g_t\|^2 = \|g\|_{1:T}^2$, $\sum_{t=1}^{T} g_t w_t = (gw)_{1:T}$ and similarly for other indexed sums. Proofs are in the appendix when they do not immediately follow the result.

3. A Frontier of Lower Bounds

In this section we give our frontier of lower bounds for online optimization without prior information. First we describe our adversarial loss sequence and lower bound frontier along the k dimension, and then we extend the argument to obtain the full two dimensional frontier parametrized by both k and γ.

3.1. Trade-offs in the multiplicative constant k

Given an algorithm, we establish a lower bound on its performance by constructing an adversarial sequence of subgradients $g_t \in \mathbb{R}$. This sequence sets $g_t = -1$ for $T-1$ iterations, where T is chosen adversarially but can be made arbitrarily large, then sets $g_T = O(k\sqrt{T})$. Perhaps surprisingly, we prove that this simple strategy forces the algorithm to experience regret that is exponential in $\sqrt{T}/k$. We then express $\sqrt{T}/k$ as a constant multiple of $\frac{1}{k^2} L_t/L_{t-1}$, where $L_t = \max_{t' \le t} |g_t|$, capturing the algorithm's sensitivity to the big jump in the gradients between $T-1$ and T in the adversarial sequence.

The cost that an algorithm pays when faced with the adversarial sequence is stated formally in the following Theorem.

Theorem 1 *For any $k > 0$, $T_0 > 0$, and any online optimization algorithm picking $w_t \in \mathbb{R}$, there exists a $T > T_0$, a $u \in \mathbb{R}$, and a fixed sequence $g_t \in \mathbb{R}$ on which the regret is:*

$$\begin{aligned} R_T(u) &= \sum_{t=1}^{T} g_t w_t - g_t u \\ &\ge k\|u\| L_{\max} \log(T\|u\| + 1)\sqrt{T} + \frac{L_{\max}}{T-1} \exp\left(\frac{\sqrt{T-1}}{8k}\right) \\ &\ge k\|u\| L_{\max} \log(T\|u\| + 1)\sqrt{T} + \max_{t \le T} L_{\max} \frac{L_{t-1}^2}{\|g\|_{1:t-1}^2} \exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288k^2}\right)\right] \end{aligned}$$

where $L_t = \max_{t' \le t} \|g_{t'}\|$, and $L_{\max} = L_T = \max_{t \le T} \|g_t\|$.

The first inequality in this bound demonstrates that it is impossible to guarantee sublinear regret without prior information while maintaining $O(L_{\max}\|u\| \log(\|u\|))$ dependence on $L_{\max}$ and $\|u\|$,[2] but the second inequality provides hope that if the loss sequence is limited to small jumps in L_t, then we might be able to obtain sublinear regret. Specifically, from the first inequality, observe that in order to bring the exponential term to lower than $O(T)$, the value of k needs to be at least $\Omega(\sqrt{T}/\log(T))$, which causes the non-exponential term to become $O(T)$. However, the second inequality emphasizes that our high regret is the result of a large jump in the value of L_t, so that we might expect to do better if there are no such large jumps. Our upper bounds are given in the form of algorithms that guarantee regret matching the second inequality of this lower bound for any k, showing that we can indeed do well without prior information so long as L_t does not increase too quickly.

2. it *is* possible to guarantee sublinear regret in exchange for $O(L_{\max}\|u\|^2)$ dependence, see Orabona and Pál (2016b)

3.2. Trade-offs in the Logarithmic exponent γ

To extend the frontier to the γ dimension, we modify our adversarial sequence by setting $g_T = O(\gamma k^{1/\gamma} T^{1-1/2\gamma})$ instead of $O(k\sqrt{T})$. This results in a penalty that is exponential in $(\sqrt{T}/k)^{1/\gamma}$, which we express as a multiple of $(L_t/\gamma k^2 L_{t-1})^{1/(2\gamma-1)}$. Since $\gamma \in (1/2, 1]$, we are getting a larger exponential penalty even though the adversarial subgradients have decreased in size, illustrating that decreasing the logarithmic factor is very expensive.

The full frontier is stated formally in the following Theorem.

Theorem 2 *For any $\gamma \in (1/2, 1]$, $k > 0$, $T_0 > 0$, and any online optimization algorithm picking $w_t \in \mathbb{R}$, there exists a $T > T_0$, a $u \in \mathbb{R}$, and a sequence $g_1, \dots, g_T \in \mathbb{R}$ with $\|g_t\| \leq \max(1, 18\gamma(4k)^{1/\gamma}(t-1)^{1-1/2\gamma})$ on which the regret is:*[3]

$$\begin{aligned} R_T(u) &= \sum_{t=1}^{T} g_t w_t - g_t u \\ &\geq k\|u\| L_{\max} \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \frac{L_{\max}}{T-1} \exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) \\ &\geq k\|u\| L_{\max} \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \max_{t \leq T} L_{\max} \frac{L_{t-1}^2}{\|g\|_{1:t-1}^2} \exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right] \end{aligned}$$

where $L_t = \max_{t' \leq t} \|g_{t'}\|$ and $L_{\max} = L_T = \max_{t \leq T} \|g_t\|$.

Again, the first inequality tells us that adversarial sequences can always deny the algorithm sublinear regret and the second inequality says that so long as L_t grows slowly, we can still hope for sublinear regret. This time, however, the second inequality appears to blow up when $\gamma \to 1/2$. In this case, $L_{\max} = O(k^2)$ regardless of T and so the value of L_t/L_{t-1} is never very large, keeping the exponent in the second inequality less than 1 so that the singularity in the exponent does not send the bound to infinity. This singularity at $\gamma = 1/2$ tells us that the adversary does not need to be "very adversarial" in order to force us to experience exponential regret.

To gain some more intuition for what happens at $\gamma = 1/2$, consider a model in which the adversary must commit ahead of time to some $L_{\max}$ (which corresponds to picking k), unknown to the optimization algorithm, such that $\|g_t\| \leq L_{\max}$ for all t. When a bound $L_{\text{bound}} \geq L_{\max}$ is known to the algorithm ahead of time, then it is possible to achieve $O(\|u\| L_{\text{bound}} \sqrt{T \log(\|u\|T)})$ regret (e.g. see Orabona and Pál (2016a)). However, note that when $\gamma = 1/2$, committing to an appropriate $L_{\max}$ would not prevent an adversary from using the sequence of Theorem 2. Therefore, Theorem 2 tells us that algorithms which achieve $O(\|u\| L_{\text{bound}} \sqrt{T \log(\|u\|T)})$ regret are inherently very fragile because if the bound is incorrect (which happens for large enough k), then the adversary can force the algorithm to suffer $L_{\max} \exp(O(T/L_{\max}))$ regret for arbitrarily large T.

Continuing with the model in which the adversary must commit to some unknown $L_{\max}$ ahead of time, suppose we are satisfied with $O(\|u\| L_{\max} \sqrt{T} \log^{\gamma}(\|u\|T))$ regret for some $\gamma > 1/2$. In this case, after some (admittedly possibly very large) number of iterations, the exponential term in the second inequality no longer grows with T, and the adversarial strategy of Theorem 2 is not available because this strategy requires a choice of $L_{\max}$ that depends on T. Therefore an algorithm

3. The same result holds with in expectation for randomized algorithms with a deterministic sequence g_t.

that guarantees regret matching the second inequality for some k and γ will obtain an asymptotic dependence on T that is only $\log^\gamma(T)\sqrt{T}$.

These lower bounds show that there is a fundamental frontier of tradeoffs the between parameters γ and k and the exponential penalty. Now we proceed to derive algorithms that match any point on the frontier without prior information.

4. Regret Analysis without Information

In this section we provide the tools used to derive algorithms whose regret matches the lower bounds in the previous section. Our algorithms make use of the Follow-the-Regularized-Leader (FTRL) framework, which is an elegant and intuitive way to design online learning algorithms (see Shalev-Shwartz (2011); McMahan (2014) for detailed discussions). After seeing the t^{th} loss of the online learning game, an FTRL algorithm chooses a function ψ_t (called a *regularizer*), and picks w_{t+1} according to:

$$w_{t+1} = \operatorname*{argmin}_{w \in W} \psi_t(w) + \sum_{t'=1}^{t} \ell_{t'}(w)$$

Careful choice of regularizers is obviously crucial to the success of such an algorithm, and in the following we provide simple conditions on ψ sufficient for FTRL to achieve optimal regret without prior information. Our analysis generalizes many previous works for online learning with unconstrained W (e.g. Orabona (2013, 2014); Cutkosky and Boahen (2016)) in which regret bounds were proved via arduous ad-hoc constructions. Further, our techniques improve the regret bound in the algorithm that does not require prior information of Cutkosky and Boahen (2016). We note that an alternative set of conditions on regularizers was given in Orabona and Pál (2016a) via an elegant reduction to coin-betting algorithms, but this prior analysis requires a known bound on $L_{\max}$.

Our regularizers ψ_t take the form $\psi_t(w) = \frac{k}{a_t \eta_t}\psi(a_t w)$ for some fixed function ψ and numbers a_t and η_t. The value k specifies the corresponding tradeoff parameter in the lower-bound frontier, while the function ψ specifies the value of γ. The values for a_t and η_t do not depend on k or ψ, but are carefully chosen functions of the observed gradients $g_1, \dots, g_t$ that guarantee the desired asymptotics in the regret bound.

4.1. Generalizing Strong Convexity

Prior analyses of FTRL often make use of strongly-convex regularizers to simplify regret analysis, but it turns out that strongly-convex regularizers cannot match our lower bounds. Fortunately, there is a simple generalization of strong-convexity that will suffice for our purposes. This generalized notion is very similar to a dual version of the "local smoothness" condition used in Orabona (2013). We define this generalization of strong-convexity below.

Definition 3 *Let W be a convex space and let $\sigma : W^2 \to \mathbb{R}$ by an arbitrary function. We say a convex function $f : W \to \mathbb{R}$ is $\sigma(\cdot,\cdot)$-strongly convex with respect to a norm $\|\cdot\|$ if for all $x, y \in W$ and $g \in \partial f(x)$ we have*

$$f(y) \geq f(x) + g \cdot (y - x) + \frac{\sigma(x,y)}{2}\|x - y\|^2$$

As a special case (and by abuse of notation), for any function $\sigma : W \to \mathbb{R}$ we define $\sigma(w, z) = \min(\sigma(w), \sigma(z))$ and define $\sigma(\cdot)$-strong convexity accordingly.

We'll usually just write σ-strongly convex instead of $\sigma(\cdot, \cdot)$-strongly convex since our definition is purely a generalization of the standard one. We will also primarily make use of the special case $\sigma(w, z) = \min(\sigma(w), \sigma(z))$.

4.2. Adaptive regularizers

Now we present a few definitions that will allow us to easily construct sequences of regularizers that achieve regret bounds without information. Intuitively, we require that our regularizers ψ_t grow super-linearly in order to ensure that $\psi_t(w)+g_{1:t}w$ always has a minimal value. However, we do not want ψ_t to grow quadratically because this will result in $O(\|u\|^2)$ regret. The formal requirements on the shape of ψ_t are presented in the following definition:

Definition 4 *Let W be a closed convex subset of a vector space such that $0 \in W$. Any differentiable function $\psi : W \to \mathbb{R}$ that satisfies the following conditions:*

1. *$\psi(0) = 0$.*
2. *$\psi(x)$ is σ-strongly-convex with respect to some norm $\|\cdot\|$ for some $\sigma : W \to \mathbb{R}$ such that $\|x\| \geq \|y\|$ implies $\sigma(x) \leq \sigma(y)$.*
3. *For any C, there exists a B such that $\psi(x)\sigma(x) \geq C$ for all $\|x\| \geq B$.*

*is called a $(\sigma, \|\cdot\|)$-*adaptive regularizer. *We also define the useful auxiliary function $h(w) = \psi(w)\sigma(w)$ and by mild abuse of notation, we define $h^{-1}(x) = \max_{h(w)\leq x} \|w\|$.*

We will use adaptive regularizers as building blocks for our FTRL regularizers ψ_t, so it is important to have examples of such functions. We will provide some tools for finding adaptive regularizers in Section 5, but to keep an example in mind for now, we remark that $\psi(w) = (\|w\| + 1)\log(\|w\| + 1) - \|w\|$ is a $\left(\frac{1}{\|\cdot\|+1}, \|\cdot\|\right)$-adaptive regularizer where $\|\cdot\|$ is the L_2 norm.

The following definition specifies the sequences η_t and a_t which we use to turn an adaptive regularizer into the regularizers used for our FTRL algorithms:

Definition 5 *Let $\|\cdot\|$ be a norm and $\|\cdot\|_\star$ be the dual norm ($\|x\|_\star = \sup_{\|y\|=1} x \cdot y$). Let $g_1, \ldots, g_T$ be a sequence of subgradients and set $L_t = \max_{t'\leq t} \|g_t\|_\star$. Define the sequences $\frac{1}{\eta_t}$ and a_t recursively by:*

$$\begin{aligned}
\frac{1}{\eta_0^2} &= 0 \\
\frac{1}{\eta_t^2} &= \max\left(\frac{1}{\eta_{t-1}^2} + 2\|g_t\|_\star^2, L_t\|g_{1:t}\|_\star\right) \\
a_1 &= \frac{1}{(L_1\eta_1)^2} \\
a_t &= \max\left(a_{t-1}, \frac{1}{(L_t\eta_t)^2}\right)
\end{aligned}$$

Suppose ψ is a $(\sigma, \|\cdot\|)$-adaptive regularizer and $k > 0$. Define

$$\psi_t(w) = \frac{k}{\eta_t a_t}\psi(a_t w)$$
$$w_{t+1} = \operatorname*{argmin}_{w \in W} \psi_t(w) + g_{1:t} \cdot w$$

Now without further ado, we give our regret bound for FTRL using these regularizers.

Theorem 6 *Suppose ψ is a $(\sigma, \|\cdot\|)$-adaptive regularizer and $g_1, \dots, g_T$ is some arbitrary sequence of subgradients. Let $k \geq 1$, and let ψ_t be defined as in Definition 5.*

Set

$$\sigma_{min} = \inf_{\|w\| \leq h^{-1}(10/k^2)} k\sigma(w)$$
$$D = \max_t \frac{L_{t-1}^2}{(\|g\|_\star^2)_{1:t-1}} h^{-1}\left(\frac{5L_t}{k^2 L_{t-1}}\right)$$
$$Q_T = 2\frac{\|g\|_{1:T}}{L_{\max}}$$

Then FTRL with regularizers ψ_t achieves regret

$$R_T(u) \leq \frac{k}{Q_T \eta_T}\psi(Q_T u) + \frac{45 L_{\max}}{\sigma_{min}} + 2L_{\max} D$$
$$\leq k L_{\max}\frac{\psi(2uT)}{\sqrt{2T}} + \frac{45 L_{\max}}{\sigma_{min}} + 2L_{\max} D$$

This bound consists of three terms, the first of which will correspond to the $\sqrt{T}$ term in our lower bounds and the last of which will correspond to the exponential penalty. The middle term is a constant independent of u and T. To unpack a specific instantiation of this bound, consider the example adaptive regularizer $\psi(w) = (\|w\| + 1)\log(\|w\| + 1) - \|w\|$. For this choice of ψ, we have $\psi(2uT)/\sqrt{2T} = O(\|u\|\sqrt{T}\log(T\|u\| + 1))$ so that the first term in the regret bound matches the $\sqrt{T}$ term in our lower bound with $\gamma = 1$. Roughly speaking, $h(w) \approx \log(w)$, so that $h^{-1}(x) \approx \exp(x)$ and the quantity $D = \max_t \frac{L_{t-1}^2}{(\|g\|_\star^2)_{1:t-1}} h^{-1}\left(\frac{5L_t}{k^2 L_{t-1}}\right)$ matches the exponential penalty in our lower bound. In the following section we formalize this argument and exhibit a family of adaptive regularizers that enable us to design algorithms whose regret matches any desired point on the lower bound frontier.

5. Optimal Algorithms

In this section we construct specific adaptive regularizers in order to obtain optimal algorithms using our regret upper bound of Theorem 6. The results in the previous section hold for arbitrary norms, but from this point on we will focus on the L_2 norm. Our regret upper bound expresses regret in terms of the function h^{-1}. Inspection of the bound shows that if $h^{-1}(x)$ is exponential in $x^{1/(2\gamma-1)}$, and $\psi(w) = O(\|w\|\log^\gamma(\|w\| + 1))$, then our upper bound will match (the second inequality in) our lower bound frontier. The following Collary formalizes this observation.

Corollary 7 *If ψ is an $(\sigma, \|\cdot\|)$-adaptive regularizer such that*

$$\psi(x)\sigma(x) \geq \Omega(\gamma \log^{2\gamma-1}(\|x\|))$$
$$\psi(x) \leq O(\|x\| \log^{\gamma}(\|x\|+1))$$

then for any $k \geq 1$, FTRL with regularizers $\psi_t(w) = \frac{k}{a_t \eta_t}\psi(a_t w)$ yields regret

$$R_T(u) \leq O\left[kL_{\max}\sqrt{T}\|u\| \log^{\gamma}(T\|u\|+1) + \max_t \frac{L_{\max}L_{t-1}^2}{\|g\|_{1:t-1}^2} \exp\left[O\left(\left(\frac{L_t}{k^2\gamma L_{t-1}}\right)^{1/(2\gamma-1)}\right)\right]\right]$$

We call regularizers that satisfy these conditions γ-optimal.

With this Corollary in hand, to match our lower bound frontier we need only construct a γ-optimal adaptive regularizer for all $\gamma \in (1/2, 1]$. Constructing adaptive regularizers is made much simpler with Proposition 8 below. This proposition allows us to design adaptive regularizers in high dimensional spaces by finding simple one-dimensional functions. It can be viewed as taking the place of arguments in prior work (McMahan and Orabona, 2014; Orabona and Pál, 2016a; Cutkosky and Boahen, 2016) that reduce high dimensional problems to one-dimensional problems by identifying a "worst-case" direction for each subgradient g_t.

Proposition 8 *Let $\|\cdot\|$ be the L_2 norm ($\|w\| = \|w\|_2 = \sqrt{w \cdot w}$). Let ϕ be a three-times differentiable function from the non-negative reals to the reals that satisfies*

1. $\phi(0) = 0$.
2. $\phi'(x) \geq 0$.
3. $\phi''(x) \geq 0$.
4. $\phi'''(x) \leq 0$.
5. $\lim_{x\to\infty} \phi(x)\phi''(x) = \infty$.

Then $\psi(w) = \phi(\|w\|)$ is a $(\phi''(\|\cdot\|), \|\cdot\|)$-adaptive regularizer.

Now we are finally ready to derive our first optimal regularizer:

Proposition 9 *Let $\|\cdot\|$ be the L_2 norm. Let $\phi(x) = (x+1)\log(x+1) - x$. Then $\psi(w) = \phi(\|w\|)$ is a 1-optimal, $(\phi''(\|\cdot\|), \|\cdot\|)$-adaptive regularizer.*

Proof We can use Proposition 8 to prove this with a few simple calculations:

$$\phi(0) = 0$$
$$\phi'(x) = \log(x+1)$$
$$\phi''(x) = \frac{1}{x+1}$$
$$\phi'''(x) = -\frac{1}{(x+1)^2}$$
$$\phi(x)\phi''(x) = (\log(x+1) - \frac{x}{x+1})$$

Now the conclusion of the Proposition is immediate from Proposition 8 and inspection of the above equations. ■

A simple application of Corollary 7 shows that FTRL with regularizers $\psi_t(w) = \frac{k}{\eta_t}((\|w\| + 1)\log(\|w\| + 1) - \|w\|)$ matches our lower bound with $\gamma = 1$ for any desired k.

In fact, the result of Proposition 9 is a more general phenomenon:

Proposition 10 *Let $\|\cdot\|$ be the L_2 norm. Given $\gamma \in (1/2, 1]$, set $\phi(x) = \int_0^x \log^\gamma(z+1)\,dz$. Then $\psi(w) = \phi(\|w\|)$ is a γ-optimal, $(\phi''(\|\cdot\|), \|\cdot\|)$-adaptive regularizer.*

Proof

$$\begin{aligned}
\phi(0) &= 0 \\
\phi'(x) &= \log^\gamma(x+1) \\
\phi''(x) &= \gamma\frac{\log^{\gamma-1}(x+1)}{x+1} \\
\phi'''(x) &= \gamma(\gamma-1)\frac{\log^{\gamma-2}(x+1)}{(x+1)^2} - \gamma\frac{\log^{\gamma-1}(x+1)}{(x+1)^2}
\end{aligned}$$

Since $\gamma \le 1$, $\phi'''(x) \le 0$ and so ϕ satisfies the first four conditions of Proposition 8. It remains to characterize $\phi(x)$ and $\phi(x)\phi''(x)$, which we do by finding lower and upper bounds on $\phi(x)$:

For a lower bound, we have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dx}x\log^\gamma(x+1) &= \frac{1}{2}\left(\log^\gamma(x+1) + \gamma\frac{x}{x+1}\log^{\gamma-1}(x+1)\right) \\
&\le \log^\gamma(x+1)
\end{aligned}$$

where the inequality follows since $\frac{x}{x+1} \le \log(x+1)$, which can be verified by differentiating both sides. Therefore $\phi(x) \ge \frac{1}{2}x\log^\gamma(x+1)$. This lower-bound implies

$$\phi(x)\phi''(x) \ge \frac{1}{2}\gamma\frac{x}{x+1}\log^{2\gamma-1}(x+1)$$

which gives us the last condition in Proposition 8, as well as the first condition for γ-optimality.

Similarly, we have

$$\begin{aligned}
\frac{d}{dx}x\log^\gamma(x+1) &= \left(\log^\gamma(x+1) + \gamma\frac{x}{x+1}\log^{\gamma-1}(x+1)\right) \\
&\ge \log^\gamma(x+1)
\end{aligned}$$

This implies $\phi(x) \le x\log(x+1)$ which gives us the second condition for γ-optimality. ■

Thus, by applying Theorem 6 to the regularizers of Proposition 10, we have a family of algorithms that matches our family of lower-bounds up to constants. The updates for these regularizers are extremely simple:

$$w_{t+1} = -\frac{g_{1:t}}{a_t\|g_{1:t}\|}\left[\exp\left((\eta_t\|g_{1:t}\|/k)^{1/\gamma}\right) - 1\right]$$

The guarantees of Theorem 6 do not make any assumptions on how k is chosen, so that we could choose k using prior knowledge if it is available. For example, if a bound on L_t/L_{t-1} is known, we can set $k \geq \sqrt{\max_t L_t/L_{t-1}}$. This reduces the exponentiated quantity $\max_t L_t/k^2 L_{t-1}$ to a constant, leaving a regret of $O(\|u\| \log(T\|u\| + 1) L_{\max} \sqrt{T \max_t L_t/L_{t-1}})$. This bound holds without requiring a bound on $L_{\max}$. Thus our algorithms open up an intermediary realm in which we have no bounds on $\|u\|$ or $L_{\max}$, and yet we can leverage some other information to avoid the exponential penalty.

6. FREEREX

Now we explicitly describe an algorithm, along with a fully worked-out regret bound. The norm $\|\cdot\|$ used in the following is the L_2 norm ($\|w\| = \sqrt{w \cdot w}$), and our algorithm uses the adaptive regularizer $\psi(w) = (\|w\| + 1)\log(\|w\| + 1) - \|w\|$. Similar calculations could be performed for arbitrary γ using the regularizers of Proposition 10, but we focus on the $\gamma = 1$ because it allows for simpler and tighter analysis through our closed-form expression for ψ. Since we do not require any information about the losses, we call our algorithm FREEREX for Information-**free** **R**egret via **ex**ponential updates.

Algorithm 1 FREEREX

Input: k.
Initialize: $\frac{1}{\eta_0^2} \leftarrow 0$, $a_0 \leftarrow 0$, $w_1 \leftarrow 0$, $L_0 \leftarrow 0$, $\psi(w) = (\|w\| + 1)\log(\|w\| + 1) - \|w\|$.
for $t = 1$ **to** T **do**
  Play w_t, receive subgradient $g_t \in \partial \ell_t(w_t)$.
  $L_t \leftarrow \max(L_{t-1}, \|g_t\|)$.
  $\frac{1}{\eta_t^2} \leftarrow \max\left(\frac{1}{\eta_{t-1}^2} + 2\|g_t\|^2, L_t\|g_{1:t}\|\right)$.
  $a_t \leftarrow \max(a_{t-1}, 1/(L_t\eta_t)^2)$.
  //Set w_{t+1} using FTRL update
  $w_{t+1} \leftarrow -\frac{g_{1:t}}{a_t\|g_{1:t}\|}\left[\exp\left(\frac{\eta_t\|g_{1:t}\|}{k}\right) - 1\right]$ // $= \operatorname{argmin}_w \left[\frac{k\psi(a_t w)}{a_t\eta_t} + g_{1:t}w\right]$
end for

Theorem 11 *The regret of* FREEREX *(Algorithm 1) is bounded by*

$$R_T(u) \leq k\|u\|\sqrt{2\|g\|_{1:T}^2 + L_{\max}\max_{t\leq T}\|g_{1:t}\|}\log\left(\frac{2\|g\|_{1:T}}{L_{\max}}\|u\| + 1\right) + \frac{45L_{\max}}{k}\exp(10/k^2 + 1)$$
$$+ 2L_{\max}\max_t \frac{L_{t-1}^2}{\|g\|_{1:t-1}^2}\left[\exp\left(\frac{5L_t}{k^2 L_{t-1}} + 1\right) - 1\right]$$

Proof Define $\phi(x) = (x+1)\log(x+1) - x$. Then $\psi(w) = (\|w\| + 1)\log(\|w\| + 1) - \|w\|$ is a $(\phi''(\|\cdot\|), \|\cdot\|)$-adaptive regularizer by Proposition 9. Therefore we can immediately apply Theorem 6 to obtain

$$R_T(u) \leq \frac{k}{Q_T\eta_T}\psi(Q_T u) + \frac{45L_{\max}}{\phi''_{\min}} + 2L_{\max}D$$

where we've defined $\phi''_{\min} = \inf_{\|w\| \leq h^{-1}(10/k^2)} k\phi''(\|w\|)$.

We can compute (for non-negative x):

$$\begin{aligned}\phi(x) &\leq (x+1)\log(x+1)\\ \phi''(x) &= \frac{1}{x+1}\\ h(w) &= \phi(\|w\|)\phi''(\|w\|) = \left(\log(\|w\|+1) - \frac{\|w\|}{\|w\|+1}\right)\\ &\geq \log(\|w\|+1) - 1\end{aligned}$$

From Proposition 19 (part 2) we have $\frac{1}{\eta_T} \leq \sqrt{2\|g\|_{1:T}^2 + L_{\max}\max_{t\leq T}\|g_{1:t}\|}$. We also have $(\|w\|+1)\log(\|w|+1) - \|w\| = \|w\|\log(\|w\|+1) + \log(\|w\|+1) - \|w\| \leq \|w\|\log(\|w\|+1)$, so we are left with

$$\begin{aligned}R_T(u) &\leq \frac{k}{\eta_T}\|u\|\log(Q_T\|u\|+1) + \sup_{\|w\|\leq h^{-1}(\frac{10}{k^2})}\frac{45(\|w\|+1)}{k} + 2L_{\max}D\\ &= k\sqrt{2\|g\|_{1:T}^2 + L_{\max}\max_{t\leq T}\|g_{1:t}\|}\|u\|\log(a_T\|u\|)+1) + \frac{45L_{\max}}{k}\left[h^{-1}\left(\frac{10}{k^2}\right)+1\right]\\ &\quad + 2L_{\max}D\end{aligned}$$

Now it remains to bound $h^{-1}(10/k^2)$ and D. From our expression for h, we have

$$h^{-1}(x/k^2) \leq \exp\left[\frac{x}{k^2}+1\right] - 1$$

Therefore we have

$$\begin{aligned}h^{-1}(10/k^2) &\leq \exp(10/k^2+1) - 1\\ D &= 2\max_t \frac{L_{t-1}^2}{(\|g\|_\star^2)_{1:t-1}} h^{-1}\left(\frac{5L_t}{k^2 L_{t-1}}\right)\\ &\leq 2\max_t \frac{L_{t-1}^2}{(\|g\|_\star^2)_{1:t-1}}\left[\exp\left(\frac{5L_t}{k^2L_{t-1}}+1\right)-1\right]\end{aligned}$$

Substituting the value $Q_T = 2\frac{\|g\|_{1:T}}{L_{\max}}$, we conclude

$$\begin{aligned}R_T(u) \leq k\sqrt{2\|g\|_{1:T}^2 + L_{\max}\max_{t\leq T}\|g_{1:t}\|}\|u\|\log\left(\frac{2\|g\|_{1:T}}{L_{\max}}\|u\|+1\right)\\ + \frac{45L_{\max}}{k}\exp(10/k^2+1) + 2L_{\max}D\end{aligned}$$

From which the result follows by substituting in our expression for D. ∎

As a specific example, for $k = \sqrt{5}$ we numerically evaluate the bound to get

$$\begin{aligned}R_T(u) \leq \|u\|\sqrt{10\|g\|_{1:T}^2 + 5L_{\max}\max_{t\leq T}\|g_{1:t}\|}\log\left(\frac{2\|g\|_{1:T}}{L_{\max}}\|u\|+1\right) + 405L_{\max}\\ + 2L_{\max}\max_t \frac{L_{t-1}^2}{\|g\|_{1:t-1}^2}\left[\exp\left(\frac{L_t}{L_{t-1}}+1\right)-1\right]\end{aligned}$$

7. Conclusions

We have presented a frontier of lower bounds on the worst-case regret of any online convex optimization algorithm without prior information. This frontier demonstrates a fundamental trade-off at work between $kuL_{\max}\log^{\gamma}(Tu+1)$ and $\exp\left[\left(\max_t \frac{L_t}{\gamma k^2 L_{t-1}}\right)^{\frac{1}{2\gamma-1}}\right]$ terms. We also present some easy-to-use theorems that allow us to construct algorithms that match our lower bound for any chosen k and γ. Note that by virtue of not requiring prior information, our algorithms are nearly hyperparameter-free. They only require the essentially unavoidable trade-off parameters k and γ. Since our analysis does not make assumptions about the loss functions or comparison point u, the parameters k and γ can be freely chosen by the user. Unlike other algorithms that require $\|u\|$ or $L_{\max}$, there are no unknown constraints on these parameters.

Our results also open a new perspective on optimization algorithms by casting using prior information as a tool to avoid the exponential penalty. Previous algorithms that require bounds on the diameter of W or $L_{\max}$ can be viewed as addressing this issue. We show that it also possible to avoid the exponential penalty by using a known bound on $\max_t L_t/L_{t-1}$, leading to a regret of $\tilde{O}(\|u\|L_{\max}\sqrt{T\max_t L_t/L_{t-1}})$.

Although we answer some important questions, there is still much to do in online learning without prior information. For example, it is possible to obtain $O(\|u\|^2 L_{\max}\sqrt{T})$ regret without prior information (Orabona and Pál, 2016b), so it should be possible to extend our lower-bound frontier beyond $\|u\|\log(\|u\|)$. Further, it would be valuable to further characterize the conditions for which the adversary can guarantee regret that is exponential in T. We showed that one such condition is that there must be a large jump in the value of L_t, but there may very well be others. Fully characterizing these conditions should allow us design algorithms that smoothly interpolate between "nice" environments that do not satisfy the conditions and fully adversarial ones that do.

Finally, while our analysis allows for the use of arbitrary norms, we focus our examples on the L_2 norm. It may be interesting to design adaptive regularizers with respect to a more diverse set of norms, or to extend our theory to encompass time-changing norms.

References

Jacob Abernethy, Peter L Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the nineteenth annual conference on computational learning theory*, 2008.

Ashok Cutkosky and Kwabena A Boahen. Online convex optimization with unconstrained domains and losses. In *Advances in Neural Information Processing Systems 29*, pages 748–756, 2016.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Learning Theory (COLT)*, 2010.

Brendan McMahan and Jacob Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, pages 2724–2732, 2013.

Brendan McMahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. In *Advances in neural information processing systems*, pages 2402–2410, 2012.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *arXiv preprint arXiv:1403.3465*, 2014.

H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory (COLT)*, pages 1020–1039, 2014.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, 2010.

Francesco Orabona. Dimension-free exponentiated gradient. In *Advances in Neural Information Processing Systems*, pages 1806–1814, 2013.

Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems 29*, pages 577–585, 2016a.

Francesco Orabona and Dávid Pál. Scale-free online learning. *arXiv preprint arXiv:1601.01974*, 2016b.

Francesco Orabona, Koby Crammer, and Nicolo Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2014.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Appendix A. Lower Bound Proof

Before getting started, we need one technical observation:

Proposition 12 *Let $k > 0$, $\gamma \in (1/2, 1]$. Set*

$$Z_t = \frac{t^{1-1/2\gamma}}{2t}\left[\exp\left(\frac{t^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right]$$

and set $r_t = Z_t - Z_{t-1}$. Then for all sufficiently large T,

$$r_T \geq \frac{Z_{T-1}}{3\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}}$$

Proof We have

$$\left.\frac{d}{dt}\right|_{t=T} Z_t = \frac{1}{4\gamma(4k)^{1/\gamma}T}\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) + \frac{1}{4\gamma}T^{-1-1/2\gamma} - \frac{1}{4\gamma}T^{-1-1/2\gamma}\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right)$$

For sufficiently large T, this quantity is positive and increasing in T. Therefore for sufficiently large T,

$$\begin{aligned}
r_T &\geq \left.\frac{d}{dt}\right|_{t=T-1} Z_t \\
&= \frac{1}{4\gamma(4k)^{1/\gamma}(T-1)}\exp\left(\frac{(T-1)^{1/2\gamma}}{(4k)^{1/\gamma}}\right) + \frac{1}{4\gamma}(T-1)^{-1-1/2\gamma} - \frac{1}{4\gamma}(T-1)^{-1-1/2\gamma}\exp\left(\frac{(T-1)^{1/2\gamma}}{(4k)^{1/\gamma}}\right) \\
&\geq \frac{1}{5\gamma(4k)^{1/\gamma}(T-1)}\exp\left(\frac{(T-1)^{1/2\gamma}}{(4k)^{1/\gamma}}\right) \\
&= \frac{2}{5\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}}\left(Z_{T-1} + \frac{(T-1)^{1-1/2\gamma}}{2(T-1)}\right) \\
&\geq \frac{1}{3\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}}Z_{T-1}
\end{aligned}$$

where the third inequality holds only for sufficiently large T. ■

Now we prove Theorem 2, restated below. Theorem 1 is an immediate consequence of Theorem 2, so we do not prove it seperately.

Theorem 2 *For any $\gamma \in (1/2, 1]$, $k > 0$, $T_0 > 0$, and any online optimization algorithm picking $w_t \in \mathbb{R}$, there exists a $T > T_0$, a $u \in \mathbb{R}$, and a sequence $g_1, \dots, g_T \in \mathbb{R}$ with $\|g_t\| \leq \max(1, 18\gamma(4k)^{1/\gamma}(t-1)^{1-1/2\gamma})$ on which the regret is:*[4]

$$\begin{aligned} R_T(u) &= \sum_{t=1}^{T} g_t w_t - g_t u \\ &\geq k\|u\|L_{\max}\log^{\gamma}(T\|u\|+1)\sqrt{T} + \frac{L_{\max}}{T-1}\exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) \\ &\geq k\|u\|L_{\max}\log^{\gamma}(T\|u\|+1)\sqrt{T} + \max_{t\leq T} L_{\max}\frac{L_{t-1}^2}{\|g\|_{1:t-1}^2}\exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right] \end{aligned}$$

where $L_t = \max_{t'\leq t}\|g_{t'}\|$ and $L_{\max} = L_T = \max_{t\leq T}\|g_t\|$.

Proof We prove the Theorem for randomized algorithms and expected regret, as this does not overly complicate the argument. Our proof technique is very similar to that of (Cutkosky and Boahen, 2016), but we use more careful analysis to improve the bound. Intuitively, the adversarial sequence foils the learner by repeatedly presenting it with the subgradient $g_t = -1$ until the learner's expected prediction $\mathbb{E}[w_t]$ crosses some threshold. If $\mathbb{E}[w_t]$ does not increase fast enough to pass the threshold, then we show that there is some large $u \gg 1$ for which $R_T(u)$ exceeds our bound. However, if $\mathbb{E}[w_t]$ crosses this threshold, then the adversary presents a large positive gradient which forces the learner to have a large $R_T(0)$.

Define $\hat{w}_t = \mathbb{E}[w_t | g_{t'} = -1 \text{ for all } t' < t]$. Without loss of generality, assume $\hat{w}_1 = 0$. Note that $\hat{w}_t$ can be computed by an adversary without access to the algorithm's internal randomness.

Let $S_n = \sum_{t=1}^{n}\hat{w}_t$. Let $Z_t = \frac{t^{1-1/2\gamma}}{2t}\left[\exp\left(\frac{t^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right]$, and set $r_t = Z_t - Z_{t-1}$ Suppose $T_1 > T_0$ is such that

1. For all $t_1 > t_2 > T_1$, $Z_{t_1} > Z_{t_2}$.
2. For all $t > T_1$, $r_t \geq \frac{Z_{t-1}}{3\gamma(4k)^{1/\gamma}(t-1)^{1-1/2\gamma}}$ (by Proposition 12).
3. For all $t > T_1$,
$$\frac{1}{4}\left[\exp\left(\frac{t^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right] \geq \frac{1}{t-1}\exp\left(\frac{(t-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right)$$
4. for all $t > T_1$,
$$\frac{1}{36\gamma(4k)^{1/\gamma}(t-1)}\left[\exp\left(\frac{(t-1)^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right] \geq \frac{1}{(t-1)}\exp\left(\frac{(t-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right)$$
5. For all $t > T_1$,
$$\frac{1}{t-1}\exp\left(\frac{(t-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) \geq \exp\left[\frac{1}{4}\left(\frac{1}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right]$$

4. The same result holds with in expectation for randomized algorithms with a deterministic sequence g_t.

6. For all $t > T_1$,

$$18\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma} \geq 1$$

We consider the quantity $\liminf_{n\to\infty} \frac{S_n}{Z_n}$. There are two cases, either the $\liminf$ is less than 1, or it is not.

Case 1: $\liminf_{n\to\infty} \frac{S_n}{Z_n} < 1$

In this case, there must be some $T > T_1$ such that $S_T < Z_T$. We use the adversarial strategy of simply giving $g_t = -1$ for all $t \leq T$. Because of this, $\mathbb{E}[w_t|g_1, \ldots, g_{t-1}] = \hat{w}_t$ so that

$$\begin{aligned}
\mathbb{E}[R_T(u)] &= \sum_{t=1}^{T} g_t \, \mathbb{E}[w_t|g_1, \ldots, g_{t-1}] - g_t u \\
&= \sum_{t=1}^{T} g_t \hat{w}_t - g_t u \\
&= Tu - S_T \\
&\geq Tu - \frac{T^{1-\frac{1}{2\gamma}}}{2T}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right] \\
&\geq Tu - \frac{1}{2}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right]
\end{aligned}$$

Set $u = \frac{1}{T}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right]$. Then clearly

$$\begin{aligned}
\mathbb{E}[R_T(u)] &\geq Tu - \frac{1}{2}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right] \\
&\geq \frac{1}{2}Tu \\
&= \frac{1}{4}Tu + \frac{1}{4}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right]
\end{aligned}$$

Now observe that we have chosen u carefully so that

$$\sqrt{T} = 4k \log^{\gamma}(Tu + 1)$$

Therefore we can write

$$\begin{aligned}
\mathbb{E}[R_T(u)] &\geq \frac{1}{4}Tu + \frac{1}{4}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right] \\
&= k\|u\| \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \frac{1}{4}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right] \\
&= k\|u\|L_{\max} \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \frac{L_{\max}}{4}\left[\exp\left(\frac{T^{1/2\gamma}}{(4k)^{1/\gamma}}\right) - 1\right]
\end{aligned}$$

where we have used $L_{\max} = 1$ to insert factors of $L_{\max}$ where appropriate.

Observing that $L_t/L_{t-1} = 1$ for all t, we can also easily conclude (using properties 3 and 5 of T_1):

$$\begin{aligned}
\mathbb{E}[R_T(u)] &\geq k\|u\|L_{\max} \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \frac{L_{\max}}{T-1}\exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) \\
&\geq k\|u\|L_{\max} \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \max_{t\leq T} L_{\max} \frac{L_{t-1}^2}{\sum_{t'=1}^{t-1}\|g_{t'}\|^2} \exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right]
\end{aligned}$$

Case 2: $\liminf_{n\to\infty} \frac{S_n}{Z_n} \geq 1$

By definition of $\liminf$, there exists some $T_2 > T_1$ and $Q \geq 1$ such that $S_{T_2} \leq \frac{3}{2}QZ_{T_2}$ and for all $t > T_2$, $S_t > \frac{3Q}{4}Z_t$.

Suppose for contradiction that $\hat{w}_t \leq \frac{Q}{2}r_t$ for all $t > T_2$. Then for all $T > T_2$,

$$\begin{aligned} S_T &= S_{T_2} + \sum_{t=T_2+1}^{T} \hat{w}_t \\ &\leq \frac{3}{2}QZ_{T_2} + \frac{Q}{2}Z_T - \frac{Q}{2}Z_{T_2} \\ &= \frac{Q}{2}Z_T + QZ_{T_2} \end{aligned}$$

Since the second term does not depend on T, this implies that for sufficiently large T, $\frac{S_T}{Z_T} \leq \frac{3}{4}QZ_T$, which contradicts our choice of T_2. Therefore $\hat{w}_t > \frac{Q}{2}r_t$ for some $t > T_2$.

Let T be the the smallest index $T > T_2$ such that $\hat{w}_T > \frac{Q}{2}r_T$. Since $\hat{w}_t \leq \frac{Q}{2}r_t$ for $t < T$, we have

$$S_{T-1} \leq \frac{Q}{2}Z_{T-1} + QZ_{T_2} \leq 2QZ_{T-1}$$

where we have used property 1 of T_1 to conclude $Z_{T_2} \leq Z_{T-1}$.

Our adversarial strategy is to give $g_t = -1$ for $t < T$, then $g_T = 18\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}$. We evaluate the regret at $u = 0$ and iteration T. Since $g_t = -1$ for $t < T$, $\mathbb{E}[w_t|g_1, \ldots, g_{t-1}] = \hat{w}_t$ for $t \leq T$ and so

$$\begin{aligned} \mathbb{E}[R_T(u)] &= -S_{T-1} + g_T w_T \\ &\geq g_T \frac{Q}{2} r_T - 2QZ_{T-1} \\ &\geq \frac{Q}{2} \frac{18\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}}{3\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}} Z_{T-1} - 2QZ_{T-1} \\ &= QZ_{T-1} \\ &\geq Z_{T-1} \end{aligned}$$

where we have used $Q \geq 1$ in the last line. Now we use the fact that $L_{\max} = 18\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}$ (by property 6 of T_1) to write

$$\begin{aligned} \mathbb{E}[R_T(u)] &\geq Z_{T-1} \\ &= \frac{1}{18\gamma(4k)^{1/\gamma)}} \frac{L_{\max}}{T-1} \left[\exp\left(\frac{(T-1)^{1/2\gamma}}{(4k)^{1/\gamma}} \right) - 1 \right] \\ &\geq \frac{L_{\max}}{T-1} \exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}} \right) \end{aligned}$$

where we have used the fourth assumption on T_1 in the last line.

Since we are considering $u = 0$, we can always insert arbitrary multiples of u:

$$\begin{aligned} \mathbb{E}[R_T(u)] &\geq \frac{L_{\max}}{T-1} \exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}} \right) \\ &= k\|u\|L_{\max} \log^{\gamma}(T\|u\| + 1)\sqrt{T} + \frac{L_{\max}}{T-1} \exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}} \right) \end{aligned}$$

Now we relate the quantity in the exponent to L_t/L_{t-1}. We have $L_T = g_T$ and $L_{T-1} = 1$ so that

$$L_T/L_{T-1} = 18\gamma(4k)^{1/\gamma}(T-1)^{1-1/2\gamma}$$

Therefore

$$\begin{aligned}(T-1)^{1/2\gamma} &= \left(\frac{L_T/L_{T-1}}{18\gamma(4k)^{1/\gamma}}\right)^{\frac{1}{2\gamma(1-1/2\gamma)}} \\ &= \left(\frac{L_T/L_{T-1}}{18\gamma(4k)^{1/\gamma}}\right)^{1/(2\gamma-1)} \\ \frac{(T-1)^{1/2\gamma}}{(4k)^{1/\gamma}} &= \left(\frac{L_T/L_{T-1}}{18\gamma(4k)^2}\right)^{1/(2\gamma-1)} \\ &= \left(\frac{L_T/L_{T-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\end{aligned}$$

Now observe that $\frac{1}{T-1} = \frac{L_{T-1}^2}{\sum_{t=1}^{T-1}\|g_t\|^2}$ so that we have

$$\frac{L_{\max}}{T-1}\exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) = L_{\max}\frac{L_{T-1}^2}{\sum_{t=1}^{T-1}\|g_t\|^2}\exp\left[\frac{1}{2}\left(\frac{L_T/L_{T-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right]$$

Further, since $\frac{1}{t-1} = \frac{L_{t-1}^2}{\sum_{t'=1}^{t-1}\|g_{t'}\|^2}$ for all $t \le T$, condition 5 on T_1 tells us that

$$\begin{aligned}\frac{L_{\max}}{T-1}\exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) &\ge L_{\max}\exp\left[\frac{1}{2}\left(\frac{1}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right] \\ &= \max_{t\le T-1} L_{\max}\frac{L_{t-1}^2}{\sum_{t'=1}^{t-1}\|g_{t'}\|^2}\exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right]\end{aligned}$$

so that

$$\frac{L_{\max}}{T-1}\exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) = \max_{t\le T} L_{\max}\frac{L_{t-1}^2}{\sum_{t'=1}^{t-1}\|g_{t'}\|^2}\exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right]$$

Therefore we can put everything together to get

$$\begin{aligned}\mathbb{E}[R_T(u)] &\ge k\|u\|L_{\max}\log^\gamma(T\|u\|+1)\sqrt{T} + \frac{L_{\max}}{T-1}\exp\left(\frac{(T-1)^{1/2\gamma}}{2(4k)^{1/\gamma}}\right) \\ &\ge k\|u\|L_{\max}\log^\gamma(T\|u\|+1)\sqrt{T} + \max_{t\le T} L_{\max}\frac{L_{t-1}^2}{\sum_{t'=1}^{t-1}\|g_{t'}\|^2}\exp\left[\frac{1}{2}\left(\frac{L_t/L_{t-1}}{288\gamma k^2}\right)^{1/(2\gamma-1)}\right]\end{aligned}$$

■

Appendix B. FTRL regret

We prove a general bound on the regret of FTRL. Our bound is not fundamentally tighter than the many previous analyses of FTRL, but we decompose the regret in a new way that makes our analysis much easier. We make use of "shadow regularizers", ψ_t^+ that can be used to characterize regret more easily. Our bound bears some similarity in form to the adaptive online mirror descent bound of (Orabona et al., 2014) and the analysis of FTRL with varying regularizers of (Cutkosky and Boahen, 2016).

Theorem 13 *Let $\ell_t, \dots, \ell_T$ be an arbitrary sequence of loss functions. Define $\ell_0(w) = 0$ for notational convenience. Let $\psi_0, \psi_1, \dots, \psi_{T-1}$ be a sequence of regularizer functions, such that ψ_t is chosen without knowledge of $\ell_{t+1}, \dots, \ell_T$. Let $\psi_1^+, \dots, \psi_T^+$ be an arbitrary sequences of regularizer functions (possibly chosen* with *knowledge of the full loss sequence). Define $w_1, \dots, w_T$ to be the outputs of FTRL with regularizers ψ_t: $w_{t+1} = \text{argmin}\, \psi_t + \ell_{1:t}$, and define w_t^+ for $t = 2, \dots, T+1$ by $w_{t+1}^+ = \text{argmin}\, \psi_t^+ + \ell_{1:t}$ Then FTRL with regularizers ψ_t obtains regret*

$$\begin{aligned} R_T(u) &= \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u) \\ &\leq \psi_T^+(u) - \psi_0(w_2^+) + \sum_{t=1}^{T} \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + \ell_t(w_t) - \ell_t(w_{t+1}^+) \\ &\quad + \sum_{t=1}^{T-1} \psi_t^+(w_{t+2}^+) - \psi_t(w_{t+2}^+) \end{aligned}$$

Proof

We define $X_t = w_{t+2}^+$ for $t < T$ and $X_T = u$. We'll use the symbols X_t as intermediate variables in our proof in an attempt to keep the algebra cleaner. By definition of w_{t+1}^+, for all $t \leq T$ we have

$$\begin{aligned} \psi_t^+(w_{t+1}^+) + \ell_{1:t}(w_{t+1}^+) &\leq \psi_t^+(X_t) + \ell_{1:t}(X_t) \\ \ell_t(w_t) &\leq \ell_t(w_t) + \ell_{1:t}(X_t) - \ell_{1:t}(w_{t+1}^+) + \psi_t^+(X_t) - \psi_t^+(w_{t+1}^+) \\ &= \ell_t(w_t) - \ell_t(w_{t+1}^+) + \ell_{1:t}(X_t) - \ell_{1:t-1}(w_{t+1}^+) \\ &\quad + \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) \\ &\quad + \psi_t^+(X_t) - \psi_{t-1}(w_{t+1}^+) \end{aligned}$$

Summing this inequality across all t we have

$$\begin{aligned} \sum_{t=1}^{T} \ell_t(w_t) &\leq \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(w_{t+1}^+) \\ &\quad + \sum_{t=1}^{T} \ell_{1:t}(X_t) - \sum_{t=1}^{T} \ell_{1:t-1}(w_{t+1}^+) \\ &\quad + \sum_{t=1}^{T} \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) \\ &\quad + \sum_{t=1}^{T} \psi_t^+(X_t) - \psi_{t-1}(w_{t+1}^+) \end{aligned}$$

Notice that $\sum_{t=1}^{T} \ell_{1:t-1}(w_{t+1}^+) = \sum_{t=2}^{T} \ell_{1:t-1}(w_{t+1}^+)$ since the first term is zero. Thus after some re-indexing we have

$$\begin{aligned}\sum_{t=1}^{T} \ell_t(w_t) \leq \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(w_{t+1}^+) \\ + \ell_{1:T}(X_T) + \sum_{t=2}^{T} \ell_{1:t-1}(X_{t-1}) - \sum_{t=2}^{T} \ell_{1:t-1}(w_{t+1}^+) \\ + \sum_{t=1}^{T} \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) \\ + \psi_T^+(X_T) - \psi_0(w_2^+) + \sum_{t=1}^{T-1} \psi_t^+(X_t) - \sum_{t=1}^{T-1} \psi_t(w_{t+2}^+)\end{aligned}$$

Now we substitute our values of $X_t = w_{t+2}^+$ for $t < T$ and $X_T = u$ to obtain

$$\begin{aligned}\sum_{t=1}^{T} \ell_t(w_t) \leq \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(w_{t+1}^+) \\ + \ell_{1:T}(u) + \psi_T^+(u) - \psi_0(w_2^+) \\ + \sum_{t=1}^{T} \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) \\ + \sum_{t=1}^{T-1} \psi_t^+(w_{t+2}^+) - \sum_{t=1}^{T-1} \psi_t(w_{t+2}^+)\end{aligned}$$

so that subtracting $\ell_{1:T}(u)$ from both sides we get a regret bound:

$$\begin{aligned}R_T(u) &= \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u) \\ &\leq \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(w_{t+1}^+) \\ &\quad + \psi_T^+(u) - \psi_0(w_2^+) \\ &\quad + \sum_{t=1}^{T} \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) \\ &\quad + \sum_{t=1}^{T-1} \psi_t^+(w_{t+2}^+) - \sum_{t=1}^{T-1} \psi_t(w_{t+2}^+)\end{aligned}$$

■

Appendix C. Facts About Strong Convexity

In this section we prove some basic facts about our generalized strong convexity.

Proposition 14 *Suppose $\psi : W \to \mathbb{R}$ is σ-strongly convex. Then:*

1. *$\psi + f$ is σ-strongly convex for any convex function f.*
2. *$c\psi$ is $c\sigma$-strongly convex for any $c \geq 0$.*
3. *Suppose $c \geq 0$ and $\phi(w) = \psi(cw)$. Let $\sigma'(x, y) = \sigma(cx, cy)$. Then ϕ is $c^2\sigma'$-strongly convex.*

Proof

1. Let $x, y \in W$ and let $g \in \partial\psi(x)$ and $b \in \partial f(x)$. Then $g + b \in \partial(\psi + f)(x)$. By convexity and strongly convexity respectively we have:

$$\psi(y) \geq \psi(x) + g \cdot (y - x) + \frac{\sigma(x, y)}{2}\|x - y\|^2$$
$$f(y) \geq f(x) + b \cdot (y - x)$$

   so that adding these equations shows that $\psi + f$ is σ-strongly convex.

2. This follows immediately by multiplying the defining equation for strong convexity of ψ by c.

3. Let $x, y \in W$ and let $g \in \partial\psi(cx)$. Then $cg \in \partial\phi(x)$.

$$\psi(cy) \geq \psi(cx) + g \cdot (cy - cx) + \frac{\sigma(cx, cy)}{2}\|cx - cy\|^2$$
$$\phi(y) \geq \phi(x) + cg \cdot (y - x) + \frac{\sigma(cx, cy)}{2}c^2\|x - y\|^2$$

■

Note that for any linear function $f(w) = g \cdot w$, if ψ is σ-strongly convex, then $\psi + f$ is also σ-strongly convex.

We show that the following lemma from (McMahan, 2014) about strongly-convex functions continues to hold under our more general definition. The proof of this lemma (and the next) are identical to the standard ones, but we include them here for completeness.

Lemma 15 *Suppose A and B are arbitrary convex functions such that $A + B$ is σ-strongly convex. Let $w_1 = \text{argmin } A$ and $w_2 = \text{argmin } A + B$ and let $g \in \partial B(w_1)$. Then*

$$\|w_1 - w_2\| \leq \frac{\|g\|_\star}{\sigma(w_1, w_2)}$$

Proof Since $w_2 \in \text{argmin } A + B$, we have $0 \in \partial(A + B)(w_2)$ and so by definition of strong convexity we have

$$\frac{\sigma(w_1, w_2)}{2}\|w_1 - w_2\|^2 \leq A(w_1) + B(w_1) - A(w_2) - B(w_2)$$

Now let $g \in \partial B(w_1)$. Consider the function $\hat{A}(w) = A(w) + B(w) - \langle g, w\rangle$. Then we must have $0 \in \partial\hat{A}(w_1)$ and so by strong-convexity again we have

$$\frac{\sigma(w_1, w_2)}{2}\|w_1 - w_2\|^2 \leq A(w_2) + B(w_2) - \langle g, w_2\rangle - A(w_1) - B(w_1) + \langle g, w_1\rangle$$

Adding these two equations yields:

$$\sigma(w_1, w_2)\|w_1 - w_2\|^2 \leq \langle g, w_1 - w_2\rangle \leq \|g\|_\star\|w_1 - w_2\|$$

and so we obtain the desired statement. ■

Finally, we have an analog of a standard way to check for strong-convexity:

Proposition 16 *Suppose $\psi : W \to \mathbb{R}$ is twice-differentiable and $v^T \nabla^2 \psi(x) v \geq \sigma(x)\|v\|^2$ for all x and v for some norm $\|\cdot\|$ and $\sigma : W \to \mathbb{R}$ where $\sigma(x + t(y-x)) \geq \min(\sigma(x), \sigma(y))$ for all $x, y \in W$ and $t \in [0,1]$. Then ψ is σ-strongly convex with respect to the norm $\|\cdot\|$.*

Proof We integrate the derivative:

$$\begin{aligned}
\psi(x) - \psi(y) &= \int_0^1 \frac{d}{dt}\psi(x + t(y-x))dt \\
&= \int_0^1 \nabla\psi(x + t(y-x)) \cdot (y-x)dt \\
&= \nabla\psi(x) \cdot (y-x) \\
&\qquad + \int_0^1 \int_0^t (y-x)^T \nabla^2 \psi(x + k(y-x))(y-x)dkdt \\
&\geq \nabla\psi(x) \cdot (y-x) + \|y-x\|^2 \int_0^1 \int_0^t \sigma(x + k(y-x))dkdt \\
&\geq \nabla\psi(x) \cdot (y-x) + \|y-x\|^2 \int_0^1 t \min(\sigma(x), \sigma(y))dt \\
&= \nabla\psi(x) \cdot (y-x) + \frac{\min(\sigma(x), \sigma(y))}{2}\|y-x\|^2
\end{aligned}$$

■

Appendix D. Proof of Theorem 8

First we prove a proposition that allows us to generate a strongly convex function easily:

Proposition 17 *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is such that $\frac{\phi'(x)}{x} \geq \phi''(x) \geq 0$ and $\phi'''(x) \leq 0$ for all $x \geq 0$. Let W be a Hilbert Space and $\psi : W \to \mathbb{R}$ be given by $\psi(w) = \phi(\|w\|)$. Then ψ is $\phi''(\|w\|)$-strongly convex with respect to $\|\cdot\|$.*

Proof Let $x, y \in W$. We have

$$\begin{aligned}
\nabla\psi(x) &= \phi'(\|x\|)\frac{x}{\|x\|} \\
\nabla^2\psi(x) &= \left(\phi''(\|x\|) - \frac{\phi'(\|x\|)}{\|x\|}\right)\frac{xx^T}{\|x\|^2} + \frac{\phi'(\|x\|)}{\|x\|}I \\
&\succeq \phi''(\|x\|)I
\end{aligned}$$

Where the last line follows since $\frac{\phi'(x)}{x} \geq \phi''(x)$ for all $x \geq 0$. Since $\phi'''(x) \leq 0$, $\phi''(x)$ is always decreasing for positive x and so we have

$$\phi''(\|x + t(y-x)\|) \geq \min(\phi''(\|x\|), \phi''(\|y\|))$$

for all $t \in [0,1]$. Therefore we can apply Proposition 16 to conclude that ψ is $\phi''(\|w\|)$-strongly convex. ■

Now we prove Proposition 8, restated below:

Proposition 8 *Let $\|\cdot\|$ be the L_2 norm ($\|w\| = \|w\|_2 = \sqrt{w \cdot w}$). Let ϕ be a three-times differentiable function from the non-negative reals to the reals that satisfies*

1. $\phi(0) = 0$.
2. $\phi'(x) \geq 0$.
3. $\phi''(x) \geq 0$.
4. $\phi'''(x) \leq 0$.
5. $\lim_{x\to\infty} \phi(x)\phi''(x) = \infty$.

Then $\psi(w) = \phi(\|w\|)$ is a $(\phi''(\|\cdot\|), \|\cdot\|)$-adaptive regularizer.

Proof

It's clear that $\psi(0) = 0$ so the first condition for being an adaptive regularizer is satisfied.

Next we will show that $\frac{\phi'(x)}{x} \geq \phi''(x)$ so that we can apply Proposition 17. It suffices to show

$$\phi'(x) - x\phi''(x) \geq 0$$

Clearly this identity holds for $x = 0$. Differentiating the right-hand-side of the equation, we have

$$\phi''(x) - x\phi'''(x) - \phi''(x) = -x\phi'''(x) \geq 0$$

since $\phi'''(x) \leq 0$ and $x \geq 0$. Thus $\phi'(x) - x\phi''(x)$ is non-decreasing and so must always be non-negative.

Therefore, by Proposition 17, ψ is $(\phi''(\|\cdot\|), \|\cdot\|)$-strongly convex. Also, since $\phi'''(x) \leq 0$, $\phi''(\|x\|) \leq \phi''(\|y\|)$ when $\|x\| \geq \|y\|$ so that ψ satisfies the second condition for being an adaptive regularizer.

Finally, observe that $\lim_{x\to\infty} \phi(x)\phi''(x)$ implies by definition that for any C there exists a B such that $\phi(x)\phi''(x) \geq C$ whenever $x \geq B$. Therefore we immediately see that $\psi(x)\phi''(\|x\|) \geq C$ for all $\|x\| \geq B$ so that the third condition is satified. ■

Appendix E. Proof of Theorem 6

First we define new regularizers ψ_t^+ analagously to ψ_t that we will use in conjunction with Theorem 13:

Definition 18 *Given a norm $\|\cdot\|$ and a sequence of subgradients $g_1, \ldots, g_T$, define L_t and $\frac{1}{\eta_t}$ as in Definition 5, and define $L_0 = L_1$. We define $\frac{1}{\eta_t^+}$ recursively by:*

$$\frac{1}{\eta_0^+} = \frac{1}{\eta_0}$$
$$\frac{1}{(\eta_t^+)^2} = \max\left(\frac{1}{\eta_{t-1}^2} + 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1}), L_{t-1}\|g_{1:t}\|_\star\right)$$

Further, given a $k \geq 1$ and a non-decreasing sequence of positive numbers a_t, define ψ_t^+ by:

$$\psi_t^+(w) = \frac{k}{\eta_t^+ a_{t-1}}\psi(a_{t-1}w)$$
$$w_{t+1}^+ = \underset{w\in W}{\textit{argmin}}\, \psi_t^+(w) + g_{1:t} \cdot w$$

Throughout the following arguments we will assume η_t and η_t^+ are the sequences defined in Definitions 5 and 18.

The next proposition establishes several identities that we will need in proving our bounds.

Proposition 19 *Suppose ψ is a $(\sigma, \|\cdot\|)$-adaptive regularizer, and $g_1, \cdots, g_T$ be some sequence of subgradients. Then the following identities hold:*

1.

$$2\|g_t\|_\star L_{t-1}\eta_t^+ \geq \left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right) \geq \|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+$$

2.

$$\frac{1}{\eta_t} \leq \sqrt{2L_t(\|g\|_\star)_{1:t}}$$
$$\frac{1}{\eta_t} \leq \sqrt{2(\|g\|_\star^2)_{1:t} + L_{\max}\max_{t'\leq t}\|g_{1:t'}\|_\star}$$

3.

$$\|w_t - w_{t+1}^+\| \leq \frac{\|g_t\|_\star\eta_t^+ + \left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{1}{L_{t-1}}}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}$$

4. Let $\hat{\psi}$ be such that $\hat{\psi}(a_{t-1}w) = \psi(a_{t-1}w)$ for $w \in W$ and $\hat{\psi}(a_{t-1}w) = \infty$ for $w \notin W$. There exists some subgradient of $\hat{\psi}$ at $a_{t-1}w_t$, which with mild abuse of notation we call $\nabla\psi(a_{t-1}w_t)$, such that:

$$|\nabla\hat{\psi}(a_{t-1}w_t)\cdot(w_t - w_{t+1}^+)| \leq 3\frac{\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}$$

5.

$$g_t\cdot(w_t - w_{t+1}^+) \leq \frac{\|g_t\|_\star^2\eta_t^+ + \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t^+}\right)\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}$$

6.

$$\frac{1}{\eta_t^+} \leq \sqrt{2L_{\max}(\|g\|_\star)_{1:T-1} + 2L_{\max}L_{t-1}}$$

Proof Let $\hat{\psi}$ be such that $\hat{\psi}(a_{t-1}w) = \psi(a_{t-1}w)$ for $w \in W$ and $\hat{\psi}(a_{t-1}w) = \infty$ for $w \notin W$. Then we can write $w_t = \text{argmin}_{w\in W}\frac{k}{a_{t-1}\eta_{t-1}}\psi(a_{t-1}w) + g_{1:t-1}\cdot w = \text{argmin}\frac{k}{a_{t-1}\eta_{t-1}}\hat{\psi}(a_{t-1}w) + g_{1:t-1}$. From this it follws that there is some subgradient of $\hat{\psi}$ at $a_{t-1}w_t$, which we refer to (by mild abuse of notation) as $\nabla\hat{\psi}(a_{t-1}w_t)$ such that

$$\nabla\hat{\psi}(a_{t-1}w_t) = -\frac{\eta_{t-1}g_{1:t-1}}{k}$$

Note that we must appeal to a subgradient rather than the actual gradient in order to encompass the case that $a_{t-1}w_t$ is on the boundary of W.

Next, observe that

$$\eta_t^+\eta_{t-1}\|g_{1:t-1}\|_\star \leq (\eta_{t-1})^2\|g_{1:t-1}\|_\star \leq \frac{1}{L_{t-1}}$$

Now we are ready to prove the various parts of the Proposition.

1. By definition of η_{t-1} and η_t^+ we have

$$\begin{aligned}
\frac{1}{(\eta_t^+)^2} - \frac{1}{(\eta_{t-1})^2} &\geq 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1}) \\
\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\left(\frac{1}{\eta_t^+} + \frac{1}{\eta_{t-1}}\right) &\geq 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1}) \\
\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\left(1 + \frac{\eta_t^+}{\eta_{t-1}}\right) &\geq 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+ \\
\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}} &\geq \|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+
\end{aligned}$$

   where in the last line we used the fact that $\eta_t^+ \leq \eta_{t-1}$ to conclude that $1 + \frac{\eta_t^+}{\eta_{t-1}} \leq 2$.
   For the other direction, we have two cases:

   1. $\frac{1}{(\eta_t^+)^2} = \frac{1}{(\eta_{t-1})^2} + 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})$.
   2. $\frac{1}{(\eta_t^+)^2} = L_{t-1}\|g_{1:t}\|_\star$.

   Case 1 $\frac{1}{(\eta_t^+)^2} = \frac{1}{(\eta_{t-1})^2} + 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})$**:**
   In this case we have

$$\begin{aligned}
\frac{1}{(\eta_t^+)^2} - \frac{1}{(\eta_{t-1})^2} &= 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1}) \\
\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\left(\frac{1}{\eta_t^+} + \frac{1}{\eta_{t-1}}\right) &= 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1}) \\
\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\left(1 + \frac{\eta_t^+}{\eta_{t-1}}\right) &= 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+ \\
\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}} &\leq 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+
\end{aligned}$$

   where in the last line we used the fact that $1 + \frac{\eta_t^+}{\eta_{t-1}} \geq 1$.
   Case 2 $\frac{1}{(\eta_t^+)^2} = L_{t-1}\|g_{1:t}\|_\star$**:**

$$\begin{aligned}
\frac{1}{(\eta_t^+)^2} - \frac{1}{(\eta_{t-1})^2} &\leq L_{t-1}\|g_{1:t}\|_\star - L_{t-1}\|g_{1:t-1}\|_\star \\
&\leq L_{t-1}\|g_t\|_\star \leq L_{t-1}\|g_t\|_\star
\end{aligned}$$

   Now we follow the exact same argument as in Case 1 to show $\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}} \leq L_{t-1}\|g_t\|_\star \eta_t^+$, which proves the desired result.

2. We proceed by induction for both claims. The statements are clear for $\frac{1}{\eta_1} = \sqrt{2}\|g_1\|_\star$. Suppose

$$\begin{aligned}
\frac{1}{\eta_t} &\leq \sqrt{2L_t(\|g\|_\star)_{1:t}} \\
\frac{1}{\eta_t} &\leq \sqrt{2(\|g\|_\star^2)_{1:t} + L_{\max}\max_{t' \leq t}\|g_{1:t'}\|_\star}
\end{aligned}$$

Then observe that $\frac{1}{\eta_t^2}+2\|g_{t+1}\|_\star^2 \leq 2L_{t+1}(\|g\|_\star)_{1:t+1}$ by the induction hypothesis, and $L_{t+1}\|g_{1:t+1}\|_\star \leq 2L_{t+1}(\|g\|_\star)_{1:t+1}$. Therefore $\frac{1}{\eta_{t+1}} \leq \sqrt{2L_{t+1}(\|g\|_\star)_{1:t+1}}$, proving the first claim.

The induction step for the second claim follows from the observations:

$$2(\|g\|_\star^2)_{1:t+1} + L_{\max}\max_{t'\leq t+1}\|g_{1:t'}\|_\star \geq 2(\|g\|_\star^2)_{1:t} + L_{\max}\max_{t'\leq t}\|g_{1:t'}\|_\star + 2\|g_{t+1}\|_\star^2$$
$$2(\|g\|_\star^2)_{1:t+1} + L_{\max}\max_{t'\leq t+1}\|g_{1:t'}\|_\star \geq L_t\|g_{1:t+1}\|_\star$$

so that $\frac{1}{\eta_{t+1}} \leq \sqrt{2(\|g\|_\star^2)_{1:t+1} + L_{\max}\max_{t'\leq t+1}\|g_{1:t'}\|_\star}$ as desired.

3. Let $I_{a_{t-1}W}(w)$ be the indicator of the set $a_{t-1}W$ - $I_{a_{t-1}W}(a_{t-1}w) = 0$ if $w \in W$ and ∞ otherwise. Observe that $\hat{\psi}(w) = \psi(w) + I_{a_{t-1}W}(w)$. Observe that $\hat{\psi}(w) = I_{a_{t-1}W}(w) + \psi(w)$.

   Now the third equation follows from Lemma 15, setting $A(w) = I_{a_{t-1}W}(w) + \frac{k}{a_{t-1}\eta_{t-1}}\psi(w) + \frac{g_{1:t-1}}{a_{t-1}}\cdot w$ and $B(w) = I_{a_{t-1}W}(w) + \frac{g_t}{a_{t-1}}\cdot w + \left(\frac{1}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\psi(w)$. Then by inspection of the definitions of w_t and w_{t+1}^+, we have $a_{t-1}w_t = \operatorname{argmin} A$ and $a_{t-1}w_{t+1}^+ = \operatorname{argmin} A + B$. Further, by Corollary 14, $A + B$ is $\frac{k\sigma}{a_{t-1}\eta_t^+}$-strongly convex. We can re-write A and B in terms of $\hat{\psi}$ by simply replacing the ψs with $\hat{\psi}$s and removing the $I_{a_{t-1}W}$s. Now we use the facts noted at the beginning of the proof:

$$\nabla\hat{\psi}(a_{t-1}w_t) = -\frac{\eta_{t-1}g_{1:t-1}}{k}$$
$$\eta_t^+\eta_{t-1}\|g_{1:t-1}\| \leq \frac{1}{L_{t-1}}$$

Applying these identities with Lemma 15 we have:

$$\begin{aligned}
\|a_{t-1}w_t - a_{t-1}w_{t+1}^+\| &\leq a_{t-1}\eta_t^+\frac{\|\frac{g_t}{a_{t-1}} + \left(\frac{k}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\nabla\hat{\psi}(a_{t-1}w_t)\|_\star}{k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\leq \frac{\eta_t^+\|g_t\|_\star}{\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} + \frac{\eta_t^+\left(\frac{k}{\eta_t^+} - \frac{k}{\eta_{t-1}}\right)\frac{\eta_{t-1}\|g_{1:t-1}\|_\star}{k}}{k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\leq \frac{\eta_t^+\|g_t\|_\star}{k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} + \frac{\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{1}{L_{t-1}}}{k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}
\end{aligned}$$

And we divide by a_{t-1} to conclude the desired identity.

4. Using the already-proved parts 1 and 3 of this Proposition and definition of dual norm, we have

$$\begin{aligned}
|\nabla\hat{\psi}(a_{t-1}w_t)\cdot(w_t - w_{t+1}^+)| &\leq \|\nabla\psi(a_{t-1}w_t)\|_\star \|w_t - w_{t+1}^+\| \\
&\leq \frac{\eta_{t-1}\|g_{1:t-1}\|_\star}{k}\frac{\eta_t^+\|g_t\|_\star}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\quad + \frac{\eta_{t-1}\|g_{1:t-1}\|_\star}{k}\frac{\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{1}{L_{t-1}}}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\leq \frac{\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} + \frac{\eta_t^+\eta_{t-1}\|g_{1:t-1}\|_\star 2L_{t-1}\|g_t\|\frac{1}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\leq \frac{\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} + \frac{\frac{1}{L_{t-1}^2}2L_{t-1}\|g_t\|_\star}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\leq 3\frac{\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}
\end{aligned}$$

5. The fifth part of the Proposition follows directly from part 3 by the definition of dual norm.

6. By part 2, we have

$$\frac{1}{\eta_{t-1}} \leq \sqrt{2L_{\max}(\|g\|_\star)_{1:t-1}}$$

We consider the two cases:

Case 1 $\frac{1}{(\eta_t^+)^2} = \frac{1}{(\eta_{t-1})^2} + 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})$**:** In this case we have

$$\begin{aligned}
\frac{1}{(\eta_t^+)^2} &\leq 2L_{\max}(\|g\|_\star)_{1:t-1} + 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1}) \\
&\leq 2L_{\max}(\|g\|_\star)_{1:t-1} + +2L_{\max}L_{t-1}
\end{aligned}$$

Case 2 $\frac{1}{(\eta_t^+)^2} = L_{t-1}\|g_{1:t}\|_\star$**:**

$$\begin{aligned}
\frac{1}{(\eta_t^+)^2} &\leq L_{t-1}\|g_{1:t}\|_\star \\
&\leq L_{t-1}\|g_{1:t-1}\| + L_{t-1}\|g_t\| \\
&\leq L_{\max}(\|g\|_\star)_{1:t-1} + L_{\max}L_{t-1}
\end{aligned}$$

■

Lemma 20 *Suppose ψ a $(\sigma, \|\cdot\|)$-adaptive regularizer and $g_1, \cdots, g_T$ is some sequence of subgradients. We use the terminology of Definition 5. Recall that we define $h(w) = \psi(w)\sigma(w)$ and $h^{-1}(x) = \max_{h(w)\leq x}\|w\|$. Suppose either of the follow holds:*

1. $\|w_{t+1}^+\| \geq \frac{h^{-1}\left(2\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ *and* $\|w_{t+1}^+\| \geq \|w_t\|$.
2. $\|w_t\| \geq \frac{h^{-1}\left(5\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ *and* $\|w_t\| \geq \|w_{t+1}^+\|$.

Then

$$\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \leq 0$$

Proof

As in Proposition 19, we use $\nabla\psi(x)$ to simply mean some particular subgradient of ψ at x.

Case 1: $\|w_{t+1}^+\| \geq \frac{h^{-1}\left(2\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ **and** $\|w_{t+1}^+\| \geq \|w_t\|$**:**

By definition of adaptive regularizer (part 2), we must have $\sigma(a_{t-1}w_{t+1}^+) \leq \sigma(a_{t-1}w_t)$ since $\|w_{t+1}^+\| \geq \|w_t\|$. Therefore $\sigma(a_{t-1}w_{t+1}^+, a_{t-1}w_t) = \sigma(a_{t-1}w_{t+1}^+)$.

By definition of h, when $\|w_{t+1}^+\| \geq \frac{h^{-1}\left(2\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ we can apply Proposition 19 (parts 1 and 5) to obtain

$$\begin{aligned}
\psi(a_{t-1}w_{t+1}^+)\sigma(a_{t-1}w_{t+1}^+) &\geq 2\frac{L_t}{k^2 L_{t-1}} \\
\left(\frac{1}{a_{t-1}\eta_t^+} - \frac{1}{a_{t-1}\eta_{t-1}}\right)\psi(a_{t-1}w_{t+1}^+) &\geq \frac{\left(\frac{1}{a_{t-1}\eta_t^+} - \frac{1}{a_{t-1}\eta_{t-1}}\right)2\frac{L_t}{L_{t-1}}}{k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
\left(\frac{k}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\psi(a_{t-1}w_{t+1}^+) &\geq \frac{\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}2\frac{L_t}{L_{t-1}} \\
\psi_t^+(w_{t+1}^+) - \psi_{t-1}(w_{t+1}^+) &\geq \frac{\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+\frac{L_t}{L_{t-1}} + \left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{L_t}{L_{t-1}}}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\geq \frac{\|g_t\|_\star^2\eta_t^+ + \left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)} \\
&\geq g_t \cdot (w_t - w_{t+1}^+)
\end{aligned}$$

We remark that in the calculations above, we showed

$$\frac{\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)2\frac{L_t}{L_{t-1}}}{a_{t-1}\sigma(a_{t-1}kw_t, a_{t-1}w_{t+1}^+)} \geq g_t(w_t - w_{t+1}^+)$$

which we will re-use in Case 2.

Case 2 $\|w_t\| \geq \frac{h^{-1}\left(5\frac{\|g_t\|_\star}{k^2 L_{t-1}}\right)}{a_{t-1}}$**, and** $\|w_t\| \geq \|w_{t+1}^+\|$**:**

Again, by definition of adaptive regularizer (part 2), we must have $\sigma(a_{t-1}w_{t+1}^+) \geq \sigma(a_{t-1}w_t)$ since $\|w_{t+1}^+\| \leq \|w_t\|$. Therefore $\sigma(a_{t-1}w_{t+1}^+, a_{t-1}w_t) = \sigma(a_{t-1}w_t)$. Let $\hat{\psi}$ be as in Proposition 19 part 4. Oberve that w_{t+1}^+ and w_t are both in W, so that we have $\psi(a_{t-1}w_{t+1}^+) = \hat{\psi}(a_{t-1}w_{t+1}^+)$ and $\psi(a_{t-1}w_t) =$

$\hat{\psi}(a_{t-1}w_t)$. Then we have:

$$\begin{aligned}
\psi_t^+(w_{t+1}^+) - \psi_{t-1}(w_{t+1}^+) &= \left(\frac{k}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\psi(a_{t-1}w_{t+1}^+) \\
&= \left(\frac{k}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\hat{\psi}(a_{t-1}w_{t+1}^+) \\
&\geq \left(\frac{k}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\left(\hat{\psi}(a_{t-1}w_t) - \left|a_{t-1}\nabla\hat{\psi}(a_{t-1}w_t)\cdot(w_{t+1}^+ - w_t)\right|\right) \\
&\geq \left(\frac{k}{\eta_t^+} - \frac{k}{\eta_{t-1}}\right)\left(\frac{\psi(a_{t-1}w_t)}{a_{t-1}} - 3\frac{\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}\right) \\
&\geq \left(\frac{k}{\eta_t^+} - \frac{k}{\eta_{t-1}}\right)\left(\frac{\psi(a_{t-1}w_t)}{a_{t-1}} - 3\frac{\frac{L_t}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}\right)]
\end{aligned}$$

Now by definition of h, when $\|w_t\| \geq \frac{h^{-1}(5\frac{L_t}{k^2L_{t-1}})}{a_{t-1}}$ we have

$$\psi(a_{t-1}w_t)\sigma(a_{t-1}w_t) \geq 5\frac{L_t}{k^2L_{t-1}}$$

$$\left(\frac{k}{a_{t-1}\eta_t^+} - \frac{k}{a_{t-1}\eta_{t-1}}\right)\psi(a_{t-1}w_t) \geq \frac{\left(\frac{k}{\eta_t^+} - \frac{k}{\eta_{t-1}}\right)5\frac{L_t}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}$$

$$\left(\frac{k}{\eta_t^+} - \frac{k}{\eta_{t-1}}\right)\left(\frac{\psi(a_{t-1}w_t)}{a_{t-1}} - 3\frac{\frac{L_t}{L_{t-1}}}{a_{t-1}k^2\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}\right) \geq \frac{\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)2\frac{L_t}{L_{t-1}}}{a_{t-1}k\sigma(a_{t-1}w_t, a_{t-1}w_{t+1}^+)}$$

$$\psi_t^+(w_{t+1}^+) - \psi_{t-1}(w_{t+1}^+) \geq g_t\cdot(w_t - w_{t+1}^+)$$

■

The next theorem is a general fact about adaptive regularizers that is useful for controlling $\psi_t^+ - \psi_t$:

Proposition 21 *Suppose $\psi : W \to \mathbb{R}$ is a $(\sigma, \|\cdot\|)$-adaptive regularizer. Then $\frac{\psi(aw)}{a}$ is an increasing function of a for all $a > 0$ for all $w \in W$.*

Proof Let's differentiate: $\frac{d}{da}\frac{\psi(aw)}{a} = \frac{\nabla\psi(aw)\cdot w}{a} - \frac{\psi(aw)}{a^2}$. Thus it suffices to show

$$\nabla\psi(aw)\cdot aw \geq \psi(aw)$$

But this follows immediately from the definition of subgradient, since $\psi(0) = 0$. ■

Lemma 22 *Suppose ψ is a $(\sigma, \|\cdot\|)$-adaptive regularizer and $g_1, \ldots, g_T$ is an arbitrary sequence of subgradients (possibly chosen adaptively). Using the terminology of Definition 5,*

$$\psi_t^+(w_{t+2}^+) - \psi_t(w_{t+1}^+) \leq 0$$

for all t

Proof

This follows from the fact that $a_{t-1} \leq a_t$, and property 4 of an adaptive regularizer ($\psi(ax)/a$ is a non-decreasing function of a). By Proposition 19 (part 1), we have $\frac{1}{\eta_t^+} \leq \frac{1}{\eta_t}$. Therefore:

$$\begin{aligned}\psi_t^+(w_{t+2}^+) &= \frac{k}{\eta_t^+ a_{t-1}}\psi(a_{t-1}w_{t+2}^+) \\ &\leq \frac{k}{\eta_t a_{t-1}}\psi(a_{t-1}w_{t+2}^+) \\ &\leq \frac{k}{\eta_t a_t}\psi(a_t w_{t+2}^+) \\ &= \psi_t(w_{t+2}^+)\end{aligned}$$

■

Lemma 23 *Suppose ψ is a $(\sigma, \|\cdot\|)$-adaptive regularizer and $g_1, \ldots, g_T$ is an arbitrary sequence of subgradients (possibly chosen adaptively). We use the regularizers of Definition 5. Recall that we define $h(w) = \psi(w)\sigma(w)$ and $h^{-1}(x) = \text{argmax}_{h(w)\leq x} \|w\|$. Define*

$$\sigma_{min} = \inf_{\|w\| \leq h^{-1}(10/k^2)} k\sigma(w)$$

and

$$D = 2\max_t \frac{h^{-1}\left(5\frac{L_t}{kL_{t-1}}\right)}{a_{t-1}}$$

Then

$$\begin{aligned}&\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \\ &\leq \begin{cases} \|g_t\|_\star \min(D, \max_t(\|w_t - w_{t+1}^+\|)) & \text{when } \|g_t\| > 2L_{t-1} \\ \frac{3\|g_t\|_\star^2 \eta_t^+}{a_{t-1}\sigma_{min}} & \text{otherwise} \end{cases}\end{aligned}$$

Proof

By Lemma 20, whenever either $\|w_{t+1}^+\| \geq \frac{h^{-1}\left(5\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}} \geq \frac{h^{-1}\left(2\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ or $\|w_t\| \geq \frac{h^{-1}\left(5\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ we must have

$$\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \leq 0$$

Therefore, we have:

$$\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \leq \begin{cases} g_t \cdot (w_t - w_{t+1}^+) & \text{when } \max(\|w_t\|, \|w_{t+1}^+\|) \leq \frac{h^{-1}\left(5\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}} \\ 0 & \text{otherwise} \end{cases}$$

When $\|g_t\|_\star \leq 2L_{t-1}$, then we have $h^{-1}\left(5\frac{L_t}{k^2 L_{t-1}}\right) \leq h^{-1}(10/k^2)$. Thus when $\max(\|w_t\|, \|w_{t+1}^+\|) \leq \frac{h^{-1}\left(5\frac{L_t}{k^2 L_{t-1}}\right)}{a_{t-1}}$ and $\|g_t\|_\star \leq 2L_{t-1}$, by Proposition 19 (part 5), we have

$$g_t(w_t - w_{t+1}^+) \leq \frac{\|g_t\|_\star^2 \eta_t^+ + \left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}\sigma_{\min}}$$

Therefore when $\|g_t\|_\star \leq 2L_{t-1}$ we have (using Proposition 19 part 1):

$$\begin{aligned}\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) &\leq \frac{\|g_t\|_\star^2\eta_t^+ + \left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right)\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}\sigma_{\min}} \\ &\leq \frac{\|g_t\|_\star^2\eta_t^+ + 2\|g_t\|_\star L_{t-1}\eta_t^+\frac{\|g_t\|_\star}{L_{t-1}}}{a_{t-1}\sigma_{\min}} \\ &\leq \frac{3\|g_t\|_\star^2\eta_t^+}{a_{t-1}\sigma_{\min}}\end{aligned}$$

so that we can improve our conditional bound to:

$$\begin{aligned}&\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \\ &\leq \begin{cases} g_t\cdot(w_t - w_{t+1}^+) & \text{when } \max(\|w_t\|, \|w_{t+1}^+\|) \leq \frac{h^{-1}\left(5\frac{L_t}{k^2L_{t-1}}\right)}{a_{t-1}} \text{ and } \|g_t\|_\star > 2L_{t-1} \\ \frac{3\|g_t\|^2\eta_t^+}{a_{t-1}\sigma_{\min}} & \text{when } \max(\|w_t\|, \|w_{t+1}^+\|) \leq \frac{h^{-1}\left(5\frac{L_t}{k^2L_{t-1}}\right)}{a_{t-1}} \text{ and } \|g_t\|_\star \leq 2L_{t-1} \\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

When both $\|w_{t+1}^+\|$ and $\|w_t\|$ are less than than $\frac{h^{-1}\left(5\frac{L_t}{k^2L_{t-1}}\right)}{a_{t-1}}$ then we also have

$$\|w_t - w_{t+1}^+\| \leq \min\left(D, \max_t \|w_t - w_{t+1}^+\|\right)$$

where we define

$$D = 2\max_t \frac{h^{-1}\left(5\frac{L_t}{k^2L_{t-1}}\right)}{a_{t-1}}$$

Therefore we have

$$\begin{aligned}&\psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \\ &\leq \begin{cases} g_t\cdot(w_t - w_{t+1}^+) & \text{when } \max(\|w_t\|, \|w_{t+1}^+\|) \leq \frac{h^{-1}\left(5\frac{L_t}{L_{t-1}}\right)}{a_{t-1}} \text{ and } \|g_t\|_\star > 2L_{t-1} \\ \frac{3\|g_t\|^2\eta_t^+}{a_{t-1}\sigma_{\min}} & \text{when } \max(\|w_t\|, \|w_{t+1}^+\|) \leq \frac{h^{-1}\left(5\frac{L_t}{k^2L_{t-1}}\right)}{a_{t-1}} \text{ and } \|g_t\|_\star \leq 2L_{t-1} \\ 0 & \text{otherwise}\end{cases} \\ &\leq \begin{cases} \|g_t\|_\star \min(D, \max_t \|w_t - w_{t+1}^+\|), & \text{when } \|g_t\| > 2L_{t-1} \\ \frac{3\|g_t\|_\star^2\eta_t^+}{a_{t-1}\sigma_{\min}} & \text{otherwise}\end{cases}\end{aligned}$$

■

Now we have three more technical lemmas:

Lemma 24 *Let $a_1, \ldots, a_M$ be a sequence of non-negative numbers such that $a_{i+1} \geq 2a_i$. Then*

$$\sum_{i=1}^M a_i \leq 2a_M$$

Proof We proceed by induction on M. For the base case, we observe that $a_1 \leq 2a_1$. Suppose $\sum_{i=1}^{M-1} a_i \leq 2a_{M-1}$. Then we have

$$\begin{aligned}\sum_{i=1}^{M} a_i &= a_M + \sum_{i=1}^{M-1} a_i \\ &\leq a_M + 2a_{M-1} \\ &\leq a_M + a_M = 2a_M\end{aligned}$$

■

The next lemma establishes some identities analogous to the bounds $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} = O(\sqrt{T})$, and $\sum_{t=1}^{T} \frac{1}{T^2} = O(1)$. These are useful for dealing with increasing a_t in our regret bounds.

Lemma 25

1.

$$\sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \|g_t\|_\star^2 \eta_t^+ \leq \frac{2}{\eta_T^+}$$

2. Suppose α_t is defined by

$$\begin{aligned}\alpha_0 &= \frac{1}{(L_1\eta_1)^2} \\ \alpha_t &= \max\left(\alpha_{t-1}, \frac{1}{(L_t\eta_t)^2}\right)\end{aligned}$$

then

$$\sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \|g_t\|_\star^2 \frac{\eta_t^+}{\alpha_{t-1}} \leq 15L_{\max}$$

Proof

1. Using part 1 from Proposition 19, and observing that $\eta_t^+ \geq \eta_t$, we have

$$\begin{aligned}\sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \|g_t\|_\star^2 \eta_t^+ &\leq \sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} 2\|g_t\|_\star \min(\|g_t\|_\star, L_{t-1})\eta_t^+ \\ &\leq \sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} 2\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}}\right) \\ &\leq \sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} 2\left(\frac{1}{\eta_t^+} - \frac{1}{\eta_{t-1}^+}\right) \\ &\leq 2\eta_T^+\end{aligned}$$

2. For the second part of the lemma, we observe that for $\|g_t\|_\star \leq 2L_{t-1}$,

$$\begin{aligned}\frac{1}{(\eta_t^+)^2} &\geq \frac{1}{(\eta_{t-1})^2} + 2\|g_t\|_\star \min(L_{t-1}, \|g_t\|_\star) \\ &\geq \frac{1}{(\eta_{t-1})^2} + \|g_t\|_\star^2 \\ &\geq (\|g\|_\star^2)_{1:t}\end{aligned}$$

Similarly, we also have $(\|g\|_\star^2)_{1:t} \leq (1 + \frac{L_t^2}{L_{t-1}^2})(\|g\|_\star^2)_{1:t-1}$ so that

$$\begin{aligned}
\frac{1}{\alpha_{t-1}} &\leq L_{t-1}^2 \eta_{t-1}^2 \\
&\leq \frac{L_{t-1}^2}{2(\|g\|_\star^2)_{1:t-1}} \\
&\leq \frac{L_{t-1}}{L_t} \frac{L_t^2}{2(\|g\|_\star^2)_{1:t-1}} \\
&\leq \frac{L_{t-1}}{L_t} \left(1 + \frac{L_t^2}{L_{t-1}^2}\right) \frac{L_t^2}{2(\|g\|_\star^2)_{1:t}} \\
&= \left(\frac{L_{t-1}}{L_t} + \frac{L_t}{L_{t-1}}\right) \frac{L_t^2}{2(\|g\|_\star^2)_{1:t}} \\
&\leq \frac{5}{4} \frac{L_t^2}{(\|g\|_\star^2)_{1:t}}
\end{aligned}$$

where in the last line we have used $L_t/L_{t-1} \leq 2$.
Combining these two calculations, we have

$$\sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \|g_t\|_\star^2 \frac{\eta_t^+}{\alpha_{t-1}} \leq \frac{5}{4} \sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}}$$

Let $T_1, T_2, \ldots, T_n$ be the indices such that $\|g_{T_i}\|_\star > 2L_{T_i-1}$, and define $T_n = T + 1$. We will show that for any i with $T_{i+1} > T_i + 1$,

$$\sum_{t=T_i+1}^{T_{i+1}-1} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} \leq 6L_{T_{i+1}-1} \tag{1}$$

Observe that for $N = T_i + 1$, we have

$$\sum_{t=T_i+1}^{N} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} \leq 6L_N - \frac{2L_N^2}{\sqrt{(\|g\|_\star^2)_{1:N}}} \tag{2}$$

We'll prove by induction that equation (2) holds for all $N \leq T_{i+1} - 1$. Suppose it holds for some $N < T_{i+1} - 1$. Then by concavity of $-\frac{1}{\sqrt{x}}$, we have

$$\left(6L_{N+1} - \frac{2L_{N+1}^2}{\sqrt{(\|g\|_\star^2)_{1:N+1}}}\right) - \left(6L_{N+1} - \frac{2L_{N+1}^2}{\sqrt{(\|g\|_\star^2)_{1:N}}}\right) \geq \frac{\|g_{N+1}\|_\star^2 L_{N+1}^2}{(\|g\|_\star^2)_{1:N+1}^{3/2}}$$

So using the inductive hypothesis:

$$\begin{aligned}
\sum_{t=1}^{N+1} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} &\leq \left(6L_N - \frac{2L_N^2}{\sqrt{(\|g\|_\star^2)_{1:N}}}\right) + \frac{\|g_{N+1}\|_\star^2 L_{N+1}^2}{(\|g\|_\star^2)_{1:N+1}^{3/2}} \\
&= \left(6L_{N+1} - \frac{2L_{N+1}^2}{\sqrt{(\|g\|_\star^2)_{1:N}}}\right) + \frac{\|g_{N+1}\|_\star^2 L_{N+1}^2}{(\|g\|_\star^2)_{1:N+1}^{3/2}} + 6(L_N - L_{N+1}) - \frac{2(L_N^2 - L_{N+1}^2)}{\sqrt{(\|g\|_\star^2)_{1:N}}} \\
&\leq 6L_{N+1} - \frac{2L_{N+1}2}{\sqrt{(\|g\|_\star^2)_{1:N+1}}} + 6(L_N - L_{N+1}) - \frac{2(L_N^2 - L_{N+1}^2)}{\sqrt{(\|g\|_\star^2)_{1:N}}}
\end{aligned}$$

To finish the induction, we show that $6(L_N - L_{N+1}) - \frac{2(L_N^2 - L_{N+1}^2)}{\sqrt{(\|g\|_\star^2)_{1:N}}} \leq 0$. We factor out the non-negative quantity $L_{N+1} - L_N$, and then observe that $L_{N+1} \leq 2L_N$ since $T_i + 1 \leq N < N + 1 \leq T_{i+1} - 1$ (and in particular, $L_{N+1} \neq T_i$ for any i).

$$\begin{aligned} -6 + \frac{2(L_N + L_{N+1})}{\sqrt{(\|g\|_\star^2)_{1:N}}} &\leq -6 + \frac{6L_N}{\sqrt{(\|g\|_\star^2)_{1:N}}} \\ &\leq 0 \end{aligned}$$

Therefore equation (2) holds for all $N \leq T_{i+1} - 1$, so that we have

$$\sum_{t=T_i+1}^{T_{i+1}-1} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} \leq 6L_{T_{i+1}-1} - \frac{2L_{T_{i+1}-1}^2}{\sqrt{(\|g\|_\star^2)_{1:T}}} \leq 6L_{T_{i+1}-1} \tag{3}$$

so that equation (1) holds. Now we write (using the convention that $\sum_{t=x}^{z} y_t = 0$ if $z < x$):

$$\begin{aligned} \sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} &= \sum_{i=1}^{n+1} \sum_{t=T_i+1}^{T_{i+1}-1} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} \\ &\leq \sum_{i=1}^{n+1} 6L_{T_{i+1}-1} \\ &\leq 12L_{\max} \end{aligned}$$

where in the last step we have observed that by definition of T_i, $L_{T_{i+1}-1} \geq 2L_{T_i-1}$ for all i and used Lemma 24.

Finally, we conclude

$$\begin{aligned} \sum_{t|\ \|g_t\|_\star \leq 2L_{t-1}} \|g_t\|_\star^2 \frac{\eta_t^+}{a_t} &\leq \frac{5}{4} \sum_{t=1}^{T} \frac{\|g_t\|_\star^2 L_t^2}{(\|g\|_\star^2)_{1:t}^{3/2}} \\ &\leq 15L_{\max} \end{aligned}$$

■

Lemma 26 *Let α_t be defined by*

$$\begin{aligned} \alpha_0 &= \frac{1}{(L_1\eta_1)^2} \\ \alpha_t &= \max\left(\alpha_{t-1}, \frac{1}{(L_t\eta_t)^2}\right) \end{aligned}$$

Then

$$\frac{2(\|g\|_\star)_{1:t}}{L_t} \geq a_t \geq \frac{2(\|g\|_\star^2)_{1:t}}{L_t^2}$$

Proof Since $\frac{1}{\eta_t^2} \geq 2(\|g\|_\star^2)_{1:t}$, we immediately recover the lower bound on a_t. The upper bound follows from Proposition 19 (part 2), which states $\frac{1}{\eta_t^2} \leq 2L_t(\|g\|_\star)_{1:t}$ ■

Now we're ready to prove Theorem 6, which we restate for reference:

Theorem 6 *Suppose ψ is a $(\sigma, \|\cdot\|)$-adaptive regularizer and $g_1, \dots, g_T$ is some arbitrary sequence of subgradients. Let $k \geq 1$, and let ψ_t be defined as in Definition 5.*

Set

$$\begin{aligned}\sigma_{min} &= \inf_{\|w\| \leq h^{-1}(10/k^2)} k\sigma(w) \\ D &= \max_t \frac{L_{t-1}^2}{(\|g\|_\star^2)_{1:t-1}} h^{-1}\left(\frac{5L_t}{k^2 L_{t-1}}\right) \\ Q_T &= 2\frac{\|g\|_{1:T}}{L_{\max}}\end{aligned}$$

Then FTRL with regularizers ψ_t achieves regret

$$\begin{aligned}R_T(u) &\leq \frac{k}{Q_T \eta_T}\psi(Q_T u) + \frac{45L_{\max}}{\sigma_{min}} + 2L_{\max}D \\ &\leq kL_{\max}\frac{\psi(2uT)}{\sqrt{2T}} + \frac{45L_{\max}}{\sigma_{min}} + 2L_{\max}D\end{aligned}$$

Proof Using Theorem 13 and Lemmas 22 and 23, our regret is bounded by

$$\begin{aligned}R_T(u) &\leq \psi_T^+(u) + \sum_{t=1}^T \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \\ &\quad + \sum_{t=1}^T \psi_t^+(w_{t+2}^+) - \psi_t(w_{t+2}^+) \\ &\leq \psi_T^+(u) + \sum_{t=1}^T \psi_{t-1}(w_{t+1}^+) - \psi_t^+(w_{t+1}^+) + g_t(w_t - w_{t+1}^+) \\ &\leq \psi_T^+(u) + \sum_{\|g_t\|_\star \leq 2L_{t-1}} \frac{3\|g_t\|^2\eta_t^+}{a_{t-1}\sigma_{\min}} + \sum_{\|g_t\|_\star > 2L_{t-1}} \|g_t\|_\star D'\end{aligned}$$

where D' is defined by

$$D' = 2\max_t \frac{h^{-1}\left(5\frac{L_t}{kL_{t-1}}\right)}{a_{t-1}}$$

Now we use Lemma 26 to conclude that

$$D' \leq D = \max_t \frac{L_{t-1}^2}{(\|g\|_\star^2)_{1:t-1}} h^{-1}\left(5\frac{L_t}{kL_{t-1}}\right)$$

so that we have

$$R_T(u) \leq \psi_T^+(u) + \sum_{\|g_t\|_\star \leq 2L_{t-1}} \frac{3\|g_t\|^2\eta_t^+}{a_{t-1}\sigma_{\min}} + \sum_{\|g_t\|_\star > 2L_{t-1}} \|g_t\|_\star D$$

Now using Lemma 25 we can simplify this to

$$R_T(u) \leq \frac{k}{a_T\eta_T^+}\psi(a_T u) + \frac{45L_{\max}}{\sigma_{\min}} + \sum_{\|g_t\|_\star > 2L_{t-1}} \|g_t\|_\star D$$

Finally, observe that each value of $\|g_t\|_\star$ in the sum $\sum_{\|g_t\|_\star > 2L_{t-1}} \|g_t\|_\star D$ is at least twice the previous value, so that by Lemma 24 we conclude

$$R_T(u) \leq \frac{k}{a_T \eta_T^+} \psi(a_T u) + \frac{45 L_{\max}}{\sigma_{\min}} + 2 L_{\max} D$$

Finally, we observe that (by Lemma 26), $a_T \leq 2 \frac{\|g\|_{1:T}}{L_T} = Q_T$, which gives the first inequality in the Theorem statement.

Using the fact that $\frac{1}{\eta_t} \leq \sqrt{2 L_{\max} (\|g\|_\star)_{1:t}}$ (from Proposition 19 part 2), we have $\eta_T^+ \geq \frac{1}{L_{\max}\sqrt{2T}}$ and it is clear that $a_T \leq 2T$, so that we recover the second inequality as well.

■