

Lipschitz and Comparator-Norm Adaptivity in Online Learning

Zakaria Mhammedi

The Australian National University and Data61

Wouter M. Koolen

Centrum Wiskunde & Informatica

ZAK.MHAMMEDI@ANU.EDU.AU

WMKOOLEN@CWI.NL

Abstract

We study Online Convex Optimization in the unbounded setting where neither predictions nor gradient are constrained. The goal is to simultaneously adapt to both the sequence of gradients and the comparator. We first develop parameter-free and scale-free algorithms for a simplified setting with hints. We present two versions: the first adapts to the squared norms of both comparator and gradients separately using $O(d)$ time per round, the second adapts to their squared inner products (which measure variance only in the comparator direction) in time $O(d^3)$ per round. We then generalize two prior reductions to the unbounded setting; one to not need hints, and a second to deal with the range ratio problem (which already arises in prior work). We discuss their optimality in light of prior and new lower bounds. We apply our methods to obtain sharper regret bounds for scale-invariant online prediction with linear models.

Keywords: Online Convex Optimization, Parameter-Free Online Learning, Scale-Invariant Online Algorithms

1. Introduction

We consider the setting of online convex optimization where the goal is to make sequential predictions to minimize a certain notion of *regret*. Specifically, at the beginning of each round $t \geq 1$, a *learner* predicts $\hat{\mathbf{w}}_t$ in some convex set $\mathcal{W} \subseteq \mathbb{R}^d$ in dimension $d \in \mathbb{N}$. The *environment* then reveals a convex loss function $f_t: \mathcal{W} \rightarrow \mathbb{R}$, and the learner suffers loss $f_t(\hat{\mathbf{w}}_t)$. The goal of the learner is to minimize the regret $\sum_{t=1}^T f_t(\hat{\mathbf{w}}_t) - \sum_{t=1}^T f_t(\mathbf{w})$ after $T \geq 1$ rounds against any “comparator” prediction $\mathbf{w} \in \mathcal{W}$. Typically, an online learning algorithm outputs a vector $\hat{\mathbf{w}}_t$, $t \geq 1$, based only on a sequence of observed sub-gradients $(\mathbf{g}_s)_{s < t}$, where $\mathbf{g}_s \in \partial f_s(\hat{\mathbf{w}}_s)$, $s < t$. In this paper, we are interested in online algorithms which can guarantee a good regret bound (by a measure which we will make precise below) against any comparator vector $\mathbf{w} \in \mathcal{W}$, even when \mathcal{W} is unbounded, and without prior knowledge of the maximum norm $L := \max_{t \leq T} \|\mathbf{g}_t\|$ of the observed sub-gradients. In what follows, we refer to L as the *Lipschitz constant*.

By assuming an upper-bound $D > 0$ on the norm of the desired comparator vector \mathbf{w} in hindsight, there exist *Lipschitz-adaptive* algorithms that can achieve a sub-linear regret of order $LD\sqrt{T}$, without knowing L in advance. A Lipschitz-adaptive algorithm is also called *scale-free* (or scale-invariant) if its predictions do not change when the loss functions (f_t) are multiplied by a factor $c > 0$; in this case, its regret bound is expected to scale by the same factor c . When L is known in advance and $\mathcal{W} = \mathbb{R}^d$, there exists another type of algorithms, so-called *parameter-free*, which can achieve a $\tilde{O}(\|\mathbf{w}\|L\sqrt{T})$ regret bound, where \mathbf{w} is the desired comparator vector in hindsight (the notation \tilde{O} hides log-factors). Up to an additive lower-order term, this type of regret bound is also achievable for bounded \mathcal{W} via the unconstrained-to-constrained reduction (Cutkosky, 2019).

The question of whether an algorithm can simultaneously be *scale-free* and *parameter-free* was posed as an open problem by Orabona and Pál (2016b). It was latter answered in the negative by Cutkosky and Boahen (2017). Nevertheless, Cutkosky (2019) recently presented algorithms which achieve a $\tilde{O}(\|w\|L\sqrt{T} + L\|w\|^3)$ regret bound, without knowing either L or $\|w\|$. This does not violate the earlier lower bound of Cutkosky and Boahen (2017), which insists on norm dependence $\tilde{O}(\|w\|)$.

Though Cutkosky (2019) designed algorithms that can to some extent adapt to both L and $\|w\|$, their algorithms are still *not* scale-free. Multiplying (f_t) , and as a result (g_t) , by a positive factor $c > 0$ changes the outputs (\hat{w}_t) of their algorithms, and their regret bounds scale by a factor c' , not necessarily equal to c . Their algorithms depend on a parameter $\epsilon > 0$ which has to be specified in advance. This parameter appears in their regret bounds as an additive term and also in a logarithmic term of the form $\log(L^\alpha/\epsilon)$, for some $\alpha > 1$. As a result of this type of dependence on ϵ and the fact that $\alpha > 1$, there is no prior choice of ϵ which can make their regret bounds scale-invariant. What is more, without knowing L , there is also no “safe” choice of ϵ which can prevent the $\log(L^\alpha/\epsilon)$ term from becoming arbitrarily large relative to L (it suffices for ϵ to be small enough relative to the “unknown” L).

Contributions. Our main contribution is a new scale-free, parameter-free learning algorithm for OCO with a regret at most $O(\|w\|\sqrt{V_T} \log(\|w\|T))$, for any comparator $w \in \mathcal{W}$ in a bounded set \mathcal{W} , where $V_T := \sum_{t=1}^T \|g_t\|^2$. When the set \mathcal{W} is unbounded, the algorithm achieves the same guarantee up to an additive $O(L\sqrt{\max_{t \leq T} B_t} + L\|w\|^3)$, where $B_t := \sum_{s=1}^t \|g_s\|/L_t$ and $L_t := \max_{s \leq t} \|g_s\|$, for all $t \in [T]$. In the latter case, we also show a matching lower bound; when \mathcal{W} is unbounded and without knowing L , any online learning algorithm which insists on a $\tilde{O}(\sqrt{T})$ bound, has regret at least $\Omega(L\sqrt{B_T} + L\|w\|^3)$. We also provide a second scale-invariant algorithm which replaces the leading $\|w\|\sqrt{V_T}$ term in the regret bound of our first algorithm by $\sqrt{w^\top V_T w \ln \det V_T}$, where $V_T := \sum_{t=1}^T g_t g_t^\top$. Our starting point for designing our algorithms is a known potential function which we show to be controlled for a unique choice of output sequence (\hat{w}_t) .

As our main application, we show how our algorithms can be applied to learn linear models. The result is an online algorithm for learning linear models whose label predictions are invariant to coordinate-wise scaling of the input feature vectors. The regret bound of the algorithm is naturally also scale-invariant and improves on the bounds of existing state-of-the-art algorithms in this setting (Kotłowski, 2017; Kempka et al., 2019).

Related Work For an overview of Online Convex Optimization in the bounded setting, we refer to the textbook (Hazan, 2016). The unconstrained case was first studied by McMahan and Streeter (2010). A powerful methodology for the unbounded case is Coin Betting by Orabona and Pál (2016a). Even though not always visible, our potential functions are inspired by this style of thinking. We build our unbounded OCO learner by targeting a specific other constrained problem. We further employ several general reductions from the literature, including gradient clipping Cutkosky (2019), the constrained-to-unconstrained reduction Cutkosky and Orabona (2018), and the restart wrapper to pacify the final-vs-initial scale ratio appearing inside logarithms by Mhammedi et al. (2019). Our analysis is, at its core, proving a certain minimax result about sufficient-statistic-based potentials reminiscent of the Burkholder approach pioneered by Foster et al. (2017, 2018). Applications for scale-invariant learning in linear models were studied by Kempka et al. (2019). For our multidimensional learner we took inspiration from the Gaussian Exp-concavity step in the analysis of the MetaGrad algorithm by Van Erven and Koolen (2016).

Outline In Section 2, we present the setting and notation, and formulate our goal. In Section 3, we present our main algorithms. In Section 4, we present new lower-bounds for algorithms which adapt to both the Lipschitz constant and the norm of the comparator. In Section 5, we apply our algorithms to online prediction with linear models.

2. Preliminaries

Our goal is to design scale-free algorithms that adapt to the Lipschitz constant L and comparator norm $\|\mathbf{w}\|$. We will first introduce the setting, then discuss existing reductions, and finally state what needs to be done to achieve our goal.

2.1. Setting and Notation

Let $\mathcal{W} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, be a convex set, and assume without loss of generality that $\mathbf{0} \in \mathcal{W}$. We allow the set \mathcal{W} to be unbounded, and we define its (possibly infinite) diameter $D := \sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} \|\mathbf{w} - \mathbf{w}'\| \in [0, +\infty]$. We consider the setting of Online Convex Optimization (OCO) where at the beginning of each round $t \geq 1$, the learner outputs a prediction $\hat{\mathbf{w}}_t \in \mathcal{W}$, before observing a convex loss function $f_t : \mathcal{W} \rightarrow \mathbb{R}$, or an element of its sub-gradient $\mathbf{g}_t \in \partial f_t(\hat{\mathbf{w}}_t)$ at $\hat{\mathbf{w}}_t$. The goal of the learner is to minimize the regret after $T \geq 0$ rounds

$$\sum_{t=1}^T f_t(\hat{\mathbf{w}}_t) - \sum_{t=1}^T f_t(\mathbf{w})$$

for any comparator vector $\mathbf{w} \in \mathcal{W}$. In this paper, we do not assume that T is known to the learner, and so we are after algorithms with so called *any-time* guarantees. By convexity, we have

$$\sum_{t=1}^T f_t(\hat{\mathbf{w}}_t) - \sum_{t=1}^T f_t(\mathbf{w}) \leq \sum_{t=1}^T \langle \mathbf{g}_t, \hat{\mathbf{w}}_t - \mathbf{w} \rangle, \quad \text{for all } \mathbf{w} \in \mathcal{W}, \quad (1)$$

and thus for the purpose of minimizing the regret, typical OCO algorithms minimize the RHS of (1), which is known as the *linearized regret*, by generating outputs $(\hat{\mathbf{w}}_t)$ based on the sequence of observed sub-gradients (\mathbf{g}_t) . Likewise, we focus our attention exclusively on linear optimization.

Given a sequence of sub-gradients (\mathbf{g}_t) , it will be useful to define the running maximum gradient norm and the clipped sub-gradients

$$L_t := \max_{s \in [t]} \|\mathbf{g}_s\| \quad \text{and} \quad \bar{\mathbf{g}}_t := \mathbf{g}_t \cdot L_{t-1} / L_t,$$

for $t \geq 1$, with the convention that $L_0 = 0$. We also drop the subscript t from L_t when $t = T$, i.e. we write L for L_T .

We denote by $A(\mathbf{g}_1, \dots, \mathbf{g}_{t-1}; h_t)$ the output in round $t \geq 1$ of an algorithm A , which uses the observed sub-gradients so far and a *hint* $h_t \geq L_t$ on the upcoming sub-gradient \mathbf{g}_t . As per Section 1, we say that an algorithm is *scale-free* (or scale-invariant) if its predictions are invariant to any common positive scaling of the loss functions (f_t) and, if applicable, the hints.

Additional Notation. Given a closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$, we denote by $\Pi_{\mathcal{X}}(\mathbf{x})$ the Euclidean projection of a point $\mathbf{x} \in \mathbb{R}^d$ on the set \mathcal{X} ; that is, $\Pi_{\mathcal{X}}(\mathbf{x}) \in \operatorname{argmin}_{\tilde{\mathbf{x}} \in \mathcal{X}} \|\mathbf{x} - \tilde{\mathbf{x}}\|$.

2.2. Helpful Reductions

The difficulty behind designing scale-free algorithms lies partially in the fact that L_t is not-known at the start of round t ; before outputting $\hat{\mathbf{w}}_t$. The following result due to [Cutkosky \(2019\)](#) quantifies the additional cost of proceeding with the plug-in estimate L_{t-1} for L_t :

Lemma 1 *Let A be an online algorithm which at the start of each round $t \geq 1$, has access to a hint $h_t \geq L_t$, and outputs $A(\mathbf{g}_1, \dots, \mathbf{g}_{t-1}; h_t) \in \mathcal{W}$, before observing \mathbf{g}_t . Suppose that A guarantees an upper-bound $R_T^A(\mathbf{w})$ on its linearized regret for the sequence (\mathbf{g}_t) and for all $\mathbf{w} \in \mathcal{W}, T \geq 1$. Then, algorithm B which at the start of each round $t \geq 1$ outputs $\hat{\mathbf{w}}_t = A(\bar{\mathbf{g}}_1, \dots, \bar{\mathbf{g}}_{t-1}; L_{t-1})$, guarantees*

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \leq R_T^A(\mathbf{w}) + \max_{t \in [T]} L_t \|\hat{\mathbf{w}}_t\| + \|\mathbf{w}\|L, \quad \forall \mathbf{w} \in \mathcal{W}, T \geq 1. \quad (2)$$

First, we note that Lemma 1 is only really useful when \mathcal{W} is bounded; otherwise, depending on algorithm A, the term $\max_{t \in [T]} L_t \|\hat{\mathbf{w}}_t\|$ on the RHS of (2) could in principle be arbitrarily large even for fixed \mathbf{w}, L , and T . The moral of Lemma 1 is that as long as the set \mathcal{W} is bounded, one does not really need to know L_t before outputting $\hat{\mathbf{w}}_t$ to guarantee a “good” regret bound against any $\mathbf{w} \in \mathcal{W}$. For example, suppose that \mathcal{W} has a bounded diameter D and algorithm A in Lemma 1 is such that $R_T^A(\mathbf{w}) = \tilde{O}(\|\mathbf{w}\|L\sqrt{T} + DL)$, for all $\mathbf{w} \in \mathcal{W}$. Then, from (2) and the fact that $\|\hat{\mathbf{w}}_t\| \leq D$ (since $\hat{\mathbf{w}}_t \in \mathcal{W}$), it is clear that algorithm B in Lemma 1 also guarantees the same regret bound $R_T^A(\mathbf{w})$ up to an additive $2DL$, despite not having had the hints (h_t) .

It is possible to extend the result of Lemma 1 so that the regret bound of algorithm B remains useful even in the case where \mathcal{W} is unbounded. An approach suggested by [Cutkosky \(2019\)](#) is to restrict the outputs $(\hat{\mathbf{w}}_t)$ of algorithm B to be in a non-decreasing sequence (\mathcal{W}_t) of *bounded* convex subsets of \mathcal{W} . In this case, the diameters $(D_t) \subset \mathbb{R}$ of (\mathcal{W}_t) need to be carefully chosen to achieve a desired regret bound. This approach, which essentially combines the idea of Lemma 1 and the unconstrained-to-constrained reduction due to [Cutkosky and Orabona \(2018\)](#), is formalized in the next lemma (essentially due to [Cutkosky \(2019\)](#)):

Lemma 2 *Let algorithm A be as in Lemma 1, and let (\mathcal{W}_t) be a sequence of non-decreasing closed convex subsets of \mathcal{W} with diameters $(D_t) \subset \mathbb{R}_{>0}$. Then, algorithm B which at the start of round $t \geq 1$ outputs $\hat{\mathbf{w}}_t = \Pi_{\mathcal{W}_t}(\tilde{\mathbf{w}}_t)$, where*

$$\tilde{\mathbf{w}}_t := A(\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_{t-1}; L_{t-1}) \quad \text{and} \quad \tilde{\mathbf{g}}_s := (\bar{\mathbf{g}}_s + \|\bar{\mathbf{g}}_s\| \cdot (\tilde{\mathbf{w}}_s - \hat{\mathbf{w}}_s) / \|\tilde{\mathbf{w}}_s - \hat{\mathbf{w}}_s\|) / 2, \quad s < t,$$

guarantees, for all $\mathbf{w} \in \mathcal{W}$ and $T \geq 1$,

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \leq R_T^A(\mathbf{w}) + \sum_{t=1}^T \|\mathbf{g}_t\| \cdot \|\mathbf{w} - \Pi_{\mathcal{W}_t}(\mathbf{w})\| + LD_T + L\|\mathbf{w}\|. \quad (3)$$

We see that compared to Lemma 1, the additional penalty that algorithm B incurs for restricting its predictions to the sets $\mathcal{W}_1, \dots, \mathcal{W}_T \subseteq \mathcal{W}$ is the sum $\sum_{t=1}^T \|\mathbf{g}_t\| \cdot \|\mathbf{w} - \Pi_{\mathcal{W}_t}(\mathbf{w})\|$. The challenge is now in choosing the diameters (D_t) to control the trade-off between this sum and the term LD_T on the RHS of (3). If T is known in advance, one could set $D_1 = \dots = D_T = \sqrt{T}$, in which case the RHS of (3) is at most

$$R_T^A(\mathbf{w}) + L(\|\mathbf{w}\|^3 + \|\mathbf{w}\|) + L\sqrt{T}. \quad (4)$$

We now instantiate the bound of Lemma 2 for another choice of (D_t) when T is unknown:

Corollary 3 *In the setting of Lemma 2, let \mathcal{W}_t be the ball of diameter $D_t := \sqrt{\max_{s \leq t} B_s}$, $t \geq 1$, where $B_t := \sum_{s=1}^t \|g_s\|^2/L_t$, and let $\mathcal{W} = \mathbb{R}^d$. Then the RHS of (3) is bounded from above by*

$$R_T^A(\mathbf{w}) + L\|\mathbf{w}\|^3 + L\sqrt{\max_{t \in [T]} B_t} + L\|\mathbf{w}\|, \quad \forall \mathbf{w} \in \mathcal{W} = \mathbb{R}^d, T \geq 1. \quad (5)$$

We see that by the more careful choice of (D_t) in Corollary 3, one can replace the $L\sqrt{T}$ term in (4) by the smaller quantity $L\sqrt{\max_{t \in [T]} B_t}$; whether this can be improved further to $\sqrt{V_T}$, where $V_T = \sum_{t=1}^T \|g_t\|^2$, was raised as an open question by Cutkosky (2019). We will answer this in the negative in Theorem 12. We will also show in Theorem 13 below that, if one insists on a regret of order $\tilde{O}(\sqrt{T})$, it is essentially not possible to improve on the penalty $L\|\mathbf{w}\|^3$ in (5).

2.3. Outlook

The conclusion that should be drawn from Lemmas 1 and 2 is the following; if one seeks an algorithm B with a regret bound of the form $\tilde{O}(\|\mathbf{w}\|L\sqrt{T})$ up to some lower-order terms in T , without knowledge of L and regardless of whether \mathcal{W} is bounded or not, it suffices to find an algorithm A which guarantees the sought type of regret whenever it has access to a sequence of hints (h_t) satisfying (as in Lemmas 1 and 2), $h_t \geq L_t$, for all $t \geq 1$. Thus, our first goal in the next section is to design a *scale-free* algorithm A which accesses such a sequence of hints and ensures that its linearized regret is bounded from above by:

$$O\left(\|\mathbf{w}\|\sqrt{V_T \ln(\|\mathbf{w}\|V_T)}\right), \quad \text{where } V_T := h_1^2 + \sum_{t=1}^T \|g_t\|^2, \quad (6)$$

for all $\mathbf{w} \in \mathbb{R}^d, T \geq 0$, and $(g_t) \subset \mathbb{R}^d$. We show an analogous “full-matrix” upgrade of order $\sqrt{\mathbf{w}^\top \mathbf{V} \mathbf{w} \ln(\mathbf{w}^\top \mathbf{V} \mathbf{w} \det \mathbf{V})}$, with $\mathbf{V} = \sum_{t=1}^T g_t g_t^\top$. We note that if Algorithm A in Lemmas 1 and 2 is scale-free, then so is the corresponding Algorithm B.

If the desired set \mathcal{W} has bounded diameter $D > 0$, then using the unconstrained-to-constrained reduction due to Cutkosky and Orabona (2018), it is straightforward to design a new algorithm based on A with regret also bounded by (6) up to an additive LD , for $\mathbf{w} \in \mathcal{W}$ (this is useful for Lemma 1).

Finally, we also note that algorithms which can access hints (h_t) such that $h_t \geq L_t$, for all $t \geq 1$, are of independent interest; in fact, it is the same algorithm A that we will use in Section 5 as a scale-invariant algorithm for learning linear models.

3. Scale-Free, Parameter-Free Algorithms for OCO

In light of the conclusions of Section 2, we will design new unconstrained scale-free algorithms which can access a sequence of hints (h_t) (as in Lemma 1) and guarantee a regret bound of the form given in (6). In this section, we will make the following assumption on the hints (h_t) :

Assumption 1 *We assume that (i) (h_t) is a non-decreasing sequence; (ii) $h_t \geq L_t$, for all $t \geq 1$; and (iii) if the sub-gradients (g_s) are multiplied by a factor $c > 0$, then the hints (h_t) are multiplied by the same factor c .*

The third item of the assumption ensures that our algorithms are scale-free. We note that Assumption 1 is satisfied by the sequence of hints that Algorithm B constructs when invoking Algorithm A in

Lemmas 1 and 2. For simplicity, we will also make the following assumption, which is without loss of generality, since the regret is zero while $\mathbf{g}_t = \mathbf{0}$.

Assumption 2 We assume that $L_1 = \|\mathbf{g}_1\| > 0$.

3.1. FREEGRAD: An Adaptive Scale-Free Algorithm

In this subsection, we design a new algorithm based on a time-varying potential function, where the outputs of the algorithm are uniquely determined by the gradients of the potential function at its iterates—an approach used in the design of many existing algorithms (Cesa-Bianchi et al., 1997).

Let $t \geq 1$, $(\mathbf{g}_s)_{s \leq t} \subset \mathbb{R}^d$ be a sequence of sub-gradients satisfying Assumption 2, and (h_t) be a sequence of hints satisfying Assumption 1. Consider the following potential function:

$$\Phi_t := S_t + \frac{h_1^2}{\sqrt{V_t}} \cdot \exp\left(\frac{\|\mathbf{G}_t\|^2}{2V_t + 2h_t\|\mathbf{G}_t\|}\right), \quad t \geq 0, \quad (7)$$

$$\text{where } S_t := \sum_{s=1}^t \langle \mathbf{g}_s, \hat{\mathbf{w}}_s \rangle, \quad \mathbf{G}_t := \sum_{s=1}^t \mathbf{g}_s, \quad V_t := h_1^2 + \sum_{s=1}^t \|\mathbf{g}_s\|^2. \quad (8)$$

This potential function has appeared as a by-product in the analyses of previous algorithms such as the ones in (Cutkosky and Orabona, 2018; Cutkosky, 2019). The expression of Φ_t in (7) is interesting to us since it can be shown via the *regret-reward duality* (McMahan and Orabona, 2014) (as we do in the proof of Theorem 5 below) that any algorithm which outputs vectors $(\hat{\mathbf{w}}_t)$ such that (Φ_t) is non-increasing for any sequence of sub-gradients (\mathbf{g}_t) , also guarantees a regret bound of the form (6). We will now design such an algorithm.

Consider the unconstrained algorithm FREEGRAD which at the beginning of round $t \geq 1$, uses the sequence of sub-gradients $(\mathbf{g}_s)_{s < t}$ seen so far and the available hint $h_t \geq L_t$ to output:

$$\hat{\mathbf{w}}_t := -\mathbf{G}_{t-1} \cdot \frac{(2V_{t-1} + h_t\|\mathbf{G}_{t-1}\|) \cdot h_1^2}{2(V_{t-1} + h_t\|\mathbf{G}_{t-1}\|)^2 \sqrt{V_{t-1}}} \cdot \exp\left(\frac{\|\mathbf{G}_{t-1}\|^2}{2V_{t-1} + 2h_t\|\mathbf{G}_{t-1}\|}\right). \quad (9)$$

where (\mathbf{G}_t) and (V_t) are as in (8). The output in (9) is obtained by setting the gradient $\nabla_{\mathbf{g}_t} \Phi_t$ at $\mathbf{g}_t = \mathbf{0}$ to the zero vector, and solving the resulting equation for $\hat{\mathbf{w}}_t \in \mathbb{R}^d$. Thus, for any $\hat{\mathbf{w}}_t$ other than the one in (9), one can find a vector \mathbf{g}_t such that $\|\mathbf{g}_t\| \leq h_t$, and $\Phi_t > \Phi_{t-1}$. Therefore, given that our aim is to find a sequence $(\hat{\mathbf{w}}_t)$ which makes (Φ_t) non-increasing, the outputs in (9) are the *unique* candidates. Our main technical contribution in this subsection is to show that, in fact, with the choice of $(\hat{\mathbf{w}}_t)_{t \geq 1}$ as in (9), the potential functions (Φ_t) are non-increasing for any sequence of sub-gradients (\mathbf{g}_t) :

Theorem 4 For $(\hat{\mathbf{w}}_t)$, and (Φ_t) as in (9), and (7), under Assumptions 1 and 2, we have:

$$\Phi_T \leq \dots \leq \Phi_0 = h_1, \quad \text{for all } T \geq 1.$$

The proof of the theorem is postponed to Appendix A. Theorem 4 and the regret-reward duality (McMahan and Orabona, 2014) yield a regret bound for the algorithm that outputs the sequence $(\hat{\mathbf{w}}_t)$. In fact, if $\Phi_T \leq \Phi_0$, then by the definition of Φ_T in (7), we have

$$\sum_{t=1}^T \langle \mathbf{g}_t, \hat{\mathbf{w}}_t \rangle \leq \Phi_0 - \Psi_T(\mathbf{G}_T), \quad \text{where } \Psi_T(\mathbf{G}) := \frac{h_1^2}{\sqrt{V_T}} \exp\left(\frac{\|\mathbf{G}\|^2}{2V_T + 2h_T\|\mathbf{G}\|}\right), \quad \mathbf{G} \in \mathbb{R}^d. \quad (10)$$

Now by Fenchel's inequality, we have $-\Psi_T(\mathbf{G}_T) \leq \langle \mathbf{w}, \mathbf{G}_T \rangle + \Psi_T^*(-\mathbf{w})$, for all $\mathbf{w} \in \mathbb{R}^d$, where $\Psi_T^*(\mathbf{w}) := \sup_{\mathbf{z} \in \mathbb{R}^d} \{\langle \mathbf{w}, \mathbf{z} \rangle - \Psi_T(\mathbf{z})\}$, $\mathbf{w} \in \mathbb{R}^d$, is the Fenchel dual of Ψ_T (Hiriart-Urruty and Lemaréchal, 2004). Combining this with (10), we obtain:

$$\sum_{t=1}^T \langle \mathbf{g}_t, \hat{\mathbf{w}}_t \rangle \leq \inf_{\mathbf{w} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \langle \mathbf{g}_t, \mathbf{w} \rangle + \Psi_T^*(-\mathbf{w}) + \Phi_0 \right\}, \quad (11)$$

Rearranging (11) for a given $\mathbf{w} \in \mathbb{R}^d$ leads to a regret bound of $\Psi_T^*(-\mathbf{w}) + \Phi_0$. Further bounding this quantity using existing results due to Cutkosky and Orabona (2018); Cutkosky (2019); McMahan and Orabona (2014), leads to the following regret bound (the proof is in Appendix B.1):

Theorem 5 *Under Assumptions 1 and 2, for $(\hat{\mathbf{w}}_t)$ as in (9), we have, with $\ln_+(\cdot) := 0 \vee \ln(\cdot)$,*

$$\sum_{t=1}^T \langle \mathbf{g}_t, \hat{\mathbf{w}}_t - \mathbf{w} \rangle \leq \left[2\|\mathbf{w}\| \sqrt{V_T \ln_+ \left(\frac{2\|\mathbf{w}\| V_T}{h_1^2} \right)} \right] \vee \left[4h_T \|\mathbf{w}\| \ln \left(\frac{4h_T \|\mathbf{w}\| \sqrt{V_T}}{h_1^2} \right) \right] + h_1,$$

for all $\mathbf{w} \in \mathcal{W} = \mathbb{R}^d, T \geq 1$.

Range-Ratio Problem. While the outputs $(\hat{\mathbf{w}}_t)$ in (9) of FREEGRAD are scale-free for the sequence of hints (h_t) satisfying Assumption (1), there remains one serious issue; the fractions V_T/h_1^2 and h_T/h_1 inside the log-terms in the regret bound of Theorem 5 could in principle be arbitrarily large if h_1 is small enough relative to h_T . Such a problematic ratio has appeared in the regret bounds of many previous algorithms which attempt to adapt to the Lipschitz constant L (Ross et al., 2013; Wintenberger, 2017; Kotłowski, 2017; Mhammedi et al., 2019; Kempka et al., 2019).

When the output set \mathcal{W} is bounded with diameter $D > 0$, this ratio can be dispensed of using a recently proposed restart trick due to Mhammedi et al. (2019), which restarts the algorithm whenever $L_t/L_1 > \sum_{s=1}^t \|\mathbf{g}_s\|/L_s$. The price to pay for this is merely an additive $O(LD)$ in the regret bound. However, this trick does not directly apply to our setting since in our case \mathcal{W} may be unbounded. Fortunately, we are able to extend the analysis of the restart trick to the unbounded setting where a sequence of hints (h_t) satisfying Assumption 1 is available; the cost we incur in the regret bound is an additive lower-order $\tilde{O}(\|\mathbf{w}\|L)$ term. Algorithm 1 displays our restart “wrapper”, FREERANGE, which uses the outputs of FREEGRAD to guarantee the following regret bound (the proof is in Appendix B):

Theorem 6 *Let $(\hat{\mathbf{w}}_t)$ be the outputs of FREERANGE (Algorithm 1). Then,*

$$\sum_{t=1}^T \langle \mathbf{g}_t, \hat{\mathbf{w}}_t - \mathbf{w} \rangle \leq 2\|\mathbf{w}\| \sqrt{2V_T \ln_+ (\|\mathbf{w}\| b_T)} + h_T \cdot (16\|\mathbf{w}\| \ln_+ (2\|\mathbf{w}\| b_T) + 2\|\mathbf{w}\| + 3),$$

for all $\mathbf{w} \in \mathbb{R}^d, T \geq 1$, and $(\mathbf{g}_t) \subset \mathbb{R}^d$, where $b_T := 2 \sum_{t=1}^T (\sum_{s=1}^{t-1} \frac{\|\mathbf{g}_s\|}{h_s} + 2)^2 \leq (T+1)^3$.

We next introduce our second algorithm, in which the variance is only measured in the comparator direction; the algorithm can be viewed as a “full-matrix” version of FREEGRAD.

Algorithm 1 FREERANGE: A Restart Wrapper for the Range-Ratio Problem (under Assumption 2).

Require: Hints (h_t) satisfying Assumption 1.

- 1: Set $\tau = 1$;
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Observe hint h_t ;
 - 4: **if** $h_t/h_\tau > \sum_{s=1}^{t-1} \|g_s\|/h_s + 2$ **then**
 - 5: Set $\tau = t$;
 - 6: **end if**
 - 7: Output \hat{w}_t as in (9) with $(h_1, V_{t-1}, \mathbf{G}_{t-1})$ replaced by $(h_\tau, h_\tau^2 + \sum_{s=\tau}^{t-1} \|g_s\|^2, \sum_{s=\tau}^{t-1} g_s)$;
 - 8: **end for**
-

3.2. MATRIX-FREEGRAD: Adapting to Directional Variance

Reflecting on the previous subsection, we see that the potential function that we ideally would like to use is $S_t + h_1 \exp\left(\frac{1}{2} \mathbf{G}_t^\top \mathbf{V}_t^{-1} \mathbf{G}_t - \frac{1}{2} \ln \det \mathbf{V}_t\right)$, $t \geq 1$, where $\mathbf{V}_t = \sum_{s=1}^t g_s g_s^\top$. However, as we saw, this is a little too greedy even in one dimension, and we need to introduce some slack to make the potential controllable. In the previous subsection we did this by increasing the scalar denominator from V to $V + \|\mathbf{G}\|$, which acts as a barrier function restricting the norm of \hat{w}_t . In this section, we will instead employ a hard norm constraint. We will further need to include a fudge factor $\gamma > 1$ multiplying \mathbf{V} to turn the above shape into a bona fide potential. To describe its effect, we define

$$\rho(\gamma) := \frac{1}{2\gamma} \left(\sqrt{(\gamma+1)^2 - 4e^{\frac{1}{2\gamma}-\frac{1}{2}} \gamma^{3/2}} + \gamma - 1 \right), \quad \text{for } \gamma \geq 1. \quad (12)$$

The increasing function ρ satisfies $\lim_{\gamma \rightarrow 1} \rho(\gamma) = 0$, $\lim_{\gamma \rightarrow \infty} \rho(\gamma) = 1$, and $\rho(2) = 0.358649$.

The potential function of this section is parameterized by a *prod factor* $\gamma > 1$ (which we will set to some universal constant). We define

$$\Psi(\mathbf{G}, \mathbf{V}, h) := \frac{h_1 \exp\left(\inf_{\lambda \geq 0} \left\{ \frac{1}{2} \mathbf{G}^\top (\gamma h_1^2 \mathbf{I} + \gamma \mathbf{V} + \lambda \mathbf{I})^{-1} \mathbf{G} + \frac{\lambda \rho(\gamma)^2}{2h^2} \right\}\right)}{\sqrt{\det\left(\mathbf{I} + \frac{1}{h_1^2} \mathbf{V}\right)}}, \quad (13)$$

where $\mathbf{G} \in \mathbb{R}^d$, $\mathbf{V} \in \mathbb{R}^{d \times d}$, and $h > 0$. Given a sequence of sub-gradients $(g_s)_{s < t}$, $t \geq 1$, and a hint $h_t \geq L_T$, we obtain the prediction at round t from the gradient of Ψ in the first argument

$$\hat{w}_t := -\nabla^{(1,0,0)} \Psi(\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t), \quad (14)$$

where $\mathbf{G}_{t-1} = \sum_{s=1}^{t-1} g_s$ and $\mathbf{V}_{t-1} := \sum_{s=1}^{t-1} g_s g_s^\top$. We can compute \hat{w}_t in $O(d^3)$ time per round by first computing an eigendecomposition of \mathbf{V}_{t-1} , followed by a one-dimensional binary search for the λ_\star which achieves the inf in (13) with $(\mathbf{G}, \mathbf{V}, h) = (\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t)$. Then the output is given by

$$\hat{w}_t = -\Psi(\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t) \cdot (\gamma h_1^2 \mathbf{I} + \gamma \mathbf{V}_{t-1} + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_{t-1}.$$

Our heavy-lifting step in the analysis is the following, which we prove in Appendix C:

Lemma 7 For any vector $g_t \in \mathbb{R}^d$ and $h_t > 0$ satisfying $\|g_t\| \leq h_t$, the vector \hat{w}_t in (14) ensures

$$g_t^\top \hat{w}_t \leq \Psi(\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t) - \Psi(\mathbf{G}_t, \mathbf{V}_t, h_t).$$

From here, we obtain our main result using telescoping and regret-reward duality:

Theorem 8 *Let $\Sigma_T^{-1} := \gamma h_1^2 \mathbf{I} + \gamma \mathbf{V}_T$. For $(\hat{\mathbf{w}}_t)$ as in (14), we have*

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \leq h_1 + \sqrt{Q_T^{\mathbf{w}} \ln_+ \left(\frac{\det(\gamma h_1^2 \Sigma_T)^{-1}}{h_1^2} Q_T^{\mathbf{w}} \right)}, \quad \text{for all } \mathbf{w} \in \mathbb{R}^d, \text{ where}$$

$$Q_T^{\mathbf{w}} := \max \left\{ \mathbf{w}^\top \Sigma_T^{-1} \mathbf{w}, \frac{1}{2} \left(\frac{h_T^2 \|\mathbf{w}\|^2}{\rho(\gamma)^2} \ln \left(\frac{\det(\gamma h_1^2 \Sigma_T)^{-1}}{h_1^2} \frac{h_T^2 \|\mathbf{w}\|^2}{\rho(\gamma)^2} \right) + \mathbf{w}^\top \Sigma_T^{-1} \mathbf{w} \right) \right\}.$$

Note in particular that the result is scale-free. Expanding the main case of the theorem (modest $\|\mathbf{w}\|$), we find regret bounded by

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \leq h_1 + h_1 \sqrt{\gamma \mathbf{w}^\top \mathbf{Q} \mathbf{w} \ln_+ (\gamma \mathbf{w}^\top \mathbf{Q} \mathbf{w} \det \mathbf{Q})} \quad \text{where} \quad \mathbf{Q} = \mathbf{I} + \mathbf{V}_T / h_1^2.$$

This bound looks almost like an ideal upgrade of that in Theorem 5, though technically, the bounds are not really comparable since the $\ln \det \mathbf{Q}$ can be as large as $d \ln T$, potentially canceling the advantage of having $\mathbf{w}^\top \mathbf{Q} \mathbf{w}$ instead of $\|\mathbf{w}\|^2 \sum_{t=1}^T \|\mathbf{g}_t\|^2$ inside the square-root. The matrix \mathbf{Q} and hence any directional variance $\mathbf{w}^\top \mathbf{Q} \mathbf{w}$ is scale-invariant. The only fudge factor in the answer is the $\gamma > 1$. We currently cannot tolerate $\gamma = 1$, for then $\rho(\gamma) = 0$ so the lower-order term would explode. We note that a bound of the form given in the previous display, with the $\ln \det \mathbf{Q}$ replaced by the larger term $d \ln \text{tr } \mathbf{Q}$, was achieved by a previous (not scale-free) algorithm due to [Cutkosky and Orabona \(2018\)](#).

Remark 9 *As Theorem 6 did in the previous subsection, our restarts method allows us to get rid of problematic scale ratios in the regret bound of Theorem 8; this can be achieved using FREERANGE with $(\hat{\mathbf{w}}_t)$ set to be as in (14) instead of (9). The key idea behind the proof of Theorem 6 is to show that the regrets from all but the last two epochs add up to a lower-order term in the final regret bound. This still holds when $(\hat{\mathbf{w}}_t)$ are the outputs of MATRIX-FREEGRAD instead of FREERANGE, since by Theorem 8, the regret bound of MATRIX-FREEGRAD is of order at most d times the regret of FREERANGE within any given epoch.*

As a final note about the algorithm, we may also develop a “one-dimensional” variant by replacing matrix inverse and determinant by their scalar analogues applied to $V_T = \sum_{t=1}^T \|\mathbf{g}_t\|^2$. One effect of this is that the minimization in λ can be computed in closed form. The resulting potential and corresponding algorithm and regret bound are very close to those of Section 3.1.

Conclusion The algorithms designed in this section can now be used in the role of algorithm A in the reductions presented in Section 2.2. This will yield algorithms which achieve our goal; they adapt to the norm of the comparator and the Lipschitz constant and are completely scale-free, for both bounded and unbounded sets, without requiring hints. We now show that the penalties incurred by these reductions are not improvable.

4. Lower Bounds

As we saw in Corollary 3, given a base algorithm A, which takes a sequence of hints (h_t) such that $h_t \geq L_t$, for all $t \geq 1$, and suffers regret $R_T^A(\mathbf{w})$ against comparator $\mathbf{w} \in \mathcal{W}$, there exists an algorithm B for the setting without hints which suffers the same regret against \mathbf{w} up to an additive penalty $L_T \|\mathbf{w}\|^3 + L_T \sqrt{\max_{t \in [T]} B_t}$, where $B_t = \sum_{s=1}^t \|\mathbf{g}_s\|/L_t$. In this section, we show that the penalty $L_T \|\mathbf{w}\|^3$ is not improvable if one insists on a regret bound of order $\tilde{O}(\sqrt{T})$. We also show that it is not possible to replace the penalty $L_T \sqrt{\max_{t \in [T]} B_t}$ by the typically smaller quantity $\sqrt{V_T}$, where $V_T = \sum_{t=1}^T \|\mathbf{g}_t\|^2$. Our starting point is the following lemma:

Lemma 10 *For all $t \geq 1$, past sub-gradients $(\mathbf{g}_s)_{s \leq t}$ and past and current outputs $(\hat{\mathbf{w}}_s)_{s \leq t} \in \mathbb{R}^d$,*

$$\exists \mathbf{g}_t \in \mathbb{R}^d, \quad \sum_{s=1}^t \langle \mathbf{g}_s, \hat{\mathbf{w}}_s \rangle \geq \|\hat{\mathbf{w}}_t\| \cdot L_t/2, \quad \text{where } L_t = \max_{s \leq t} \|\mathbf{g}_s\|.$$

Proof We want to find \mathbf{g}_t such that $\langle \mathbf{g}_t, \hat{\mathbf{w}}_t \rangle \geq \|\mathbf{w}_t\| L_t/2 - S_{t-1}$, where $S_{t-1} := \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \hat{\mathbf{w}}_s \rangle$. By restricting \mathbf{g}_t to be aligned with $\hat{\mathbf{w}}_t$, the problem reduces to finding $x = \|\mathbf{g}_t\|$ such that

$$x \|\hat{\mathbf{w}}_t\| - \|\hat{\mathbf{w}}_t\| \cdot (L_{t-1} \vee x)/2 - S_{t-1} \geq 0. \quad (15)$$

The LHS of (15) is a piece-wise linear function in x which goes to infinity as $x \rightarrow \infty$. Therefore, there exists a large enough $x \geq 0$ which satisfies (15). \blacksquare

Observe that if $\|\hat{\mathbf{w}}_t\| \geq D_t > 0$, for $t \geq 1$, then by Lemma 10, there exists a sub-gradient \mathbf{g}_t which makes the regret against $\mathbf{w} = \mathbf{0}$ at round t at least $D_t L/2$. This essentially means that if the sub-gradients (\mathbf{g}_t) are unbounded, then the outputs $(\hat{\mathbf{w}}_t)$ must be in a bounded set whose diameter will depend on the desired regret bound; if one insists on a regret of order $\tilde{O}(\sqrt{T})$, then the norm of the outputs $\hat{\mathbf{w}}_t, t \geq 1$, must be in a ball of radius at most $\tilde{O}(\sqrt{T})$.

Cutkosky (2019) posed the question of whether there exists an algorithm which can guarantee a regret bound of order $L \|\mathbf{w}\|^3 + (\|\mathbf{w}\| + 1) \sqrt{V_T \ln T}$, with $V_T = \sum_{t=1}^T \|\mathbf{g}_t\|^2$, while adapting to both L and $\|\mathbf{w}\|$. Here we ask the question whether $L \|\mathbf{w}\|^\nu + (\|\mathbf{w}\| + 1) \sqrt{V_T \ln T}$ is possible for any $\nu \geq 1$. If such an algorithm exists, then by Lemma 10, there exists a constant $b > 0$ such that its outputs $(\hat{\mathbf{w}}_t)$ satisfy $\|\hat{\mathbf{w}}_t\| \leq b \sqrt{V_t \ln t}/L_t$, for all $t \geq 1$. The next lemma, when instantiated with $\alpha = 2$, gives us a regret lower-bound on such algorithms (the proof is in Appendix D):

Lemma 11 *For all $b, c, \beta \geq 0$, $\nu \geq 1$, and $\alpha \in [1, 2]$, there exists $(\mathbf{g}_t) \in \mathbb{R}^d$, $T \geq 1$, and $\mathbf{w} \in \mathbb{R}^d$, such that for any sequence $(\hat{\mathbf{w}}_t)$ satisfying $\|\hat{\mathbf{w}}_t\| \leq b \cdot \sqrt{V_{\alpha,t} \ln(t)}/L_t^\alpha$, for all $t \in [T]$, where $V_{\alpha,t} := \sum_{s=1}^t \|\mathbf{g}_s\|^\alpha$, we have*

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \geq c \cdot \ln(1 + \|\mathbf{w}\| T)^\beta \cdot (L_T \|\mathbf{w}\|^\nu + L_T^{1-\alpha/2} (\|\mathbf{w}\| + 1) \sqrt{V_{\alpha,T} \ln T}).$$

By combining the results of Lemma 10 and 11, we have the following regret lower bound for algorithms which can adapt to both L and $\|\mathbf{w}\|$:

Theorem 12 *For any $\alpha \in [1, 2]$, $c > 0$ and $\nu \geq 1$, there exists no algorithm that guarantees, up to log-factors in $\|\mathbf{w}\|$ and T , a regret bound of the form $c \cdot (L_T \|\mathbf{w}\|^\nu + L_T^{1-\alpha/2} (\|\mathbf{w}\| + 1) \sqrt{V_{\alpha,T} \ln T})$, for all $T \geq 1$, $\mathbf{w} \in \mathbb{R}^d$, and $(\mathbf{g}_t) \subset \mathbb{R}^d$, where $V_{\alpha,T} := \sum_{t=1}^T \|\mathbf{g}_t\|^\alpha$.*

Proof By Lemma 10, the only candidate algorithms are those whose outputs $(\hat{\mathbf{w}}_t)$ satisfy $\|\hat{\mathbf{w}}_t\| \leq b\sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha}$, for all $t \geq 1$, for some constant $b > 0$. By Lemma 11, no such algorithms can achieve the desired regret bound. \blacksquare

The regret lower bound in Theorem 12 does not apply to the case where $\alpha = 1$. In fact, thanks to Corollary 3 and our main algorithm in Section 3 (which can play the role of Algorithm A in Corollary 3), we know that there exists an algorithm B which guarantees a regret bound of order $\tilde{O}(L_T \|\mathbf{w}\|^3 + \|\mathbf{w}\| \sqrt{V_T} + L_T \sqrt{\max_{t \in [T]} B_t})$, where $B_t = \sum_{s=1}^t \|\mathbf{g}_s\|/L_t$. Next we show that if one insists on a regret bound of order $\sqrt{B_T}$, or even \sqrt{T} (up to log-factors), the exponent in $\|\mathbf{w}\|^3$ is unimprovable (the proof of Theorem 13 is in Appendix D.2).

Theorem 13 *For any $\nu \in [1, 3[$ and $c > 0$, there exists no algorithm that guarantees, up to log-factors in $\|\mathbf{w}\|$ and T , a regret bound of the form $c \cdot (L_T \|\mathbf{w}\|^\nu + L_T(\|\mathbf{w}\| + 1)\sqrt{T \ln T})$, for all $T \geq 1$, $\mathbf{w} \in \mathbb{R}^d$, and $(\mathbf{g}_t) \subset \mathbb{R}^d$.*

5. Application to Learning Linear Models with Online Algorithms

In this section, we consider the setting of online learning of linear models which is a special case of OCO. At the start of each round $t \geq 1$, a learner receives a feature vector $\mathbf{x}_t \in \mathcal{W} = \mathbb{R}^d$, then issues a prediction $\hat{y}_t \in \mathbb{R}$ in the form of an inner product between \mathbf{x}_t and a vector $\hat{\mathbf{u}}_t \in \mathbb{R}^d$, i.e. $\hat{y}_t = \hat{\mathbf{u}}_t^\top \mathbf{x}_t$. The environment then reveals a label $y_t \in \mathbb{R}$ and the learner suffers loss $\ell(y_t, \hat{y}_t)$, where $\ell: \mathbb{R}^2 \rightarrow \mathbb{R}$ is a fixed loss function which is convex and 1-Lipschitz in its second argument; this covers popular losses such as the logistic, hinge, absolute and Huberized squared loss. (Technically, the machinery developed so far and the reductions in Section 2.2 allow us to handle the non-Lipschitz case).

In the current setting, the regret is measured against the best fixed “linear model” $\mathbf{w} \in \mathbb{R}^d$ as

$$\text{REGRET}_T(\mathbf{w}) := \sum_{t=1}^T \ell(y_t, \hat{y}_t) - \sum_{t=1}^T \ell(y_t, \mathbf{w}^\top \mathbf{x}_t) \leq \sum_{t=1}^T \delta_t \langle \mathbf{x}_t, \hat{\mathbf{u}}_t - \mathbf{w} \rangle, \quad (16)$$

where the last inequality holds for any sub-gradients $\delta_t \in \partial^{(0,1)} \ell(y_t, \hat{y}_t)$, $t \geq 1$, due to the convexity of ℓ in its second argument, which in turn makes the function $f_t(\mathbf{w}) := \ell(y_t, \mathbf{w}^\top \mathbf{x}_t)$ convex for all $\mathbf{w} \in \mathcal{W} = \mathbb{R}^d$. Here, $\partial^{(0,1)} \ell$ denotes the sub-differential of ℓ with respect to its second argument. Thus, minimizing the regret in (16) fits into the OCO framework described in Section 2. In fact, we will show how our algorithms from Section 3 can be applied in this setting to yield scale-free, and even *rotation-free*, (all with respect to the feature vectors (\mathbf{x}_t)) algorithms for learning linear models. These algorithms can, without any prior knowledge on \mathbf{w} or $(\mathbf{w}^\top \mathbf{x}_t)$, achieve regret bounds against any $\mathbf{w} \in \mathbb{R}^d$ matching (up to log-factors) that of OGD with optimally tuned learning rate.

As in Section 3, we focus on algorithms which make predictions based on observed sub-gradients (\mathbf{g}_t) ; in this case, $\mathbf{g}_t = \mathbf{x}_t \delta_t \in \mathbf{x}_t \cdot \partial^{(0,1)} \ell(y_t, \hat{y}_t) = \partial f_t(\hat{\mathbf{u}}_t)$, $t \geq 1$, where $f_t(\mathbf{w}) = \ell(y_t, \mathbf{w}^\top \mathbf{x}_t)$. Since the loss ℓ is 1-Lipschitz, we have $|\delta_t| \leq 1$, for all $\delta_t \in \partial^{(0,1)} \ell(y_t, \hat{y}_t)$ and $t \geq 1$, and so $\|\mathbf{g}_t\| \leq \|\mathbf{x}_t\|$. Since \mathbf{x}_t is revealed at the beginning of round $t \geq 1$, the hint

$$h_t = \max_{s \leq t} \|\mathbf{x}_s\| \geq L_T = \max_{s \leq t} \|\mathbf{g}_s\| \quad (17)$$

is available ahead of outputting $\hat{\mathbf{u}}_t$, and so our algorithms from Section 3 are well suited for this setting.

Improvement over Current Algorithms. We improve on current state-of-the-art algorithms in two ways; First, we provide a (coordinate-wise) scale-invariant algorithm which guarantees regret bound, against any $w \in \mathbb{R}^d$, of order

$$\sum_{i=1}^d |w_i| \sqrt{V_{T,i} \ln(|w_i| \sqrt{V_{T,i} T})} + |w_i| \ln_+(|w_i| \sqrt{V_{T,i} T}), \quad (18)$$

where $V_{T,i} := |x_{1,i}|^2 + \sum_{t=1}^T \delta_t^2 |x_{t,i}|^2$, $i \in [d]$, which improves the regret bound of the current state-of-the-art scale-invariant algorithm SCLNOL_1 (Kempka et al., 2019) by a $\sqrt{\ln(\|w\|T)}$ factor. Second, we provide an algorithm that is both scale and rotation invariant with respect to the input feature vectors (x_t) with a state-of-the-art regret bound; by scale and rotation invariance we mean that, if the sequence of feature vectors (x_t) is multiplied by cO , where $c > 0$ and O is any special orthogonal matrix in $\mathbb{R}^{d \times d}$, the outputs (\hat{y}_t) of the algorithm remain unchanged. Arguably the closest algorithm to ours in the latter case is that of Kotłowski (2017) whose regret bound is essentially of order $\tilde{O}(\sqrt{w^\top S_T w})$ for any comparator $w \in \mathbb{R}^d$, where $S_T = \sum_{t=1}^T x_t x_t^\top$. However, in our case, instead of the matrix S_T , we have $V_T := \|x_1\|^2 I + \sum_{t=1}^T x_t x_t^\top \delta_t^2$, where $\delta_t \in \partial^{(0,1)} \ell(y_t, \hat{y}_t)$, $t \geq 1$, which can yield a much smaller bound for small (δ_t) (this typically happens when the algorithm starts to “converge”).

A Scale-Invariant Algorithm. To design our first scale-invariant algorithm, we will use the outputs (\hat{w}_t) of FREEGRAD in (9) with (h_t) as in (17), and a slight modification of FREERANGE (see Algorithm 2). This modification consists of first scaling the outputs (\hat{w}_t) of FREEGRAD by the initial hint of the current epoch to make the predictions (\hat{y}_t) scale-invariant. By Theorem 18 below, the regret bound corresponding to such scaled outputs will have a lower-order term which, unlike in the regret bound of Theorem 5, does not depend on the initial hint. This breaks our current analysis of FREERANGE in the proof of Theorem 6 which we used to overcome the range-ratio problem. To solve this issue, we further scale the output \hat{w}_t at round $t \geq 1$ by the sum $\sum_{s=1}^\tau \|x_s\|/h_s$, where τ denotes the first index of the current epoch (see Algorithm 2). Due to this change, the proof of the next theorem differs slightly from that of Theorem 6.

First, we study the regret bound of Algorithm 2 in the case where $\mathcal{W} = \mathbb{R}$.

Theorem 14 *Let $d = 1$ and (h_t) be as in (17). If (\hat{u}_t) are the outputs of Algorithm 2, then for all $w \in \mathbb{R}$; $T \geq 1$; $(x_t, y_t) \subset \mathbb{R}^2$, s.t. $h_1 = |x_1| > 0$; and $\delta_t \in \partial^{(0,1)} \ell(y_t, x_t \hat{u}_t)$, $t \in [T]$,*

$$\begin{aligned} \sum_{t=1}^T \delta_t x_t \cdot (\hat{u}_t - w) &\leq 2|w| \sqrt{V_T \ln_+(2|w|^2 V_T c_T)} \\ &\quad + h_T |w| (14 \ln_+(2|w| \sqrt{2V_T c_T}) + 1) + 2 + \ln B_T, \end{aligned} \quad (19)$$

where $V_T := |x_1|^2 + \sum_{t=1}^T \delta_t^2 x_t^2$, $c_T := 2B_T^2 \sum_{t=1}^T \left(\sum_{s=1}^t \frac{|x_s|}{h_s} \right)^2 \leq T^5$, and $B_T = \sum_{s=1}^T \frac{|x_s|}{h_s} \leq T$.

The proof of Theorem 14 is in Appendix E. If (\hat{u}_t) are the outputs of Algorithm 2 in the one-dimensional case, then by Theorem 14 and (16), the algorithm which, at each round $t \geq 1$, predicts $\hat{y}_t = x_t \hat{u}_t$ has regret bounded from above by the RHS of (19). Note also that the outputs (\hat{y}_t) are scale-invariant.

Now consider an algorithm A which at round $t \geq 1$ predicts $\hat{y}_t = \sum_{i=1}^d x_{t,i} \hat{u}_{t,i}$, where $(\hat{u}_{t,i})$, $i \in [d]$, are the outputs of Algorithm 2 when applied to coordinate i ; in this case, we will have a sequence

Algorithm 2 Modified FREERANGE for the setting of online learning of linear models.

Require: The hints (h_t) as in (17).

- 1: Set $\tau = 1$;
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Observe hint h_t ;
 - 4: **if** $h_t/h_\tau > \sum_{s=1}^{t-1} \|x_s\|/h_s + 1$ **then**
 - 5: Set $\tau = t$;
 - 6: **end if**
 - 7: Output $\hat{w}_t = \hat{w}_\tau \cdot \left(h_\tau \cdot \sum_{s=1}^\tau \frac{\|x_s\|}{h_s} \right)^{-1}$, where \hat{w}_t is as in (9) with $(h_1, V_{t-1}, \mathbf{G}_{t-1})$ replaced by $(h_\tau, h_\tau^2 + \sum_{s=\tau}^{t-1} \|g_s\|^2, \sum_{s=\tau}^{t-1} g_s)$;
 - 8: **end for**
-

of hints $(h_{t,i})$ for each coordinate i satisfying $h_{t,i} = \max_{s \leq t} |x_{t,i}|$, for all $t \geq 1$. Algorithm A is coordinate-wise scale-invariant, and due to (16) and Theorem 14, it guarantees a regret bound of the form (18). We note, however, that a factor d will appear multiplying the lower-order term $(2 + \ln B_T)$ in (19) (since the regret bounds for the different coordinates are added together). To avoid this, at the cost of a factor d appearing inside the logarithms in (18), it suffices to divide the outputs of algorithm A by d . To see why this works, see Theorem 18 in the appendix.

A Rotation-Invariant Algorithm. To obtain a rotation and scale-invariant online algorithm for learning linear models we will make use of the outputs of MATRIX-FREEGRAD instead of FREEGRAD. Let (\hat{y}_t) be the sequence of predictions defined by

$$\hat{y}_t = x_t^\top \hat{w}_t / h_1, \quad t \geq 1, \quad (20)$$

with (h_t) as in (17) and where \hat{w}_t are the predictions of a variant of MATRIX-FREEGRAD, where the leading h_1 in the potential (13) is replaced by 1 (we analyze this variant in Appendix C.1).

Theorem 15 *Let $\gamma > 0$ and (h_t) be as in (17). If (\hat{y}_t) are as in (20), then*

$$\forall w \in \mathbb{R}^d, \forall T \geq 1, \forall (g_t) \subset \mathbb{R}^d, \quad \text{REGRET}_T(w) \leq 1 + \sqrt{Q_T^w \ln_+ \left(\det(\gamma h_1^2 \Sigma_T)^{-1} Q_T^w \right)}, \quad \text{where}$$

$$Q_T^w := \max \left\{ w^\top \Sigma_T^{-1} w, \frac{1}{2} \left(\frac{h_T \|w\|^2}{\rho(\gamma)^2} \ln \left(\frac{h_T \|w\|^2}{\rho(\gamma)^2} \det(\gamma h_1^2 \Sigma_T)^{-1} \right) + w^\top \Sigma_T^{-1} w \right) \right\},$$

$$\text{and } \Sigma_T^{-1} := \gamma h_1^2 \mathbf{I} + \gamma \sum_{t=1}^T g_t g_t^\top.$$

Proof It suffices to use (16) and instantiate the regret bound in Theorem 19 with $(\epsilon, \sigma^{-2}) = (1, \gamma h_1^2)$. ■

The range-ratio problem manifests itself again in Theorem 15 through the term $\det(\gamma h_1^2 \Sigma_T)^{-1}$. This can be solved using the outputs of Algorithm 2, where in Line 7, \hat{w}_t is taken to be as in (20) (see Remark 9).

References

- Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.
- Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pages 874–894, 2019.
- Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on Learning Theory*, pages 643–677, 2017.
- Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1493–1529, 2018.
- Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3666–3674, 2016.
- Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Zigzag: A new approach to adaptive online learning. In *Proc. of the 2017 Annual Conference on Learning Theory (COLT)*, pages 876–924, 2017.
- Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Online learning: Sufficient statistics and the Burkholder method. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 3028–3064, 2018.
- Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.
- Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Science & Business Media, 2004.
- Michał Kempka, Wojciech Kotłowski, and Manfred K. Warmuth. Adaptive scale-invariant online algorithms for learning linear models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3321–3330, 2019.
- Wojciech Kotłowski. Scale-invariant unconstrained online learning. In *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, pages 412–433, 2017.
- H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pages 1020–1039, 2014.
- H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 244–256, 2010.

- Zakaria Mhammedi, Wouter M. Koolen, and Tim van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 2490–2511, 2019.
- Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, pages 577–585, 2016a.
- Francesco Orabona and Dávid Pál. Open problem: Parameter-free and scale-free online algorithms. In *Conference on Learning Theory*, pages 1659–1664, 2016b.
- Stephane Ross, Paul Mineiro, and John Langford. Normalized online learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 537–545, 2013.
- Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Appendix A. Proof of Theorem 4

The proof of Theorem 4 relies on the following key lemma:

Lemma 16 For $G, g \in \mathbb{R}$ and, $V > 0$, define

$$\Theta(G, V, g) := \frac{\sqrt{V}}{\sqrt{V + g^2}} \cdot \exp \left(\frac{(G + g)^2}{2V + 2g^2 + 2|G + g|} - \frac{G^2}{2V + 2|G|} \right) - \frac{gG(|G| + 2V)}{2(|G| + V)^2} - 1.$$

It holds that $\Theta(G, V, g) \leq 0$, for all $G \in \mathbb{R}$, $V > 0$, and $g \in [-1, 1]$.

Proof For notational simplicity we assume $G \geq 0$. Let us look at

$$\Gamma(G, V, g) := \frac{1}{2} \frac{(g + G)^2}{V + g^2 + |G + g|} - \frac{1}{2} \frac{G^2}{G + V} - \ln \left(1 + \frac{gG(G + 2V)}{2(G + V)^2} \right) - \frac{1}{2} \ln \left(1 + \frac{g^2}{V} \right).$$

Since \ln is increasing, we have that $\Theta \leq 0$, if and only if, $\Gamma \leq 0$, and so we want to show $\Gamma \leq 0$ for all $V > 0$, $G \geq 0$, and $g \in [-1, 1]$. Our approach will be to show that Γ is increasing in V . The result then follows from $\lim_{V \rightarrow \infty} \Gamma = 0$. It remains to study the derivative

$$\begin{aligned} \frac{\partial \Gamma}{\partial V} = & -\frac{1}{2} \frac{(g + G)^2}{(|g + G| + g^2 + V)^2} + \frac{2gGV}{(G + V)(2(G + V)^2 + gG(G + 2V))} \\ & + \frac{1}{2} \frac{G^2}{(G + V)^2} + \frac{1}{2} \frac{g^2}{g^2V + V^2}. \end{aligned}$$

Factoring this as a ratio of polynomials, we obtain:

$$\frac{\partial \Gamma}{\partial V} = \frac{\alpha_0 + \alpha_1 V + \alpha_2 V^2 + \alpha_3 V^3 + \alpha_4 V^4 + \alpha_5 V^5}{2V(g^2 + V)(G + V)^2((g + 2)G^2 + 2(g + 2)GV + 2V^2)(|g + G| + g^2 + V)^2},$$

where $\alpha_i, i \in [5]$, are polynomials in g and G whose explicit (yet gruesome) expressions are:

$$\begin{aligned} \alpha_0 &= g^2(g + 2)G^4(|g + G| + g^2)^2 \\ \alpha_1 &= g^2(g + 2)G^3(2(g^2(G + 4) + G)|g + G| + g^4(G + 4) + 2g^2(G + 2) + 8gG + 4G^2) \\ \alpha_2 &= g^2G^2 \left(\begin{aligned} &2(g^3(2G + 9) + 4g^2(G + 3) + 2gG(G + 2) + 4G(G + 2))|g + G| \\ &+ 3g^2(3g + 4)(g^2 + 1) - 2(g + 2)G^3 \\ &+ (g(g(3g + 2) + 2) + 14)G^2 + 2g(g(g(g + 2) + 3) + 15) + 12)G \end{aligned} \right) \\ \alpha_3 &= G \left(\begin{aligned} &2(6g^5 + 2g^4(G + 4) + g^3G(4G + 13) + 4g^2G(2G + 3) + gG^3 + 2G^3)|g + G| \\ &+ (g(g + 2)(2g + 13) + 20)g^3G + 2(g(3g^2 + g - 6) + 10)g^2G^2 \\ &+ 2(3g + 4)(g^2 + 1)g^4 - 2(g + 2)G^4 - 2(g(g + 4) + 2)gG^3 \end{aligned} \right) \\ \alpha_4 &= 2 \left(\begin{aligned} &2(g^4 + (5g + 4)g^2G + (g + 2)G^3 + 2(g + 1)gG^2)|g + G| + g^6 + g^4 \\ &+ (g(7g + 4) + 4)g^3G + ((g - 2)g + 6)g^2G^2 - 2(g + 2)G^4 + (g(2g - 3) - 8)gG^3 \end{aligned} \right) \\ \alpha_5 &= 2 \left(2|g + G|^3 + g^4 + 4g^3G - 6gG^2 - 2G^3 \right) \end{aligned}$$

Under our assumptions $V > 0$, $G > 0$ and $g \in [-1, 1]$, the denominator of $\partial \Gamma / \partial V$ above is positive. Furthermore, its numerator, regarded as a polynomial in V , has exclusively positive

```

In[*]:= $Assumptions = V > 0 && Element[G, Reals] && -1 ≤ g ≤ 1;
Let us call "gap" the difference between what we have and the upper bound we want to establish.


$$\Phi[G, V] = -\frac{G^2}{2(V + \sqrt{G^2})} + \frac{1}{2} \log[V]$$


In[*]:= gap = -Log[1 - g D[Φ[G, V], G]] - (Φ[G + g, V + g^2] - Φ[G, V]) // FullSimplify
We want to show that the gap is ≤ 0. Our approach will be to show that gap is increasing, so that we
can then bound it by the limit

In[*]:= Limit[gap, V → Infinity] // FullSimplify
So why is gap increasing? Let us take the derivative

In[*]:= dgap = D[gap, V] // Simplify
and write it as a ratio of polynomials

In[*]:= {num, den} = With[{rpoly = Factor[dgap]}, NumeratorDenominator[rpoly] // Simplify]
Now the denominator is always positive

In[*]:= den ≥ 0 // FullSimplify
We will show that the numerator is positive by showing that it is a polynomial with only positive
coefficients.
Here are the coefficients on the monomials Vi for i=0,1,...

In[*]:= coefs = CoefficientList[num, V] // Simplify
And all the coefficients are positive

Map[FullSimplify[# ≥ 0] &, coefs]

Out[*]:= {True, True, True, True, True, True}

```

Figure 1: Mathematica notebook in support of Theorem 4.

coefficients $\alpha_i \geq 0$, as can be verified using computer algebra software (we used Mathematica's `FullSimplify`—see Figure 1). This implies that $\partial\Gamma/\partial V \geq 0$, for all $G \in \mathbb{R}$, $V > 0$, and $g \in [-1, 1]$, and so $\Gamma \leq \lim_{V \rightarrow \infty} \Gamma = 0$. ■

We are ourselves a bit disgruntled about the opacity of the above proof. On the one hand, it is just a tedious verification of an analytic statement about a function of three scalar variables, and one might expect that tighter statements require more sophisticated techniques (c.f. [Kotłowski, 2017](#), Appendix F). It is quite plausible that positivity may be established in a somewhat more streamlined fashion using a Sum-of-Squares techniques. Yet on the other hand, we were hoping to gain, from the proof, a deeper insight into the design of potential functions. Unfortunately this did not materialize. In particular, we still do not know how to address the multi-dimensional case of our Section 3.2 with

a similar potential. Controlling the intuitive upgrade of (7) where the exponential is replaced by

$$\exp \left(\sup_{\mathbf{u}} \frac{(\mathbf{u}^\top \mathbf{G})^2}{2(\mathbf{u}^\top \mathbf{V} \mathbf{u} + L \|\mathbf{u}\| \cdot \|\mathbf{u}^\top \mathbf{G}\|)} - \frac{1}{2} \ln \det \mathbf{V} \right)$$

is impossible, as witnessed by numerical counterexamples returned by random search, already in dimension 2.

We need one more result before we prove of Theorem 4:

Lemma 17 *Let $G, s \in \mathbb{R}$ and $V, h \geq 0$. Then, the function*

$$g \mapsto \frac{1}{\sqrt{V + g^2}} \exp \left(\frac{g^2 + 2s + G^2}{2V + 2g^2 + 2h\sqrt{g^2 + 2s + G^2}} \right),$$

is non-increasing on $\{g \geq 0 \mid g^2 + 2s + G^2 \geq 0\}$.

Proof It suffices to show that the function

$$\Xi(g) := \frac{g^2 + 2s + G^2}{2V + 2g^2 + 2h\sqrt{g^2 + 2s + G^2}} - \frac{1}{2} \ln(V + g^2),$$

is non-increasing on $\{g \geq 0 \mid g^2 + 2s + G^2 \geq 0\}$. Evaluating the derivative of Ξ , we find that

$$\frac{d\Xi}{dg}(g) = -\frac{gN \cdot (2h^2N + 2VN + 2g^2N + 3g^2h + 3hV)}{2(g^2 + V)(hN + g^2 + V)^2}, \quad (21)$$

where $N := \sqrt{g^2 + G^2 + 2s}$. The derivative in (21) is non-positive for all $V, h \geq 0$ and $g \geq 0$ such that $g^2 + 2s + G^2 \geq 0$. \blacksquare

Proof of Theorem 4. We will proceed by induction. By the fact that $\|\mathbf{G}_0\| = 0$ and the definition of the potential in (7), we have $\Phi_0 = h_1$. Now let $t \geq 0$, $h_1 > 0$, and $(S_t, V_t, h_t, \mathbf{G}_t) \in \mathbb{R} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \times \mathbb{R}^d$. We will show that $\Phi_{t+1} - \Phi_t \leq 0$. First, note that for any $h_{t+1} \geq h_t$, we have

$$S_t + \frac{h_1^2}{\sqrt{V_t}} \cdot \exp \left(\frac{\|\mathbf{G}_t\|^2}{2V_t + 2h_{t+1}\|\mathbf{G}_t\|} \right) \leq \Phi_t = S_t + \frac{h_1^2}{\sqrt{V_t}} \cdot \exp \left(\frac{\|\mathbf{G}_t\|^2}{2V_t + 2h_t\|\mathbf{G}_t\|} \right).$$

Thus, for any $\mathbf{g}_{t+1} \in \mathbb{R}^d$ such that $\|\mathbf{g}_{t+1}\| \leq h_{t+1}$, and $(S, V, h, \mathbf{G}, \mathbf{g}) := (S_t, V_t, h_{t+1}, \mathbf{G}_t, \mathbf{g}_{t+1})$,

$$\begin{aligned} \Phi_{t+1} - \Phi_t &\leq \frac{h_1^2}{\sqrt{V + \|\mathbf{g}\|^2}} \cdot \exp \left(\frac{\|\mathbf{g} + \mathbf{G}\|^2}{2V + 2\|\mathbf{g}\|^2 + 2h\|\mathbf{g} + \mathbf{G}\|} \right) \\ &\quad - \left(1 + \frac{\langle \mathbf{g}, \mathbf{G} \rangle \cdot (2h\|\mathbf{G}\| + 2V)}{2(h\|\mathbf{G}\| + V)^2} \right) \cdot \frac{h_1^2}{\sqrt{V}} \cdot \exp \left(\frac{\|\mathbf{G}\|^2}{2V + 2h\|\mathbf{G}\|} \right). \end{aligned} \quad (22)$$

Let \mathbf{g}_\star be the vector $\mathbf{g} \in \mathcal{B}_h$ which maximizes the RHS of (22), where \mathcal{B}_h is the ball in \mathbb{R}^d of radius h . Suppose that $\mathbf{G} \neq \mathbf{0}$, and let $\mathcal{H} := \{\mathbf{g} \in \mathbb{R}^d \mid \langle \mathbf{g}, \mathbf{G} \rangle = \langle \mathbf{g}_\star, \mathbf{G} \rangle\}$. Note that within the hyperplane \mathcal{H} , only the first term on the RHS of (22) varies. Since \mathbf{g}_\star is the maximizer of the RHS of (22) within \mathcal{B}_h , instantiating Lemma 17 with $s := \langle \mathbf{g}_\star, \mathbf{G} \rangle$ and $G := \|\mathbf{G}\|$, implies that $\mathbf{g}_\star \in \operatorname{argmin}\{\|\mathbf{g}\| \mid \mathbf{g} \in \mathcal{H}\}$.

Adding this to the fact that \mathcal{H} is a hyperplane orthogonal to \mathbf{G} implies that \mathbf{g}_\star and \mathbf{G} must be aligned, *i.e.* there exists a $c_\star \in \mathbb{R}$ such that $\mathbf{g}_\star = c_\star \mathbf{G} / \|\mathbf{G}\|$. Therefore, we have $\|\mathbf{g}_\star + \mathbf{G}\| = |g_\star + G|$, where

$$g_\star := \begin{cases} c_\star, & \text{if } G > 0; \\ \|\mathbf{g}_\star\|, & \text{otherwise.} \end{cases}$$

Further, note that $|g_\star| \leq h$. Thus, the RHS of (22) is bounded from above by

$$\begin{aligned} \Delta &:= \frac{h_1^2}{\sqrt{V + g_\star^2}} \cdot \exp\left(\frac{(g_\star + G)^2}{2V + 2g_\star^2 + 2h|g_\star + G|}\right) \\ &\quad - \left(1 + \frac{g_\star G \cdot (2V + 2h|G|)}{2(V + h|G|)^2}\right) \cdot \frac{h_1^2}{\sqrt{V}} \cdot \exp\left(\frac{G^2}{2V + 2h|G|}\right). \end{aligned} \quad (23)$$

Note that Δ in (23) can be written in terms of the function Θ in Lemma 16 as:

$$\Delta = \frac{h_1^2}{\sqrt{V}} \cdot \exp\left(\frac{G^2}{2V + 2hG}\right) \cdot \Theta\left(\frac{G}{h}, \frac{V}{h^2}, \frac{g_\star}{h}\right).$$

Since $(G/h, V/h^2, g_\star/h) \in \mathbb{R} \times \mathbb{R}_{>0} \times [-1, 1]$, Lemma 16 implies that $\Theta(G/h, V/h^2, g_\star/h) \leq 0$, and so due to (23), we also have $\Delta \leq 0$. Since Δ is an upper-bound on the RHS of (22), it follows that $\Phi_{t+1} - \Phi_t \leq 0$ as desired. \blacksquare

Appendix B. Proofs of Section 3.1

B.1. Proof of Theorem 5

The proof of Theorem 5 follows from the next theorem by setting $\epsilon = 1$. Theorem 18 essentially gives the regret bound of FREEGRAD if its outputs $(\hat{\mathbf{w}}_t)$ are scaled by a constant $\epsilon > 0$. This will be useful to us later.

Theorem 18 *Let $\epsilon > 0$, and $\hat{\mathbf{u}}_t := \hat{\mathbf{w}}_t/\epsilon$, for $(\hat{\mathbf{w}}_t)$ as in (9). Then, under Assumptions 1 and 2:*

$$\sum_{t=1}^T \langle \mathbf{g}_t, \hat{\mathbf{u}}_t - \mathbf{w} \rangle \leq \left[2\|\mathbf{w}\| \sqrt{V_T \ln_+ \left(\frac{2\epsilon\|\mathbf{w}\|V_T}{h_1^2} \right)} \right] \vee \left[4h_T\|\mathbf{w}\| \ln \left(\frac{4h_T\epsilon\|\mathbf{w}\|\sqrt{V_T}}{h_1^2} \right) \right] + \frac{h_1}{\epsilon},$$

for all $\mathbf{w} \in \mathcal{W} = \mathbb{R}^d, T \geq 1$.

Proof Since the assumptions of Theorem 4 are satisfied, we have

$$\Phi_T = \sum_{t=1}^T \hat{\mathbf{w}}_t^\top \mathbf{g}_t + \frac{h_1^2}{\sqrt{V_T}} \cdot \exp\left(\frac{\|\mathbf{G}_T\|^2}{2V_T + 2h_T\|\mathbf{G}_T\|}\right) \leq \Phi_0 = h_1, \quad (24)$$

Dividing both sides of (24) by $\epsilon > 0$ and rearranging yields

$$\sum_{t=1}^T \hat{\mathbf{u}}_t^\top \mathbf{g}_t \leq \frac{h_1}{\epsilon} - \Theta_T(\mathbf{G}), \quad \text{where } \Theta_T(\mathbf{G}) := \frac{h_1^2}{\epsilon\sqrt{V_T}} \cdot \exp\left(\frac{\|\mathbf{G}\|^2}{2V_T + 2h_T\|\mathbf{G}\|}\right), \mathbf{G} \in \mathbb{R}^d,$$

By duality, we further have that

$$\sum_{t=1}^T \hat{\mathbf{u}}_t^\top \mathbf{g}_t \leq \frac{h_1}{\epsilon} + \mathbf{w}^\top \mathbf{G}_T + \Theta_T^*(-\mathbf{w}), \quad \text{for all } \mathbf{w} \in \mathbb{R}^d. \quad (25)$$

Since $\Theta_T(\mathbf{G}) = \Psi_T(\mathbf{G})/\epsilon$, for all $\mathbf{G} \in \mathbb{R}^d$, where Ψ_T is the function defined in (10), we have by the properties of the Fenchel dual (Hiriart-Urruty and Lemaréchal, 2004, Prop. 1.3.1) that

$$\Theta_T^*(\mathbf{w}) = \Psi_T^*(\epsilon \mathbf{w})/\epsilon, \quad \text{for all } \mathbf{w} \in \mathbb{R}^d. \quad (26)$$

We now bound $\Psi_T^*(-\mathbf{w})$ from above, for $\mathbf{w} \in \mathbb{R}^d$. For this, note that $\Psi_T(\mathbf{G}) = \psi_T(\|\mathbf{G}\|/h_T)$, for $\mathbf{G} \in \mathbb{R}^d$, where

$$\psi_T(x) := \frac{h_1^2}{\sqrt{V_T}} \cdot \exp\left(\frac{x^2}{2V_T/h_T^2 + 2|x|}\right).$$

Thus, according to (McMahan and Orabona, 2014, Lemma 3) and the properties of duality (Hiriart-Urruty and Lemaréchal, 2004, Prop. E.1.3.1), we have

$$\Psi_T^*(-\mathbf{w}) = \psi_T^*(h_T \|\mathbf{w}\|). \quad (27)$$

On the other hand, (Cutkosky and Orabona, 2018, Lemma 18, 19) and (Orabona and Pál, 2016a, Lemma 18) provides the following upper-bound on $\psi_T^*(u)$, $u \in \mathbb{R}$, using the Lambert function W (where $W(x)$ is defined as the principal solution to $W(x)e^{W(x)} = x$):

$$\psi_T^*(u) \leq \Lambda_T(u) \vee \left(4u \cdot \ln\left(\frac{4u\sqrt{V_T}}{h_1^2}\right)\right), \quad (28)$$

$$\text{where } \Lambda_T(y) := y \sqrt{\frac{2V_T}{h_T}} \cdot \left((W(c_T^2 y^2))^{1/2} - (W(c_T^2 y^2))^{-1/2} \right), y \in \mathbb{R},$$

and $c_T := \sqrt{2}V_T/(h_T h_1^2)$. Using the fact that the Lambert function satisfies $(W(x))^{1/2} - (W(x))^{-1/2} \leq \sqrt{\ln_+ x}$, for all $x \geq 0$ (see Lemma 20), together with (28) and (27) implies that

$$\Psi_T^*(-\mathbf{w}) \leq \left[2\|\mathbf{w}\| \sqrt{V_T \ln_+ \left(\frac{2\|\mathbf{w}\|V_T}{h_1^2} \right)} \right] \vee \left[4h_T \|\mathbf{w}\| \ln \left(\frac{4h_T \|\mathbf{w}\| \sqrt{V_T}}{h_1^2} \right) \right],$$

for all $\mathbf{w} \in \mathbb{R}^d$. Combining this with (25) and (26) leads to the desired regret bound. ■

Proof of Theorem 5 Invoke Theorem 18 with $\epsilon = 1$. ■

B.2. Proof of Theorem 6

Proof Fix $\mathbf{w} \in \mathbb{R}^d$ and let $k \geq 1$ be the total number of epochs. We denote by $\tau_i \geq 1$ the start index of epoch $i \in [k]$. Further, for $\tau, \tau' \in \mathbb{N}$, we define $\tilde{\tau} := \tau - 1$ and $V_{\tau:\tau'} := h_\tau^2 + \sum_{s=\tau}^{\tau'} \|\mathbf{g}_s\|^2$ (note

how the upper index is exclusive). Recall that at epoch $i \in [k]$, the restart condition in Algorithm 1 is triggered at $t = \tau_{i+1} > \tau_i$ only if

$$\frac{h_t}{h_{\tau_i}} > \sum_{s=1}^{t-1} \frac{\|g_s\|}{h_s} + 2 \geq \sum_{s=1}^t \frac{\|g_s\|}{h_s}, \quad (29)$$

where the last inequality follows by Assumption 1. We note that (29) also implies that

$$h_{\tau_{i+1}} > 2h_{\tau_i}, \quad \text{for all } i \in [k]. \quad (30)$$

On the other hand, within epoch $i \in [k]$, $\frac{h_t}{h_{\tau_i}} \leq \sum_{s=1}^{t-1} \frac{\|g_s\|}{h_s} + 2$, for all $\tau_i \leq t \leq \tilde{\tau}_{i+1}$, and thus

$$\frac{h_{\tilde{\tau}_{i+1}}}{h_{\tau_i}} \leq \frac{\sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}} \leq \sqrt{\sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} \left(\sum_{s=1}^{t-1} \frac{\|g_s\|}{h_s} + 2 \right)^2} \leq \sqrt{\frac{b_T}{2}}, \quad (31)$$

where $b_T := 2 \sum_{t=1}^T (\sum_{s=1}^{t-1} \|g_s\|/h_s + 2)^2$. Therefore, by the regret bound of Theorem 5 and (31):

$$\sum_{s=\tau_i}^{\tilde{\tau}_{i+1}} \langle g_s, \hat{w}_s - w \rangle \leq 2\|w\| \sqrt{V_{\tau_i:\tau_{i+1}}} \ln_+(\|w\|b_T) + (4\|w\| \ln(2\|w\|b_T) + 1)h_{\tau_{i+1}}, \quad i \in [k]. \quad (32)$$

Summing this inequality over $i = 1, \dots, k-2$, we get:

$$\sum_{s=1}^{\tilde{\tau}_{k-1}} \langle g_s, \hat{w}_s - w \rangle \leq 2\|w\| \sum_{i=1}^{k-2} \sqrt{V_{\tau_i:\tau_{i+1}}} \ln_+(\|w\|b_T) + \sum_{i=1}^{k-2} (4\|w\| \ln(2\|w\|b_T) + 1)h_{\tau_{i+1}}. \quad (33)$$

Now using (29) at $t = \tau_{i+2}$, we have for all $i \in [k]$,

$$\begin{aligned} V_{\tau_i:\tau_{i+1}} &\leq h_{\tau_i}^2 + \sum_{s=1}^{\tilde{\tau}_{i+1}} \frac{\|g_s\|^2}{h_{\tau_{i+1}}^2} \cdot h_{\tau_{i+1}}^2 \leq h_{\tau_i}^2 + \sum_{s=1}^{\tilde{\tau}_{i+2}} \frac{\|g_s\|^2}{h_s^2} \cdot h_{\tau_{i+1}}^2, \\ &\stackrel{(*)}{\leq} h_{\tau_i}^2 + \sum_{s=1}^{\tilde{\tau}_{i+2}} \frac{\|g_s\|}{h_s} \cdot h_{\tau_{i+1}}^2, \\ &\stackrel{(29)}{\leq} h_{\tau_i}^2 + h_{\tau_{i+1}} h_{\tau_{i+2}} \leq 2h_{\tau_{i+2}}^2, \end{aligned} \quad (34)$$

where the inequality $(*)$ follows by Assumption 1. Now by (30), we also have

$$\sum_{i=1}^{k-2} h_{\tau_{i+1}} \leq \left(\sum_{i=1}^k \frac{1}{2^i} \right) h_{\tau_k} \leq h_{\tau_k}. \quad (35)$$

Thus, substituting (35) and (34) into (33), and using the fact that $h_{\tau_k} \leq h_T$, we get:

$$\begin{aligned} \sum_{s=1}^{\tilde{\tau}_{k-1}} \langle g_s, \hat{w}_s - w \rangle &\leq 4\|w\|h_T \sqrt{2 \ln_+(\|w\|b_T)} + h_T \cdot (4\|w\| \ln(2\|w\|b_T) + 1), \\ &\leq h_T \cdot (8\|w\| \ln_+(2\|w\|b_T) + 2\|w\| + 1), \end{aligned} \quad (36)$$

where in the last inequality, we used the fact that $\sqrt{2x} \leq x + 1/2$, for all $x \geq 0$. Now, summing (32) over the last two epochs, yields

$$\sum_{s=\tau_{k-1}}^T \langle \mathbf{g}_s, \hat{\mathbf{w}}_s - \mathbf{w} \rangle \leq 2\|\mathbf{w}\| \sqrt{2V_T \ln_+(\|\mathbf{w}\|b_T)} + 2h_T \cdot (4\|\mathbf{w}\| \ln_+(2\|\mathbf{w}\|b_T) + 1). \quad (37)$$

Adding (36) and (37) together leads to

$$\sum_{s=1}^T \langle \mathbf{g}_s, \hat{\mathbf{w}}_s - \mathbf{w} \rangle \leq 2\|\mathbf{w}\| \sqrt{2V_T \ln_+(\|\mathbf{w}\|b_T)} + (16\|\mathbf{w}\| \ln_+(2\|\mathbf{w}\|b_T) + 2\|\mathbf{w}\| + 3)h_T.$$

This concludes the proof. ■

Appendix C. Proofs for Section 3.2

In this section we work on a version of the potential function that does not have the tuning for Section 3.2 substituted in yet, so that we can prove the result necessary for Section 5 in one go. The potential is parameterized by a *prior variance* $\sigma^2 > 0$, *initial wealth* $\epsilon > 0$ and, as before *prod factor* $\gamma > 1$. It is defined by

$$\Psi(\mathbf{G}, \mathbf{V}, h) := \frac{\epsilon \exp \left(\inf_{\lambda \geq 0} \left\{ \frac{1}{2} \mathbf{G}^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V} + \lambda \mathbf{I})^{-1} \mathbf{G} + \frac{\lambda \rho(\gamma)^2}{2h^2} \right\} \right)}{\sqrt{\det(\mathbf{I} + \sigma^2 \gamma \mathbf{V})}}, \quad (38)$$

C.1. Proof of Lemma 7

We prove the claim in Lemma 7 for the more general potential (38). Let $\lambda_\star \geq 0$ be the minimizer in the problem $\Psi(\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t)$. With that notation, we see that $\hat{\mathbf{w}}_t = -\Psi(\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t) \cdot (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V} + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_{t-1}$. To prove the lemma, it suffices to prove the stronger statement obtained by picking the sub-optimal choice $\lambda = \lambda_\star$ for the problem $\Psi(\mathbf{G}_t, \mathbf{V}_t, h_t)$, and dividing by $\Psi(\mathbf{G}_{t-1}, \mathbf{V}_{t-1}, h_t) > 0$, *i.e.*

$$\begin{aligned} & -\mathbf{g}_t \cdot (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_{t-1} + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_{t-1} \\ & \leq 1 - \frac{\exp \left(\frac{1}{2} \mathbf{G}_t^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_t + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_t + \frac{\lambda_\star \rho(\gamma)^2}{2h_t^2} - \frac{1}{2} \ln \det(\mathbf{I} + \sigma^2 \gamma \mathbf{V}_t) \right)}{\exp \left(\frac{1}{2} \mathbf{G}_{t-1}^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_{t-1} + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_{t-1} + \frac{\lambda_\star \rho(\gamma)^2}{2h_t^2} - \frac{1}{2} \ln \det(\mathbf{I} + \sigma^2 \gamma \mathbf{V}_{t-1}) \right)}. \end{aligned}$$

Let us abbreviate $\Sigma^{-1} = \sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_{t-1} + \lambda_\star \mathbf{I}$. The matrix determinant lemma and monotonicity of matrix inverse give

$$\ln \frac{\det(\mathbf{I} + \sigma^2 \gamma \mathbf{V}_t)}{\det(\mathbf{I} + \sigma^2 \gamma \mathbf{V}_{t-1})} = \ln \left(1 + \gamma \mathbf{g}_t^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_{t-1})^{-1} \mathbf{g}_t \right) \geq \ln(1 + \gamma \mathbf{g}_t^\top \Sigma \mathbf{g}_t).$$

Then Sherman-Morrison gives

$$\mathbf{G}_t^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_t + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_t = \mathbf{G}_t^\top \Sigma \mathbf{G}_t - \gamma \frac{(\mathbf{g}_t^\top \Sigma \mathbf{G}_t)^2}{1 + \gamma \mathbf{g}_t^\top \Sigma \mathbf{g}_t}$$

and splitting off the last round $\mathbf{G}_t = \mathbf{G}_{t-1} + \mathbf{g}_t$ gives

$$\mathbf{G}_t^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_t + \lambda_\star \mathbf{I})^{-1} \mathbf{G}_t = \mathbf{G}_{t-1}^\top \Sigma \mathbf{G}_{t-1} + \frac{2\mathbf{G}_{t-1}^\top \Sigma \mathbf{g}_t + \mathbf{g}_t^\top \Sigma \mathbf{g}_t - \gamma(\mathbf{g}_t^\top \Sigma \mathbf{G}_{t-1})^2}{1 + \gamma \mathbf{g}_t^\top \Sigma \mathbf{g}_t}.$$

All in all, it suffices to show

$$-\mathbf{g}_t^\top \Sigma \mathbf{G}_{t-1} \leq 1 - \exp \left(\frac{2\mathbf{G}_{t-1}^\top \Sigma \mathbf{g}_t + \mathbf{g}_t^\top \Sigma \mathbf{g}_t - \gamma(\mathbf{g}_t^\top \Sigma \mathbf{G}_{t-1})^2}{2(1 + \gamma \mathbf{g}_t^\top \Sigma \mathbf{g}_t)} - \frac{1}{2} \ln(1 + \gamma \mathbf{g}_t^\top \Sigma \mathbf{g}_t) \right).$$

Introducing scalars $r = \mathbf{g}_t^\top \Sigma \mathbf{G}_{t-1}$ and $z = \mathbf{g}_t^\top \Sigma \mathbf{g}_t$, this simplifies to

$$-r \leq 1 - \exp \left(\frac{2r + z - \gamma r^2}{2(1 + \gamma z)} - \frac{1}{2} \ln(1 + \gamma z) \right)$$

Being a square, $z \geq 0$ is positive. In addition, optimality of λ_\star ensures that $\|\Sigma \mathbf{G}_{t-1}\| = \frac{\rho(\gamma)}{h_t}$; this follows from the fact that $\frac{d}{d\lambda} \mathbf{G}_{t-1}^\top (\sigma^{-2} \mathbf{I} + \gamma \mathbf{V} + \lambda \mathbf{I})^{-1} \mathbf{G}_{t-1} \big|_{\lambda=\lambda_\star} = \|\Sigma \mathbf{G}_{t-1}\|^2$. In combination with $\|\mathbf{g}_t\| \leq h_t$, we find $|r| \leq \rho(\gamma) \leq 1$. The above requirement may hence be further reorganized to

$$2r - \gamma r^2 \leq -z + (1 + \gamma z) (\ln(1 + \gamma z) + 2 \ln(1 + r)).$$

The convex right hand side is minimized subject to $z \geq 0$ at

$$z = \max \left\{ 0, \frac{e^{\frac{1}{\gamma} - 1 - 2 \ln(1+r)} - 1}{\gamma} \right\}$$

so it remains to show

$$2r - \gamma r^2 \leq \begin{cases} \frac{1}{\gamma} - (1+r)^{-2} e^{\frac{1}{\gamma} - 1}, & \text{if } \frac{1}{\gamma} - 1 \geq 2 \ln(1+r); \\ 2 \ln(1+r), & \text{otherwise.} \end{cases}$$

The function ρ in (12) is designed to satisfy the hardest case, where $r = -\rho(\gamma)$, with equality.

C.2. Proof of Theorem 8

We restate the claim for the potential (38) before tuning:

Theorem 19 (Theorem 8 rephrased) *Let $\Sigma_T^{-1} := \sigma^{-2} \mathbf{I} + \gamma \mathbf{V}_T$. For $(\hat{\mathbf{w}}_t)$ as in (14), we have*

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \leq \epsilon + \sqrt{Q_T^{\mathbf{w}} \ln_+ \left(\frac{\det(\sigma^2 \Sigma_T^{-1})}{\epsilon^2} Q_T^{\mathbf{w}} \right)}, \quad \text{for all } \mathbf{w} \in \mathbb{R}^d, \text{ where}$$

$$Q_T^{\mathbf{w}} := \max \left\{ \mathbf{w}^\top \Sigma_T^{-1} \mathbf{w}, \frac{1}{2} \left(\frac{h_T^2 \|\mathbf{w}\|^2}{\rho(\gamma)^2} \ln \left(\frac{\det(\sigma^2 \Sigma_T^{-1})}{\epsilon^2} \frac{h_T^2 \|\mathbf{w}\|^2}{\rho(\gamma)^2} \right) + \mathbf{w}^\top \Sigma_T^{-1} \mathbf{w} \right) \right\}.$$

Using that $\Psi(\mathbf{G}, \mathbf{V}, h)$ is decreasing in h , we can telescope to obtain

$$\sum_{t=1}^T \mathbf{g}_t^\top \hat{\mathbf{w}}_t \leq \Psi(\mathbf{0}, \mathbf{0}, h_1) - \Psi(\mathbf{G}_T, \mathbf{V}_T, h_T)$$

Using the definition reveals $\Psi(\mathbf{0}, \mathbf{0}, h_1) = \epsilon$, yielding

$$\sum_{t=1}^T \mathbf{g}_t^\top \hat{\mathbf{w}}_t \leq \epsilon - \frac{\epsilon \exp \left(\inf_{\lambda \geq 0} \frac{1}{2} \mathbf{G}_T^\top (\boldsymbol{\Sigma}_T^{-1} + \lambda \mathbf{I})^{-1} \mathbf{G}_T + \frac{\lambda \rho(\gamma)^2}{2h_T^2} \right)}{\sqrt{\det(\sigma^2 \boldsymbol{\Sigma}_T^{-1})}}. \quad (39)$$

To transform this into a regret bound, it remains to compute the convex conjugate of the RHS of (39) in \mathbf{G}_T . To this end, let

$$f(\mathbf{G}) = \exp \left(\inf_{\lambda \geq 0} \frac{1}{2} \mathbf{G}^\top (\mathbf{Q} + \lambda \mathbf{I})^{-1} \mathbf{G} + \frac{\lambda Z}{2} \right).$$

The Fenchel dual of this function is

$$\begin{aligned} f^*(\mathbf{w}) &= \sup_{\mathbf{G}} \mathbf{w}^\top \mathbf{G} - \exp \left(\inf_{\lambda \geq 0} \frac{1}{2} \mathbf{G}^\top (\mathbf{Q} + \lambda \mathbf{I})^{-1} \mathbf{G} + \frac{\lambda Z}{2} \right) \\ &= \sup_{\mathbf{G}, \lambda \geq 0} \mathbf{w}^\top \mathbf{G} - \exp \left(\frac{1}{2} \mathbf{G}^\top (\mathbf{Q} + \lambda \mathbf{I})^{-1} \mathbf{G} + \frac{\lambda Z}{2} \right) \\ &= \sup_{\alpha, \lambda \geq 0} \alpha \mathbf{w}^\top (\mathbf{Q} + \lambda \mathbf{I}) \mathbf{w} - \exp \left(\frac{\alpha^2}{2} \mathbf{w}^\top (\mathbf{Q} + \lambda \mathbf{I}) \mathbf{w} + \frac{\lambda Z}{2} \right) \\ &= \sup_{\lambda \geq 0} \sqrt{\mathbf{w}^\top (\mathbf{Q} + \lambda \mathbf{I}) \mathbf{w}} X \left(\mathbf{w}^\top (\mathbf{Q} + \lambda \mathbf{I}) \mathbf{w} e^{-\lambda Z} \right), \end{aligned}$$

where the model complexity is measured for $\theta \geq 0$ through the function $X(\theta) := \sup_{\alpha} \alpha - e^{\frac{\alpha^2}{2} - \frac{1}{2} \ln \theta}$. One can write $X(\theta) = W(\theta)^{1/2} - W(\theta)^{-1/2}$ in terms of the Lambert function W (where $W(x)$ is defined as the principal solution to $W(x)e^{W(x)} = x$). We will further use that $X(\theta)$ is increasing, and that it satisfies $X(\theta) \leq \sqrt{\ln_+ \theta}$ (see Lemma 20). Zero derivative of the above objective for λ occurs at the pleasantly explicit

$$\lambda = \frac{\ln \frac{\|\mathbf{w}\|^2}{Z}}{2Z} - \frac{\mathbf{w}^\top \mathbf{Q} \mathbf{w}}{2\|\mathbf{w}\|^2},$$

and hence the optimum for λ is either at that point or at zero, whichever is higher, with the crossover point at $\frac{\|\mathbf{w}\|^2}{Z} \ln \frac{\|\mathbf{w}\|^2}{Z} = \mathbf{w}^\top \mathbf{Q} \mathbf{w}$. Plugging that in, we find that

$$f^*(\mathbf{w}) = \begin{cases} \sqrt{\frac{1}{2} (C + \mathbf{w}^\top \mathbf{Q} \mathbf{w})} X \left(\frac{1}{2} (C + \mathbf{w}^\top \mathbf{Q} \mathbf{w}) e^{-\frac{\ln \frac{\|\mathbf{w}\|^2}{Z}}{2} + \frac{Z \mathbf{w}^\top \mathbf{Q} \mathbf{w}}{2\|\mathbf{w}\|^2}} \right), & \text{if } C \geq \mathbf{w}^\top \mathbf{Q} \mathbf{w}; \\ \sqrt{\mathbf{w}^\top \mathbf{Q} \mathbf{w}} X(\mathbf{w}^\top \mathbf{Q} \mathbf{w}), & \text{otherwise,} \end{cases}$$

where $C := \frac{\|\mathbf{w}\|^2}{Z} \ln \frac{\|\mathbf{w}\|^2}{Z}$. Using that $X(\theta)$ is increasing, we may drop the exponential in its argument in the first case, and obtain

$$f^*(\mathbf{w}) \leq \sqrt{Q_T^w} X(Q_T^w) \quad \text{where} \quad Q_T^w := \max \left\{ \mathbf{w}^\top \mathbf{Q} \mathbf{w}, \frac{1}{2} \left(\frac{\|\mathbf{w}\|^2}{Z} \ln \frac{\|\mathbf{w}\|^2}{Z} + \mathbf{w}^\top \mathbf{Q} \mathbf{w} \right) \right\}.$$

Note that this is a curious maximum between $\mathbf{w}^\top \mathbf{Q} \mathbf{w}$ (the larger for modest \mathbf{w}), and the *average* between that very same term and another quantity that grows super-linearly with $\|\mathbf{w}\|^2$ (so this is the winner for extreme \mathbf{w}).

Okay, now let's collect everything for the final result and undo the abbreviations. We have

$$\begin{aligned} \sum_{t=1}^T \mathbf{g}_t^\top \hat{\mathbf{w}}_t &\leq \epsilon + \inf_{\mathbf{w}} \mathbf{w}^\top \mathbf{G}_T + \frac{\epsilon}{\sqrt{\det(\sigma^2 \Sigma_T^{-1})}} f^* \left(-\frac{\sqrt{\det(\sigma^2 \Sigma_T^{-1})}}{\epsilon} \mathbf{w} \right), \\ &\leq \epsilon + \inf_{\mathbf{w}} \mathbf{w}^\top \mathbf{G}_T + \sqrt{Q_T^w} X \left(\frac{\det(\sigma^2 \Sigma_T^{-1})}{\epsilon^2} Q_T^w \right), \end{aligned}$$

where

$$Q_T^w := \max \left\{ \mathbf{w}^\top \Sigma_T^{-1} \mathbf{w}, \frac{1}{2} \left(\frac{h_T^2 \|\mathbf{w}\|^2}{\rho(\gamma)^2} \ln \left(\frac{\det(\sigma^2 \Sigma_T^{-1})}{\epsilon^2} \frac{h_T^2 \|\mathbf{w}\|^2}{\rho(\gamma)^2} \right) + \mathbf{w}^\top \Sigma_T^{-1} \mathbf{w} \right) \right\}.$$

To complete the proof of Theorem 19, it remains to prove the following result.

Lemma 20 *For $\theta \geq 0$, define $X(\theta) := \sup_{\alpha} \alpha - e^{\frac{\alpha^2}{2} - \frac{1}{2} \ln \theta}$. Then $X(\theta) = (W(\theta))^{1/2} - (W(\theta))^{-1/2} = \sqrt{\ln \theta} + o(1)$.*

Proof The fact that $X(\theta) = (W(\theta))^{1/2} - (W(\theta))^{-1/2}$ follows from (Orabona and Pál, 2016a, Lemma 18). Recall that

$$\sup_x yx - e^x = y \ln y - y$$

Hence

$$\begin{aligned} X(\theta) &= \sup_{\alpha} \alpha - e^{\frac{\alpha^2}{2} - \frac{1}{2} \ln \theta} \\ &= \sup_{\alpha} \inf_{\eta} \alpha - \eta \left(\frac{\alpha^2}{2} - \frac{1}{2} \ln \theta \right) + \eta \ln \eta - \eta \\ &= \inf_{\eta} \frac{1}{2\eta} + \frac{\eta}{2} \ln \theta + \eta \ln \eta - \eta \\ &\leq \min \left\{ \sqrt{\ln \theta} - \frac{1 + \frac{1}{2} \ln \ln \theta}{\sqrt{\ln \theta}}, \frac{\sqrt{\theta}}{2} - \frac{1}{\sqrt{\theta}} \right\} \\ &\leq \sqrt{\ln_+ \theta} \end{aligned}$$

where we plugged in the sub-optimal choices $\eta = \frac{1}{\sqrt{\ln \theta}}$ (this requires $\theta \geq 1$) and $\eta = \frac{1}{\sqrt{\theta}}$. When we stick in $\eta = \frac{1}{\sqrt{\ln(e^{e^{-2}} + \theta)}}$ we find

$$X(\theta) \leq \frac{\ln(e^{e^{-2}} + \theta) + \ln \theta - \ln(\ln(e^{e^{-2}} + \theta)) - 2}{2\sqrt{\ln(e^{e^{-2}} + \theta)}} \leq \sqrt{\ln(e^{e^{-2}} + \theta)}$$

Note that $e^{e^{-2}} = 1.14492$. This is less than 2, the value of θ where $\sqrt{\theta}/2 - 1/\sqrt{\theta}$ becomes positive. ■

Appendix D. Proofs for Section 4

D.1. Proof of Lemma 11

Let $c, b, \beta \geq 0$, $\nu \geq 1$, $\alpha \in]1, 2]$, and $\gamma \in]-1, -\alpha^{-1}[$. We consider the 1-dimensional case (*i.e.* $d = 1$) and set $g_t = t^\gamma$, for all $t \geq 1$. Since $-1 < \gamma < -1/\alpha$, we have $L_t = L_1 = 1$, for all $t \geq 1$, and so the sequence $(\sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha})$ is increasing. Further, there exists $p, q > 0$ such that,

$$\forall t \geq 1, \quad p\sqrt{\ln t} \leq \sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha} = \sqrt{\ln t \sum_{s=1}^t s^{\alpha\gamma}} \leq q\sqrt{\ln t}. \quad (40)$$

Thus, given any sequence $(\hat{w}_t) \in \mathbb{R}$ satisfying

$$|\hat{w}_t| \leq b\sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha}, \quad t \geq 1,$$

we have, for $T \geq 1$ and $w = -2b\sqrt{V_{\alpha,T} \ln(T)/L_T^\alpha}$,

$$\begin{aligned} \sum_{t=1}^T g_t \cdot (\hat{w}_t - w) &\geq b\sqrt{V_{\alpha,T} \ln(T)/L_T^\alpha} \cdot \sum_{t=1}^T g_t, \\ &\stackrel{(40)}{\geq} bp\sqrt{\ln T} \cdot \sum_{t=1}^T t^\gamma, \\ &\geq \frac{bp\sqrt{\ln T}}{\gamma + 1} \cdot ((T + 1)^{\gamma+1} - 1). \end{aligned} \quad (41)$$

Now by the choice of w and (40), we have $|w| \leq 2bq\sqrt{\ln T}$, and so by (41),

$$\sum_{t=1}^T g_t \cdot (\hat{w}_t - w) \geq L_T |w|^\nu \cdot \frac{p \cdot ((T + 1)^{\gamma+1} - 1)}{(\gamma + 1)(2q)^\nu b^{\nu-1} (\ln T)^{\nu/2-1/2}}. \quad (42)$$

Using again the fact that $|w| \leq 2bq\sqrt{\ln T}$ and (40), we have $L_T^{1-\alpha/2}(|w| + 1)\sqrt{V_{\alpha,T} \ln T} \leq 2bq^2 \ln T + q\sqrt{\ln T}$, and so due to (41), we have

$$\sum_{t=1}^T g_t \cdot (\hat{w}_t - w) \geq L_T^{1-\alpha/2}(|w| + 1)\sqrt{V_{\alpha,T} \ln T} \cdot \frac{(T + 1)^{\gamma+1} - 1}{(\gamma + 1) \left(\frac{2q^2}{p} \sqrt{\ln T} + \frac{q}{bp} \right)}. \quad (43)$$

Since $\gamma > -1$, there exists $T \geq 1$ such that

$$2c \cdot \ln(1 + |w|T)^\beta \leq \min \left(\frac{(T + 1)^{\gamma+1} - 1}{(\gamma + 1) \left(\frac{2q^2}{p} \sqrt{\ln T} + \frac{q}{bp} \right)}, \frac{p \cdot ((T + 1)^{\gamma+1} - 1)}{(\gamma + 1)(2q)^\nu b^{\nu-1} (\ln T)^{\nu/2-1/2}} \right),$$

and so for such a choice of T , (42) and (43) imply the desired result.

D.2. Proof of Theorem 13

We need the following lemma in the proof of Theorem 13:

Lemma 21 *For all $b, c, \beta \geq 0$ and $\nu \in [1, 3[$, there exists $(\mathbf{g}_t) \in \mathbb{R}^d$, $T \geq 1$, and $\mathbf{w} \in \mathbb{R}^d$, such that for any sequence $(\hat{\mathbf{w}}_t)$ satisfying $\|\hat{\mathbf{w}}_t\| \leq b \cdot \sqrt{t \ln t}$, for all $t \geq 1$, we have*

$$\sum_{t=1}^T \langle \hat{\mathbf{w}}_t - \mathbf{w}, \mathbf{g}_t \rangle \geq c \cdot \ln(1 + \|\mathbf{w}\|T)^\beta \cdot (L_T \|\mathbf{w}\|^\nu + L_T(\|\mathbf{w}\| + 1)\sqrt{T \ln T}).$$

Proof Let $c, b, \beta \geq 0$, $\nu \in [1, 3[$, and $\alpha \in]1, 2]$. We consider the 1-dimensional case (*i.e.* $d = 1$) and set $g_t = 1$, for all $t \geq 1$. In this case, we have $L_t = 1$, for all $t \geq 1$. Given any sequence $(\hat{w}_t) \in \mathbb{R}$ satisfying

$$|\hat{w}_t| \leq b\sqrt{t \ln t}, \quad t \geq 1, \quad (44)$$

we have, for $T \geq 1$ and $w = -2b\sqrt{T \ln T}$,

$$\begin{aligned} \sum_{t=1}^T g_t \cdot (\hat{w}_t - w) &\stackrel{(44)}{\geq} b\sqrt{T \ln T} \cdot \sum_{t=1}^T g_t, \\ &= b\sqrt{T \ln T} \cdot T. \end{aligned} \quad (45)$$

Now since $|w| = 2b\sqrt{T \ln T}$, we have, by (45),

$$\sum_{t=1}^T g_t \cdot (\hat{w}_t - w) \geq L_T |w|^\nu \cdot \frac{T^{3/2-\nu/2}}{2^\nu b^{\nu-1} (\ln T)^{\nu/2-1/2}}. \quad (46)$$

Using again the fact that $|w| = 2b\sqrt{T \ln T}$ and $L_T = 1$, we have $L_T(|w| + 1)\sqrt{T \ln T} = 2bT \ln T + \sqrt{T \ln T}$, and so due to (45),

$$\sum_{t=1}^T g_t \cdot (\hat{w}_t - w) \geq L_T(|w| + 1)\sqrt{T \ln T} \cdot \frac{T}{2\sqrt{T \ln T} + 1/b}. \quad (47)$$

Since $\nu \in [1, 3[$, there exists $T \geq 1$ such that

$$2c \cdot \ln(1 + |w|T)^\beta \leq \min \left(\frac{T}{2\sqrt{T \ln T} + 1/b}, \frac{T^{3/2-\nu/2}}{2^\nu b^{\nu-1} (\ln T)^{\nu/2-1/2}} \right),$$

and so for such a choice of T , (46) and (47) imply the desired result. \blacksquare

Proof of Theorem 13 By Lemma 10, the only candidate algorithms are the ones whose outputs $(\hat{\mathbf{w}}_t)$ satisfy $\|\hat{\mathbf{w}}_t\| \leq b\sqrt{t \ln t}$, for all $t \geq 1$, for some constant $b > 0$. By Lemma 21, no such algorithms can achieve the desired regret bound. \blacksquare

Appendix E. Proof of Section 5

Proof of Theorem 14 The proof is similar to that of Theorem 6 except for some changes to account for the fact that the modified FREERANGE wrapper scales the outputs of FREEGRAD.

First, let us review some notation. Let $k \geq 1$ be the total number of epochs and denote by $\tau_i \geq 1$ the start index of epoch $i \in [k]$. Further, for $\tau, \tau' \in \mathbb{N}$, we define $\tilde{\tau} := \tau - 1$, $V_{\tau:\tau'} := |x_\tau|^2 + \sum_{s=\tau}^{\tilde{\tau}'} |g_s|^2$, and $B_\tau := \sum_{s=1}^\tau |x_s|/h_s$. In what follows, let $w \in \mathbb{R}$ be fixed.

Let (\hat{u}_t) be the outputs of algorithm 2 and $i \in [k]$. In this case, we have $\hat{u}_t = \hat{w}_t/(h_{\tau_i} B_{\tau_i})$, for all $t \in \{\tau_i, \dots, \tilde{\tau}_{i+1}\}$, and by Theorem 18, with $d = 1$, $\epsilon = h_{\tau_i} B_{\tau_i}$, and $g_t \in x_t \cdot \partial^{(0,1)} \ell(y_t, x_t \hat{u}_t)$:

$$\begin{aligned} \sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} g_t \cdot (\hat{u}_t - w) &\leq 2|w| \sqrt{V_{\tau_i:\tau_{i+1}}} \ln_+ \left(\frac{2h_{\tau_i} B_{\tau_i} |w| V_{\tau_i:\tau_{i+1}}}{h_{\tau_i}^2} \right) \\ &\quad + 4h_{\tau_{i+1}} |w| \ln \left(\frac{4h_{\tilde{\tau}_{i+1}} h_{\tau_i} B_{\tau_i} |w| \sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}^2} \right) + \frac{h_{\tau_i}}{h_{\tau_i} B_{\tau_i}}, \\ &\leq 2|w| \sqrt{V_{\tau_i:\tau_{i+1}}} \ln_+ \left(\frac{2B_{\tau_i} |w| V_{\tau_i:\tau_{i+1}}}{h_{\tau_i}} \right) \\ &\quad + 4h_{\tau_{i+1}} |w| \ln \left(\frac{4h_{\tilde{\tau}_{i+1}} B_{\tau_i} |w| \sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}} \right) + \frac{1}{B_{\tau_i}}, \end{aligned} \quad (48)$$

Recall that at epoch $i \in [k]$, the restart condition in Algorithm 2 is triggered at $t = \tau_{i+1} \geq \tau_i$ only if

$$\frac{h_t}{h_{\tau_i}} \stackrel{(*)}{>} \sum_{s=1}^{t-1} \frac{|x_s|}{h_s} + 1 = \sum_{s=1}^t \frac{|x_s|}{h_s}, \quad (49)$$

where the equality follows by the fact that when $(*)$ is satisfied for the first time, it must hold that $|x_t| = h_t$ (recall that the hints (h_t) satisfy (17)); in fact, we have,

$$h_{\tau_i} = |x_{\tau_i}|, \quad \text{for all } i \in [k]. \quad (50)$$

From (49), we get that

$$\frac{h_{\tilde{\tau}_{i+1}}}{h_{\tau_i}} \leq \frac{\sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}} \leq \sqrt{\sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} \left(\sum_{s=1}^t \frac{|x_s|}{h_s} \right)^2} \leq \sqrt{\frac{b_T}{2}}, \quad (51)$$

where $b_T := 2 \sum_{t=1}^T (\sum_{s=1}^t |x_s|/h_s)^2$. Plugging (51) into (48), and letting $c_T := B_T^2 b_T$, we get:

$$\begin{aligned} \sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} g_t \cdot (\hat{u}_t - w) &\leq 2|w| \sqrt{V_{\tau_i:\tau_{i+1}}} \ln_+ (|w| \sqrt{2V_T c_T}) \\ &\quad + 4h_{\tau_{i+1}} |w| \ln \left(2|w| \sqrt{2V_T c_T} \right) + \frac{1}{B_{\tau_i}}, \end{aligned} \quad (52)$$

Summing this inequality over $i = 1, \dots, k-2$, we get:

$$\begin{aligned}
\sum_{t=1}^{\tilde{\tau}_{k-1}} g_t \cdot (\hat{u}_t - w) &\leq 2|w| \sqrt{k \sum_{i=1}^{k-2} V_{\tau_i: \tau_{i+1}} \ln_+ \left(|w| \sqrt{2V_T c_T} \right)} \\
&\quad + \sum_{i=1}^{k-2} 4h_{\tau_{i+1}} |w| \ln \left(2|w| \sqrt{2V_T c_T} \right) + \sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}}, \\
&\leq 2|w| \sqrt{k \cdot \left(\sum_{s=1}^{\tau_{k-1}} |x_s|^2 + \sum_{i=1}^{k-2} h_{\tau_i}^2 \right) \ln_+ \left(|w| \sqrt{2V_T c_T} \right)} \\
&\quad + \sum_{i=1}^{k-2} 4h_{\tau_{i+1}} |w| \ln \left(2|w| \sqrt{2V_T c_T} \right) + \sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}}, \\
&\stackrel{(50)}{\leq} 2|w| \sqrt{2k \sum_{s=1}^{\tau_{k-1}} |x_s|^2 \ln_+ \left(|w| \sqrt{2V_T c_T} \right)} \\
&\quad + \sum_{i=1}^{k-2} 4h_{\tau_{i+1}} |w| \ln \left(2|w| \sqrt{2V_T c_T} \right) + \sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}}. \tag{53}
\end{aligned}$$

Using (49) again, we get that

$$\begin{aligned}
\sum_{s=1}^{\tau_{k-1}} |x_s|^2 &= \sum_{s=1}^{\tau_{k-1}} \frac{|x_s|^2}{h_{\tau_{k-1}}^2} \cdot h_{\tau_{k-1}}^2 \leq \sum_{s=1}^{\tau_k} \frac{|x_s|^2}{h_s^2} \cdot h_{\tau_{k-1}}^2 \leq \sum_{s=1}^{\tau_k} \frac{|x_s|}{h_s} \cdot h_{\tau_{k-1}}^2 \stackrel{(*)}{\leq} \left(\sum_{s=1}^{\tau_k} \frac{|x_s|}{h_s} \right)^2 \cdot \frac{h_{\tau_{k-1}}^2}{k}, \\
&\stackrel{(49)}{\leq} \frac{h_{\tau_k}^2}{k}, \tag{54}
\end{aligned}$$

where the inequality $(*)$ follows by the fact that $\sum_{s=1}^{\tau_k} |x_s|/h_s \geq k$ due to (50). We also have

$$\sum_{i=1}^{k-2} h_{\tau_{i+1}} \leq \sum_{s=1}^{\tau_{k-1}} |x_s| = \sum_{s=1}^{\tau_{k-1}} \frac{|x_s|}{h_{\tau_{k-1}}} \cdot h_{\tau_{k-1}} \leq \sum_{s=1}^{\tau_k} \frac{|x_s|}{h_s} \cdot h_{\tau_{k-1}} \stackrel{(49)}{\leq} h_{\tau_k}. \tag{55}$$

$$\sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}} \stackrel{(50)}{=} \sum_{i=1}^{k-2} \frac{|x_{\tau_i}|}{B_{\tau_i}} = \sum_{i=1}^{k-2} \frac{|x_{\tau_i}|}{\sum_{s=1}^{\tau_i} |x_s|} \leq \sum_{t=1}^{\tau_{k-2}} \frac{|x_t|}{\sum_{s=1}^t |x_s|} \leq \ln B_T. \tag{56}$$

Thus, substituting (56), (55), and (54) into (53), and using the fact that $h_{\tau_k} \leq h_T$, we get:

$$\begin{aligned}
\sum_{t=1}^{\tilde{\tau}_{k-1}} g_t \cdot (\hat{u}_t - w) &\leq 2|w| h_T \sqrt{2 \ln_+ \left(|w| \sqrt{2V_T c_T} \right)} \\
&\quad + 4h_T |w| \ln \left(2|w| \sqrt{2V_T c_T} \right) + \ln B_T, \\
&\leq h_T |w| \cdot (6 \ln_+ (2|w| \sqrt{2V_T c_T}) + 1) + \ln B_T, \tag{57}
\end{aligned}$$

where in the last inequality, we used the fact that $\sqrt{2x} \leq x + 1/2$, for all $x \geq 0$. Now, summing (52) over the last two epochs, yields

$$\sum_{t=\tau_{k-1}}^T g_t \cdot (\hat{u}_t - w) \leq 2|w| \sqrt{2V_T \ln_+(|w| \sqrt{2V_T c_T})} + 8h_T |w| \ln(2|w| \sqrt{2V_T c_T}) + 2. \quad (58)$$

Adding (57) and (58) together implies the desired result. ■