Implicit Parameter-free Online Learning with Truncated Linear Models

Keyi Chen   KEYICHEN@BU.EDU
Department of Computer Science, Boston University, Boston, MA, USA
Ashok Cutkosky   ASHOK@CUTKOSKY.COM
Francesco Orabona   FRANCESCO@ORABONA.COM
Department of Electrical & Computer Engineering, Boston University, Boston, MA, USA

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

Parameter-free algorithms are online learning algorithms that do not require setting learning rates. They achieve optimal regret with respect to the distance between the initial point and *any* competitor. Yet, parameter-free algorithms do not take into account the geometry of the losses. Recently, in the stochastic optimization literature, it has been proposed to instead use *truncated* linear lower bounds, which produce better performance by more closely modeling the losses. In particular, truncated linear models greatly reduce the problem of overshooting the minimum of the loss function. Unfortunately, truncated linear models cannot be used with parameter-free algorithms because the updates become very expensive to compute. In this paper, we propose new parameter-free algorithms that can take advantage of truncated linear models through a new update that has an "implicit" flavor. Based on a *novel decomposition of the regret*, the new update is efficient, requires only one gradient at each step, never overshoots the minimum of the truncated model, and retains the favorable parameter-free properties. We also conduct an empirical study demonstrating the practical utility of our algorithms.

Keywords: Online convex optimization, Regret, Truncated linear models, Parameter-free

1. Introduction

In this paper, we study Online Convex Optimization (OCO) (Gordon, 1999; Zinkevich, 2003). In this setting, for each of T steps, a learner produces a prediction $\boldsymbol{w}_t \in V$ in each step t, where $V \subseteq \mathbb{R}^d$ is the feasible convex set. After each prediction, an adversary reveals a convex loss function $\ell_t : V \to \mathbb{R}$ and the learner pays $\ell_t(\boldsymbol{w}_t)$. The aim of the learner is to minimize its *regret* with respect to any fixed prediction $\boldsymbol{u} \in V$, defined as

$$\text{Regret}_T(\boldsymbol{u}) \triangleq \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \ .$$

Depending on the assumptions on the feasible set and the losses, there are many OCO algorithms that achieve optimal regret. The two main families of OCO algorithms are based on Online Mirror Descent (OMD) (Nemirovsky and Yudin, 1983; Warmuth and Jagota, 1997) and Follow-The-Regularized-Leader (FTRL) (Shalev-Shwartz, 2007; Abernethy et al., 2008b; Hazan and Kale, 2008). For the particular case where $V \equiv \mathbb{R}^d$, *parameter-free* algorithms are minimax optimal (Orabona and Pál, 2016; Cutkosky and Orabona, 2018). The key feature of parameter-free algorithms is that the $\text{Regret}_T(\boldsymbol{u})$ scales nearly linearly in $\|\boldsymbol{u}\|$, and is *constant* for $\boldsymbol{u} = 0$. This guarantee is only obtainable by popular strategies like Online Subgradient Descent (Zinkevich, 2003) if the learning

rate is carefully tuned to the (unknown!) value of $\|\boldsymbol{u}\|$. This lack of tuning learning rates motivates the name "parameter-free". Yet, even this favorable minimax optimality might not be satisfactory.

In particular, most OCO algorithms simply approximate the losses using linear functions, ignoring their geometry. This approach is justified by the fact that the worst-case losses are indeed just linear ones. However, in the extremely common case that the losses are not actually linear, that is they are not worst-case, the algorithm is wasting potentially useful information. More generally, too much focus on worst-case analyses and asymptotic rates can prevent the design of better algorithms.

In an effort to address this issue and go beyond focusing only on asymptotic rates, Asi and Duchi (2019) have proposed the use of *truncated linear models* instead of linear models to obtain better stochastic Mirror Descent algorithms with negligible additional computational complexity. Truncated linear models are tighter lower bounds to the original function that do not require additional curvature while still yielding a closed form update for Mirror Descent algorithms. For example, a common issue with standard gradient descent methods is that they can *overshoot* the minimum of the loss during any given iteration. The use of truncated linear models significantly mitigates this concern by providing a "signal" that gradient descent might overshoot, allowing the learning algorithm to take a more conservative step. The same idea can be applied to the online (rather than stochastic) setting, but only for OMD. It is unknown how to use truncated linear models in parameter-free algorithms without having an explosion in the computational time.

In this work, we propose new parameter-free algorithms that are able to take advantage of truncated linear models. Note that any optimization algorithm based on linear models can overshoot the optimum, but parameter-free algorithms may be even more prone to overshooting because their iterates can move *exponentially far* between iterations. Instead, our new algorithms effectively alleviate this problem, see Figure 1. Our algorithms are based on a *new decomposition of the regret that takes advantage of the geometry of truncated linear losses* that might be of independent interest.

In summary, our primary contribution is a new algorithm that maintains optimal parameter-free regret bounds *but also* incorporates additional geometric information about the loss functions. While such an improvement is not visible in *worst-case* rates, we demonstrate through an "implicit-style" regret bound that the algorithm could perform *significantly better* in practice, and verify this behavior in an empirical study.

Figure 1: Coin-Betting with truncated linear models (Red), Coin-Betting with linear models (Green), OGD with large constant stepsizes (Grey), and OGD with small constant stepsizes (Black).

The rest of the paper is organized as follows: in Section 2 we discuss related work and in Section 3 we review some definitions and background knowledge. In Section 4, we show the difficulties in using truncated linear models in parameter-free algorithms. In Section 5, we present our solution and prove a bound on its regret. Since this algorithm does not have a closed form update rule, in Section 6 we propose a more efficiently computable variant while still retaining the same theoretical guarantee. In Section 7, we present a coordinate-wise extension that obtains a tighter bound as well as better empirical performance. Finally, in Section 8, we empirically validate our algorithm.
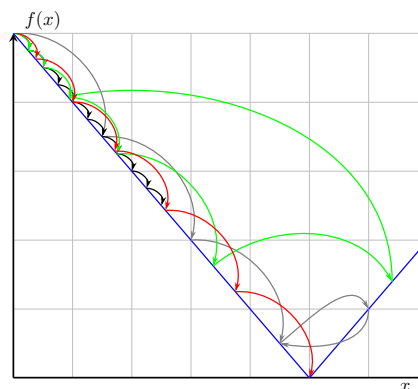
2. Related work

Parameter-free OCO Algorithms Parameter-free OCO algorithms are motivated by a desire to avoid choosing a step size and can achieve optimal theoretical regret bounds (e.g., McMahan and Orabona, 2014; Orabona, 2014; Orabona and Pál, 2016; Cutkosky and Boahen, 2017; Foster et al., 2018; Cutkosky and Orabona, 2018; Kotłowski, 2020; Kempka et al., 2019; Cutkosky and Sarlos, 2019; Jun and Orabona, 2019; van der Hoeven, 2019; Mhammedi and Koolen, 2020; Orabona and Pál, 2021; Chen et al., 2021). Some of them are based on the FTRL framework (Shalev-Shwartz, 2007; Abernethy et al., 2008b; Hazan and Kale, 2008) (sometimes indirectly through methods such as coin-betting). The closest work to our algorithms is the CODE algorithm (Chen et al., 2022) which is the first attempt to combine parameter-free methods with truncated losses. Inspired by the Importance Weight Awareness updates in Karampatziakis and Langford (2011), CODE models the optimization algorithm with an ODE, and solves the ODE in a closed form to make infinitely many infinitesimal parameter-free updates on truncated losses. While CODE solves the ODE in closed form, it does not have any theoretical guarantee. In our work, in each step t, we consider the loss in two points only: on the current prediction and the updated one. This gives rise to an implicit equation that we can solve for truncated losses and to an optimal regret guarantee.

Truncated Linear Models and Implicit Updates Truncated linear models were proposed in Asi and Duchi (2019) to create a tighter surrogate model for optimization. While the use of convex linear lower bounds is also the core method in OCO algorithms (see, e.g., Orabona, 2019), we are not aware of any other online learning algorithm with a regret guarantee based on truncated linear models. Asi and Duchi (2019) incorporate truncated linear models into the Mirror Descent update (Nemirovsky and Yudin, 1983), forming a *proximal/implicit* update (Moreau, 1965; Martinet, 1970; Rockafellar, 1976; Kivinen and Warmuth, 1997; Parikh and Boyd, 2014). In online learning, Kulis and Bartlett (2010) provides the first regret bounds for implicit updates that match those of OMD, while McMahan (2010) makes the first attempt to quantify the advantage of the implicit updates in the regret bound. Song et al. (2018) generalize the results in McMahan (2010) to Bregman divergences and strongly convex functions, and quantify the gain differently in the regret bound. Finally, Campolongo and Orabona (2020) show that implicit updates give rise to regret guarantees that depend on the temporal variability of the losses as well. We will match the dependency on the subgradients in our final results to the one of FTRL with implicit updates (McMahan, 2010), which underlines the "implicit" nature of our algorithm.

3. Preliminary

In this section, we introduce some of the needed background and definition.

Convex Analysis Definitions For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we define a subgradient of f in $\boldsymbol{x} \in \mathbb{R}^d$ as a vector $\boldsymbol{g} \in \mathbb{R}^d$ that satisfies $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle, \ \forall \boldsymbol{y} \in \mathbb{R}^d$. We denote the set of subgradients of f at $\boldsymbol{x}$ by $\partial f(\boldsymbol{x})$. A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is μ-strongly convex over a convex set $V \subseteq \operatorname{int} \operatorname{dom} f$ w.r.t. $\| \cdot \|$ if $\forall \boldsymbol{x}, \boldsymbol{y} \in V$, we have $\boldsymbol{g} \in \partial f(\boldsymbol{x})$, $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$. The Fenchel conjugate $f^\star$ of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as $f^\star(\boldsymbol{\theta}) = \sup_{\boldsymbol{x}} \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle - f(\boldsymbol{x})$. We denote the projection of a vector $\boldsymbol{x}$ onto a convex set B as $\Pi_B(\boldsymbol{x}) \triangleq \operatorname{argmin}_{\boldsymbol{u} \in B} \|\boldsymbol{x} - \boldsymbol{u}\|^2$.

Coin-Betting and Online Learning We now explain the coin-betting framework for parameter-free algorithm design (Orabona and Pál, 2016), which operates through convex duality. We consider a vector-valued "coin" $\boldsymbol{c}_t \in \mathbb{R}^d$ with $\|\boldsymbol{c}_t\| \leq 1$ provided to a gambler in response to a "bet" $\boldsymbol{x}_t \in \mathbb{R}^d$.

The gambler earns $\langle \boldsymbol{c}_t, \boldsymbol{x}_t \rangle$ dollars, for a total wealth of $\text{Wealth}_t = \epsilon + \sum_{i=1}^{t} \langle \boldsymbol{c}_i, \boldsymbol{x}_i \rangle$ at time t, assuming an initial endowment of ϵ. We enforce that $\boldsymbol{x}_t = \boldsymbol{\beta}_t \text{Wealth}_{t-1}$ for some *betting fraction* $\|\boldsymbol{\beta}_t\| \leq 1$, which intuitively corresponds to preventing the gambler from betting more money than the gambler has: $\text{Wealth}_t \geq 0$ for all t. The goal of the gambler is of course to make the wealth as high as possible. To use this gambling game in online learning, set $\boldsymbol{c}_t \in -\partial \ell_t(\boldsymbol{x}_t)$, and let the learner's $\boldsymbol{w}_t \in \mathbb{R}^d$ be simply equal to the gambler's $\boldsymbol{x}_t$. To analyze the regret, suppose that $\text{Wealth}_T \geq H\left(\sum_{t=1}^{T} \boldsymbol{c}_t\right)$ for some arbitrary function H. Then we have:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \sum_{t=1}^{T} \langle \boldsymbol{c}_t, \boldsymbol{u} - \boldsymbol{w}_t \rangle = \epsilon + \sum_{t=1}^{T} \langle \boldsymbol{c}_t, \boldsymbol{u} \rangle - \text{Wealth}_T$$

$$\leq \epsilon + \sum_{t=1}^{T} \langle \boldsymbol{c}_t, \boldsymbol{u} \rangle - H\left(\sum_{t=1}^{T} \boldsymbol{c}_t\right) \leq \epsilon + \sup_{G \in \mathbb{R}^d} \langle G, \boldsymbol{u} \rangle - H(G) = \epsilon + H^\star(\boldsymbol{u})$$

where in the first inequality we use the definition of the subgradient, in the second the assumption on H, and the last equality the definition of Fenchel conjugate $H^\star$.

Critically, notice that the wealth lower-bound $\text{Wealth}_T \geq H\left(\sum_{t=1}^{T} \boldsymbol{c}_t\right)$ does not involve $\boldsymbol{u}$. Instead, $\boldsymbol{u}$ appears only in analysis through Fenchel duality, which provides the parameter-free property. Hence, *we can use any betting algorithm that guarantees a high wealth to design a parameter-free optimization algorithm.*

Truncated Linear model For any $\hat{\ell}_t$ such that $\hat{\ell}_t(\boldsymbol{w}_t) = \ell_t(\boldsymbol{w}_t)$ and $\hat{\ell}_t(\boldsymbol{w}) \leq \ell_t(\boldsymbol{w})$, $\forall \boldsymbol{w}$, we have $\sum_{t=1}^{T}(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})) \leq \sum_{t=1}^{T}(\hat{\ell}_t(\boldsymbol{w}_t) - \hat{\ell}_t(\boldsymbol{u}))$. Linear models $\hat{\ell}_t(\boldsymbol{w}) = \ell_t(\boldsymbol{w}_t) + \langle \boldsymbol{g}_t, \boldsymbol{w} - \boldsymbol{w}_t \rangle$ satisfy this property with $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$, which motivates the popularity of online *linear* optimization. However, we might think to design tighter approximations. In particular, Asi and Duchi (2019) proposed truncated linear models:

$$\hat{\ell}_t(\boldsymbol{w}) \triangleq \max\{\ell_t(\boldsymbol{w}_t) + \langle \boldsymbol{g}_t, \boldsymbol{w} - \boldsymbol{w}_t \rangle, \inf_{\boldsymbol{w}} \ell_t(\boldsymbol{w})\}\ .$$
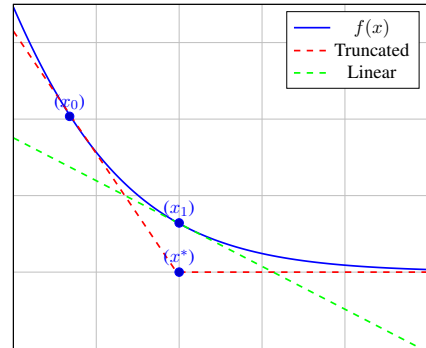

Figure 2: Models of the function $f(x) = \log(1 + e^{-x})$: a truncated linear model (Red) built around the point x_0, and a linear model (Green) built around the point x_1. x^* is the hinge corner.

Besides the property above, truncated linear models satisfy: i) $\hat{\ell}_t(\boldsymbol{w})$ is convex and subdifferentiable on the domain; ii) For any $\boldsymbol{w}$, we have $\boldsymbol{g}^+ = h\boldsymbol{g}_t$ where $\boldsymbol{g}^+ \in \partial \hat{\ell}_t(\boldsymbol{w})$ and $h \in [0, 1]$; iii) $\hat{\ell}_t(\boldsymbol{w}) \geq \inf_{\boldsymbol{w}} \ell_t(\boldsymbol{w})$. In the following, we will assume that $\inf_{\boldsymbol{w}} \ell_t(\boldsymbol{w}) = 0$, w.l.o.g. for loss functions bounded from below.

4. Difficulties in Using Truncated Models in Parameter-free Algorithms

Many parameter-free algorithms are based on FTRL. Hence, it is natural to ask whether it is possible to directly use truncated linear models instead of linear models in FTRL to utilize truncated linear models in parameter-free algorithms. This approach immediately runs into significant problems. Specifically, FTRL algorithms usually maintain the sum of the losses observed so far. One can

easily store this sum when the losses are linear, but truncated losses would require $\mathcal{O}(T)$ space and $\mathcal{O}(poly(T))$ time for every update. Thus, using truncated linear models with FTRL has the same computational cost as using the original cost functions - the simplification to using truncated linear models does not appear to help. In contrast, our solution has the same computational and space complexity of online gradient descent.

Another possibility is to adapt the coin-betting design of parameter-free algorithms (Orabona and Pál, 2016) to truncated linear models. A moment of thinking should convince the reader this is far from simple: The reduction from optimization to coin-betting described in Section 3 works by transforming subgradients into coin outcomes, but the subgradient of a truncated loss is exactly the same as the subgradient of the original function! Thus, simply using the reduction as-is on the truncated linear model would provide *no benefit* over using the simpler linear model.

Another method that seems possible is using truncated linear models in online gradient descent, and then make online gradient descent parameter-free via some application of the doubling trick (Shalev-Shwartz, 2012, 2.3.1). Perhaps surprisingly, employing the doubling trick in this way is quite difficult. The only known application of the doubling trick is in Streeter and McMahan (2012), but it does not achieve the optimal regret bound, and more importantly, employs a delicate identity relating the regret and sums of gradients that may fail for the truncated linear model. Furthermore, the doubling trick usually has terrible empirical performance, which completely defeats the purpose of using truncated linear models.

5. Parameter-free OCO with Truncated Linear Models

In this section, we introduce our novel parameter-free algorithms for truncated linear models.

We overcome the difficulties described above through a multi-step process, during which we will introduce three separate algorithms. First, we introduce a new regret decomposition to take advantage of truncated losses while still only requiring storage of a few vectors. As an illustration of the key principles, we will use this new regret decomposition to design Algorithm 1. While the theoretical guarantee of Algorithm 1 matches our desiderata, the update does not have a closed form. Hence, we then show how to slightly change our algorithm to obtain a closed form update in Algorithm 2. Finally, in Section 7, we consider each coordinate as a separate 1-d problem to obtain a coordinate-wise variant that achieves better performance both theoretically and empirically.

As mentioned previously, using truncated linear models in an FTRL-based parameter-free algorithm would result in an inefficient update. Therefore, in the following, we show a different approach inspired by the idea of implicit updates (Kivinen et al., 2006; Kulis and Bartlett, 2010; McMahan, 2010). Our method introduces a new variation on the standard approach to bounding online convex optimization with online linear optimization, and an accompanying update to the regret/reward duality.

A New Regret Decomposition We are interested in upper bounding the terms $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})$ for any $\boldsymbol{u} \in \mathbb{R}^d$. The usual method (Zinkevich, 2003) is to upper bound the regret by linear terms, and then proceed to bound the regret on the linear losses as follows:

$$\text{Regret}_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle,$$

where $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$. While this approach gives worst-case optimal upper bounds, it completely ignores the geometry of the loss functions ℓ_t. In contrast, we consider upper bounding the term

---

Algorithm 1 Parameter-free OCO with Truncated Linear Models

1: Initialize $\boldsymbol{\beta}_1 \leftarrow \mathbf{0}, \mathrm{Wealth}_0 \leftarrow 1, \eta_1 \leftarrow 1/3$
2: **for** $t = 1$ **to** T **do**
3: Predict $\boldsymbol{w}_t \leftarrow \boldsymbol{\beta}_t \mathrm{Wealth}_{t-1}$
4: Receive $\ell_t(\boldsymbol{w}_t)$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$
5: Calculate $\boldsymbol{g}_t^+$ (see Section 5.1). $\boldsymbol{g}_t^+ = h_t \boldsymbol{g}_t$ by property (ii) of truncated linear models
6: $\hat{\boldsymbol{\beta}}_{t+1} \leftarrow \boldsymbol{\beta}_t - \frac{\boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)}{1/\eta_1 + 2\sum_{i=1}^{t-1}(\|\boldsymbol{g}_i\|^2 - \|\boldsymbol{g}_i^+ - \boldsymbol{g}_i\|^2)}$
7: $\boldsymbol{\beta}_{t+1} \leftarrow \hat{\boldsymbol{\beta}}_{t+1} / \max\left(1, 2\|\hat{\boldsymbol{\beta}}_{t+1}\|\right)$
8: $\mathrm{Wealth}_t \leftarrow \mathrm{Wealth}_{t-1} \frac{1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle}{1 + (h_t - 1)\langle \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle}$
9: **end for**

$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})$ with the truncated linear loss $\hat{\ell}_t$, and decompose the regret on the truncated linear losses from an "implicit" point of view for a tighter bound. Specifically, for any $\boldsymbol{u} \in \mathbb{R}^d$ we have

$$
\begin{aligned}
\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \hat{\ell}_t(\boldsymbol{w}_t) - \hat{\ell}_t(\boldsymbol{u}) &= \hat{\ell}_t(\boldsymbol{w}_t) - \hat{\ell}_t(\boldsymbol{w}_{t+1}) + \hat{\ell}_t(\boldsymbol{w}_{t+1}) - \hat{\ell}_t(\boldsymbol{u}) \\
&\leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + \langle \boldsymbol{g}_t^+, \boldsymbol{w}_{t+1} - \boldsymbol{u} \rangle,
\end{aligned}
\tag{1}
$$

where $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t), \boldsymbol{g}_t \in \partial \hat{\ell}_t(\boldsymbol{w}_t), \boldsymbol{g}_t^+ \in \partial \hat{\ell}_t(\boldsymbol{w}_{t+1})$, the first inequality is true by the property of the truncated linear model, and the second inequality is from the convexity. This decomposition can *take into account part of the geometry of the function through* $\boldsymbol{g}_t^+$, which quantifies how far we are from the infimum of ℓ_t. Note that the decomposition itself is very general and does *not* require $\hat{\ell}$ to be a truncated linear loss: this structure is primarily used to form more efficient algorithms.

Implicit Coin-Betting To leverage this decomposition, we now define a modified notion of the *wealth* quantity described in Section 3. Our key idea is to realize that *the regret/reward duality is more general than previously thought*. In particular, we define $\mathrm{Wealth}_0 = \epsilon$ and $\mathrm{Wealth}_T \triangleq \mathrm{Wealth}_{T-1} - \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle - \langle \boldsymbol{g}_t^+, \boldsymbol{w}_{t+1} \rangle$, to have

$$
\mathrm{Wealth}_T = \epsilon - \sum_{t=1}^{T}(\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + \langle \boldsymbol{g}_t^+, \boldsymbol{w}_{t+1} \rangle).
\tag{2}
$$

This implies $\mathrm{Regret}_T(\boldsymbol{u}) \leq \epsilon + \left\langle -\sum_{t=1}^{T} \boldsymbol{g}_t^+, \boldsymbol{u} \right\rangle - \mathrm{Wealth}_T$. Suppose that we obtain a bound $\mathrm{Wealth}_T \geq \psi_T\left(-\sum_{t=1}^{T} \boldsymbol{g}_t^+\right)$ for some ψ_T. Then, we can still use the Fenchel conjugate:

$$
\mathrm{Regret}_T(\boldsymbol{u}) - \epsilon \leq -\left\langle \sum_{t=1}^{T} \boldsymbol{g}_t^+, \boldsymbol{u} \right\rangle - \psi_T\left(-\sum_{t=1}^{T} \boldsymbol{g}_t^+\right) \leq \sup_{\boldsymbol{y}} \langle \boldsymbol{y}, \boldsymbol{u} \rangle - \psi_T(\boldsymbol{y}) = \psi_T^\star(\boldsymbol{u}).
$$

Hence, it suffices to design an algorithm that guarantees a lower bound on Wealth_T to achieve a regret upper bound, *even for our modified notion of wealth*. Moreover, given that our regret decomposition takes into account the geometry of the truncated linear losses, we can expect a regret guarantee that becomes tighter when we are close to the infimum of the functions ℓ_t.

We designed Algorithm 1 to maximize the wealth in (2), yielding a regret bound in Theorem 1.

Theorem 1 *Assume $\ell_t(\boldsymbol{x}), t = 1, \ldots, T$, to be convex functions. Set* $\text{Wealth}_0 = \epsilon = 1$ *and assume that* $\|\boldsymbol{g}_t\| \leq 1$ *and* $\boldsymbol{g}_t^+ = h_t \boldsymbol{g}_t$ *where* $h_t \in [0, 1]$*. Then, Algorithm 1 satisfies*

$$
\begin{aligned}
\text{Regret}_T(\boldsymbol{u}) = \mathcal{O}\Bigg(\max\Bigg\{ &\|\boldsymbol{u}\| \ln\left(\|\boldsymbol{u}\| \left(1 + \sum_{t=1}^T \|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| \right) \right), \\
&\|\boldsymbol{u}\| \sqrt{ \sum_{t=1}^T (2\|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2) \cdot \ln\left(1 + \|\boldsymbol{u}\| \sum_{t=1}^T (2\|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2) \right) } \Bigg\} \Bigg).
\end{aligned}
$$

To convey the main ideas, here we present a proof sketch, the full proof is included in the Appendix.

Proof [Proof sketch] We first lower bound the wealth of the algorithm. From the definition of the wealth (2) and the fact that the algorithm predicts with $\boldsymbol{w}_t = \boldsymbol{\beta}_t \text{Wealth}_{t-1}$, we have

$$
\text{Wealth}_t = \text{Wealth}_{t-1} - \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle - \langle \boldsymbol{g}_t^+, \boldsymbol{w}_{t+1} \rangle \Rightarrow \text{Wealth}_t = \frac{\text{Wealth}_{t-1}(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle)}{1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle}.
$$

This implies that $\ln \text{Wealth}_T = \ln \epsilon + \sum_{t=1}^T (\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle))$. It is possible to show that $\ln \text{Wealth}_T - \ln \epsilon$ can be lower bounded as

$$
\begin{aligned}
\sum_{t=1}^T &\left(\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle) \right) \\
&\geq \sum_{t=1}^T \left[-\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|\boldsymbol{\beta}_t\|^2 - 2\|\boldsymbol{g}_t\| \|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \right] .
\end{aligned}
\tag{3}
$$

So, $\boldsymbol{\beta}_t$ is designed to be the output of running OGD (Online Gradient Descent) on μ_t strongly-convex losses $f_t(\boldsymbol{\beta}) \triangleq \langle \boldsymbol{g}_t^+, \boldsymbol{\beta} \rangle + \frac{\mu_t}{2}\|\boldsymbol{\beta}_t\|^2$, where $\mu_t = 2(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t - \boldsymbol{g}_t^+\|^2)$ with $\boldsymbol{\beta} \in B$, $B = \{\boldsymbol{x} | \|\boldsymbol{x}\| \leq 1/2\}$, and stepsizes $\eta_t = \frac{1}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i}$. Standard OGD analysis provides the lower bound for $\sum_{t=1}^T \langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle + (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|\boldsymbol{\beta}_t\|^2$. Besides, $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|$ is upper bounded by $\frac{3\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^t \|\boldsymbol{g}_i\| \|\boldsymbol{g}_i^+\|}$. Combining all pieces together leads to the lower bound

$$
\ln \text{Wealth}_T \geq -3/2 - 7.25 \ln\left(1 + 2\sum_{t=1}^T \|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| \right) + \min\left\{ \frac{\left\| \sum_{t=1}^T \boldsymbol{g}_t^+ \right\|}{4}, \frac{\left\| \sum_{t=1}^T \boldsymbol{g}_t^+ \right\|^2}{2\sum_{t=1}^T \mu_t} \right\} .
$$

A lower bound on Wealth_T indicates an upper bound on regret. Now, we derive the upper bound on the regret from the Fenchel conjugate of the function above. ∎

Note that the bounded subgradient assumption is a known requirement shared by all parameter-free algorithms, see lower bound in Cutkosky and Boahen (2017). However, the limitation is milder than it seems at first blush: this Lipschitz bound can actually be over-estimated by a factor of $\sqrt{T}$ before significant damage is done to the regret bound. This can be seen by observing that other than an $O(\log T)$ term, our regret bounds scale with the observed norms of the gradients. Thus, the limitation is actually rather benign - we simply assume a bound of 1 to simplify equations.

Comparison with Parameter-Free Bounds Previous work of Cutkosky and Orabona (2018) achieved a regret bound of $\mathcal{O}\left(\|\boldsymbol{u}\|\sqrt{\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2}\right)$ which has the optimal worst-case dependence on $\|\boldsymbol{g}_t\|$ (Abernethy et al., 2008a; Cutkosky, 2018). In Theorem 1, we obtain a regret bound depending on $\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2 = \|\boldsymbol{g}_t\|^2(2h_t - h_t^2)$. So, as long as the algorithm goes close to the hinge corner of the truncated linear model, it will yield an $h_t < 1$ and a smaller regret. Intuitively, this should be expected to occur whenever it is possible to obtain small loss as obtaining small loss requires reaching the hinge of the truncated linear model.

Relation to Implicit Updates Truncated linear models were introduced as an approximation of the implicit updates (Asi and Duchi, 2019). In this view, it is instructive to compare the dependency on the subgradients in Theorem 1 and the regret bounds for implicit updates. For example, McMahan (2010, Theorem 2) gives a regret guarantee for FTRL with implicit updates and non-adaptive regularizer that depends on $\langle \boldsymbol{g}_t - \frac{1}{2}\boldsymbol{g}_t^+, \boldsymbol{g}_t^+ \rangle$. This quantity is exactly $\frac{1}{2}(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)$ that appears in Theorem 1. This supports the idea that the decomposition in (1) "emulates" the idea of implicit updates in parameter-free algorithms. However, there is a subtle difference: in standard implicit updates $\boldsymbol{g}_t^+ \in \partial \ell_t(\boldsymbol{w}_{t+1})$ is a subgradient of the original loss function. Instead, here $\boldsymbol{g}_t^+ \in \partial \hat{\ell}_t(\boldsymbol{w}_{t+1})$, so it is a subgradient of the truncated linear model. We can see this as a price we pay to obtain a smaller computational complexity compared to standard implicit updates.

Comparison with OMD with truncated linear model To the best of our knowledge, there are actually no regret guarantees with OMD with truncated linear models in the literature (Asi and Duchi (2019) do not consider the adversarial setting). However, it is quite likely that OMD with truncated models can achieve an implicit regret similar to that reported by McMahan (2010) *subject to oracle tuning of the learning rates*. Our results match this benchmark in the dependency on $\boldsymbol{g}_t$ and $\boldsymbol{g}_t^+$ and improve in the dependency on $\|\boldsymbol{u}\|$ since we do not require oracle tuning of the learning rate.

No Overshooting Property We now prove that the proposed algorithm never overshoots the minimum of the truncated linear loss. Moreover, in the case that the minimum of $\hat{\ell}_t$ coincides with the minimum of ℓ_t, we end up exactly in the minimum, as illustrated by Figure 1.

Theorem 2 *Under the assumptions of Theorem 1 and the notation of Algorithm 1, assume that $\boldsymbol{g}_t \neq \boldsymbol{0}$. Then, the update $\boldsymbol{w}_{t+1}$ can never land on the flat part of the loss $\hat{\ell}_t$, but only on its linear part or in the corner.*

Proof The statement is equivalent to showing that $\ell_t(\boldsymbol{w}_t) + \langle \boldsymbol{g}_t, \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle \geq 0$ from the definition of $\hat{\ell}_t$. We prove it by contradiction. Let's assume that $\ell_t(\boldsymbol{w}_t) + \langle \boldsymbol{g}_t, \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle < 0$. Then, we would have $\hat{\ell}_t(\boldsymbol{w}_{t+1}) = 0$ and $\boldsymbol{g}_t^+ = \boldsymbol{0}$ (equivalently $h_t = 0$). In turn, this would imply $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t$ and $\text{Wealth}_t = \text{Wealth}_{t-1}$. So, we would have $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$ which is impossible because $\boldsymbol{g}_t \neq \boldsymbol{0}$. ∎

Note that we assume $\boldsymbol{g}_t \neq \boldsymbol{0}$ in Theorem 2 since when $\boldsymbol{g}_t = \boldsymbol{0}$ the algorithm is already in the corner.

5.1. Computation of h_t

The next challenge is how to find h_t. This is the only part of the Algorithm that uses the truncated linear model structure: the analysis Theorem 1 actually applies to *any* losses for which $\boldsymbol{g}_t^+ = h_t \boldsymbol{g}_t$. Truncated linear models combine this favorable property with the additional property that it is possible to efficiently compute h_t. The argument of Theorem 2 shows that we cannot be in the flat

---

Algorithm 2 Parameter-free OCO with Truncated Linear Models – closed form update

---

1: Initialize $\boldsymbol{\beta}_1 \leftarrow \mathbf{0}, \eta_1 \leftarrow \frac{1}{2C}, C \leftarrow 9, \text{Wealth}_0 \leftarrow 1$
2: **for** $t = 1$ **to** T **do**
3:      Predict $\boldsymbol{w}_t \leftarrow \boldsymbol{\beta}_t \text{Wealth}_{t-1}$
4:      Receive $\ell_t(\boldsymbol{w}_t)$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$
5:      Calculate h_t (see Section 6.1)
6:      **if** $\|\boldsymbol{\beta}_t\| < \frac{3}{8}$ **then**
7:          $\boldsymbol{\beta}_{t+1} \leftarrow \boldsymbol{\beta}_t - \eta_t(\boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2))$
8:          $\frac{1}{\eta_{t+1}} \leftarrow \frac{1}{\eta_t} + 2(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)$
9:      **else**
10:          $\boldsymbol{\beta}_{t+1} \leftarrow \boldsymbol{\beta}_t - \eta_t 2C\|\boldsymbol{g}_t^+\|\boldsymbol{\beta}_t$
11:          $\frac{1}{\eta_{t+1}} \leftarrow \frac{1}{\eta_t} + 2C\|\boldsymbol{g}_t^+\|$
12:      **end if**
13:      $\text{Wealth}_t \leftarrow \text{Wealth}_{t-1} \frac{1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle}{1 + (h_t - 1)\langle \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle}$
14: **end for**

---

region. By inspection of the updates, there are two achievable cases: in the first case $h_t = 1$ and we are not in the corner of the truncated model, while in the second case, we are in the corner. Hence, as a first step, we posit that $h_t = 1$, calculate $\boldsymbol{w}_{t+1}$ and see if indeed $\boldsymbol{g}_t^+ = \boldsymbol{g}_t$. If this is not the case, then the solution must be in the corner and $h_t \in [0, 1)$. By definition $\boldsymbol{w}_{t+1} = \boldsymbol{\beta}_{t+1} \text{Wealth}_t$, where

$$\boldsymbol{\beta}_{t+1} = \prod_B (\boldsymbol{\beta}_t - \eta_t(h_t\boldsymbol{g}_t + 2\boldsymbol{\beta}_t\|\boldsymbol{g}_t\|^2(2h_t - h_t^2))) \text{ and } \text{Wealth}_t = \text{Wealth}_{t-1} \frac{1 - \boldsymbol{g}_t\boldsymbol{\beta}_t}{1 + (\boldsymbol{g}_t^+ - \boldsymbol{g}_t)\boldsymbol{\beta}_{t+1}} .$$

Thus, $\boldsymbol{w}_{t+1}$ is a function of h_t. Assuming $\inf_{\boldsymbol{w}} \ell_t(\boldsymbol{w}) = 0$ w.l.o.g. for loss functions bounded from below, we are looking for h_t that makes

$$\ell_t(\boldsymbol{w}_t) + \langle \boldsymbol{g}_t, \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle = 0 . \tag{4}$$

Although we could solve for h_t via bisection, there is no closed form solution due to the projection of $\boldsymbol{\beta}_{t+1}$. Thus, we next propose a more complex algorithm with a closed form equation for h_t.

6. Variant with Closed-form Update

In this section, we introduce the Algorithm 2: a variant of Algorithm 1 that has a closed form update. The key steps are still the same, but here we want to remove the projection step on $\boldsymbol{\beta}_t$. In this way, the expression of $\boldsymbol{w}_{t+1}$ depends on a simple polynomial in h_t. In turn, to remove the projection step, we change the update of $\boldsymbol{\beta}_t$ so that its norm is always assured to be bounded. For Algorithm 2, we can prove the following guarantee. We present a proof sketch, while the full proof is in the Appendix.

Theorem 3 *Assume $\ell_t(x), t = 1, \ldots, T$, to be convex functions. Set $\text{Wealth}_0 = \epsilon = 1$, $C = 9$ and $1/\eta_1 = 2C$, and assume that $\|\boldsymbol{g}_t\| \leq 1$ and $\boldsymbol{g}_t^+ = h_t\boldsymbol{g}_t$ where $h_t \in [0, 1]$. Then, for all $\boldsymbol{u} \in \mathbb{R}^d$, Algorithm 2 satisfies the same bound as in Theorem 1, up to constants hidden in the big O notation.*

Proof [Proof sketch] As we stated above, due to the projection step in line 6 of Algorithm 1, h_t can not be solved in a closed form. To overcome this, we design a new update rule that guarantees that $\boldsymbol{\beta}_t$ will always end up in the ball $B = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq \frac{1}{2}\}$ to avoid the projection step.

In Algorithm 1, β_t is the output of running OGD on strongly-convex losses $f_t(\beta)$. However, when β_t is close to 1/2, the next iteration, β_{t+1}, could go too far so that a projection step can be necessary. To avoid this, we intricately design an update rule that when $\|\beta_t\| \geq 3/8$ indicating that β_t is close to the boundary of the ball, the next iteration will shrink it a little bit, to make sure that it stays in the ball B. The new update rule is the output of running OGD on strongly-convex losses $\phi_t(\beta)$. In the following, we introduce the sketch of the proof.

β_t is the output of OGD with $\eta_t = \frac{1}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i}$ on the strongly-convex losses $\phi_t(\beta)$:

$$
\phi_t(\beta) = \begin{cases} f_t(\beta), & \text{if } \|\beta_t\| < \frac{3}{8}, \\ C\|g_t^+\|\|\beta\|^2, & \text{if } \frac{3}{8} \leq \|\beta_t\| \leq \frac{1}{2}, \end{cases}
$$

where $\beta_t \in B$ and $B = \{x : \|x\| \leq \frac{1}{2}\}$. $\phi_t(\beta)$ is μ_t strongly convex. $\beta_1^\star \triangleq \arg\min_{\beta \in B_1} \sum_{t=1}^{T} f_t(\beta)$, where $B_1 = \{x : \|x\| \leq \frac{1}{4}\}$. The intricate design of $\phi_t(\beta)$ allows to say that if $\beta_t \geq 3/8$, $\|\beta_{t+1}\| = (1 - \eta_t 2C\|g_t^+\|)\|\beta_t\|$ will shrink if $\eta_t \leq 1/2C$; if $\beta_t \leq 3/8$, $\|\beta_{t+1}\|$ will stay in B if η_t is small enough. Therefore, Algorithm 2 guarantees $\|\beta_t\| \leq \frac{1}{2}$ for all $t = 1, \ldots, T$, which removes the projection step on β_t, and gives rise to the closed form updates.

Furthermore, for $C \geq 9$, the regret of $\phi_t(\beta_t)$ upper bounds the regret of $f_t(\beta_t)$: $f_t(\beta_t) - f_t(\beta_1^\star) \leq \phi_t(\beta_t) - \phi_t(\beta_1^\star)$. We upper bound $\sum_{t=1}^{T} \phi_t(\beta_t) - \phi_t(\beta_1^\star)$ by standard OGD analysis, which implies a bound on $\sum_{t=1}^{T} f_t(\beta_t)$. Combining the upper bound of $\|\beta_{t+1} - \beta_t\|$ and $\sum_{t=1}^{T} f_t(\beta_t)$ lower bounds wealth by (3). The rest of the proof is similar to the proof of Theorem 1. ∎

We also note that the non-overshooting property holds for this algorithm too. The proof is exactly the same as before and it is omitted.

Theorem 4 *Under the assumptions of Theorem 3 and the notation of Algorithm 2, assume that $g_t \neq 0$. Then, the update w_{t+1} can never land on the flat part of the loss $\hat{\ell}_t$, but only on its linear part or on the corner.*

6.1. Computation of h_t with Closed Form Solution

Now, we show how to obtain a closed form expression for the update in Algorithm 2. As before, first we tentatively set $h_t = 1$ and check if $g_t^+ = g_t$. If yes, then $h_t = 1$ and we can compute w_{t+1}. If not, thanks to Theorem 4, we know that w_{t+1} lands in the corner of $\hat{\ell}_t$ and we need to compute $h_t \in [0, 1)$. In this case we are looking for the h_t such that w_{t+1} satisfies (4).

Let $A = \langle g_t, w_t \rangle - \ell_t(w_t)$, $B = \text{Wealth}_t(1 - \langle g_t, \beta_t \rangle)$. We consider two cases based on β_t.

If $\|\beta_t\| \leq \frac{3}{8}$, we have that $\langle g_t, \beta_{t+1} \rangle = \langle g_t, \beta_t - \eta_t(g_t^+ + 2\beta_t(2h_t\|g_t\|^2 - h_t^2\|g_t\|^2)) \rangle$. Let $D = 2\eta_t\|g_t\|^2\langle g_t, \beta_t \rangle$, so (4) becomes a cubic equation of h_t that has closed form solution:

$$
\begin{aligned}
&- ADh_t^3 + (2AD + A\eta_t\|g_t\|^2 + (A+B)D)h_t^2 + (-(A+B)\eta_t\|g_t\|^2 - 2(A+B)D \\
&- A\langle g_t, \beta_t \rangle)h_t + (A+B)\langle g_t, \beta_t \rangle - A = 0 \,.
\end{aligned}
$$

If $\frac{3}{8} \leq \|\beta_t\| \leq \frac{1}{2}$, we have $\langle g_t, \beta_{t+1} \rangle = \langle g_t, \beta_t \rangle(1 - 2C\eta_t\|g_t\|h_t)$. Let $D = 2C\eta_t\|g_t\|\langle g_t, \beta_t \rangle$. So, (4) can be rewritten as the following quadratic equation of h_t and again it has closed form solution: $ADh_t^2 + (-A\langle g_t, \beta_t \rangle - (A+B)D)h_t + (A+B)\langle g_t, \beta_t \rangle - A = 0$.

Algorithm 3 Parameter-free OCO with Truncated Linear Models – coordinate-wise update

1: Initialize $\boldsymbol{\beta}_1 \leftarrow \mathbf{0}, \eta_1 \leftarrow \frac{1}{2C} \cdot \mathbf{1}, C \leftarrow 9, \text{Wealth}_0 \leftarrow \mathbf{1} \in \mathbb{R}^d$
2: **for** $t = 1$ **to** T **do**
3:      Predict $\boldsymbol{w}_t \leftarrow \boldsymbol{\beta}_t \odot \text{Wealth}_{t-1}$
4:      Receive $\ell_t(\boldsymbol{w}_t)$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$
5:      Calculate h_t s.t. $\ell_t(\boldsymbol{w}_t) + \langle \boldsymbol{g}_t, \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle = 0$ (see Section 5.1)
6:      **for** $i = 1$ **to** d **do**
7:          **if** $|\beta_{t,i}| < \frac{3}{8}$ **then**
8:              $\beta_{t+1,i} \leftarrow \beta_{t,i} - \eta_{t,i}(g_{t,i}^+ + 2\beta_{t,i}(2g_{t,i}g_{t,i}^+ - (g_{t,i}^+)^2))$
9:              $\frac{1}{\eta_{t+1,i}} \leftarrow \frac{1}{\eta_{t,i}} + 2(g_{t,i}^2 - (g_{t,i}^+ - g_{t,i})^2)$
10:         **else**
11:             $\beta_{t+1,i} \leftarrow \beta_{t,i} - \eta_{t,i}2C|g_{t,i}^+|\beta_{t,i},$
12:             $\frac{1}{\eta_{t+1,i}} \leftarrow \frac{1}{\eta_{t,i}} + 2C|g_{t,i}^+|$
13:         **end if**
14:         $\text{Wealth}_{t,i} \leftarrow \text{Wealth}_{t-1,i} \frac{1 - g_{t,i}\beta_{t,i}}{1 + (h_t - 1)g_{t,i}\beta_{t+1,i}}$
15:     **end for**
16: **end for**

7. Tighter Regret Guarantee through Coordinate-wise Updates

In this section, we introduce a coordinate-wise variant of Parameter-free OCO with truncated linear models. This is a simple extension of Algorithm 2 by considering each coordinate as a different 1-d OCO algorithm. The advantage is that this regret bound is even *tighter* than the bound of Theorem 1. Here we present the Theorem 5. (The proof can be found in the Appendix.) We use $\beta_{t,i}$, u_i, $g_{t,i} \in \mathbb{R}$ to represent the i-th element of the vector correspondingly.

Theorem 5 *Assume $\ell_t(x), t = 1, \ldots, T$, to be convex functions. Set* $\text{Wealth}_0 = \epsilon = \mathbf{1}$, $C = 9$. *For* $i = 1, \ldots, d$, $1/\eta_{1,i} = 2C$ *and assume that* $|g_{t,i}| \leq 1$, $\boldsymbol{g}_t^+ = h_t \boldsymbol{g}_t$ *where* $h_t \in [0, 1]$. *Then, for all* $\boldsymbol{u} \in \mathbb{R}^d$, *Algorithm 3 guarantees*

$$
\text{Regret}_T(\boldsymbol{u}) = \sum_{i=1}^d \epsilon + \mathcal{O}\left(\max\left\{ |u_i| \ln\left(|u_i|\left(1 + \sum_{t=1}^T |g_{t,i}||g_{t,i}^+| \right) \right), \right.\right.
$$
$$
\left.\left. |u_i|\sqrt{ \sum_{t=1}^T (2|g_{t,i}||g_{t,i}^+| - |g_{t,i}^+|^2) \cdot \ln\left(1 + |u_i| \sum_{t=1}^T (2|g_{t,i}||g_{t,i}^+| - |g_{t,i}^+|^2) \right)} \right\}\right) .
$$

To see how this bound is more desirable than Theorem 1, first notice that the bound obtains some adaptivity to L_1 geometry: if $G_\infty = \max \|g_t\|_\infty$, then the bound is at most $\tilde{\mathcal{O}}(\|u\|_1 G_\infty \sqrt{T})$. Furthermore, by application of Cauchy-Schwarz (twice!) we can see that the bound is at most an additive ϵd larger than Theorem 1 - and could be much smaller if either application of Cauchy-Schwarz is loose. Thus, this bound is never much worse than that of Theorem 1, but has the further desirable property that one can add a large number of "irrelevant" dimensions for which $u_i = 0$ without harming the bound.

Unfortunately, we no longer have a closed form expression for h_t due to the coupling of the coordinates. However, at each iteration, we can find a δ-approximation to h_t using $\mathcal{O}(\log(1/\delta))$
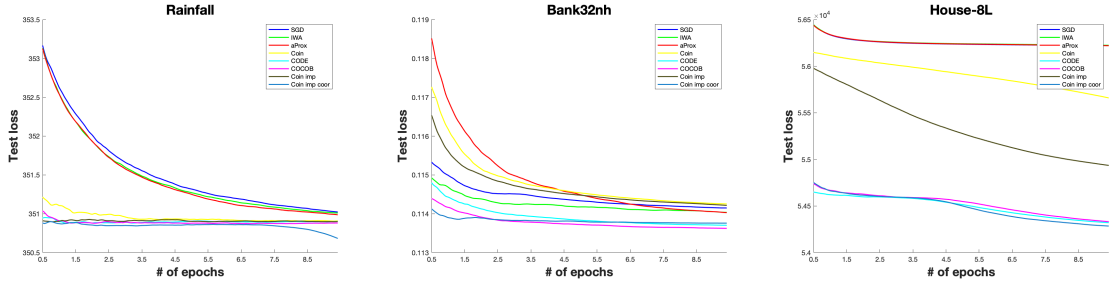
Figure 3: Regression tasks: Mean test loss versus epochs.

steps of bisection and a single gradient oracle call. By "bisection" we mean a binary-search style algorithm: given a guess for h_t, we can compute if the true value is lower or higher than the guess by computing what the update would be if the guess were correct and checking if we have overshot the corner of the truncated bound. Though this inflates the cost of an update by $\mathcal{O}(\log(1/\delta))$, this is still *significantly* more efficient than the $\text{poly}(T)$ oracle calls required to run FTRL with truncated linear models directly at each step. This small extra computation cost is a price we pay for better theoretical as well as empirical results.

8. Empirical Evaluation

While our main contribution is theoretical, here we evaluate the empirical performance of Algorithm 2 and Algorithm 3 to show their practical potential. We will denote the algorithms as Implicit Coin, and Coordinate-wise Implicit Coin. We would also like to stress that we used Algorithm 2 and Algorithm 3 as they are, with the choice of the hyperparameters directly given by theory, i.e., η_0, C, Wealth_0. It is quite possible that these choices were not optimal. We do this on purpose: we want to demonstrate how robust parameter-free algorithms are, even with theory-derived constants.

We compare SGD, SGD with truncated models (aProx) (Asi and Duchi, 2019), SGD with Importance Weight Aware updates (IWA) (Karampatziakis and Langford, 2011), Coin-betting algorithm (Coin) (Orabona and Pál, 2016), Coin-betting with ODE updates (CODE) (Chen et al., 2022), COntinuous COin Betting (COCOB) (Orabona and Tommasi, 2017).

We tested the algorithms on real-world datasets from the LIBSVM website (Chang and Lin, 2001) and OpenML (Vanschoren et al., 2013). 2dPlane, CPU-act, and Houses are classification tasks, Rainfall, Bank32nh, and House-8L are regression tasks. (More information about datasets is in Appendix). We standardize and pre-process the samples, normalizing them to unit norm vectors. We shuffle the data and separate into a training set (70%), validation set (15%), and test set (15%).

For SGD, aProx and IWA, we tune the initial step size η_0 and consider stepsize sequence of the form: $\eta_k = \eta_0/\sqrt{k}$. All the parameter-free algorithms do not have parameters to tune. Implicit Coin has a closed form solution for h_t, so the implementation is as efficient as SGD. For CODE and Coordinate-wise Implicit Coin, we used a bisection algorithm to find h_t.

We considered linear predictors trained with the hinge loss for classification, and with the absolute loss for regression. We trained all algorithms for 10 epochs over the training data. Each epoch requires running through the total shuffled training set. All the experiments are repeated 3 times, we report the average of the 3 repetitions. For all the algorithms that require a learning rate, on every
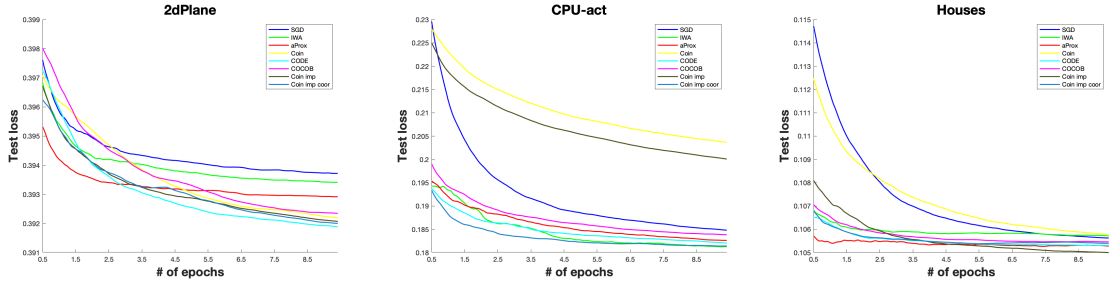
Figure 4: Classification tasks: Mean test loss versus epochs.

dataset and for each repetition, we use the validation loss to choose the best learning rate, train using that learning rate, test on the test set, and report the test loss averaged over 3 repetitions.

Figure 3 and Figure 4 show the average of test losses versus number of epochs. In all experiments, Coordinate-wise Implicit Coin has a performance that is superior or comparable to all the competitors. It is worth remembering that the algorithms with a learning rate were tuned on the validation set, while the parameter-free algorithms were not tuned in any way. Moreover, in all cases, Implicit Coin has a better performance than Coin. Given that their key difference is the truncated models in Implicit coin, this directly supports the advantage of these updates over linearized ones.

More in detail, on regression tasks (Figures 3), the performance of Coordinate-wise Implicit Coin is superior to the other competitors at the end of the training on Rainfall and House-8L. COCOB, CODE, and Coordinate-wise Implicit Coin outperform the other competitors significantly on Houses-8L, and their performances are close to each other on Bank32nh. On the classification tasks (Figure 4), Coordinate-wise Implicit Coin and Implicit Coin achieve essentially the optimal performance on CPU-act and Houses correspondingly. On 2dPlane, Coordinate-wise Implicit Coin, Implicit Coin, and CODE perform almost equally satisfying, and they outperform the other competitors.

9. Conclusion

In this paper, we present new parameter-free algorithms utilizing a better convex lower bound: the truncated linear model. *We overcome the theoretical difficulties of using truncated linear models in parameter-free algorithms with a new regret decomposition.* Our regret bounds are analogous to bounds achieved by implicit methods. Besides, we propose a variant of our algorithm that has a very efficient closed form update rule, matching the runtime of gradient descent. Finally, we provide a coordinate-wise variant with tighter regret bounds.

In the future, we would like to explore the additional possibilities offered by the new decomposition of regret. For example, we would like to overcome the limitation of the current per-coordinate formulation and explore the possibility to design a per-coordinate variant that uses truncated linear models, retaining the closed form update. Besides, considering the good empirical performance of CODE and the similarity in the spirit of CODE and Implicit Coin, we are interested in the possibility of studying the theoretical properties of CODE.

13

Acknowledgments

References

J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 414–424. Omnipress, 2008a.

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

H. Asi and J. C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

N. Campolongo and F. Orabona. Temporal variability in implicit online learning. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

K. Chen, J. Langford, and F. Orabona. Better parameter-free stochastic optimization with ODE updates for coin-betting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

L. Chen, H. Luo, and C.-Y. Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Proc. of the Conference on Learning Theory*, 2021.

A. Cutkosky. *Algorithms and Lower Bounds for Parameter-free Online Learning*. PhD thesis, Stanford University, 2018.

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

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

A. Cutkosky and T. Sarlos. Matrix-free preconditioning in online learning. In *International Conference on Machine Learning*, pages 1455–1464. PMLR, 2019.

D. J. Foster, A. Rakhlin, and K. Sridharan. Online learning: Sufficient statistics and the Burkholder method. In *Proc. of the Conference on Learning Theory*, 2018.

G. J. Gordon. Regret bounds for prediction problems. In *Proc. of the twelfth annual conference on Computational learning theory (COLT)*, pages 29–40, 1999.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. In *Proc. of the 21st Conference on Learning Theory*, 2008.

K.-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Proc. of the Conference on Learning Theory (COLT)*, 2019.

N. Karampatziakis and J. Langford. Online importance weight aware updates. In *Proc. of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 392–399, Arlington, Virginia, USA, 2011. AUAI Press.

M. Kempka, W. Kotłowski, and M. K. Warmuth. Adaptive scale-invariant online algorithms for learning linear models. In K. Chaudhuri and R. Salakhutdinov, editors, *Proc. of the 36th International Conference on Machine Learning*, volume 97 of *Proc. of Machine Learning Research*, pages 3321–3330, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, January 1997.

J. Kivinen, M. K. Warmuth, and B. Hassibi. The p-norm generalization of the LMS algorithm for adaptive filtering. *IEEE Transactions on Signal Processing*, 54(5):1782–1793, 2006.

W. Kotłowski. Scale-invariant unconstrained online learning. *Theoretical Computer Science*, 808: 139–158, 2020.

B. Kulis and P. L. Bartlett. Implicit online learning. In *International Conference on Machine Learning*, pages 575–582, 2010.

B. Martinet. Régularisation dinéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

H. B. McMahan. A unified view of regularized dual averaging and mirror descent with implicit updates. *arXiv preprint arXiv:1009.3240*, 2010.

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

Z. Mhammedi and W. M Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, pages 2858–2887. PMLR, 2020.

J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

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

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

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

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

F. Orabona and D. Pál. Parameter-free stochastic optimization of variationally coherent functions. *arXiv preprint arXiv:2102.00236*, 2021.

F. Orabona and T. Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems*, pages 2160–2170, 2017.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

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

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2), 2012.

C. Song, J. Liu, H. Liu, Y. Jiang, and T. Zhang. Fully implicit online learning. *arXiv preprint arXiv:1809.09350*, 2018.

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

D. van der Hoeven. User-specified local differential privacy in unconstrained adaptive online learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

M. K. Warmuth and A. K. Jagota. Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *Electronic proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics*, volume 326, 1997.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. of the International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A. Proof of Theorem 1

Before we prove Theorem 1, we first introduce some technical Lemmas that will be used in the proof.

Lemma 6 *Under the assumptions of Theorem 1 and the notation of Algorithm 1, we have* $\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \leq \frac{3\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^{t} \|\boldsymbol{g}_i\| \|\boldsymbol{g}_i^+\|}$ *for all* $t = 1, \dots, T$.

Proof By the definition of $\boldsymbol{\beta}_{t+1}$, $\hat{\boldsymbol{\beta}}_{t+1}$, $\boldsymbol{\beta}_t$, we have:

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \le \|\hat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}_t\| = \frac{\|\boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|}{3 + 2\sum_{i=1}^{t-1}(\|\boldsymbol{g}_i\|^2 - \|\boldsymbol{g}_i^+ - \boldsymbol{g}_i\|^2)}$$

$$\le \frac{\|\boldsymbol{g}_t^+\| + 2\|\boldsymbol{\beta}_t\|(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)}{3 + 2\sum_{i=1}^{t-1}(2\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\| - \|\boldsymbol{g}_i^+\|^2)} \le \frac{\|\boldsymbol{g}_t^+\| + 2\|\boldsymbol{\beta}_t\|(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)}{3 + 2\sum_{i=1}^{t-1}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|}$$

$$\le \frac{\|\boldsymbol{g}_t^+\| + 2\|\boldsymbol{\beta}_t\|\|\boldsymbol{g}_t^+\|(2\|\boldsymbol{g}_t\| - h_t\|\boldsymbol{g}_t\|)}{1 + 2\sum_{i=1}^{t}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} \le \frac{3\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^{t}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|},$$

where we used the fact that $\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\| \le 1$ in second to last inequality. ∎

Lemma 7 *Let $a_0 \ge 0$ and $f : [0, +\infty) \to [0, +\infty)$ a nonincreasing function. Then,*

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

Proof Denote by $s_t = a_0 + \sum_{i=1}^{t} a_i$.

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

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

Lemma 8 *Let V a non-empty closed convex set in $\mathbb{R}^d$. Assume that the functions $f_t : \mathbb{R} \to (-\infty, \infty]$ are μ_t-strongly convex w.r.t $\|\cdot\|_2$ over $V \subset \cap_{t=1}^{T} \text{int dom } f_t$, where $\mu_t > 0$. Assume we receive subgradients $\boldsymbol{v}_t \in \partial f_t(\boldsymbol{w}_t)$ and set $\boldsymbol{w}_t$ using Online Gradient Descent with stepsizes: $\eta_t = \frac{1}{1/\eta_1 + \sum_{i=1}^{t-1}\mu_i}$. Then, for any $\boldsymbol{u} \in V$, we have the following regret guarantee:*

$$\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})) \le \frac{\|\boldsymbol{w}_1 - \boldsymbol{u}\|^2}{2\eta_1} + \sum_{t=1}^{T}\frac{\eta_t \|\boldsymbol{v}_t\|^2}{2} + \sum_{t=1}^{T}\frac{\mu_t}{2}\|\boldsymbol{w}_{t+1} + \boldsymbol{w}_t - 2\boldsymbol{u}\|\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| .$$

Proof From the strongly convexity of the function f_t, we have that

$$f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \le \langle \boldsymbol{v}_t, \boldsymbol{w}_t - \boldsymbol{u}\rangle - \frac{\mu_t}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\| ,$$

where $\boldsymbol{v}_t \in \partial f_t(\boldsymbol{w}_t)$. From the fact that $\eta_t = \frac{1}{1/\eta_1 + \sum_{i=1}^{t-1}\mu_i}$, we have

$$\frac{1}{2\eta_t} + \frac{\mu_t}{2} = \frac{1}{2\eta_{t+1}}, \qquad t = 1, \ldots, T .$$

Also, observe that

$$
\begin{aligned}
\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 &= \|\boldsymbol{w}_{t+1} - \boldsymbol{u} + \boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \\
&= \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + 2\langle \boldsymbol{w}_{t+1} - \boldsymbol{u}, \boldsymbol{w}_t - \boldsymbol{w}_{t+1}\rangle + \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \\
&= \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \langle 2\boldsymbol{w}_{t+1} - 2\boldsymbol{u} + \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, \boldsymbol{w}_t - \boldsymbol{w}_{t+1}\rangle \\
&\ge \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_{t+1} + \boldsymbol{w}_t - 2\boldsymbol{u}\|\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| .
\end{aligned}
$$

Thus, we have

$$\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})) \le \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2\eta_t}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 - \frac{\mu_t}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 + \frac{\eta_t}{2}\|\boldsymbol{v}_t\|^2\right)$$

$$\le \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2\eta_t}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 - \frac{\mu_t}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2\right.$$

$$\left. + \frac{\mu_t}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|\|\boldsymbol{w}_{t+1} + \boldsymbol{w}_t - 2\boldsymbol{u}\| + \frac{\eta_t}{2}\|\boldsymbol{v}_t\|^2\right)$$

$$= \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2\eta_{t+1}}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2\right)$$

$$+ \sum_{t=1}^{T}\frac{\mu_t}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|\|\boldsymbol{w}_{t+1} + \boldsymbol{w}_t - 2\boldsymbol{u}\| + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\boldsymbol{v}_t\|^2 .$$

Summing and telescoping, we obtain

$$\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})) \le \frac{\|\boldsymbol{w}_1 - \boldsymbol{u}\|^2}{2\eta_1} + \sum_{t=1}^{T}\frac{\mu_t}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|\|\boldsymbol{w}_{t+1} + \boldsymbol{w}_t - 2\boldsymbol{u}\| + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\boldsymbol{v}_t\|^2 .$$

∎

Corollary 9 *Consider an OCO problem with losses $f_t(\boldsymbol{\beta}) = \langle \boldsymbol{g}_t^+, \boldsymbol{\beta}\rangle + \left(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t - \boldsymbol{g}_t^+\|^2\right)\|\boldsymbol{\beta}\|^2$, with $\boldsymbol{\beta} \in B$, $B = \{\boldsymbol{x}|\|\boldsymbol{x}\| \le 1/2\}$. Then, from Lemma 8, we have*

$$\sum_{t=1}^{T}(f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^{\star})) \le 3/2 + 17/4\sum_{t=1}^{T}\frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^{t}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} .$$

Proof From Lemma 8,

$$\sum_{t=1}^{T}(f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^{\star})) \le \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^{\star}\|^2}{2\eta_1} + \sum_{t=1}^{T}\frac{\mu_t}{2}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|\|\boldsymbol{\beta}_{t+1} + \boldsymbol{\beta}_t - 2\boldsymbol{\beta}^{\star}\| + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\boldsymbol{v}_t\|^2 .$$

Observe that $\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}^{\star} \in B$, so $\|\boldsymbol{\beta}_{t+1} + \boldsymbol{\beta}_t - 2\boldsymbol{\beta}^{\star}\| \le 2$. Moreover, we have $\mu_t \le 2\|\boldsymbol{g}_t\|$. So, summing the second term and using Lemma 6, we obtain

$$\sum_{t=1}^{T}\frac{\mu_t}{2}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|\|\boldsymbol{\beta}_{t+1} + \boldsymbol{\beta}_t - 2\boldsymbol{\beta}^{\star}\| \le \sum_{t=1}^{T}\mu_t\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \le 3\sum_{t=1}^{T}\frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^{t}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} .$$

Noting that $\boldsymbol{v}_t \in \partial f_t(\boldsymbol{\beta}_t) = \boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)$, we have

$$\|\boldsymbol{v}_t\|^2 \le \|\boldsymbol{g}_t^+\|^2 + 4\|\boldsymbol{\beta}_t\|\|\boldsymbol{g}_t^+\|\|\boldsymbol{g}_t\| + 4\|\boldsymbol{\beta}_t\|^2\|\boldsymbol{g}_t\|(2\|\boldsymbol{g}_t^+\|) \le 5\|\boldsymbol{g}_t^+\|\|\boldsymbol{g}_t\| .$$

Hence, summing the third term, we have

$$\sum_{t=1}^{T}\frac{\eta_t}{2}\|\boldsymbol{v}_t\|^2 \le \frac{5}{4}\sum_{t=1}^{T}\frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^{t}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} .$$

18

Now using the fact that $\frac{1}{\eta_1} = 3$, we obtain

$$\sum_{t=1}^{T} f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^\star) \le 3/2 + \frac{17}{4} \sum_{t=1}^{T} \frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^{t} \|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} \ .$$

∎

Lemma 10 *Consider an OCO problem with losses $f_t(\boldsymbol{\beta}) = \langle \boldsymbol{g}_t^+, \boldsymbol{\beta} \rangle + \frac{\mu_t}{2}\|\boldsymbol{\beta}\|^2$ and $\mu_t = 2\left(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t - \boldsymbol{g}_t^+\|^2\right)$, with $\boldsymbol{\beta} \in B = \{\boldsymbol{x}|\|\boldsymbol{x}\| \le r\}$. Define $\boldsymbol{\beta}_T^* := \arg\min_{\boldsymbol{\beta} \in B} \sum_{t=1}^{T} f_t(\boldsymbol{\beta})$. Then we have*

$$\sum_{t=1}^{T} f_t(\boldsymbol{\beta}_T^\star) \le \max\left\{ \frac{-r\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{2}, \frac{-\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t} \right\} \ .$$

Proof We consider 2 cases: By definition $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_{T+1} = \prod_B(\hat{\boldsymbol{\beta}}_{T+1})$, f_t is μ_t strongly convex, where $\mu_t = 2(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)$.

- If $\hat{\boldsymbol{\beta}}_{T+1} \in B$, that is $\boldsymbol{\beta}_{T+1} = \hat{\boldsymbol{\beta}}_{T+1} = \frac{-\sum_{t=1}^{T}\boldsymbol{g}_t^+}{\sum_{t=1}^{T}\mu_t}$. In this case,

$$\sum_{t=1}^{T} f_t(\boldsymbol{\beta}^*) = \left(\sum_{t=1}^{T}\boldsymbol{g}_t^+\right)\frac{-\sum_{t=1}^{T}\boldsymbol{g}_t^+}{\sum_{t=1}^{T}\mu_t} + \frac{\left\|-\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{\left(\sum_{t=1}^{T}\mu_t\right)^2}\frac{\sum_{t=1}^{T}\mu_t}{2} = -\frac{1}{2}\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{\sum_{t=1}^{T}\mu_t} \ .$$

- If $\hat{\boldsymbol{\beta}}_{T+1} \notin B$, $\boldsymbol{\beta}_{T+1} = \prod_B(\hat{\boldsymbol{\beta}}_{T+1}) = r\frac{-\sum_{t=1}^{T}\boldsymbol{g}_t^+}{\|-\sum_{t=1}^{T}\boldsymbol{g}_t^+\|}$. In this case,

$$\|\hat{\boldsymbol{\beta}}_{T+1}\| = \left\|\frac{-\sum_{t=1}^{T}\boldsymbol{g}_t^+}{\sum_{t=1}^{T}\mu_t}\right\| \ge r \ .$$

That is,

$$\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\| \ge r\sum_{t=1}^{T}\mu_t \ .$$

Therefore, we have

$$\sum_{t=1}^{T} f_t(\boldsymbol{\beta}^*) = -r\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\| + r^2\sum_{t=1}^{T}\frac{\mu_t}{2}\left(\frac{\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\|}{\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\|}\right)^2 \le -r\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\| + \frac{r^2}{2}\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{r}$$

$$= -\frac{r}{2}\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\| \ .$$

In conclusion,

$$\sum_{t=1}^{T} f_t(\boldsymbol{\beta}^\star) \le \max\left\{ \frac{-r\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{2}, \frac{-\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t} \right\} \ .$$

∎

Lemma 11 *Let $\|\boldsymbol{g}_t\| \le 1$, $\boldsymbol{g}_t^+ = h_t \boldsymbol{g}_t$ where $h_t \in [0,1]$, for $t = 1, \ldots, T$. Then, we have*

$$\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle) \ge -\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|\boldsymbol{\beta}_t\|^2$$
$$- 2\|\boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| .$$

Proof

$$\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle)$$
$$\ge \ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle + \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|)$$
$$= \ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln\left(1 + \frac{\|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|}{1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle}\right)$$
$$\ge \ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \frac{\|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|}{1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle}$$
$$\ge \ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \frac{\|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|}{1 - 1/2}$$
$$\ge \ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - 2\|\boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|,$$

where in the last inequality we used the fact that $\boldsymbol{g}_t^+ = h_t \boldsymbol{g}_t$, $0 \le h_t \le 1$, $\|\boldsymbol{\beta}_t\| \le 1/2$. The sum of these last terms is upper bounded by a logarithmic term.

Now, considering the first two terms, we have

$$\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) = f(\langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle),$$

where $f(x) = \ln(1 - x) - \ln(1 + (h_t - 1)x)$ for $|x| \le 1/2$.

To have the tightest inequality, we consider two cases separately.

Case $0 \le x \le 1/2$. We consider $g(x) = f(x) - h_t x$. We have that the first derivative of g is negative for $0 \le x \le 1/2$. Moreover, the function $\phi(x) = x^2(2h_t - h_t^2)$ is increasing for $0 \le x \le 1/2$. Hence, for $0 \le x \le 1/2$, we have

$$\frac{\ln(1 - x) - \ln(1 + (h_t - 1)x) - h_t x}{x^2(2h_t - h_t^2)} = \frac{g(x)}{\phi(x)} \ge \frac{g(1/2)}{\phi(1/2)} = \frac{-4\ln(1 + h_t) - 2h_t}{2h_t - h_t^2} \ge -1,$$

where in the last inequality we used the fact that $0 \le h_t \le 1$.

Case $-1/2 \le x < 0$. Here, we lower bound f using a Taylor expansion:

$$f(x) = f(0) + x f'(0) + \frac{x^2}{2} f''(y),$$

where y is between 0 and x. Denoting $a = h_t - 1$, we have

$$f(x) = -x h_t - \frac{x^2}{2} \frac{(a+1)(2ay - a + 1)}{(y-1)^2(ay+1)^2} .$$

Now, consider the quadratic term in the above expression. Dividing and multiplying by $1 - a \ge 0$, we have

$$\frac{(1 - a^2)(2ay - a + 1)}{(1 - a)(y - 1)^2(ay + 1)^2} .$$

20

We are interested in keeping the term $1 - a^2$ and upper bounding the rest. Given that we are considering the case $-1/2 \le x < 0$, we have $-1/2 \le y < 0$. So, we have

$$2ay - a + 1 \le 1 - a - y + ay = (1 - a)(1 - y) \,.$$

and

$$\frac{2ay - a + 1}{(1 - a)(y - 1)^2(ay + 1)^2} \le \frac{(1 - a)(1 - y)}{(1 - a)(y - 1)^2(ay + 1)^2} = \frac{1}{(1 - y)(ay + 1)^2} \le 1 \,.$$

Hence, putting all together, we have

$$\begin{aligned}
\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) &\ge -h_t \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle - (1 - a^2)(\langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle)^2 \\
&\ge -h_t \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle - (1 - a^2)\|\boldsymbol{g}_t\|^2\|\boldsymbol{\beta}_t\|^2 \\
&= -\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t\|^2(h_t - 1)^2)\|\boldsymbol{\beta}_t\|^2 \\
&= -\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|\boldsymbol{\beta}_t\|^2 \,.
\end{aligned}$$

$\blacksquare$

Now, we present the full proof of Theorem 1. Note that a part of the proof of Theorem 3 is similar to the proof of Theorem 1.

Proof [Proof of Theorem 1] From the (2), we get

$$\text{Wealth}_t = \text{Wealth}_{t-1} - \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle - \langle \boldsymbol{g}_t^+, \boldsymbol{w}_{t+1} \rangle \,.$$

Using the fact that the algorithm predicts with $\boldsymbol{w}_t = \boldsymbol{\beta}_t \, \text{Wealth}_{t-1}$, we obtain

$$\text{Wealth}_t = \text{Wealth}_{t-1} \frac{1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle}{1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle} \,.$$

This implies that $\ln \text{Wealth}_T = \ln \epsilon + \sum_{t=1}^T \ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle)$.

Using Lemma 11 and Lemma 6, we have

$$\begin{aligned}
\sum_{t=1}^T &\left(\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle) \right) \\
&\ge \sum_{t=1}^T \left[-\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|\boldsymbol{\beta}_t\|^2 - 2\|\boldsymbol{g}_t\|\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \right] \\
&\ge \sum_{t=1}^T \left[-\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\|\boldsymbol{\beta}_t\|^2 - \frac{6\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1 + 2\sum_{i=1}^t \|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} \right] \\
&\ge -3\ln\left(1 + 2\sum_{t=1}^T \|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right) - \sum_{t=1}^T f_t(\boldsymbol{\beta}_t) \,.
\end{aligned}$$

Where f_t is as defined in Corollary 9. Applying Corollary, and defining μ_t as in Lemma 10, we have

$$\geq -3/2 - 7.25 \ln\left(1 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right) - \sum_{t=1}^{T} f_t(\boldsymbol{\beta}^\star)$$

$$\geq -3/2 - 7.25 \ln\left(1 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right) - \max\left\{\frac{-\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{4}, \frac{-\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t}\right\}$$

$$= -3/2 - 7.25 \ln\left(1 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right) + \min\left\{\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{4}, \frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t}\right\}.$$

The last inequality is from Lemma 10. Next, we perform a case analysis to derive the upper bound for Regret_T.

If $\min\left\{\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{4}, \frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t}\right\} = \frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{4}$, then

$$\text{Wealth}_T \geq \frac{e^{-3/2}}{\left(1 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right)^{7.25}} \exp\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{4}.$$

By Cutkosky and Orabona (2018, Lemma 18), we have

$$\text{Regret}_T(\boldsymbol{u}) \leq 4\|\boldsymbol{u}\|\left(\ln\frac{4\|\boldsymbol{u}\|\left(1 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right)^{7.25}}{e^{-3/2}} - 1\right).$$

If $\min\left\{\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|}{4}, \frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t}\right\} = \frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t}$, then

$$\text{Wealth}_T \geq \frac{e^{-3/2}}{\left(1 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right)^{7.25}} \exp\frac{\left\|\sum_{t=1}^{T}\boldsymbol{g}_t^+\right\|^2}{2\sum_{t=1}^{T}\mu_t}.$$

By Orabona and Tommasi (2017, Lemma 1), we have

$$\text{Regret}_T(\boldsymbol{u}) \leq -\frac{e^{-3/2}}{\left(1 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right)^{7.25}}$$

$$+ \|\boldsymbol{u}\|\sqrt{2\sum_{t=1}^{T}(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)\ln\left(1 + \frac{\left(\sum_{t=1}^{T}\mu_t\right)\|\boldsymbol{u}\|^2\left(1 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right)^{14.5}}{e^{-3}}\right)}.$$

Finally, combining the two cases result gives the regret bound

$$\text{Regret}_T(\boldsymbol{u}) = \mathcal{O}\left(\max\left\{\|\boldsymbol{u}\|\ln\left(\|\boldsymbol{u}\|\left(1 + \sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right)\right),\right.\right.$$
$$\left.\left.\|\boldsymbol{u}\|\sqrt{\sum_{t=1}^{T}(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2) \cdot \ln\left(1 + \|\boldsymbol{u}\|\sum_{t=1}^{T}(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)\right)}\right\}\right).$$

$\blacksquare$

Appendix B. Proof of Theorem 3

In this section, we present the proofs of Theorem 3 and the Lemmas required for its proof.

Lemma 13 shows that the regret of $\phi_t(\boldsymbol{\beta}_t)$ upper bounds the regret of $f_t(\boldsymbol{\beta}_t)$, and Lemma 12 gives the upper bound of the regret of running $\phi_t(\boldsymbol{\beta}_t)$. Thus we are able to obtain similar results as in Corollary 9. Lemma 14 proves that Algorithm 2 guarantees $\|\boldsymbol{\beta}_t\| \leq 1/2$ which removes the projection step, and leads to the closed form updates.

Lemma 12 *For $t = 1, \ldots, T$, $\boldsymbol{\beta}_t$ are outputs of running OGD with stepsizes: $\eta_t = \frac{1}{1/\eta_1 + \sum_{i=1}^{t-1}\mu_i}$ on the strongly convex losses $\phi_t(\boldsymbol{\beta})$ defined as following:*

- *If $\|\boldsymbol{\beta}_t\| < \frac{3}{8}$, $\phi_t(\boldsymbol{\beta}) = f_t(\boldsymbol{\beta})$,*

- *If $\frac{3}{8} \leq \|\boldsymbol{\beta}_t\| \leq \frac{1}{2}$, $\phi_t(\boldsymbol{\beta}) = C\|\boldsymbol{g}_t^+\|\|\boldsymbol{\beta}\|^2$,*

where $\boldsymbol{\beta}_t \in B$ and $B = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq \frac{1}{2}\}$. $\phi_t(\boldsymbol{\beta})$ is μ_t strongly convex. $\boldsymbol{\beta}_1^\star \triangleq \arg\min_{\boldsymbol{\beta}\in B_1}\sum_{t=1}^{T}f_t(\boldsymbol{\beta})$, where $B_1 = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq \frac{1}{4}\}$. Then, for $C \geq 3$ we have

$$\sum_{t=1}^{T}(\phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star)) \leq -\frac{5C^2}{4}\ln(1/\eta_1 - 2) + \frac{1}{2\eta_1} + \frac{5C^2}{4}\ln\left(1/\eta_1 - 2 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right).$$

Proof Applying Lemma 8, we have

$$\sum_{t=1}^{T}(\phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star)) \leq \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^\star\|^2}{2\eta_1} + \sum_{t=1}^{T}\left(\mu_t\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| + \frac{\eta_t\|\nabla\phi_t(\boldsymbol{\beta}_t)\|^2}{2}\right).$$

We now consider the two possible cases for h_t.

If $\phi_t(\boldsymbol{\beta}) = f_t(\boldsymbol{\beta})$, then

$$\|\nabla\phi_t(\boldsymbol{\beta}_t)\| = \|\boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2)\| \leq 3\|\boldsymbol{g}_t^+\|.$$

Hence, $\|\nabla\phi_t(\boldsymbol{\beta}_t)\|^2 \leq 9\|\boldsymbol{g}_t^+\|^2 \leq 9\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|$. Moreover, $\mu_t = 2(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2) \geq 2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|$.

If $\phi_t(\boldsymbol{\beta}) = C\|\boldsymbol{g}_t^+\|\|\boldsymbol{\beta}_t\|^2$, then

$$\|\nabla\phi_t(\boldsymbol{\beta}_t)\| = \|2C\boldsymbol{\beta}_t\|\boldsymbol{g}_t^+\|\| \leq C\|\boldsymbol{g}_t^+\|.$$

Hence, $\|\nabla\phi_t(\boldsymbol{\beta}_t)\|^2 = C^2\|\boldsymbol{g}_t^+\|^2 \leq C^2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|$. Moreover, assuming $C \geq 1$, we have $\mu_t = 2C\|\boldsymbol{g}_t^+\| \geq 2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|$.

Consequently, with $C \geq 3$, we have

$$\sum_{t=1}^{T} \frac{\eta_t\,\|\nabla\phi_t(\boldsymbol{\beta}_t)\|^2}{2} \leq \frac{1}{2}\sum_{t=1}^{T}\frac{C^2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1/\eta_1 + \sum_{i=1}^{t-1}\mu_i} \leq \frac{C^2}{4}\sum_{t=1}^{T}\frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1/\eta_1 + 2\sum_{i=1}^{t-1}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|}\ .$$

Applying Lemma 15 and $C \geq 3$, we thus obtain

$$\sum_{t=1}^{T}\phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star) \leq \frac{1}{2\eta_1} + \left(\max\left(C^2,3\right) + \frac{C^2}{4}\right)\sum_{t=1}^{T}\frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1/\eta_1 + 2\sum_{i=1}^{t-1}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|}$$

$$\leq \frac{1}{2\eta_1} + \left(\max\left(C^2,3\right) + \frac{C^2}{4}\right)\sum_{t=1}^{T}\frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1/\eta_1 - 2 + 2\sum_{i=1}^{t}\|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|}$$

$$\leq \frac{1}{2\eta_1} + \left(\max\left(C^2,3\right) + \frac{C^2}{4}\right)\left(\ln\left(1/\eta_1 - 2 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right) - \ln(1/\eta_1 - 2)\right)$$

$$= \frac{5C^2}{4}\ln\left(1/\eta_1 - 2 + 2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|\right) + \frac{1}{2\eta_1} - \frac{5C^2}{4}\ln(1/\eta_1 - 2)\ .$$

$\blacksquare$

Lemma 13 *Under conditions and notations of Lemma 12 and for any $C \geq 9$, we have*

$$f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}_1^\star) \leq \phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star)\ .$$

Proof We consider the two cases for the losses h_t.

When $\frac{3}{8} \leq \|\boldsymbol{\beta}_t\| \leq \frac{1}{2}$, we have $\phi_t(\boldsymbol{\beta}_t) = C\|\boldsymbol{g}_t^+\|\|\boldsymbol{\beta}_t\|^2$. Hence, we have

$$\phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star) = C\|\boldsymbol{g}_t^+\|\left(\|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2\right)\ .$$

Moreover,

$$f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}_1^\star) = \langle\boldsymbol{g}_t^+, \boldsymbol{\beta}_t - \boldsymbol{\beta}_1^\star\rangle + (2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)\|\boldsymbol{\beta}_t\|^2 - (2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)\|\boldsymbol{\beta}_1^\star\|^2$$

$$\leq \|\boldsymbol{g}_t^+\|(\|\boldsymbol{\beta}_t\| + \|\boldsymbol{\beta}_1^\star\|) + \|\boldsymbol{g}_t^+\|\|\boldsymbol{g}_t\|(2 - h_t)(\|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2)$$

$$\leq \|\boldsymbol{g}_t^+\|(\|\boldsymbol{\beta}_t\| + \|\boldsymbol{\beta}_1^\star\|) + \|\boldsymbol{g}_t^+\|(\|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2)\ .$$

Hence, we have $f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}_1^\star) \leq \phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star)$ iff

$$\|\boldsymbol{\beta}_t\| + \|\boldsymbol{\beta}_1^\star\| + \|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2 \leq C\left(\|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2\right),$$

that is

$$C \geq \frac{\|\boldsymbol{\beta}_t\| + \|\boldsymbol{\beta}_1^\star\| + \|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2}{\|\boldsymbol{\beta}_t\|^2 - \|\boldsymbol{\beta}_1^\star\|^2} = \frac{1}{\|\boldsymbol{\beta}_t\| - \|\boldsymbol{\beta}_1^\star\|} + 1\ .$$

Using the fact that $\frac{1}{\|\boldsymbol{\beta}_t\| - \|\boldsymbol{\beta}_1^\star\|} + 1 \leq 9$, gives the stated value for C. When $\|\boldsymbol{\beta}_t\| \leq \frac{3}{8}$, we have $\phi_t(\boldsymbol{\beta}) = f_t(\boldsymbol{\beta})$ and $f_t(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}_1^\star) \leq \phi_t(\boldsymbol{\beta}_t) - \phi_t(\boldsymbol{\beta}_1^\star)$ is trivially true. $\blacksquare$

Lemma 14 *Under the assumptions of Theorem 3 and the notation of Algorithm 2, let $1/\eta_1 \geq \max(2C, 16)$. Then, $\|\boldsymbol{\beta}_t\| \leq \frac{1}{2}$ for all $t = 1, \ldots, T$.*

Proof We prove this by induction: **Base case**: $\boldsymbol{\beta}_0 = \mathbf{0}$, $\|\boldsymbol{\beta}_0\| \leq \frac{1}{2}$ is trivially true.
Induction step: Suppose $\|\boldsymbol{\beta}_t\| \leq \frac{1}{2}$, we prove $\|\boldsymbol{\beta}_{t+1}\| \leq \frac{1}{2}$. We perform the following case analysis:
If $\frac{3}{8} \leq \|\boldsymbol{\beta}_t\| \leq \frac{1}{2}$,

$$\|\boldsymbol{\beta}_{t+1}\| = \left\| \boldsymbol{\beta}_t - \frac{2C\boldsymbol{\beta}_t \|\boldsymbol{g}_t^+\|}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i} \right\| \leq \|\boldsymbol{\beta}_t\| \left(1 - \frac{2C\|\boldsymbol{g}_t^+\|}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i} \right) \leq \|\boldsymbol{\beta}_t\| \leq \frac{1}{2} .$$

If $\|\boldsymbol{\beta}_t\| < \frac{3}{8}$

$$\|\boldsymbol{\beta}_{t+1}\| = \left\| \boldsymbol{\beta}_t - \frac{\boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i} \right\| \leq \|\boldsymbol{\beta}_t\| + \left\| \frac{\boldsymbol{g}_t^+ + 2\boldsymbol{\beta}_t(2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| - \|\boldsymbol{g}_t^+\|^2)}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i} \right\|$$
$$\leq \frac{3}{8} + 2\eta_1 \leq \frac{1}{2} .$$

$\blacksquare$

Lemma 15 *Under the assumptions of Theorem 3 and the notation of Algorithm 2, we have, for any $C \geq 1$, we have*

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \leq \max(C, 3) \frac{\|\boldsymbol{g}_t^+\|}{1/\eta_1 + 2\sum_{i=1}^{t-1} \|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} .$$

and

$$\mu_t \|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \leq \max(C^2, 3) \frac{2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|}{1/\eta_1 + 2\sum_{i=1}^{t-1} \|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} .$$

Proof We consider the two cases for h_t.

- If $\phi_t(\boldsymbol{\beta}) = C\|\boldsymbol{g}_t^+\|\|\boldsymbol{\beta}\|^2$, we have $\mu_t = 2C\|\boldsymbol{g}_t^+\|$. Assuming $C \geq 1$ we have $2C\|\boldsymbol{g}_t^+\| \geq 2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|$. Moreover, $\|\nabla\phi_t(\boldsymbol{\beta}_t)\| = 2C\|\boldsymbol{g}_t^+\|\|\boldsymbol{\beta}_t\| \leq C\|\boldsymbol{g}_t^+\|$. So, we have

$$\mu_t\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \leq 2C\|\boldsymbol{g}_t^+\| \frac{C\|\boldsymbol{g}_t^+\|}{1/\eta_1 + \sum_{i=1}^{t-1} \mu_i} \leq C^2 \frac{2\|\boldsymbol{g}_t^+\|\|\boldsymbol{g}_t\|}{1/\eta_1 + 2\sum_{i=1}^{t-1} \|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|} .$$

- If $\phi_t(\boldsymbol{\beta}) = f_t(\boldsymbol{\beta})$, we have

$$2\|\boldsymbol{g}_t\| \geq 2\|\boldsymbol{g}_t\|^2 \geq \mu_t = 2(\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2) = 2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\|(2 - h_t) \geq 2\|\boldsymbol{g}_t\|\|\boldsymbol{g}_t^+\| .$$

  Moreover, $\|\nabla\phi_t(\boldsymbol{\beta}_t)\| \leq 3\|\boldsymbol{g}_t^+\|$. So, we obtain $\mu_t\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \leq 3\frac{2\|\boldsymbol{g}_t^+\|\|\boldsymbol{g}_t\|}{1/\eta_1 + 2\sum_{i=1}^{t-1} \|\boldsymbol{g}_i\|\|\boldsymbol{g}_i^+\|}$.

Taking the maximum of the two cases, we obtain the stated upper bounds. $\blacksquare$

We can now present the proof of Theorem 3.

Proof [Proof of Theorem 3] Using Lemma 11 in the first inequality, and Lemma 15 in the second inequality, we have

$$
\sum_{t=1}^{T} \left(\ln(1 - \langle \boldsymbol{g}_t, \boldsymbol{\beta}_t \rangle) - \ln(1 + \langle \boldsymbol{g}_t^+ - \boldsymbol{g}_t, \boldsymbol{\beta}_{t+1} \rangle) \right)
$$

$$
\geq \sum_{t=1}^{T} \left(-\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2) \|\boldsymbol{\beta}_t\|^2 - 2\|\boldsymbol{g}_t\| \|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| \right)
$$

$$
\geq \sum_{t=1}^{T} \left(-\langle \boldsymbol{g}_t^+, \boldsymbol{\beta}_t \rangle - (\|\boldsymbol{g}_t\|^2 - \|\boldsymbol{g}_t^+ - \boldsymbol{g}_t\|^2) \|\boldsymbol{\beta}_t\|^2 - C \frac{2\|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\|}{1/\eta_1 - 2 + 2\sum_{i=1}^{t} \|\boldsymbol{g}_i\| \|\boldsymbol{g}_i^+\|} \right)
$$

$$
\geq -C \ln \left(1/\eta_1 - 2 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| \right) - \sum_{t=1}^{T} f_t(\boldsymbol{\beta}_t)
$$

$$
\geq -\frac{1}{2\eta_1} + \frac{5C^2}{4} \ln(1/\eta_1 - 2) - \left(\frac{5C^2}{4} + C \right) \ln \left(1/\eta_1 - 2 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| \right) - \sum_{t=1}^{T} f_t(\boldsymbol{\beta}_1^{\star})
$$

$$
\geq -\left(\frac{5C^2}{4} + C \right) \ln \left(1/\eta_1 - 2 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| \right) - \sum_{t=1}^{T} f_t(\boldsymbol{\beta}_1^{\star})
$$

$$
= -110.25 \ln \left(16 + 2\sum_{t=1}^{T} \|\boldsymbol{g}_t\| \|\boldsymbol{g}_t^+\| \right) - \sum_{t=1}^{T} f_t(\boldsymbol{\beta}_1^{\star}),
$$

The second to the last inequality is from Lemma 13 and Lemma 12. $C = 9$ gives rise to the last inequality and the last equation. We now use Lemma 10 to obtain

$$
-\sum_{t=1}^{T} f_t(\boldsymbol{\beta}_1^{\star}) \geq -\max \left\{ \frac{-\left\| \sum_{t=1}^{T} \boldsymbol{g}_t^+ \right\|}{8}, \frac{-\left\| \sum_{t=1}^{T} \boldsymbol{g}_t^+ \right\|^2}{2\sum_{t=1}^{T} \mu_t} \right\} = \min \left\{ \frac{\left\| \sum_{t=1}^{T} \boldsymbol{g}_t^+ \right\|}{8}, \frac{\left\| \sum_{t=1}^{T} \boldsymbol{g}_t^+ \right\|^2}{2\sum_{t=1}^{T} \mu_t} \right\}.
$$

The rest proof is similar to the proof of Theorem 1. Finally, we obtain the same bound as in Theorem 1, up to constants hidden in big O notation.

∎

Appendix C. Proof of Theorem 5

Proof In each coordinate, we perform the regret decomposition as following:

$$
\text{Regret}_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) = \sum_{t=1}^{T} \hat{\ell}_t(\boldsymbol{w}_t) - \hat{\ell}_t(\boldsymbol{w}_{t+1}) + \hat{\ell}_t(\boldsymbol{w}_{t+1}) - \hat{\ell}_t(\boldsymbol{u})
$$

$$
\leq \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + \langle \boldsymbol{g}_t^+, \boldsymbol{w}_{t+1} - \boldsymbol{u} \rangle = \sum_{t=1}^{T} \sum_{i=1}^{d} g_{t,i}(w_{t,i} - w_{t+1,i}) + g_{t,i}^+(w_{t+1,i} - u_i)
$$

$$
= \sum_{i=1}^{d} \sum_{t=1}^{T} g_{t,i}(w_{t,i} - w_{t+1,i}) + g_{t,i}^+(w_{t+1,i} - u_i)
$$

$$
= \sum_{i=1}^{d} u_i \left(-\sum_{t=1}^{T} g_{t,i}^+ \right) - \left(-\sum_{t=1}^{T} (g_{t,i}(w_{t,i} - w_{t+1,i}) + g_{t,i}^+ w_{t+1,i}) \right) ,
$$

where $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$, $\boldsymbol{g}_t \in \partial \hat{\ell}_t(\boldsymbol{w}_t)$, $\boldsymbol{g}_t^+ \in \partial \hat{\ell}_t(\boldsymbol{w}_{t+1})$.

Define $\text{Wealth}_{T,i} = \epsilon - \sum_{t=1}^{T}(g_{t,i}(w_{t,i} - w_{t+1,i}) + g_{t,i}^+ w_{t+1,i})$. Suppose we obtain a bound $\text{Wealth}_{T,i} \geq \psi_T \left(\sum_{i=1}^{T} g_{t,i}^+ \right)$ for some ψ_T. Using the definition of Fenchel conjugate, and the lower bound on the wealth in each coordinate, we have

$$
\text{Regret}_T(\boldsymbol{u}) \leq \sum_{i=1}^{d} u_i \left(-\sum_{t=1}^{T} g_{t,i}^+ \right) - (\text{Wealth}_{T,i} - \epsilon) \leq \sum_{i=1}^{d} \epsilon + u_i \left(-\sum_{t=1}^{T} g_{t,i}^+ \right) - \psi_T \left(\sum_{i=1}^{T} g_{t,i}^+ \right)
$$

$$
\leq \sum_{i=1}^{d} \epsilon + \sup_y y \cdot u_i - \psi_T(y) = \sum_{i=1}^{d} \epsilon + \psi_T^*(u_i) .
$$

In each coordinate, Algorithm 3 is a specific case of running Algorithm 2 with $d = 1$. Applying the result in Theorem 3, we obtain the final bound. ∎

Appendix D. List of Datasets

In our empirical evaluation, we used 3 regression datasets and 3 classification datasets from the LIBSVM website (Chang and Lin, 2001) and OpenML (Vanschoren et al., 2013), randomly selected among the ones with a large number of samples. For the OpenML datasets, categorical features are one-hot-encoded. A short summary of the datasets is in Table 1.

Table 1: Datasets in experiments.

DATASET	TYPE	NUMBER OF SAMPLES	NUMBER OF FEATURES
CPU-ACT	CLASSIFICATION	8192	21
2DPLANE	CLASSIFICATION	40768	10
HOUSES	CLASSIFICATION	20640	8
RAINFALL	REGRESSION	16755	3
BANK32NH	REGRESSION	8192	32
HOUSES-8L	REGRESSION	22784	8