Generalized Implicit Follow-The-Regularized-Leader

Keyi Chen [1] **Francesco Orabona** [1]

Abstract

We propose a new class of online learning algorithms, generalized implicit Follow-The-Regularized-Leader (FTRL), that expands the scope of FTRL framework. Generalized implicit FTRL can recover known algorithms, such as FTRL with linearized losses and implicit FTRL, and it allows the design of new update rules, as extensions of aProx and Mirror-Prox to FTRL. Our theory is constructive in the sense that it provides a simple unifying framework to design updates that directly improve the worst-case upper bound on the regret. The key idea is substituting the linearization of the losses with a Fenchel-Young inequality. We show the flexibility of the framework by proving that some known algorithms, like the Mirror-Prox updates, are instantiations of the generalized implicit FTRL. Finally, the new framework allows us to recover the temporal variation bound of implicit OMD, with the same computational complexity.

1. Introduction

Online learning is a setting where the learner receives an arbitrary sequence of loss functions, selects points before knowing the loss functions, and is evaluated on the values of the loss functions on the points it selects (Cesa-Bianchi & Lugosi, 2006; Orabona, 2019; Cesa-Bianchi & Orabona, 2021). More in detail, at round t the learner outputs a point $\boldsymbol{x}_t$ in a feasible set $V \subseteq \mathbb{R}^d$. Then, it receives a loss function $\ell_t : V \to \mathbb{R}$ and it pays the value $\ell_t(\boldsymbol{x}_t)$. Given the arbitrary nature of the losses, the learner cannot guarantee to have a small cumulative loss, $\sum_{t=1}^T \ell_t(\boldsymbol{x}_t)$. On the other hand, it is possible to minimize the *regret*, that is the difference between the cumulative loss of the algorithm and the one of

any arbitrary comparator $\boldsymbol{u} \in V$:

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

In particular, a successful online learning algorithm must guarantee a regret that grows sublinearly in time for any $\boldsymbol{u} \in V$. In this way, its average performance approaches the one of the best comparator in hindsight.

There are two families of online learning algorithms: Online Mirror Descent (OMD) (Nemirovskij & Yudin, 1983; Warmuth & Jagota, 1997) and Follow-the-Regularized-Leader (FTRL) (Shalev-Shwartz, 2007; Abernethy et al., 2008; Hazan & Kale, 2008). They stem from two similar but complementary approaches: the update of OMD aims at minimizing a linearization of the current loss without going too far from its previous prediction $\boldsymbol{x}_t$, while FTRL minimizes the sum of all the losses (or their linear approximation) plus a regularization term. On the contrary to the first approaches in online learning that focused on specific algorithms (e.g., the Winnow algorithm (Littlestone, 1988)), the theory of these two frameworks is particularly interesting because it allows *both the design and the analysis of generic online learning algorithms*.

While FTRL and OMD provide similar bounds in most situations, they are not completely equivalent. For example, FTRL has an advantage over OMD in unbounded domains, where it allows to use time-varying regularizers. In fact, OMD allows the use of time-varying stepsizes only in domains where its associated Bregman divergence is bounded. On the other hand, in the cases where we can use time-varying stepsizes, OMD can achieve a superior adaption to the gradients (see, e.g., Theorem 2 in Streeter & McMahan (2010) versus Theorem 2 in Orabona & Pál (2015)). In this view, these two frameworks are complementary.[1] Moreover, there exists another orthogonal axis on the use of the actual loss functions or a linear surrogate for both frameworks. We summarize all the variants of OMD and FTRL in Table 1.

Our motivation stems from the fact that in practical cases, all the variants that use full losses offer a big advantage in terms of empirical performance at the cost of a higher

---

[1]Boston University, Boston, MA, USA. Correspondence to: Keyi Chen <keyichen@bu.edu>, Francesco Orabona <francesco@orabona.com>.

---

[1]See also the blog post on this topic by Tim van Erven at https://www.timvanerven.nl/blog/ftrl-vs-omd/.

Table 1. Summary of implicit and linearized updates for FTRL and OMD. (The Bregman divergence $B_\psi(\boldsymbol{x}; \boldsymbol{y})$ is defined as $\psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \nabla\psi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle$. The * denotes the Fenchel conjugate.)

Algorithm	Update
OMD (Warmuth & Jagota, 1997)	$\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \; B_\psi(\boldsymbol{x}; \boldsymbol{x}_t) + \eta_t(\ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t\rangle)$
Implicit OMD (Warmuth & Jagota, 1997)	$\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \; B_\psi(\boldsymbol{x}; \boldsymbol{x}_t) + \eta_t \ell_t(\boldsymbol{x})$
FTRL (linearized) (Abernethy et al., 2008)	$\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \; \psi_{t+1}(\boldsymbol{x}) + \sum_{i=1}^{t}(\ell_i(\boldsymbol{x}_i) + \langle \boldsymbol{g}_i, \boldsymbol{x} - \boldsymbol{x}_i\rangle)$
FTRL (full losses) (McMahan, 2017)	$\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \; \psi_{t+1}(\boldsymbol{x}) + \sum_{i=1}^{t} \ell_i(\boldsymbol{x})$
Implicit FTRL (McMahan, 2010)	$\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \; \psi_{t+1}(\boldsymbol{x}) + \ell_t(\boldsymbol{x}) + \sum_{i=1}^{t-1}(\ell_i(\boldsymbol{x}_i) + \langle \boldsymbol{g}_i, \boldsymbol{x} - \boldsymbol{x}_i\rangle)$
Generalized Implicit FTRL [This work]	$\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \; \psi_{t+1}(\boldsymbol{x}) + \sum_{i=1}^{t} \langle \boldsymbol{z}_i, \boldsymbol{x}\rangle$ $\boldsymbol{z}_i$ such that $\psi_{i+1,V}^\star(\sum_{j=1}^{i} \boldsymbol{z}_j) + \ell_i^\star(\boldsymbol{z}_i) \le \psi_{i+1,V}^\star(\sum_{j=1}^{i-1} \boldsymbol{z}_j - \boldsymbol{g}_i) + \ell_i^\star(\boldsymbol{g}_i)$

computational complexity. On the theoretical side, the situation is not so clear given that in the worst case using the full losses can be equivalent to their linearized version, as it should be clear considering linear losses. In particular, the standard theoretical framework for FTRL does not allow a clear analysis of the implicit case. Moreover, while for implicit OMD it has been proven that one can achieve lower regret if the temporal variation of the losses is small, it is unclear if the same guarantee can be achieved for FTRL without the computational cost of using full losses.

In this paper, we aim at bridging this gap by proposing a *generalized* version of implicit FTRL. We go beyond implicit and linearized updates: *we directly construct the update rule in a way that minimizes an upper bound on the regret*. Our framework effectively expands the scope of the FTRL framework, fully retaining its coupling between design and analysis. Also, our updates come with a worst-case guarantee to never be worse than the standard linearized ones.

We show the flexibility of our framework recovering known update schemes, like the Mirror-Prox update (Nemirovski, 2004), or extending updates specifically designed for OMD to the FTRL case, like the aProx one (Asi & Duchi, 2019). Moreover, for the first time, we show an implicit version of FTRL that recovers the temporal variation bound of implicit OMD (Campolongo & Orabona, 2020), but with the same computational complexity of implicit OMD.

Related Work While there are many works on implicit mirror descent in both the online and offline setting (see, e.g., Moreau, 1965; Martinet, 1970; Rockafellar, 1976; Kivinen & Warmuth, 1997; Parikh & Boyd, 2014; Campolongo & Orabona, 2020; Shtoff, 2022), the number of works that deal with implicit updates for FTRL is quite limited. We are only aware of McMahan (2010), which quantifies a gain only for specific regularizers. However, the framework in McMahan (2010) is non-constructive in the sense that it is difficult to see how to generalize implicit updates. Joulani et al. (2017) extends this last result, but it does not provide a link with the maximization of the dual function that governs the regret upper bound.

The closest approach to our framework is the one of Shalev-Shwartz & Singer (2007a;b), which develop a theory of FTRL updates as maximization of a dual function. However, their framework is limited to a specific shape of regularizers and it does not deal with implicit updates.

For implicit OMD, Campolongo & Orabona (2020) showed that implicit updates give rise to regret guarantees that depend on the temporal variability of the losses, so that constant regret is achievable if the variability of the losses is zero. They suggest that FTRL with full losses can achieve the same guarantee, but they also point out that given its computational complexity it would be "not worth pursuing." Here, we show how to achieve the same bound of implicit OMD with our generalized implicit FTRL, while retaining the same computational complexity of implicit OMD.

Proximal updates on truncated linear models were introduced in Asi & Duchi (2019) for the OMD algorithm. Chen et al. (2022b) used gradient flow on the same truncated linear models with a coin-betting algorithm (Orabona & Pál, 2016), but their approach does not seem to satisfy a regret guarantee. Chen et al. (2022a) have used truncated linear models in an FTRL-based parameter-free algorithm (Orabona & Pál, 2021) with a novel decomposition of the regret. However, their approach is ad-hoc and seems difficult to generalize.

2. Definitions and Basic Tools

We define here some basic concepts and tools of convex analysis, we refer the reader to, e.g., Rockafellar (1970); Bauschke & Combettes (2011) for a complete introduction to this topic. We will consider extended value function that can assume infinity values too. A function f is *proper* if it is nowhere $-\infty$ and finite somewhere. A function $f : V \subseteq \mathbb{R}^d \to [-\infty, +\infty]$ is *closed* if $\{\boldsymbol{x} : f(\boldsymbol{x}) \le \alpha\}$ is closed for every $\alpha \in \mathbb{R}$. For a proper function $f : \mathbb{R}^d \to (-\infty, +\infty]$, we define a *subgradient* of f in $\boldsymbol{x} \in \mathbb{R}^d$ as a vector $\boldsymbol{g} \in \mathbb{R}^d$ that satisfies $f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle, \; \forall \boldsymbol{y} \in \mathbb{R}^d$. We denote the set of subgradients of f in $\boldsymbol{x}$ by $\partial f(\boldsymbol{x})$. The *indicator function of the set* V, $i_V : \mathbb{R}^d \to (-\infty, +\infty]$,

has value 0 for $x \in V$ and $+\infty$ otherwise. We denote the *dual norm* of a norm $\|\cdot\|$ by $\|\cdot\|_\star$. A proper function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is *μ-strongly convex* over a convex set $V \subseteq \text{int dom } f$ w.r.t. $\|\cdot\|$ if $\forall x, y \in V$ and $\forall g \in \partial f(x)$, we have $f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2}\|x - y\|^2$. A function $f : V \to \mathbb{R}$, differentiable in an open set containing V, is *L-smooth* w.r.t. $\|\cdot\|$ if $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$ for all $x, y \in V$. For a function $f : \mathbb{R}^d \to [-\infty, \infty]$, we define the *Fenchel conjugate* $f^\star : \mathbb{R}^d \to [-\infty, \infty]$ as $f^\star(\theta) = \sup_{x \in \mathbb{R}^d} \langle \theta, x \rangle - f(x)$. From this definition, we immediately have the Fenchel-Young inequality: $f(x) + f^\star(\theta) \geq \langle \theta, x \rangle, \ \forall x, \theta$. We will also make use of the following properties of Fenchel conjugates.

Theorem 2.1 ((Orabona, 2019, Theorem 5.7)). *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper. Then, the following conditions are equivalent:*

(a) $\theta \in \partial f(x)$.
(b) $\langle \theta, y \rangle - f(y)$ achieves its supremum in y at $y = x$.
(c) $f(x) + f^\star(\theta) = \langle \theta, x \rangle$.
Moreover, if f is also convex and closed, we have an additional equivalent condition
(d) $x \in \partial f^\star(\theta)$.

Theorem 2.2 ((Orabona, 2019, Theorem 6.11)). *Let $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, closed, convex function, and $\text{dom } \partial \psi$ be non-empty. Then, ψ is $\lambda > 0$ strongly convex w.r.t. $\|\cdot\|$ iff $\psi^\star$ is $\frac{1}{\lambda}$-smooth w.r.t. $\|\cdot\|_\star$ on $\mathbb{R}^d$.*

3. Generalized Implicit FTRL

In this section, we introduce our novel generalized formulation of the implicit FTRL algorithm. The main idea is to depart from the implicit or linearized updates, and directly design updates that improve the upper bound on the regret. More in detail, the basic analysis of most of the online learning algorithms is based on the definition of subgradients:

$$\ell_t(x_t) - \ell_t(u) \leq \langle g_t, x_t - u \rangle, \ \forall g_t \in \partial \ell_t(x_t). \quad (1)$$

This allows to study the regret on the linearized losses as a proxy for the regret on the losses ℓ_t. However, we can do better. We introduce a new fundamental and more general strategy: using the Fenchel-Young inequality, we have

$$\ell_t(x_t) - \ell_t(u) \leq \ell_t(x_t) - \langle z_t, u \rangle + \ell_t^\star(z_t), \ \forall z_t.$$

In particular, the algorithm will choose z_t to make a certain upper bound involving this quantity to be tighter. This is a better inequality than (1) because when we select $z_t = g_t \in \partial \ell_t(x_t)$, using Theorem 2.1, we recover (1). So, this inequality subsumes the standard one for subgradients, but, using $z_t \in \ell_t(x_{t+1})$, it also subsumes the similar inequality used in the implicit case, as we show in Section 3.1. Moreover, we will see in Section 6 that it covers cases where z_t is *not* a subgradient of ℓ_t.

Algorithm 1 Generalized Implicit FTRL

Require: Non-empty closed set $V \subseteq \mathbb{R}^d$, a sequence of regularizers $\psi_1, \ldots, \psi_T : \mathbb{R}^d \to (-\infty, +\infty]$
1: $\theta_1 = \mathbf{0}$
2: **for** $t = 1$ **to** T **do**
3:      Output $x_t \in \text{argmin}_{x \in V} \ \psi_t(x) - \langle \theta_t, x \rangle$
4:      Receive $\ell_t : V \to \mathbb{R}$ and pay $\ell_t(x_t)$
5:      Set $g_t \in \partial \ell_t(x_t)$
6:      Set z_t such that $H_t(z_t) \leq H_t(g_t)$ or $H_t'(z_t) \leq H_t'(g_t)$ where H_t and H_t' are defined in (2) and (3)
7:      Set $\theta_{t+1} = \theta_t - z_t$
8: **end for**

The analysis shows that the optimal setting of z_t is the one that minimizes the function

$$H_t(z) \triangleq \psi_{t+1,V}^\star(\theta_t - z) + \ell_t^\star(z) \quad (2)$$

or

$$H_t'(z) \triangleq \psi_{t,V}^\star(\theta_t - z) + \ell_t^\star(z), \quad (3)$$

where $\psi_{t,V}$ is the restriction of the regularizer used at time t on the feasible set V, i.e., $\psi_{t,V} \triangleq \psi_t + i_V$. However, we can show that any setting of z_t that guarantees $H_t(z_t) < H_t(g_t)$ (or $H_t'(z_t) < H_t'(g_t)$) guarantee a strict improvement in the worst-case regret w.r.t. using the linearized losses.

One might wonder why the need for two different updates using H_t or H_t'. The reason is that when using time-varying regularizers that depend on the data, like in the FTRL version of AdaGrad (McMahan & Streeter, 2010; Duchi et al., 2011), if λ_{t+1} depends on z_t it might make the calculation of the update particularly difficult. This can be avoided using the update involving H_t'.

Once we have the z_t, we treat them as the subgradient of surrogate linear losses. So, putting it all together, Algorithm 1 shows the final algorithm. We now show a regret guarantee for this algorithm. First, we state a general Lemma and then instantiate it in a few interesting cases.

Theorem 3.1. *Let $V \subseteq \mathbb{R}^d$ be closed and non-empty and $\psi_t : V \to \mathbb{R}$. With the notation in Algorithm 1, define by $F_t(x) = \psi_t(x) + \sum_{i=1}^{t-1} \langle z_i, x \rangle$, so that $x_t \in \text{argmin}_{x \in V} F_t(x)$. Finally, assume that $\text{argmin}_{x \in V} F_t(x)$ and $\partial \ell_t(x_t)$ are not empty for all t.*

- *For any $z_t \in \mathbb{R}^d$ and any $u \in \mathbb{R}^d$, we have*

$$\text{Regret}_T(u) \leq \psi_{T+1}(u) - \min_{x \in V} \psi_1(x)$$

$$+ \sum_{t=1}^{T} [\psi_{t+1,V}^\star(\theta_t - g_t) - \psi_{t,V}^\star(\theta_t) + \langle x_t, g_t \rangle - \delta_t]$$

$$+ F_{T+1}(x_{T+1}) - F_{T+1}(u),$$

where $\delta_t \triangleq H_t(g_t) - H_t(z_t)$.

- If $\psi_{t+1}(\boldsymbol{x}) \geq \psi_t(\boldsymbol{x})$ for any $\boldsymbol{x} \in V$, then, for any $\boldsymbol{z}_t \in \mathbb{R}^d$, we have

$$\mathrm{Regret}_T(\boldsymbol{u}) \leq \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in V} \psi_1(\boldsymbol{x})$$
$$+ \sum_{t=1}^{T} [\psi_{t,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) - \psi_{t,V}^{\star}(\boldsymbol{\theta}_t) + \langle \boldsymbol{x}_t, \boldsymbol{g}_t \rangle - \delta'_t]$$
$$+ F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u}),$$

where $\delta'_t \triangleq H'_t(\boldsymbol{g}_t) - H'_t(\boldsymbol{z}_t)$.

Proof. The proof is composed of simple but not obvious steps. The first important observation is that the definition of $\boldsymbol{x}_t$ in the algorithm corresponds exactly to the one of FTRL on the linear losses $\langle \boldsymbol{z}_t, \cdot \rangle$. Hence, we can use the FTRL equality in Orabona (2019, Lemma 7.1):

$$-\sum_{t=1}^{T} \langle \boldsymbol{z}_t, \boldsymbol{u} \rangle = \sum_{t=1}^{T} [F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1})] + \psi_{T+1}(\boldsymbol{u})$$
$$- \min_{\boldsymbol{x} \in V} \psi_1(\boldsymbol{x}) + F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u}),$$

where we have cancelled out the terms $\langle \boldsymbol{z}_t, \boldsymbol{x}_t \rangle$ on both sides.

Now, use Fenchel-Young inequality, to have $\langle \boldsymbol{z}_t, \boldsymbol{u} \rangle \leq \ell_t(\boldsymbol{u}) + \ell_t^{\star}(\boldsymbol{z}_t)$. Hence, we have

$$-\sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leq \sum_{t=1}^{T} [F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t^{\star}(\boldsymbol{z}_t)]$$
$$+ \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in V} \psi_1(\boldsymbol{x})$$
$$+ F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u}).$$

Observe that

$$F_t(\boldsymbol{x}_t) = \min_{\boldsymbol{x} \in V} \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \langle \boldsymbol{z}_i, \boldsymbol{x} \rangle$$
$$= -\max_{\boldsymbol{x} \in V} \langle \boldsymbol{\theta}_t, \boldsymbol{x} \rangle - \psi_t(\boldsymbol{x}) = -\psi_{t,V}^{\star}(\boldsymbol{\theta}_t).$$

In the same way, we have $-F_{t+1}(\boldsymbol{x}_{t+1}) = \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_{t+1})$. Also, for any $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$, by Theorem 2.1 we have $\ell_t^{\star}(\boldsymbol{g}_t) = \langle \boldsymbol{x}_t, \boldsymbol{g}_t \rangle - \ell_t(\boldsymbol{x}_t)$. Hence, each term in the sum can be written as

$$F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t^{\star}(\boldsymbol{z}_t)$$
$$= \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_{t+1}) - \psi_{t,V}^{\star}(\boldsymbol{\theta}_t) + \ell_t^{\star}(\boldsymbol{z}_t)$$
$$= H_t(\boldsymbol{z}_t) - \psi_{t,V}^{\star}(\boldsymbol{\theta}_t).$$

Now, we just add and subtract $H_t(\boldsymbol{g}_t) = \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle - \ell_t(\boldsymbol{x}_t)$ to obtain the stated bound.

The second case is similar. We just have to observe that if $\psi_{t+1,V} \geq \psi_{t,V}$, then $\psi_{t+1,V}^{\star} \leq \psi_{t,V}^{\star}$. Hence, each term in

the sum can be upper bounded as

$$F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t^{\star}(\boldsymbol{z}_t)$$
$$\leq \psi_{t,V}^{\star}(\boldsymbol{\theta}_{t+1}) - \psi_{t,V}^{\star}(\boldsymbol{\theta}_t) + \ell_t^{\star}(\boldsymbol{z}_t)$$
$$= H'_t(\boldsymbol{z}_t) - \psi_{t,V}^{\star}(\boldsymbol{\theta}_t).$$

As before, adding and subtracting $H'_t(\boldsymbol{g}_t) = \psi_{t,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \langle \boldsymbol{x}_t, \boldsymbol{g}_t \rangle - \ell_t(\boldsymbol{x}_t)$ gives the stated bound. □

The Theorem is stated with very weak assumption to show its generality, but it is immediate to obtain concrete regret guarantees just assuming, for example, strongly convex regularizers and convex and Lipschitz losses and using well-known methods as Orabona (2019, Lemma 7.8)

However, we can already understand why this is an interesting guarantee. Let's first consider the case that $\boldsymbol{z}_t = \boldsymbol{g}_t$. In this case, we exactly recover the linearized FTRL algorithm. Even the guarantee in the Theorem exactly recovers the best known one (Orabona, 2019, Corollary 7.9), with $\delta_t = 0$ and $\delta'_t = 0$. Now, if we set $\boldsymbol{z}_t$ such that $H_t(\boldsymbol{z}_t) < H_t(\boldsymbol{g}_t)$ or $H'_t(\boldsymbol{z}_t) < H'_t(\boldsymbol{g}_t)$ we will have that $\delta_t > 0$ or $\delta'_t > 0$. Hence, in each single term of the sum we have a negative factor that makes the regret bound smaller. While it might be difficult to give a lower bound to δ_t and δ'_t without additional assumptions, the main value of this analysis is in giving a *unifying way to design generalized implicit updates for FTRL*. In fact, in the next sections we will show a number of possibilities that this framework enables.

Next, we will gain more understanding on the updates in Algorithm 1, comparing them to implicit OMD.

3.1. Comparison with Implicit Online Mirror Descent

In this section, we show that when $\boldsymbol{z}_t$ is set to minimize $H_t(\boldsymbol{z})$ or $H'_t(\boldsymbol{z})$, we recover different variants of implicit updates.

Assume that the ℓ_t are closed and convex. Also, assume that $\psi_{t,V}^{\star}$ is differentiable, that is true, for example, when ψ_t is strongly convex by Theorem 2.2. Then, observe that by the first-order optimality condition and Theorem 2.1, we have

$$\boldsymbol{z}_t = \underset{\boldsymbol{z}}{\mathrm{argmin}} \ H_t(\boldsymbol{z})$$
$$\Leftrightarrow \nabla \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) \in \partial \ell_t^{\star}(\boldsymbol{z}_t)$$
$$\Leftrightarrow \boldsymbol{z}_t \in \partial \ell_t(\nabla \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t)) = \partial \ell_t(\boldsymbol{x}_{t+1}). \quad (4)$$

Hence, in this case, we have that the optimal $\boldsymbol{z}_t$ is the gradient at the *next* point $\boldsymbol{x}_{t+1}$. This is exactly what happens in the implicit updates. This condition that depends on the gradient in the next iterate is similar to other implicit updates.

Under the same assumptions, we also have

$$\boldsymbol{z}_t = \operatorname*{argmin}_{\boldsymbol{z}} H_t'(\boldsymbol{z}) \Leftrightarrow \nabla\psi_{t,V}^\star(\boldsymbol{\theta}_t - \boldsymbol{z}_t) \in \partial\ell_t^\star(\boldsymbol{z}_t)$$

$$\Leftrightarrow \boldsymbol{z}_t \in \partial\ell_t(\nabla\psi_{t,V}^\star(\boldsymbol{\theta}_{t+1})) . \quad (5)$$

In this other case, the update also has an implicit flavor but the subgradient is queried on a point different from the next point, where the difference depends on how much $\nabla\psi_{t,V}^\star$ differs from $\nabla\psi_{t+1,V}^\star$.

Let's see this connection even more precisely, considering *proximal updates*. Hence, for simplicity, let's consider the case that $V = \mathbb{R}^d$, similar considerations hold in the constrained case. Consider the case that $\psi_t(\boldsymbol{x}) = \frac{\lambda_t}{2}\|\boldsymbol{x}\|_2^2$. In this case, the update can be written with the *proximal operator* of the loss functions. In particular, the proximal operator of ηf, is defined as

$$\operatorname{Prox}_{\eta f}(\boldsymbol{y}) \triangleq \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \eta f(\boldsymbol{x}) .$$

If the function f is differentiable we have that $\operatorname{Prox}_{\eta f}(\boldsymbol{y}) = \boldsymbol{y} - \eta\nabla f(\operatorname{Prox}_{\eta f}(\boldsymbol{y}))$. In words, the proximal update moves by a quantity that depends on the gradient on the updated point. The implicit nature of these updates justifies the name "implicit updates" used in the online learning literature. More generally, we have that $\operatorname{Prox}_{\eta f}(\boldsymbol{y}) \in \boldsymbol{y} - \eta\partial f(\operatorname{Prox}_{\eta f}(\boldsymbol{y}))$. We list some common proximal operators in Appendix A.

Assuming λ_{t+1} does not depend on $\boldsymbol{z}_t$, using the proximal operator we can rewrite the update in (4) as

$$\boldsymbol{x}_{t+1} = \frac{\boldsymbol{\theta}_{t+1}}{\lambda_{t+1}} = \operatorname{Prox}_{\frac{\ell_t}{\lambda_{t+1}}}\left(\frac{\boldsymbol{\theta}_t}{\lambda_{t+1}}\right)$$

$$= \operatorname{Prox}_{\frac{\ell_t}{\lambda_{t+1}}}\left(\frac{\lambda_t\boldsymbol{x}_t}{\lambda_{t+1}}\right) . \quad (6)$$

Similarly, we can rewrite the update in (5) as

$$\frac{\boldsymbol{\theta}_{t+1}}{\lambda_t} = \frac{\boldsymbol{\theta}_t}{\lambda_t} - \frac{\boldsymbol{z}_t}{\lambda_t} = \boldsymbol{x}_t - \frac{\boldsymbol{z}_t}{\lambda_t} \in \boldsymbol{x}_t - \frac{1}{\lambda_t}\partial\ell_t(\nabla\psi_{t,V}^\star(\boldsymbol{\theta}_{t+1}))$$

$$= \boldsymbol{x}_t - \frac{1}{\lambda_t}\partial\ell_t\left(\frac{\boldsymbol{\theta}_{t+1}}{\lambda_t}\right) .$$

Hence, we have that $\frac{\boldsymbol{\theta}_{t+1}}{\lambda_t} = \operatorname{Prox}_{\frac{\ell_t}{\lambda_t}}(\boldsymbol{x}_t)$ and we get

$$\boldsymbol{x}_{t+1} = \frac{\boldsymbol{\theta}_{t+1}}{\lambda_{t+1}} = \frac{\lambda_t}{\lambda_{t+1}}\operatorname{Prox}_{\frac{\ell_t}{\lambda_t}}(\boldsymbol{x}_t) . \quad (7)$$

It is instructive to compare both updates with the one of Implicit Online Mirror Descent using $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ as distance generating function and stepsizes $\frac{1}{\lambda_t}$. In this case, we would update with

$$\boldsymbol{x}_{t+1} = \operatorname*{argmin}_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}\|_2^2 + \frac{1}{\lambda_t}\ell_t(\boldsymbol{x})$$

$$= \operatorname{Prox}_{\frac{\ell_t}{\lambda_t}}(\boldsymbol{x}_t) . \quad (8)$$

Comparing (4) and (5) to (8), we see, when $\lambda_t \le \lambda_{t+1}$ as it is usual, the two updates above shrink a bit towards the zero vector, that is the initial point $\boldsymbol{x}_1$, before or after the proximal operator. This shrinking is given by the FTRL update and it is the key difference with Implicit OMD update. The different update also corresponds to a different guarantee: the regret of the generalized implicit FTRL holds for unbounded domains too, while in Implicit OMD with time-varying stepsizes can have linear regret on unbounded domains (Orabona & Pál, 2018). Interestingly, a similar shrinking has been proposed in Fang et al. (2020) to fix the unbounded issue in OMD. Clearly, the updates (4) and (5) become equivalent to (8) for λ_t constant in t, that is exactly the only case when implicit/proximal online mirror descent works for unbounded domains.

4. Temporal Variability Bound

In this section, we quantify the advantage of the generalized implicit FTRL updates in the case of slow temporal variability of the loss functions.

It was observed in Campolongo & Orabona (2020) that implicit OMD satisfies regret guarantees that depend on the temporal Variability V_T:

$$V_T \triangleq \sum_{t=2}^{T} \max_{x\in V} \ell_t(\boldsymbol{x}) - \ell_{t-1}(\boldsymbol{x}) .$$

In Campolongo & Orabona (2020, Appendix E) they also show that FTRL with full losses guarantees a similar guarantee, but at a much higher computational price. Indeed, FTRL with full losses requires solving a finite sum optimization problem at each step, whose size increases with the number of iterations. Such computational burden induced Campolongo & Orabona (2020) to say that such approach is "not worth of pursuing."

Here, we show that the Algorithm 1 can satisfy the same guarantee of implicit OMD with the same computational complexity too. First, we show the following Lemma.

Lemma 4.1. *Under the assumptions of Theorem 3.1, further assume V to be convex, $\psi_t : V \to \mathbb{R}$ closed, λ_t-strongly convex w.r.t. $\|\cdot\|$, and subdifferentiable in V, ℓ_t closed, convex, and subdifferentiable in V, and $\lambda_{t+1} \ge \lambda_t$. Set $\boldsymbol{z}_t \in \operatorname{argmin}_{\boldsymbol{z}} H_t(\boldsymbol{z})$. Then, we have*

$$\operatorname{Regret}_T(\boldsymbol{u}) \le \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V} \psi_1(\boldsymbol{x})$$

$$+ \sum_{t=1}^{T}\left(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}_{t+1}) - \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\right), \forall\boldsymbol{u}\in V.$$

Proof. First of all, the existence and unicity of $\boldsymbol{x}_t$ is guaranteed by ψ_t being closed and strongly convex (see, e.g., Orabona, 2019, Theorem 6.8).

From Theorem 2.1, for any $\boldsymbol{g}'_t \in \partial\ell_t(\boldsymbol{x}_{t+1})$, we have $\ell_t^\star(\boldsymbol{g}'_t) = \langle \boldsymbol{x}_{t+1}, \boldsymbol{g}'_t \rangle - \ell_t(\boldsymbol{x}_{t+1})$. Hence, from (4), we have

$$\psi_{t+1,V}^\star(\boldsymbol{\theta}_{t+1}) - \psi_{t,V}^\star(\boldsymbol{\theta}_t) + \ell_t^\star(\boldsymbol{z}_t)$$
$$= \psi_{t+1,V}^\star(\boldsymbol{\theta}_t - \boldsymbol{z}_t) - \psi_{t,V}^\star(\boldsymbol{\theta}_t) + \langle \boldsymbol{x}_{t+1}, \boldsymbol{z}_t \rangle - \ell_t(\boldsymbol{x}_{t+1}) .$$

Using this identity, we have

$$\psi_{t+1,V}^*(\boldsymbol{\theta}_t - \boldsymbol{z}_t) - \psi_{t,V}^*(\boldsymbol{\theta}_t) + \langle \boldsymbol{x}_{t+1}, \boldsymbol{z}_t \rangle$$
$$= \langle \boldsymbol{\theta}_t - \boldsymbol{z}_t, \boldsymbol{x}_{t+1} \rangle - \psi_{t+1}(\boldsymbol{x}_{t+1}) - \langle \boldsymbol{\theta}_t, \boldsymbol{x}_t \rangle + \psi_t(\boldsymbol{x}_t)$$
$$\quad + \langle \boldsymbol{x}_{t+1}, \boldsymbol{z}_t \rangle$$
$$\leq \psi_t(\boldsymbol{x}_t) - \psi_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\theta}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle .$$

From the first-order optimality condition of $\boldsymbol{x}_t$, we have that $\boldsymbol{\theta}_t \in \partial\psi_t(\boldsymbol{x}_t) + \partial i_V(\boldsymbol{x}_t)$. Moreover, for all $\boldsymbol{g}''_t \in \partial i_V(\boldsymbol{x}_t)$, by definition we have $\langle \boldsymbol{g}''_t, \boldsymbol{y} - \boldsymbol{x}_t \rangle \leq 0$ for all $\boldsymbol{y} \in V$. Hence, for $\boldsymbol{g}'_t \in \partial\psi_t(\boldsymbol{x}_t)$ and $\boldsymbol{g}''_t \in \partial i_V(\boldsymbol{x}_t)$ such that $\boldsymbol{\theta}_t = \boldsymbol{g}'_t + \boldsymbol{g}''_t$, we have

$$\psi_t(\boldsymbol{x}_t) - \psi_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{\theta}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle$$
$$= \psi_t(\boldsymbol{x}_t) - \psi_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}'_t + \boldsymbol{g}''_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle$$
$$\leq -\frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2,$$

where in the inequality we also used the strong convexity of ψ_t. Using this inequality in Theorem 3.1 and summing over time, we have

$$\sum_{t=1}^T \ell_t(\boldsymbol{x}_{t+1}) - \sum_{t=1}^T \ell_t(\boldsymbol{u})$$
$$\leq \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V} \psi_1(\boldsymbol{x}) - \sum_{t=1}^T \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 .$$

By adding and subtracting $\sum_{t=1}^T \ell_t(\boldsymbol{x}_t)$ to both sides and reordering the terms, we have the stated bound. □

This Lemma mirrors Theorem 5.2 in Campolongo & Orabona (2020), with the important difference that here we do not need the Bregman divergence to be bounded on the feasible set V, thanks to the use of FTRL instead of OMD. We can now state the immediate corollary on a regret bound that depends on the temporal variation.

Corollary 4.2. *Under the assumptions of Lemma 4.1, for any $\boldsymbol{u} \in V$, we have*

$$\text{Regret}_T(\boldsymbol{u}) \leq \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V} \psi_1(\boldsymbol{x})$$
$$+ \ell_1(\boldsymbol{x}_1) - \ell_T(\boldsymbol{x}_{T+1}) + V_T .$$

From this result, following (Campolongo & Orabona, 2020), it is relatively easy to obtain the following adaptive regret guarantee. The only difficulty is the fact that we need ψ_{t+1} to be independent of $\boldsymbol{z}_t$ to have a simpler update rule. We

solve this problem using an increasing regularizer that is "behind of two steps". In this way, we have that λ_{t+1} depends on quantities that are all known at the beginning of round t. The proof is in Appendix B.

Corollary 4.3. *Under the assumptions of Lemma 4.1, further assume $\|\boldsymbol{g}_t\|_\star \leq G$ for all t. Define $\gamma_t = \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}_{t+1}) - \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2$ and $\lambda_t = \frac{1}{\beta^2}\left(G\beta + \sum_{i=1}^{t-2}\gamma_i\right)$. Assume that ψ is closed and 1-strongly convex w.r.t. $\|\cdot\|$ and set $\psi_t = \lambda_t\psi$. Then, for any $\boldsymbol{u} \in V$, we have*

$$\text{Regret}_T(\boldsymbol{u}) \leq \min\left(\frac{1}{\beta}(\ell_1(\boldsymbol{x}_1) - \ell_T(\boldsymbol{x}_{T+1}) + V_T),\right.$$
$$\left. G + \sqrt{\frac{5}{4}\sum_{t=1}^T \|\boldsymbol{g}_t\|_\star^2}\right)\left(\frac{\psi(\boldsymbol{u})}{\beta} + \beta\right) .$$

5. Two-step Updates

The choice of $\boldsymbol{z}_t$ that minimizes the regret upper bound requires solving the optimization problem $\min_{\boldsymbol{z}} H_t(\boldsymbol{z})$ or $\min_{\boldsymbol{z}} H'_t(\boldsymbol{z})$. We have seen in Section 3.1 that this corresponds to (some variant) of a implicit/proximal update and, depending on ℓ_t, it can be of difficult calculation. However, as we said, any choice better than $\boldsymbol{g}_t$ will cause a provable gain. Hence, a viable solution is to *approximately* solve for the optimal $\boldsymbol{z}_t$.

Here, we propose a simple approximation: set $\boldsymbol{z}_t$ as

$$\boldsymbol{z}_t \in \partial\ell_t(\nabla\psi_{t+1,V}^\star(\boldsymbol{\theta}_t - \boldsymbol{g}_t)) \tag{9}$$

or as

$$\boldsymbol{z}_t \in \partial\ell_t(\nabla\psi_{t,V}^\star(\boldsymbol{\theta}_t - \boldsymbol{g}_t)) . \tag{10}$$

In words, we set $\boldsymbol{z}_t$ to be a subgradient after one fake update. This is exactly the approach used in the Mirror-Prox algorithm (Nemirovski, 2004), an offline optimization algorithm. In the next theorem, when the loss functions ℓ_t are smooth and the regularizer is chosen appropriately, we show that this choice can be used in the generalized implicit FTRL too and it cannot be worse than using $\boldsymbol{g}_t$.

Theorem 5.1. *Assume $\psi_t(\boldsymbol{x})$ proper, closed, and λ_t-strongly convex with respect to $\|\cdot\|$. Assume $\ell_t(\boldsymbol{x})$ closed and λ_t-smooth w.r.t. $\|\cdot\|_\star$ for all t. Then, using (9) and assuming $\lambda_{t+1} \geq L_t$, we have $H_t(\boldsymbol{z}_t) \leq H_t(\boldsymbol{g}_t)$. On the other hand, when using (10) and assuming $\lambda_t \geq L_t$ we have $H'_t(\boldsymbol{z}_t) \leq H'_t(\boldsymbol{g}_t)$.*

Proof. We only prove that statement for (9), the other one is similar. We would like to prove that

$$H_t(\boldsymbol{z}_t) = \psi_{t+1,V}^\star(\boldsymbol{\theta}_t - \boldsymbol{z}_t) + \ell_t^\star(\boldsymbol{z}_t)$$
$$\leq \psi_{t+1,V}^\star(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \ell_t^\star(\boldsymbol{g}_t) = H_t(\boldsymbol{g}_t) .$$

This is equivalent to prove

$$\psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) - \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) \leq \ell_t^{\star}(\boldsymbol{g}_t) - \ell_t^{\star}(\boldsymbol{z}_t)\,.$$

Given that $\psi_{t+1}(\boldsymbol{x}_t)$ is λ_{t+1}-strongly convex, by Theorem 2.2, we have $\psi_t^{\star}(\boldsymbol{\theta})$ is $1/\lambda_{t+1}$-smooth with respect to $\|\cdot\|_{\star}$. By the definition of smoothness, we have

$$\psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) - \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t)$$
$$\leq \langle \nabla\psi_{t+1}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t), \boldsymbol{g}_t - \boldsymbol{z}_t \rangle + \frac{1}{2\lambda_{t+1}}\|\boldsymbol{g}_t - \boldsymbol{z}_t\|_{\star}^2\,.$$

Given that $\ell_t(\boldsymbol{x}_t)$ is L_t-smooth w.r.t $\|\cdot\|_{\star}$, by Theorem 2.2 $\ell_t^{\star}(\boldsymbol{g})$ is $1/L_t$ strongly convex w.r.t. $\|\cdot\|$. So, by the definition of the strong convexity, we have

$$\ell_t^{\star}(\boldsymbol{g}_t) - \ell_t^{\star}(\boldsymbol{z}_t) \geq \langle \boldsymbol{q}_t, \boldsymbol{g}_t - \boldsymbol{z}_t \rangle + \frac{1}{2L_t}\|\boldsymbol{g}_t - \boldsymbol{z}_t\|_{\star}^2,$$

for all $\boldsymbol{q}_t \in \partial\ell_t^{\star}(\boldsymbol{z}_t)$. Defining $\boldsymbol{x}_{t+1}' \triangleq \nabla\psi_{t+1}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t)$, by Theorem 2.1, we have $\boldsymbol{x}_{t+1}' \in \partial\ell_t^{\star}(\boldsymbol{z}_t)$. Hence, we can select $\boldsymbol{q}_t$ such that $\boldsymbol{x}_{t+1}' = \boldsymbol{q}_t$. Finally, using the assumption on $\lambda_{t+1} \geq L_t$, we have the stated bound. $\square$

6. Going Beyond Subgradients with aProx

Till now, in all the updates we have considered $\boldsymbol{z}_t$ was set to be a subgradient of ℓ_t in a specific point. In this section, we show that we can go beyond this idea.

Asi & Duchi (2019) introduced aProx updates, that is proximal updates on surrogate loss functions. In particular, they used truncated linear lower bounds to the loss functions as surrogate functions. These simple surrogates are motivated by the fact that they are strictly better than linear approximation and at the same time they allow writing the proximal update in a closed form. Moreover, they showed empirically that in certain situations the performance of the algorithms becomes much more resistant to the tuning of the stepsizes.

One might just use the same truncated lower bounds in implicit FTRL, but it would not be clear why this should give any advantage in the theoretical bound. Indeed, even in Asi & Duchi (2019) it is not completely clear what part of the theory tells us that we should expect a better performance from these updates.

Here, we show how the *updates in the generalized implicit FTRL are actually a generalization of the aProx ones*. In particular, we generalize the aProx updates to arbitrary regularizers and show that all of them satisfy $H_t(\boldsymbol{z}_t) \leq H_t(\boldsymbol{g}_t)$ and $H_t'(\boldsymbol{z}_t) \leq H_t'(\boldsymbol{g}_t)$. In words, the aProx updates are guaranteed to be at least as good as the subgradient $\boldsymbol{g}_t$ in minimizing the worst-case regret.

In order to consider truncated linear lower bounds to the functions ℓ_t, in this section we will assume that the loss functions ℓ_t are lower bounded. Given that the regret is invariant to additive constants in the losses, without loss of generality we can assume the lower bound to be 0 for all the loss functions. Hence, define the truncated linear model $\hat{\ell}_t : V \to \mathbb{R}$ around $\boldsymbol{x}_t$ to be

$$\hat{\ell}_t(\boldsymbol{x}) \triangleq \max(\ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle, 0),$$

where $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$. For brevity of notation, our notation does not stress the fact that the truncated linear model depends on $\boldsymbol{x}_t$ and the specific subgradient $\boldsymbol{g}_t$.

The idea to extend aProx to the case of generalized implicit FTRL, we use the truncated linear lower bound in the update of $\boldsymbol{z}_t$. So, we define

$$\boldsymbol{z}_t = \arg\min_{\boldsymbol{z}} \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) + \hat{\ell}_t^{\star}(\boldsymbol{z}_t) \qquad (11)$$

or

$$\boldsymbol{z}_t = \arg\min_{\boldsymbol{z}} \psi_{t,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) + \hat{\ell}_t^{\star}(\boldsymbol{z}_t)\,. \qquad (12)$$

Theorem 6.1. *Assume the loss functions $\ell_t : V \to \mathbb{R}$ to be convex, closed, and subdifferentiable in V for all t. Set $\boldsymbol{z}_t$ using (11) or (12). Then, we have that $H_t(\boldsymbol{z}_t) \leq H_t(\boldsymbol{g}_t)$ or $H_t'(\boldsymbol{z}_t) \leq H_t'(\boldsymbol{g}_t)$ respectively.*

Proof. We consider the update (11), the other case is very similar and we omit it.

First, we derive some inequalities on the quantities of interest. From Theorem 2.1, given that $\boldsymbol{g}_t \in \partial\hat{\ell}_t(\boldsymbol{x}_t)$ and $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$ we have both $\ell_t(\boldsymbol{x}_t) + \ell^{\star}(\boldsymbol{g}_t) = \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle$ and $\hat{\ell}_t(\boldsymbol{x}_t) + \hat{\ell}^{\star}(\boldsymbol{g}_t) = \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle$. Moreover, given that $\hat{\ell}_t(\boldsymbol{x}) \leq \ell_t(\boldsymbol{x})$ for any $\boldsymbol{x}$, we have $\hat{\ell}_t^{\star}(\boldsymbol{z}) \geq \ell_t^{\star}(\boldsymbol{z})$ for any $\boldsymbol{z}$. Finally, by the definition of truncated linear lower bound, we have $\ell_t(\boldsymbol{x}_t) = \hat{\ell}_t(\boldsymbol{x}_t)$.

Hence, we have

$$\psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) + \ell_t^{\star}(\boldsymbol{z}_t)$$
$$\leq \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}_t) + \hat{\ell}_t^{\star}(\boldsymbol{z}_t)$$
$$= \min_{\boldsymbol{z}} \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{z}) + \hat{\ell}_t^{\star}(\boldsymbol{z})$$
$$\leq \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \hat{\ell}_t^{\star}(\boldsymbol{g}_t)$$
$$= \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle - \hat{\ell}_t(\boldsymbol{x}_t)$$
$$= \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle - \ell_t(\boldsymbol{x}_t)$$
$$= \psi_{t+1,V}^{\star}(\boldsymbol{\theta}_t - \boldsymbol{g}_t) + \ell_t^{\star}(\boldsymbol{g}_t) = H_t(\boldsymbol{g}_t)\,. \quad \square$$

We can also immediately write closed form updates for generalized implicit FTRL with regularizer $\psi_t(\boldsymbol{x}) = \frac{\lambda_t}{2}\|\boldsymbol{x}\|^2$, that mirror the ones of aProx. The proof is in Section C.

Corollary 6.2. *Set $\psi_t = \frac{\lambda_t}{2}\|\boldsymbol{x}\|_2^2$ and $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$. Setting $\boldsymbol{z}_t$ as in (11), we have that the update of generalized implicit FTRL is*

$$\boldsymbol{x}_{t+1} = \frac{\lambda_t}{\lambda_{t+1}}\boldsymbol{x}_t - \min\left(\frac{1}{\lambda_{t+1}}, \frac{\ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}\right)\boldsymbol{g}_t\,.$$
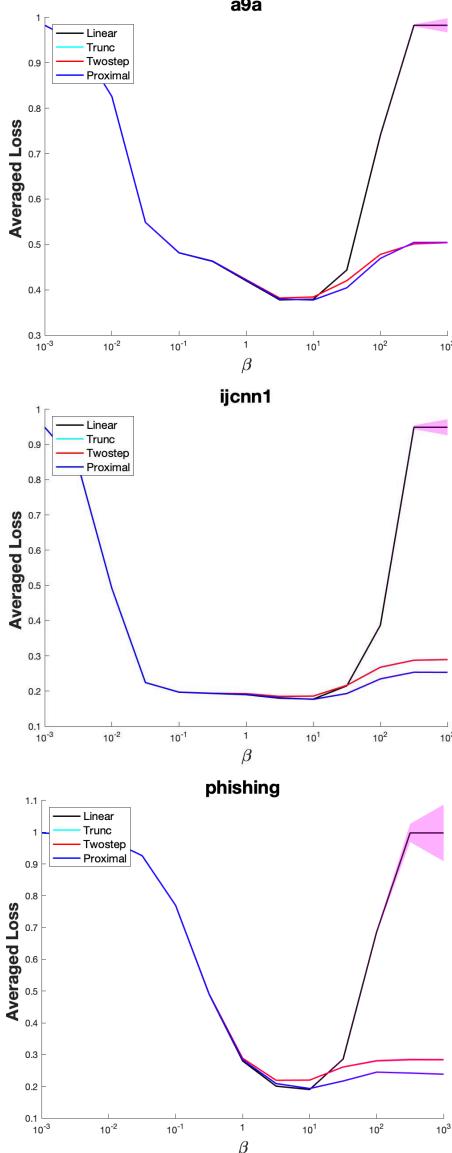
Figure 1. Hinge loss, averaged loss vs. hyperparameter β.

Figure 2. Logistic loss, averaged loss vs. hyperparameter β.

On the other hand, setting $\boldsymbol{z}_t$ as in (12), *the update is*

$$\boldsymbol{x}_{t+1} = \frac{\lambda_t}{\lambda_{t+1}} \boldsymbol{x}_t - \min\left(\frac{1}{\lambda_{t+1}}, \frac{\lambda_t}{\lambda_{t+1}} \frac{\ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}\right) \boldsymbol{g}_t .$$

7. Empirical Evaluation

As we said, in the worst case scenario any kind of implicit update cannot give any advantage over the usual updates. However, in practice it is well-known that things are vastly different. Hence, in this section, we compare the performance of different choices of $\boldsymbol{z}_t$ in Algorithm 1 when $\psi_t(\boldsymbol{x}) = \frac{\lambda_t}{2}\|\boldsymbol{x}\|_2^2$. In particular, we consider:

- FTRL with linearized losses (Linear): $\boldsymbol{z}_t = \boldsymbol{g}_t$;
- Implicit FTRL with aProx updates (Trunc): $\boldsymbol{z}_t =$

$\min\left\{1, \frac{\lambda_t \ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}\right\} \boldsymbol{g}_t$;
- Implicit FTRL with two-step updates (Twostep): $\boldsymbol{z}_t = \partial\ell_t(\boldsymbol{x}_t - \boldsymbol{g}_t/\lambda_t)$;
- Implicit FTRL with (6) when the proximal operator has a closed form (Proximal).

We adopt the choice of λ_t from Corollary 4.3.

We conduct linear prediction experiments on datasets from LibSVM (Chang & Lin, 2011). We show here experiments on classification tasks using the hinge loss and the logistic loss, and regression tasks with absolute loss. We normalize the datasets and added a constant bias term to the features. Given that in the online learning setting, we do not have the training data and validation data to tune the β, we will plot the averaged loss, $\frac{1}{t}\sum_{i=1}^{t} \ell_i(\boldsymbol{x}_i)$, versus different choice of
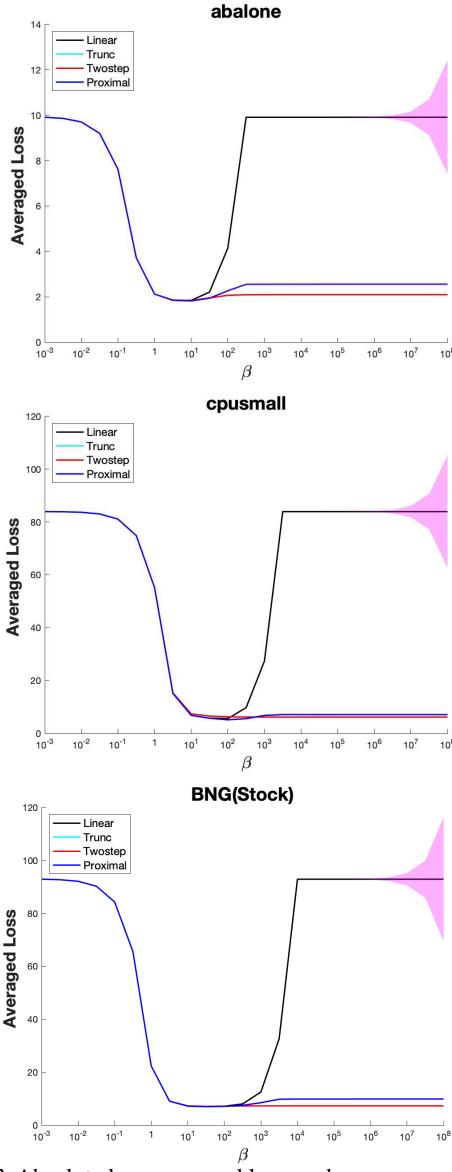
Figure 3. Absolute loss, averaged loss vs. hyperparameter β.

β, that at the same time show the algorithms' sensitivity to the hyperparameter β and their best achievable performance. We consider $\beta \in [10^{-3}, 10^3]$ for hinge loss and logistic loss, and $\beta \in [10^{-3}, 10^8]$ for the absolute loss. Each algorithm is run 15 times, we plot the average of the averaged losses and the 95% confidence interval. Note that the confidence intervals so small to be invisible, but for the larger values of the β for the Linear updates.

Figure 1 and Figure 2 show the averaged loss versus different selections of hyperparameter β for classification tasks with hinge loss and logistic loss respectively. Note that with the hinge loss aProx updates and proximal updates are completely equivalent. In all experiments, FTRL with linearized updates is more sensitive to the setting of β, and

its performance is almost uniformly worse than all the other generalized implicit updates. This is in line with previous results in Asi & Duchi (2019) in the offline setting. With the logistic loss, the proximal operator does not have a closed-form solution. In all the classification experiments, the performance of generalized implicit FTRL with two-step updates seems remarkable and a possible viable alternative to aProx. The confidence intervals for all implicit updates have a width smaller than 0.01, making them too narrow to be visible in the figures. In contrast, when using hinge loss, the performance of FTRL with linear models exhibits significant fluctuations across different repetitions when a large learning rate is used. This observation provides evidence supporting our assertion that the selection of hyperparameter β greatly affects the performance of FTRL with linear models, while implicit updates demonstrate robustness.

Figure 3 shows that FTRL with linearized updates is very sensitive to the choice of the hyperparameter β, while the implicit FTRL updates are robust. Again, Implicit FTRL with two-step updates achieves essentially the best performance. The confidence intervals in the regression tasks lead to a similar conclusion as in the classification tasks.

8. Conclusion and Future Work

In this work, we propose a new framework: generalized implicit Follow-the-Regularized-Leader. We show that generalized implicit FTRL can not only recover known algorithms, e.g., implicit FTRL and FTRL with linearized losses, but it also provides a theoretical guideline to design new algorithms, such as the extensions of aProx and Mirror-Prox. Indeed, we believe that the main contribution of our work lies precisely in the fact that it provides a unifying framework that is general, flexible, and theoretically grounded.

In the future, we plan to explore further this framework designing new z_t with low computational complexity. This is a promising direction because the two-step update seems to be already a valid alternative to the aProx updates, even if it comes at the computational expense of querying an additional gradient in each round.

Acknowledgements

References

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

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

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

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

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Cesa-Bianchi, N. and Orabona, F. Online learning algorithms. *Annual Review of Statistics and Its Application*, 8:165–190, 2021.

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, K., Cutkosky, A., and Orabona, F. Implicit parameter-free online learning with truncated linear models. In *International Conference on Algorithmic Learning Theory*, pp. 148–175. PMLR, 2022a.

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

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

Duchi, J. C., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

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

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

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

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

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

Littlestone, N. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

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

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

McMahan, H. B. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

McMahan, H. B. and Streeter, M. J. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.

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

Nemirovski, A. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

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

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. Version 6.

Orabona, F. and Pál, D. Scale-free algorithms for online linear optimization. In *International Conference on Algorithmic Learning Theory*, pp. 287–301. Springer, 2015.

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

Orabona, F. and Pál, D. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. Special Issue on ALT 2015.

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

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

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, 1970.

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

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

Shalev-Shwartz, S. and Singer, Y. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69:115–142, 2007a.

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

Shtoff, A. Efficient implementation of incremental proximal-point methods. *arXiv preprint arXiv:2205.01457*, 2022.

Streeter, M. and McMahan, H. B. Less regret via online conditioning. *arXiv preprintarXiv:1002.4862*, 2010.

Warmuth, M. K. and Jagota, A. K. 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.

A. Update Rule for Common Losses

In this section, we report the proximal operator of common losses for easy referencing. These formulas are well-known and they can be found, for example, in Crammer et al. (2006); Kulis & Bartlett (2010).

$$\ell_t(\boldsymbol{x}) = \max(1 - y_t\langle \boldsymbol{s}_t, \boldsymbol{x}\rangle, 0) \Rightarrow \mathrm{Prox}_{\frac{\ell_t}{\lambda}}(\boldsymbol{x}) = \boldsymbol{x} + \min\left(\frac{1}{\lambda}, \frac{\max(1 - y_t\langle \boldsymbol{s}_t, \boldsymbol{x}\rangle, 0)}{\|\boldsymbol{s}_t\|^2}\right)y_t\boldsymbol{s}_t$$

$$\ell_t(\boldsymbol{x}) = |\langle \boldsymbol{s}_t, \boldsymbol{x}\rangle - y_t| \Rightarrow \mathrm{Prox}_{\frac{\ell_t}{\lambda}}(\boldsymbol{x}) = \boldsymbol{x} - \min\left(\frac{1}{\lambda}, \frac{|\langle \boldsymbol{s}_t, \boldsymbol{x}\rangle - y_t|}{\|\boldsymbol{s}_t\|^2}\right)\boldsymbol{s}_t$$

$$\ell_t(\boldsymbol{x}) = \frac{1}{2}(\langle \boldsymbol{s}_t, \boldsymbol{x}\rangle - y_t)^2 \Rightarrow \mathrm{Prox}_{\frac{\ell_t}{\lambda}}(\boldsymbol{x}) = \boldsymbol{x} - \frac{(\langle \boldsymbol{s}_t, \boldsymbol{x}\rangle - y_t)\boldsymbol{s}_t}{\lambda + \|\boldsymbol{s}_t\|_2^2}\ .$$

B. Proof of Corollary 4.3

Proof. From the regret guarantee in Lemma 4.1, we have that

$$\mathrm{Regret}_T(\boldsymbol{u}) \le \psi_{T+1}(\boldsymbol{u}) + \beta^2\sum_{t=1}^{T}\gamma_t \le \lambda_{T+1}(\psi(\boldsymbol{u}) + \beta^2),\ \forall \boldsymbol{u} \in V\ .$$

Now, we upper bound $\sum_{t=1}^{T}\gamma_t$ in two different ways. In the first upper bound, we have

$$\sum_{t=1}^{T}\gamma_t = \sum_{t=1}^{T}\left(\ell_t(\boldsymbol{x}_t) - \ell(\boldsymbol{x}_{t+1}) - \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\right) = \ell_1(\boldsymbol{x}_1) - \ell_T(\boldsymbol{x}_{T+1}) + \sum_{t=2}^{T}(\ell_t(\boldsymbol{x}_t) - \ell_{t-1}(\boldsymbol{x}_t))$$

$$\le \ell_1(\boldsymbol{x}_1) - \ell_T(\boldsymbol{x}_{T+1}) + V_T\ .$$

For the second upper bound, we have

$$\gamma_t = \ell_t(\boldsymbol{x}_t) - \ell(\boldsymbol{x}_{t+1}) - \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \le \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle - \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2$$

$$\le \frac{\|\boldsymbol{g}_t\|_\star^2}{2\lambda_t} + \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 - \frac{\lambda_t}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 = \frac{\|\boldsymbol{g}_t\|_\star^2}{2\lambda_t} \le \beta\frac{\|\boldsymbol{g}_t\|_\star}{2},$$

where we used Fenchel-Young inequality and the second lower bound is obtained by using the fact that $\lambda_t \ge \lambda_1 = \frac{G}{\beta}$. Hence, we have

$$\lambda_{t+1} = \lambda_t + \frac{\gamma_t}{\beta^2} \le \lambda_t + \min\left(\frac{\|\boldsymbol{g}_t\|_\star}{2\beta}, \frac{\|\boldsymbol{g}_t\|_\star^2}{2\beta^2\lambda_t}\right)\ .$$

Using Lemma 6.1 in Campolongo & Orabona (2020) and taking into account the fact that $\lambda_1 = \frac{G}{\beta}$, we have

$$\lambda_{T+1} \le \frac{G}{\beta} + \sqrt{\frac{5}{4\beta^2}\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_\star^2}\ .$$

Putting all together, we have the stated bound. □

C. Proof of Corollary 6.2

Proof. The proximal operator of $\frac{\hat{\ell}_t}{\lambda}$ is

$$\mathrm{Prox}_{\frac{\hat{\ell}_t}{\lambda}}(\boldsymbol{x}) = \boldsymbol{x} - \min\left(\frac{1}{\lambda}, \frac{\ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}\right)\boldsymbol{g}_t\ .$$

Hence, from (7), we have

$$\boldsymbol{x}_{t+1} = \frac{\lambda_t}{\lambda_{t+1}} \left(\boldsymbol{x}_t - \min\left(\frac{1}{\lambda_t}, \frac{\ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2} \right) \boldsymbol{g}_t \right) = \frac{\lambda_t}{\lambda_{t+1}} \boldsymbol{x}_t - \min\left(\frac{1}{\lambda_{t+1}}, \frac{\lambda_t}{\lambda_{t+1}} \frac{\ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2} \right) \boldsymbol{g}_t \ .$$

Instead, from (6), we have

$$\boldsymbol{x}_{t+1} = \frac{\lambda_t}{\lambda_{t+1}} \boldsymbol{x}_t - \min\left(\frac{1}{\lambda_{t+1}}, \frac{\ell_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2} \right) \boldsymbol{g}_t \ . \qquad\qquad \square$$