A Unified Modular Analysis of Online and Stochastic Optimization: Adaptivity, Optimism, Non-Convexity

Pooria Joulani[1] **András György**[2] **Csaba Szepesvári**[1]
[1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada
{pooria,szepesva}@ualberta.ca
[2]Department of Electrical and Electronic Engineering, Imperial College London, UK
a.gyorgy@imperial.ac.uk

Abstract

We present a simple unified analysis of adaptive Mirror Descent (MD) and Follow-the-Regularized-Leader (FTRL) algorithms for online and stochastic optimization in (possibly infinite-dimensional) Hilbert spaces. The analysis is modular in the sense that it completely decouples the effect of possible assumptions on the loss functions (such as smoothness, strong convexity, and non-convexity) and on the optimization regularizers (such as strong convexity, non-smooth penalties in composite-objective learning, and non-monotone step-size sequences). We demonstrate the power of this decoupling by obtaining generalized algorithms and improved regret bounds for the so-called "adaptive optimistic online learning" setting. In addition, we simplify and extend a large body of previous work, including several various AdaGrad formulations, composite-objective and implicit-update algorithms. In all cases, the results follow as simple corollaries within few lines of algebra. Finally, the decomposition enables us to obtain preliminary global guarantees for limited classes of non-convex problems.

1 Introduction

Online and stochastic optimization algorithms are state-of-the-art techniques for optimization in machine learning, analyzed under specific assumptions on the loss functions and the regularizers used by the algorithm. Previous work [10, 13] have attempted to obtain a unified understanding of these analyses. Nevertheless, a complete view that decouples the effect of these various assumptions is missing. The present paper provides this unified view. Our analysis captures, generalizes and considerably simplifies the scattered analysis techniques used under different assumptions, such as smoothness or strong convexity of losses (e.g., Agarwal and Duchi [1], Dekel et al. [5], [4]; see also Bubeck [3]), adaptive and composite-objective regularization (Duchi et al. [6, 7], McMahan and Streeter [11]), non-monotone regularization [14], and optimistic online learning [12]. [1]

The main idea is to decompose the regret into two parts: the first part only depends on the performance of the algorithm based on the first-order information that it receives, while the second part connects the assumptions about the losses to their first-order approximation given to the algorithm. Lemma 1 in Section 2.1 provides such a decomposition, which can be viewed as a refined version of the so-called "be-the-leader" style of analysis. Then, in Theorem 1, we bound the linear (first) part, using a careful analysis of the linear regret of generalized adaptive Follow-The-Regularized-Leader (FTRL) and Mirror Descent (MD) algorithms. The analysis applies to inifinite-dimensional Hilber spaces, and completely decouples the effect of assumptions about the problem setting, the losses, and the algorithm, allowing us to recover, improve, and considerably extend previous results.

[1]Due to space constraints, we have only provided the main parts of the framework in detail. An overview of the applications is given in Section 4, with further details provided in the extended version of the paper [9].

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

We use $\{c_t\}_{t=i}^{j}$ to denote the sequence $c_i, c_{i+1}, \ldots, c_j$, and $c_{i:j}$ to denote the sum $\sum_{t=i}^{j} c_t$, with $c_{i:j} := 0$ for $i > j$. We will work with a (possibly infinite-dimensional) Hilbert space $\mathcal{H}$ over the reals. We denote the extended real line by $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, and work with functions of the form $f : \mathcal{H} \to \overline{\mathbb{R}}$. Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be proper. Let $x \in \mathrm{dom}(f)$, and $z \in \mathcal{H}$. The *directional derivative* of f at x in the direction of z is defined as $f'(x; z) := \lim_{\alpha \downarrow 0} \frac{f(x+\alpha z) - f(x)}{\alpha}$, provided that the limit exists in $[-\infty, +\infty]$. The function f is *locally sub-differentiable* at x if there exists $g_x \in \mathcal{H}$ such that $\langle g_x, z \rangle \leq f'(x; z)$ for all $z \in \mathcal{H}$. The function f is called *directionally differentiable* if $f'(x; z)$ exists in $[-\infty, +\infty]$ for all $z \in \mathcal{H}$ at every $x \in \mathrm{dom}(f)$. Given a directionally differentiable f, we define a generalized[2] notion of Bregman divergence:

Definition 1 (Bregman divergence)**.** *Let f be directionally differentiable and $x \in \mathrm{dom}(f)$. The f-induced* Bregman divergence *from x is a function from $\mathcal{H} \to \overline{\mathbb{R}}$, given by*

$$\mathcal{B}_f(y, x) := \begin{cases} f(y) - f(x) - f'(x; y - x)\,. & \text{if } y \in \mathrm{dom}(f)\,, \\ +\infty & \text{if } y \notin \mathrm{dom}(f)\,. \end{cases} \tag{1}$$

2 Problem setting: online optimization

We study a general first-order iterative optimization setting that encompasses several common optimization scenarios, including online, stochastic, and full-gradient optimization. Consider a convex set $\mathcal{X} \subset \mathcal{H}$, a sequence of locally sub-differentiable functions $f_1, f_2, \ldots, f_T$ from $\mathcal{H}$ to $\overline{\mathbb{R}}$ with $\mathcal{X} \subset \mathrm{dom}(f_t)$ for all $t = 1, 2, \ldots, T$, and a first-order iterative optimization algorithm. The algorithm starts with an initial point x_1. Then, in each iteration $t = 1, 2, \ldots, T$, the algorithm suffers a loss $f_t(x_t)$ from the latest point x_t, receives a local sub-gradient g_t of f_t at x_t, and comes up with the next point x_{t+1}. Unlike Online Convex Optimization (OCO), at this stage we do not assume that f_t are convex. Our goal is to minimize the *regret* $R_T(x^*)$ against any $x^* \in \mathcal{X}$, defined as $R_T(x^*) = \sum_{t=1}^{T} f_t(x_t) - f_t(x^*)$. We will also discuss the special case of stochastic optimization setting, in which $f_t(\cdot) = F(\cdot, \xi_t)$ for i.i.d. $\xi_t \sim \mathcal{D}$. Defining $f(\cdot) := \mathbb{E}_{\xi \sim D} F(\cdot, \xi)$, our goal is to minimize the expected stochastic regret, $\bar{R}_T(x^*) = \mathbb{E}\Big\{\sum_{t=1}^{T} f(x_t) - f(x^*)\Big\}$.

2.1 Regret decomposition

Below, we provide a decomposition of $R_T(x^*)$, in terms of the *forward linear regret* of the algorithm.

Lemma 1 (regret decomposition)**.** *Let $x_1, x_2, \ldots, x_{T+1}$ be any sequence of points in $\mathcal{X}$. For $t = 1, 2, \ldots, T$, let $f_t : \mathcal{H} \to \overline{\mathbb{R}}$ be locally sub-differentiable with $\mathcal{X} \subset \mathrm{dom}(f_t)$, and let g_t be a local-subgradient of f_t at x_t. Let $\tilde{R}_T^+(x^*) = \sum_{t=1}^{T} \langle g_t, x_{t+1} - x^* \rangle$. Then,*

$$R_T(x^*) \leq \tilde{R}_T^+(x^*) + \sum_{t=1}^{T} \langle g_t, x_t - x_{t+1} \rangle - \sum_{t=1}^{T} \mathcal{B}_{f_t}(x^*, x_t)\,. \tag{2}$$

In particular, if $f_t, t = 1, 2, \ldots, T$ are differentiable, then $g_t = \nabla f_t(x_t)$, and

$$R_T(x^*) = \tilde{R}_T^+(x^*) + \sum_{t=1}^{T} \langle g_t, x_t - x_{t+1} \rangle - \sum_{t=1}^{T} \mathcal{B}_{f_t}(x^*, x_t)\,, \tag{3}$$

Proof. Expand both sides and use the definition of $B_{f_t}(x^*, x_t)$, and the fact that $-f'(x_t; x^* - x_t) \leq -\langle g_t, x^* - x_t \rangle$. □

3 The algorithms: ADA-FTRL and ADA-MD

The ADA-FTRL algorithm works with two sequences of *regularizers*, $p_1, p_2, \ldots, p_T$ and $q_0, q_1, q_2, \ldots, q_T$, where each p_t and q_t is a function from $\mathcal{H}$ to $\overline{\mathbb{R}}$, and can be built by the algorithm in an online manner using the information received up to time step t. We use p_t to distinguish

[2] If f is differentiable at x, then (1) matches the traditional definition of Bregman divergence.

the "proximal" part of ADA-FTRL's regularization: for all $t = 1, 2, \ldots, T$, p_t (but not necessarily q_t) is minimized over $\mathcal{X}$ at x_t, that is, $p_t(x_t) = \inf_{x \in \mathcal{X}} p_t(x)$, where x_t is the action taken by ADA-FTRL at time t. We define $r_0 := q_0$ and $r_t := p_t + q_t$ for $t = 1, 2, \ldots, T$. Similarly, the ADA-MD algorithm uses a sequence of regularizers $r_0, r_1, r_2, \ldots, r_T$, each a function from $\mathcal{H}$ to $\overline{\mathbb{R}}$. With the definitions above, the first choice x_1 of ADA-FTRL and ADA-MD is given by

$$x_1 \in \operatorname{argmin}_{x \in \mathcal{X}} \; r_0(x) \,. \tag{4}$$

Then, at $t = 1, 2, \ldots, T$, ADA-FTRL receives $g_t \in \mathcal{H}$ and selects the next point x_{t+1} such that

$$x_{t+1} \in \operatorname{argmin}_{x \in \mathcal{X}} \langle g_{1:t}, x \rangle + r_{0:t}(x) = \operatorname{argmin}_{x \in \mathcal{X}} \langle g_{1:t}, x \rangle + p_{1:t}(x) + q_{0:t}(x) \,, \tag{5}$$

whereas ADA-MD receives $g_t \in \mathcal{H}$ and selects

$$x_{t+1} \in \operatorname{argmin}_{x \in \mathcal{X}} \langle g_t, x \rangle + \mathcal{B}_{r_{0:t}}(x, x_t) \,. \tag{6}$$

For ADA-FTRL and ADA-MD, respectively, we only make the following assumption:

Assumption 1. *In* ADA-FTRL, *q_0 and $p_t, q_t, t = 1, 2, \ldots, T$, are proper, and for all $t = 1, 2, \ldots, T$, the function $p_t + r_{0:t-1}$ is directionally differentiable. In addition, for all $t = 0, 1, \ldots, T$, $\mathcal{X} \cap \operatorname{dom}(r_{0:t}) \neq \emptyset$, and the* argmin *sets that define x_{t+1} in* (4) *and* (5) *are non-empty.*

Assumption 2. *In* ADA-MD, *for all $t = 0, 1, \ldots, T$, r_t is proper, $S_t := \operatorname{dom}(r_{0:t})$ is convex, $x_t \in \operatorname{dom}(r_t)$, $r_{0:t}$ is directionally differentiable, and the* argmin *sets that define x_{t+1} in* (4) *and* (6) *are non-empty. In addition, for all $t = 0, 1, \ldots, T$, for all $y \in S_t$, the directional derivative $r'_{0:t}(y; \cdot)$ is real-valued and concave on $S_t - \{y\}$, i.e., for all $x_1, x_2 \in S_t$ and all $\alpha \in [0, 1]$,*

$$+\infty > r'_{0:t}(y; \alpha x_1 + (1-\alpha) x_2 - y) \geq \alpha r'_{0:t}(y; x_1 - y) + (1-\alpha) r'_{0:t}(y; x_2 - y) > -\infty \,.$$

Theorem 1 (forward regret of ADA-FTRL and ADA-MD)**.** *For any $x^* \in \mathcal{X}$, against any sequence of linear losses $\langle g_t, \cdot \rangle, t = 1, 2, \ldots, T$, the forward regret of* ADA-FTRL *under Assumption 1 satisfies*

$$\tilde{R}_T^+(x^*) \leq \sum_{t=0}^{T} (q_t(x^*) - q_t(x_{t+1})) + \sum_{t=1}^{T} (p_t(x^*) - p_t(x_t)) - \sum_{t=1}^{T} \mathcal{B}_{p_{1:t} + q_{0:t-1}}(x_{t+1}, x_t) \tag{7}$$

whereas the forward regret of ADA-MD *under Assumption 2 satisfies*

$$\tilde{R}_T^+(x^*) \leq \mathcal{B}_{r_0}(x^*, x_1) + \sum_{t=1}^{T} \mathcal{B}_{r_t}(x^*, x_t) - \sum_{t=1}^{T} \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t) \,. \tag{8}$$

4 Applications

Assumption 3. *For all $t = 1, 2, \ldots, T$, $\mathcal{B}_{f_t}(x^*, x_t) \geq 0$ (e.g., when f_t is convex).*

Assumption 4. *In stochastic optimization, for all $x \in \mathcal{X}$, $\mathcal{B}_f(x^*, x) \geq 0$ (e.g., when f is convex).*

Assumption 5. *In stochastic optimization, f is differentiable and 1-smooth w.r.t. some norm $\|\cdot\|$. Furthermore, $\mathbb{E}\left[\|\sigma_t\|_*^2 \middle| x_t\right] \leq \sigma^2$ for all t, where $\sigma_t = \nabla f(x_t) - g_t$.*

Assumption 6. *In* ADA-FTRL, *$p_t, t = 1, 2, \ldots, T$ and $q_t, t = 0, 1, \ldots, T$ are non-negative.*

Assumption 7. *In* ADA-FTRL *(respectively,* ADA-MD*), for all $t = 0, 1, \ldots, T$, $p_{1:t} + q_{0:t-1}$ (respectively, $r_{0:t}$) is 1-strongly convex w.r.t. some norm $\|\cdot\|_{(t)}$.*

Assumption 8. *In* ADA-MD*, the losses are 1-strongly convex w.r.t. the regularizers, that is, $\mathcal{B}_{f_1}(x^*, x_1) \geq \mathcal{B}_{r_{0:1}}(x^*, x_1)$, and $\mathcal{B}_{f_t}(x^*, x_t) \geq \mathcal{B}_{r_t}(x^*, x_t)$ for $t = 2, \ldots, T$.*

Assumption 9. *In stochastic optimization with* ADA-MD*, the objective f is 1-strongly convex w.r.t. the regularizers, that is, $\mathcal{B}_f(x^*, x_1) \geq \mathcal{B}_{r_{0:1}}(x^*, x_1)$, and $\mathcal{B}_f(x^*, x_t) \geq \mathcal{B}_{r_t}(x^*, x_t)$ for $t = 2, \ldots, T$.*

Table 1 provides a summary of the standard results that are recovered and generalized using our framework (in particular, we obtain generalized ADAGRAD [6] for $p_t \equiv 0$, and FTRL-PROX [11] for $r_t \equiv 0$). Detailed derivations, typically a few lines of algebra, are provided in the full paper [9]. Compared to previous work (e.g., [10, 13]), our analysis recovers all previous results, applies to infinite-dimensional Hilbert spaces, relaxes several assumptions on the regularizers and losses, extends ADAGRAD-style algorithms to stochastic optimization with smooth losses, and obtains the best known constants in the regret bound. Furthermore, the complete decomposition of assumptions

Setting / Algorithms	Assumptions	Regret / Expected Stochastic Regret Bound
Online / Stochastic Optimization ADA-FTRL	1, 3/4, 6, 7 (Theorem 2)	$q_{0:T-1}(x^*) + p_{1:T}(x^*) + \sum_{t=1}^T \frac{1}{2}\|g_t\|^2_{(t),*}$
Online / Stochastic Optimization ADA-MD	2, 3/4, 7 (Theorem 2)	$\mathcal{B}_{r_0}(x^*, x_1) + \sum_{t=1}^T \mathcal{B}_{r_t}(x^*, x_t) + \frac{1}{2}\|g_t\|^2_{(t),*}$
Strongly-convex Online / Stochastic Optimization ADA-MD	2, (3/4), 7, 8/9 (Theorem 3)	$\sum_{t=1}^T \frac{1}{2}\|g_t\|^2_{(t),*}$
Smooth Stochastic Optimization ADA-FTRL	1, 4, 5, 6, 7 (Theorem 4)	$\frac{1}{2}\|x^*\|^2 + q_{0:T-1}(x^*) + p_{1:T-1}(x^*)$ $+ \sum_{t=1}^{T-1} \frac{1}{2}\|\sigma_t\|^2_{(t),*} + f(x_1) - f(x^*)$
Smooth Stochastic Optimization ADA-MD	2, 4, 5, 7 (Theorem 4)	$\frac{1}{2}\|x^* - x_1\|^2 + \mathcal{B}_{r_0}(x^*, x_1) + \sum_{t=1}^{T-1} \mathcal{B}_{r_t}(x^*, x_t)$ $+ \sum_{t=1}^{T-1} \frac{1}{2}\|\sigma_t\|^2_{(t),*} + f(x_1) - f(x^*)$

Table 1: Recovered and generalized standard results. A number in parentheses indicates that the assumption is not directly required, but is implied by the other assumptions.

makes the analysis modular, and thus simpler to extend, as we show in the following sections. Finally, note that for ADA-FTRL, composite-objective bounds simply follow by adding the non-smooth penalty to q_t. The case of ADA-MD is more detailed, and is provided in the full paper [9].

Implicit-update online convex optimization and Follow-The-Leader: Suppose that f_t is convex, and add $\mathcal{B}_{f_t}(\cdot, x_t)$ to $p_t(\cdot)$ in ADA-FTRL (respectively, add f_t to r_t in ADA-MD). Then, the optimization problems (5) and (6) will be equivalent to implicit-update ADA-FTRL and ADA-MD (which use the whole loss f_t rather than the lienarized loss $\langle g_t, x\rangle$ in the optimization). For $r_t \equiv 0$, this gives the Follow-The-Leader algorithm. Canceling the resulting p_t terms with the negative Bregman divergences in (2) extends the bounds in Table 1 to implicit-update online learning, but under relaxed assumptions. See the extended version of the paper [9] for details.

Working with non-convex functions: Note that in stochastic optimization, we do not need each f_t to be convex; it suffices for f to be convex. As such, all of our bounds for stochastic optimization work for stochastic minimization of convex sums of non-convex functions, which are useful in certain applications [8]. In addition, in fact Assumptions 3 and 4 are more general than convexity, and hold, in particular, for certain classes of quasi-convex functions. We provide examples of such functions, as well as other possible routes for non-convex optimization, in the extended version of the paper [9].

Adaptive optimistic online learning: Let $\tilde{g}_t, t = 1, 2, \ldots, T+1$, be any sequence of vectors in $\mathcal{H}$. Suppose that we run ADA-FTRL, with regularizers satisfying the corresponding assumptions (Assumption 1 or 2), on a sequence of convex losses $f_1, f_2, \ldots, f_T$, producing the sequence of points $x_{t+1}, t = 0, 1, 2, \ldots, T$ given by $x_{t+1} \in \operatorname{argmin}_{x\in\mathcal{X}} \langle g_{1:t} + \tilde{g}_{t+1}, x\rangle + p_{1:t}(x) + q_{0:t}(x)$, where $g_t \in \delta f_t(x_t), t = 1, 2, \ldots, T$. That is, we run ADA-FTRL, but we also incorporate $\tilde{g}_{t+1}$ as a "guess" of the future loss g_{t+1} that the algorithm will suffer. It is easy to see that this is exactly equivalent to running ADA-FTRL on the sequence $g_t - \tilde{g}_t$ with regularizers $\tilde{q}_t = q_t + \langle \tilde{g}_{t+1}, \cdot\rangle$, i.e., $x_{t+1} \in \operatorname{argmin}_{x\in\mathcal{X}} \langle g_{1:t} - \tilde{g}_{1:t}, x\rangle + p_{1:t}(x) + \tilde{q}_{0:t}(x)$. Then, Theorem 1 will upper-bound the linear cheating regret $\sum_{t=1}^T \langle g_t - \tilde{g}_t, x_{t+1} - x^*\rangle$. Rearranging and combining with Lemma 1, the exact same manipulations as for Table 1 give $R_T(x^*) \le \sum_{t=0}^{T-1} (q_t(x^*) - q_t(x_{t+1})) + \sum_{t=1}^T (p_t(x^*) - p_t(x_t)) + \sum_{t=1}^T \frac{1}{2}\|g_t - \tilde{g}_t\|^2_{(t),*}$. This bound recovers all the regret bounds in Theorems 1-7 of Mohri and Yang [12] (and hence their corollaries). In particular, the modular analysis makes the modifications discussed in the previous section, e.g., stochastic optimization and embedding composite-objective regularizers in q_t, naturally extend to the bound above. Instead of analyzing each case from scratch, our framework enables us to considerably simplify the analysis of this new setting, relax the assumptions and generalize the results, and improve the sub-optimal constants in the regret bounds to the best-known constant $(1/2)$.

It should also be noted that the decomposition of assumptions has also helped us avoid certain proof artifacts. For example, in their Theorem 3, Mohri and Yang [12] obtained a sub-optimal bound

for proximal regularizers, because the effects of proximal and non-proximal regularizers were not separated. Our corresponding bound, however, retains the best bound for proximal regularizers, since the analysis separates the effect of these two types of regularizers early on.

5 Conclusion and future work

We provided a generalized, unified and modular framework for analyzing online and stochastic optimization algorithms, and demonstrated its flexibility on several examples. Our framework is also applicable to other problem settings, such as adaptive algorithms like AdaDelay [14]. Exploring these and other applications of this framework is left for future work.

Acknowledgments

This work was supported by AICML and NSERC.

References

[1] Alekh Agarwal and John Duchi. Distributed delayed stochastic optimization. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 873–881, 2011.

[2] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

[3] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[4] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[5] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 13(1):165–202, January 2012. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2503308.2188391`.

[6] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July 2011.

[7] John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, pages 14–26. Citeseer, 2010.

[8] Shripad Gade and Nitin H Vaidya. Distributed optimization of convex sum of non-convex functions. *arXiv preprint arXiv:1608.05401*, 2016.

[9] Pooria Joulani, Csaba Szepesvári, and András György. A unified modular analysis of online and stochastic optimization: Adaptivity, optimism, non-convexity. Technical report, 2016. URL `http://ualberta.ca/~pooria/modular-opt.pdf`.

[10] H. Brendan McMahan. Analysis techniques for adaptive online learning. *CoRR*, abs/1403.3465, 2014.

[11] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory*, 2010.

[12] Mehryar Mohri and Scott Yang. Accelerating online convex optimization via adaptive prediction. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 848–856, 2016.

[13] Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

[14] Suvrit Sra, Adams Wei Yu, Mu Li, and Alexander J Smola. Adadelay: Delay adaptive distributed stochastic convex optimization. *arXiv preprint arXiv:1508.05003*, 2015.