Parameter-free projected gradient descent

Evgenii Chzhen, Christophe Giraud, Gilles Stoltz

Parameter-free projected gradient descent

Evgenii Chzhen **Christophe Giraud** **Gilles Stoltz**
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
{evgenii.chzhen, christophe.giraud, gilles.stoltz}@universite-paris-saclay.fr

Abstract

We consider the problem of minimizing a convex function over a closed convex set, with Projected Gradient Descent (PGD). We propose a fully parameter-free version of AdaGrad, which is adaptive to the distance between the initialization and the optimum, and to the sum of the square norm of the subgradients. Our algorithm is able to handle projection steps, does not involve restarts, reweighing along the trajectory or additional gradient evaluations compared to the classical PGD. It also fulfills optimal rates of convergence for cumulative regret up to logarithmic factors. We provide an extension of our approach to stochastic optimization and conduct numerical experiments supporting the developed theory.

1 Introduction

In this work we study the problem of minimizing a convex function f over a closed, possibly unbounded, convex set $\Theta \subseteq \mathbb{R}^d$. Our main goal is to provide a variant of AdaGrad [SM10, DHS11] which is adaptive to the distance $\|x_1 - x_*\|$ between the initialization $x_1 \in \Theta$ and a minimizer $x_* \in \Theta$, which is assumed to exist. More precisely, we provide a Projected Gradient Descent (PGD) algorithm of the form

$$x_{t+1} = \mathrm{Proj}_{\Theta}\left(x_t - \eta_t g_t\right) \quad \text{with} \quad \eta_t = \frac{2^{k_t}}{H\big(\sum_{s\leqslant t}\|g_s\|^2\big)}\,,$$

where $g_t \in \partial f(x_t)$ is a sub-gradient of f at x_t, $\mathrm{Proj}_{\Theta}(\cdot)$ is the Euclidean projection operator onto closed convex Θ, $H(x) = \sqrt{(x+1)\log(\mathrm{e}(1+x))}$ and k_t is an automatically tuned sequence by Algorithm 1. Unlike recent works on the subject [DM23, CH23], we provide bounds on the cumulative regret of the form

$$R_T := \sum_{t=1}^{T}\big(f(x_t) - f(x_*)\big)\,,$$

where x_* is any minimizer of f over Θ. Using standard online-to-batch conversion, we also have by convexity $f(\bar{x}_T) - f(x_*) \leqslant R_T/T$, for $\bar{x}_T$ being the average of $x_1, \dots, x_T$.

In the classical case where f is assume to be L-Lipschitz, it is well known that setting $\eta_t = \frac{\|x_1 - x_*\|}{L\sqrt{T}}$ gives the optimal rate of convergence [N+18]:

$$R_T \leqslant \|x_1 - x_*\| L\sqrt{T}\,.$$

However, such a choice requires f to be Lipschitz, and the knowledge of three quantities: *1)* distance to the optimum $\|x_1 - x_*\|$; *2)* Lipschitz constant L; *3)* optimization horizon T. Should the distance $\|x_1 - x_*\|$ be known, one could set $\eta_t = \frac{\|x_1 - x_*\|}{\sqrt{\sum_{s=1}^{t}\|g_s\|^2}}$, resulting in AdaGrad algorithm [SM10, DHS11].

Preprint. Under review.

Algorithm 1: FREE ADAGRAD

Input: $x_1 \in \mathbb{R}^d, \Theta \subset \mathbb{R}^d, \gamma_0 > 0$

Initialization: $\Gamma_1^2 = 0, k_0 = 1, S_0 = 0, \gamma_k = \gamma_0 2^k$ for $k \geqslant 1$

for $t \geqslant 1$ **do**

- $g_t \in \partial f(x_t)$ // get subgradient
- $S_t = S_{t-1} + \|g_t\|^2$ // cumulative grad-norm
- $h_t = \sqrt{(S_t + 1)\log(\mathrm{e}(1 + S_t))}$ // update $h_t \geqslant h_{t-1}$
- $B_{t+1}(k) = \dfrac{2\gamma_k}{\sqrt{k}} + \sqrt{\Gamma_t^2 + \gamma_k^2 \|g_t\|^2 / h_t^2}$ // define the threshold
- $x_t^+(k) = \mathrm{Proj}_\Theta\left(x_t - \frac{\gamma_k}{h_t} g_t\right)$ // probing step
- $k_t = \min\left\{k \geqslant k_{t-1} \,:\, \|x_t^+(k) - x_1\| \leqslant B_{t+1}(k)\right\}$ // find the step size
- $x_{t+1} = x^+(k_t)$ // make the step
- $\Gamma_{t+1}^2 = \Gamma_t^2 + \gamma_{k_t}^2 \frac{\|g_t\|^2}{h_t^2}$ // update Γ_t^2

end

Output : Trajectory $(x_t)_{t \geqslant 1}$

For this choice of η_t, without Lipschitz assumption, we have the upper bound on the regret

$$R_T \leqslant c\|x_1 - x_*\| \sqrt{\sum_{t=1}^T \|g_t\|^2}\,. \tag{1}$$

In practice the distance $\|x_1 - x_*\|$ is unknown. When an upper bound D_* on $\|x_1 - x_*\|$ is available, typically the diameter of Θ when Θ is bounded, $\|x_1 - x_*\|$ can be replaced by D_* in η_t. The ADAGRAD algorithm then fulfills (1) with $\|x_1 - x_*\|$ replaced by D_*. This bound can be very sub-optimal yet, when $\|x_1 - x_*\|$ is much smaller than D_*. Worse, when Θ is unbounded, no bound D_* on $\|x_1 - x_*\|$ is available, without additional information.

Our objective is to provide a variant of the ADAGRAD step-size tuning, not requiring f to be Lipschitz, nor any knowledge on $\|x_1 - x_*\|$ or T, while still fulfilling the regret bound (1) up to a log factor. Our contribution can be placed alongside the ever expanding literature of parameter-free optimization algorithm [DM23, CH23, Cut19, MS12, MO14a, MK20, OP21, OT17, ZCP22, JC22, OP16], discussed below Theorem 1.

Main contributions. Let us describe our three main contributions

1. we propose a simple tuning of PGD, we call FREE ADAGRAD, with no line-search, no cold-restart, no gradient transformation, and no computations of extra gradients;
2. we handle any finite convex function f (no Lipschitz condition), over any possibly unbounded constraint set Θ;
3. we provide regret bounds like $R_T = \widetilde{\mathcal{O}}\big((\|x_1 - x_*\| + 1)\sqrt{1 + \sum_{t \leqslant T} \|g_t\|^2}\big)$, where $\tilde{\mathcal{O}}$ hides log-factor, but no additional terms.

We also partially extend our results to the Stochastic Gradient Descent setting.

Notation. For any $a, b \in \mathbb{R}$ we denote by $a \vee b$ (*resp.* $a \wedge b$) the largest (*resp.* the smallest) of the two. We denote by $\|\cdot\|$ the Euclidean norm and by $\langle \cdot, \cdot \rangle$ the standard inner product in $\mathbb{R}^d$. For $\Theta \subset \mathbb{R}^d$, we denote by $\mathrm{Proj}_\Theta(\cdot)$ the Euclidean projection operator onto Θ. We denote by $\log_2(\cdot)$ and $\log(\cdot)$ the base 2 and the natural logarithms respectively. The base of the natural logarithm is denoted by e.

2 Main result

We make the following assumption, which is necessary for the meaningful treatment of the problem.

Assumption 1. *The set $\Theta \subseteq \mathbb{R}^d$ is closed convex, $\mathcal{D} \subseteq \mathbb{R}^d$ is open such that $\Theta \subseteq \mathcal{D}$. The function $f : \mathcal{D} \to \mathbb{R}$ is convex on Θ and there exists a bounded minimizer $x_* \in \arg\min_{x \in \Theta} f(x)$.*

Let us highlight that we do not assume that the subgradients g_t are uniformly bounded, that is, we do not require f to be globally Lipschitz. This stays in contrast with the literature on online

convex optimization (OCO). Indeed, OCO lower bounds imply that without any prior knowledge a regret $\widetilde{\mathcal{O}}(\|x_1 - x_*\|(\sum_{t\leqslant T}\|g_t\|^2)^{1/2})$ is not achievable [CB17, Cut19, MK20] and higher order terms either in T or in $\|x_1 - x_*\|$ are necessary. For example, we are able to handle $\Theta = [0, +\infty)^d$ and $f(x) = \sum_{i=1}^n \exp(\|x - a_i\|/\sigma_i)$ for some $a_i \in \mathbb{R}^d$ and $\sigma_i > 0$.

The proposed method, that we call FREE ADAGRAD, is summarized in Algorithm 1. It consists in simple projected gradient steps at every round $t \geqslant 1$, but with additional cheap condition on Line 7 that is checked on every iteration. If the condition on this line is satisfied for $k = k_{t-1}$, then the algorithm makes (almost) the usual ADAGRAD step, otherwise, the step size is doubled and the condition is checked again. We underline that the sequence of integers $\{k_t : t \geqslant 1\}$ in Algorithm 1 is non-decreasing, and we prove in Eq. (12) that it is upper-bounded by $2(\log_2(1 + \|x_1 - x_*\|/\gamma_0) + 1)$. Thus, there are only a finite number of doublings in the sequence. The only input parameter of the algorithm is γ_0 that can be seen as an initial lower-bound guess for $\|x_1 - x_*\|$ and can be taken arbitrary small only incurring additional $\sqrt{\log(1/\gamma_0)}$ factor. Note that once the step-size is doubled at some time $t \geqslant 1$, Algorithm 1 continues the optimization from x_t without cold-restart.

Theorem 1. *Let Assumption 1 be satisfied. Let $S_T = \sum_{t=1}^T \|g_t\|^2$. For any $\gamma_0 > 0$, let $D_{\gamma_0} := \|x_1 - x_*\| \vee \gamma_0$, Algorithm 1 satisfies for some universal $c > 0$*

$$R_T \leqslant cD_{\gamma_0}\sqrt{(S_{T+1} + 1)\log(1 + S_{T+1})\log(1 + D_{\gamma_0})}\log\log(1 + S_T)\,.$$

A large part of the literature on parameter-free optimization considers the context of online convex optimization with L-Lipschitz functions. A series of papers [MO14a, ZCP22, JC22, CO18] have produced algorithms, mainly based on coin betting, enjoying regret bounds $\mathcal{O}(D_{\gamma_0}\sqrt{S_T\log(1 + D_{\gamma_0}S_T/\gamma_0)})$, up to lower order terms. Such a regret bound is not achievable in online convex optimization when L is unknown as shown in [CB17]. Some papers [Cut19, MK20, JC22] consider yet the case where L is unknown, and provide regret bound including additional terms depending on L and on higher order of D_{γ_0}.

If only the optimization error is of concern (and not the regret), bounds with log-factors replaced by log-log factors have been produced by [CH23, DM23], breaking the barrier of online-to-batch conversion, but still requiring some knowledge about the Lipschitz constant L. Relying on binary search, [CH23] construct an adaptive algorithm for the problem of stochastic optimization, while [DM23] provide an adaptive version of dual averaging and gradient descent algorithms, without allowing for projection step and requiring a careful weighted averaging along the trajectory to obtain the final solution. In contrast, we do not require Lipschitz condition, we handle the projection step and our bounds are valid for the usual average.

The closer to us, in a setting of online convex optimization with L-Lipschitz functions, [MS12] propose a tuning of GD (without projection) which is based on a doubling trick with cold-restarts and which requires the knowledge of L. This algorithm is shown to be adaptive to $\|x_1 - x_*\|$ at the price of loosing a log factor $\log(D_{\gamma_0}T/\gamma_0)$ in the regret bound. In Appendix D, we show that the algorithm of [MS12] can be seen as a specific instantiation of Algorithm 1, with the major difference that cold-restart are performed when doubling the step-size.

3 Warm-up: simple analysis and intuition

Before proceeding to the analysis of the FREE ADAGRAD algorithm 1, we explain the main ideas behind our step-size scheme in the following simpler setup.

Simple warm-up setup

1) the norm of the subgradients are uniformly bounded by some known L, i.e. $\|g_t\| \leqslant L$,

2) the optimization is unconstrained, i.e. $\Theta = \mathbb{R}^d$,

3) the time horizon T is fixed in advance.

In this case, we can replace $(h_t)_{t\geqslant 1}$ set on Line 4 of Algorithm 1 by the constant sequence $h_t = L\sqrt{T}$, and the choice $\gamma = \|x_1 - x_*\|$ is known to achieve the optimal rates for the regret $\|x_1 - x_*\|L\sqrt{T}$, see e.g. [N+18]. In this context, the overall strategy of Algorithm 1 is to start from a small value

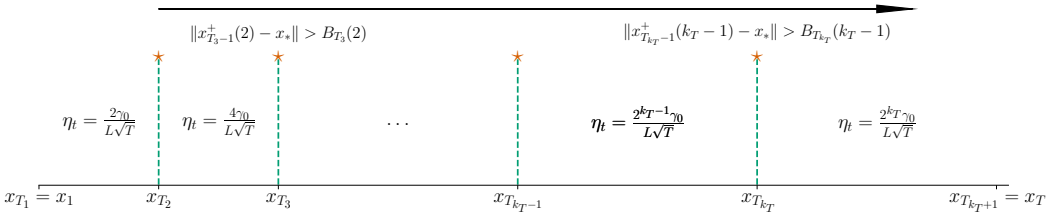

Figure 1: Schematic illustration of the algorithm in the simplest case.

γ_0 for γ, and then track $\|x_t - x_1\|$ in order to detect if $\gamma < \|x_1 - x_*\|$. If so, γ is doubled. The algorithm then increases the value γ until reaching the level $\|x_t - x_1\|$.

In order to keep the analysis simple in this warm-up section, we replace the threshold $B_{t+1}(k)$ Line 5 of Algorithm 1 by $B^{\text{simple}}_{t+1}(k) = 3\gamma_k$ (recall that $\gamma_k = \gamma_0 2^k$), what eventually leads to a slightly worse bound. The gradient step and the step-size choice are then simply

$$x_t^+(k) = x_t - \frac{\gamma_k}{L\sqrt{T}} g_t \quad \text{and} \quad k_t = \min\left\{k \geqslant k_{t-1} \,:\, \|x_t^+(k) - x_1\| \leqslant 3\gamma_k\right\}. \tag{2}$$

Below, we sketch the main arguments, and we refer to Appendix A for all the details.
The first ingredient is the text-book decomposition using subgradient upper-bound: for any $k \geqslant 1$

$$\begin{aligned} 0 \leqslant f(x_t) - f(x_*) \leqslant \langle g_t, x_t - x_*\rangle &= \frac{\gamma_k}{2L\sqrt{T}}\|g_t\|^2 + \frac{L\sqrt{T}}{2\gamma_k}\left(\|x_* - x_t\|^2 - \|x_t^+(k) - x_*\|^2\right) \\ &\leqslant \frac{\gamma_k L}{2\sqrt{T}} + \frac{L\sqrt{T}}{2\gamma_k}\left(\|x_* - x_t\|^2 - \|x_t^+(k) - x_*\|^2\right). \end{aligned} \tag{3}$$

It follows from this bound, a one-step deviation upper-bound

$$\|x_t^+(k) - x_*\|^2 \leqslant \|x_t - x_*\|^2 + \gamma_k^2/T\,.$$

Summing this bound over t, we get a first important bound on the distance to optimum

$$\|x_t^+(k) - x_*\|^2 \leqslant \|x_1 - x_*\|^2 + \sum_{s=1}^{t-1}\frac{\gamma_{k_s}^2}{T} + \frac{\gamma_k^2}{T} \;\leqslant\; \|x_1 - x_*\|^2 + \gamma_k^2, \quad \text{for all } k \geqslant k_{t-1}\,, \tag{4}$$

and then another important bound on the distance to initialization

$$\|x_t^+(k) - x_1\| \leqslant \|x_1 - x_*\| + \|x_t^+(k) - x_*\| \;\leqslant\; 2\|x_1 - x_*\| + \gamma_k, \quad \text{for all } k \geqslant k_{t-1}\,, \tag{5}$$

where the last inequality follows from (4) and the sub-additivity of square-root.

Controlling the number of phases. The bound (5) plays a central role in our step-size tuning. Indeed, we observe that if $\|x_t^+(k_{t-1}) - x_1\| > 3\gamma_{k_{t-1}}$, then it means that $\gamma_{k_{t-1}} < \|x_1 - x_*\|$, and our step-size tuning then increases k until the condition $\|x_t^+(k) - x_1\| \leqslant 3\gamma_k$ is met. In addition, we check below that the design of $B^{\text{simple}}_{t+1}(k)$ ensures that we have $k_t \leqslant k^*$ for all $t \leqslant T$, where $k^* \geqslant 1$ is the integer defined by $\gamma_{k^*-1} \leqslant D_{\gamma_0} := \|x_* - x_1\| \vee \gamma_0 < \gamma_{k^*}$, and fulfilling

$$k^* \leqslant 1 + \log_2\left(\frac{\|x_* - x_1\|}{\gamma_0} \vee 1\right) = \log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right), \quad \text{and} \quad \gamma_{k^*} \leqslant 2D_{\gamma_0}\,. \tag{6}$$

Indeed, if $k_{t-1} \leqslant k^*$, then (5) ensures that

$$\|x_t^+(k^*) - x_1\| \leqslant 2\gamma_{k^*} + \gamma_{k^*} = B^{\text{simple}}_{t+1}(k^*)\,,$$

so $k_t \leqslant k^*$, and by induction the property holds for all $t \leqslant T$.

Bounding the regret. Let us now upper-bound the regret. We denote by $[T_k, T_{k+1} - 1]$ the interval where $k_t = k$, with the convention $T_{k+1} = T_k$ if we never have $k_t = k$, see Figure 1 for schematic

illustration. Summing the central equation (3), the regret can then be decomposed as follows

$$\begin{aligned} R_T &= \sum_{k=1}^{k_T} \sum_{t=T_k}^{T_{k+1}-1} (f(x_t) - f(x_*)) \\ &\leqslant \sum_{k=1}^{k^*} \left(\frac{\gamma_k L}{2\sqrt{T}}(T_{k+1} - T_k) + \frac{L\sqrt{T}}{2\gamma_k} \left(\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2 \right) \right) \\ &\leqslant \frac{L\sqrt{T}}{2} \left(\gamma_{k^*} + \sum_{k=1}^{k^*} \frac{1}{\gamma_k} \left(\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2 \right) \right). \end{aligned} \tag{7}$$

From the step-size rule, we have that $\|x_{T_{k+1}} - x_1\| \leqslant B_{T_{k+1}}(k_{T_{k+1}-1}) = 3\gamma_k$, and from (4) we have $\|x_{T_k} - x_*\|^2 \leqslant \|x_1 - x_*\|^2 + \gamma_{k-1}^2$, so we can upper bound the last term in the right-hand side of (7)

$$\begin{aligned} \|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2 &\leqslant \|x_1 - x_*\|^2 + \gamma_{k-1}^2 - \left[\|x_1 - x_*\| - \|x_{T_{k+1}} - x_1\| \right]_+^2 \\ &\leqslant \gamma_{k-1}^2 + \|x_1 - x_*\|^2 - \left[\|x_1 - x_*\| - 3\gamma_k \right]_+^2 \\ &\leqslant \frac{1}{4}\gamma_k^2 + 6\gamma_k \|x_1 - x_*\|, \end{aligned} \tag{8}$$

where the last inequality follows from the basic inequality $\Delta^2 - [\Delta - B]_+^2 \leqslant 2\Delta B$, for all $\Delta, B \geqslant 0$. Substituting (8) in (7) and using the bound (6), we end with the upper-bound

$$R_T \leqslant \frac{L\sqrt{T}}{2} \left[\gamma_{k^*} + \frac{\gamma_{k^*+1}}{4} + 6k^* \|x_1 - x_*\| \right] \leqslant L\sqrt{T} \left[3\|x_1 - x_*\| \log_2 \left(\frac{2D_{\gamma_0}}{\gamma_0} \right) + 2D_{\gamma_0} \right]. \tag{9}$$

The bound (9) for Algorithm 1, then matches the optimal rate $\|x_1 - x_*\| L\sqrt{T}$ obtained with the oracle step size $\eta = \|x_1 - x_*\|/(L\sqrt{T})$, up to a factor $\log_2(D_{\gamma_0}/\gamma_0)$.

It turns out that, in this L-Lipschitz setting, it is possible to adapt to $\|x_1 - x_*\|$ with a bound $\mathcal{O}\big(D_{\gamma_0} L\sqrt{T \log_2(D_{\gamma_0}/\gamma_0)}\big)$ on the regret, by, for example, using coin betting [MO14b, OP16]. We achieve such tighter bound with PGD with a better tuning of the threshold $B_{t+1}(k)$ which is explained in the next section.

3.1 Improving log factor by better tuning of $B_{t+1}(k)$

Previous section gave the basic intuition, explaining why such a doubling strategy works. Yet, our choice of B_{t+1} on Line 5 of Algorithm 1 differs from $B_{t+1}^{\text{simple}}(k) = 3\gamma_k$. Remaining in the simple setup of warmup, let us explain two key ingredients, which eventually lead to our choice of B_{t+1} on Line 5 of Algorithm 1. The first ingredient is to track more tightly the upper-bound on $\|x_t^+(k) - x_1\|$. Indeed, we can improve the Bound (4) by keeping $\|x_t^+(k) - x_*\|^2 \leqslant \|x_1(k) - x_1\|^2 + \Gamma_t^2 + \gamma_k^2/T$ instead of relying on the last bound in (4). Hence, we can replace (5) by

$$\|x_t^+(k) - x_1\| \leqslant 2\|x_1 - x_*\| + \sqrt{\Gamma_t^2 + \gamma_k^2/T}, \tag{10}$$

in order to implicitly track the value $\|x_1 - x_*\|$. This improved tracking alone is not enough in order to improve the log factor. Indeed, choosing $B_{t+1}(k) = 2\gamma_k + \sqrt{\Gamma_t^2 + \gamma_k^2/T}$, still introduces $\log(D_{\gamma_0})$ term. To improve the log factor, our second ingredient is to choose a slightly smaller threshold B_{t+1}, at the price of possibly moderately increasing the number k_T of doubling. In particular, setting

$$B_{t+1}(k) = \frac{2\gamma_k}{\sqrt{k}} + \sqrt{\Gamma_t^2 + \gamma_k^2/T}, \tag{11}$$

we get that $k_T \leqslant k^* + 0.5\log_2(k^*) + 1.25$, and instead of $\log(D_{\gamma_0})$ we have $\sqrt{\log(D_{\gamma_0})}$ in the regret bound (see Appendix A for details). Combining everything together, we get the bound

$$R_T \leqslant 10 D_{\gamma_0} L\sqrt{T} \sqrt{2\log_2(2D_{\gamma_0}/\gamma_0)},$$

for algorithm in (2) with $3\gamma_k$ replaced by $B_{t+1}(k)$ in (11) (see Theorem 4 in Appendix A). While, the above discussion was still assuming the simple setup of L-Lipschitz function f, known L and T, we are able to generalize the above argument to nearly arbitrary convex f and unknown T.

4 Meta theorem: a general case of Algorithm 1

In this section, we provide a unified analysis of Algorithm 1, that is valid under the minimal Assumption 1, and for the choice of *any positive non-decreasing sequence* $(h_t)_{t\geqslant 1}$ on Line 4 of Algorithm 1. Our main result, stated in Theorem 1, is obtained as a consequence of this general result, and is made precise in Corollary 1.

Theorem 2. *Let Assumption 1 be satisfied. For any $\gamma_0 > 0$, for any positive non-decreasing $(h_t)_{t\geqslant 1}$ on Line 4 of Algorithm 1, Algorithm 1 satisfies*

$$\sum_{t=1}^{T}(f(x_t)-f(x_*)) \leqslant h_{T+1}\left(2\left\|x_1-x_*\right\|\sqrt{k_T}\left(2+\sqrt{\frac{1}{3}\sum_{t=1}^{T}\frac{\|g_t\|^2}{h_t^2}}\right)+\gamma_{k_T}\sum_{t=1}^{T}\frac{\|g_t\|^2}{h_t^2}\right).$$

It is interesting to observe that the term $\sum_{t\leqslant T}\|g_t\|^2/h_t$, that often appears in the analysis of AdaGrad is absent in our bound. Instead, we have $\sum_{t\leqslant T}\|g_t\|^2/h_t^2$ which behaves slightly worse and hence requires additional correction of h_t (extra log factor) to ensure convergence. The proof of Theorem 2, which can be found in Appendix B, is based on the following general lemma.

Lemma 1. *Let Assumption 1 be satisfied. Consider the following algorithm for $t \geqslant 1$*

$$x_{t+1} = \mathrm{Proj}_{\Theta}\left(x_t - \frac{\gamma}{h_t}g_t\right),$$

where $g_t \in \partial f(x_t)$, $(h_t)_{t\geqslant 1}$ is non-decreasing and positive, and $x_1 \in \mathbb{R}^d$. For all $T > 1$, and all $x_1 \in \mathbb{R}^d$, we have

$$\sum_{t=1}^{T}(f(x_t)-f(x_*)) \leqslant h_{T+1}\left(\frac{\|x_1-x_*\|^2-\|x_{T+1}-x_*\|^2}{2\gamma}+\frac{\gamma}{2}\sum_{t=1}^{T}\frac{\|g_t\|^2}{h_t^2}\right).$$

The above lemma replaces the key inequality (3) that was available for one step of PGD in the simplest case. However, since the step-size in our case is time-varying, we rather need a variant of this inequality over the whole trajectory. While simple to prove, it seems that this result is novel and could be of independent interest.

Finally, to obtain the bound of Theorem 1, we only need to bound the number of phases k_T. Note that the intuition of the previous section still applies in this case, yet, the actual bound on k_T is more refined—it gives better constants and improves logarithmic factors.

Lemma 2. *Let Assumption 1 be satisfied. For any $\gamma_0 > 0$, and any non-decreasing positive $(h_t)_{t\geqslant 1}$, Algorithm 1 satisfies for $T \geqslant 2$*

$$k_T \leqslant k^* + \frac{1}{2}\log_2(k^*) + \frac{5}{4} \qquad \text{and} \qquad \gamma_{k_T} \leqslant \frac{5}{2}\sqrt{k^*}\left(\gamma_0 2^{k^*}\right),$$

where k^* is such that $\gamma_0 2^{k^*-1} \leqslant \|x_1-x_*\| \vee \gamma_0 \leqslant \gamma_0 2^{k^*}$. Furthermore, $k_T = 1$ if $k^* = 1$.

As a direct consequence of the above lemma and recalling that $D_{\gamma_0} = \|x_1-x_*\| \vee \gamma_0$, we obtain

$$\sqrt{k_T} \leqslant \sqrt{2}\sqrt{\log_2\left(\frac{D_{\gamma_0}}{\gamma_0}\right)+1} \qquad \text{and} \qquad \gamma_{k_T} \leqslant 5D_{\gamma_0}\sqrt{\log_2\left(\frac{D_{\gamma_0}}{\gamma_0}\right)+1}, \tag{12}$$

that is, k_t takes at most $2(\log_2(D_{\gamma_0}/\gamma_0)+1)$ values.

Proof of Lemma 2. Lemma 4 in Appendix, applied by phases, implies that for all $t \geqslant 1$ and $k \geqslant 1$ we have

$$\left\|x_t^+(k)-x^*\right\|^2 \leqslant \|x_1-x_*\|^2 + \Gamma_t^2 + \gamma_k^2\frac{\|g_t\|^2}{h_t^2}.$$

Thus, the triangle inequality, yields

$$\left\|x_t^+(k)-x_1\right\| \leqslant 2\|x_1-x_*\| + \sqrt{\Gamma_t^2+\gamma_k^2\frac{\|g_t\|^2}{h_t^2}} \leqslant 2D_{\gamma_0} + \sqrt{\Gamma_t^2+\gamma_k^2\frac{\|g_t\|^2}{h_t^2}},$$

where $D_{\gamma_0} = \|x_1 - x_*\| \vee \gamma_0$. Let $\bar{k}$ be the smallest integer such that $2^{\bar{k}}/\sqrt{\bar{k}} \geqslant 2^{k^*}$. Then, for any $k \geqslant 1$ and any $t \geqslant 1$

$$\left\|x_t^+(k) - x_1\right\| \leqslant \frac{2\gamma_{\bar{k}}}{\sqrt{\bar{k}}} + \sqrt{\Gamma_t^2 + \gamma_k^2 \frac{\|g_t\|^2}{h_t^2}}\,.$$

In particular, the above implies that $\left\|x_t^+(\bar{k}) - x_1\right\| \leqslant B_{t+1}(\bar{k})$ for all $t \geqslant 1$. Thus, once k_t reaches $\bar{k}$ on Line 7 of Algorithm 1, it never changes its value. That is, $k_T \leqslant \bar{k}$. Lemma 12 in Appendix shows that $\bar{k} \leqslant k^* + 0.5\log_2(k^*) + 1.25$ and $\bar{k} = 1$ if $k^* = 1$, which concludes the proof. □

4.1 Applications of Theorem 2: specific choices of $(h_t)_{t\geqslant 1}$

Theorem 2 and Lemma 2 yield the main result of this work—theorem announced in Section 2.

Corollary 1. *Under assumptions of Theorem 2. Let* $H(x) = \sqrt{(x+1)\log(\mathrm{e}(x+1))}$. *Setting* $h_t = H(S_t)$ *and* $D_{\gamma_0} = \|x_1 - x_*\| \vee \gamma_0$, *Algorithm 1 satisfies*

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant D_{\gamma_0} H(S_{T+1}) \sqrt{\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)} \left[6\log(\log(\mathrm{e}(1+S_T))) + 6.5\right].$$

While the above choice of $(h_t)_{t\geqslant 1}$ gives nearly optimal rates, it is not standard in the literature. Let us highlight the usefulness of Theorem 2 by providing some instantiations which correspond to other, more common, but less optimal, examples.

The standard AdaGrad corresponds to $h_t = \sqrt{S_t}$ [SM10]. The main inconvenience of this choice, is that the term $\sum_{t\leqslant T}\|g_t\|^2/S_t$ is not bounded uniformly by a non-decreasing function of S_T. Indeed, assume that $\|g_t\|^2 = 1/T$ for all $t = 1, \dots, T$, then $S_t = t/T \leqslant 1$ and $\sum_{t\leqslant T}\|g_t\|^2/S_t \approx \log(T)$. It is possible, however, to write $\sum_{t\leqslant T}\|g_t\|^2/S_t \leqslant 1 + \log(S_T/\|g_1\|^2)$, which involves additional dependency on the gradient at initialization. All in all, we can state the following corollary.

Corollary 2. *Under assumptions of Theorem 2. Setting* $h_t = \sqrt{S_t}$ *and* $D_{\gamma_0} = \|x_1 - x_*\| \vee \gamma_0$, *Algorithm 1 satisfies*

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant D_{\gamma_0} \sqrt{S_{T+1}\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)} \left[6\log\left(\frac{\mathrm{e}S_T}{\|g_1\|^2}\right) + 6.5\right].$$

An attractive feature of this bound is its scale-invariance—multiplying f by some constant, multiplies the bound by the same constant.

The dependency on the initial gradient can be avoided setting $h_t = \sqrt{\varepsilon + S_t}$ with arbitrary $\varepsilon > 0$, as it is usually done in practice with AdaGrad, and initially proposed in [DHS11].

Corollary 3. *Under assumptions of Theorem 2. Let* $h_t = \sqrt{\varepsilon + S_t}$, *for some* $\varepsilon > 0$. *Setting* $D_{\gamma_0} = \|x_1 - x_*\| \vee \gamma_0$, *Algorithm 1 satisfies*

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant D_{\gamma_0} \sqrt{(S_{T+1}+\varepsilon)\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)} \left[6\log\left(1 + \frac{S_T}{\varepsilon}\right) + 6.5\right].$$

Note that compared to Corollary 1, the above bound contains an additional $\sqrt{\log(1+S_T)}$ multiplicative factor, but it improves upon that of Corollary 2. Finally, we can also recover the results claimed in the end of Section 3, where f is assumed to be L-Lipschitz, see Appendix B.2 for details.

5 An extension to stochastic optimization

In this section we demonstrate that at least the warm-up analysis provided in Section 3 extends to the setup of stochastic optimization (see Algorithm 2), where the objective function takes the form

$$f(x) = \mathbb{E}[F(x, \xi)]\,,$$

and where we only have access to $g_t \in \partial F(x_t, \xi_t)$, for some i.i.d. $(\xi_t)_{t\geqslant 1}$. As in [CH23], we make the following standard assumption on the regularity of $F(\cdot, \xi)$.

Algorithm 2: Stochastic case

Input: $x_1 \in \mathbb{R}^d, \Theta \subset \mathbb{R}^d, \gamma_0 > 0, L > 0, T, \delta > 0$

Initialization: $\Gamma_1^2 = 0, k_0 = 1, S_0 = 0, h : \mathbb{R} \to \mathbb{R}, \gamma_k = \gamma_0 2^k$ for $k \geqslant 1, \ell_T(\delta) := 1 \vee \log(\log_2(2T)/\delta)$

for $t = 1, \dots, T$ **do**

- $g_t \in \partial F(x_t, \xi_t)$ // get subgradient
- $h_t = L\sqrt{T\ell_T(\delta/(1+k_{t-1})^2)}$ // update h_t
- $x_t^+(k) = \mathrm{Proj}_\Theta\left(x_t - \frac{\gamma_k}{h_t} g_t\right)$ // probing step
- $k_t = \min\left\{k \geqslant k_{t-1} \,:\, \|x_t^+(k) - x_1\| \leqslant 38\gamma_k\right\}$ // find the step size
- $x_{t+1} = x^+(k_t)$ // make the step

end

Output : Trajectory $(x_t)_{t=1}^T$

Assumption 2. *The mapping $x \mapsto F(x, \xi)$ is L-Lipschitz almost surely.*

In some applications (e.g., linear contextual bandits), L is actually known and the control of regret is necessary. For example, in linear Contextual Bandits with Knapsacks (lin-CBwK), having PGD strategy for unbounded Θ, while still controlling the regret is needed [see e.g., AD16]. Thus, our Algorithm 2, could bring new results in CBwK and related contexts.

We can state the following result concerning Algorithm 2.

Theorem 3. *Let Assumptions 1 and 2 be satisfied. Define $\ell_T(\delta) = 1 \vee \log(\log_2(2T)/\delta)$ and $D_{\gamma_0} = \|x_1 - x_*\| \vee \gamma_0$. For any $\gamma_0 > 0$, Algorithm 2 satisfies with probability at least $1 - \delta$*

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant 3500 D_{\gamma_0} L\sqrt{T}\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)\ell_T^{1/2}\left(\delta/\log_2^2(4D_{\gamma_0}/\gamma_0)\right).$$

The above bound is of order $\mathcal{O}(LD_{\gamma_0}\sqrt{T}\log(D_{\gamma_0})\log\log(TD_{\gamma_0}))$. Note that if only the optimization error is of concern, and one does not wish to control the regret, [CH23] provide a bound without $\log_2(2D_{\gamma_0}/\gamma_0)$ using bisection algorithm and several restarted runs of SGD.

6 Experiments

We have implemented our FREE ADAGRAD (with $\gamma_0 = 1$ throughout) algorithm and compared it to the ADAGRAD that requires the knowledge of $\|x_1 - x_*\|$ and to the Oracle choice of step $\|x_1 - x_*\|/(L\sqrt{T})$.

We consider three functions $f(x) = \|x\|_p$ for $p \in \{1, 2\}$ and $f(x) = n^{-1}\sum_{i\leqslant n} |\langle a_i, x\rangle|$ where $a_i \in \mathbb{R}^d$ are generated i.i.d. from standard multivariate Gaussian. The initialization point is picked the same for the three algorithms and is sampled from uniform distribution on $[-1, 1]^d$. For our experiments, we set $d = 625$ and $n = 1000$. Note that in the first case, the considered function is Lipschitz with $L = 1$ and for the second one $L \leqslant \frac{1}{n}(\|a_1\| + \ldots + \|a_n\|)$. A subgradient at $x \in \mathbb{R}^d$ in the second case is given by $n^{-1}\sum_{i\leqslant n} a_i \mathrm{sign}(\langle a_i, x\rangle)$ and since a_i's are i.i.d. Gaussian, it is expected that $\|g\| \ll \frac{1}{n}(\|a_1\| + \ldots + \|a_n\|)$—algorithms that are adaptive to the norm of the gradient should perform better in this case. For all three functions, a global minimizer is given by $x_* = (0, \ldots, 0)^\top$. All the algorithms run for $T = 10000$ iterations.

All the plots are reported on $\log - \log$ scale. The first results are reported on Figure 2. The second column displays the step-sizes used by the three algorithms. As a sanity check, we observe that the step size of ADAGRAD decreases over time and the step size of the ORACLE remains constant. One can also observe the characteristic jumps of the proposed FREE ADAGRAD method—the step size decreases withing a fixed phase and is doubled from one phase to the other. On the first row of Figure 2 we display the regret, on initial stages our algorithm behaves similarly to the ORACLE one, while surpassing the performance of the ADAGRAD on the later stages. The third row of Figure 2 displays the case of the averages. Note that in this case the ORACLE algorithm performs worse than the other two, since it takes the worst-case Lipschitz constant and does not adapt to the actual norms of the seen gradients.

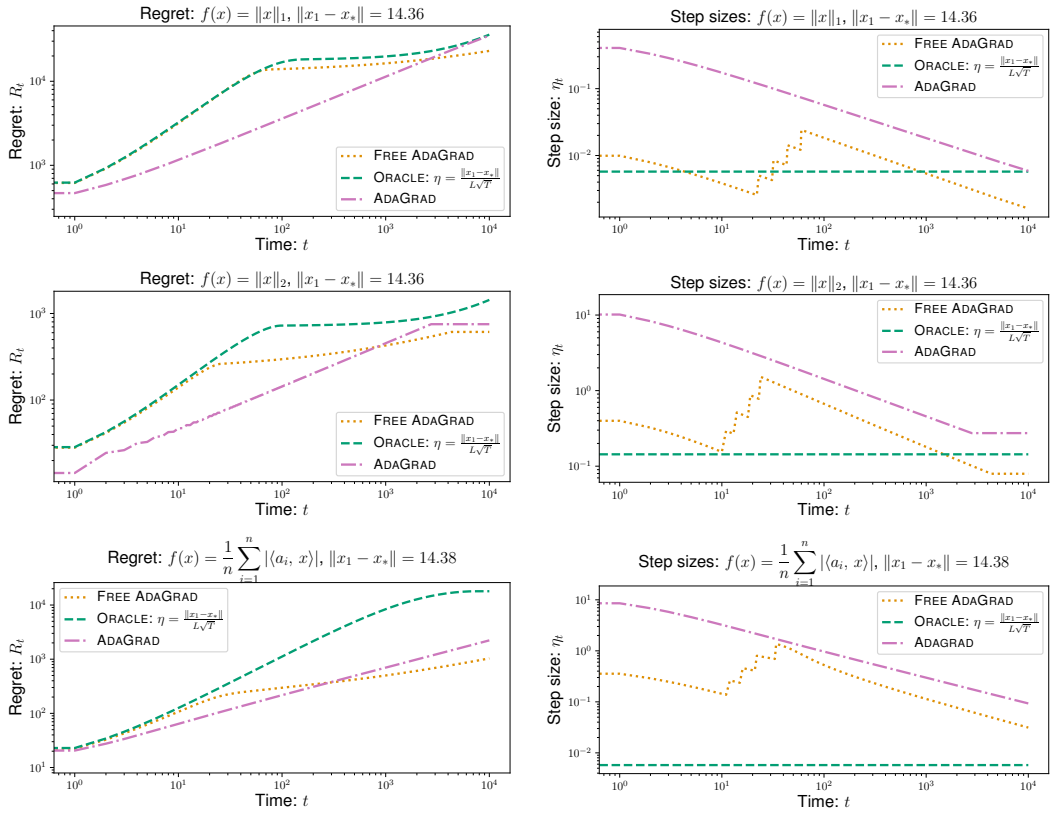

Figure 2: Regret (left) and step-sizes (right) of three algorithms on $\log - \log$ scale.

7 Discussion

We have introduced FREE ADAGRAD—a simple fully adaptive version of ADAGRAD, that does not rely on any prior information about the objective function. Our bounds are optimal up to logarithmic factors and are applicable to non-globally Lipschitz functions. We have extended our approach to stochastic optimization in a Lipschitz context, at the cost of the knowledge of the Lipschitz constant and sub-optimal logarithmic factors. Numerical illustrations suggest that FREE ADAGRAD performs on par or outperforms ADAGRAD with knowledge of $\|x_1 - x_*\|$ and the ORACLE choice of step-size.
Limitations. Let us also list the main limitations and future directions of our work.

1) We are only dealing with batch optimization. The extension of our analysis to the case of Online Convex Optimization (OCO) seems non-trivial, since the bounds that we obtain are known to be unachievable without prior knowledge in the OCO context [CB17]. The investigation of FREE ADAGRAD in the OCO setting is left for future work;

2) If f is assumed to be L-Lipschitz with known constant L, slightly better bounds—with improved log-factors—can be obtained in OCO setting [see e.g., OP16, CO18]. It remains an open question wether such bounds, can be obtained in batch optimization in the non-Lipschitz (or unknown L) and unknown $\|x_1 - x_*\|$ case;

3) Concerning stochastic optimization, we require f to be L-Lipschitz for some known L. We note, however, that even in the state-of-the-art bound of [CH23], the knowledge of L is required.

4) When translating our regret bound on a rate for the optimization error, using $\bar{x}_T$—average along the trajectory—we have additional log-factors compared to [CH23, DM23], which is an artifact of online-to-batch conversion [MS12, Theorem 7]. Contrary to us, the algorithms in [CH23, DM23] require yet some knowledge about the Lipschitz constant L.

References

[AD16] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

[CB17] Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on learning theory*, pages 643–677. PMLR, 2017.

[CBLS05] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label-efficient prediction. *IEEE Transactions on Information Theory*, 51:2152–2162, 2005.

[CH23] Yair Carmon and Oliver Hinder. Making sgd parameter-free, 2023.

[CO18] Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529. PMLR, 2018.

[Cut19] Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 874–894. PMLR, 25–28 Jun 2019.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[DM23] Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by D-adaptation. *arXiv preprint arXiv:2301.07733*, 2023.

[JC22] Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. In *Conference on Learning Theory*, pages 4160–4211. PMLR, 2022.

[MK20] Zakaria Mhammedi and Wouter M. Koolen. Lipschitz and comparator-norm adaptivity in online learning. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2858–2887. PMLR, 09–12 Jul 2020.

[MO14a] H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pages 1020–1039. PMLR, 2014.

[MO14b] H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1020–1039, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[MS12] Brendan Mcmahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. *Advances in neural information processing systems*, 25, 2012.

[N+18] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[OP16] Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[OP21] Francesco Orabona and Dávid Pál. Parameter-free stochastic optimization of variationally coherent functions. *arXiv preprint arXiv:2102.00236*, 2021.

[OT17] Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. *Advances in Neural Information Processing Systems*, 30, 2017.

[SM10] Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

[ZCP22] Zhiyu Zhang, Ashok Cutkosky, and Ioannis Paschalidis. Pde-based optimal strategy for unconstrained online learning. In *International Conference on Machine Learning*, pages 26085–26115. PMLR, 2022.

Supplementary material for "Parameter-free projected gradient descent"

Appendix A provides details for the proof of Section 3 of the main body. Appendix B contains all the proof for Theorem 2 and Lemma 1. Appendix B.2 provides proof for corollaries in Section 4. Appendix C deals with stochastic version of our algorithm and contains the proof of Theorem 3. Appendix D gives detailed connection with the reward doubling algorithm of [MS12]. Finally, Appendix E contains auxiliary results that are used in different parts of the proofs.

Below we provide a basic python implementation of our FREE ADAGRAD.

```
import numpy as np

class ObjectiveFunction():
    """
        Class for objective function has get_subgradient method
    """

    def get_subgradient(self, x):
        # To implement
        pass


def step(x, eta, g):
    return x - eta * g


def free_adagrad(stopping_criteria, obj_func, x1, gamma0=1.):
    """
        x1: initialization
        gamma0: initial guess for |x_1 - x_*|
        stopping_criteria: stopping criteria (e.g., max_iter)
        obj_func: objective function with get_subgrad() method
    """
    S = 0.
    Gamma = 0.
    k = 1
    gamma = gamma0

    x = np.copy(x1)
    trajectory = [x1]

    while not stopping_criteria:
        g = obj_func.get_subgrad(x)
        norm_g = np.linalg.norm(g)
        S += norm_g ** 2
        h = np.sqrt((S + 1.) * (1. + np.log(1. + S)))
        while True:
            x_plus = step(x, gamma / h, g)
            B = (2. / np.sqrt(k)) * gamma \
                + np.sqrt(Gamma + (gamma * norm_g / h) ** 2)
            if np.linalg.norm(x_plus - x1) > B:
                k += 1
                gamma *= 2
            else:
                Gamma += (gamma * (norm_g / h)) ** 2
                break
        x = x_plus
        trajectory.append(x.copy())

return trajectory
```

A Proofs of the results of the warm-up Section 3

We prove here with full details the results of the warm-up Section 3, in the setting where the norm of the subgradients are bounded by a known constant L, and where the time horizon T is known. We set $h_t = L\sqrt{T}$, and we analyze simultaneously FREE ADAGRAD algorithm 1 with this choice of h, and the simple variant, where we set $B_{t+1}^{\text{simple}}(k) = 3\gamma_k$ for the threshold, as in Section 3.

Theorem 4. *Assume that f is a convex L-Lipschitz function, such that there exists $x_* \in \arg\min_{x\in\Theta} f(x)$ bounded. Let $\gamma_0 > 0$ and $D_{\gamma_0} = \|x_1 - x_*\| \vee \gamma_0$. The* FREE ADAGRAD *algorithm 1 with $h_t = L\sqrt{T}$ and $B_{t+1}(k) = \gamma_k \left(\frac{2}{\sqrt{k}} + \frac{1}{\sqrt{T}}\right) + \Gamma_t$ fulfills*

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant 10 D_{\gamma_0} L\sqrt{T}\sqrt{2\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)}\,.$$

The simple variant with $h_t = L\sqrt{T}$ and $B_{t+1}^{\text{simple}}(k) = 3\gamma_k$ fulfills

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant 3\|x_1 - x_*\| L\sqrt{T}\ \log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right) + 2D_{\gamma_0} L\sqrt{T}\ .$$

Proof of Theorem 4.

We start by emphasizing that the algorithm runs without diverging, in the sense that

$$k_t := \min\left\{k \geqslant k_{t-1} \text{ such that } \|x_t^+(k) - x_1\| \leqslant B_{t+1}(k)\right\} \tag{13}$$

is finite for any t. Indeed, we observe that $\|x_t^+(k) - x_1\|$ grows at most like $\gamma_k/\sqrt{T}$ when k goes to infinity, while $B_{t+1}(k)$ grows faster than $\gamma_k\left(2/\sqrt{k} + 1/\sqrt{T}\right)$ and $B_{t+1}^{\text{simple}}(k)$ grows like $3\gamma_k$. In fact, we will prove below that k_t remains upper-bounded by a quantity independent of T.

The starting point of the proof is the classical analysis for a projected gradient step $\text{Proj}_\Theta\left(x_t - \eta g_t\right)$

$$\begin{aligned} f(x_t) - f(x_*) \leqslant \langle g_t, x_t - x_*\rangle &= \frac{\eta}{2}\|g_t\|^2 + \frac{1}{2\eta}\left(\|x_t - x_*\|^2 - \|x_t - \eta g_t - x_*\|^2\right) \\ &\leqslant \frac{\eta}{2}L^2 + \frac{1}{2\eta}\left(\|x_t - x_*\|^2 - \|\,\text{Proj}_\Theta\left(x_t - \eta g_t\right) - x_*\|^2\right), \end{aligned}$$

where the last inequality follows from the fact that $x_* \in \Theta$. Since x_* is a minimizer of f in Θ, the left hand side is non-negative, so the above inequality with $\eta = \gamma_k/(L\sqrt{T})$ gives that for any $k \geqslant 1$

$$0 \leqslant f(x_t) - f(x_*) \leqslant \langle g_t, x_t - x_*\rangle \leqslant \frac{\gamma_k L}{2\sqrt{T}} + \frac{L\sqrt{T}}{2\gamma_k}\left(\|x_* - x_t\|^2 - \|x_t^+(k) - x_*\|^2\right). \tag{14}$$

It follows from this bound, a one-step deviation upper-bound

$$\|x_t^+(k) - x_*\|^2 \leqslant \|x_t - x_*\|^2 + \frac{\gamma_k^2}{T}\ .$$

Summing this bound over t, we get a bound on the distance to optimum

$$\|x_t^+(k) - x_*\|^2 \leqslant \|x_1 - x_*\|^2 + \sum_{s=1}^{t-1}\frac{\gamma_{k_s}^2}{T} + \frac{\gamma_k^2}{T} = \|x_1 - x_*\|^2 + \Gamma_t^2 + \frac{\gamma_k^2}{T}\,, \tag{15}$$

and then a bound on the distance to initialization

$$\begin{aligned} \|x_t^+(k) - x_1\| &\leqslant \|x_1 - x_*\| + \|x_t^+(k) - x_*\| \\ &\leqslant \|x_1 - x_*\| + \sqrt{\|x_1 - x_*\|^2 + \Gamma_t^2 + \frac{\gamma_k^2}{T}} \\ &\leqslant 2\|x_1 - x_*\| + \sqrt{\Gamma_t^2 + \frac{\gamma_k^2}{T}}\,, \end{aligned} \tag{16}$$

where the two inequalities follow from (15) and the sub-additivity of square-root.

Controlling the number k_T of phases. The Inequality (16) is the key to get an upper-bound on $k_T = \max\{k_t : 1 \leqslant t \leqslant T\}$. Let us define the integer $k^* \geqslant 1$ by $\gamma_{k^*-1} \leqslant D_{\gamma_0} := \|x_* - x_1\| \vee \gamma_0 < \gamma_{k^*}$, which fulfills

$$k^* \leqslant 1 + \log_2\left(\frac{\|x_* - x_1\|}{\gamma_0} \vee 1\right) = \log\left(\frac{2D_{\gamma_0}}{\gamma_0}\right), \quad \text{and} \quad \gamma_{k^*} \leqslant 2D_{\gamma_0}\,. \tag{17}$$

To upper bound k_T, we rely on the following estimate derived from (16)

$$\|x_t^+(k) - x_1\| \leqslant 2\gamma_{k^*} + \sqrt{\Gamma_t^2 + \frac{\gamma_k^2}{T}}\,. \tag{18}$$

- **Simple case:** $B_{t+1}^{\text{simple}}(k) = 3\gamma_k$ **.** A basic induction ensures that $k_t \leqslant k^*$ for all $t \leqslant T$. Indeed, if the property $k_{t-1} \leqslant k^*$ holds, then, since $\Gamma_t^2 \leqslant \frac{t-1}{T}\gamma_{k_{t-1}}^2$, we have

  $$\|x_t^+(k^*) - x_1\| \leqslant 2\gamma_{k^*} + \gamma_{k^*} = B_{t+1}^{\text{simple}}(k^*)\,,$$

  which, in turn, ensures that $k_t \leqslant k^*$. So k^* is an upper-bound on $k_T = \max\{k_t : 1 \leqslant t \leqslant T\}$ in this case.

- **FREE ADAGRAD case:** $B_{t+1}(k) = \frac{2\gamma_{\bar{k}}}{\bar{k}^{1/2}} + \sqrt{\Gamma_t^2 + \frac{\gamma_{\bar{k}}^2}{T}}$ **.** Let us define $\bar{k}$ as the smallest integer fulfilling $\bar{k}^{-1/2}\gamma_{\bar{k}} \geqslant \gamma_{k^*}$. Then, from (18), we have

  $$\|x_t^+(\bar{k}) - x_1\| \leqslant \frac{2\gamma_{\bar{k}}}{\bar{k}^{1/2}} + \sqrt{\Gamma_t^2 + \frac{\gamma_{\bar{k}}^2}{T}} = B_{t+1}(\bar{k})\,,$$

  so, by induction, we get $k_t \leqslant \bar{k}$ for all $t \leqslant T$. In addition, we prove in Lemma 12 page 27 that

  $$\bar{k} \leqslant k^* + 0.5\log_2(k^*) + 1.25\,. \tag{19}$$

Bounding the regret on phase $k_t = k$. We denote by $[T_k, T_{k+1} - 1]$ the interval where $k_t = k$, with the convention $T_{k+1} = T_k$ if we never have $k_t = k$. For $t \in [T_k, T_{k+1} - 1]$, we have $x_{t+1} = x_t^+(k)$. So from (14), we get

$$\begin{aligned}\sum_{t=T_k}^{T_{k+1}-1}(f(x_t) - f(x_*)) &\leqslant \sum_{t=T_k}^{T_{k+1}-1}\left(\frac{\gamma_k L}{2\sqrt{T}} + \frac{L\sqrt{T}}{2\gamma_k}\left(\|x_* - x_t\|^2 - \|x_{t+1} - x_*\|^2\right)\right)\\ &= \frac{\gamma_k L}{2\sqrt{T}}(T_{k+1} - T_k) + \frac{L\sqrt{T}}{2\gamma_k}\left(\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2\right)\,,\end{aligned} \tag{20}$$

with the convention that $\sum_{t=T_k}^{T_k-1} = 0$. We bound the second term in the right-hand side of (20), with (15)

$$\|x_{T_k} - x_*\|^2 \leqslant \|x_1 - x_*\|^2 + \Gamma_{T_k-1}^2 + \frac{\gamma_k}{T} = \|x_1 - x_*\|^2 + \Gamma_{T_k}^2\,,$$

and for the third term, we combine (13) with a triangular inequality to get

$$\|x_{T_{k+1}} - x_*\|^2 \geqslant \left[\|x_1 - x_*\| - \|x_{T_{k+1}} - x_1\|\right]_+^2 \geqslant \left[\|x_1 - x_*\| - B_{T_{k+1}}\right]_+^2\,,$$

where we used the condensed notation $B_{T_{k+1}} := B_{T_{k+1}}(k_{T_{k+1}-1}) = B_{T_{k+1}}(k)$. Plugging these two upper and lower bounds in (20), and applying the simple inequality

$$\Delta^2 - [\Delta - B]_+^2 \leqslant 2\Delta B, \quad \text{for all } \Delta, B \geqslant 0, \tag{21}$$

we get

$$\begin{aligned}\sum_{t=T_k}^{T_{k+1}-1}(f(x_t) - f(x_*)) &\leqslant \frac{\gamma_k L}{2\sqrt{T}}(T_{k+1} - T_k) + \frac{L\sqrt{T}}{2\gamma_k}\left(\|x_1 - x_*\|^2 + \Gamma_{T_k}^2 - \left[\|x_1 - x_*\| - B_{T_{k+1}}\right]_+^2\right)\\ &\leqslant \frac{L\sqrt{T}}{2}\left(\gamma_k\frac{T_{k+1} - T_k}{T} + \frac{\Gamma_{T_k}^2}{\gamma_k} + \frac{2B_{T_{k+1}}}{\gamma_k}\|x_1 - x_*\|\right)\,,\end{aligned}$$

using $\Gamma_{T_k}^2 \leqslant \gamma_{k-1}^2 T_k/T/ \leqslant \gamma_{k-1}\gamma_k/2$ for $k \geqslant 1$, we get

$$\sum_{t=T_k}^{T_{k+1}-1}(f(x_t)-f(x_*)) \leqslant \frac{L\sqrt{T}}{2}\left(\gamma_k\frac{T_{k+1}-T_k}{T}+\frac{\gamma_{k-1}}{2}+\frac{2B_{T_{k+1}}}{\gamma_k}\|x_1-x_*\|\right).$$

Bounding the total regret. Summing the above inequality over k, we get the upper-bound on the total regret

$$\begin{aligned}\sum_{t=1}^{T}(f(x_t)-f(x_*)) &\leqslant \frac{L\sqrt{T}}{2}\sum_{1\leqslant k\leqslant k_T}\left(\gamma_{k_T}\frac{T_{k+1}-T_k}{T}+\frac{\gamma_{k-1}}{2}+\frac{2B_{T_{k+1}}}{\gamma_k}\|x_1-x_*\|\right)\\ &\leqslant \frac{L\sqrt{T}}{2}\left(\frac{3\gamma_{k_T}}{2}+\sum_{1\leqslant k\leqslant k_T}\frac{2B_{T_{k+1}}}{\gamma_k}\|x_1-x_*\|\right). \qquad (22)\end{aligned}$$

We point out that the bound (22) is valid for any choice of $B_{t+1}(k)$. Let us treat apart the two cases.

- **Simple case:** $B_{t+1}^{\text{simple}}(k) = 3\gamma_k$ **.** Using that $k_T \leqslant k^*$ in this case,

  $$B_{T_{k+1}} = B_{T_{k+1}}(k) = 3\gamma_k\ ,$$

  and recalling the upper bound (17) on k^* and γ_{k^*}, we get from (22)

  $$\begin{aligned}\sum_{t=1}^{T}(f(x_t)-f(x_*)) &\leqslant \frac{L\sqrt{T}}{2}\left(\frac{3}{2}\gamma_{k^*}+\sum_{1\leqslant k\leqslant k^*}6\|x_1-x_*\|\right)\\ &= \frac{L\sqrt{T}}{2}\left(\frac{3}{2}\gamma_{k^*}+6k^*\|x_1-x_*\|\right)\\ &\leqslant L\sqrt{T}\left(3\|x_1-x_*\|\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)+2D_{\gamma_0}\right).\end{aligned}$$

- **FREE ADAGRAD case:** $B_{t+1}(k) = \frac{2\gamma_k}{k^{1/2}} + \sqrt{\Gamma_t^2 + \frac{\gamma_k^2}{T}}$ **.** We have proved in this case that $k_T \leqslant \bar{k}$ with $\bar{k}$ upper bounded by (19). Combining (19) and (17), the first term in the right-hand side of (22) can be readily bounded by

  $$\frac{3}{4}\gamma_{\bar{k}} \leqslant \frac{3}{4}\gamma_{k^*+0.5\log_2(k^*)+1.25} \leqslant 4D_{\gamma_0}\sqrt{\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)}.$$

  The last term in the right-hand side of (22), can be bounded as follows. We notice that

  $$\Gamma_{T_{k+1}-1}^2 + \frac{\gamma_{k_{T_{k+1}-1}}^2}{T} = \Gamma_{T_{k+1}}^2\ ,$$

  so we have

  $$\sum_{1\leqslant k\leqslant\bar{k}}\frac{B_{T_{k+1}}}{\gamma_k} = \sum_{1\leqslant k\leqslant\bar{k}}\left(\frac{2}{\sqrt{k}}+\frac{\Gamma_{T_{k+1}}}{\gamma_k}\right) \leqslant 4\sqrt{\bar{k}} + \sum_{1\leqslant k\leqslant\bar{k}}\frac{\Gamma_{T_{k+1}}}{\gamma_k}\,.$$

  For the last term, we observe that

  $$\begin{aligned}\sum_{1\leqslant k\leqslant\bar{k}}\frac{\Gamma_{T_{k+1}}}{\gamma_k} &= \sum_{1\leqslant k\leqslant\bar{k}}\gamma_k^{-1}\sqrt{\sum_{j\leqslant k}\gamma_j^2\Delta T_j/T} \leqslant \sum_{1\leqslant k\leqslant\bar{k}}\sum_{j\leqslant k}\gamma_k^{-1}\gamma_j\sqrt{\Delta T_j/T}\\ &\leqslant \sum_{1\leqslant j\leqslant\bar{k}}\gamma_j\sqrt{\Delta T_j/T}\sum_{k:k\geqslant j}\gamma_k^{-1} = 2\sum_{1\leqslant j\leqslant\bar{k}}\sqrt{\Delta T_j/T} \leqslant 2\sqrt{\bar{k}}\,,\end{aligned}$$

where the last inequality follows from Cauchy Schwarz. Then, plugging these bounds in (22), and using that $\bar{k} = 1$ when $k^* = 1$, and

$$\bar{k} \leqslant k^* + 0.5 \log_2(k^*) + 1.25 \leqslant 2k^*, \quad \text{for } k^* \geqslant 2\,,$$

we get from (17)

$$\begin{aligned}\sum_{t=1}^{T}(f(x_t) - f(x_*)) &\leqslant L\sqrt{T}\left[\frac{3}{4}\gamma_{\bar{k}} + 6\|x_1 - x_*\|\sqrt{\bar{k}}\right] \\ &\leqslant 10 D_{\gamma_0} L\sqrt{T}\sqrt{2\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)}\,.\end{aligned}$$

which concludes the proof of Theorem 4.

B Proofs for Section 4

Proof of Theorem 2. First of all, observe that the algorithm in question can be written as

$$x_{t+1} = \mathrm{Proj}_\Theta\left(x_t - \frac{\gamma_{k_t}}{h_t} g_t\right),$$

where we recall that $(h_t)_{t\geqslant 1}$ is assumed to be non-decreasing and positive. As before, we denote by $[T_k, T_{k+1}-1]$ the interval where $k_t = k$. In particular, $T_{k_T+1}-1 = T$. On the interval $[T_k, T_{k+1}-1]$, the algorithm is simply AdaGrad (slightly modified) started from the point x_{T_k} and with the final point at $x_{T_{k+1}}$. Thus, within each phase, we can apply the analysis of the AdaGrad that we recall and slightly adapt in Appendix B.1, page 18. The proof closely follows that of the warm-up setup: observing that

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) = \sum_{k=1}^{k_T}\sum_{t=T_k}^{T_{k+1}-1}(f(x_t) - f(x_*)),$$

1. We start with one phase analysis, using the results of Appendix B.1, page 18, which contains Lemma 1 displayed in the main body;
2. Then, we sum-up the total regret over k_T phases, using the previous analysis, and bound the key quantities;

One phase analysis. Fix some $k \leqslant k_T$ and assume that the the kth phase is non-empty, that is, $T_{k+1} > T_k$. Thus, in view of the above discussion, Lemma 5, page 18, yields

$$\begin{aligned}\sum_{t=T_k}^{T_{k+1}-1}(f(x_t) - f(x_*)) &\leqslant h_{T_{k+1}}\left(\frac{\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2}{2\gamma_k} + \frac{\gamma_k}{2}\underbrace{\sum_{t=T_k}^{T_{k+1}-1}\frac{\|g_t\|^2}{h_t^2}}_{=\frac{\Gamma^2_{T_{k+1}} - \Gamma^2_{T_k}}{\gamma_k^2}}\right) \\ &= h_{T_{k+1}}\left(\frac{\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2}{2\gamma_k} + \frac{\Gamma^2_{T_{k+1}} - \Gamma^2_{T_k}}{2\gamma_k}\right).\end{aligned} \tag{23}$$

Note that by design $\|x_{T_{k+1}} - x_*\| \geqslant \left[\,\|x_1 - x_*\| - B_{T_{k+1}}(k)\right]_+$. Furthermore, iteratively applying Lemma 4 by phases, we deduce that

$$\|x_{T_k} - x_*\|^2 \leqslant \|x_1 - x^*\|^2 + \Gamma^2_{T_k}.$$

That is, we have

$$\frac{\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2}{2\gamma_k} \leqslant \frac{\|x_1 - x_*\|^2 - \left[\,\|x_1 - x_*\| - B_{T_{k+1}}(k)\right]_+^2}{2\gamma_k} + \frac{\Gamma^2_{T_k}}{2\gamma_k}.$$

Furthermore, recalling that $\Delta^2 - [\Delta - B]_+^2 \leqslant 2\Delta B$, the above can be further bounded as

$$\frac{\|x_{T_k} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2}{2\gamma_k} \leqslant \|x_1 - x_*\|\frac{B_{T_{k+1}}(k)}{\gamma_k} + \frac{\Gamma^2_{T_k}}{2\gamma_k}. \tag{24}$$

Substitution of (24) into (23), yields

$$\sum_{t=T_k}^{T_{k+1}-1}(f(x_t) - f(x_*)) \leqslant h_{T_{k+1}}\|x_1 - x_*\|\frac{B_{T_{k+1}}(k)}{\gamma_k} + h_{T_{k+1}}\frac{\Gamma^2_{T_{k+1}}}{2\gamma_k}. \tag{25}$$

Summing up over phases. Summing up all the inequalities (25) for k_T phases, we obtain

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant \|x_1 - x_*\|\underbrace{\sum_{k\leqslant k_T}\left(h_{T_{k+1}}\frac{B_{T_{k+1}}(k)}{\gamma_k}\right)}_{=:\mathrm{I}} + \underbrace{\sum_{k\leqslant k_T} h_{T_{k+1}}\frac{\Gamma^2_{T_{k+1}}}{2\gamma_k}}_{=:\mathrm{II}}. \tag{26}$$

Bounding the sum of $h\frac{\Gamma}{2\gamma}$ terms (I). Observe that, by definition of thereof,

$$\Gamma_{T_{k+1}}^2 = \sum_{j \leqslant k} \gamma_j^2 \left(\sum_{t=T_j}^{T_{j+1}-1} \frac{\|g_t\|^2}{h_t^2} \right) . \tag{27}$$

Hence, using trivial bound $h_{T_{k+1}} \leqslant h_{T+1}$, we deduce

$$\begin{aligned}
\mathsf{I} = \sum_{k \leqslant k_T} h_{T_{k+1}} \frac{\Gamma_{T_{k+1}}^2}{2\gamma_k} &\leqslant \frac{h_{T+1}}{2} \sum_{k \leqslant k_T} \frac{\Gamma_{T_{k+1}}^2}{\gamma_k} \\
&= \frac{h_{T+1}}{2} \sum_{k \leqslant k_T} \sum_{j \leqslant k} \gamma_k^{-1} \gamma_j^2 \left(\sum_{t=T_j}^{T_{j+1}-1} \frac{\|g_t\|^2}{h_t^2} \right) \\
&= \frac{h_{T+1}}{2} \sum_{j \leqslant k_T} \sum_{k \geqslant j} \gamma_k^{-1} \gamma_j^2 \left(\sum_{t=T_j}^{T_{j+1}-1} \frac{\|g_t\|^2}{h_t^2} \right) \\
&\leqslant h_{T+1} \sum_{j \leqslant k_T} \gamma_j \left(\sum_{t=T_j}^{T_{j+1}-1} \frac{\|g_t\|^2}{h_t^2} \right) \\
&\leqslant h_{T+1} \gamma_{k_T} \sum_{t=1}^{T} \frac{\|g_t\|^2}{h_t^2} ,
\end{aligned} \tag{28}$$

where the penultimate inequality is due to the fact that $\sum_{k \geqslant j} 2^{-k} \leqslant 2^{-j+1}$ and the last one holds since $\gamma_k \leqslant \gamma_{k_T}$.

Bounding the sum of B terms (II). We observe that by definition of $B_t(k)$, we have

$$B_{T_{k+1}}(k) = B_{T_{k+1}-1+1}(k) = \frac{2\gamma_k}{\sqrt{k}} + \Gamma_{T_{k+1}}^2 .$$

Hence, the term of interest is bounded as

$$\begin{aligned}
\mathsf{II} = \sum_{k \leqslant k_T} h_{T_{k+1}} \frac{B_{T_{k+1}}(k)}{\gamma_k} &= 2 \sum_{k \leqslant k_T} \frac{h_{T_{k+1}}}{\sqrt{k}} + \sum_{k \leqslant k_T} h_{T_{k+1}} \gamma_k^{-1} \Gamma_{T_{k+1}} \\
&\leqslant 4 h_{T+1} \sqrt{k_T} + h_{T+1} \sum_{k \leqslant k_T} \gamma_k^{-1} \Gamma_{T_{k+1}} .
\end{aligned}$$

For the third term, similarly to the previous paragraph, but additionally invoking Jensen's inequality, we can write

$$\begin{aligned}
\sum_{k \leqslant k_T} \gamma_k^{-1} \Gamma_{T_{k+1}} &= k_T \sum_{k \leqslant k_T} \frac{1}{k_T} \sqrt{\sum_{j \leqslant k} \gamma_j^2 \gamma_k^{-2} \left(\sum_{t=T_j}^{T_{j+1}-1} \frac{\|g_t\|^2}{h_t^2} \right)} \\
&\leqslant \sqrt{k_T} \sqrt{\sum_{j \leqslant k_T} \sum_{k \geqslant j} \gamma_j^2 \gamma_k^{-2} \left(\sum_{t=T_j}^{T_{j+1}-1} \frac{\|g_t\|^2}{h_t^2} \right)} \\
&\leqslant \frac{2\sqrt{k_T}}{\sqrt{3}} \sqrt{\sum_{t=1}^{T} \frac{\|g_t\|^2}{h_t^2}} ,
\end{aligned}$$

where in the last inequality we used the fact that $\sum_{k=a}^{b} 2^{-2k} \leqslant \frac{4}{3} 2^{-2a}$. Thus, overall, we have

$$\mathsf{II} = \sum_{k \leqslant k_T} h_{T_{k+1}} \frac{B_{T_{k+1}}(k)}{\gamma_k} \leqslant 2 h_{T+1} \sqrt{k_T} \left(2 + \sqrt{\frac{1}{3} \sum_{t=1}^{T} \frac{\|g_t\|^2}{h_t^2}} \right) \tag{29}$$

The end (combining bounds for I and II). Substituting (28) and (29) into (26), we deduce that for any non-decreasing $(h_t)_{t\geqslant 1}$

$$\sum_{t=1}^{T}(f(x_t)-f(x_*)) \leqslant h_{T+1}\left(2\left\|x_1-x_*\right\|\sqrt{k_T}\left(2+\sqrt{\frac{1}{3}\sum_{t=1}^{T}\frac{\|g_t\|^2}{h_t^2}}\right)+\gamma_{k_T}\sum_{t=1}^{T}\frac{\|g_t\|^2}{h_t^2}\right) . \quad \square$$

B.1 Basic analysis for AdaGrad and Proof of Lemma 1

In this section we extend the standard analysis of AdaGrad for our purposes and prove Lemma 1 restated below. Throughout, we consider the following algorithm for $t \geqslant 1$

$$x_{t+1} = \mathrm{Proj}_{\Theta}\left(x_t - \frac{\gamma}{h_t}g_t\right) , \tag{30}$$

where $g_t \in \partial f(x_t)$, $S_t = \sum_{s=1}^{t}\left\|g_s\right\|^2$ and $(h_t)_{t\geqslant 1}$ is non-decreasing and positive, and $x_1 \in \mathbb{R}^d$.

We start with some elementary results.

Lemma 3. *For all $t \geqslant 1$ and all $x_1 \in \mathbb{R}$*

$$0 \leqslant \frac{2\gamma}{h_t}(f(x_t)-f(x_*)) \leqslant \left\|x_t-x_*\right\|^2 - \left\|x_{t+1}-x_*\right\|^2 + \frac{\gamma^2}{h_t^2}\left\|g_t\right\|^2 .$$

Proof. By the property of projection

$$\begin{aligned}\left\|x_{t+1}-x_*\right\|^2 &\leqslant \left\|x_t-x_*\right\|^2 - \frac{2\gamma}{h_t}\left\langle x_t-x_*,\, g_t\right\rangle + \frac{\gamma^2}{h_t^2}\left\|g_t\right\|^2 \\ &\leqslant \left\|x_t-x_*\right\|^2 - \frac{2\gamma}{h_t}(f(x_t)-f(x_*)) + \frac{\gamma^2}{h_t^2}\left\|g_t\right\|^2 ,\end{aligned} \tag{31}$$

where we used the fact that f is convex. The result follows after re-arranging. $\square$

Lemma 3 applied iteratively yields the following result.

Lemma 4. *For all $T \geqslant 1, \bar{\gamma} > 0$ and all $x_1 \in \mathbb{R}$, Algorithm (30) satisfies*

$$\left\|\mathrm{Proj}_{\Theta}\left(x_T - \frac{\bar{\gamma}}{h_T}g_T\right) - x_*\right\|^2 \leqslant \left\|x_1-x_*\right\|^2 + \gamma^2\sum_{t=1}^{T-1}\frac{\|g_t\|^2}{h_t^2} + \bar{\gamma}^2\frac{\|g_T\|^2}{h_T^2} .$$

Finally, we are in position to prove Lemma 1 brought up in the main body of the paper.

Lemma 5 (Restated Lemma 1 from Section 4)**.** *For all $T > 1$ and all $x_1 \in \mathbb{R}^d$ we have*

$$\sum_{t=1}^{T-1}(f(x_t)-f(x_*)) \leqslant h_T\left(\frac{\left\|x_1-x_*\right\|^2 - \left\|x_T-x_*\right\|^2}{2\gamma} + \frac{\gamma}{2}\sum_{t=1}^{T-1}\frac{\|g_t\|^2}{h_t^2}\right) .$$

Proof. Using Lemma 3, we deduce that

$$\sum_{t=1}^{T-1}(f(x_t)-f(x_*)) \leqslant \frac{1}{2\gamma}\sum_{t=1}^{T-1}h_t\left(\left\|x_t-x_*\right\|^2 - \left\|x_{t+1}-x_*\right\|^2\right) + \frac{\gamma}{2}\sum_{t=1}^{T-1}\frac{\|g_t\|^2}{h_t} . \tag{32}$$

Let us bound the first sum on the right hand side, adding and subtracting $h_{t+1}\left\|x_{t+1}-x_*\right\|^2$ and using telescoping summation, we obtain

$$\begin{aligned}\sum_{t=1}^{T-1}h_t\left(\left\|x_t-x_*\right\|^2 - \left\|x_{t+1}-x_*\right\|^2\right) &= h_1\left\|x_1-x_*\right\|^2 - h_T\left\|x_T-x_*\right\|^2 \\ &\quad + \sum_{t=1}^{T-1}\left(h_{t+1}-h_t\right)\left\|x_{t+1}-x_*\right\|^2 .\end{aligned} \tag{33}$$

Furthermore, by Lemma 4 with $\bar{\gamma} = \gamma$ and the fact that $(h_t)_{t\geqslant 1}$ is non-decreasing, we get

$$\sum_{t=1}^{T-1} (h_{t+1} - h_t) \|x_{t+1} - x_*\|^2 \leqslant \sum_{t=1}^{T-1} (h_{t+1} - h_t) \left(\|x_1 - x_*\|^2 + \gamma^2 \sum_{s=1}^{t} \frac{\|g_s\|^2}{h_s^2} \right)$$
$$\leqslant (h_T - h_1) \|x_1 - x_*\|^2 + \gamma^2 \sum_{t=1}^{T-1} (h_{t+1} - h_t) \sum_{s=1}^{t} \frac{\|g_s\|^2}{h_s^2} .$$

For the second term in the above bound, we can write

$$\sum_{t=1}^{T-1} (h_{t+1} - h_t) \sum_{s=1}^{t} \frac{\|g_s\|^2}{h_s^2} = \sum_{s=1}^{T-1} \frac{\|g_s\|^2}{h_s^2} \sum_{t=s}^{T-1} (h_{t+1} - h_t)$$
$$= \sum_{s=1}^{T-1} \frac{\|g_s\|^2}{h_s^2} (h_T - h_s)$$
$$= h_T \sum_{t=1}^{T-1} \frac{\|g_t\|^2}{h_t^2} - \sum_{t=1}^{T-1} \frac{\|g_t\|^2}{h_t} .$$

Substitution of the above into the penultimate inequality yields

$$\sum_{t=1}^{T-1} (h_{t+1} - h_t) \|x_{t+1} - x_*\|^2 \leqslant (h_T - h_1) \|x_1 - x_*\|^2 + \gamma^2 \left(h_T \sum_{t=1}^{T-1} \frac{\|g_t\|^2}{h_t^2} - \sum_{t=1}^{T-1} \frac{\|g_t\|^2}{h_t} \right) , \quad (34)$$

Substituting (34) into (33), we deduce that

$$\sum_{t=1}^{T-1} h_t \left(\|x_t - x_*\|^2 - \|x_{t+1} - x_*\|^2 \right) \leqslant h_T \left(\|x_1 - x_*\|^2 - \|x_T - x_*\|^2 \right)$$
$$+ \gamma^2 h_T \sum_{t=1}^{T-1} \frac{\|g_t\|^2}{h_t^2} - \gamma^2 \sum_{t=1}^{T-1} \frac{\|g_t\|^2}{h_t} .$$

Combination of the above with (32) concludes the proof. □

B.2 Proofs of corollaries in Section 4

In this section we provide proofs of four corollaries presented in Section 4.

Proof of Corollary 1. Substituting our choice of h_t into Theorem 2, we prove in Lemma 9 in Appendix E, page 27, that

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{h_t^2} \leqslant \log(\log(\mathrm{e}(1 + S_T))) .$$

Substituting the above into Theorem 2 and using (12), we deduce that

$$R_T \leqslant \sqrt{(S_{T+1} + 1) \log(\mathrm{e}(S_{T+1} + 1))} \Bigg[\sqrt{8} \|x_1 - x_*\| \sqrt{\log_2 \left(\frac{D_{\gamma_0}}{\gamma_0} \right) + 1} \left(2 + \sqrt{\frac{1}{3} \log(\log(\mathrm{e}(1 + S_T)))} \right)$$
$$+ 5 D_{\gamma_0} \log(\log(\mathrm{e}(1 + S_T))) \sqrt{\log_2 \left(\frac{D_{\gamma_0}}{\gamma_0} \right) + 1} \Bigg] .$$

The proof is concluded after re-arranging and using $2ab \leqslant a^2 + b^2$. □

Proof of Corollary 2. Theorem 2 and Lemma 2 (rather Eq. (12)) and Lemma 10 give

$$\begin{aligned}\sum_{t=1}^{T}(f(x_t)-f(x_*)) &\leqslant \sqrt{S_{T+1}\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)}\left[2\left\|x_1-x_*\right\|\sqrt{2}\left(2+\sqrt{\frac{1}{3}\log\left(\mathrm{e}\left(\frac{S_T}{\|g_1\|^2}\right)\right)}\right)\right.\\ &\qquad\left.+5D_{\gamma_0}\log\left(\mathrm{e}\left(\frac{S_T}{\|g_1\|^2}\right)\right)\right]\\ &\leqslant D_{\gamma_0}\sqrt{S_{T+1}\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)}\left[2\sqrt{2}\left(2+\sqrt{\frac{1}{3}\log\left(\mathrm{e}\left(\frac{S_T}{\|g_1\|^2}\right)\right)}\right)\right.\\ &\qquad\left.+5\log\left(\mathrm{e}\left(\frac{S_T}{\|g_1\|^2}\right)\right)\right].\end{aligned}$$

The proof is concluded after re-arranging and using $2ab \leqslant a^2+b^2$. □

Proof of Corollary 3. From Lemma 11 we have

$$\sum_{t=1}^{T}\frac{\|g_t\|^2}{\varepsilon+S_t} \leqslant \log\left(1+\frac{S_T}{\varepsilon}\right).$$

Hence, substituting the above into Theorem 2 and using (12), we obtain

$$R_T \leqslant \sqrt{\varepsilon+S_{T+1}}\sqrt{\log_2\left(\frac{D_{\gamma_0}}{\gamma_0}\right)+1}\left[\sqrt{8}\left\|x_1-x_*\right\|\left(2+\sqrt{\frac{1}{3}\log\left(1+\frac{S_T}{\varepsilon}\right)}\right)+5D_{\gamma_0}\log\left(1+\frac{S_T}{\varepsilon}\right)\right].$$

The proof is concluded after re-arranging and using $2ab \leqslant a^2+b^2$. □

As promised in Section 4, Theorem 4 of Appendix A can be obtained as a corollary of Theorem 2.

Corollary 4. *Under assumptions of Theorem 2, with f an L-Lipschitz function. Setting $h_t = L\sqrt{T}$ and $D_{\gamma_0} = \|x_1-x_*\| \vee \gamma_0$, Algorithm 1 satisfies*

$$\sum_{t=1}^{T}(f(x_t)-f(x_*)) \leqslant 12.3 D_{\gamma_0} L\sqrt{T\log_2\left(\frac{2D_{\gamma_0}}{\gamma_0}\right)}.$$

Proof of Corollary 4. Substituting $h \equiv L\sqrt{T}$ into Theorem 2 gives

$$R_T \leqslant L\sqrt{T}\left(2\left\|x_1-x_*\right\|\sqrt{k_T}\left(2+\sqrt{\frac{1}{3}}\right)+\gamma_{k_T}\right)+L\sqrt{T}\frac{\gamma_{k_T}}{2},$$

Eq. (12) applied to the above, yields

$$\begin{aligned}R_T &\leqslant L\sqrt{T}\left[\sqrt{8}\left\|x_1-x_*\right\|\sqrt{\log_2\left(\frac{D_{\gamma_0}}{\gamma_0}\right)+1}\left(2+\sqrt{\frac{1}{3}}\right)+5D_{\gamma_0}\sqrt{\log_2\left(\frac{D_{\gamma_0}}{\gamma_0}\right)+1}\right]\\ &\leqslant L\sqrt{T}\underbrace{\left[\sqrt{8}\left(2+\sqrt{\frac{1}{3}}\right)+5\right]}_{\leqslant 12.3}D_{\gamma_0}\sqrt{\log_2\left(\frac{D_{\gamma_0}}{\gamma_0}\right)+1}.\end{aligned}$$

The proof is concluded. □

C Analysis for stochastic PGD in Section 5: proof of Theorem 3

Proof of Theorem 3. We recall that $\ell(\delta) = 1 \vee \log(\log_2(T)/\delta)$. As in all the previous sections, we denote by $[T_k, T_{k+1} - 1]$, the interval where $k_t = k$. The regret of the algorithm can be expressed as

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) = \sum_{k=1}^{k_T} \underbrace{(f(x_{T_k}) - f(x_*))}_{=:\mathtt{T}_1(k)} + \sum_{k=1}^{k_T} \underbrace{\sum_{t=T_k+1}^{T_{k+1}-1} (f(x_t) - f(x_*))}_{=:\mathtt{T}_2(k)} .$$

We are going to apply Lemma 7 of Appendix C.1 on page 23, to the second term and provide deterministic bound on the first one. First let us explain the reason of such separation of terms. Observe that for $t = T_k$

$$x_{t+1} = \mathrm{Proj}_\Theta \left(x_t - \frac{\gamma_{k_t}}{L\sqrt{T\ell(\delta/(1+k_{t-1})^2)}} g_t \right) .$$

Meanwhile, for $t \in [T_k + 1, T_{k+1} - 1]$ we have

$$x_{t+1} = \mathrm{Proj}_\Theta \left(x_t - \frac{\gamma_{k_t}}{L\sqrt{T\ell(\delta/(1+k_t)^2)}} g_t \right) .$$

That is, the x_{T_k+1} step is outside the pattern and requires additional splitting. The proof proceeds as follows

1. First we give deterministic bound on $\mathtt{T}_1(k)$ terms using Lipschitzness of the objective function f;
2. Then we use Lemma 7 to bound each $\mathtt{T}_2(k)$;
3. Finally, we show that k_T is still bounded by k^* as in the warmup analysis.

Analysis for $\mathtt{T}_1(k)$. Since f is assumed to be Lipschitz, we can write

$$\mathtt{T}_1(k) \leqslant L \left\| x_{T_k} - x_* \right\| .$$

Then, simply by the triangle inequality and property of the Euclidean projection, we deduce that for all $k \geqslant 2$

$$\begin{aligned} \left\| x_{T_k} - x_* \right\| &\leqslant \left\| x_{T_{k-1}+1} - x_{T_{k-1}} \right\| + \left\| x_{T_{k-1}} - x_* \right\| + (T_k - T_{k-1} - 1)\frac{\gamma_{k-1}}{\sqrt{T\ell(\delta/k^2)}} \\ &\leqslant \left\| x_{T_{k-1}+1} - x_{T_{k-1}} \right\| + \left\| x_{T_{k-1}} - x_* \right\| + (T_k - T_{k-1} - 1)\frac{\gamma_{k_T-1}}{\sqrt{T\ell(\delta)}} . \end{aligned}$$

Furthermore, we have

$$\left\| x_{T_{k-1}+1} - x_{T_{k-1}} \right\| \leqslant \frac{\gamma_{k-1}}{\sqrt{T\ell(\delta/(k-1)^2)}} \leqslant \frac{\gamma_{k_T-1}}{\sqrt{T\ell(\delta)}} .$$

Hence, we have

$$\left\| x_{T_k} - x_* \right\| \leqslant \left\| x_{T_{k-1}} - x_* \right\| + (T_k - T_{k-1})\frac{\gamma_{k_T-1}}{\sqrt{T\ell(\delta)}} .$$

Unfolding the above recursion, we deduce that

$$\left\| x_{T_k} - x_* \right\| \leqslant \left\| x_{T_1} - x_* \right\| + T_k \frac{\gamma_{k_T-1}}{\sqrt{T\ell(\delta)}} \leqslant \left\| x_1 - x_* \right\| + 0.5\gamma_{k_T}\sqrt{T} .$$

We conclude that

$$\begin{aligned} \sum_{k=1}^{k_T}(f(x_{T_k}) - f(x_*)) &\leqslant Lk_T \left(\left\| x_1 - x_* \right\| + 0.5\gamma_{k_T}\sqrt{T} \right) \\ &\leqslant L\sqrt{T\ell_T(\delta/(1+k_T)^2)}k_T \left(\left\| x_1 - x_* \right\| + 0.5\gamma_{k_T} \right) . \end{aligned} \tag{35}$$

Analysis for $\mathtt{T}_2(k)$. Let us first fix the high-probability event on which we are going to work. Note that $\rho = T_k + 1$ is a stopping time and $T_{k+1} - 1 - \rho \leqslant T$. Thus, by Lemma 7 with probability at least $1 - \delta/(1+k)^2$, it holds that

$$\mathtt{T}_2(k) \leqslant L\sqrt{T\ell_T(\delta/(1+k)^2)}\left(\frac{\gamma_k}{2} + \frac{\|x_{T_k+1} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2}{2\gamma_k} + 10\|x_{T_k+1} - x_*\| + 68\gamma_k\right).$$

We observe that

$$\begin{aligned}\|x_{T_k+1} - x_*\|^2 - \|x_{T_{k+1}} - x_*\|^2 &\leqslant \|x_{T_k+1} - x_*\|^2 - \left[\|x_{T_k+1} - x_*\| - \|x_{T_k+1} - x_{T_{k+1}}\|\right]_+^2 \\ &\leqslant 2\|x_{T_k+1} - x_*\|\|x_{T_k+1} - x_{T_{k+1}}\|\,.\end{aligned}$$

Let us bound each term of the product. By the design of the rule,

$$\|x_{T_k+1} - x_{T_{k+1}}\| \leqslant \|x_{T_k+1} - x_1\| + \|x_1 - x_{T_{k+1}}\| \leqslant 38\gamma_k + 38\gamma_k \leqslant 76\gamma_k\,,$$

where $B = 28$. Furthermore, by Lemma 7 and the fact that the $\|x_{T_k+1} - x_{T_k}\| \leqslant \gamma_k$ for all $k \geqslant 1$

$$\begin{aligned}\|x_{T_k+1} - x_*\| &\leqslant \|x_{T_k} - x_*\| + 2\gamma_{k-1} \\ &\leqslant \|x_{T_{k-1}+1} - x_*\| + 18\gamma_{k-1} \\ &\leqslant \|x_2 - x_*\| + 18\sum_{j=1}^{k-1}\gamma_j \\ &\leqslant \|x_2 - x_*\| + 18\gamma_k \\ &\leqslant \|x_1 - x_*\| + 18\gamma_k + \gamma_1 \\ &\leqslant \|x_1 - x_*\| + 19\gamma_k\,.\end{aligned} \tag{36}$$

Thus, we have shown that with probability at least $1 - \delta/(1+k)^2$

$$\begin{aligned}\mathtt{T}_2(k) &\leqslant L\sqrt{T\ell_T(\delta/(1+k)^2)}\left(\frac{\gamma_k}{2} + 76\left(\|x_1 - x_*\| + 19\gamma_k\right) + 10(\|x_1 - x_*\| + 19\gamma_k) + 68\gamma_k\right) \\ &\leqslant L\sqrt{T\ell_T(\delta/(1+k_T)^2)}\left(86\|x_1 - x_*\| + (259 + 38^2)\gamma_k\right) \\ &\leqslant L\sqrt{T\ell_T(\delta/(1+k_T)^2)}\left(86\|x_1 - x_*\| + 1703\gamma_{k_T}\right)\,.\end{aligned}$$

Overall, by the union bound, we have with probability at least $1 - \sum_{k=1}^{\infty}\delta/(1+k)^2 \geqslant 1 - \delta$

$$\sum_{k=1}^{k_T}\mathtt{T}_2(k) \leqslant L\sqrt{T\ell_T(\delta/(1+k_T)^2)}k_T\left(86\|x_1 - x_*\| + 1703\gamma_{k_T}\right) \tag{37}$$

Regret bound Putting together (35) and (37), we obtain with probability $1 - \delta$

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant L\sqrt{T\ell_T(\delta/(1+k_T)^2)}k_T\left(87\|x_1 - x_*\| + 1704\gamma_{k_T}\right). \tag{38}$$

Bounding the number of phases It remains to bound the number of phases k_T. Fix some $t \geqslant 1$ and $k \geqslant k_{t-1}$. Observe that for any $k \geqslant 1$, $\|x_t^+(k) - x_t\| \leqslant \gamma_k$. Then thanks to Lemma 7 and Eq. (36) (which hold on exact same event that we consider in (38)), we have

$$\begin{aligned}\|x_t^+(k) - x_*\| &\leqslant \|x_t - x_*\| + \gamma_k \leqslant \|x_{T_{k_t}+1} - x_*\| + 16\gamma_{k_t} + \gamma_k \\ &\leqslant \|x_1 - x_*\| + 35\gamma_{k_t} + \gamma_k \leqslant \|x_1 - x_*\| + 36\gamma_k\,.\end{aligned}$$

Hence, for all $k \geqslant k_{t-1}$

$$\|x_t^+(k) - x_1\| \leqslant 2\|x_1 - x_*\| + 36\gamma_k \leqslant 2\gamma_{k^*} + 36\gamma_k\,.$$

Implying that $k_T \leqslant k^*$.

Concluding In view of the bound $k_T \leqslant k^* \leqslant \log_2(2D_{\gamma_0}/\gamma_0)$ and $\gamma_{k^*} \leqslant 2D_{\gamma_0}$, we conclude that with probability at least $1 - \delta$

$$\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leqslant 3500LD_{\gamma_0}\sqrt{T\ell_T(\delta/(1+k^*)^2)}\log_2(2D_{\gamma_0}/\gamma_0)\,.$$

Note that the constant 3500 is certainly extremely pessimistic as we did not attempt to optimize it.

□

C.1 High probability bound

We slightly adapt a version of the Bernstein-Freedman inequality, derived in [CBLS05, Corolary 16], improving $\log(T)$ dependency to $\log\log(T)$.

Lemma 6 (a version of the Bernstein-Freedman inequality by [CBLS05]). *Let $X_1, X_2, \dots$ be a martingale difference with respect to the filtration $\mathcal{F} = (\mathcal{F}_s)_{s\geqslant 0}$ and with increments bounded in absolute values by K. For all $t \geqslant 1$, let*

$$\mathfrak{S}_t^2 = \sum_{\tau=1}^{t} \mathbb{E}\big[X_\tau^2 \,\big|\, \mathcal{F}_{\tau-1}\big]$$

denote the sum of the conditional variances of the first t increments. Then, for all $\delta \in (0,1)$ and $T \geqslant 1$, with probability at least $1-\delta$,

$$\max_{t\leqslant T} \sum_{\tau=1}^{t} X_\tau \leqslant 2\mathfrak{S}_T \sqrt{\ln\left(\frac{\log_2(2T)}{\delta}\right)} + 3K \ln\left(\frac{\log_2(2T)}{\delta}\right).$$

Proof. Let $X_T^* = \max_{t\leqslant T} \sum_{s=1}^{t} X_s$. For $k \geqslant 1$ we have

$$\begin{aligned}
&\mathbb{P}\left[X_T^* > \sqrt{4(\mathfrak{S}_T^2 + K^2)\ell} + \sqrt{2}K\ell/3;\ K^{-2}\mathfrak{S}_T^2 \in [2^{k-1}-1, 2^k]\right] \\
&\leqslant \mathbb{P}\left[X_T^* > \sqrt{2^{k+1}K^2\ell} + \sqrt{2}K\ell/3;\ K^{-2}\mathfrak{S}_T^2 \in [2^{k-1}-1, 2^k]\right] \\
&\leqslant \mathbb{P}\left[X_T^* > \sqrt{2^{k+1}K^2\ell} + \sqrt{2}K\ell/3;\ K^{-2}\mathfrak{S}_T^2 \leqslant 2^k\right] \ \leqslant\ e^{-\ell},
\end{aligned}$$

where the last inequality follows from Lemma 15 in [CBLS05]. Since $0 \leqslant K^{-2}\mathfrak{S}_T^2 \leqslant T$, we take a union bound over $k = 1, \dots, \lceil \log_2(T) \rceil$ and notice that

$$\sqrt{4(\mathfrak{S}_T^2 + K^2)\ell} + \sqrt{2}K\ell/3 \leqslant 2\sqrt{\mathfrak{S}_T^2\ell} + 3K\ell\,. \qquad \square$$

The next result is a version of Lemma 1, that was used to analyze deterministic setup, which accounts for the stochasticity.

Lemma 7. *Let ρ be a bounded stopping time with respect to the filtration $\mathcal{F}$ of the stochastic gradients. Let $\delta \in (0,1)$, $T \geqslant 1$, $x_1' \in \mathcal{F}_\rho$ and consider $\ell_T(\delta) := 1 \vee \log(\log_2(2T)/\delta)$,*

$$x'_{t+1} = x'_t - \frac{\gamma}{L\sqrt{T\ell_T(\delta)}} g_{\rho+t}\,.$$

Assume that T, δ are such that $T \geqslant 1$, then with probability at least $1-\delta$ we have for all $\tau \leqslant T$

$$\begin{aligned}
&\|x'_\tau - x_*\| \leqslant \|x'_1 - x_*\| + 16\gamma\,, \\
&\sum_{t=1}^{\tau} (f(x'_t) - f_*) \leqslant L\sqrt{T\ell_T(\delta)} \left(\frac{\gamma}{2} + \frac{\|x'_1 - x_*\|^2 - \|x'_{\tau+1} - x_*\|^2}{2\gamma} + 10\|x'_1 - x_*\| + 68\gamma\right)
\end{aligned}$$

simultaneously.

Proof. To simplify the expressions, we drop the primes, writing x_t for x'_t, and we set $\eta = \frac{\gamma}{L\sqrt{T\ell_T(\delta)}}$.

Using classical analysis of projected gradient descent, we obtain

$$\sum_{t\leqslant\tau} \langle g_{\rho+t},\, x_t - x_*\rangle \leqslant \frac{\eta L^2}{2} + \frac{\|x_1 - x_*\|^2 - \|x_{\tau+1} - x_*\|^2}{2\eta}\,. \tag{39}$$

Introducing the following martingale difference

$$X_t = \langle \nabla f(x_t) - g_{t+\rho},\, x_t - x_*\rangle\,,$$

we deduce from the above, and the fact that $\tau \leqslant T$

$$\sum_{t=1}^{\tau} (f(x_t) - f(x_*)) \leqslant \frac{\eta L^2 T}{2} + \frac{1}{2\eta}\left(\|x_1 - x_*\|^2 - \|x_{\tau+1} - x_*\|^2\right) + \sum_{t=1}^{\tau} X_t\,. \tag{40}$$

Dealing with randomness. Now we are in position to apply Freedman-Bernstein inequality recalled in Lemma 6.

To this end, we need to bound X_t and get an appropriate expression for $\mathfrak{S}_t$. First observe that each martingale difference satisfies

$$|X_t| \leqslant 2L\|x_t - x_*\| \qquad \text{almost surely}\,. \tag{41}$$

Furthermore, by the triangle inequality, property of projection and the fact that $\|g_t\| \leqslant L$ almost surely, we obtain

$$\|x_t - x_*\| \leqslant \|x_{t-1} - \eta g_{\rho+t-1} - x_*\| \leqslant \|x_{t-1} - x_*\| + \eta L \leqslant \|x_1 - x_*\| + \eta L T, \quad \forall t \leqslant T+1\,.$$

Hence,

$$|X_t| \leqslant K := 2L\,\|x_1 - x_*\| + 2L^2 T\eta, \qquad \forall t \leqslant T+1.$$

The conditional variance $\mathfrak{S}_T^2$ can be bounded using (41) as

$$\mathfrak{S}_T \leqslant 2L\sqrt{\sum_{t=1}^{T}\|x_t - x_*\|^2} \leqslant 2L\sqrt{T}\max_{t\leqslant T}\|x_t - x_*\| \qquad \text{almost surely}\,.$$

Thus, invoking Lemma 6, for any $\delta \in (0,1)$ with probability at least $1-\delta$,

$$\max_{\tau\leqslant T}\sum_{t=1}^{\tau} X_t \leqslant 4L\max_{t\leqslant T}\|x_t - x_*\|\sqrt{T\ell_T(\delta)} + 6L(\|x_1 - x_*\| + L\eta T)\ell_T(\delta)\,. \tag{42}$$

From now on, we work on this event which holds with probability at least $1-\delta$.

Substituting (42) into (40), we get for all $\tau \leqslant T$

$$\begin{aligned}\sum_{t=1}^{\tau}(f(x_t) - f(x_*)) \leqslant &\frac{\eta L^2 T}{2} + \frac{\|x_1 - x_*\|^2 - \|x_{\tau+1} - x_*\|^2}{2\eta} + 4L\Phi_T\sqrt{T\ell_T(\delta)} \\ &+ 6L(\|x_1 - x_*\| + L\eta T)\ell_T(\delta)\,,\end{aligned} \tag{43}$$

where $\Phi_T = \max_{t\leqslant T}\|x_t - x_*\|$.

Bounding the trajectory. Observing that the left hand side of (43) is non-negative and that it holds for all $\tau \leqslant T$, we deduce

$$\begin{aligned}\max_{0\leqslant\tau\leqslant T}\|x_{\tau+1} - x_*\|^2 \leqslant &\left(\|x_1 - x_*\|^2 + 12L\eta\ell_T(\delta)\|x_1 - x_*\|\right) + (1 + 12\ell_T(\delta))\,\eta^2 L^2 T \\ &+ 8L\eta\Phi_T\sqrt{T\ell_T(\delta)}\,.\end{aligned}$$

Solving the above inequality, we deduce that

$$\begin{aligned}\Phi_T &\leqslant \sqrt{(\|x_1 - x_*\| + 6L\eta\ell_T(\delta))^2 + (1 + 28\ell_T(\delta))\,\eta^2 L^2 T} + 4L\eta\sqrt{T\ell_T(\delta)} \\ &\leqslant \|x_1 - x_*\| + 6L\eta\ell_T(\delta) + \eta L\sqrt{(1 + 28\ell_T(\delta))\,T} + 4L\eta\sqrt{T\ell_T(\delta)}\,.\end{aligned}$$

Substituting the value of η, we further deduce that

$$\begin{aligned}\max_{t\leqslant T}\|x_t - x_*\| &\leqslant \|x_1 - x_*\| + \gamma\left(\sqrt{\frac{1 + 28\ell_T(\delta)}{\ell_T(\delta)}} + 4 + 6\sqrt{\frac{\ell_T(\delta)}{T}}\right) \\ &\leqslant \|x_1 - x_*\| + \gamma\underbrace{\left(\sqrt{29} + 4 + 6\right)}_{\leqslant 15.5}\,.\end{aligned} \tag{44}$$

Bounding the regret On the other hand, substituting (42) into (40), we obtain

$$\sum_{t=1}^{\tau}(f(x_t)-f_*) \leqslant \frac{\eta L^2 T}{2} + \frac{\|x_1-x_*\|^2 - \|x_{\tau+1}-x_*\|^2}{2\eta} + 4L\max_{t\leqslant T}\|x_t-x_*\|\sqrt{T\ell_T(\delta)} + 6(L\|x_1-x_*\|+L^2\eta T)\ell_T(\delta)\,.$$

Substitution of (44) into the above inequality, yields

$$\sum_{t=1}^{\tau}(f(x_t)-f_*) \leqslant \frac{\eta L^2 T}{2} + \frac{\|x_1-x_*\|^2 - \|x_{\tau+1}-x_*\|^2}{2\eta} + 4L\left(\|x_1-x_*\|+15.5\gamma\right)\sqrt{T\ell_T(\delta)} + 6(L\|x_1-x_*\|+L^2\eta T)\ell_T(\delta)\,.$$

Recalling that $\eta = \gamma/(L\sqrt{T\ell_T(\delta)})$ and using some rough bounds, we deduce that

$$\sum_{t=1}^{\tau}(f(x_t)-f_*) \leqslant L\sqrt{T\ell_T(\delta)}\left(\frac{\gamma}{2} + \frac{\|x_1-x_*\|^2 - \|x_{\tau+1}-x_*\|^2}{2\gamma} + \underbrace{(4+6)}_{=10}\|x_1-x_*\| + \underbrace{(4\times 15.5+6)}_{=68}\gamma\right)\,.$$

The proof is complete. □

D On a relation with [MS12]

In case where there exists a known bound $\|g_t\| \leqslant L$ on the norms of the subgradients, MacMahan and Streeter [MS12] propose to tune the step size of gradient descent with a scheme based on a reward doubling argument and cold-restart. Their theory works in a setup of unconstrained online convex optimization with L-bounded subgradients. Since we do not require Lipschitz functions, and we additionally handle the projection step, the two results cannot be directly compared. Nevertheless, there are some similarities and, in a specific instantiation of our Algorithm 1, we recover that of [MS12].

Below, we sketch the relation between the two, considering the setting of MacMahan and Streeter [MS12], where the norms of the subgradients are bounded by some known L and the optimization is unconstrained, i.e. $\Theta = \mathbb{R}^d$. We also assume that the time horizon T is known, since unknown T is handled in MacMahan and Streeter [MS12] by a time-doubling trick. Using our notation, their analysis starts with a simple bound

$$\sum_{t=1}^{T}(f_t(x_t) - f_t(x_*)) \leqslant \sum_{t=1}^{T}\langle g_t, x_t - x_*\rangle = \sum_{t=1}^{T}\langle g_t, x_1 - x_*\rangle + \sum_{t=1}^{T}\langle g_t, x_t - x_1\rangle$$
$$\leqslant \|\underbrace{\sum_{t=1}^{T} g_t}_{=:G_T}\|\|x_1 - x_*\| - \underbrace{\sum_{t=1}^{T}\langle g_t, x_1 - x_t\rangle}_{=:Q_T} = \|G_T\|\|x_1 - x_*\| - Q_T\,.$$

They observe, using a duality argument, that it is sufficient to show that

$$Q_T \geqslant a^{-1}\exp\left(\|G_T\|/(bL\sqrt{T})\right) - c, \tag{45}$$

in order to derive

$$\sum_{t=1}^{T}(f_t(x_t) - f_t(x_*)) \leqslant b\|x_1 - x_*\|L\sqrt{T}\log\left(ab\|x_1 - x_*\|L\sqrt{T}\right) + c\,.$$

The principle of their algorithm is to perform gradient descent by phases and, during a phase, to track the reward Q_t relative to this phase, and restart with a doubled step-size when the condition $Q_t > \eta L^2 t$ is met. This step-size doubling ensures that the Condition (45) is met at the time horizon T.

Let us relate this algorithm with a specific instantiation of our Algorithm 1. When the algorithm is the simple Gradient Descent (GD), that does not involve the projection step, we have $x_{T+1} - x_1 = -\eta G_T$ for the GD with a fixed step size η. Hence, it holds that

$$Q_T = \sum_{t=1}^{T}\langle g_t, x_1 - x_t\rangle = \eta\sum_{t=1}^{T}\langle g_t, G_{t-1}\rangle = \frac{\eta}{2}\sum_{t=1}^{T}\left(\|G_t\|^2 - \|G_{t-1}\|^2 - \|g_t\|^2\right)$$
$$= \frac{\eta}{2}\left(\|G_T\|^2 - \sum_{1}^{T}\|g_t\|^2\right) = \frac{1}{2\eta}\|x_{T+1} - x_1\|^2 - \frac{1}{2\eta}\tilde{\Gamma}_{T+1}^2$$

where $\tilde{\Gamma}_{T+1}^2 = \sum_1^T \eta^2\|g_t\|^2$. Thus, if in Algorithm 1 we allow cold restarts (the exact thing that we want to avoid), then the condition

$$\|x_T^+(\eta) - x_1\|^2 \leqslant 2\eta^2 L^2 T + \tilde{\Gamma}_T^2 + \eta^2\|g_T\|^2$$

is equivalent to their doubling condition

$$Q_T \leqslant \eta L^2 T\,.$$

In our notation, the algorithm of MacMahan and Streeter [MS12] corresponds to a variant of Algorithm 1 with

$$\tilde{B}_{t+1}(k) = \sqrt{\left(2 + \frac{\|g_t\|^2}{L^2 T}\right)\gamma_k^2 + \Gamma_t^2 - \Gamma_{T_k - 1}^2}\,,$$

with the major difference that a cold-restart is performed when k_t is increased, and the minor difference that step-size doubling happens after (and not before) the condition $\left\|x_t^+(k) - x_1\right\| \leqslant \tilde{B}_{t+1}(k)$ is broken.

E Auxiliary results

Let $(a_t)_{t\geqslant 1}$ be a non-negative sequence, and $S_t = \sum_{\tau=1}^t a_\tau$. For any concave function F on $[0, +\infty)$, we have

$$\sum_{t=1}^T a_t F'(S_t) \leqslant \sum_{t=1}^T (F(S_t) - F(S_{t-1})) = F(S_T) - F(0). \tag{46}$$

Applying (46) with $F(x) = 2\sqrt{x}$, $F'(x) = 1/\sqrt{x}$, we get the following bound.

Lemma 8. *Let $(a_t)_{t\geqslant 1}$ be a non-negative sequence and $S_t = \sum_{\tau=1}^t a_\tau$, then for all $\varepsilon > 0$*

$$\sum_{t=1}^T \frac{a_t}{\sqrt{S_t}} \leqslant 2\sqrt{S_T}\,.$$

The inequality (46) with $F(x) = \log(\log(e(1+x)))$, $F'(x) = ((1+x)\log(e(1+x)))^{-1}$ gives the next lemma (the first inequality follows directly from Lemma 8).

Lemma 9. *Let $(a_t)_{t\geqslant 1}$ be a non-negative sequence and $S_t = \sum_{\tau=1}^t a_\tau$, then for all $\varepsilon > 0$*

$$\sum_{t=1}^T \frac{a_t}{\sqrt{(S_t+1)\log(\mathrm{e}(1+S_t))}} \leqslant 2\sqrt{S_T}\,,$$
$$\sum_{t=1}^T \frac{a_t}{(S_t+1)\log(\mathrm{e}(1+S_t))} \leqslant \log(\log(\mathrm{e}(1+S_T)))\,.$$

Using (46), with $F(x) = \log(x)$, we deduce that

Lemma 10. *Let $(a_t)_{t\geqslant 1}$ be a non-negative sequence and $S_t = \sum_{\tau=1}^t a_\tau$, then*

$$\sum_{t=1}^T \frac{a_t}{S_t} = 1 + \sum_{t=2}^T \frac{a_t}{S_t} \leqslant 1 + \log(S_T/S_1)\,. \tag{47}$$

Using (46), with $F(x) = \log(\varepsilon + x)$, we deduce that

Lemma 11. *Let $(a_t)_{t\geqslant 1}$ be a non-negative sequence and $S_t = \sum_{\tau=1}^t a_\tau$, then*

$$\sum_{t=1}^T \frac{a_t}{\varepsilon + S_t} \leqslant \log(1 + S_T/\varepsilon)\,. \tag{48}$$

Lemma 12. *Let us define $\bar{k}$ as the smallest integer fulfilling $\bar{k}^{-1/2} 2^{\bar{k}} \geqslant 2^{k^*}$. Then*

$$\bar{k} \leqslant k^* + 0.5\log_2(k^*) + 1.25\ .$$

Proof. We observe that $\bar{k} = 1$ for $k^* = 1$. We will prove that

$$\bar{k} \leqslant \lceil k^* + 0.5\log_2(k^*) + 0.25\rceil \leqslant k^* + 0.5\log_2(k^*) + 1.25\,.$$

For proving the first inequality, we only need to prove that $2^y/\sqrt{y} \geqslant 2^{k^*}$ for $y = k^* + 0.5\log_2(k^*) + 0.25$. Plugging the value of y and taking the square, we get

$$\frac{2^y}{\sqrt{y}} \geqslant 2^{k^*} \Leftrightarrow \frac{2^{1/4} 2^{k^*}\sqrt{k^*}}{\sqrt{k^* + 0.5\log_2(k^*) + 0.25}} \geqslant 2^{k^*} \Leftrightarrow \sqrt{2}\,k^* \geqslant k^* + 0.5\log_2(k^*) + 0.25\ .$$

So, all we need is to prove by induction that

$$\sqrt{2}\,k^* \geqslant k^* + 0.5\log_2(k^*) + 0.25\,.$$

For $k^* = 2$, the inequality holds. By induction hypothesis, for $k^* \geqslant 2$

$$\sqrt{2}(k^*+1) \geqslant (k^*+1) + \frac{1}{2}\log_2(k^*) + \frac{1}{4} + \sqrt{2} - 1\,.$$

Thus, it suffices to show that

$$\frac{1}{2}\log_2(k^*) + \sqrt{2} - 1 \geqslant \frac{1}{2}\log_2(k^*+1), \quad \text{for all } k^* \geqslant 2\,,$$

or equivalently

$$\frac{1}{2}\log_2\left(1+\frac{1}{k^*}\right) + 1 \leqslant \sqrt{2}, \quad \text{for all } k^* \geqslant 2\,,$$

to prove the induction. The concavity of $\log_2$ ensures that $\log_2(1+x) \leqslant x/\ln(2)$ for all $x > -1$. Thus, for all $k^* \geqslant 2$

$$\frac{1}{2}\log_2\left(1+\frac{1}{k^*}\right) + 1 \leqslant \frac{1}{2\ln(2)k^*} + 1 \leqslant \frac{1}{4\ln(2)} + 1 \leqslant \sqrt{2}\,,$$

which concludes the proof of Lemma 12. □