Università degli Studi di Padova

**Marcello Bullo**
**Francesco Mandruzzato**
MSc. students
University of Padova
ID: 1204533- 1204532
marcello.bullo@studenti.unipd.it
francesco.mandruzzato.1@studenti.unipd.it

# Final Project:
# Accelegrad and UnixGrad: a Comparison Analysis

Optimization, Francesco Rinaldi
19th February 2021

## Contents

# Introduction

In the context of Stochastic Convex Optimization (SCO), we analyze the performance of two new algorithms, AcceleGrad and UnixGrad for unconstrained and constrained optimization respectively. We study theoretical results in order to derive the convergence rate for different settings and we analyze their performance on some SVM problems.

# General definitions

**Definition 0.1.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is $\beta$-smooth if:*

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|x - y\|^2, \forall x, y \in \mathbb{R}^n$$

**Definition 0.2.** *$f$ is G-Lipschitz if $\forall w \in \Omega$ we have $\|\nabla f_i(w)\| \leq G$*

**Definition 0.3.** *Given a real-valued convex function $f$ with elements $x_1, \dots x_n$ in its domain and positive weights $a_i$, Jensen's inequality is stated as:*

$$f\left(\frac{\sum a_i x_i}{\sum a_i}\right) \leq \frac{\sum a_i f(x_i)}{\sum a_i}$$

# UnixGrad

UnixGrad is a universal method for constrained SCO that obtains the optimal rates in both smooth and non-smooth cases, $\mathcal{O}(GD/\sqrt{T})$ and $\mathcal{O}(LD^2/T^2 + \sigma D/\sqrt{T})$ respectively, without any prior knowledge regarding the smoothness of the problem $L$, nor the noise magnitude $\sigma$. Such algorithm is build on top of the Mirror-Prox method by Nemrovski [1] using some techniques taken from online learning literature and an adaptive learning rate. Hence, before explore the details of UnixGrad method, let us recall Mirror-Prox which represents the core block for this algorithm.

## UnixGrad algorithm setting

UnixGrad algorithm focuses on (approximately) solving the following constrained problem,

$$\min_{x \in \mathcal{K}} f(x), \tag{1}$$

where $f : \mathcal{K} \mapsto \mathbb{R}$ and $\mathcal{K} \subset \mathbb{R}^d$ is a compact convex set.
For the analysis, two settings will be keeping in consideration: a deterministic setting where exact gradients are accessible and a stochastic one where only unbiased (noisy) gradient estimates are provided. This means that when such an oracle is queried with a point $x \in \mathcal{K}$, we receive $\tilde{\nabla} f(x) \in \mathbb{R}^d$ such that

$$\mathbb{E}\left[\tilde{\nabla} f(x)|x\right] = \nabla f(x).$$

Furthermore, from now on we assume that the norm of the (sub)-gradient estimates is bounded by $G$, i.e.,

$$\|\tilde{\nabla} f(x)\|_* \leq G.$$

## Mirror-Prox

Mirror-Prox algorithm is a variant of Mirror Descent (MD) introduced by Nemirovski [1] in 2004. Due to the fact that it extends MD, such algorithm is designed to work with arbitrary norm definition and

extend the optimization framework to a more general situation in some Banach space $\mathcal{B}$. In such a space the gradient descent strategy make no sense since the gradient is a linear mapping from $\mathcal{B} \to \mathbb{R}$(or $\mathbb{C}$) and it belongs to $\mathcal{B}^*$ (dual Banach space) and the starting point is in $\mathcal{B}$ (no such problem in Hilbert space because $\mathcal{H}$ is isometric to $\mathcal{H}^*$). The general solution to this problem is mapping a point $x \in \mathcal{X} \cap \mathcal{B}$ into $\mathcal{B}^*$, perform the gradient update in $\mathcal{B}^*$ (dual space), mapping the resulting point $\tilde{x}_{k+1}$ backward to the primal space and, if $\tilde{x}_{k+1} \notin \mathcal{X} \cap \mathcal{B}$, then project it into the feasible subset. Clearly, to deal with this situation mirror maps have been used and from such combination Mirror Descent have been created.

**Mirror descent**

1. A point $x \in \mathcal{X} \cap \mathcal{B}$ is mapped to $\nabla\Phi(x)$(from primal to dual space, since $\nabla\Phi(x) \in \mathcal{B}^*$);
2. Take a gradient step in the dual space, that is
$$\nabla\Phi(x) - \eta g_t.$$
Notice that a function in order to be a mirror map must satisfy $\nabla\Phi(\mathcal{D}) = \mathbb{R}^n$ (take all possible values) and for this reason we can write
$$\nabla\Phi(y) = \nabla\Phi(x) - \eta g_t;$$
3. Since the primal point $y \in \mathcal{D}$ may lie out of $\mathcal{X}$, a projection onto $\mathcal{X}$ is needed. Such projection is done via Bregman divergence associated with $\Phi$. Precisely:
$$\Pi^{\Phi}_{\mathcal{X}}(y) = \text{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} D_\Phi(x, y)$$

So the updating rule for MD is $x_{t+1} = \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}} D_\Phi(x, y_{t+1})$ where $\nabla\Phi(y_{t+1}) = \nabla\Phi(x_t) - \eta g_t$. Notice that MD strategy can be rewritten in the following way:

$$\begin{aligned}
\underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ D_\Phi(x, y_{t+1}) &= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ \Phi(x) - \Phi(y_{t+1}) - \nabla\Phi(y_{t+1})^\top (x - y_{t+1}) \\
&= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ \Phi(x) - \Phi(y_{t+1}) - (\nabla\Phi(x_t) - \eta g_t)^\top (x - y_{t+1}) \quad \text{(definition of } \nabla\Phi(y_{t+1})) \\
&= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ \Phi(x) - (\nabla\Phi(x_t) - \eta g_t)^\top x \quad \text{(remove constant term w.r.t. } x) \\
&= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ \eta g_t^\top x + \Phi(x) - \nabla\Phi(x_t)^\top x + \nabla\Phi(x_t)^\top x_t - \Phi(x_t) \quad \text{(add constant terms)} \\
&= \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ \eta g_t^\top x + D_\Phi(x, x_t)
\end{aligned}$$

which give a proximal point of view in fact it tries to minimize the local linearization of the function while moving not too far away from the previous step.

Mirror-Prox follows the same passages of MD but add one more inner step for the updating rule, that is

1. $\nabla\Phi(y'_{t+1}) = \nabla\Phi(x_t) - \eta\nabla f(x_t)$
2. $y_{t+1} \in \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ D_\Phi(x, y'_{t+1})$
3. $\nabla\Phi(x'_{t+1}) = \nabla\Phi(x_t) - \eta\nabla f(y_{t+1})$
4. $x_{t+1} \in \underset{x \in \mathcal{X} \cap \mathcal{D}}{\text{argmin}}\ D_\Phi(x, x'_{t+1})$

Again, following the same passages taken for MD we can obtain a proximal view of the algorithm which is the following.

**Algorithm 1:** Mirror-Prox Template

**Input :** #Iterations T, $y_0 \in \mathcal{X}$, learning rate $\{\eta_t\}_{t\in[T]}$

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

$$x_t = \underset{x\in\mathcal{X}\cap\mathcal{D}}{\operatorname{argmin}} \ \langle x, M_t\rangle + \frac{1}{\eta_t} D_{\mathcal{R}}(x, y_{t-1}) \qquad (M_t = \nabla f(y_{t-1}))$$

$$y_t = \underset{y\in\mathcal{X}\cap\mathcal{D}}{\operatorname{argmin}} \ \langle x, g_t\rangle + \frac{1}{\eta_t} D_{\mathcal{R}}(y, y_{t-1}) \qquad (g_t = \nabla f(x_t))$$

## The Algorithm

As outlined before, UnixGrad algorithm is built upon Mirror-Prox template (Algorithm 1) but with some changes on the spot where gradients are computed and the adaptivity of the learning rate. In particular the following changes are taken into account:

- in order to have a more stable behavior, the notion of averaging is implemented, taken inspiration from different interpretation of acceleration. Let's define $\alpha_t = t$, then

$$\bar{x}_t = \frac{\alpha_t x_t + \sum_{i=1}^{t-1} \alpha_i x_i}{\sum_{i=1}^{t} \alpha_i}, \qquad \tilde{z}_t = \frac{\alpha_t y_{t-1} + \sum_{i=1}^{t-1} \alpha_i x_i}{\sum_{i=1}^{t} \alpha_i};$$

- gradients are computed at weighted average of iterates, that is $g_t = \nabla f(\bar{x}_t)$ and $M_t = \nabla f(\tilde{z}_t)$;
- an adaptive learning rate is adopted in particular the so called lag-one-behind learning rate, that is

$$\eta_t = \frac{2D}{\sqrt{1 + \sum_{i=1}^{t-1} \alpha_i^2 \left\| g_i - M_i \right\|_*^2}},$$

where $D^2 = \sup_{x,y\in\mathcal{K}} D_{\mathcal{R}}(x, y)$ is the diameter of the compact set $\mathcal{K}$ with respect to the Bregman divergences.

Algorithm 2 summarizes UnixGrad idea.

**Algorithm 2:** UnixGrad

**Input :** #Iterations T, $y_0 \in \mathcal{X}$, weight $\alpha_t = t$, learning rate $\{\eta_t\}_{t\in[T]}$

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

$$x_t = \underset{x\in\mathcal{K}}{\operatorname{argmin}} \ \langle x, M_t\rangle + \frac{1}{\eta_t} D_{\mathcal{R}}(x, y_{t-1}) \qquad (M_t = \nabla f(\tilde{z}_t))$$

$$y_t = \underset{y\in\mathcal{X}}{\operatorname{argmin}} \ \langle x, g_t\rangle + \frac{1}{\eta_t} D_{\mathcal{R}}(y, y_{t-1}) \qquad (g_t = \nabla f(\bar{x}_t))$$

**Output:** $\bar{x}_T$

## Convergence results

In this section convergence results is shown for both smooth and non-smooth cases. Proofs of convergence rates are reported in the appendices in the UnixGrad section while, here, the results are simply reported. Anyway the convergence analysis is performed in the sense of bounding "weighted regret". Then a simple conversion strategy is adopted to directly convert the weighted regret to convergence rate

### Non-smooth setting

- **Deterministic setting**

**Theorem 1.** *Consider the constrained optimization setting where $f : \mathcal{K} \mapsto \mathbb{R}$ is a proper, convex and G-Lipschitz function defined over compact, convex set athcal$K$. Let $x^* \in \min_{x\in\mathcal{K}} f(x)$. Then, Algorithm 2 guarantees*

$$f(\bar{x}_T) - \min_{x\in\mathcal{K}} f(x) \leq \frac{7D\sqrt{1+\sum_{t=1}^{T}\alpha_t^2 \|g_t - M_t\|_*^2} - D}{T^2} \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}$$

- **Stochastic setting**

**Theorem 2.** *Consider the optimization setting where $f$ is non-smooth, convex and G-Lipschitz. Let $\{x_t\}_{t=1,..,T}$ be a sequence generated by UnixGrad such that $g_t = \tilde{\nabla} f(\bar{x}_t)$ and $M_t = \tilde{\nabla} f(\tilde{z}_t)$. With $\alpha_t = t$ and the lag-behind-one learning rate, it holds that*

$$\mathbb{E}\left[f(\bar{x}_T)\right] - \min_{x\in\mathcal{K}} f(x) \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}$$

**Smooth setting**

- **Deterministic setting**

**Theorem 3.** *Consider the constrained optimization setting where $f : \mathcal{K} \to \mathbb{R}$ is a proper, convex and $L$ -smooth function defined over compact, convex set $\mathcal{K}$. Let $x^* \in \min_{x\in\mathcal{K}} f(x)$. Then, Algorithm 2 ensures the following*

$$f(\bar{x}_T) - \min_{x\in\mathcal{K}} f(x) \leq \frac{20\sqrt{7}D^2L}{T^2}$$

- **Stochastic setting**

**Theorem 4.** *Consider the optimization setting where $f$ is $L$-smooth and convex. Let $\{x_t\}_{t=1,...,T}$ be a sequence generated by UnixGrad such that $g_t = \tilde{\nabla} f(\bar{x}_t)$ and $M_t = \tilde{\nabla} f(\tilde{z}_t)$ With $\alpha_t = t$ and the lag-behind-one learning rate, it holds that*

$$\mathbb{E}\left[f(\bar{x}_T)\right] - \min_{x\in\mathcal{K}} f(x) \leq \frac{112\sqrt{14}D^2L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}}$$

# Accelegrad

## Previous related work

The Accelerated Gradient Descent method devised by Nesterov is a popular optimization technique which provides faster convergence rate $\mathcal{O}(1/T^2)$ in the smooth setting, guaranteeing standard rate $\mathcal{O}(1/\sqrt{T})$ for a general convex objective function $f$. The structure of the algorithm is divided in two different steps:

1. **Extrapolation step**: We move along the direction of the difference between the last two iterates:

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \tag{2}$$

   were $\beta_k$ is chosen depending on the properties of $f$. Ideally this parameter regulates the amount of information that we want to keep regarding past iterates.

2. **Gradient step** We performe a step likewise the GD algorithm in $y_k$ to get the next point $x_{k+1}$:

$$x_{k+1} = y_k - \alpha_k \nabla f(y_k), \tag{3}$$

   where $\alpha_k = 1/L$

Despite its efficiency, Accelerated Gradient Descent is inappropriate for handling noisy feedback because of the choice of the fixed step size $\alpha_k$, moreover acceleration requires the knowledge of the objective's smoothness.

One of the main popular adaptive first order methods comes with AdaGrad [2]. Informally, AdaGrad exploits the geometry of the dataset in the sense that the adaptation of the step size $\alpha_k$ vary with the dimension considered, in other words, it associates a small step size to frequently occurring features and a large step size to infrequent features. This adaptation is encoded in its update rule:

$$\eta_t = \frac{D}{\sqrt{2\sum_{\tau=1}^{T} \|g_\tau\|^2}} \tag{4}$$

where $g_\tau$ are the subgradients at time step $\tau$, and $D$ is the diameter of the considered set. With this adaption AdaGrad may handle noisy feedback, however there is no guarantee that AdaGrad can ensure acceleration, moreover it was unknown whether AdaGrad is able to exploit the smoothness in order to converge faster. Considering the advantage of handling noisy feedback to ensure a solid acceleration, Accelegrad takes inspiration from the update rule of AdaGrad.

Regarding acceleration, an elegant way of interpreting it comes from [3] and we provide the high level concept which is useful to understand the algorithm steps.

In *Allen-Zhu* and *Orecchia*, they combine standard Gradient Descent and Mirror Descent to yield a new and simple accelerated gradient method for smooth convex optimization problems. Given an $L$-smooth function $f$ we are able to find a quadratic upper bound of the form:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2. \quad \forall y$$

This is particularly useful to find an upper bound to the maximal objective decrease given as:

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2.$$

From this point it should be clear that the larger the magnitude of the gradient the better may be the convergence speed of Gradient Descent. Mirror Descent instead peeks a uniformly distributed sequences of points $(x_1, \ldots, x_n)$ and it combines them to construct a stronger lower bound to the function $f$ which is not provided by the standard Gradient Descent, in particular this lower bound is obtained as:

$$f(y) \geqslant \frac{1}{n}\sum_{t=0}^{n-1} f(x_t) + \frac{1}{n}\sum_{t=0}^{n-1} \langle \nabla f(x_t), y - x_t \rangle \quad \forall y.$$

To construct a good upper bound instead we consider the mean of the queried points and by the convexity property we get $f(\hat{x}) \leq \frac{1}{n}\sum_{t=1}^{n} f(t_i)$. Combining this with the previous result we get:

$$f(\hat{x}) - f(y) \leq \frac{1}{n}\sum_{t=0}^{n-1} \langle \nabla f(x_t), y - x_t \rangle,$$

which is known as the regret bound $R_n(y)$ for the sequence $\{x_t\}_{t=0}^{n-1}$. In the Appendix we provide a rigorous analysis which finds an upper bound on the the regret and show that Mirror Descent converges with $\mathcal{O}(\rho^2/\epsilon^2)$ iterations, where $\rho^2$ is equal to the average value of $\|\nabla f(x_k)\|^2$ across the iterations. In particular we demonstrate that the smaller the queries gradients, the faster is Mirror descent to converge.

In light of this consideration, they propose a linear combination of the two algorithms at each step, in particular following Nesterov, given the gradient step $y_k$ and the mirror step $z_k$ the next step will be given by:

$$x_{k+1} = \alpha z_k + (1 - \alpha) y_k, \tag{5}$$

where $\alpha_k$ is a tuneable parameter. It is possible to prove that this technique provides the same convergence guarantees of the accelerated gradient method of Nesterov in a nicer and simpler way, for this reason Accelegrad incorporates the notion of acceleration taking inspiration from this linear coupling framework.

## The Algorithm

In the Accelegrad paper [4], there is no assumption about the smoothness parameter $\beta$ (See Def.0.1 of smoothness), but they assumed to be given a bound between some initial point $x_0$ and a global minimum of $f$ for a given compact convex set $K$ which contains this global minimizer. This distance is bounded by the diameter of $K$ i.e. $D = \max_{x,y \in K} \|x - y\|$. Finally they assume that the function $f$ is G-Lipschitz for which we provide the definition in (Def.0.2).
First, we discuss the offline setting where we have always access to the exact gradients of $f$, the algorithm is presented in Algo.3. As we previously said, Accelegrad takes inspiration from [3] and linearly couples between two sequences $\{y_t\}_t$ and $\{z_t\}_t$ into $\{x_{t+1}\}_t$. These two sequences are updated in two different manners, the former $(y_{t+1})$, takes a step starting from $x_{t+1}$, the latter $(z_{t+1})$, takes a step starting from $z_t$. Both sequences are updated with the same step size but for $z_{t+1}$ the gradient is scaled by a factor of $\alpha_t$ and the final result is projected into $K$. The step size is defined taking inspiration from AdaGrad:

$$\eta_t = \frac{2D}{(G^2 + \sum_{\tau=0}^{t} \alpha_\tau^2 \|g_t\|^2)^2}, \tag{6}$$

where the only difference are the importances weights $\alpha_t$ defined as:

$$\alpha_t = \begin{cases} 1 & 0 \leq t \leq 2 \\ \frac{1}{4}(t+1) & t \geqslant 3 \end{cases}, \tag{7}$$

which increase with $t$, putting emphasis on recent queries of the gradient. As we are considering the smooth setting it is not required the function to be $G$-Lipschitz which usually becomes important in the non smooth setting.

**Algorithm 3:** Accelerated Adaptive Gradient Method (AcceleGrad)

**Input :** #Iterations T, $x_0 \in K$, diameter $D$, weights $\alpha_{t \in T}$, learning rate $\eta_t$
Set $y_0 = x_0 = z_0$
**for** $t = 1 \dots T$ **do**
Set $\tau_t = 1/\alpha_t$
Update:

$$x_{t+1} = \tau_t z_t + (1 - \tau_t) y_t, \quad g_t = \nabla f(x_{t+1})$$

$$z_{t+1} = \Pi_K(z_t - \alpha_t \eta_t g_t)$$

$$y_{t+1} = x_{t+1} - \eta_t g_t$$

**Output:** $\hat{y}_T \propto \sum_{t=0}^{T-1} \alpha_t y_{t+1}$

### Convergence results

For the smooth setting Accelegrad requires a number of iterations to get a solution within $\epsilon$ of $\mathcal{O}(\frac{c}{\epsilon})^{\frac{1}{2}}$ with $c = DG + \beta D^2 log(\beta D/G)$, which means that the algorithm is able to converge as fast as the Accelerated gradient method asymptotically. Actually, since we are dealing with smooth functions we don't need to consider $G$, which means that the constant $c$ is actually equal to $c = \beta D log(\beta D/\|g_0\|)$.
Considering the non smooth setting the previous algorithm is able to ensure standard sublinear convergence rate of $\mathcal{O}(1/\sqrt{T})$ approximately. With this rate of convergence Accelegrad requires a number of iterations to get a solution within $\epsilon$ of $\mathcal{O}(1/\epsilon^2)$ approximately, this result is also true for the stochastic optimization setting where usually state of art algorithms rely on a line search procedure which is inappropriate for noisy feedback. In the Appendix we provide a short version of the proofs regarding the convergence rate for these three cases. It is important to mark the fact that Accelegrad is able to adapt to the smoothness parameter as it is independent of it, therefore it is referred as universal algorithm.

## Online to Batch conversion, brief overview

Online learning is the process of answering a sequence of questions given knowledge of the correct answers to the previous ones. As described in [5], given a sequence of questions $S = (x_1, \ldots, x_t)$, for each one of them sequentially our algorithm makes a prediction and then after the prediction the correct answer is revealed incurring in a loss. The goal of the learner is to minimize the loss also exploiting past information. If there is no correlation between samples, learning is hopeless.
Intuitively, for each sample passed to the learner at stage $t$, the learner chose an hypothesis $h$ from the hypothesis class $H$ and predicts the label associated with the sample. For this reason we can define a regret function which measure "how sorry" the learner is not to have followed the best predictor $h^*$. Formally when running on a sequence of $T$ examples:

$$R_T(h^*) = \sum_{t=1}^{T} l(p_t, y_t) - \sum_{t=1}^{T} l(h^*(x_t), y_t),$$

where $p_t$ is the prediction obtained by $h$ at stage $t$, and the regret w.r.t the hypothesis class $H$ is defined as:

$$R_T(H) := \max_{h^* \in H} R_T(h^*).$$

Usually we look for an algorithm which converge sub-linearly to the lowest possible regret.
What we described in short say that an online learning algorithm wants performance close to the single best hypothesis chosen in hindsight given all the data.
Different from this practice, an online to batch converter needs to construct a model based on the set of modules generated by the online learner, this can be done in different ways but one of the most classical way is to take the average of these models. This conversion is needed because the individual iterates of an online algorithm do not come with individual guarantees. The standard online to batch conversion algorithm format is shown in Algo.4. In particular since the algorithm outputs an average of the iterates, we can apply Jensen's inequality to show the following:

$$\mathbb{E}[L(\hat{w}) - L(x^*)] \leq \frac{\mathbb{E}[R_T(x^*)]}{T},$$

and as long as the algorithm obtains sublinear regret the left side part of the inequality will approach zero an average.
In [6] to avoid the guarantees problems of single iterates it is provided a black box "Anytime" Online to Batch conversion algorithm. The "Anytime property" comes from the fact that the last iterate is always a good estimate of $x^*$. The algorithm is very similar to the classical online to batch previously shown. The key difference is that they evaluate the stochastic gradient oracle at $x_t$ rather than the iterates provided by the algorithm $A$, furthermore they incorporate the importance weights $\alpha_t$ which are useful to achieve faster convergence rates on smooth losses, similarly to what is done in Algorithm.3.

**Algorithm 4:** Online to batch conversion

**Input :** #Iterations $T$, Params $S$, Cost function $l$, Algorithm $A$
**for** $t = 1 \dots T$ **do**
  let $w_t$ be the prediction of $A$, provide the cost function $l(w_t, g_t)$ to $A$ where $\mathbb{E}[g_t] = \nabla F(w_t)$
**Output:** $\hat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

**Algorithm 5:** Anytime Online to batch conversion

**Input :** #Iterations $T$, Online Algorithm $A$ with convex domain $D$, Non-negative weights $\alpha_t$ with $\alpha_t > 0$
**for** $t = 1 \dots T$ **do**
  Get $x_t = \frac{\sum_{i=1}^{t} \alpha_i w_i}{\sum_{i=1}^{t} \alpha_i}$
  Play $x_t$ and receive subgradient $g_t$
  Send $l_t(x) = \langle \alpha_t g_t, x \rangle$ to $a$ as the $t$th loss
  Get $w_{t+1}$ from $A$
**Output:** $x_T$

Different from the standard online to batch conversion, this algorithm is able to achieve the following guarantees for every $x^* \in D$:

$$\mathbb{E}[L(x_T) - L(x^*)] \leq \frac{\mathbb{E}[R_T(x^*)]}{\sum_{t=1}^{T} \alpha_t},$$

in particular this bound is valid also for loss functions with some known non-linearity. In the Appendix we provide a proof sketch of this bound.
We now consider the analysis of *A.Cutkosky* regarding an extension of the Accelegrad algorithm, in particular we show that Algorithm.5 can adapt to both smoothness and variance optimally which is a step further Accelegrad as it adapts also to variance. Unlike in [4] we require the objective function $L$ to be defined on the entire vector space rather than the compact closed set $K$ and we assume $\|x^*\| \leq B/2$ for some parameter $B$. The pseudocode is shown in Algorithm.6.

**Algorithm 6:** Adaptive Stochastic Acceleration

**Input :** Bound $B \geq 2\|x^*\|$, value $c$, Online Learning Algorithm $A$ with domain $D = \{\|w\| \leq B/2\}$
  Get initial point $w_1 \in D$
**for** $t = 1 \dots T$ **do**
  Set $\alpha_t = t$
  $\tau_t = \frac{\alpha_t}{\sum_{i=1}^{t} \alpha_i}$
  $x_t = \tau_t w_t + (1 - \tau_t) y_{t-1}$
  Play $x_t$ receive subgradient $g_t$
  $\eta_t = \frac{cB}{\sqrt{1 + \sum_{i=1}^{t} \alpha_{1:i} \|g_i\|^2}}$
  $y_t = x_t - \eta_t g_t$
  Send $l(x_t) = \langle \alpha_t g_t, x \rangle$ to $A$ as the $th$ loss
  Get $w_{t+1}$ from $A$
**Output:** $x_T$

In the appendix we provide the detailed analysis for this algorithm (See Theorem 6) Finally they show that Algorithm 6 in the non-smooth setting can recover the convergence rate of $\mathcal{O}(1/\sqrt{T})$ (See Theorem 7).

# Results

In this section we provide a comparison between AcceleGrad and UnixGrad. We test both algorithms on SVM (Support Vector Machine) problem over two datasets taken from LIBSVM, that is, given $y \in \{-1, 1\}^n, X \in \mathbb{R}^{n \times p}$ with rows $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$

$$\min_{\boldsymbol{w}, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{n} \xi_i$$
$$\text{subject to } \xi_i \geq, y_i(\boldsymbol{x}_i^\top \boldsymbol{w}) \geq 1 - \xi_i, \quad i = 1, \ldots, n.$$

By rewriting the constraints as follows

$$\xi \geq \max\{0, 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w})\}$$

and we plug it in the SVM problem definition we obtain the so-called *regularized hinge loss*

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{n} \max\{0, 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w})\}.$$

Actually, in order to test both UnixGrad and AcceleGrad we decided to use the *regularized squared hinge loss*

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{n} \left(\max\{0, 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w})\}\right)^2.$$

In particular we worked with *breast-cancer* and *adult* datasets. We try to minimize the L2-regularized squared hinge loss using random mini-batches of size 5, with random weights initialization. For both UnixGrad and Accelegrad we measure the performance with respect to the average iterate which results in a more stable behaviour. As stated in [7] for both algorithm we consider the performance with respect to the average iterate as it shows a more stable behaviour. In what follows we provide the same analysis done in [7] for the breast cancer dataset and for the adult dataset:

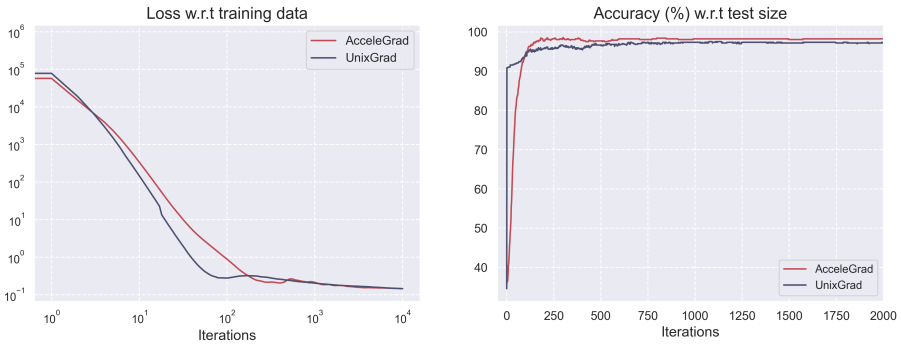


**Figure 1:** Comparison results w.r.t the breast-cancer dataset

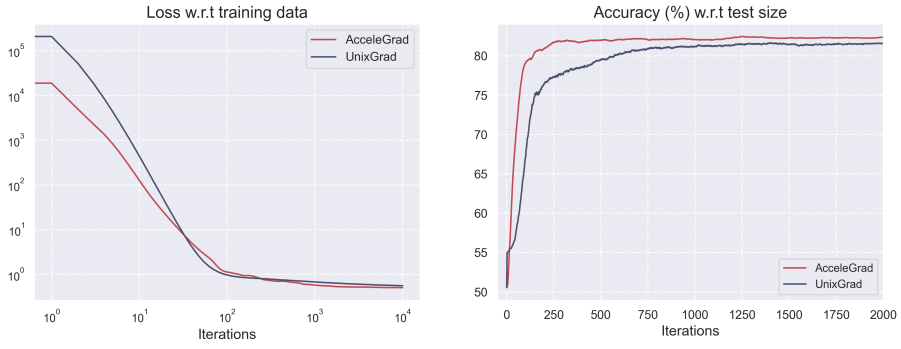


**Figure 2:** Comparison results w.r.t the adult dataset

From the results obtained we can see that UnixGrad and AcceleGrad achieve comparable generalization performances to each other. Regarding Accelegrad we show an interesting but expected phenomena which is reported in Fig.3. As we increase the batch size the performance improves, the intuition comes from the fact that upon using small batches $b$ the gradient estimate is noisy and Accelegrad obtain the slow rate $\mathcal{O}(1/\sqrt{T})$, while for large batches we can exploit the acceleration $\mathcal{O}(1/T^2)$. More precisely, given the number of gradient calculations as $N = bT$, we get $T = N/b$ which translates in a convergence rate of $\mathcal{O}(\sqrt{b}/\sqrt{N})$ for the stochastic setting and $\mathcal{O}(b^2/N^2)$ in the accelerated one.

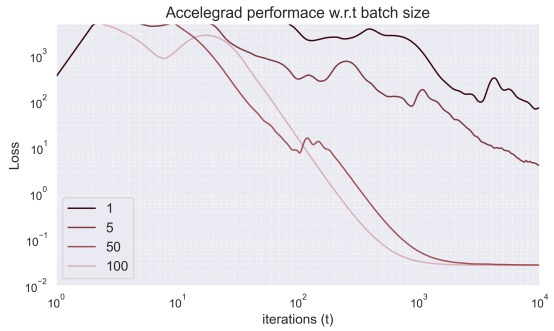


**Figure 3:** Accelegrad performance changing the batch size

# References

[1] A. Nemirovski, "Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2004.

[2] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 07 2011.

[3] Z. A. Zhu and L. Orecchia, "A novel, simple interpretation of nesterov's accelerated method as a combination of gradient and mirror descent," *CoRR*, vol. abs/1407.1537, 2014.

[4] K. Y. Levy, A. Yurtsever, and V. Cevher, "Online adaptive methods, universality and acceleration," 2018.

[5] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and TrendsÂ® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[6] A. Cutkosky, "Anytime online-to-batch conversions, optimism, and acceleration," 2019.

[7] A. Kavis, K. Y. Levy, F. Bach, and V. Cevher, "Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization," 2019.

# Appendices

## Anytime online-to-batch

### Algorithm guarantees

**Theorem 5.** *Suppose $g_1, \dots, g_t$ satisfy $\mathbb{E}[g_t|x_t] \in \delta L(x_t)$ for some objective function $L$ and $g_t$ is independent of all other quantities given $x_t$. Let $R_T(x^*)$ be a bound on the linearized regret of the Anytime Online to Batch algorithm.*

$$R_T(x^*) \geqslant \sum_{t=1}^{T} \langle \alpha_t g_t, w_t - x^* \rangle$$

*Then for all $x^* \in D$, the algorithm guarantees:*

$$\mathbb{E}[L(x_t) - L(x^*)] \leq \mathbb{E}\left[\frac{R_T(x^*)}{\sum_{t=1}^{T} \alpha_t}\right]$$

*Furthermore, suppose that $D$ has a diameter $B$ and $\|g_t\|_* \leq G$ with probability 1 for some $G$. Then with probability at least $1 - \delta$:*

$$L(x_T) - L(x^*) \leq \frac{R_T(x^*) + 2BG\sqrt{\sum_{t=1}^{T} \alpha_t^2 log(2/\delta)}}{\alpha_{1:T}}$$

*Proof.* In order to prove this theorem we observe that $\alpha_t(x_t - w_t) = \alpha_{1:t-1}(x_{t-1} - x_t)$, where $\alpha_{1:t-1} = \sum_{i=1}^{t-1} \alpha_i$. First we use the property of convexity to bound $\mathbb{E}[\alpha_t(L(x_t) - L(x^*))]$with $\mathbb{E}[\sum_{t=1}^{T} \alpha_t \langle g_t, x_t - x^* \rangle]$ and we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \left(\mathcal{L}\left(x_t\right) - \mathcal{L}\left(x^\star\right)\right)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \left\langle g_t, x_t - x^\star \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \left\langle g_t, x_t - w_t \right\rangle + \alpha_t \left\langle g_t, w_t - x^\star \right\rangle\right] \quad \text{(Add and subtract } w_t\text{)} \\
&\leq \mathbb{E}\left[R_T\left(x^\star\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \left\langle g_t, x_t - w_t \right\rangle\right] \quad \text{(Bound of the linearized regret)} \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} \langle g_t, x_{t-1} - x_t \rangle\right] + \mathbb{E}[R_T(x^*)] \quad \text{(Assumption on the alphas)} \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} \langle g_t, L(x_{t-1}) - L(x_t) \rangle\right] + \mathbb{E}[R_T(x^*)] \quad \text{(Convexity)}
\end{aligned}$$

Now we subtract both sides $\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t L(x_t)\right]$ to obtain:

$$\mathbb{E}\left[-\alpha_{1:T}\mathcal{L}\left(x^\star\right)\right] \quad \leq \mathbb{E}\left[R_T\left(x^\star\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1}\mathcal{L}\left(x_{t-1}\right) - \alpha_{1:t}\mathcal{L}\left(x_t\right)\right]$$

Now it is sufficient to apply the telescope series to obtain:

$$\mathbb{E}[L(x_T) - L(x^*)] \leq \mathbb{E}\left[\frac{R_T(x^*)}{\alpha_{1:T}}\right],$$

and conclude the first part of the proof. Furthermore, let's define $H_{t-1}$ as the history $g_{t-1}, x_{t-1}, \dots, g_1, x_1$ and define $G_t = \mathbb{E}[g_t|H_{t-1}, x_t, w_t]$ which is the expected value of the subgradient in relation to the past

history. Let's define $\epsilon_t = \alpha_t\langle G_t, w_t - x^*\rangle - \langle g_t, w_t - x^*\rangle$ since $g_t$ is independent of all other quantities given $x_t$, we have that $\mathbb{E}[\epsilon_t|H_{t-1}, x_t, w_t] = 0$. If we sum all over the t's we obtain:

$$\sum_{t=1}^{T}\epsilon_t = \sum_{t=1}^{T}\alpha_t\langle \underbrace{G_t}_{(\leq G)}, \underbrace{w_t - x^*}_{(\leq B)}\rangle - \sum_{t=1}^{T}\alpha_t\langle \underbrace{g_t}_{(\leq G)}, \underbrace{w_t - x^*}_{(\leq B)}\rangle,$$

where $|\epsilon_t| \leq 2\alpha_t BG$ with probability 1, and using the Azuma-Hoeffding bound with probability at least $1-\delta$ we get $\sum_{t=1}^{T}\epsilon_t \leq 2BG\sqrt{\sum_{t=1}^{T}\alpha_t^2 log(2/\delta)}$. This bound is useful to conclude our proof, the steps are similar to the previous part of the proof and with some simple math adjustments we obtain the final result of the theorem. The generalized case where we don't rely on the linearized regret is pretty similar with some prior math considerations, the final bound is exactly the same. □

## Acceleration setting

**Theorem 6.** *Suppose $\mathbb{E}[g_t] = \nabla L(x_t)$ for some L-smooth function L with domain on the entire Hilbert space H. Suppose $\|g_t\| \leq G$ with probability 1 and $g_t$ has variance at most $\sigma^2$ for all t. Suppose $\|x^*\| \leq B/2$. Let D be the ball of radius $B/2$ in H and suppose A guarantees retret:*

$$R_T(x^\star) \leq kB\sqrt{\sum_{t=1}^{T}\alpha_t \|g_t\|^2}$$

*For some k. Then with $c = \sqrt{2}k$, Algorithm 5 guarantees.*

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(y_T) - \mathcal{L}(x^\star)\right] \leq & \frac{2\sqrt{2}kB + 2kLB^2\log\left(1+G^2T^3\right)}{T^2} \\ & + \frac{2kB\sigma\sqrt{2\log\left(1+G^2T^3\right)}}{\sqrt{T}}\end{aligned}$$

*Proof.* The opening of the proof is very similar to the first proof of *A.Cutkosky* (Theorem 5), we again observe that:

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\left(\mathcal{L}(x_t) - \mathcal{L}(x^\star)\right)\right] \leq \mathbb{E}\left[R_T(x^\star) + \sum_{t=1}^{T}a_{1:t-1}\langle g_t, y_{t-1} - x_t\rangle\right]$$

We use convexity argument to argue $\mathbb{E}\left[\langle g_t, y_{t-1} - x_t\rangle\right] \leq \mathbb{E}\left[\mathcal{L}(y_{t-1}) - \mathcal{L}(x_t)\right]$, and then we subtract $\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\mathcal{L}(x_t)\right]$ from both sides:

$$\mathbb{E}\left[-\alpha_{1:T}\mathcal{L}(x^\star)\right] \leq \mathbb{E}\left[R_T(x^\star) + \sum_{t=1}^{T}\alpha_{1:t-1}\mathcal{L}(y_{t-1}) - \alpha_{1:t}\mathcal{L}(x_t)\right]$$

Now we use smoothness to relate $\mathcal{L}(y_t)$ to $\mathcal{L}(x_t)$. Defining $\zeta_t = g_t - \nabla\mathcal{L}(x_t)$ and $\beta_t = \alpha_{1:t}$, we have:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(y_t)\right] &\leq \mathbb{E}\left[\mathcal{L}(x_t) + \nabla\mathcal{L}(x_t)(y_t - x_t) + \frac{L}{2}\|x_t - y_t\|^2\right] \\ &= \mathbb{E}[L(x_t) - \eta_t\|g_t\|^2 + \eta_t\|g_t\|^2 - \eta_g g_t\nabla L(x_t) + \frac{L\eta_t^2\|x_t - y_t\|^2}{2}\Big] \\ &= \mathbb{E}\left[\mathcal{L}(x_t) - \eta_t\|g_t\|^2 + \eta_t\langle\zeta_t, g_t\rangle + \frac{L\eta_t^2\|g_t\|^2}{2}\right]\end{aligned}$$

Where we used the updating rule $x_t - y_t = \eta_t g_t$. Then multiply by $\beta_t$ both sides and open the learning rate to obtain:

$$\mathbb{E}\left[\beta_t\left(\mathcal{L}(y_t) - \mathcal{L}(x_t)\right)\right] \leq \mathbb{E}\left[-\frac{cB\beta_t\|g_t\|^2}{\sqrt{1+\sum_{i=1}^{t}\beta_i\|g_i\|^2}} + \frac{L\beta_t\eta_t^2\|g_t\|^2}{2} + \beta_t\eta_t\langle\zeta_t, g_t\rangle\right]$$

Next, we borrow Lemma.9: for positive numbers $x_1, \ldots, x_n$ and we observe from concavity of log that:

$$\sum_{i=1}^{n} \frac{x_i}{1+\sum_{i'=1}^{i} x_{i'}} \leq \log\left(1+\sum_{i=1}^{n} x_i\right)$$

Using this we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \beta_t \left(\mathcal{L}\left(y_t\right)-\mathcal{L}\left(x_t\right)\right)\right] \leq \mathbb{E}\left[-cB\sqrt{1+\sum_{t=1}^{T}\beta_t \left\|g_t\right\|^2}+\frac{c^2B^2L\log\left(1+G^2\beta_{1:T}\right)}{2}+cB+\sum_{t=1}^{T}\left\langle\zeta_t,\beta_t g_t\right\rangle\eta_t\right],$$

where the first term comes from Lemma.9 and the second from the concavity of *log*. Next, use Cauchy-Schwarz:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\zeta_t,\beta_t g_t\right\rangle\eta_t\right] \leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\beta_t\left\|\zeta_t\right\|^2}\sqrt{\sum_{t=1}^{T}\beta_t\left\|g_t\right\|^2\eta_t^2}\right] \leq \mathbb{E}\left[cB\sqrt{\sum_{t=1}^{T}\beta_t\|\zeta_t\|^2}\sqrt{\log\left(1+\sum_{t=1}^{t}\beta_t\left\|g_t\right\|^2\right)}\right]$$

where in the last inequality we used again the concavity of *log*. Then using Jensen's inequality:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\zeta_t,\beta_t g_t\right\rangle\eta_t\right] \leq \mathbb{E}\left[cB\sqrt{\sum_{t=1}^{T}\beta_t\|\zeta_t\|^2}\sqrt{log(1+G^2\beta_{1:T})}\right] \leq cB\sigma\sqrt{\beta_{1:T}\log\left(1+G^2\beta_{1:T}\right)}$$

Where in the last line we use the fact that the variance of the sugradient $\mathbb{E}\left[\left\|\zeta_t\right\|^2\right] = \mathbb{E}\left[\left\|g_t-\nabla L(x_t)\right\|^2\right] \leq \sigma^2$. Combining everything, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}-\alpha_t\mathcal{L}\left(x^\star\right)\right] \leq \mathbb{E}\left[R_T\left(x^\star\right)+\sum_{t=1}^{T}\alpha_{1:t-1}\mathcal{L}\left(y_{t-1}\right)-\alpha_{1:t}\mathcal{L}\left(y_t\right)\right]$$

$$+\mathbb{E}\left[\frac{c^2LB^2\log\left(1+G^2\beta_{1:T}\right)}{2}-cB\sqrt{1+\sum_{t=1}^{T}\alpha_{1:t}\left\|g_t\right\|^2} \quad -cB+cB\sigma\sqrt{\beta_{1:t}\log\left(1+G^2\beta_{1:t}\right)}\right]$$

Now observe that $\alpha_{1:t} > \alpha_t^2/2$ and recall $R_T\left(x^\star\right) \leq kB\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|g_t\right\|^2}$. Therefore since $c=\sqrt{2}k$ we cancel the $R_T\left(x^\star\right)$, observe $\beta_{1:T} \leq \sum_{t=1}^{T} t^2 \leq T^3$, where $\sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6}$ and telescope to obtain:

$$\mathbb{E}\left[\alpha_{1:T}\left(\mathcal{L}\left(y_T\right)-\mathcal{L}\left(x^\star\right)\right)\right] \leq cB+\frac{c^2B^2L\log\left(1+G^2T^3\right)}{2} + cBT^{3/2}\sigma\sqrt{\log\left(1+G^2T^3\right)}$$

and dividing by $\alpha_{1:T}=\frac{T(T+1)}{2}$ completes the proof. (Recall that $\alpha_{1:0}=0$ when telescoping the previous series. □

**Theorem 7.** *Suppose $\mathbb{E}[g_t] = L(x_t)$ for some convex function $L$. Then Algorithm 6 guarantees:*

$$\mathbb{E}\left[\left[\mathcal{L}\left(y_T\right)-\mathcal{L}\left(x^\star\right)\right] \leq \mathbb{E}\left[\frac{2R_T\left(x^\star\right)+B\sqrt{2\sum_{t=1}^{T}t^2\left\|\nabla\mathcal{L}\left(y_t\right)\right\|^2}\sqrt{\log\left(1+G^3T^3\right)}}{T^2}\right]\right.$$

*Note that in the setting with $\|g_t\| \leq G$ and $R_T\left(x^\star\right) = O\left(\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|g_t\right\|^2}\right)$, Theorem 5 implies a convergence rate of $O(\sqrt{\log(T)/T})$*

*Proof.* We start from $\mathbb{E}\left[-\alpha_{1:T}\mathcal{L}\left(x^\star\right)\right] \leq \mathbb{E}\left[R_T\left(x^\star\right)+\sum_{t=1}^{T}\alpha_{1:t-1}\mathcal{L}\left(y_{t-1}\right)-\alpha_{1:t}\mathcal{L}\left(x_t\right)\right]$, and again proceed to relate $\mathcal{L}\left(y_t\right)$ to $\mathcal{L}\left(x_t\right)$, this time without the aid of smoothness:

$$\mathbb{E}\left[\mathcal{L}\left(y_t\right)-\mathcal{L}\left(x_t\right)\right] \leq \mathbb{E}\left[\left\langle\nabla\mathcal{L}\left(y_t\right),y_t-x_t\right\rangle\right] \quad \leq \mathbb{E}\left[\left\|\nabla\mathcal{L}\left(y_t\right)\right\|\left\|g_t\right\|\eta_t\right]$$

So by Cauchy-Schwarz, again defining $\beta_t = \alpha_{1:t}$ we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\beta_t\left(\mathcal{L}\left(y_t\right)-\mathcal{L}\left(x_t\right)\right)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\beta_t\left\|\nabla\mathcal{L}\left(y_t\right)\right\|\left\|g_t\right\|\eta_t\right] \\ &\leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\beta_t\left\|\nabla\mathcal{L}\left(y_t\right)\right\|^2}\sqrt{\sum_{t=1}^{T}\beta_t\left\|g_t\right\|^2\eta_t^2}\right] \\ &\leq \mathbb{E}\left[B\sqrt{\sum_{t=1}^{T}\beta_t\left\|\nabla\mathcal{L}\left(y_t\right)\right\|^2\sqrt{\log\left(1+G^3T^3\right)}}\right]\end{aligned}$$

And combining everything yields

$$\mathbb{E}\left[-\alpha_{1:T}\mathcal{L}\left(x^\star\right)\right] \leq \mathbb{E}\Bigg[B\sqrt{\sum_{t=1}^{T}\beta_t\|\nabla L(x_t)\|^2}\sqrt{log(1+G^3T^3)}+\sum_{t=1}^{T}\alpha_{1:t-1}\mathcal{L}\left(y_{t-1}\right)-\alpha_{1:t}\mathcal{L}\left(y_t\right)\Bigg]$$

Telescope the sum and rearrange to prove the theorem. □

# Unixgrad

## Regret-to-rate conversion

In order to provide the mathematical analysis of the convergence rate of UnixGrad in the different optimization scenarios, classical tools in the online learning literature are taken into account an the convergence analysis is performed in the sense of bounding "weighted regret". Hence, a conversion strategy is adopted to directly translate the weighted regret to convergence rate.

**Lemma 8** (Regret-to-rate conversion)**.** *Consider weighted average*

$$\bar{x}_t = \frac{\alpha_t x_t + \sum_{i=1}^{t-1}\alpha_i x_i}{\sum_{i=1}^{t}\alpha_i}.$$

*Let* $R_T\left(x_*\right) = \sum_{t=1}^{T}\alpha_t\left\langle x_t - x_*, g_t\right\rangle$ *denote the weighted regret after* $T$ *iterations,* $\alpha_t = t$ *and* $g_t = \nabla f\left(\bar{x}_t\right)$*. Then,*

$$f\left(\bar{x}_T\right) - f\left(x_*\right) \leq \frac{2R_T\left(x_*\right)}{T^2}$$

*Proof.* Proof. Let's define $A_t = \sum_{i=1}^{t}\alpha_i$. Then, by definition, we could express $x_t$ as

$$x_t = \frac{A_t}{\alpha_t}\bar{x}_t - \frac{A_{t-1}}{\alpha_t}\bar{x}_{t-1} \tag{8}$$

Then, use Eq. (8) and replace $g_t$ by $\nabla f(\bar{x}_t)$ in the weighted regret expression, i.e.

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t \langle x_t - x_*, \nabla f(\bar{x}_t)\rangle &= \sum_{t=1}^{T} \alpha_t \left\langle \frac{A_t}{\alpha_t}\bar{x}_t - \frac{A_{t-1}}{\alpha_t}\bar{x}_{t-1} - x_*, \nabla f(\bar{x}_t)\right\rangle \\
&= \sum_{t=1}^{T} \alpha_t \left\langle \frac{A_t}{\alpha_t}(\bar{x}_t - x_*) - \frac{A_{t-1}}{\alpha_t}(\bar{x}_{t-1} - x_*), \nabla f(\bar{x}_t)\right\rangle \qquad \left(-x_* = \frac{A_{t-1}}{\alpha_t}x_* - \frac{A_t}{\alpha_t}x_*\right) \\
&= \sum_{t=1}^{T} A_t \langle \bar{x}_t - x_*, \nabla f(\bar{x}_t)\rangle - A_{t-1}\langle \bar{x}_{t-1} - x_*, \nabla f(\bar{x}_t)\rangle \\
&= \sum_{t=1}^{T} (\alpha_t + A_{t-1})\langle \bar{x}_t - x_*, \nabla f(\bar{x}_t\rangle) - A_{t-1}\langle \bar{x}_{t-1} - x_*, \nabla f(\bar{x}_t)\rangle \\
&= \sum_{t=1}^{T} \alpha_t \langle \bar{x}_t - x_*, \nabla f(\bar{x}_t)\rangle + A_{t-1}\langle \bar{x}_t - \bar{x}_{t-1}, \nabla f(\bar{x}_t)\rangle \\
&\geq \sum_{t=1}^{T} \alpha_t (f(\bar{x}_t) - f(x_*)) + \sum_{t=1}^{T}\sum_{i=1}^{t-1} \alpha_i (f(\bar{x}_t) - f(\bar{x}_{t-1})) \qquad \big(f(x) - f(y) \leq \nabla f(x)^\top (x-y)\big)
\end{aligned}$$

where in the last line we use $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ and multiply by $(-1)$. We also take $\alpha_0 = 0$ and $A_0 = 0$. Then, we telescope the double summation and reorganize the terms

$$\begin{aligned}
&= \sum_{t=1}^{T} \alpha_t (f(\bar{x}_t) - f(x_*)) + \sum_{t=1}^{T-1} \alpha_t (f(\bar{x}_T) - f(\bar{x}_t)) \\
&= \alpha_T (f(\bar{x}_T) - f(x_*)) + \sum_{t=1}^{T-1} \alpha_t (f(\bar{x}_t) - f(x_*) + f(\bar{x}_T) - f(\bar{x}_t)) \\
&= \sum_{t=1}^{T} \alpha_t (f(\bar{x}_T) - f(x_*))
\end{aligned}$$

Having simplified the expression, we divide both sides by $A_T$ and conclude the proof. Observe that $A_T \geq \frac{T^2}{2}$, hence,

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t (f(\bar{x}_T) - f(x_*)) &\leq \sum_{t=1}^{T} \alpha_t \langle x_t - x_*, \nabla f(\bar{x}_t)\rangle \\
\frac{1}{A_T}\sum_{t=1}^{T} \alpha_t (f(\bar{x}_T) - f(x_*)) &\leq \frac{1}{A_T}\sum_{t=1}^{T} \alpha_t \langle x_t - x_*, \nabla f(\bar{x}_t)\rangle \\
f(\bar{x}_T) - f(x_*) &\leq \frac{2R_T(x_*)}{T^2}
\end{aligned}$$

□

## Non-smooth Setting

Before starting with the mathematical analysis of the convergence rate of UnixGrad in the non-smooth setting, we recall some useful lemmas.

**Lemma 9.** *Let $\{a_i\}_{i=1,\dots,n}$ be a sequence of non negative numbers. Then, it holds that*

$$\sqrt{\sum_{i=1}^{n} a_i} \leq \sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{i} a_j} \leq 2\sqrt{\sum_{i=1}^{n} a_i}$$

*Proof.* Please refer to [4] for the proof. □

**Lemma 10.**

$$\alpha_t \|g_t - M_t\|_* \|x_t - y_t\| = \inf_{\rho>0} \left\{ \frac{\rho}{2} \|g_t - M_t\|_*^2 + \frac{\alpha_t^2}{2\rho} \|x_t - y_t\|^2 \right\} \tag{9}$$

*Proof.* Given two nonnegative real numbers $a \geq 0, b \geq 0$ and two real number $p > 0$ and $q > 0$ such that $1/p + 1/q = 1$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \qquad \text{(Young's inequality)}$$

By using it with $p, q = 2$ we have that

$$\alpha_t \|g_t - M_t\|_* \|x_t - y_t\| \leq \frac{1}{2} \|g_t - M_t\|_*^2 + \frac{\alpha_t^2}{2} \|x_t - y_t\|^2 .$$

Now, since the equality holds when $a^p = b^q$ we can use mathematical trick and formulate the statement as

$$\alpha_t \|g_t - M_t\|_* \|x_t - y_t\| = \inf_{\rho>0} \left\{ \frac{\rho}{2} \|g_t - M_t\|_*^2 + \frac{\alpha_t^2}{2\rho} \|x_t - y_t\|^2 \right\}.$$

□

**Deterministic setting**

**Theorem 11.** *Consider the constrained optimization setting, where $f : \mathcal{K} \to \mathbb{R}$ is a proper, convex and G-Lipschitz function defined over compact, convex set $\mathcal{K}$. Let $x^* \in \min_{x\in\mathcal{K}} f(x)$. Then, UnixGrad algorithm guarantees*

$$f(\bar{x}_T) - \min_{x\in\mathcal{K}} f(x) \leq \frac{7D\sqrt{1 + \sum_{t=1}^T \alpha_t^2 \|g_t - M_t\|_*^2} - D}{T^2} \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}$$

*Proof.*

$$\sum_{t=1}^T \alpha_t \langle x_t - x_*, g_t \rangle = \sum_{t=1}^T \underbrace{\alpha_t \langle x_t - y_t, g_t - M_t \rangle}_{\text{(A)}} + \underbrace{\alpha_t \langle x_t - y_t, M_t \rangle}_{\text{(B)}} + \underbrace{\alpha_t \langle y_t - x^*, g_t \rangle}_{\text{(C)}}$$

**Bounding (A)**

$$\begin{aligned}\sum_{t=1}^T \alpha_t \langle x_t - y_t, g_t - M_t \rangle &\leq \sum_{t=1}^T \alpha_t \|g_t - M_t\|_* \|x_t - y_t\| \qquad \text{(Holder's Inequality)} \\ &\leq \sum_{t=1}^T \frac{\rho}{2} \|g_t - M_t\|_*^2 + \frac{\alpha_t^2}{2\rho} \|x_t - y_t\|^2 \qquad \text{(Equation (9))}\end{aligned}$$

By setting $\rho = \alpha_t^2 \eta_{t+1}$ $(\alpha_t, \eta_{t+1} > 0)$, we get the following upper bound for term (A),

$$\sum_{t=1}^T \alpha_t \langle x_t - y_t, g_t - M_t \rangle \leq \sum_{t=1}^T \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \frac{1}{2\eta_{t+1}} \|x_t - y_t\|^2$$

**Bounding (B)**

In order to find an upper-bound for this term, we have to exploit the first order optimality condition of $x_t$, that is $\nabla f(x_t)^\top (y - x_t) \geq 0$, $\forall y$ feasible (all feasible directions from $x$ are aligned with increasing gradient). Since $\nabla f(x_t) = \alpha_t M_t + \frac{1}{\eta_t} \nabla_x \mathcal{D}_\mathcal{R}(x_t, y_{t-1})$, then

$$\alpha_t \langle y_t - x_t, M_t \rangle + \frac{1}{\eta_t} \nabla_x \mathcal{D}_\mathcal{R}(x_t, y_{t-1})^\top (y_t - x_t) \geq 0.$$

By simply rearranging the above inequality we get

$$\frac{1}{\eta_t} \nabla_x \mathcal{D}_\mathcal{R}(x_t, y_{t-1})^\top (y_t - x_t) \geq \alpha_t \langle x_t - y_t, M_t \rangle .$$

Now the upper-bound becomes straightforward.

$$\sum_{t=1}^{T} \alpha_t \langle x_t - y_t, M_t \rangle \leq \sum_{t=1}^{T} \frac{1}{\eta_t} \nabla_x D_{\mathcal{R}} (x_t, y_{t-1})^\top (y_t - x_t) \quad \text{(Optimality for } x_t\text{)}$$

Now considering the fact that $\nabla_x \mathcal{D}_{\mathcal{R}}(x, y) = \nabla \mathcal{R}(x) - \nabla \mathcal{R}(y)$ we can rewrite the right-hand side as follows

$$\sum_{t=1}^{T} \alpha_t \langle x_t - y_t, M_t \rangle \leq \sum_{t=1}^{T} \frac{1}{\eta_t} \left( D_{\mathcal{R}} (y_t, y_{t-1}) - D_{\mathcal{R}} (x_t, y_{t-1}) - D_{\mathcal{R}} (y_t, x_t) \right).$$

**Bounding (C)**
Following the same procedure used in the previous bound we find the upper-bound also for the third term.

$$\begin{aligned}\sum_{t=1}^{T} \alpha_t \langle y_t - x^*, g_t \rangle &\leq \sum_{t=1}^{T} \frac{1}{\eta_t} \nabla_x D_{\mathcal{R}} (y_t, y_{t-1})^T (x^* - y_t) \quad \text{(Optimality for } y_t\text{)} \\ &= \sum_{t=1}^{T} \frac{1}{\eta_t} \left( D_{\mathcal{R}} (x^*, y_{t-1}) - D_{\mathcal{R}} (y_t, y_{t-1}) - D_{\mathcal{R}} (x^*, y_t) \right)\end{aligned}$$

**Final Bound**

$$\begin{aligned}\sum_{t=1}^{T} \alpha_t \langle x_t - x_*, g_t \rangle \leq & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \frac{1}{2\eta_{t+1}} \|x_t - y_t\|^2 \\ & + \frac{1}{\eta_t} \left( D_{\mathcal{R}} (x^*, y_{t-1}) - D_{\mathcal{R}} (x^*, y_t) - D_{\mathcal{R}} (x_t, y_{t-1}) - D_{\mathcal{R}} (y_t, x_t) \right)\end{aligned}$$

Then using the fact that $D_{\mathcal{R}}(x, y) \geq \frac{1}{2}\|x - y\|^2$ for all $x, y \in \mathcal{K}$ due to the strong convexity of $\mathcal{R}$, we have

$$\begin{aligned}\leq & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \frac{1}{2\eta_{t+1}} \|x_t - y_t\|^2 \\ & + \frac{1}{\eta_t} \left( D_{\mathcal{R}} (x^*, y_{t-1}) - D_{\mathcal{R}} (x^*, y_t) - \frac{1}{2} \left( \|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) \right).\end{aligned}$$

Now, let's consider that $-1/2 \|x_t - y_{t-1}\|^2$ is a negative term and call $\mathcal{D}_{\mathcal{R}}(x^*, y_t) = \delta_t$. We obtain

$$\begin{aligned}\leq & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \sum_{t=1}^{T} \frac{1}{\eta_t} (\delta_{t-1} - \delta_t) + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|x_t - y_t\|^2 \\ = & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \frac{D_{\mathcal{R}(x, y_0)}}{\eta_1} + \sum_{t=1}^{T-1} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \delta_t - \frac{\delta_T}{\eta_T} + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|x_t - y_t\|^2\end{aligned}$$

Due to positiveness of $\delta_T$ and $\eta_T$ and the definition of $D^2 = \sup_{x,y \in \mathcal{K}} \mathcal{D}_{\mathcal{R}}(x, y)$ we can build an upper-bound

$$\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + D^2 \sum_{t=1}^{T-1} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|x_t - y_t\|^2 + \frac{1}{\eta_1} D^2$$

At this point we telescope the sum with $T - 1$ terms, obtaining

$$\begin{aligned}= & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + D^2 \left( \frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|x_t - y_t\|^2 + \frac{1}{\eta_1} D^2 \\ = & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|x_t - y_t\|^2 + \frac{D^2}{\eta_T} \\ \leq & \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|x_t - y_t\|^2 + \frac{D^2}{\eta_T} + \frac{D}{2}\end{aligned}$$

Re-using $\|x_t - y_t\|^2 \leq 2\mathcal{D}_{\mathcal{R}}(x_t, y_t) \leq 2D^2$ and telescope the sum we have

$$\begin{aligned}
&\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + \frac{2D^2}{\eta_{T+1}} + \frac{D^2}{\eta_T} + \frac{D}{2} \\
&= \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_{t+1}}{2} \|g_t - M_t\|_*^2 + D^2 \left( \frac{2}{\eta_{T+1}} + \frac{1}{\eta_T} \right) + \frac{D}{2}
\end{aligned}$$

Using the definition of $\eta_t = \frac{2D}{\sqrt{1+\sum_{i=1}^{t-1} \alpha_i^2 \|g_i - M_i\|_*^2}}$ and $\frac{1}{\eta_T} \leq \frac{1}{\eta_{T+1}}$

$$\begin{aligned}
&\leq D \sum_{t=1}^{T} \frac{\alpha_t^2 \|g_t - M_t\|_*^2}{\sqrt{1 + \sum_{i=1}^{t} \alpha_i^2 \|g_i - M_i\|_*^2}} + \frac{3}{2} D \sqrt{1 + \sum_{t=1}^{T} \alpha_t^2 \|g_t - M_t\|_*^2} + \frac{D}{2} \\
&\leq \frac{7}{2} D \sqrt{1 + \sum_{t=1}^{T} \alpha_t^2 \|g_t - M_t\|_*^2} - \frac{D}{2} \\
&\leq 3D + 7GD \sqrt{\sum_{t=1}^{T} \alpha_t^2} \\
&\leq 3D + 7GDT^{3/2}
\end{aligned}$$

We obtain the rate by applying Lemma 8 to the weighted regret bound above. □

**Stochastic setting**

**Theorem 12.** *Consider the optimization setting where $f$ is non-smooth, convex and G-Lipschitz. Let $\{x_t\}_{t=1,..,T}$ be a sequence generated by UnixGrad such that $g_t = \tilde{\nabla} f(\bar{x}_t)$ and $M_t = \tilde{\nabla} f(\tilde{z}_t)$. With $\alpha_t = t$ and the lag-behind-one learning rate, it holds that*

$$\mathbb{E}\left[f\left(\bar{x}_T\right)\right] - \min_{x \in \mathcal{K}} f(x) \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}$$

*Proof.* Similar to $\nabla f(x) \leftrightarrow \tilde{\nabla} f(x)$ notation, $\tilde{g}_t$ denotes a stochastic but unbiased estimate of $g_t$ for any $t \in [0, .., T]$. Also note that $x^* \in \min_{x \in \mathcal{K}} f(x)$. We start with weighted regret bound,

$$R_T\left(x_*\right) = \sum_{t=1}^{T} \alpha_t \left\langle x_t - x^*, g_t \right\rangle$$

We separate $g_t$ as $\tilde{g}_t + (g_t - \tilde{g}_t)$ and re-write the above term as

$$\sum_{t=1}^{T} \alpha_t \left\langle x_t - x^*, g_t \right\rangle = \underbrace{\sum_{t=1}^{T} \alpha_t \left\langle x_t - x^*, \tilde{g}_t \right\rangle}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T} \alpha_t \left\langle x_t - x^*, g_t - \tilde{g}_t \right\rangle}_{\text{(B)}}.$$

Due to unbiasedness of the gradient estimates, expected value of $\alpha_t \left\langle x_t - x^*, g_t - \tilde{g}_t \right\rangle$, conditioned on the average iterate $\bar{x}_t$ evaluates to 0. We will only need to bound the first summation whose analysis is identical to its deterministic counterpart up to replacing $g_t$ with $\tilde{g}_t$, and $M_t$ with $\tilde{M}_t$. Hence, term (A) is upper bounded by $6D + 14GDT^{3/2}$.
In addition to the setup in the deterministic setting, we put forth the assumption that stochastic gradients have bounded norms, which is natural in the constrained optimization framework. Using Lemma 8 , we translate the regret bound into the convergence rate, i.e,

$$\mathbb{E}\left[f\left(\bar{x}_T\right)\right] - \min_{x \in \mathcal{K}} f(x) \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}$$

□

## Smooth Setting

We will now introduce an additional assumption that f is L-smooth. In this section, we provide the weighted regret analysis for smooth functions in the presence of deterministic and stochastic oracles and convert these bound into suboptimality gap via our regret-to-rate scheme (Lemma 8).

### Deterministic setting

**Theorem 13.** *Consider the constrained optimization setting where $f : \mathcal{K} \to \mathbb{R}$ is a proper, convex and L -smooth function defined over compact, convex set $\mathcal{K}$. Let $x^* \in \min_{x\in\mathcal{K}} f(x)$. Then, Algorithm 2 ensures the following*

$$f\left(\bar{x}_T\right) - \min_{x\in\mathcal{K}} f(x) \leq \frac{20\sqrt{7}D^2L}{T^2}$$

*Proof.* Recall the regret analysis for the non-smooth, convex objective.

$$\begin{aligned}
R_T\left(x_*\right) \leq & \frac{1}{2}\sum_{t=1}^{T}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2 + \frac{1}{\eta_{t+1}}\left\|x_t - y_t\right\|^2 \\
& + \sum_{t=1}^{T}\frac{1}{\eta_t}\left(D_{\mathcal{R}}\left(x^*, y_{t-1}\right) - D_{\mathcal{R}}\left(x^*, y_t\right) - \frac{1}{2}\left(\left\|x_t - y_t\right\|^2 + \left\|x_t - y_{t-1}\right\|^2\right)\right) \\
\leq & \frac{1}{2}\sum_{t=1}^{T}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\left\|x_t - y_t\right\|^2 + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)D_{\mathcal{R}}\left(x^*, y_t\right) \\
& - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\left\|x_t - y_{t-1}\right\|^2 + \frac{D^2}{\eta_1}
\end{aligned}$$

By adding and subtract the term $\frac{1}{\eta_{t+1}}\left\|x_t - y_{t-1}\right\|^2$ we get

$$\begin{aligned}
= & \frac{1}{2}\sum_{t=1}^{T}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\left\|x_t - y_t\right\|^2 + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\left\|x_t - y_{t-1}\right\|^2 \\
& + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)D_{\mathcal{R}}\left(x^*, y_t\right) - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\left\|x_t - y_{t-1}\right\|^2 + \frac{D^2}{\eta_1} \\
& \leq \frac{1}{2}\sum_{t=1}^{T}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\left\|x_t - y_{t-1}\right\|^2 + D^2\left(\frac{2}{\eta_{T+1}} + \frac{1}{\eta_T} + \frac{1}{\eta_1}\right)
\end{aligned}$$

The key challenge in this analysis is to exploit the negative term, i.e., $-\frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\left\|x_t - y_{t-1}\right\|^2$, such that we could tighten the regret bound from non-smooth analysis. Using the smoothness of $f$ and that $\alpha_t = t, A_t = \sum_{i=1}^{t}\alpha_t, g_t = \nabla f\left(\bar{x}_t\right)$ and $M_t = \nabla f\left(\tilde{z}_t\right)$ we obtain

$$\begin{aligned}
\left\|g_t - M_t\right\|_*^2 & \leq L^2\left\|\bar{x}_t - \bar{y}_{t-1}\right\|^2 \\
& = L^2\left\|\frac{\alpha_t x_t + \sum_{i=1}^{t-1}\alpha_i x_i}{A_t} - \frac{\alpha_t y_{t-1} + \sum_{i=1}^{t-1}\alpha_i x_i}{A_t}\right\|^2 \\
& = \frac{L^2\alpha_t^2}{A_t^2}\left\|x_t - y_{t-1}\right\|^2 \\
& = \frac{4L^2t^2}{t^2(t+1)^2}\left\|x_t - y_{t-1}\right\|^2 \quad \text{(sum of first } t \text{ number and } \alpha_t = t\text{)} \\
& = \frac{4L^2}{\alpha_{t+1}^2}\left\|x_t - y_{t-1}\right\|^2 \\
& \leq \frac{4L^2}{\alpha_t^2}\left\|x_t - y_{t-1}\right\|^2 \quad (\alpha_t \text{ increasing in } t)
\end{aligned}$$

Hence,

$$-\frac{1}{\eta_{t+1}}\left\|x_t - y_{t-1}\right\|^2 \leq -\frac{\alpha_t^2}{4L^2\eta_{t+1}}\left\|g_t - M_t\right\|_*^2$$

After applying this upper bound and regrouping the terms we have

$$R_T\left(x_*\right) \leq \frac{1}{2}\sum_{t=1}^{T}\left(\eta_{t+1} - \frac{1}{4L^2\eta_{t+1}}\right)\alpha_t^2\left\|g_t - M_t\right\|_*^2 + D^2\left(\frac{2}{\eta_{T+1}} + \frac{1}{\eta_T} + \frac{1}{\eta_1}\right)$$

Define that $\tau^* = \max\left\{t \in \{1,\ldots,T\} : \frac{1}{\eta_{t+1}^2} \leq 7L^2\right\}$ such that $\forall t > \tau^*, \eta_{t+1} - \frac{1}{4L^2\eta_{t+1}} \leq -\frac{3}{4}\eta_{t+1}$ We can rewrite the above term as

$$\begin{aligned}
R_T\left(x_*\right) \leq & \frac{1}{2}\left(\sum_{t=1}^{\tau^*}\left(\eta_{t+1} - \frac{1}{4L^2\eta_{t+1}}\right)\alpha_t^2\left\|g_t - M_t\right\|_*^2 + \sum_{t=\tau^*+1}^{T}\left(\eta_{t+1} - \frac{1}{4L^2\eta_{t+1}}\right)\alpha_t^2\left\|g_t - M_t\right\|_*^2\right) \\
& + \frac{3D^2}{\eta_{T+1}} + \frac{D^2}{\eta_1} \\
\leq & \underbrace{\frac{1}{2}\sum_{t=1}^{\tau^*}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2 + \frac{D}{2}}_{\text{(A)}} + \underbrace{\frac{3D^2}{\eta_{T+1}} - \frac{3}{4}\sum_{t=\tau^*+1}^{T}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2}_{\text{(B)}}
\end{aligned}$$

**Bounding (A)**: We will simply need to use the definition of $\tau^*$ and Lemma 8

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{\tau^*}\eta_{t+1}\alpha_t^2\left\|g_t - M_t\right\|_*^2 + \frac{D}{2} &= D\sum_{t=1}^{\tau^*}\frac{\alpha_t^2\left\|g_t - M_t\right\|_*^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2\left\|g_i - M_i\right\|_*^2}} + \frac{D}{2} \\
&\leq 2D\sqrt{1 + \sum_{t=1}^{\tau^*}\alpha_t^2\left\|g_t - M_t\right\|_*^2} \\
&= \frac{4D^2}{\eta_{\tau^*+1}} \\
&\leq 4\sqrt{7}D^2L
\end{aligned}$$

**Bounding (B)**: using the definition of $\eta_t$ and Lemma 9

$$\begin{aligned}
\text{(B)} \leq & \frac{3D}{2}\left(\sqrt{1 + \sum_{t=1}^{T}\alpha_t^2\left\|g_t - M_t\right\|_*^2} - \sum_{t=\tau^*+1}^{T}\frac{\alpha_t^2\left\|g_t - M_t\right\|_*^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2\left\|g_i - M_i\right\|_*^2}}\right) \quad \text{(Lemma 9)} \\
\leq & \frac{3D}{2} + \frac{3D}{2}\left(\sum_{t=1}^{T}\frac{\alpha_t^2\left\|g_t - M_t\right\|_*^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2\left\|g_i - M_i\right\|_*^2}} - \sum_{t=\tau^*+1}^{T}\frac{\alpha_t^2\left\|g_t - M_t\right\|_*^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2\left\|g_i - M_i\right\|_*^2}}\right) \\
= & \frac{3D}{2} + \frac{3D}{2}\sum_{t=1}^{\tau^*}\frac{\alpha_t^2\left\|g_t - M_t\right\|_*^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2\left\|g_i - M_i\right\|_*^2}} \quad \text{(Lemma 9)} \\
\leq & 3D\sqrt{1 + \sum_{t=1}^{\tau^*}\alpha_i^2\left\|g_i - M_i\right\|_*^2} \quad \text{(definition of } \eta_t\text{)} \\
= & \frac{6D^2}{\eta_{\tau^*+1}} \quad \text{(definition of } \tau^*\text{)} \\
\leq & 6\sqrt{7}D^2L
\end{aligned}$$

**Final Bound**:

$$R_T(x_*) \leq \frac{1}{2}\sum_{t=1}^{\tau^*}\eta_{t+1}\alpha_t^2\|g_t - M_t\|_*^2 + \frac{D}{2} + \frac{3D^2}{\eta_{T+1}} - \frac{3}{4}\sum_{t=\tau^*+1}^{T}\eta_{t+1}\alpha_t^2\|g_t - M_t\|_*^2$$
$$\leq 10\sqrt{7}D^2L$$

We conclude the proof by applying Lemma 1 and get $f(\bar{x}_T) - \min_{x\in\mathcal{K}} f(x) \leq \frac{20\sqrt{7}D^2L}{T^2}$.

□

**Stochastic setting**

In this setting, we will make an additional, but classical, bounded variance assumption on the stochastic gradient oracles that is

$$E\left[\|\nabla f(x) - \tilde{\nabla} f(x)\|_*^2 \mid x\right] \leq \sigma^2, \quad \forall x \in \mathcal{K}.$$

Let us define $\xi_t = \left(\tilde{g}_t - \tilde{M}_t\right) - (g_t - M_t)$. Since $\|\xi_t\|_*^2 \leq 2\|\tilde{g}_t - g_t\|_*^2 + 2\left\|\tilde{M}_t - M_t\right\|_*^2$, we can write,

$$\mathbb{E}\left[\|\xi_t\|_*^2 \mid \bar{x}_t\right] \leq 4\sigma^2$$

Next, we will present the final convergence theorem.

**Theorem 14.** *Consider the optimization setting where $f$ is $L$-smooth and convex. Let $\{x_t\}_{t=1,\dots,T}$ be a sequence generated by UnixGrad such that $g_t = \tilde{\nabla} f(\bar{x}_t)$ and $M_t = \tilde{\nabla} f(\tilde{z}_t)$ With $\alpha_t = t$ and the lag-behind-one learning rate, it holds that*

$$\mathbb{E}[f(\bar{x}_T)] - \min_{x\in\mathcal{K}} f(x) \leq \frac{112\sqrt{14}D^2L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}}$$

*Proof.* We start out with weighted regret, the same way as before

$$\sum_{t=1}^{T}\alpha_t\langle x_t - x^*, g_t\rangle = \underbrace{\sum_{t=1}^{T}\alpha_t\langle x_t - x^*, \tilde{g}_t\rangle}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T}\alpha_t\langle x_t - x^*, g_t - \tilde{g}_t\rangle}_{\text{(B)}}.$$

We already know that term (B) is zero in expectation. Following the proof steps of the deterministic smooth case and considering that $\frac{1}{\eta_T} \leq \frac{1}{\eta_{T+1}}$ we could upper bound term (A) as

$$\leq \frac{1}{2}\sum_{t=1}^{T}\eta_{t+1}\alpha_t^2\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\|x_t - y_{t-1}\|^2 + D^2\left(\frac{3}{\eta_{T+1}} + \frac{1}{\eta_1}\right)$$
$$= D\sum_{t=1}^{T}\frac{\alpha_t^2\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2}} - \frac{1}{2}\sum_{t=1}^{T}\frac{\|x_t - y_{t-1}\|^2}{\eta_{t+1}} + \frac{3D}{2}\sqrt{1 + \sum_{t=1}^{T}\alpha_t^2\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2} + \frac{D}{2}$$
$$\leq \frac{7D}{2}\sqrt{1 + \sum_{t=1}^{T}\alpha_t^2\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\|x_t - y_{t-1}\|^2 \quad \text{(Lemma 9)}$$

Now let's denote,

$$B_t^2 := \min\left\{\|g_t - M_t\|_*^2, \left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2\right\}$$

as well as an auxiliary learning rate which we will only use for the analysis

$$\tilde{\eta}_t = \frac{2D}{\sqrt{1 + \sum_{i=1}^{t-1}\alpha_i^2 B_i^2}} \tag{10}$$

Clearly, for any $t \in [T]$ we have $1/\tilde{\eta}_t \leq 1/\eta_t$, and therefore,

$$-\frac{1}{\eta_{t+1}}\left\|g_t - M_t\right\|_*^2 \leq -\frac{1}{\tilde{\eta}_{t+1}}B_t^2 \tag{11}$$

Also, for $\xi_t = \left(\tilde{g}_t - \tilde{M}_t\right) - (g_t - M_t)$, we can write,

$$\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 \leq 2\left\|g_t - M_t\right\|_*^2 + 2\left\|\xi_t\right\|_*^2 \tag{12}$$

Thus, by adding and subtracting $B_t^2$

$$\begin{aligned}
\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 &= B_t^2 + \left(\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 - \min\left\{\left\|g_t - M_t\right\|_*^2, \left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2\right\}\right) \\
&= B_t^2 + \max\left\{0, \left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 - \left\|g_t - M_t\right\|_*^2\right\} \\
&\leq B_t^2 + B_t^2 + 2\left\|\xi_t\right\|_*^2 \\
&= 2B_t^2 + 2\left\|\xi_t\right\|_*^2
\end{aligned}$$

where the last inequality is due to the fact that if $\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 \geq \left\|g_t - M_t\right\|_*^2$, then $B_t^2 := \left\|g_t - M_t\right\|_*^2$. Then, we combine this with Eq. (12) to deduce that $\left\|\tilde{g}_t - \tilde{M}_t\right\|_*^2 - \left\|g_t - M_t\right\|_*^2 \leq B_t^2 + 2\left\|\xi_t\right\|_*^2$. We will take conditional expectation after we simplify the expression. Now, we plug Eq. (11) and (12) into above bound,

$$\begin{aligned}
&\leq \frac{7D}{2}\sqrt{1 + 2\sum_{t=1}^{T}\alpha_t^2 B_t^2 + \alpha_t^2\left\|\xi_t\right\|_*^2} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\tilde{\eta}_{t+1}}\alpha_t^2 B_t^2 \quad (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}) \\
&\leq \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|\xi_t\right\|_*^2} + \frac{7D}{2}\sqrt{1 + 2\sum_{t=1}^{T}\alpha_t^2 B_t^2} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\tilde{\eta}_{t+1}}\alpha_t^2 B_t^2 \quad \text{(Lemma 9)} \\
&\leq \frac{7D}{2} + \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|\xi_t\right\|_*^2} + 7D\sum_{t=1}^{T}\frac{\alpha_t^2 B_t^2}{\sqrt{1 + 2\sum_{i=1}^{t}\alpha_i^2 B_i^2}} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\tilde{\eta}_{t+1}}\alpha_t^2 B_t^2 \\
&\leq \frac{7D}{2} + \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|\xi_t\right\|_*^2} + 7D\sum_{t=1}^{T}\frac{\alpha_t^2 B_t^2}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2 B_i^2}} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\tilde{\eta}_{t+1}}\alpha_t^2 B_t^2 \\
&= \sum_{t=1}^{T}\left(\frac{7}{2}\frac{2D}{\sqrt{1 + \sum_{i=1}^{t}\alpha_i^2 B_i^2}} - \frac{1}{8L^2\tilde{\eta}_{t+i}}\right)\alpha_t^2 B_t^2 + \frac{7D}{2} + \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|\xi_t\right\|_*^2} \\
&= \underbrace{\frac{7}{2}\sum_{t=1}^{T}\left(\tilde{\eta}_{t+1} - \frac{1}{28L^2\tilde{\eta}_{t+1}}\right)\alpha_t^2 B_t^2 + \frac{7D}{2}}_{\text{(A)}} + \underbrace{\frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\left\|\xi_t\right\|_*^2}}_{\text{(B)}}
\end{aligned}$$

**Bounding (A)**: We will make use of the exact same approach as we did in the smooth deterministic case, where we defined an auxiliary time variable $\tau^*$ to characterize the behavior of the learning rate. Now, let us denote $\tau^* = \max\left\{t \in \{1, \ldots, T\} : \frac{1}{\tilde{\eta}_{t+1}^2} \leq 56L^2\right\}$. It implies that

$$\tilde{\eta}_{t+1} - \frac{1}{28L^2\tilde{\eta}_{t+1}} \leq -\tilde{\eta}_{t+1}, \quad \forall t > \tau^*$$

Then, we could proceed as

$$\begin{aligned}
(\mathrm{A}) &= \frac{7}{2}\sum_{t=1}^{\tau^*}\left(\tilde{\eta}_{t+1} - \frac{1}{28L^2\tilde{\eta}_{t+1}}\right)\alpha_t^2 B_t^2 + \frac{7}{2}\sum_{t=\tau^*+1}^{T}\left(\tilde{\eta}_{t+1} - \frac{1}{28L^2\tilde{\eta}_{t+1}}\right)\alpha_t^2 B_t^2 + \frac{7D}{2} \\
&\leq \frac{7}{2}\sum_{t=1}^{\tau^*}\tilde{\eta}_{t+1}\alpha_t^2 B_t^2 - \frac{7}{2}\sum_{t=\tau^*+1}^{T}\tilde{\eta}_{t+1}\alpha_t^2 B_t^2 + \frac{7D}{2} \\
&\leq \frac{7}{2}\sum_{t=1}^{\tau^*}\tilde{\eta}_{t+1}\alpha_t^2 B_t^2 + \frac{7D}{2} \\
&= 7D\sum_{t=1}^{\tau^*}\frac{\alpha_t^2 B_t^2}{\sqrt{1+\sum_{i=1}^{t}\alpha_i^2 B_i^2}} + \frac{7D}{2} \\
&\leq 14D\sqrt{1+\sum_{t=1}^{\tau^*}\alpha_t^2 B_t^2} \\
&\leq \frac{28D^2}{\tilde{\eta}_{\tau^*+1}} \quad \text{(definition of } \tau^*\text{)} \\
&\leq 56\sqrt{14}D^2 L
\end{aligned}$$

**Bounding (B)**: Following bounded variance definition, we can write $\mathbb{E}\left[\|\xi_t\|_*^2\right] \leq 4\sigma^2$. After taking expected value conditioned on $\bar{x}_t$, we simply use Jensen's inequality to complete the proof

$$\begin{aligned}
\mathbb{E}\left[\frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\xi_t\|_*^2} \mid \bar{x}_t\right] &\leq \frac{7D}{\sqrt{2}}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\xi_t\|_*^2 \mid \bar{x}_t\right]} \\
&= \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\mathbb{E}\left[\|\xi_t\|_*^2 \mid \bar{x}_t\right]} \\
&\leq \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}4\alpha_t^2\sigma^2} \\
&\leq \frac{14D\sigma}{\sqrt{2}}\sqrt{T^3} \\
&= \frac{14\sigma D T^{3/2}}{\sqrt{2}}
\end{aligned}$$

Finally, we combine all these bounds together and feed them through Lemma 8 to obtain the final rate.

$$\mathbb{E}\left[f\left(\bar{x}_T\right)\right] - \min_{x\in\mathcal{K}} f(x) \leq \frac{112\sqrt{14}D^2 L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}}$$

□

# Accelegrad

## Mirror Descent intuition, in depth analysis for linear coupling

Here we provide detailed analysis to understand why Mirror descent needs small gradients to converge faster. Recovering the previous initial analysis from Mirror-Prox template, the mirror step is defined as $x_{k+1} = \mathrm{Mirr}_x(\epsilon) = \mathrm{argmin}_{y\in Q}\{V_x(y) + \langle\epsilon, y - x\rangle\}$. In the following we provide a mirror descent guarantee given by choosing those step size.

**Theorem 15.** *If we choose as a mirror step $Mirr_{x_k}(\alpha\delta f(x_k))$ then we get:*

$$\forall u \in Q, \quad \alpha\left(f\left(x_k\right) - f(u)\right) \leq \alpha\left\langle\partial f\left(x_k\right), x_k - u\right\rangle \leq \frac{\alpha^2}{2}\left\|\partial f\left(x_k\right)\right\|^2 + V_{x_k}(u) - V_{x_{k+1}}(u),$$

*Proof.* We start by adding and subtracting $x_{t+1}$ from the first term:

$$\begin{aligned}
\alpha \left\langle \partial f\left(x_k\right), x_k - u\right\rangle &= \left\langle \alpha \partial f\left(x_k\right), x_k - x_{k+1}\right\rangle + \left\langle \alpha \partial f\left(x_k\right), x_{k+1} - u\right\rangle \\
&\leq \left\langle \alpha \partial f\left(x_k\right), x_k - x_{k+1}\right\rangle + \left\langle -\nabla V_{x_k}\left(x_{k+1}\right), x_{k+1} - u\right\rangle \\
&= \left\langle \alpha \partial f\left(x_k\right), x_k - x_{k+1}\right\rangle + V_{x_k}(u) - V_{x_{k+1}}(u) - V_{x_k}\left(x_{k+1}\right) \\
&\leq \left(\left\langle \alpha \partial f\left(x_k\right), x_k - x_{k+1}\right\rangle - \frac{1}{2}\left\|x_k - x_{k+1}\right\|^2\right) + \left(V_{x_k}(u) - V_{x_{k+1}}(u)\right) \\
&\leq \frac{\alpha^2}{2}\left\|\partial f\left(x_k\right)\right\|^2 + \left(V_{x_k}(u) - V_{x_{k+1}}(u)\right)
\end{aligned}$$

In the second line we used the minimality of the mirror step $x_{k+1} = \operatorname{argmin}_{x \in Q}\{V_{x_k}(x) + \langle \alpha \delta f(x_k), x\rangle\}$ which is equal to $\langle \nabla V_{x_k}(x_{k+1} + \alpha \delta f(x_k), u - x_{k+1}\rangle \geq 0 \; \forall u \in Q$, this is the first order optimality condition. In the third line we used the triangle equality of the Bregman Divergence given as:

$$\begin{aligned}
\forall x, y \geq 0, \quad \langle -\nabla V_x(y), y - u\rangle &= \langle \nabla w(x) - \nabla w(y), y - u\rangle \\
&= (w(u) - w(x) - \langle \nabla w(x), u - x\rangle) - (w(u) - w(y) - \langle \nabla w(y), u - y\rangle) \\
&\quad - (w(y) - w(x) - \langle \nabla w(x), y - x\rangle) \\
&= V_x(u) - V_y(u) - V_x(y)
\end{aligned}$$

In the third line we use the $1-$strong convexity property of the distance generating function $V_x(y)$ and in the final line we use Cauchy-Schwartz inequality $|\langle \alpha \partial f\left(x_k\right), x_k - x_{k+1}\rangle| \leq \|\alpha \delta f(x_k)\| \|x_k - x_{k+1}\|$ to obtain the final bound. □

We can now telescope the result of the theorem from $k = 0, \ldots, T-1$, set $\hat{x} = \sum_{k=0}^{T-1} x_k$ and choosing $u = x^*$ to obtain:

$$\alpha T\left(f(\bar{x}) - f\left(x^*\right)\right) \leq \sum_{k=0}^{T-1} \alpha \left\langle \partial f\left(x_k\right), x_k - x^*\right\rangle \leq \frac{\alpha^2}{2}\sum_{k=0}^{T-1}\left\|\partial f\left(x_k\right)\right\|^2 + V_{x_0}\left(x^*\right) - V_{x_T}\left(x^*\right)$$

From this bound on the regret we can deduce the motivation behind the idea of faster convergence with small gradient magnitudes. Now, choosing the step size $\alpha = \frac{2\Theta}{\rho\sqrt{T}}$, where $\Theta$ is any upper bound on $V_{x_0}(x^*)$ we can obtain the convergence rate of mirror descent:

$$f(\bar{x}) - f\left(x^*\right) \leq \frac{\sqrt{2\Theta} \cdot \rho}{\sqrt{T}} \text{ or equivalently } T \geq \frac{2\Theta \cdot \rho^2}{\varepsilon^2} \Rightarrow f(\bar{x}) - f\left(x^*\right) \leq \varepsilon,$$

where $\rho^2$ is the average magnitude of the queries gradients, which give us the intuition behind mirror descent convergence speed up.

## Smooth setting

**Theorem 16.** *Assume that $f$ is convex and $\beta$-smooth. Let $K$ be a convex set with bounded diameter $D$, and assume there exists a global minimizer for $f$ in $K$. Then Algorithm 3 with weights equal to Equation 7 and learning rate as in Equation 6 ensures:*

$$f(\hat{y}_t) - f(z) \leq \mathcal{O}\left(\frac{DG + \beta D^2 log(\beta D/G)}{T^2}\right),$$

*where $z$ is the minimum of our objective function $f$.*

Since the proof of this Theorem is very long, we provide the most meaningful parts which guides us into deriving the upper bound on the error.

*Proof.* Our objective is to find a bound of the left-side in order to precisely state the convergence rate of Accelegrad in the smooth case. Given the fact that the algorithm outputs a weighted average of gradients we may apply the Jensen Inequality defined in 0.3 as:

$$f(\hat{y}_t) - f(z) \leq \frac{1}{\sum_{t=0}^{T-1} a_i} \sum_{t=0}^{T-1} a_i (f(y_{t+1}) - f(z))$$

Given the fact that $\sum_{t=0}^{T-1} a_i \geqslant \Omega(T^2)$ it is sufficient to bound the numerator by a constant in order to get the sublinear converge rate $\mathcal{O}(C/T^2)$ and the proof is focused precisely at this point.

**Lemma 17.** *Assume that $f$ is convex and $\beta$-smooth. Then for any sequence of non-negative weights $\{\alpha_t\}_{t\geq 0}$, and learning rates $\{\eta_t\}_{t\geq 0}$, Algorithm 2 ensures the following to hold:*

$$\alpha_t \left(f\left(y_{t+1}\right) - f(z)\right) \leq \left(\alpha_t^2 - \alpha_t\right)\left(f\left(y_t\right) - f\left(y_{t+1}\right)\right) + \frac{\alpha_t^2}{2}\left(\beta - \frac{1}{\eta_t}\right)\left\|y_{t+1} - x_{t+1}\right\|^2 + \frac{1}{2\eta_t}\left(\left\|z_t - z\right\|^2 - \left\|z_{t+1} - z\right\|^2\right)$$

Equipped with the following Lemma we can proceed as follows:

$$\sum_{t=0}^{T-1} a_i(f(y_{t+1}) - f(z)) \leq \underbrace{(\alpha_t^2 - \alpha_t)(f(y_t) - f(y_{t+1})}_{\text{(A)}} + \underbrace{\frac{\alpha_t^2}{2}\left(\beta_t - \frac{1}{\eta_t}\right)\|y_{t+1} - x_{t+1}\|^2}_{\text{(B)}} + \underbrace{\frac{1}{2\eta_t}(\|z_t - z\|^2 - \|z_{t+1} - z\|^2)}_{\text{(C)}}.$$

We now derive for each term of the sum an upper bound:

**Bounding (A)** Since $\{1/\eta_t\}_{t\in T}$ is monotonically increasing in $t$ we may bound A as follows:

$$\sum_{t=0}^{T-1} \frac{1}{2\eta_t}\left(\left\|z_t - z\right\|^2 - \left\|z_{t+1} - z\right\|^2\right) \leq \frac{1}{2}\sum_{t=1}^{T-1}\left\|z_t - z\right\|^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{\left\|z_0 - z\right\|^2}{2\eta_0} \leq \frac{D^2}{2\eta_{T-1}}$$

where in the second line we do the following trick, we define $\delta_t = \|z_t - z\|^2$ where we now that $\|z_t - z\|^2$, then we have:

$$\left(\frac{1}{\eta_1} - \frac{1}{\eta_0}\right)\delta_1 + \cdots + \left(\frac{1}{\eta_{T-2}} - \frac{1}{\eta_{T-1}}\right)\delta_{T-1}$$

since $\delta_t$ is bounded by the diameter of the considered set $D$ we have:

$$\left(\frac{1}{\eta_1} - \frac{1}{\eta_0}\right)D^2 + \cdots + \left(\frac{1}{\eta_{T-2}} - \frac{1}{\eta_{T-1}}\right)D^2.$$

No we can bring upside the $D^2$ from the sum and the remaining term is $-1/\eta_0 + 1/\eta_{T-1}$ and we obtain the final result with simple math adjustments.

**Bounding (B)**

To bound this sequence we require the following lemma to hold:

**Lemma 18.** *The following holds for the $\alpha_t$ which are described in Eq.7:*

$$\left(\alpha_t^2 - \alpha_t\right) - \left(\alpha_{t-1}^2 - \alpha_{t-1}\right) \leq \alpha_{t-1}/2$$

We can now bound (B), to do that we recall the fact that $z$ is the minimizer of $f$. and let denote the suboptimality of $y_t$ with $\delta_t = f(y_t) - f(z)$ which is greater than or equal to zero, we have:

$$\begin{aligned}
&\textstyle\sum_{t=0}^{T-1}\left(\alpha_t^2 - \alpha_t\right)\left(f\left(y_t\right) - f\left(y_{t+1}\right)\right)\\
&= \textstyle\sum_{t=0}^{T-1}\left(\alpha_t^2 - \alpha_t\right)\left(\delta_t - \delta_{t+1}\right)\\
&= \textstyle\sum_{t=1}^{T-1}\underbrace{\left(\left(\alpha_t^2 - \alpha_t\right) - \left(\alpha_{t-1}^2 - \alpha_{t-1}\right)\right)}_{\text{(Apply Lemma 3)}}\delta_t + \underbrace{\left(\alpha_0^2 - \alpha_0\right)}_{(=0)}\delta_0 - \underbrace{\left(\alpha_{T-1}^2 - \alpha_{T-1}\right)}_{(\geq 0)}\delta_T\\
&\leq \tfrac{1}{2}\textstyle\sum_{t=1}^{T-1}\alpha_{t-1}\delta_t\\
&\leq \tfrac{1}{2}\textstyle\sum_{t=1}^{T-1}\alpha_{t-1}\delta_t + \tfrac{1}{2}\alpha_{T-1}\delta_T\\
&= \tfrac{1}{2}\textstyle\sum_{t=0}^{T-1}\alpha_t\delta_{t+1}\\
&= \tfrac{1}{2}\textstyle\sum_{t=0}^{T-1}\alpha_t\left(f\left(y_{t+1}\right) - f(z)\right)
\end{aligned},$$

which conclude the proof of this bound.

**Bounding (C)**

We denote $\tau_*$ as:

$$\tau_\star = \max\left\{t \in \{0,\dots,T-1\} : 2\beta \geq 1/\eta_t\right\},$$

and we split the term (C) as follows:

$$\begin{aligned}
&\textstyle\sum_{t=0}^{T-1} \frac{\alpha_t^2}{2}\left(\beta - \frac{1}{\eta_t}\right)\|y_{t+1}-x_{t+1}\|^2 \\
&\quad= \textstyle\sum_{t=0}^{\tau_\star} \underbrace{\frac{\alpha_t^2}{2}\left(\beta - \frac{1}{\eta_t}\right)\|y_{t+1}-x_{t+1}\|^2}_{(2\beta \geq \frac{1}{\eta_t})} + \sum_{t=\tau_\star+1}^{T-1} \underbrace{\frac{\alpha_t^2}{2}\left(\beta - \frac{1}{\eta_t}\right)\|y_{t+1}-x_{t+1}\|^2}_{(2\beta \leq \frac{1}{\eta_t})} \\
&\leq \textstyle\frac{\beta}{2}\sum_{t=0}^{\tau_\star} \alpha_t^2 \underbrace{\|y_{t+1}-x_{t+1}\|^2}_{(=\eta_t\|g_t\|)} - \frac{1}{4}\sum_{t=\tau_\star+1}^{T-1} \frac{\alpha_t^2}{\eta_t} \underbrace{\|y_{t+1}-x_{t+1}\|^2}_{(=\eta_t\|g_t\|)} \\
&= \textstyle\frac{\beta}{2}\sum_{t=0}^{\tau_\star} \eta_t^2\alpha_t^2\|g_t\|^2 - \frac{1}{4}\sum_{t=\tau_\star+1}^{T-1} \eta_t\alpha_t^2\|g_t\|^2,
\end{aligned}$$

where the third line inequality comes from the fact that $\beta - \frac{1}{\eta_t} \leq -\frac{1}{2\eta_t}$.

Combining all the above bounds and doing some re-arrangements yields the new compact bound:

$$\frac{1}{2}\sum_{t=0}^{T-1} a_i(f(y_{t+1}) - f(z)) \leq \underbrace{\frac{D^2}{2\eta_{T-1}} - \frac{1}{4}\sum_{t=\tau_*+1}^{T-1}\eta_t\alpha_t^2\|g_t\|^2}_{(*)} + \underbrace{\frac{\beta}{2}\sum_{t=0}^{\tau_*}\eta_t^2\alpha_t^2\|g_t\|^2}_{(**)}.$$

Now by fixing the learning rate equal to the value in Equation.6, we provide a final bound which is constant and we conclude the proof. These steps requires many computations and Lemmas that we skip in order to provide the higher point of view.

The final result is of the form:

$$\frac{1}{2}\sum_{t=0}^{T-1} a_i(f(y_{t+1}) - f(z)) \leq DG/4 + 2\beta D^2 + 2\beta D^2(1 + 2log(4\beta D/D)$$

Combining this result with our initial inequality we get by Jensen Inequality the following final bound:

$$\begin{aligned}
f(\hat{y}_t) - f(z) &\leq \frac{1}{\sum_{t=0}^{T-1} a_i}\sum_{t=0}^{T-1} a_i(f(y_{t+1}) - f(x^*)) \\
&\leq \frac{DG/2 + 8\beta D^2(1 + log(4\beta D/D)}{T^2/32} \\
&\approx \mathcal{O}\left(\frac{DG + \beta D^2 log(4\beta D/D)}{T^2}\right),
\end{aligned}$$

where we used the fact that $\alpha_t \geq \frac{1}{4}(t+1)$ and $\sum_{t=0}^{T-1} a_i \geqslant T^2/32$ for the "Gauss sum", which conclude the proof for the smooth case. □

## Non-smooth setting

**Theorem 19.** *Assume that $f$ is convex and $G$-Lipschitz. Let $K$ be a convex set with bounded diameter $D$, and assume there exists a global minimizer for $f$ in $K$. Then Algorithm 3 with weights equal to Equation 7 and learning rate as in Equation 6 ensures:*

$$f(\hat{y}_t) - f(x^*) \leq \mathcal{O}\left(GD\sqrt{logT}/\sqrt{T}\right)$$

*Proof.* The proof follows the same considerations of the previous one, this time we need to prove that $\sum_{t=0}^{T-1} a_i(f(y_{t+1}) - f(z))$ is bounded by $\mathcal{O}(T^{3/2})$. As in the previous case we can bound our term of interest using the following Lemma which is slightly different from the previous one:

**Lemma 20.** *Assume that $f$ is convex and G-Lipschitz. Then for any sequence of non-negative weights $\{\alpha_t\}_{t\geq 0}$, and learning rates $\{\eta_t\}_{t\geq 0}$, Algorithm.3 ensures the following to hold:*

$$\begin{aligned}&\alpha_t \left(f\left(y_{t+1}\right) - f(z)\right)\\ &\quad\leq \eta_t \alpha_t^2 \left\|g_t\right\|^2 + \eta_t \alpha_t^2 \left\|g_t\right\| G + \tfrac{1}{2\eta_t}\left(\left\|z_t - z\right\|^2 - \left\|z_{t+1} - z\right\|^2\right) + \left(\alpha_t^2 - \alpha_t\right)\left(f\left(y_t\right) - f\left(y_{t+1}\right)\right)\end{aligned}$$

From now on we can bound each term of this sum and combine the results to conclude our proof similarly to what we have done in the smooth setting, for this reason in this review we do not re-write all the analysis. □

## Stochastic setting

**Theorem 21.** *Assume that $f$ is convex and G-Lipschitz. Let $K$ be a convex set with bounded diameter $D$, and assume there exists a global minimizer for $f$ in $K$. Assume that we involve Algorithm 3 with weights equal to Equation 7 and learning rate as in Equation 6 ensures, but this time we provide it with noisy gradient estimates, then our algorithm ensures:*

$$\mathbb{E}[f(\hat{y}_t)] - f(x^*) \leq \mathcal{O}\left(GD\sqrt{logT}/\sqrt{T}\right)$$

*Proof.* In order to provide the proof of this theorem we assume that upon querying a first order oracle with a point $x$, we receive a bounded and unbiased gradient estimate $\tilde{g}$ such that $\mathbb{E}[\tilde{g}|x] = \nabla f(x)$ with $\|\tilde{g}\| \leq G$ . As in the previous theorems we need to find an upper bound for $\mathbb{E}[f(\hat{y}_t)] - f(x^*)$ to obtain the claimed convergence rate. The procedure is again similar to the previous ones and we start by bounding $a_i \left(f\left(y_{t+1}\right) - f\left(z\right)\right)$ as follows:

**Lemma 22.** *Assume that $f$ is convex and G-Lipschitz. Assume that we invoke Algorithm.3 but provide it with noisy gradient estimates rather then the exact ones. Then for any sequence of non-negative weights $\{\alpha_t\}_{t\geq 0}$, and learning rates $\{\eta_t\}_{t\geq 0}$, the following holds*

$$\begin{aligned}&\alpha_t \left(f\left(y_{t+1}\right) - f(z)\right)\\ &\quad\leq \eta_t \alpha_t^2 \left\|\tilde{g}_t\right\|^2 + \eta_t \alpha_t^2 \left\|\tilde{g}_t\right\| G + \tfrac{1}{2\eta_t}\left(\left\|z_t - z\right\|^2 - \left\|z_{t+1} - z\right\|^2\right) + \left(\alpha_t^2 - \alpha_t\right)\left(f\left(y_t\right) - f\left(y_{t+1}\right)\right)\\ &\qquad+\alpha_t \left(g_t - \tilde{g}_t\right)\cdot\left(z_t - z\right)\end{aligned}$$

*Proof.* First we note that due to the unbiassdness of the queried gradients the last term $\alpha_t \left(g_t - \tilde{g}_t\right) \cdot \left(z_t - z\right)$ is zero. The others terms are retrieved following the same procedure of the first bound that we found on the non-smooth setting which are derived consequently in a similar way of the smooth setting. Since finding each bound requires many computations which are quite similar to the previous ones we don't provide the detailed proof. Removing the last term we again find the same bound of the non-smooth setting upon replacing $\tilde{g}_t$ with $g_t$ and this guides us towards the same convergence rate. □

□