UNDERGRAD: A Universal Black-Box Optimization Method with Almost Dimension-Free Convergence Rate Guarantees

Kimon Antonakopoulos [1 2] **Dong Quan Vu** [3 2] **Volkan Cevher** [1] **Kfir Y. Levy** [4 5] **Panayotis Mertikopoulos** [6 7]

Abstract

Universal methods for optimization are designed to achieve theoretically optimal convergence rates without any prior knowledge of the problem's regularity parameters or the accurarcy of the gradient oracle employed by the optimizer. In this regard, existing state-of-the-art algorithms achieve an $\mathcal{O}(1/T^2)$ value convergence rate in Lipschitz smooth problems with a perfect gradient oracle, and an $\mathcal{O}(1/\sqrt{T})$ convergence rate when the underlying problem is non-smooth and/or the gradient oracle is stochastic. On the downside, these methods do not take into account the problem's dimensionality, and this can have a catastrophic impact on the achieved convergence rate, in both theory and practice. Our paper aims to bridge this gap by providing a scalable universal gradient method – dubbed UNDERGRAD – whose oracle complexity is almost dimension-free in problems with a favorable geometry (like the simplex, linearly constrained semidefinite programs and combinatorial bandits), while retaining the order-optimal dependence on T described above. These "best-of-both-worlds" results are achieved via a primal-dual update scheme inspired by the dual exploration method for variational inequalities.

1. Introduction

The analysis of first-order methods for convex minimization typically revolves around the following basic regularity conditions: *a*) Lipschitz continuity of a problem's objective function and / or *b*) Lipschitz continuity of the objective's gradients. Depending on these conditions and the quality of the gradient oracle available to the optimizer, the optimal convergence rates that can be obtained by an iterative first-order algorithm after T oracle queries are:

1. $\mathcal{O}\big(\|\mathcal{X}\|\sqrt{(G^2+\sigma^2)/T}\big)$ if the problem's objective is G-Lipschitz continuous and the oracle's variance is σ.
2. $\mathcal{O}\big(L\|\mathcal{X}\|^2/T^2 + \sigma\|\mathcal{X}\|/\sqrt{T}\big)$ if the objective is L-Lipschitz smooth.

[In both cases, $\|\mathcal{X}\| \coloneqq \sup_{x,x'\in\mathcal{X}}\|x'-x\|$ denotes the diameter of the problem's domain $\mathcal{X}\subseteq\mathbb{R}^d$; for an in-depth treatment, see [11, 38] and references therein.]

This stark separation of black-box guarantees has led to an intense search for *universal* methods that are capable of interpolating smoothly between these rates without any prior knowledge of the problem's regularity properties or the oracle's noise profile. As far as we are aware, the first algorithm with order-optimal rate guarantees for unconstrained problems and no knowledge of the problem's smoothness parameters was the ACCELEGRAD proposal of Levy et al. [28]. Subsequently, in the context of *constrained* convex problems (the focus of our work), Kavis et al. [24] combined the extra-gradient / mirror-prox algorithmic template of Korpelevich [25] and Nemirovski [36] with an "iterate averaging" scheme introduced by Cutkosky [16] to change the query structure of the base algorithm and make it more amenable to acceleration. In this way, Kavis et al. [24] obtained a universal extra-gradient algorithm – dubbed UNIXGRAD – which interpolates between the optimal rates mentioned above, without requiring any tuning.

Our contributions. The starting point of our paper is the observation that, even though the rates in question are optimal in T, they may be highly supoptimal in d, the problem's dimensionality. For example, if the noise in the oracle has unit variance, σ would scale as $\mathcal{O}(\sqrt{d})$; this represents a hidden dependence on d which could have a catastrophic impact on the method's actual convergence rate. Likewise, in problems with a favorable geometry (like the L^1-ball, trace-constrained semidefinite programs, combinatorial bandits, etc.), methods based on the mirror descent [37] and mirror-

[1]Laboratory for Information and Inference Systems, IEL STI EPFL, 1015 Lausanne, Switzerland. [2]This work was done when KA and DQV were with Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France. [3]Safran Tech, Magny-Les-Hameaux, France. [4]Technion, Haifa, Israel. [5]A Viterbi Fellow. [6]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France. [7]Criteo AI Lab. Correspondence to: Kimon Antonakopoulos <kimon.antonakopoulos@epfl.ch>, Dong Quan Vu <dong-quan.vu@safrangroup.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

prox [36] templates can achieve rates with a *logarithmic* (instead of polynomial) dependence on d.

Importantly, the UNIXGRAD algorithm of Kavis et al. [24] is itself based on the mirror-prox blueprint, so it would seem ideally suited to achieve convergence rates that are simultaneously optimal in T and d. However, the method's guarantees depend crucially on the *Bregman diameter* of the problem's domain, a quantity which becomes infinite when the method is used with a regularization setup that can lead to almost dimension-free guarantees. This would seem to suggest that universality comes at the cost of scalability, leading to the following open question:

Is it possible to achieve almost dimension-free convergence rates while retaining an order-optimal dependence on T?

In this paper, we develop a novel adaptive algorithm, which we call *universal dual extrapolation with reweighted gradients* (UNDERGRAD), and which provides a positive answer to this question. Specifically, the value convergence rate of UNDERGRAD scales in terms of G, σ, L and T as:

1. $\mathcal{O}\big(\sqrt{R_h(G^2+\sigma^2)/T}\big)$ in non-smooth problems.
2. $\mathcal{O}\big(R_h L/T^2 + \sigma\sqrt{R_h/T}\big)$ in smooth problems.

In the above, the method's "range parameter" R_h scales as $\mathcal{O}(\|\mathcal{X}\|^2)$ in the worst case and as $\mathcal{O}(\log d)$ in problems with a favorable geometry – that is, in problems where it is possible to attain almost dimension-free convergence rates [11, 38]. In this regard, UNDERGRAD seems to be the first method in the literature that concurrently achieves order-optimal rates in both T and d, without any prior knowledge on the problem's level of smoothness.

To achieve this result, the UNDERGRAD algorithm combines the following basic ingredients:

1. A modified version of the dual extrapolation method of Nesterov [39] for solving variational inequalities.
2. A gradient "reweighting" scheme that allows gradients to enter the algorithm with *increasing* weights.
3. An iterative averaging scheme in the spirit of Cutkosky [16] which allows us to obtain an accelerated rate of convergence by means of an online-to-batch conversion.

The glue that holds these elements together is an adaptive learning rate inspired by Rakhlin & Sridharan [41, 42] which automatically rescales aggregated gradients by *a*) a small, constant amount when the method approaches a solution where gradient differences vanish (as in the smooth, deterministic case); and *b*) a factor of $\mathcal{O}(\sqrt{T})$ otherwise (thus providing the desired interpolation between smooth and non-smooth problems). In so doing, the proposed policy achieves the correct step-size scaling and achieves the desired optimal rates.

Related work. The term "universality" was coined by Nesterov [40] whose *universal primal gradient descent* (UPGD) algorithm interpolates between the $\mathcal{O}(1/T^2)$ and $\mathcal{O}(1/\sqrt{T})$ rates for smooth and non-smooth problems respectively (assuming access to noiseless gradients in both cases). On the downside, UPGD relies on an Armijo-like line search to interpolate between smooth and non-smooth objectives, so it is not applicable to stochastic environments.

A partial work-around to this issue was achieved by the accelerated stochastic approximation (AC-SA) algorithm of Lan [26] which uses a mirror descent template and guarantees order-optimal rates for both noisy *and* noiseless oracles. However, to attain these rates, the AC-SA algorithm requires a precise estimate of the smoothness modulus of the problem's objective, so it is not universal in this respect. Subsequent works on the topic have focused on attaining universal guarantees for composite problems [21], non-convex objectives [29, 46], preconditioned methods [17, 21], non-Lipschitz settings [2–4], specific applications [45], or variational inequalities / min-max problems [4, 5, 7, 20].

Of the generalist works above, some employ a Bregman regularization setup [2, 7], but the guarantees obtained therein either fall short of an accelerated $\mathcal{O}(1/T^2)$ convergence rate for Lipschitz smooth problems, or they depend on the problem's Bregman diameter – so they cannot be associated with a Bregman setup leading to almost dimension-free convergence rate guarantees. To the best of our knowledge, UNDERGRAD is the first method that manages to combine the "best of both worlds" in terms of universality with respect to T and scalability with respect to d.

2. Preliminaries

2.1. Notation and basic definitions

Let $\mathcal{V}$ be a d-dimensional space with norm $\|\cdot\|$. In what follows, we will write $\mathcal{Y} \equiv \mathcal{V}^*$ for the dual of $\mathcal{V}$, $\langle y, x\rangle$ for the pairing between $y \in \mathcal{Y}$ and $x \in \mathcal{V}$, and $\|y\|_* \equiv \sup\{\langle y, x\rangle : \|x\| \leq 1\}$ for the dual norm on $\mathcal{Y}$. Given an extended-real-valued convex function $f : \mathcal{V} \to \mathbb{R}\cup\{\infty\}$, we will write $\operatorname{dom} f \equiv \{x \in \mathcal{V} : f(x) < \infty\}$ for its *effective domain* and $\partial f(x) \equiv \{y \in \mathcal{Y} : f(x') - f(x) - \langle y, x' - x\rangle \geq 0 \text{ for all } x' \in \mathcal{V}\}$ for the *subdifferential* of f at $x \in \operatorname{dom} f$. Any element $g \in \partial f(x)$ will be called a *subgradient* of f at x, and we will write $\operatorname{dom}\partial f \equiv \{x \in \operatorname{dom} f : \partial f \neq \varnothing\}$ for the *domain of subdifferentiability* of f.

2.2. Problem setup and blanket assumptions

The main focus of our paper is the solution of convex minimization problems of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \mathcal{X} \end{aligned} \tag{Opt}$$

where $\mathcal{X}$ is a closed convex subset of $\mathcal{V}$ and $f\colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is a convex function with $\operatorname{dom} f = \operatorname{dom} \partial f = \mathcal{X}$. To avoid trivialities, we will assume throughout that the solution set $\mathcal{X}^* := \arg\min f$ of (Opt) is non-empty, and we will write x^* for a generic minimizer of f.

Other than this blanket assumption, our main regularity requirements for f will be as follows:

1. *Lipschitz continuity:*

$$|f(x') - f(x)| \leq G\|x' - x\| \tag{LC}$$

for some $G \geq 0$ and for all $x, x' \in \mathcal{X}$.

2. *Lipschitz smoothness:*

$$f(x') \leq f(x) + \langle \nabla f(x), x' - x\rangle + \frac{L}{2}\|x' - x\|^2 \tag{LS}$$

for some $L \geq 0$ and for all $g \in \partial f(x)$, $x, x' \in \mathcal{X}$.

Since $\operatorname{dom} \partial f = \mathcal{X}$, the above requirements are respectively equivalent to assuming that f admits a selection of subgradients $\nabla f(x) \in \partial f(x)$ with the properties below:

1. *Bounded (sub)gradient selection:*

$$\|\nabla f(x)\|_* \leq G \tag{BG}$$

for some $G \geq 0$ and for all $x \in \mathcal{X}$.

2. *Lipschitz (sub)gradient selection:*

$$\|\nabla f(x') - \nabla f(x)\|_* \leq L\|x' - x\| \tag{LG}$$

for some $L \geq 0$ and for all $x, x' \in \mathcal{X}$.

In the rest of our paper, we will assume that f satisfies at least one of (BG) or (LG).

Remark 1. For posterity, we note here that the requirement (LG) *does not* imply that $\partial f(x)$ is a singleton.[1] In any case, the directional derivative $f'(x; z) = d/dt|_{t=0} f(x + tz)$ of f at $x \in \mathcal{X}$ along $z \in \mathcal{V}$ exists and is equal to $\langle \nabla f(x), z\rangle$ for all vectors of the form $z = x' - x$, $x' \in \mathcal{X}$. We will use this fact freely in the sequel. ¶

2.3. The oracle model

To solve (Opt), we will consider iterative methods and algorithms with access to a *stochastic first-order oracle* (SFO), i.e., a black-box device that returns a (possibly random) estimate of a subgradient of f at the point at which it was queried. Formally, following Nesterov [38], an SFO for f is a measurable function $\mathsf{G}\colon \mathcal{X} \times \Omega \to \mathcal{Y}$ such that

$$\mathbb{E}[\mathsf{G}(x; \omega)] = \nabla f(x) \quad \text{for all } x \in \mathcal{X} \tag{SFO}$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a complete probability space and $\nabla f(x)$ is a selection of subgradients of f as per (BG)/(LG). The oracle's statistical precision will then be measured by the associated *noise level* $\sigma := \operatorname{ess\,sup}_{\omega,x} \|\mathsf{G}(x; \omega) - \nabla f(x)\|_*$ (assumed finite). In particular, if $\sigma = 0$, G will be called *perfect* (or *deterministic*); otherwise, G will be called *noisy*.

In practice, the oracle is called repeatedly at a sequence of query points x_t with a different random seed ω_t drawn according to $\mathbb{P}$ at each time.[2] In this way, at the t-th query to (SFO), the oracle G returns the gradient signal

$$g_t = \mathsf{G}(x_t; \omega_t) = \nabla f(x_t) + U_t \tag{1}$$

where U_t denotes the "gradient noise" of the oracle (obviously, $U_t \equiv 0$ if the oracle is perfect). For measurability purposes, we will write $\mathcal{F}_t$ for the history (adapted filtration) of x_t, so x_t is $\mathcal{F}_t$-measurable (by definition) but ω_t, g_t and U_t are not. In particular, conditioning on $\mathcal{F}_t$, we have $\mathbb{E}[g_t \,|\, \mathcal{F}_t] = \nabla f(x_t)$ and $\mathbb{E}[U_t \,|\, \mathcal{F}_t] = 0$, justifying in this way the terminology "gradient noise" for U_t.

Remark 2. The oracle model detailed above is not the only one possible, but it is very widely used in the analysis of parameter-agnostic and adaptive methods, cf. [2, 24, 28, 46] and references therein. In view of this, we will not examine either finer or coarser assumptions for (SFO). ¶

We close this section by noting that the best convergence rates that can be achieved by an iterative algorithm that outputs a candidate solution $\bar{x}_T \in \mathcal{X}$ after T queries to (SFO) are:[3]

1. $f(\bar{x}_T) - \min f = \mathcal{O}(1/\sqrt{T})$ if f satisfies (BG) and G is deterministic.
2. $f(\bar{x}_T) - \min f = \mathcal{O}(1/T^2)$ if f satisfies (LG) and G is deterministic.
3. $\mathbb{E}[f(\bar{x}_T) - \min f] = \mathcal{O}(1/\sqrt{T})$ if G is stochastic.

In general, without finer assumptions on f or G, the dependence of these rates on T cannot be improved [11, 38]; we will revisit this issue several times in the sequel.

3. Regularization, universality, and the curse of dimensionality

To set the stage for the analysis to come, we discuss below the properties of two algorithmic frameworks – one non-adaptive, the other adaptive – based on the mirror-prox template [36]. Our aim in doing this will be to set a baseline for the sequel as well as to explore the impact of the problem's dimensionality on the attained rates of convergence.

[1] Consider for example the case of $f(x) = x$ for $x \in [0, 1]$ and $f(x) = \infty$ otherwise: f clearly satisfies (BG)/(LS), even though its $\partial f(0)$ and $\partial f(1)$ are infinite sets.

[2] In the sequel, t may take both integer and half-integer values.

[3] In general, the query and output points – x_T and $\bar{x}_T$ respectively – need not coincide, hence the different notation. The only assumption for the rates provided below is that the output point $\bar{x}_T$ is an affine combination of $x_1, g_1, \dots, x_T, g_T$ [11, 38].

3.1. Motivating examples

As a first step, we present three archetypal problems to motivate and illustrate the general setup that follows.

Example 1 (Resource allocation)**.** Consider a set of computing resources (GPUs in a cluster, servers in a computing grid, ...) indexed by $s \in \mathcal{S} = \{1, \dots, d\}$. Each resource is capable of serving a stream of computing demands that arrive at a rate of ρ units per time: if the optimizer assigns a load of $x_s \geq 0$ to the s-th resource, the marginal cost incurred is $c_s(x_s)$ per unit served, where $c_s \colon [0, \rho] \to \mathbb{R}_+$ is the cost function of the s-th resource (assumed convex, differentiable, and increasing in x_s). Taking $\rho = 1$ for simplicity, the goal of the optimizer is to minimize the aggregate cost $f(x) = \sum_{s=1}^{d} x_s c_s(x_s)$, leading to a convex minimization problem over the unit d-dimensional simplex $\mathcal{X} = \Delta(\mathcal{S}) = \{x \in \mathbb{R}_+^d : \sum_s x_s = 1\}$. ¶

Example 2 (Input covariance matrix optimization)**.** Consider a Gaussian vector channel in the spirit of [44, 47]: the encoder controls the covariance matrix $\mathbf{X} = \mathbb{E}[\mathbf{x}\mathbf{x}^\dagger]$ of the Gaussian input signal $\mathbf{x} \in \mathbb{C}^M$ and seeks to maximize the transfer rate of the output signal $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}$, where $\mathbf{z} \in \mathbb{C}^N$ is the ambient noise in the channel and $\mathbf{H} \in \mathbb{C}^{N \times M}$ is the channel's transfer matrix. By the Shannon–Telatar formula [44], this boils down to maximizing the capacity function

$$R(\mathbf{X}) = \mathbb{E}\big[\log\det\big(\mathbf{I} + \mathbf{H}\mathbf{X}\mathbf{H}^\dagger\big)\big] \tag{2}$$

subject to the constraint $\operatorname{tr}(\mathbf{X}) \leq P$, where P denotes the encoder's maximum input power and the expectation in (2) is taken over the statistics of the (possibly deterministic) matrix $\mathbf{H}$. Since R is concave in $\mathbf{X}$ [10, 47], we obtain a minimization problem of the form (Opt) over the spectrahedron $\mathcal{D} = \{\mathbf{X} \succcurlyeq 0 : \operatorname{tr}(\mathbf{X}) \leq P\}$. Since $\mathbf{X}$ is Hermitian, $\mathcal{D}$ can be seen as a convex body of $\mathbb{R}^d$ where $d = M^2$; in the optimization literature, this is sometimes referred to as the "spectrahedron setup" [22]. ¶

Example 3 (Combinatorial bandits)**.** In bandit linear optimization problems, the optimizer is given a finite set of n possible *actions* $\mathcal{A} \subseteq \{0,1\}^d$, i.e., each action $\alpha \in \mathcal{A}$ is a d-dimensional binary vector indicating whether the i-th component is "on" or "off". The optimizer then chooses an action $\alpha \in \mathcal{A}$ based on a mixed strategy $p \in \Delta(\mathcal{A})$ and incurs the mean loss

$$\ell(p; \omega) = \mathbb{E}\Big[\sum\nolimits_{\alpha \in \mathcal{A}} p_\alpha \langle \alpha, \omega \rangle\Big] \tag{3}$$

where ω is a random vector with values in $[0,1]^d$ (but otherwise unknown distribution). In many cases of interest – such as slate recommendation and shortest-path problems – the cardinality of $\mathcal{A}$ is exponential in d, so it is computationally prohibitive to state the resulting minimization problem in terms of p. Instead, writing $x_i = \sum_{\alpha \in \mathcal{A}} p_\alpha \alpha_i$ for the probability of the i-th component being "on" under p, the optimizer's objective can be rewritten more compactly as $f(x) = \mathbb{E}[\langle x, \omega \rangle]$ with x constrained to lie on the d-dimensional convex hull $\mathcal{X} = \operatorname{conv}(\mathcal{A})$ of $\mathcal{A}$ in $\mathbb{R}^d$. In the literature on multi-armed bandits, this setup is known as a *combinatorial bandit*; for an in-depth treatment, see [13, 14, 19, 27] and the many references cited therein. ¶

Examples 1–3 all suffer from the "curse of dimensionality": for instance, the dimensionality of a vector Gaussian channel with $M = 256$ input entries is $d \approx 6.5 \times 10^4$, while a combinatorial bandit for recommendation systems may have upwards of several million arms. Nonetheless, these examples also share a number of geometric properties that make it possible to design scalable optimization algorithms with (almost) dimension-free convergence rate guarantees. We elaborate on this in the next section.

3.2. The mirror-prox template

We begin by considering the well-known *mirror-prox* (MP) method of Nemirovski [36]. Following [22, 35], this is defined via the recursion

$$\begin{aligned} X_{t+1/2} &= P_{X_t}(-\gamma_t g_t) \\ X_{t+1} &= P_{X_t}(-\gamma_t g_{t+1/2}) \end{aligned} \tag{MP}$$

where

1. $t = 1, 2, \dots$ denotes the method's iteration counter (for the origins of the half-integer notation, see Facchinei & Pang [18] and references therein).
2. $\gamma_t > 0$ is the algorithm's step-size sequence.
3. g_t and $g_{t+1/2}$ are stochastic gradients of f obtained by querying the oracle G at X_t and $X_{t+1/2}$ respectively.
4. $P_{X_t}(\cdot)$ is a generalized projection operator known as the method's "prox-mapping" (more on this later).

The most elementary instance of (MP) is the *extra-gradient* (EG) algorithm of Korpelevich [25], in which case the method's prox-mapping is the Euclidean projector

$$P_x(y) = \Pi_{\mathcal{X}}(x + y) \coloneqq \arg\min\nolimits_{x' \in \mathcal{X}} \|x + y - x'\|_2 \tag{4}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. More generally, the prox-mapping in (MP) is defined in terms of a *Bregman regularizer* as follows:

Definition 1. A *Bregman regularizer* on $\mathcal{X}$ is a convex function $h \colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ such that

1. $\operatorname{dom} h = \mathcal{X}$ and h is continuous on $\mathcal{X}$.
2. The subdifferential of h admits a *continuous selection*, i.e., there exists a continuous mapping $\nabla h \colon \operatorname{dom} \partial h \to \mathcal{Y}$ with $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom} \partial h$.
3. h is *strongly convex* on $\mathcal{X}$, i.e.,

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \tfrac{1}{2} K_h \|x' - x\|^2 \tag{5}$$

	Domain ($\mathcal{X}$)	**Breg. Diameter** (B_h)	**Range** (R_h)	**Shape** (χ)	**Rate** ($L=\infty$)	**Rate** ($L<\infty, \sigma=0$)
EUCLIDEAN	any below	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\sqrt{d}$	$\mathcal{O}\big(\sqrt{d/T}\big)$	$\mathcal{O}(d/T)$
ENTROPIC	simplex	∞	$\log d$	1	$\mathcal{O}\big(\sqrt{\log d/T}\big)$	$\mathcal{O}(\log d/T)$
VON NEUMANN	spectrahedron	∞	$\log d$	1	$\mathcal{O}\big(\sqrt{\log d/T}\big)$	$\mathcal{O}(\log d/T)$
COMBAND	$\operatorname{conv}(\mathcal{A})$	∞	$\mathcal{O}(\log d)$	1	$\mathcal{O}\big(\sqrt{\log d/T}\big)$	$\mathcal{O}(\log d/T)$

Table 1: The convergence rate of (MP) in terms of d and T for different regularizers. In the combinatorial setup of Example 3, the unnormalized entropy has $R_h = m(1+\log(d/m))$, where $m = \max_{\alpha\in\mathcal{A}}\|\alpha\|_1$ is the maximum number of elements of $\{1,\dots,d\}$ that can be simultaneously "on" [27, Chap. 30]. In many applications, m does not scale with d, so it has been absorbed in the $\mathcal{O}(\cdot)$ notation; other than that, $\mathcal{O}(\cdot)$ contains only universal constants.

for some $K_h > 0$ and all $x \in \operatorname{dom}\partial h$, $x' \in \mathcal{X}$.

We also define the *Bregman divergence* of h as

$$D(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x\rangle \tag{6}$$

and the induced *prox-mapping* as

$$P_x(y) = \arg\min_{x'\in\mathcal{X}}\{\langle y, x - x'\rangle + D(x', x)\} \tag{7}$$

for all $x \in \mathcal{X}_h$, $x' \in \mathcal{X}$ and all $y \in \mathcal{Y}$.

Remark. The set $\mathcal{X}_h := \operatorname{dom}\partial h$ is often referred to as the *prox-domain* of h; by standard results in convex analysis, we have $\operatorname{ri}\mathcal{X} \subseteq \mathcal{X}_h \subseteq \mathcal{X}$ [43, Chap. 26].

In terms of output, the candidate solution returned by (MP) after T iterations is the so-called "ergodic average"

$$\bar{X}_T = \frac{\sum_{t=1}^T \gamma_t X_{t+1/2}}{\sum_{t=1}^T \gamma_t}. \tag{8}$$

Then, assuming the method's step-size γ_t is chosen appropriately (more on this below), $\bar{X}_T$ enjoys the following guarantees [22, 42]:

a) If f satisfies (BG), then

$$\mathbb{E}[f(\bar{X}_T) - \min f] = \mathcal{O}\left(\sqrt{\frac{G^2+\sigma^2}{K_h}\frac{D_1}{T}}\right) \tag{9a}$$

b) If f satisfies (LG), then

$$\mathbb{E}[f(\bar{X}_T) - \min f] = \mathcal{O}\left(\frac{LD_1}{K_hT} + \sigma\sqrt{\frac{D_1}{K_hT}}\right) \tag{9b}$$

In the above, $D_1 = D(x^*, X_1)$ is the minimum Bregman divergence between a solution x^* of (Opt) and the initial state X_1 of (MP). In particular, if (MP) is initialized at the *prox-center* $x_c = \arg\min h$ of $\mathcal{X}$, we have

$$D_1 \leq h(x^*) - \min h \leq \max h - \min h =: R_h. \tag{10}$$

We will refer to $R_h = \max h - \min h$ as the *range* of h.

To quantify the interplay betwen the problem's dimensionality and the rate guarantees (9) for (MP), it will be convenient to introduce the normalized regularity parameters

$$G_h = \frac{G}{\sqrt{K_h}} \quad L_h = \frac{L}{K_h} \quad \text{and} \quad \sigma_h = \frac{\sigma}{\sqrt{K_h}} \tag{11}$$

and the associated *shape factor*

$$\chi = \begin{cases} \sqrt{G_h^2 + \sigma_h^2} & \text{if } L = \infty, \\ \sqrt{L_h} & \text{if } L < \infty \text{ and } \sigma = 0, \\ \sigma_h & \text{if } L < \infty \text{ and } \sigma > 0. \end{cases} \tag{12}$$

Since at least one of the terms $G/\sqrt{K_h}$, L/K_h and $\sigma/\sqrt{K_h}$ appears in (9), it follows that the leading term in T scales as $\mathcal{O}(\chi\sqrt{D_1/T})$ in non-smooth / stochastic environments, and as $\mathcal{O}(\chi^2 D_1/T)$ in smooth, deterministic problems.

The importance of the normalized parameters G_h, L_h, σ_h and the shape factor χ lies in the fact that they do not depend on the ambient norm $\|\cdot\|$ (a choice which, to a certain extent, is arbitrary). Indeed, if $\|\cdot\|$ and $\|\cdot\|'$ are two norms on $\mathcal{V}$ that are related as $\|\cdot\| \leq \mu\|\cdot\|'$ for some $\mu > 0$, it is straightforward to verify that h is $(\mu^2 K_h)$-strongly convex relative to $\|\cdot\|'$. Likewise, in terms of dual norms we have $\|\cdot\|_* \geq (1/\mu)\|\cdot\|'_*$, so the constants G, σ and L would respectively become μG, $\mu\sigma$ and $\mu^2 L$ when computed under $\|\cdot\|'$. In general, these inequalities are all tight, so a change in norm does not affect the shape factor χ; accordingly, any dependence of χ on d will be propagated verbatim to the guarantees (9).

In Table 1, we provide the values of R_h and χ for the following cases:

1. The *Euclidean regularizer* $h(x) = \|x\|_2^2/2$ that gives rise to the extra-gradient algorithm (4).
2. The *entropic regularizer* $h(x) = \sum_{i=1}^d x_i \log x_i$ for the simplex setup of Example 1.
3. The *von Neumann regularizer* $h(\mathbf{X}) = \operatorname{tr}(\mathbf{X}\log\mathbf{X}) + (1 - \operatorname{tr}\mathbf{X})\log(1-\operatorname{tr}\mathbf{X})$ for the spectrahedron setup of Example 2.
4. The *unnormalized entropy* $h(x) = \sum_{i=1}^d (x_i \log x_i - x_i)$ for the combinatorial setup of Example 3.

These derivations are standard, so we omit the details. For posterity, we only note that the logarithmic dependence on d is asymptotically optimal, cf. [12, 13] and references therein.

3.3. The UNIXGRAD algorithm

As can be seen from Table 1, the mirror-prox algorithm achieves an almost dimension-free rate of convergence when used with a suitable regularizer. However, this comes with two important caveats: First, the algorithm's rate in the smooth case falls short of the optimal $\mathcal{O}(1/T^2)$ dependence in T, so (MP) is suboptimal in this regard. Second, to achieve the rates presented in Eq. (9), the algorithm's step-size γ_t must be tuned with prior knowledge of the problem's parameters: in particular, under (BG), the algorithm must be run with step-size $\gamma_t \propto 1/\sqrt{(G^2+\sigma^2)T}$ while, under (LG), the algorithm requires $\gamma_t = K_h/L$ if $\sigma = 0$ and $\gamma_t \propto 1/(\sigma\sqrt{T})$ otherwise. This creates an undesirable state of affairs because the parameters G, L and σ are usually not known in advance, and (MP) can – and *does* – fail to converge if run with an untuned step-size.

In the rest of this section, we briefly discuss the UNIXGRAD algorithm of Kavis et al. [24] which expands on the mirror-prox template in the following two crucial ways: *a*) it introduces an iterate-averaging mechanism in the spirit of Cutkosky [16] to enable acceleration; and *b*) it employs an adaptive step-size policy that does not require any tuning by the optimizer. In so doing, UNIXGRAD interpolates smoothly between the optimal convergence rates described in Section 2 without requiring any prior knowledge of G, L or σ.

Concretely, UNIXGRAD proceeds as (MP), but instead of querying G at X_t and $X_{t+1/2}$, it introduces the weighted query states

$$\begin{aligned}\bar{X}_t &= \frac{\alpha_t X_t + \sum_{s=1}^{t-1} \alpha_s X_{s+1/2}}{\sum_{s=1}^{t} \alpha_s} \\ \bar{X}_{t+1/2} &= \frac{\alpha_t X_{t+1/2} + \sum_{s=1}^{t-1} \alpha_s X_{s+1/2}}{\sum_{s=1}^{t} \alpha_s}\end{aligned} \tag{13}$$

where α_t is a "gradient weighting" parameter. Then, building on an idea by Rakhlin & Sridharan [41, 42], the oracle queries $g_t \leftarrow \mathsf{G}(\bar{X}_t;\omega_t)$ and $g_{t+1/2} \leftarrow \mathsf{G}(\bar{X}_{t+1/2};\omega_{t+1/2})$ are used to update the method's step-size as

$$\gamma_t = \frac{B_h \alpha_t}{\sqrt{1 + \sum_{s=1}^{t-1} \alpha_s^2 \|g_{s+1/2} - g_s\|_*^2}} \tag{14}$$

where

$$B_h = \sup_{x\in\mathcal{X}, x'\in\mathcal{X}_h} \sqrt{2D(x,x')} \tag{15}$$

is the so-called *Bregman diameter* of $\mathcal{X}$.

With all this in hand, Kavis et al. [24] provide the following bounds if UNIXGRAD is run with $\alpha_t = t$:

a) If f satisfies (BG), then

$$\mathbb{E}[f(\bar{X}_{T+1/2}) - \min f] = \mathcal{O}\left(\frac{B_h\sqrt{G^2+\sigma^2}}{\sqrt{K_h T}}\right) \tag{16a}$$

b) If f satisfies (LG), then

$$\mathbb{E}[f(\bar{X}_{T+1/2}) - \min f] = \mathcal{O}\left(\frac{B_h^2 L}{K_h T^2} + \frac{B_h\sigma}{\sqrt{K_h T}}\right) \tag{16b}$$

As we mentioned in Section 2, the bounds (16) cannot be improved in terms of T without further assumptions, so UNIXGRAD is universally optimal in this regard.

That being said, these guarantees also uncover an important limitation of UNIXGRAD, namely that the bounds (16) become void when the method is used in conjunction with one of the non-Euclidean frameworks of Examples 1–3. For example, the Bregman diameter of the simplex under the entropic regularizer is $B_h = \sup_{x,x'} \sum_i x_i \log(x_i/x_i') = \infty$, so the multiplicative constants in (16) become infinite (and the bounds themselves become meaningless). However, since the use of these regularizers is crucial to obtain the scalable, dimension-free convergence rates reported in Table 1, [4] we are led to the open question we stated before:

Is it possible to achieve almost dimension-free convergence rates while retaining an order-optimal dependence on T?

We address this question in the next section.

4. Universal dual extrapolation

The point of departure of our analysis is the observation that gradient queries enter (MP) with *decreasing* weights. Specifically, if UNIXGRAD is run with $\alpha_t = t$ (a choice which is necessary to have a shot at acceleration), the denominator of (14) may grow as fast as $\Theta(t^{3/2})$ in the non-smooth/stochastic case, leading to an asymptotic $\mathcal{O}(1/\sqrt{t})$ worst-case behavior for γ_t. In fact, even under the ansatz that the algorithm's query points converge to a minimizer of f at an accelerated rate, the denominator of (14) may still grow as $\Theta(t)$, indicating that γ_t will, at best, stabilize to a positive value as $t \to \infty$. This feature of the step-size rule (14) is somewhat counterintuitive because conventional wisdom would suggest that *a*) recent queries are more useful than older, potentially obsolete ones; and *b*) gradients should be "inflated" as the method's query points approach a zero-gradient solution in order to maintain a fast rate of convergence.

[4] In particular, since the shape factor of the Euclidean regularizer is $\chi = \sqrt{d}$, employing UNIXGRAD with ordinary Euclidean projections would not lead to scalable guarantees.

The problem with a vanishing step-size becomes especially pronounced if the method is used with a non-Euclidean regularizer (which is what one would wish to do in order to obtain scalable convergence guarantees). To see this, consider the iterates of the mirror-prox template generated by the regularizer $h(x) = x\log x$ on $\mathcal{X} = [0,\infty)$.[5] In this case, the induced prox-mapping is $P_x(y) = x\exp(y)$, leading to the recursion

$$x^+ = P_x(-\gamma v) = x\exp(-\gamma v). \tag{17}$$

Therefore, if the problem's objective function attains its minimum at 0, the actual steps of the method scale as $x^+ - x = \mathcal{O}(x)$ for small x, so it is imperative to maintain a large step-size to avoid stalling the algorithm.

This scaling issue is at the heart of the *dual extrapolation* (DE) method of Nesterov [39]. Originally designed to solve variational inequalities and related problems, the method proceeds by (*i*) using a prox-step to generate the method's leading state and get a "look-ahead" gradient query; (*ii*) aggregating gradient information with a *constant* weight; and, finally, (*iii*) using a "primal-dual" mirror map to update the method's base state. Formally, the algorithm follows the iterative update rule

$$\begin{aligned} X_{t+1/2} &= P_{X_t}(-\gamma_t g_t) \\ Y_{t+1} &= Y_t - g_{t+1/2} \\ X_{t+1} &= Q(\gamma_{t+1} Y_{t+1}) \end{aligned} \tag{DE}$$

where the so-called "mirror map" $Q\colon \mathcal{Y} \to \mathcal{X}$ is defined as

$$Q(y) = \arg\max_{x\in\mathcal{X}} \{\langle y, x\rangle - h(x)\}. \tag{18}$$

Unfortunately, the template (DE) is not sufficient for our purposes, for two main reasons: First, the method still couples a prox-step with a variable (decreasing) step-size update; this is not problematic for the application of the method to VIs (where the achievable rates are different), but it is not otherwise favorable for acceleration.

In addition to the above, the method's gradient pre-multiplier is the same as its post-multiplier (γ_t in both cases), and it is not possible to differentiate these parameters while maintaining optimal rates [39]. However, this differentiation is essential for acceleration, especially when γ_t cannot be tuned with prior knowledge of the problem's parameters.

Our approach to overcome this issue consists of: *a*) eliminating the prox-step altogether in favor of a mirror step; and *b*) separating the weights used for introducing new gradients to the algorithm versus those used to generate the base and leading states. To formalize this, we introduce below the

[5] Strictly speaking this regularizer is not strongly convex over $[0,\infty)$ but this detail is not relevant for the question at hand.

Algorithm 1: Universal dual extrapolation with reweighted gradients (UNDERGRAD)

Parameters $a \leftarrow \sqrt{K_h}$; $b \leftarrow \sqrt{K_h(R_h + K_h\|\mathcal{X}\|^2)}$
Initialize $Y_1 \leftarrow 0$; $Z_1 \leftarrow 0$; $S_1 \leftarrow a^2$
for $t = 1, 2, \dots, T$ **do**
- $\eta_t \leftarrow b/\sqrt{S_t}$ // set learning rate
- $X_t \leftarrow Q(\eta_t Y_t)$ // mirror step
- $\bar{X}_t \leftarrow (\alpha_t X_t + Z_t)/\sum_{s=1}^t \alpha_s$ // mixing
- $g_t \leftarrow \mathsf{G}(\bar{X}_t; \omega_t)$ // oracle query
- $Y_{t+1/2} \leftarrow Y_t - \alpha_t g_t$ // dual step
- $X_{t+1/2} \leftarrow Q\big(\eta_t Y_{t+1/2}\big)$ // mirror step
- $\bar{X}_{t+1/2} \leftarrow \big(\alpha_t X_{t+1/2} + Z_t\big)/\sum_{s=1}^t \alpha_s$ // mixing
- $g_{t+1/2} \leftarrow \mathsf{G}(\bar{X}_{t+1/2}; \omega_{t+1/2})$ // oracle query
- $Y_{t+1} \leftarrow Y_t - \alpha_t g_{t+1/2}$ // dual step
- $S_{t+1} \leftarrow S_t + \alpha_t^2 \|g_{t+1/2} - g_t\|_*^2$ // precondition
- $Z_{t+1} \leftarrow Z_t + \alpha_t X_{t+1/2}$ // update mixing state

return $\bar{x}_T \leftarrow \bar{X}_{T+1/2}$

"universal" dual extrapolation template:

$$\begin{aligned} Y_{t+1/2} &= Y_t - \alpha_t g_t & X_{t+1/2} &= Q(\eta_t Y_{t+1/2}) \\ Y_{t+1} &= Y_t - \alpha_t g_{t+1/2} & X_{t+1} &= Q(\eta_{t+1} Y_{t+1}) \end{aligned} \tag{UDE}$$

In the above, the gradient signals g_t and $g_{t+1/2}$ are considered generic and the query points are not specified. To get a concrete algorithm, we will use the weighting scheme of Kavis et al. [24] and query the oracle at the averaged states $\bar{X}_t$ and $\bar{X}_{t+1/2}$ introduced previously in (13). Finally, regarding the algorithm's gradient weighting and averaging parameters (α_t and η_t respectively), we will use an *increasing* weight for the method's step-size $\alpha_t = t$ and the adaptive rule

$$\eta_t = \frac{b}{\sqrt{a^2 + \sum_{s=1}^{t-1} \alpha_s^2 \|g_{s+1/2} - g_s\|_*^2}} \tag{19}$$

for the method's learning rate (the parameters a and b are discussed below, and we are using the standard convention that empty sums are taken equal to zero).

The resulting method – which we call *universal dual extrapolation with reweighted gradients* (UNDERGRAD) – is encoded in pseudocode form in Algorithm 1 and represented schematically in Fig. 1. Its main guarantees are as follows:

Theorem 1. *Suppose that* UNDERGRAD (*Algorithm 1*) *is run for T iterations with η_t given by (19), $\alpha_t = t$ for all $t = 1, 2, \dots$, and $a = \sqrt{K_h}$, $b = C_h\sqrt{K_h}$ with $C_h = \sqrt{R_h + K_h\|\mathcal{X}\|^2}$. Then the algorithm's output state $\bar{x}_T \equiv \bar{X}_{T+1/2}$ simultaneously enjoys the following guarantees:*

a) If f satisfies (BG), then

$$\mathbb{E}[f(\bar{x}_T)] \leq \min f + 2C_h\sqrt{\frac{K_h + 8(G^2+\sigma^2)}{K_h T}} \tag{20a}$$

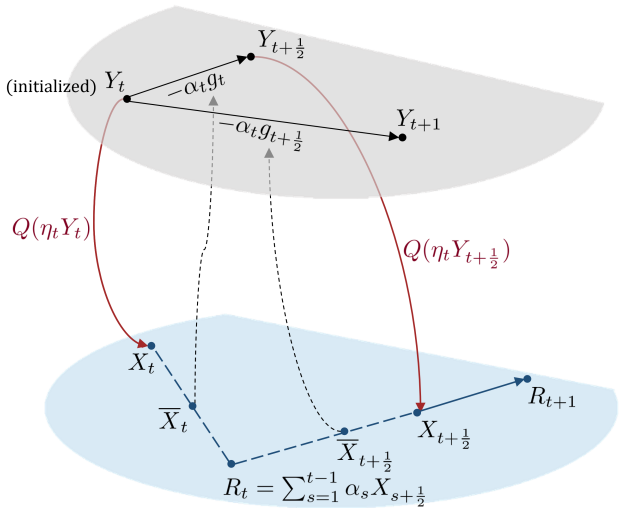

Figure 1: Schematic representation of the UNDERGRAD algorithm (Algorithm 1). The light blue area represents the problem's domain ($\mathcal{X}$), whereas the grey area represents the dual space ($\mathcal{Y}$).

b) If f satisfies (LG), then

$$\mathbb{E}[f(\bar{x}_T)] \leq \min f + \frac{32\sqrt{2}C_h^2 L}{K_h T^2} + \frac{8\sqrt{2}C_h\sigma}{\sqrt{K_h T}} \quad (20b)$$

Theorem 1 is our main result so, before discussing its proof (which we carry out in detail in the appendix), some remarks are in order.

The first point of note concerns the dependence of the anytime bounds (20) on the problem's dimensionality. To that end, let $C_{\text{fast}} = C_h^2$ and $C_{\text{slow}} = C_h$, so UNDERGRAD's rate of convergence scales as $\mathcal{O}(C_{\text{fast}}\chi^2/T^2)$ in smooth, deterministic problems, and as $\mathcal{O}(C_{\text{slow}}\chi/\sqrt{T})$ in non-smooth and/or stochastic environments. Thus, to compare the algorithm's rate of convergence to that of mirror-prox and UNIXGRAD (and up to universal constants), we have to compare C_h to R_h and B_h respectively.

To that end, we calculate below the values of C_{fast} and C_{slow} in the three archetypal examples of Section 3:

1. In the simplex setup of Example 1, we have $R_h = \log d$, $\|\mathcal{X}\| = 1$ and $K_h = 1$, so $C_{\text{slow}} = \mathcal{O}(\sqrt{\log d})$ and $C_{\text{fast}} = \mathcal{O}(\log d)$.
2. In the spectrahedron setup of Example 2, we have again $R_h = \log d$, $\|\mathcal{X}\| = 1$ and $K_h = 1$, so $C_{\text{slow}} = \mathcal{O}(\sqrt{\log d})$ and $C_{\text{fast}} = \mathcal{O}(\log d)$. [For a detailed discussion, see [9, 23, 34] and references therein.]
3. Finally, in the combinatorial setup of Example 3, we have $R_h = m(1 + \log(d/m))$, $\|\mathcal{X}\| = m$ and $K_h = 1$ [27]. Thus, if $m = \mathcal{O}(1)$ in d, we get again $C_{\text{slow}} = \mathcal{O}(\sqrt{\log d})$ and $C_{\text{fast}} = \mathcal{O}(\log d)$.

The above shows that UNDERGRAD achieves the desired almost dimension-free rates of the *non-adaptive* mirror-prox algorithm, as well as the *universal* order-optimal guarantees of UNIXGRAD. The only discrepancy with the rates presented in Table 1 is the additive constant K_h that appears in the numerator of (20a): this constant is an artifact of the analysis and it only becomes relevant when $G \to 0$ and $\sigma \to 0$. Since the scaling of the algorithm's convergence rate concerns the large G regime, this difference is not relevant for our purposes.

An additional difference between UNDERGRAD and UNIXGRAD is that the latter involves the prox-mapping (7), whereas the former involves the mirror map (18). To compare the two in terms of their per-iteration complexity, note that if we apply the prox-mapping (7) to the prox-center $x_c \leftarrow \arg\min h$ of $\mathcal{X}$, we get

$$\begin{aligned} P_{x_c}(y) &= \arg\min_{x\in\mathcal{X}}\{\langle y, x_c - x\rangle + D(x, x_c)\} \\ &= \arg\min_{x\in\mathcal{X}}\{h(x) - \langle \nabla h(x_c) + y, x\rangle\} \\ &= Q(\nabla h(x_c) + y) \end{aligned} \quad (21)$$

so, in particular, $Q(y) = P_{x_c}(y)$ whenever $x_c \in \operatorname{ri}\mathcal{X}$ (which is the case for most regularizers used in practice, including the Legendre regularizers used in Examples 1–3 above). This shows that the calculation of the mirror map $Q(y) = P_{x_c}(y - \nabla h(x_c))$ is at least as simple as the calculation of the prox-mapping $P_x(y)$ for a general base point $x \in \mathcal{X}$ – and, typically, calculating the mirror map is strictly lighter because there is no need to vary the base point over different iterations of the algorithm. In this regard, the per-iteration overhead of (UDE) is actually *lighter* compared to (MP) or (DE).

Finally, we should note that all our results above implicitly assume that the problem's domain is bounded (otherwise the range parameter R_h of the problem becomes infinite). Thus, in addition to these convergence properties of UNDERGRAD, we also provide below an asymptotic guarantee for problems with an *unbounded* domain:

Theorem 2. *Suppose that* UNDERGRAD *is run with perfect oracle feedback with η_t given by (19) and $\alpha_t = t$. If f satisfies (LG), the algorithm's output state $\bar{x}_T = \bar{X}_{T+1/2}$ enjoys the rate $f(\bar{x}_T) - \min f = \mathcal{O}(1/T^2)$.*

This result provides an important extension of Theorem 1 to problems with unbounded domains. It remains an open question for the future to derive the precise constants in the convergence rate presented in Theorem 2.

Main ideas of the proof. The detailed proof of Theorem 1 is fairly long so we defer it to the appendix and only present here the main ideas.

The main ingredient of our proof is a specific *template inequality* used to derive an "appropriate" upper bound of the term $\tilde{\mathcal{R}}_T(x) := \sum_{t=1}^T \alpha_t \langle g_{t+1/2}, X_{t+1/2} - x\rangle$.

Importantly, to prove the dimension-free properties of UNDERGRAD, such an upper-bound *cannot involve Bregman divergences*: even though this is common practice in previous papers [1, 24], these terms would ultimately lead to the Bregman diameter B_h that we seek to avoid. This is a principal part of the reason for switching gears to the DE template for UNDERGRAD: in so doing, we are able to employ the notion of the *Fenchel coupling* [31, 32], which is a "primal-dual distance" as opposed to the Bregman divergence which is a "primal-primal distance" (cf. Appendix A.1). This poses another challenge on connecting the Fenchel coupling of targeted points before and after a mirror step, for which we need to employ a primal-dual version of the "three-point identity" (Lemma A.3). These elements lead to the following proposition:

Proposition 1. *For all $x \in \mathcal{X}$, we have*

$$\tilde{\mathcal{R}}_T(x) \leq \frac{R_h}{\eta_{T+1}} + \sum_{t=1}^{T} \alpha_t \langle g_{t+1/2} - g_t, X_{t+1/2} - X_{t+1} \rangle - K_h \sum_{t=1}^{T} \frac{\|X_{t+1} - X_{t+1/2}\|^2 + \|X_{t+1/2} - X_t\|^2}{2\eta_t} \tag{22}$$

With (22) in hand, (20a) comes from the application of the Fenchel-Young inequality to upper-bound the right-hand-side of (22) as $\sum_{t=1}^{T} \alpha_t^2 \eta_{t+1} \|g_{t+1/2} - g_t\|_*$ (plus a constant term involving $\|\mathcal{X}\|$). The challenge here is to notice and successfully prove that this summation is actually upper-bounded by η_{T+1}^{-1} (due to our special choice of the learning rate update). Finally, by its definition, η_{T+1}^{-1} can be bounded by G, σ and K_h as described in the statement of Theorem 1.

The proof of (20b) is far more complex. The main challenge is to manipulate the terms in (22) to derive an upper-bound of the form $\sum_{t=1}^{T} \alpha_t^2 g(\eta_{t+1}) \|\nabla f(\bar{X}_{t+1/2}) - \nabla f \bar{X}_t\|_*^2$ (plus a term involving the noise level σ) where $g(\eta_{t+1})$ is a function of the learning rate chosen such that only the first $T_0 \ll T$ elements of this summation are positive. Once this has been achieved, the quantity $\|\nabla f(\bar{X}_{t+1/2}) - \nabla f \bar{X}_t\|_*$ is connected to $\|\mathcal{X}\|$ via (LG) and our claim is obtained.

5. Numerical Experiments

For the experimental validation of our results, we focus on the simplex setup of Example 1 with linear losses and $d = 100$. Our first experiment concerns the *perfect* SFO case and tracks down the convergence properties of UNDERGRAD run with the *entropic regularizer* adapted to the simplex. As a baseline, we ran UNIXGRAD, also with the entropic regularizer. A first challenge here is that the Bregman diameter B_h of the simmplex is infinite, so UNIXGRAD is not well-defined. On that account, we choose the step-size update rule of UNIXGRAD such that its initial step-size γ_1 equals the initial learning rate η_1 of UNDERGRAD. We also ran UNIXGRAD with the update rule such that γ_1 is smaller or larger than η_1. Finally, for comparison purposes, we also present a non-adaptive *accelerated entropic gradient* (AEG) algorithm, and we report the results in Fig. 2.

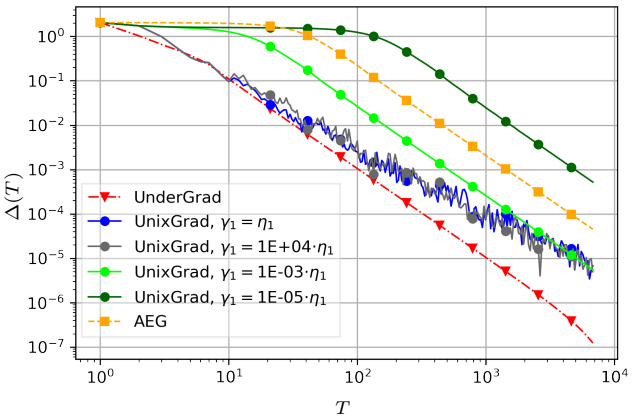

Figure 2: Convergence of UNDERGRAD and UNIXGRAD in the simplex setup with a perfect SFO. The y-axis corresponds to the differences between the f-value of the relevant point of each algorithm and $\min f$. The code is available at https://github.com/dongquan-vu/UnDerGrad_Universal_CnvOpt_ICML2022.

Fig. 2 confirms that UNDERGRAD successfully converges with an accelerated rate of $\mathcal{O}(1/T^2)$. Perhaps surprisingly, it also shows that when UNIXGRAD's initial step-size is small (10E-3 or smaller), UNIXGRAD also achieves an $\mathcal{O}(1/T^2)$, but at a much more conservative pace, trailing UNDERGRAD by one or two orders of magnitude. On the other hand, when UNIXGRAD's initial step-size is of the same magnitude as the UNDERGRAD's learning rate (or larger), UNIXGRAD eventually destabilizes and its rate drops from $\mathcal{O}(1/T^2)$ to approximately $\mathcal{O}(1/T)$. We also conducted experiments in the setup with a noisy SFO; these are reported in Appendix C.

Acknowledgments

PM is grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grant ALIAS (ANR-19-CE48-0018-01). KYL is grateful for financial support by Israel Science Foundation (grant No. 447/20), by Israel PBC-VATAT, and by the Technion Center for Machine Learning and Intelligent Systems (MLIS). VC and KA were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data) and from the Swiss National Science Foundation (SNSF) under grant number 200021_205011.

References

[1] Allen-Zhu, Z. and Orecchia, L. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017. Full version available at http://arxiv.org/abs/1407.1537.

[2] Antonakopoulos, K. and Mertikopoulos, P. Adaptive first-order methods revisited: Convex optimization without Lipschitz requirements. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[3] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. An adaptive mirror-prox algorithm for variational inequalities with singular operators. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[4] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[5] Antonakopoulos, K., Pethick, T., Kavis, A., Mertikopoulos, P., and Cevher, V. Sifting through the noise: Universal first-order methods for stochastic variational inequalities. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[6] Auer, P., Cesa-Bianchi, N., and Gentile, C. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[7] Bach, F. and Levy, K. Y. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *COLT '19: Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

[8] Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

[9] Bilenne, O., Mertikopoulos, P., and Belmega, E. V. Fast optimization with zeroth-order feedback in distributed multi-user MIMO systems. *IEEE Trans. Signal Process.*, 68:6085–6100, October 2020.

[10] Boyd, S. P. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[11] Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

[12] Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[13] Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[14] Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 78:1404–1422, 2012.

[15] Chen, G. and Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993.

[16] Cutkosky, A. Anytime online-to-batch, optimism and acceleration. In *ICML '19: Proceedings of the 36th International Conference on Machine Learning*, 2019.

[17] Ene, A., Nguyen, H. L., and Vladu, A. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *AAAI '21: Proceedings of the 35th Conference on Artificial Intelligence*, 2021.

[18] Facchinei, F. and Pang, J.-S. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

[19] György, A., Linder, T., Lugosi, G., and Ottucsák, G. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007.

[20] Hsieh, Y.-G., Antonakopoulos, K., and Mertikopoulos, P. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *COLT '21: Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

[21] Joulani, P., Raj, A., György, A., and Szepesvári, C. A simpler approach to accelerated stochastic optimization: Iterative averaging meets optimism. In *ICML '20: Proceedings of the 37th International Conference on Machine Learning*, 2020.

[22] Juditsky, A., Nemirovski, A. S., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[23] Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13:1865–1890, 2012.

[24] Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. UnixGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[25] Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

[26] Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, June 2012.

[27] Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

[28] Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[29] Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[30] McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. In *COLT '10: Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

[31] Mertikopoulos, P. and Sandholm, W. H. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, November 2016.

[32] Mertikopoulos, P. and Staudigl, M. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, January 2018.

[33] Mertikopoulos, P. and Zhou, Z. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

[34] Mertikopoulos, P., Belmega, E. V., Negrel, R., and Sanguinetti, L. Distributed stochastic optimization via matrix exponential learning. *IEEE Trans. Signal Process.*, 65(9): 2277–2290, May 2017.

[35] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

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

[37] Nemirovski, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983.

[38] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[39] Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[40] Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2): 381–404, 2015.

[41] Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

[42] Rakhlin, A. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013.

[43] Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[44] Telatar, I. E. Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications and Related Technologies*, 10(6):585–596, 1999.

[45] Vu, D. Q., Antonakopoulos, K., and Mertikopoulos, P. Fast routing under uncertainty: Adaptive learning in congestion games with exponential weights. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[46] Ward, R., Wu, X., and Bottou, L. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. In *ICML '19: Proceedings of the 36th International Conference on Machine Learning*, 2019.

[47] Yu, W., Rhee, W., Boyd, S. P., and Cioffi, J. M. Iterative water-filling for Gaussian vector multiple-access channels. *IEEE Trans. Inf. Theory*, 50(1):145–152, 2004.

A. Bregman regularizers and several preliminary results

A.1. Bregman regularizers and their properties

We begin by clarifying and recalling some of the notational convetions used throughout the paper. We also give the formal definition of the Fenchel coupling (a key notion for the proof our main results) and we present some preliminary results to prepare the ground for the sequel.

The convex conjugate $h^*\colon \mathcal{Y} \to \mathbb{R}$ of h is then defined as

$$h^*(y) = \sup_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}. \tag{A.1}$$

As a result, the supremum in (A.1) is always attained, and $h^*(y)$ is finite for all $y \in \mathcal{Y}$ [8]. Moreover, by standard results in convex analysis [43, Chap. 26], h^* is differentiable on $\mathcal{Y}$ and its gradient satisfies the identity

$$\nabla h^*(y) = \arg\max_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}. \tag{A.2}$$

Thus, recalling the definition of the mirror map $Q\colon \mathcal{Y} \to \mathcal{X}$, we readily get

$$Q(y) = \nabla h^*(y). \tag{A.3}$$

Lemma A.1. *Let h be a Bregman regularizer on $\mathcal{X}$. Then, for all $x, \in^+ \operatorname{dom}\partial h$ and all $y, v \in \mathcal{Y}$, we have:*

$$a)\ \ x = Q(y) \iff y \in \partial h(x). \tag{A.4a}$$

$$b)\ \ x^+ = Q(\nabla h(x) + v) \iff \nabla h(x) + v \in \partial h(x^+) \tag{A.4b}$$

Finally, if $x = Q(y)$ and $x^* \in \mathcal{X}$, we have

$$\langle \nabla h(x), x - x^*\rangle \leq \langle y, x - x^*\rangle. \tag{A.5}$$

Proof of Lemma A.1. To prove (A.4a), note that x solves (A.2) if and only if $y - \partial h(x) \ni 0$, i.e., if and only if $y \in \partial h(x)$. Eq. (A.4b) is then obtained in the same manner.

For the inequality (A.5), it suffices to show it holds for all $x^* \in \mathcal{X}_h \equiv \operatorname{dom}\partial h$ (by continuity). To do so, let

$$\phi(t) = h(x + t(x^* - x)) - [h(x) + \langle y, x + t(x^* - x)\rangle]. \tag{A.6}$$

Since h is strongly convex relative to g and $y \in \partial h(x)$ by (A.4a), it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Moreover, note that $\psi(t) = \langle \nabla h(x + t(x^* - x)) - y, x^* - x\rangle$ is a continuous selection of subgradients of ϕ. Given that ϕ and ψ are both continuous on $[0, 1]$, it follows that ϕ is continuously differentiable and $\phi' = \psi$ on $[0, 1]$. Thus, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0, 1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, x^* - x\rangle \geq 0$. ∎

As we mentioned earlier, much of our analysis revolves around a "primal-dual" divergence between a target point $x^* \in \mathcal{X}$ and a dual vector $y \in \mathcal{Y}$, called the *Fenchel coupling*. Following [33], this is defined as follows for all $x^* \in \mathcal{X}$, $y \in \mathcal{Y}$:

$$F(x^*, y) = h(x^*) + h^*(y) - \langle y, x^*\rangle. \tag{A.7}$$

The following lemma illustrates some basic properties of the Fenchel coupling:

Lemma A.2. *Let h be a Bregman regularizer on $\mathcal{X}$ with convexity modulus K_h. Then, for all $x^* \in \mathcal{X}$ and all $y \in \mathcal{Y}$, we have:*

1. *$F(x^*, y) = D(x^*, Q(y))$ if $Q(y) \in \mathcal{X}_h$ (but not necessarily otherwise).*
2. *$F(x^*, y) \geq \frac{K_h}{2}\|Q(y) - x^*\|^2$*

Proof. For our first claim, let $x = Q(y)$. Then, by definition we have:

$$F(x^*, y) = h(x^*) - \langle y, Q(y)\rangle - h(Q(y)) - \langle y, x^*\rangle = h(x^*) - h(x) - \langle y, x^* - x\rangle. \tag{A.8}$$

Since $y \in \partial h(x)$, we have $h'(x; x^* - x) = \langle y, x^* - x\rangle$ whenever $x \in \mathcal{X}_h$, thus proving our first claim. For our second claim, working in the previous spirit we get that:

$$F(x^*, y) = h(x^*) - h(x) - \langle y, x^* - x\rangle \tag{A.9}$$

Thus, we obtain the result by recalling the strong convexity assumption for h with respect to the respetive norm $\|\cdot\|$. ■

We continue with some basic relations connecting the Fenchel coupling relative to a target point before and after a gradient step. The basic ingredient for this is a primal-dual analogue of the so-called "three-point identity" for Bregman functions [15]:

Lemma A.3. *Let h be a Bregman regularizer on $\mathcal{X}$. Fix some $x^* \in \mathcal{X}$ and let $y, y^+ \in \mathcal{Y}$. Then, letting $x = Q(y)$, we have*

$$F(x^*, y^+) = F(x^*, y) + F(x, y^+) + \langle y^+ - y, x - x^*\rangle. \tag{A.10}$$

Proof. By definition, we get:

$$\begin{aligned} F(x^*, y^+) &= h(x^*) + h^*(y^+) - \langle y^+, x^*\rangle \\ F(x^*, y) &= h(x^*) + h^*(y) - \langle y, x^*\rangle. \end{aligned} \tag{A.11}$$

Then, by subtracting the above we get:

$$\begin{aligned} F(x^*, y^+) - F(x^*, y) &= h(x^*) + h^*(y^+) - \langle y^+, x^*\rangle - h(x^*) - h^*(y) + \langle y, x^*\rangle \\ &= h^*(y^+) - h^*(y) - \langle y^+ - y, x^*\rangle \\ &= h^*(y^+) - \langle y, Q(y)\rangle + h(Q(y)) - \langle y^+ - y, x^*\rangle \\ &= h^*(y^+) - \langle y, x\rangle + h(x) - \langle y^+ - y, x^*\rangle \\ &= h^*(y^+) + \langle y^+ - y, x\rangle - \langle y^+, x\rangle + h(x) - \langle y^+ - y, x^*\rangle \\ &= F(x, y^+) + \langle y^+ - y, x - x^*\rangle \end{aligned} \tag{A.12}$$

and our proof is complete. ■

A.2. Numerical sequence inequalities

In this section, we provide some necessary inequalities on numerical sequences that we require for the convergence rate analysis of the previous sections. Most of the lemmas presented below already exist in the literature, and go as far back as Auer et al. [6] and McMahan & Streeter [30]; when appropriate, we note next to each lemma the references with the statement closest to the precise version we are using in our analysis.

Lemma A.4 (30, 28)**.** *For all non-negative numbers $a_1, \dots a_t$, the following inequality holds:*

$$\sqrt{a^2 + \sum_{t=1}^{T} a_t} \leq a + \sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}} \leq 2\sqrt{a^2 + \sum_{t=1}^{T} a_t} \tag{A.13}$$

B. Analysis and proofs of the main results

The proof of the template inequality. We first prove the template inequality of UNDERGRAD; this is the primary element of our proof of Theorem 1:

Proposition 1. *For all $x \in \mathcal{X}$, we have*

$$\begin{aligned} \tilde{\mathcal{R}}_T(x) \leq \frac{R_h}{\eta_{T+1}} &+ \sum_{t=1}^{T} \alpha_t \langle g_{t+1/2} - g_t, X_{t+1/2} - X_{t+1}\rangle \\ &- K_h \sum_{t=1}^{T} \frac{\|X_{t+1} - X_{t+1/2}\|^2 + \|X_{t+1/2} - X_t\|^2}{2\eta_t} \end{aligned} \tag{22}$$

Proof. First, we set $\tilde{Y}_t = \eta_t Y_t$. For all $x \in \mathcal{X}$ we have:

$$\begin{aligned}
&\alpha_t\langle g_{t+1/2}, X_{t+1} - x\rangle\\
=&\langle \frac{1}{\eta_t}\tilde{Y}_t - \frac{1}{\eta_{t+1}}\tilde{Y}_{t+1}, X_{t+1} - x\rangle\\
=&\frac{1}{\eta_t}\langle \tilde{Y}_t - \tilde{Y}_{t+1}, X_{t+1} - x\rangle + \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right]\langle 0 - \tilde{Y}_{t+1}, X_{t+1} - x\rangle\\
=&\frac{1}{\eta_t}\left[F(x,\tilde{Y}_t) - F(x,\tilde{Y}_{t+1}) - F(X_{t+1},\tilde{Y}_t)\right] + \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right]\Big(F(x,0) - F(x,\tilde{Y}_{t+1}) - F(X_{t+1},0)\Big)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \#\ \textit{from Lemma A.3}\\
\le&\frac{1}{\eta_t}F(x,\tilde{Y}_t) - \frac{1}{\eta_{t+1}}F(x,\tilde{Y}_{t+1}) + \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right]R_h - \frac{1}{\eta_t}F(X_{t+1},\tilde{Y}_t). \qquad \text{(B.1)}
\end{aligned}$$

Here, the last inequality comes from the facts that $F(x,0) = h(x) - h(Q(0)) = h(x) - \min_{x\in\mathcal{X}} h \le R_h$ and $F(\cdot,\cdot) \ge 0$ and that η_t is decreasing.

As a consequence of (B.1), we have:

$$\begin{aligned}
&\alpha_t\langle g_{t+1/2}, X_{t+1/2} - x\rangle\\
=&\alpha_t\langle g_{t+1/2}, X_{t+1/2} - X_{t+1}\rangle + \alpha_t\langle g_{t+1/2}, X_{t+1} - x\rangle\\
\le&\alpha_t\langle g_{t+1/2}, X_{t+1/2} - X_{t+1}\rangle + \frac{1}{\eta_t}F(x,\tilde{Y}_t) - \frac{1}{\eta_{t+1}}F(x,\tilde{Y}_{t+1}) + \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right]R_h - \frac{1}{\eta_t}F(X_{t+1},\tilde{Y}_t) \qquad \text{(B.2)}
\end{aligned}$$

On the other hand, let us define $\tilde{Y}_{t+1/2} \coloneqq \eta_t Y_{t+1/2}$, we have

$$\begin{aligned}
&\alpha_t\langle g_t, X_{t+1/2} - X_{t+1}\rangle = \frac{1}{\eta_t}\langle \tilde{Y}_t - \tilde{Y}_{t+1/2}, X_{t+1/2} - X_{t+1}\rangle = \frac{1}{\eta_t}\Big[F(X_{t+1},\tilde{Y}_t) - F(X_{t+1},\tilde{Y}_{t+1/2}) - F(X_{t+1/2},\tilde{Y}_t)\Big]\\
\Rightarrow&\frac{1}{\eta_t}F(X_{t+1},\tilde{Y}_t) = \alpha_t\langle g_t, X_{t+1/2} - X^{t+1}\rangle + \frac{1}{\eta_t}F(X_{t+1},\tilde{Y}_{t+1/2}) + \frac{1}{\eta_t}F(X_{t+1/2},\tilde{Y}_t). \qquad \text{(B.3)}
\end{aligned}$$

Replace (B.3) into (B.2), we get:

$$\begin{aligned}
&\alpha_t\langle g_{t+1/2}, X_{t+1/2} - x\rangle\\
\le&\alpha_t\langle g_{t+1/2} - g_t, X_{t+1/2} - X_{t+1}\rangle + \frac{1}{\eta_t}F(x,\tilde{Y}_t) - \frac{1}{\eta_{t+1}}F(x,\tilde{Y}_{t+1}) - \frac{1}{\eta_t}F(X_{t+1},\tilde{Y}_{t+1/2})\\
&\qquad - \frac{1}{\eta_t}F(X_{t+1/2},\tilde{Y}_t) + \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right]R_h. \qquad \text{(B.4)}
\end{aligned}$$

Now, recall the definitions $X_{t+1/2} = Q(\tilde{Y}_{t+1/2})$ and $X_{t+1} = Q(Y_{t+1})$ in Algorithm 1, apply Lemma A.2 to $F(X_{t+1},\tilde{Y}_{t+1/2})$ and $F(X_{t+1/2},\tilde{Y}_t)$ then combine with (B.4), we get:

$$\begin{aligned}
\alpha_t\langle g_{t+1/2}, X_{t+1/2} - x\rangle \le&\ \alpha_t\langle g_{t+1/2} - g_t, X_{t+1/2} - X_{t+1}\rangle + \frac{1}{\eta_t}F(x,\tilde{Y}_t) - \frac{1}{\eta_{t+1}}F(x,\tilde{Y}_{t+1})\\
&- \frac{K_h}{2\eta_t}\|X_{t+1} - X_{t+1/2}\|^2 - \frac{K_h}{2\eta_t}\|X_{t+1/2} - X_t\|^2 + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)R_h \qquad \text{(B.5)}
\end{aligned}$$

Hence, after telescoping $t = 1,\dots,T$ and recalling the notation $\tilde{\mathcal{R}}_T(x) \coloneqq \sum_{t=1}^T \alpha_t\langle g_{t+1/2}, X_{t+1/2} - x\rangle$, we get:

$$\tilde{\mathcal{R}}_T(x) \le \frac{1}{\eta_1}F(x,\tilde{Y}_1) + \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_1}\right)R_h$$

$$+\sum_{t=1}^{T}\alpha_t\langle g_{t+1/2}-g_t, X_{t+1/2}-X_{t+1}\rangle-\sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1}-X_{t+1/2}\|^2-\sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1/2}-X_t\|^2 \quad \text{(B.6)}$$

Finally, by our initial choice of $Y_1 = 0$, we have $F(x, \tilde{Y}_1) = h(x) - \min_{x\in\mathcal{X}} h(x) \le R_h$ and (22) follows (B.6). This concludes the proof of Proposition 1. ■

Regret-to-rate conversion lemma. The next element in our proof is the following lemma that will be used to connect the term $\tilde{\mathcal{R}}_T(x)$ (which, in intuition, is similar to a regret term) and the term $\mathbb{E}\big[f(\bar{X}_{T+1/2}) - \min f\big]$ whose bounds will characterize the convergence rate of UNDERGRAD.

Lemma B.1. *For any $x^* \in \mathcal{X}^*$, for any T, we have:*

$$\mathbb{E}\big[f(\bar{X}_{T+1/2}) - \min f\big] \le \mathbb{E}\left[\frac{2}{T^2}\sum_{t=1}^{T}\alpha_t\langle\nabla f\big(\bar{X}_{t+1/2}\big), X_{t+1/2}-x^*\rangle\right] = \frac{2}{T^2}\,\mathbb{E}\Big[\tilde{\mathcal{R}}_T(x^*)\Big]. \quad \text{(B.7)}$$

Note that a version of Lemma B.1 appears previously in [16, 24]; for the sake of completeness, we provide its proof below.

Proof. Let us denote $H_t := \sum_{s=1}^{t}\alpha_s$. From the update rule of Algorithm 1, we have can rewrite $X_{t+1/2} = \frac{H_t}{\alpha_t}\bar{X}_{t+1/2} - \frac{H_{t-1}}{\alpha_t}\bar{X}_{t-1/2}$. Therefore,

$$X_{t+1/2}-x^* = \frac{H_t}{\alpha_t}(\bar{X}_{t+1/2}-x^*) - \frac{H_{t-1}}{\alpha_t}(\bar{X}_{t-1/2}-x^*) = \frac{1}{\alpha_t}\big[\alpha_t(\bar{X}_{t+1/2}-x^*) + H_{t-1}(\bar{X}_{t+1/2}-\bar{X}_{t-1/2})\big]. \quad \text{(B.8)}$$

As a consequence, we have:

$$\begin{aligned}
&\sum_{t=1}^{T}\alpha_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2}-x^*\rangle\\
=&\sum_{t=1}^{T}\big[\alpha_t\big\langle\nabla f(\bar{X}_{t+1/2}), \bar{X}_{t+1/2}-x^*\big\rangle + H_{t-1}\big\langle\nabla f(\bar{X}_{t+1/2}), \bar{X}_{t+1/2}-\bar{X}_{t-1/2}\big\rangle\big]\\
\ge&\sum_{t=1}^{T}\alpha_t\big[f(\bar{X}_{t+1/2}) - f(x^*)\big] + \sum_{t=1}^{T}H_{t-1}\big[f(\bar{X}_{t+1/2}) - f(\bar{X}_{t-1/2})\big]\\
=&\sum_{t=1}^{T}\alpha_t\big[f(\bar{X}_{t+1/2}) - f(x^*)\big] + \sum_{t=1}^{T-1}\alpha_t\big[f(\bar{X}_{T+1/2}) - f(\bar{X}_{t+1/2})\big] \qquad \#\ \textit{since } H_t - H_{t-1} = \alpha_t\\
=&\big[f(\bar{X}_{T+1/2}) - f(x^*)\big]\sum_{t=1}^{T}\alpha_t. \qquad \text{(B.9)}
\end{aligned}$$

Divide two sides of (B.9) by $H_t > 0$ and choose α_t such that $H_t > \frac{T^2}{2}$ (for example, choose $\alpha_t = \alpha$), we obtain that:

$$\frac{2}{T^2}\sum_{t=1}^{T}\alpha_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2}-x^*\rangle \ge f(\bar{X}_{T+1/2}) - f(x^*) = f(\bar{X}_{T+1/2}) - \min f. \quad \text{(B.10)}$$

Finally, we recall that by definition, $g_{t+1/2} = \mathsf{G}(\bar{X}_{t+1/2}; \omega_{t+1/2}) = \nabla f(\bar{X}_{t+1/2}) + U_{t+1/2}$ where $\mathbb{E}\big[U_{t+1/2}\,\big|\,\mathcal{F}_{t+1/2}\big] = 0$. Therefore, by the law of total expectation, we have:

$$\tilde{\mathcal{R}}_T(x^*) = \mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2}-x^*\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\langle U_{t+1/2}, X_{t+1/2}-x^*\rangle\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^* \rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \, \mathbb{E}\big[\langle U_{t+1/2}, X_{t+1/2} - x^* \rangle \,\big|\, \mathcal{F}_{t+1/2}\big]\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^* \rangle\right]. \tag{B.11}$$

Take the expectation of the two sides of (B.10) then combine it with (B.11), we conclude the proof. ■

Proof of (20a): convergence of UNDERGRAD under (LC)/(BG). Our starting point is Eq. (22) that we established in Proposition 1 that leads to the following inequality:

$$\tilde{\mathcal{R}}_T(x) \le \frac{R_h}{\eta_{T+1}} + \sum_{t=1}^{T} \alpha_t \langle g_{t+1/2} - g_t, X_{t+1/2} - X_{t+1} \rangle - \sum_{t=1}^{T} \frac{K_h}{2\eta_t} \|X_{t+1} - X_{t+1/2}\|^2 \tag{B.12}$$

We now focus on the second term in the right hand side of (B.12). From the Cauchy-Schwarz inequality and the fact that $\|Y - Y'\|_* \|X - X'\| = \min_{a>0} \left\{ \frac{1}{2a} \|Y - Y'\|_*^2 + \frac{a}{2} \|X - X'\|^2 \right\}$ for any $X, X', Y, Y' \in \mathbb{R}^d$,[6] we have:

$$\sum_{t=1}^{T} \alpha_t \langle g_{t+1/2} - g_t, X_{t+1/2} - X_{t+1} \rangle \le \sum_{t=1}^{T} \alpha_t \|g_{t+1/2} - g_t\|_* \|X_{t+1/2} - X_{t+1}\|$$
$$\le \frac{1}{2K_h} \sum_{t=1}^{T} \alpha_t^2 \eta_{t+1} \|g_{t+1/2} - g_t\|_*^2 + \frac{K_h}{2} \sum_{t=1}^{T} \frac{1}{\eta_{t+1}} \|X_{t+1} - X_{t+1/2}\|^2. \tag{B.13}$$

Moreover, from the definition of η_{t+1} and by applying Lemma A.4, we have:

$$\frac{1}{2K_h} \sum_{t=1}^{T} \alpha_t^2 \eta_{t+1} \|g_{t+1/2} - g_t\|_*^2 = \frac{b}{2K_h} \sum_{t=1}^{T} \frac{\alpha_t^2 \|g_{t+1/2} - g_t\|_*^2}{\sqrt{a^2 + \sum_{s=1}^{t} \|g_{s+1/2} - g_s\|_*^2}}$$
$$\le \frac{b}{K_h} \sqrt{a^2 + \sum_{t=1}^{T} \alpha_t^2 \|g_{t+1/2} - g_t\|_*^2} - \frac{b\sqrt{a^2}}{2K_h}$$
$$= \frac{b^2}{K_h \cdot \eta_{T+1}} - \frac{b\sqrt{a^2}}{2K_h}. \tag{B.14}$$

Combine (B.13) and (B.14) with (B.12) and by the compactness of the feasible region $\mathcal{X}$, we get:

$$\tilde{\mathcal{R}}_T(x) \le \frac{R_h}{\eta_{T+1}} + \frac{1}{2K_h} \sum_{t=1}^{T} \alpha_t^2 \eta_{t+1} \|g_{t+1/2} - g_t\|_*^2$$
$$+ \frac{K_h}{2} \sum_{t=1}^{T} \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right] \|X_{t+1} - X_{t+1/2}\|^2$$
$$\le \frac{R_h}{\eta_{T+1}} + \frac{b^2}{K_h \cdot \eta_{T+1}} - \frac{b\sqrt{a^2}}{2K_h} + \frac{K_h \|\mathcal{X}\|^2}{2} \sum_{t=1}^{T} \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right]$$
$$= \frac{1}{\eta_{T+1}} \left(R_h + \frac{b^2}{K_h} + \frac{K_h \|\mathcal{X}\|^2}{2}\right) - b\sqrt{a^2} \left(\frac{1}{2K_h} + \frac{K_h \|\mathcal{X}\|^2}{2}\right). \tag{B.15}$$

[6] This can be proved trivially: $a^* = \|Y - Y'\|_* \|X - X'\|$ is a minimizer of the function $\psi(a) = \frac{1}{2a} \|Y - Y'\|_*^2 + \frac{a}{2} \|X - X'\|^2$.

Hence, by invoking Lemma B.1, we have:

$$\mathbb{E}\left[f(\overline{X}_{T+1/2}) - \min_{x\in\mathcal{X}} f(x)\right] \leq \frac{2\,\mathbb{E}\left[\tilde{\mathcal{R}}_T(x^*)\right]}{T^2} = \frac{2}{T^2}\left[\left(R_h + \frac{b^2}{K_h} + \frac{K_h\|\mathcal{X}\|^2}{2}\right)\mathbb{E}\left[\frac{1}{\eta_{T+1}}\right] - b\sqrt{a^2}\left(\frac{1}{2K_h} + \frac{K_h\|\mathcal{X}\|^2}{2}\right)\right]. \tag{B.16}$$

On the other hand, by the definition of η_{T+1}, we get:

$$\mathbb{E}\left[\frac{1}{\eta_{T+1}}\right] \leq \frac{1}{b}\,\mathbb{E}\left[\sqrt{a^2 + \sum_{t=1}^{T}\alpha_t^2\|g_{t+1/2} - g_t\|_*^2}\right] \leq \frac{1}{b}\sqrt{a^2 + \sum_{t=1}^{T}\alpha_t^2\,\mathbb{E}\left[\|g_{t+1/2} - g_t\|_*^2\right]}. \tag{B.17}$$

Moreover, we have that:

$$\begin{aligned}
\mathbb{E}\left[\|g_{t+1/2} - g_t\|_*^2\right] &= \mathbb{E}\left[\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t) - (U_{t+1/2} - U_t)\|_*^2\right] \\
&\leq \mathbb{E}\left[2\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2 + 2\|U_{t+1/2} - U_t\|_*^2\right] \\
&\leq \mathbb{E}\left[4(\|\nabla f(\overline{X}_{t+1/2})\|_*^2 + \|\nabla f(\overline{X}_t)\|_*^2) + 4(\|U_{t+1/2}\|_*^2 + \|U_t\|_*^2)\right] \\
&\leq \mathbb{E}\left[8G^2 + 4(\|U_{t+1/2}\|_*^2 + \|U_t\|_*^2)\right] && \text{\# \textit{from} (BG)} \\
&= 8\left[G^2 + \sigma^2\right]. && \text{(B.18)}
\end{aligned}$$

Therefore, when we choose the step-size parameters $\alpha_t = t, \forall t$ as indicated in Theorem 1, we have:

$$\begin{aligned}
\left(R_h + \frac{b^2}{K_h} + \frac{K_h\|\mathcal{X}\|^2}{2}\right)\mathbb{E}\left[\frac{1}{\eta_{T+1}}\right] &\leq \left(R_h + \frac{b^2}{K_h} + \frac{K_h\|\mathcal{X}\|^2}{2}\right)\frac{1}{b}\sqrt{a^2 + 8\left(G^2+\sigma^2\right)\sum_{t=1}^{T}\alpha_t^2} \\
&\leq \left(\frac{R_h}{b} + \frac{b}{K_h} + \frac{K_h\|\mathcal{X}\|^2}{2b}\right)\sqrt{a^2 + 8(G^2+\sigma^2)T^3}. && \text{(B.19)}
\end{aligned}$$

Finally, substituting (B.19) into (B.16), we get:

$$\begin{aligned}
\mathbb{E}\left[f(\overline{X}_{T+1/2}) - \min_{x\in\mathcal{X}} f(x)\right] &\leq 2\left(\frac{R_h}{b} + \frac{b}{K_h} + \frac{K_h\|\mathcal{X}\|^2}{2b}\right)\frac{\sqrt{a^2 + 8(G^2+\sigma^2)T^3}}{T^2} \\
&\quad - \frac{b\sqrt{a^2}}{T^2}\left(\frac{1}{2K_h} + \frac{K_h\|\mathcal{X}\|^2}{2}\right). && \text{(B.20)}
\end{aligned}$$

Then, from our choice for b and a^2 in Theorem 1, we obtain:

$$\mathbb{E}\left[f(\overline{X}_{T+1/2}) - \min_{x\in\mathcal{X}} f(x)\right] \leq 2\frac{C_h}{\sqrt{K_h}}\frac{\sqrt{K_h + 8(G^2+\sigma^2)}}{\sqrt{T}}. \tag{B.21}$$

Proof of (20b): convergence of UNDERGRAD under (LG)/(LS). From (22) and (B.13), we have:

$$\begin{aligned}
\tilde{\mathcal{R}}_T(x) &\leq \frac{R_h}{\eta_{T+1}} + \frac{1}{2K_h}\sum_{t=1}^{T}\alpha_t^2\eta_{t+1}\|g_{t+1/2} - g_t\|_*^2 \\
&\quad + \frac{K_h}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_{t+1} - X_{t+1/2}\|^2 - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1/2} - X_t\|^2. && \text{(B.22)}
\end{aligned}$$

We analyze the terms in the right-hand-side of (B.22). First, we have:

$$\frac{K_h}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|X_{t+1} - X_{t+1/2}\|^2 \leq \frac{K_h\|\mathcal{X}\|^2}{2}\left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_1}\right). \tag{B.23}$$

Second, we have:

$$\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\|X_{t+1/2}-X_t\|^2 = \frac{K_h}{2}\sum_{t=1}^{T}\left[\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right]\|X_{t+1/2}-X_t\|^2 + \frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\|X_{t+1/2}-X_t\|^2$$
$$\leq \frac{K_h\|\mathcal{X}\|^2}{2}\left(\frac{1}{\eta_{T+1}}-\frac{1}{\eta_1}\right)+\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\|X_{t+1/2}-X_t\|^2 \tag{B.24}$$

Hence,

$$-\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\|X_{t+1/2}-X_t\|^2 \leq \frac{K_h\|\mathcal{X}\|^2}{2}\left(\frac{1}{\eta_{T+1}}-\frac{1}{\eta_1}\right)-\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\|X_{t+1/2}-X_t\|^2. \tag{B.25}$$

Combine (B.23) and (B.25) with (B.22), we have:

$$\tilde{\mathcal{R}}_T(x) \leq \frac{R_h+K_h\|\mathcal{X}\|^2}{\eta_{T+1}}-\frac{K_h\|\mathcal{X}\|^2}{\eta_1}+\frac{1}{2K_h}\sum_{t=1}^{T}\alpha_t^2\eta_{t+1}\|g_{t+1/2}-g_t\|_*^2-\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_{t+1}}\|X_{t+1/2}-X_t\|^2. \tag{B.26}$$

We will analyze the terms in the right-hand-side of (B.26). To do this, we first introduce the quantities

$$B_t^2 = \min\{\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2, \|g_{t+1/2}-g_t\|_*^2\} \tag{B.27a}$$

and

$$\xi_t = \left[g_{t+1/2}-g_t\right]-\left[\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\right]. \tag{B.27b}$$

We also define

$$\tilde{\eta}_t = \frac{b}{\sqrt{a^2+\sum_{s=1}^{t-1}\alpha_s^2B_s^2}}. \tag{B.28}$$

By these definitions, we obtain that

$$\begin{aligned}\|g_{t+1/2}-g_t\|_*^2 &\leq B_t^2+\left[\|g_{t+1/2}-g_t\|_*^2-\min\{\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2,\|g_{t+1/2}-g_t\|_*^2\}\right]\\ &\leq B_t^2+\max\{0,\|g_{t+1/2}-g_t\|_*^2-\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2\}\\ &\leq B_t^2+B_t^2+2\|\xi_t\|_*^2\\ &= 2B_t^2+2\|\xi_t\|_*^2.\end{aligned} \tag{B.29}$$

Here, the last inequality is obtained by the fact that if $\|g_{t+1/2}-g_t\|_*^2 \geq \|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2$ then it yields:

$$\|g_{t+1/2}-g_t\|_*^2-\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2 \leq B_t^2+2\|\xi_t\|_*^2. \tag{B.30}$$

Therefore, we have:

$$\begin{aligned}\frac{1}{2K_h}\sum_{t=1}^{T}\alpha_t^2\eta_{t+1}\|g_{t+1/2}-g_t\|_*^2 &= \frac{b}{2K_h}\sum_{t=1}^{T}\frac{\alpha_t^2\|g_{t+1/2}-g_t\|_*^2}{\sqrt{a^2+\sum_{s=1}^{t}\alpha_s^2\|g_{s+1/2}-g_s\|^2}}\\ &\leq \frac{b}{K_h}\sqrt{a^2+\sum_{t=1}^{T}\alpha_t^2\|g_{t+1/2}-g_t\|_*^2}-\frac{b\sqrt{a^2}}{2K_h} && \text{\# from Lemma A.4}\\ &\leq \frac{b}{K_h}\sqrt{a^2+2\sum_{t=1}^{T}\alpha_t^2B_t^2+2\sum_{t=1}^{T}\alpha_t^2\|\xi_t\|_*^2}-\frac{b\sqrt{a^2}}{2K_h} && \text{\# from (B.29)}\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{b\sqrt{2}}{K_h}\sqrt{a^2+\sum_{t=1}^T \alpha_t^2 B_t^2} + \frac{b\sqrt{2}}{K_h}\sqrt{\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2} - \frac{b\sqrt{a^2}}{2K_h} \\
&\leq \frac{b\sqrt{2}}{K_h}\left(\sqrt{a^2}+\sum_{t=1}^T \frac{\alpha_t^2 B_t^2}{\sqrt{a^2+\sum_{s=1}^t \alpha_s^2 B_s^2}}\right) + \frac{b\sqrt{2}}{K_h}\sqrt{\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2} - \frac{b\sqrt{a^2}}{2K_h} \\
&\qquad\qquad \text{\# } \textit{from Lemma A.4} \\
&= \frac{\sqrt{2}}{K_h}\sum_{t=1}^T \alpha_t^2\tilde{\eta}_{t+1}B_t^2 + \frac{b\sqrt{2}}{K_h}\sqrt{\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2} + \frac{b\sqrt{a^2}}{K_h}\left(\sqrt{2}-\frac{1}{2}\right). \qquad \text{(B.31)}
\end{aligned}
$$

On the other hand, by the update rule in Algorithm 1 and our choice of $\alpha_t = t, \forall t$ as in Theorem 1, we also have $X_t - X_{t+1/2} = \frac{\sum_{s=1}^t \alpha_s}{\alpha_t}\big(\bar{X}_t - \bar{X}_{t+1/2}\big) = \frac{\alpha_{t+1}}{2}\big(\bar{X}_t - \bar{X}_{t+1/2}\big)$. Use this and recall that $\frac{1}{\tilde{\eta}_t} \leq \frac{1}{\eta_t}$ for any t and that f is L-smooth over $\mathcal{X}$, we have:

$$
\begin{aligned}
-\frac{K_h}{2}\sum_{t=1}^T \frac{1}{\eta_{t+1}}\|X_t - X_{t+1/2}\|^2 &\leq -\frac{K_h}{2}\sum_{t=1}^T \frac{1}{\tilde{\eta}_{t+1}}\|X_t - X_{t+1/2}\|^2 \\
&= -\frac{K_h}{8}\sum_{t=1}^T \frac{\alpha_{t+1}^2}{\tilde{\eta}_{t+1}}\|\bar{X}_t - \bar{X}_{t+1/2}\|^2 \\
&\leq -\frac{K_h}{8}\sum_{t=1}^T \frac{1}{\tilde{\eta}_{t+1}}\frac{1}{L^2}\|\nabla f(\overline{X}_t) - \nabla f(\overline{X}_{t+1/2})\|_*^2 \\
&\leq -\frac{K_h}{8L^2}\sum_{t=1}^T \frac{\alpha_t^2 B_t^2}{\tilde{\eta}_{t+1}}. \qquad \text{(B.32)}
\end{aligned}
$$

Finally, letting $C_h = \sqrt{R_h + K_h\|\mathcal{X}\|^2}$, (B.29) yields:

$$
\begin{aligned}
\frac{C_h^2}{\eta_{T+1}} &= \frac{C_h^2}{b}\sqrt{a^2+\sum_{t=1}^T \alpha_t^2\|g_{t+1/2}-g_t\|_*^2} \\
&\leq \frac{C_h^2}{b}\sqrt{a^2+2\sum_{t=1}^T \alpha_t^2 B_t^2 + 2\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2} \\
&\leq \frac{\sqrt{2}C_h^2}{b}\sqrt{a^2+\sum_{t=1}^T \alpha_t^2 B_t^2} + \frac{\sqrt{2}C_h^2}{b}\sqrt{\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2} \\
&\leq \frac{\sqrt{2}C_h^2}{b}\left(\sqrt{a^2}+\sum_{t=1}^T \frac{\alpha_t^2 B_t^2}{\sqrt{a^2+\sum_{s=1}^t \alpha_s^2 B_s^2}}\right) + \frac{\sqrt{2}C_h^2}{b}\sqrt{\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2} \\
&\leq \frac{\sqrt{2}C_h^2}{b^2}\sum_{t=1}^T \alpha_t^2\tilde{\eta}_{t+1}B_t^2 + \frac{\sqrt{2}C_h^2\sqrt{a^2}}{b} + \frac{\sqrt{2}C_h^2}{b}\sqrt{\sum_{t=1}^T \alpha_t^2\|\xi_t\|_*^2}. \qquad \text{(B.33)}
\end{aligned}
$$

Combine (B.31), (B.32) and (B.33) into (B.26), we have:

$$
\tilde{\mathcal{R}}_T(x) \leq \sqrt{2}\sum_{t=1}^T \alpha_t^2 B_t^2\left[\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)\tilde{\eta}_{t+1} - \frac{K_h}{8L^2\tilde{\eta}_{t+1}}\right]
$$

$$+\sqrt{2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\xi_t\|_*^2}+\left[\frac{b\sqrt{a^2}}{K_h}\left(\sqrt{2}-\frac{1}{2}\right)-\frac{K_h\|\mathcal{X}\|^2\sqrt{a^2}}{b}+\frac{\sqrt{2}C_h^2\sqrt{a^2}}{b}\right]. \quad \text{(B.34)}$$

Now, define T_0 as follows:

$$T_0=\max\left\{1\le t\le T:\tilde{\eta}_{t+1}\ge\frac{\sqrt{K_h}}{\sqrt{8L^2\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)}}\right\} \quad \text{(B.35)}$$

In other words, for any $t>T_0$, $\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)\tilde{\eta}_{t+1}-\frac{K_h}{8L^2\tilde{\eta}_{t+1}}<0$. As a consequence,

$$\begin{aligned}
&\sqrt{2}\sum_{t=1}^{T}\alpha_t^2B_t^2\left[\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)\tilde{\eta}_{t+1}-\frac{K_h}{8L^2\tilde{\eta}_{t+1}}\right]\\
=&\sqrt{2}\sum_{t=1}^{T_0}\alpha_t^2B_t^2\left[\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)\tilde{\eta}_{t+1}-\frac{K_h}{8L^2\tilde{\eta}_{t+1}}\right]\\
\le&\sqrt{2}\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)\sum_{t=1}^{T_0}\alpha_t^2B_t^2\tilde{\eta}_{t+1}\\
=&\sqrt{2}\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)\sum_{t=1}^{T_0}b\frac{\alpha_t^2B_t^2}{\sqrt{a^2+\sum_{s=1}^{t}\alpha_s^2B_s^2}}\\
\le&\sqrt{2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\left(2\sqrt{a^2+\sum_{t=1}^{T_0}\alpha_t^2B_t^2}-\sqrt{a^2}\right)\\
=&2\sqrt{2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\frac{b}{\tilde{\eta}_{T_0+1}}-\sqrt{2a^2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\\
\le&8\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)^{3/2}\frac{b^2L}{\sqrt{K_h}}-\sqrt{2a^2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right).
\end{aligned} \quad \text{(B.36)}$$

Combine (B.34) with (B.36) and use the fact that $\mathbb{E}\left[\|\xi\|_*^2\right]\le 4\sigma^2$, we have:

$$\begin{aligned}
\mathbb{E}\left[\tilde{\mathcal{R}}_T(x)\right]\le&\,8\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)^{3/2}\frac{b^2L}{\sqrt{K_h}}-\sqrt{2a^2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\\
&+2\sqrt{2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\sigma\sqrt{\sum_{t=1}^{T}\alpha_t^2}+\left[\frac{b\sqrt{a^2}}{K_h}\left(\sqrt{2}-\frac{1}{2}\right)-\frac{K_h\|\mathcal{X}\|^2\sqrt{a^2}}{b}+\frac{\sqrt{2}C_h^2\sqrt{a^2}}{b}\right]
\end{aligned} \quad \text{(B.37)}$$

Recall the choice $\alpha_t=t,\forall t$, apply Lemma B.1, we have:

$$\begin{aligned}
\mathbb{E}\left[f(\overline{X}_{T+1/2})-\min_{x\in\mathcal{X}}f(x)\right]\le&\frac{16}{T^2}\left(\frac{R_h}{b^2}+\frac{K_h\|\mathcal{X}\|^2}{b^2}+\frac{1}{K_h}\right)^{3/2}\frac{b^2L}{\sqrt{K_h}}+\frac{4\sqrt{2}}{\sqrt{T}}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right)\sigma\\
&+\frac{2C(b,a^2)}{T^2}.
\end{aligned} \quad \text{(B.38)}$$

where we set

$$C(b,a^2) \coloneqq \frac{b\sqrt{a^2}}{K_h}\left(\sqrt{2}-\frac{1}{2}\right) - \frac{K_h\|\mathcal{X}\|^2\sqrt{a^2}}{b} + \frac{\sqrt{2}C_h^2\sqrt{a^2}}{b} - \sqrt{2a^2}\left(\frac{R_h}{b}+\frac{K_h\|\mathcal{X}\|^2}{b}+\frac{b}{K_h}\right). \tag{B.39}$$

Finally, replace $b = \sqrt{K_h C_h^2}$ and $a^2 = K_h$ as chosen in Theorem 1 into (B.38) and note that with these choices, $C(b,a^2) \le -\frac{1}{2}C_h\sqrt{K_h} \le 0$; we rewrite (B.38) as follows:

$$\mathbb{E}\left[f(\overline{X}_{T+1/2}) - \min_{x\in\mathcal{X}} f(x)\right] \le \frac{32\sqrt{2}L}{T^2}\left(\frac{C_h^2}{K_h}\right) + \frac{8\sqrt{2}\sigma}{\sqrt{T}}\frac{C_h}{\sqrt{K_h}}. \tag{B.40}$$

Convergence of UNDERGRAD in unbounded domains. Finally, we give the proof of Theorem 2 concerning the deterministic SFO in the (LG) case with a possibly *unbounded* domain $\mathcal{X}$.

Proof. Since the respective learning rate η_t is non-increasing and non-negative, we have that its limit exists. Particularly,

$$\lim_{t\to\infty}\eta_t = \inf_{t\in\mathbb{N}}\eta_t \ge 0 \tag{B.41}$$

Let us assume that $\inf_{t\in\mathbb{N}}\eta_t = 0$. Then, by applying Proposition 1 we have:

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t\langle\nabla f(\overline{X}_{t+1/2}), X_{t+1/2}-x\rangle &\le \frac{h(x)-\min_{x\in\mathcal{X}}h(x)}{\eta_{T+1}}\\ &+\sum_{t=1}^{T}\alpha_t\langle\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_{t+1/2}), X_t - X_{t+1}\rangle - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1}-X_{t+1/2}\|^2 - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1/2}-X_t\|^2\end{aligned} \tag{B.42}$$

Now for the term: $\sum_{t=1}^{T}\alpha_t\langle\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t), X_{t+1/2}-X_{t+1}\rangle - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1}-X_{t+1/2}\|^2$ we have:

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t\langle\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_{t+1/2}), X_t - X_{t+1}\rangle - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1}-X_{t+1/2}\|^2 &\le \frac{1}{2K_h}\sum_{t=1}^{T}\alpha_t^2\eta_t\|\nabla f(\overline{X}_{s+1/2})-\nabla f(\overline{X}_s)\|_*^2\\ &+\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\|X_{t+1}-X_{t+1/2}\|^2 - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1}-X_{t+1/2}\|^2\end{aligned} \tag{B.43}$$

which readily yields:

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t\langle\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t), X_{t+1/2}-X_{t+1}\rangle - \sum_{t=1}^{T}\frac{K_h}{2\eta_t}\|X_{t+1}-X_{t+1/2}\|^2 &\le \frac{1}{2K_h}\sum_{t=1}^{T}\alpha_t^2\eta_t\|\nabla f(\overline{X}_{s+1/2})-\nabla f(\overline{X}_s)\|_*^2\end{aligned} \tag{B.44}$$

Hence, putting everything together, we get:

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t\langle\nabla f(\overline{X}_{t+1/2}), X_{t+1/2}-x\rangle &\le \frac{1}{2K_h}\sum_{t=1}^{T}\alpha_t^2\eta_t\|\nabla f(\overline{X}_{s+1/2})-\nabla f(\overline{X}_s)\|_*^2\\ &-\frac{K_h}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\|X_t - X_{t+1/2}\|^2\end{aligned} \tag{B.45}$$

Moreover, since f is smooth we have:

$$\begin{aligned}\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2 &\leq L^2\|\overline{X}_{t+1/2} - \overline{X}_t\|^2 \\ &\leq L^2\frac{\alpha_t^2}{\left(\sum_{t=1}^T \alpha_t\right)^2}\|X_{t+1/2} - X_t\|^2 \\ &= L^2\frac{4t^2}{t^2(t+1)^2}\|X_{t+1/2} - X_t\|^2 \\ &\leq \frac{4L^2}{\alpha_t^2}\|X_{t+1/2} - X_t\|^2 \end{aligned} \tag{B.46}$$

Combining this with the fact that η_t is a decreasing sequence, we can rewrite (B.45) as follows:

$$\begin{aligned}\sum_{t=1}^T \alpha_t\langle\nabla f(\overline{X}_{t+1/2}), X_{t+1/2} - x\rangle \leq \frac{\eta_1}{2K_h}\sum_{t=1}^T \alpha_t^2\|\nabla f(\overline{X}_{s+1/2}) - \nabla f(\overline{X}_s)\|_*^2 \\ - \frac{K_h}{8L^2}\sum_{t=1}^T \frac{\alpha_t^2}{\eta_t}\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2] \end{aligned} \tag{B.47}$$

In the sequel, we look for the appropriate bounds of the two terms in the right-hand-side of (B.47). We start with the second term. From (B.9), we also have $\sum_{t=1}^T \alpha_t\langle\nabla f(\overline{X}_{t+1/2}), X_{t+1/2} - x\rangle \geq 0$. Combine this with (B.47), we have:

$$0 \leq \frac{1}{2K_h}\sum_{t=1}^T \alpha_t^2\eta_t\|\nabla f(\overline{X}_{s+1/2}) - \nabla f(\overline{X}_s)\|_*^2 - \frac{K_h}{8L^2}\sum_{t=1}^T \frac{\alpha_t^2}{\eta_t}\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2 \tag{B.48}$$

Hence by rearranging we have:

$$\begin{aligned}\frac{K_h}{16L^2}\sum_{t=1}^T \frac{\alpha_t^2}{\eta_t}\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2 &\leq \frac{1}{2K_h}\sum_{t=1}^T \alpha_t^2\eta_t\|\nabla f(\overline{X}_{s+1/2}) - \nabla f(\overline{X}_s)\|_*^2 \\ &\quad - \frac{K_h}{16L^2}\sum_{t=1}^T \frac{\alpha_t^2}{\eta_t}\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2 \\ &= \frac{1}{2}\sum_{t=1}^T \alpha_t^2\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2\left[\frac{\eta_t}{K_h} - \frac{K_h}{8\eta_t L^2}\right] \end{aligned} \tag{B.49}$$

Now, since we assumed that η_t converges to 0, there exists some $t_0 \in \mathbb{N}$ such that:

$$\eta_t \leq \frac{K_h}{\sqrt{8}L} \quad \text{for all} \quad t > t_0 \tag{B.50}$$

which directly yields that $\left[\frac{\eta_t}{K_h} - \frac{K_h}{8\eta_t L^2}\right] \leq 0$ for all $t > T_0$. Hence, we have:

$$\frac{K_h}{16L^2}\sum_{t=1}^T \frac{\alpha_t^2}{\eta_t}\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2 \leq \frac{1}{2}\sum_{t=1}^{T_0} \alpha_t^2\|\nabla f(\overline{X}_{t+1/2}) - \nabla f(\overline{X}_t)\|_*^2\left[\frac{\eta_t}{K_h} - \frac{K_h}{8\eta_t L^2}\right] \tag{B.51}$$

On the other hand, we have:

$$\begin{aligned}\frac{1}{\eta_t} &= \frac{1}{\sqrt{K_h}}\sqrt{K_h + \sum_{s=1}^{t-1} \alpha_s^2\|\nabla f(\overline{X}_{s+1/2}) - \nabla f(\overline{X}_s)\|_*^2} \\ &\geq \frac{1}{\sqrt{K_h}}\sqrt{K_h} = 1 \end{aligned} \tag{B.52}$$

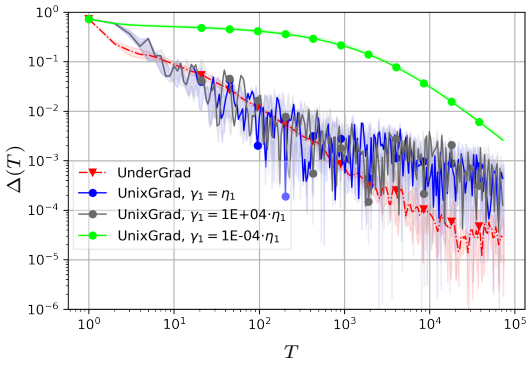

Figure 3: Convergence of UNDERGRAD and UNIXGRAD in the simplex setup with a noisy SFO. The plot is drawn in log-log scale. The y-axis corresponds to the differences between the f-value of the relevant point of each algorithm and $\min f$.

and hence

$$\begin{aligned}\frac{K_h}{16L^2}\sum_{t=1}^{T}\frac{\alpha_t^2}{\eta_t}\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2 &\geq \frac{K_h}{16L^2}\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2\\ &= \frac{K_h^2}{16L^2K_h}\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2\\ &= \frac{K_h^2}{16\eta_{T-1}^2L^2}\end{aligned} \tag{B.53}$$

So, summarizing we have:

$$\frac{K_h^2}{16\eta_{T-1}^2L^2}\leq\frac{1}{2}\sum_{t=1}^{T_0}\alpha_t^2\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2\left[\frac{\eta_t}{K_h}-\frac{K_h}{8\eta_tL^2}\right] \tag{B.54}$$

We now focus on the first term of the right-hand-size of (B.47). If one lets T to infinity and recalling the fact that we assumed that η_t converges to 0 (and so $1/\eta_t^2\to\infty$), we have that:

$$\infty\leq\frac{1}{2}\sum_{t=1}^{T_0}\alpha_t^2\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2\left[\frac{\eta_t}{K_h}-\frac{K_h}{8\eta_tL^2}\right]<\infty \tag{B.55}$$

a contradiction. This shows that $\inf_{t\in\mathbb{N}}\eta_t>0$ and hence

$$\begin{aligned}\sum_{t=1}^{+\infty}\alpha_t^2\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2 &= \lim_{T\to+\infty}\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\overline{X}_{t+1/2})-\nabla f(\overline{X}_t)\|_*^2\\ &= \lim_{T\to\infty}\frac{K_h}{\eta_t^2}-K_h\\ &= \frac{K_h}{\inf_t\eta_t}^2-K_h<\infty\end{aligned} \tag{B.56}$$

so our proof is complete. ■

C. Additional Numerical Experiments

In this last section, we report another numerical experiment highlighting the universality of UNDERGRAD. In this experiment, we also focus on the simplex setup as presented in Section 5. However, this time, we work with a noisy SFO that returns first-order feedback that is perturbed by a noise generated from a pre-determined zero-mean normal distribution. We

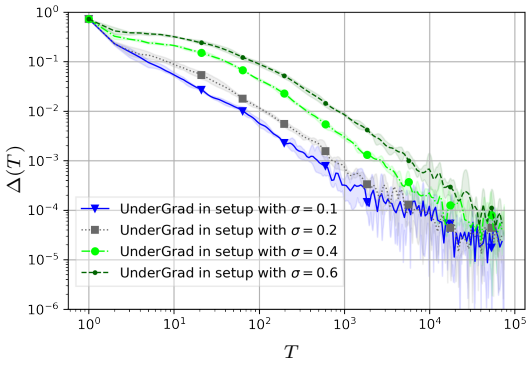

Figure 4: Convergence of UNDERGRAD in the simplex setup with different noise levels of the SFO.

compare the performances of UNDERGRAD and UNIXGRAD, both run with the entropic regularizer. The result of this experiment is reported in Fig. 3.

Fig. 3 shows that UNDERGRAD obtains the optimal rate $\mathcal{O}(1/\sqrt{T})$ in this set-up. UNIXGRAD can also obtain the same rate but only when its step-size update rule is chosen appropriately (note again that with entropic regularizer, the update rule (14) of UNIXGRAD is not available due to the fact that $B_h = \infty$): when γ_1 is chosen with the same or larger magnitude of UNDERGRAD's initial learning rate, UNIXGRAD converges with the rate $\mathcal{O}(1/\sqrt{T})$; but if γ_1 is too small (e.g., when $\gamma_1 = 10^{-3} \cdot \eta_1$), UNIXGRAD can have a very long warming up phase. This experiment reasserts that in cases where the step-size update rule (14) is unavailable, it is non-trivial to choose an appropriate step-size update rule of UNIXGRAD: small γ_1 might lead to better performances under perfect SFO (cf. Section 5) but might create unwanted behaviors in noisy SFO setups. On the contrary, UNDERGRAD does not encounter such issues in our experiments.

Finally, we conduct another experiment to confirm the dependency of the convergence rates of UNDERGRAD on the noise level σ. The result is reported in Fig. 4.