

OBLIVIOUS STOCHASTIC CONVEX OPTIMIZATION

CLÉMENT LEZANE^{1,*}, ALEXANDRE D'ASPREMONT²

¹*Department of Statistics, University of Twente, Netherlands*

²*CNRS & ENS, Paris*

Abstract. In stochastic convex optimization problems, most existing adaptive methods rely on prior knowledge about the diameter bound D when the smoothness or the Lipschitz constant is unknown. This often significantly affects performance as only a rough approximation of D is usually known in practice. Here, we bypass this limitation by combining mirror descent with dual averaging techniques and we show that, under oblivious step-sizes regime, our algorithms converge without any prior knowledge on the parameters of the problem. We introduce three oblivious stochastic algorithms to address different settings. The first algorithm is designed for objectives in relative scale, the second one is an accelerated version tailored for smooth objectives, whereas the last one is for relatively-smooth objectives. All three algorithms work without prior knowledge of the diameter of the feasible set, the Lipschitz constant or smoothness of the objective function. We use these results to revisit the problem of solving large-scale semidefinite programs using randomized first-order methods and stochastic smoothing. We extend our framework to relative scale and demonstrate the efficiency and robustness of our methods on large-scale semidefinite programs.

Keywords. Convergence analysis; Mirror descent; Regularization; Stochastic convex problems.

1. INTRODUCTION

We address stochastic convex optimization problems where regularity parameters are unknown a priori. Our goal is to solve the population loss minimization problem

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E}_{\xi}[f_{\xi}(x)]$$

with ξ being a random variable, f_{ξ} being convex functions, and \mathcal{X} being a convex set. In the adaptive setting ([2, 14]) three key assumptions are fundamental in all our problem settings. We first assume access to an unbiased stochastic oracle $\nabla f_{\xi}(x)$ estimating $\nabla F(x)$, i.e.

$$\forall x \in \mathcal{X}, \quad \mathbb{E}_{\xi}[\nabla f_{\xi}(x)] = \nabla F(x),$$

which provides noisy gradient information for the population loss. The second assumption is the availability of a strictly convex function H that serves as a regularizer. For example, we could define H as the squared norm $H = \|\cdot\|_p^2$ for ℓ_p spaces when $1 < p \leq 2$. The last

*Corresponding author.

E-mail address: c.w.lezane@utwente.nl (C. Lezane).

Received 7 October 2025; Accepted 3 December 2025; Published online 1 January 2026.

©2026 Journal of Nonlinear and Variational Analysis

assumption concerns various bounds on the problem depending on the setting, that will be detailed in the next subsection.

1.1. Problem settings. We focus on oblivious algorithms in multiple problem settings, namely optimization in relative scale, standard smoothness and relative smoothness assumptions. In what follows, we write the Bregman divergence of a continuously derivable function f by D^f ,

$$\forall x, y \in \mathcal{X}, \quad D^f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Optimization in relative scale. In the relative scale setting, we extend the results from [22] which imposed a relative constraint on the stochastic oracle and a lower bound on F , written

$$\begin{aligned} \forall x \in \mathcal{X}, \quad \mathbb{E}_\xi \left[\|\nabla f_\xi(x)\|^2 \right] &\leq \mathcal{M}F(x) \\ \exists H, \exists \hat{x}_0 \in \mathcal{X}, \forall x \in \mathcal{X}, \quad F(x) &\geq \Gamma \left(H(x) - H(\hat{x}_0) \right), \end{aligned}$$

where $\mathcal{M}, \Gamma > 0$ are positive constants and $\hat{x}_0 \in \mathcal{X}$ a fixed point in the space. In the original setting [22], the condition is satisfied with $H(x) := \|x - \hat{x}_0\|_2^2$ being the Euclidean norm. As a consequence, the objective function F must be positive. The authors used prior knowledge on $(\mathcal{M}, \Gamma, \hat{x}_0)$ and designed an algorithm that guarantees the following convergence rate,

$$\mathbb{E}[F(x_{T+1})] - F_\star \leq F_\star \sqrt{\frac{2\mathcal{M}}{\Gamma T}},$$

where x_{T+1} is the final output of their algorithm.

Minimizing smooth convex functions. For the standard smooth case, we require the objective function F to be L -smooth with respect to an arbitrary norm $\|\cdot\|$ and the stochastic noise to have a finite variance σ^2 , i.e.,

$$\begin{aligned} \forall x, y \in \mathcal{X}, \quad D^F(x, y) &\leq \frac{L}{2} \|x - y\|^2 \\ \forall x \in \mathcal{X}, \quad \mathbb{E}_\xi \left[\|\nabla f_\xi(x) - \nabla F(x)\|_*^2 \right] &\leq \sigma^2, \end{aligned}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. In that case, we can use the accelerated framework in [13] to obtain

$$\mathbb{E} \left[F(x_{T+1}) - F_\star \right] \lesssim \frac{LD^2}{T^2} + \frac{\sigma D}{\sqrt{T}},$$

where x_{T+1} is the output of the accelerated algorithm. The convergence rate is optimal for the smooth case [18, 13], but the algorithm requires prior knowledge on L and D .

Optimization for relatively-smooth functions. In the relatively-smooth setting [6, 15], the smoothness is related to a convex function H instead of a norm. We note that, in the relatively smooth case, H is only required to be strictly convex and twice continuously differentiable. We assume that the stochastic oracle satisfies

$$\begin{aligned} \forall x, y \in \mathcal{X}, \quad D^F(x, y) &\leq L_\star D^H(x, y) \\ \forall \eta < \eta_0, \quad \mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] &\leq 2\eta^2 \sigma^2. \end{aligned}$$

For that relative smooth case, we can use the mirror descent algorithm in [6] to get

$$\mathbb{E}[F(x_{T+1}) - F_\star] \lesssim \frac{\sigma D_0}{\sqrt{T}},$$

where x_{T+1} is the final output of the algorithm. Note that we need an extra assumption in relative smoothness to show convergence.

Assumption 1.1. For all $y \in \mathbb{R}^d$, the following problem

$$\min_{x \in \mathcal{X}} H(x) - \langle y, x \rangle$$

has an unique solution in the interior of \mathcal{X} .

1.2. Related work. Previous methods required prior knowledge of different parameters. Having no or partial information on the problem's parameters has a direct impact on the performance of most first order algorithms, as they are often used to select step-sizes for example. Even if one does have a rough estimate of some parameters (such as the Lipschitz constant), over or under-estimation makes these algorithms slow and inefficient. The key challenge is developing algorithms that can either handle parameter uncertainty or learn these parameters from scratch.

1.2.1. Polyak step-size and Line search. For the non-smooth case, the Polyak step-size offers a viable alternative [23], but this method requires knowledge of $F_\star := \min F$. For the smooth case, line search methods help if the gradient is known exactly [21, 24], but designing a stochastic counterpart of line search is still an open question.

1.2.2. Grid search. When direct evaluation of the objective function F is feasible, an alternative strategy involves running multiple parallel algorithm instances, each configured with different parameter settings. The final solution is selected as the output that achieves the minimum value of F . This Grid search approach was originally proposed by Nemirovski for smooth problems where the gradient is deterministic and known, recommending $\log(T)$ parallel sessions to achieve optimal performance guarantees. We have extended this idea for the stochastic case when evaluating the objective is still possible but might be costly. More details can be found in Section 4.

1.2.3. Adagrad and online algorithms. For the smooth case, different variations on the Adagrad step-size [7] prove useful if we have prior knowledge on the diameter D of the feasible set \mathcal{X} , but L remains unknown. Writing $g_t = \nabla f_{\xi_t}(x_t)$ the stochastic oracle at x_t , the Adagrad step-size becomes

$$\eta_t = \frac{2D}{\sqrt{D^2 + \sum_{\tau=1}^{t-1} \alpha_\tau \|g_\tau\|^2}}.$$

We mention here a general adaptive method [14] that works for both non-smooth and smooth cases, and has inspired our proof technique. In the smooth case, the convergence rate becomes

$$\mathbb{E}[F(x_{T+1}) - F_\star] \lesssim \frac{LD^2}{T} + \frac{\sigma D}{\sqrt{T}},$$

which is the standard non-accelerated rate. In comparison, the optimism in online learning scheme [2] provides a sharper accelerated rate written

$$\mathbb{E}[F(x_{T+1}) - F_*] \lesssim \frac{LD^2}{T\sqrt{T}} + \frac{\sigma D}{\sqrt{T}},$$

but the optimism algorithm requires implicitly prior knowledge on D . In the same way, another rate $O\left(\frac{LD^2 \log(\mathcal{MT})}{T^2} + \frac{\sigma D \log(\mathcal{MT})}{\sqrt{T}}\right)$ is derived in [2]; however, this method requires an almost sure bound on the stochastic gradients, aside from knowledge of D_0 .

More bounds have been provided in later works [11], but they require additional conditions or have a convergence rate explicitly depending on the dimension d , so we omit them here. On the other hand, a parameter free online mirror descent algorithm has been proposed in [10], assuming extra conditions on the second derivative of the proximal function, which do not cover the smooth case described here.

1.3. Our contributions. Our objective here is to develop algorithms capable of handling diverse stochastic convex optimization scenarios where key problem parameters remain unknown.

1.3.1. Composite Mirror Descent with Oblivious Step-sizes. When confronting optimization problems with limited knowledge of regularity parameters, we introduce a regularization framework H based on complementary composite stochastic optimization techniques [3, 5, 20]. We then construct a family of oblivious step-sizes $\{\alpha_t, \beta_t, \gamma_t\}$ that enable convergence without requiring prior knowledge of Lipschitz or smoothness constants.

1.3.2. Applications Across Different Problem Classes. Our algorithmic framework applies to three important non-smooth optimization settings: relative-scale precision targets [22], smooth objectives [8], and relatively-smooth functions [6]. Relative-scale optimization measures complexity in terms of digits of precision gained rather than absolute precision targets, often yielding substantial computational savings. Relative smoothness generalizes the objective function's geometric properties by relating them to arbitrary regularization functions instead of standard norms.

1.3.3. Practical Impact and Theoretical Foundations. Our oblivious algorithms combine ease of implementation with strong empirical performance while providing interesting theoretical convergence guarantees. We summarize our contribution with respect to existing methods in Table 1.

1.4. Applications to semidefinite programming. Consider the problem of minimizing the maximum eigenvalue of a symmetric matrix x , over a convex set $\mathcal{X} \subseteq S^d$, written

$$\min_{x \in \mathcal{X}} \lambda_{\max}(x), \tag{1.1}$$

in the variable $x \in S^d$ where S^d is the set of symmetric real $d \times d$ matrices.

While semidefinite programs (SDP) can be efficiently solved by interior point methods, their algorithmic complexity becomes prohibitive for large-scale problems. First-order methods offer a viable alternative, albeit with a potential sacrifice in accuracy. For many large-scale applications however, achieving an approximate solution within an acceptable

	Oblivious			
Class	D	$\mathcal{M}/L/L_*$	Unbounded noise	Convergence rate
Relative scale	\times	\times	\checkmark	$O\left(\sqrt{\frac{\mathcal{M}}{\Gamma}} \frac{F_*}{\sqrt{T}}\right)$ [14]
	\checkmark	\checkmark	\checkmark	$\tilde{O}\left(\sqrt{\frac{\mathcal{M}}{\Gamma}} \frac{F_*}{\sqrt{T}} + \frac{1}{T^{n/2}}\right)$
L -smooth	\times	\checkmark	\checkmark	$O\left(\frac{LD^2}{T} + \frac{\sigma D}{\sqrt{T}}\right)$ [14]
	\times	\checkmark	\checkmark	$O\left(\frac{LD^2 \log(T)^2}{T} + \frac{\sigma \log(T)}{\sqrt{T}}\right)$ [2]
	\times	\checkmark	\times	$\tilde{O}\left(\left(\frac{LD^2}{T^2} + \frac{\sigma D}{\sqrt{T}}\right)\right)$ [2]
	\checkmark	\checkmark	\checkmark	$\tilde{O}\left(\frac{\sigma D}{\sqrt{T}} + \frac{1}{T^n}\right)^{(*)}$
L_* -relatively-smooth	\times	\times	\checkmark	$O\left(\frac{\sigma D_0}{\sqrt{T}}\right)$ [6]
L_* -relatively-smooth	\checkmark	\checkmark	\times	$\tilde{O}\left(\frac{\sigma D_0}{\sqrt{T}} + \frac{1}{T^n}\right)^{(*)}$

TABLE 1. Summary of convergence rates with limited prior knowledge on parameters (absolute constants omitted). The exponent n is the degree of the oblivious step size which can be chosen in an arbitrary way. $(*)$ are the contributions of this paper. Results given with \tilde{O} have an additional polylog factor.

range of the optimum is often sufficient. This has fueled a growing interest in the development of efficient first-order algorithms to handle large-scale semidefinite programs [4, 25].

Three common approaches to solve problem (1.1) will serve as a starting point for our work. The first approach is the celebrated mirror prox algorithm [16], whose main limitation is the requirement of a full spectral decomposition of a symmetric matrix at each iteration.

A second approach is based on a technique known as *stochastic smoothing*. In [4], for instance, the authors use stochastic smoothing with k random rank one perturbations and target precision $\epsilon > 0$ to produce a smooth approximation $F_{k,\epsilon}$, written as follows

$$F_{k,\epsilon}(x) := \mathbb{E}_{z_i \sim \mathcal{N}(0, I_d)} \left[\max_{i=1, \dots, k} \lambda_{\max} \left(x + (\epsilon/d) z_i z_i^T \right) \right].$$

More recently, a different type of stochastic smoothing based on matrix powers with exponent $p \geq 1$ was formulated in [22] as

$$F_p(x) := \mathbb{E}_{u \in \mathcal{U}([0,1]^d)} \left[\langle x^p u, u \rangle^{1/p} \right].$$

Computing the minimum of an approximate objective function, as above, can be more effective depending on problem structure. A more detailed comparison is given in the introduction of [4] and in [22]. For solving large scale SDPs, we will call F the approximate objective given by a form of stochastic smoothing (it could be $F_{k,\epsilon}$ or F_p) and we will suppose F convex (which is the case above).

2. PRELIMINARIES

We recall the framework of the Stochastic Convex Optimization problem

$$\min_{x \in \mathcal{X}} F(x) = \mathbb{E}_\xi[f_\xi(x)].$$

We suppose that we have a stochastic gradient oracle $\nabla f_\xi(x)$, and define the composite mirror descent step as

$$x_{t+1} = \arg \min_x \{ \alpha_t \langle \nabla f_{\xi_t}(x_t), x \rangle + \beta_t H(x) + \gamma_t D^H(x, x_t) \}$$

where $\{\alpha_t, \beta_t, \gamma_t\}_{t \geq 1}$ is our sequence of step-sizes. Inspired by previous works on the composite setting [1, 3, 8, 19], we focus on the following regime

$$\begin{cases} \gamma_{t+1} - \gamma_t & \leq \beta_t \\ \gamma_t / \alpha_t & \leq \gamma_{t+1} / \alpha_{t+1}. \end{cases} \quad (2.1)$$

Particularly, when the information on different parameters of the problem is not available, we will consider general oblivious step-sizes that do not take any parameters into account. More precisely, we use a family of polynomial step-sizes $\alpha_t \propto t^n, \beta_t \propto \mu t^n, \gamma_t \propto \mu t^{n+1} / (n+1)$ with a constant $\mu > 0$ and a degree $n \geq 0$. As we will see later in Section 3, the parameters (μ, n) are oblivious to the instance, as they are selected a priori, in a problem-independent way.

2.1. Useful Lemmas. We introduce important lemmas in this subsection. First, we present an extension of the three-points identity.

Lemma 2.1. *Let f be a convex function and ν continuously differentiable. If we consider*

$$u^* = \arg \min_{u \in \mathcal{X}} \{f(u) + D^\nu(u, y)\},$$

with the Bregman divergence $D^\nu(a, b) := \nu(a) - \nu(b) - \langle \nabla \nu(b), a - b \rangle$, then, for all u ,

$$f(u^*) + D^\nu(u^*, y) + D^f(u, u^*) \leq f(u) + D^\nu(u, y) - D^\nu(u, u^*).$$

Proof. From the first order optimality conditions, for all $u \in \mathcal{X}$:

$$\langle \nabla f(u^*) + \nabla_u D^\nu(u^*, y), u - u^* \rangle \geq 0,$$

with the gradient taken with respect to the first entry. We also apply the three-points identity

$$\langle \nabla D^\nu(u^*, y), u - u^* \rangle = D^\nu(u, y) - D^\nu(u, u^*) - D^\nu(u^*, y).$$

Then

$$\begin{aligned} f(u) - f(u^*) - D^f(u, u^*) &= \langle \nabla f(u^*), u - u^* \rangle \\ &\geq D^\nu(u, u^*) - D^\nu(u, y) + D^\nu(u^*, y). \end{aligned}$$

□

We note that Lan's work [8] follows a similar route, except for a negative Bregman divergence term upper bounded by zero, that we maintain here. We now introduce an extension of Lemma 4 in [6] in the case of relative smoothness.

Lemma 2.2. *Let f be a convex function and ν continuously differentiable. If we consider*

$$u^* = \arg \min_{u \in \mathcal{X}} \{f(u) + D^\nu(u, y)\},$$

with the Bregman divergence $D^\nu(a, b) := \nu(a) - \nu(b) - \langle \nabla \nu(b), a - b \rangle$ and u^ is in the interior of \mathcal{X} , then, for all $u \in \mathcal{X}$,*

$$f(u^*) - f(u) + D^\nu(u^*, y) + D^f(u, u^*) = D^\nu(u, y) - D^\nu(u, u^*).$$

Proof. The proof is similar as before with only difference of u^* being in the interior of \mathcal{X} ,

$$\nabla f(u^*) + \nabla D^\nu(u^*, y) = 0.$$

The rest follows. \square

For the relative smooth case, it is also to show the Lemma 3 from [6] on the property and cocoercivity of the Bregman divergence.

Lemma 2.3. *Consider two convex functions $F, H : \mathcal{X} \rightarrow \mathbb{R}$ where F is L_\star -relatively smooth H . For any $0 < \eta < \frac{1}{L_\star}$ and $x, y \in \text{int}(\mathcal{X})$,*

$$D^{H_\star}(\nabla H(x) - \eta(\nabla F(x) - \nabla F(y)), \nabla H(x)) \leq \eta D^F(x, y)$$

where H_\star is the convex conjugate of H .

Proof. For all $y \in \text{int}(\mathcal{X})$, we note

$$D^F(\cdot, y) = F(\cdot) - F(y) - \langle \nabla F(y), \cdot - y \rangle$$

Therefore, for any $\eta \leq \frac{1}{L_\star}$, $y, x, u \in \text{int}(\mathcal{X})$ we know by the definition of relative smoothness

$$D^F(u, x) = D^{D^F(\cdot, y)}(u, x) = D^F(u, y) - D^F(x, y) - \langle \nabla_x D^F(x, y), u - x \rangle \leq \frac{1}{\eta} D^H(u, x)$$

We notice that the function $Q_{y,x} : u \rightarrow \frac{1}{\eta} D^H(u, x) + D^F(x, y) + \langle \nabla_x D^F(x, y), u - x \rangle$ is convex, greater than $D^F(u, y)$. We note u^+ as the minimum of $Q_{y,x}$, the gradients cancel at this point:

$$\begin{aligned} \frac{1}{\eta} \left(\nabla H(u^+) - \nabla H(x) \right) + \nabla_x D^F(x, y) &= 0 \\ \Leftrightarrow \nabla H(u^+) &= \nabla H(x) - \eta \nabla_x D^F(x, y) = \nabla H(x) - \eta \nabla F(x) + \eta \nabla F(y). \end{aligned}$$

By the previous inequality, we know that, for any $\eta \leq \frac{1}{L_\star}$,

$$\begin{aligned} 0 &\leq D^F(u^+, y) \leq Q_{y,x}(u^+) \\ &= \frac{1}{\eta} D^H(u^+, x) + D^F(x, y) + \langle \nabla_x D^F(x, y), u^+ - x \rangle \\ &= \frac{1}{\eta} D^H(u^+, x) + D^F(x, y) + \frac{1}{\eta} \langle \nabla H(x) - \nabla H(u^+), u^+ - x \rangle \\ &= D^F(x, y) - \frac{1}{\eta} D^H(x, u^+) \end{aligned}$$

We conclude by noticing that

$$D^H(x, u^+) = D^{H_\star}(\nabla H(u^+), \nabla H(x)) = D^{H_\star}(\nabla H(x) - \eta \nabla F(x) + \eta \nabla F(y), \nabla H(x)).$$

□

3. COMPOSITE MIRROR DESCENT

Our approach combines stochastic mirror descent with dual averaging to minimize convex functions without prior knowledge about the problem. The general analysis takes inspiration from existing methods like minimization in relative scale [22], the composite setting [3], and relative smoothness [6]. In the following part, we will present three variations of our algorithm in these various settings.

3.1. Relative scale setting. We recall that in the relative scale setting, there exist constants $\Gamma, \mathcal{M} > 0$ and a fixed point \hat{x}_0 such that, for all $x \in \mathcal{X}$,

$$\begin{aligned}\Gamma|H(x) - H(\hat{x}_0)| &\leq F(x) \\ \mathbb{E}\left[\|\nabla f_\xi(x, \xi)\|_*^2\right] &\leq \mathcal{M}F(x).\end{aligned}$$

We notice that in the Euclidean case where $H(x) = \|\cdot\|_2^2$, our setting recovers that described in [22], where

$$\Gamma\|x - x_\star\|_2^2 \leq F(x).$$

The first line is an extension of the quadratic lower bound assumption over the objective function and the second line controls the stochastic oracle noise using the objective function. In order to solve the previous problem without prior knowledge on the constants Γ, \mathcal{M} or the location of \hat{x}_0 , we have designed the following composite mirror descent algorithm.

Algorithm 1 Composite Mirror Descent for Relative Scale

Require: Number of iterations $T \geq 0$, starting point $x_1 \in \mathcal{X}$, step-sizes $(\alpha_t, \beta_t, \gamma_t)_{t \geq 1}$, regularizer H

for $1 \leq t \leq T$ **do**

$$x_{t+1} = \arg \min_{x \in \mathcal{X}} \{\alpha_t \langle \nabla f_{\xi_t}(x_t), x \rangle + \beta_t H(x) + \gamma_t D^H(x, x_t)\}$$

end for

Output $x_t^{ag} := \frac{\sum_{t=1}^T \alpha_t x_t}{\sum_{t=1}^T \alpha_t}.$

Unlike in the original setting, we will also suppose that F is explicitly bounded by an a priori unknown constant \mathcal{L} . Our convergence result is similar to existing results in [3, 12, 17] with some extra terms due to the new constraints.

Theorem 3.1. *Assume that H is 1-strongly-convex with regard to $\|\cdot\|$ and there exist $\Gamma, \mathcal{M}, \mathcal{L} > 0$ such that for all $x \in \mathcal{X}$,*

$$\begin{aligned}\Gamma(H(x) - \min H) &\leq F(x), \\ \mathbb{E}\left[\|\nabla f_\xi(x, \xi)\|_*^2\right] &\leq \mathcal{M}F(x), \\ F(x) - F_\star &\leq \mathcal{L}.\end{aligned}$$

Suppose Algorithm 1 runs under the oblivious step-size (2.1). Then, for any μ satisfying $2\mathcal{M}n/T \leq \mu \leq \frac{\Gamma}{2}$,

$$\mathbb{E}[F(x_T^{ag})] - F_\star = O\left(\left(\frac{\mathcal{M}n}{\mu T} + \frac{\mu}{\Gamma}\right)F_\star + \frac{\mu(H(\hat{x}_0) - H_\star)}{T} + \mathcal{L}\left(\frac{\mathcal{M}n}{\mu T}\right)^{n+1} + \frac{\mu D^H(x_F, x_1)}{T^{n+1}}\right).$$

We note that the first two terms with F_\star and H_\star are part of the standard result for optimization in relative scale. A large part of the proof follows the same scheme as [3, 9, 22].

Proof. To simplify notation, we write $g_t = \nabla f_{\xi_t}(x_t)$. By the proximal lemma (Lemma 2.1) applied to $u = x$, $y = x_t$, $u^\star = x_{t+1}$, $f(\cdot) = \alpha_t \langle g_t, \cdot \rangle + \beta_t H(\cdot)$, and $\nu(\cdot) = \mu \gamma_t H(\cdot)$, we have, for all $x \in \mathcal{X}$,

$$\begin{aligned} & \alpha_t \langle g_t, x_{t+1} - x \rangle + \beta_t H(x_{t+1}) - \beta_t H(x) + \beta_t D^H(x, x_{t+1}) \\ & \leq \gamma_t \left(D^H(x, x_t) - D^H(x, x_{t+1}) \right) - \gamma_t D^H(x_{t+1}, x_t). \end{aligned}$$

We write the stochastic noise by $\Delta_t(x_t) := g_t - \nabla F(x_t)$. By convexity of F , we obtain

$$\begin{aligned} \langle g_t, x_{t+1} - x \rangle &= \langle \nabla F(x_t) + \Delta_t(x_t), x_t - x \rangle + \langle g_t, x_{t+1} - x_t \rangle \\ &\geq F(x_t) - F(x) + \langle \Delta_t(x_t), x_t - x \rangle + \langle g_t, x_{t+1} - x_t \rangle. \end{aligned}$$

Combining the two previous equations, we see, for all $x \in \mathcal{X}$,

$$\begin{aligned} & \alpha_t \left(F(x_t) - F(x) + \langle \Delta_t(x_t), x_t - x \rangle + \langle g_t, x_{t+1} - x_t \rangle \right) + \beta_t \left[H(x_{t+1}) - H(x) \right] \\ & \quad + \beta_t D^H(x, x_{t+1}) \\ & \leq \gamma_t \left(D^H(x, x_t) - D^H(x, x_{t+1}) \right) - \gamma_t D^H(x_{t+1}, x_t). \end{aligned}$$

After rearranging the terms, we have

$$\begin{aligned} \alpha_t \left(F(x_t) - F(x) \right) &\leq \alpha_t \langle g_t, x_t - x_{t+1} \rangle - \gamma_t D^H(x_{t+1}, x_t) \\ &\quad + \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) \\ &\quad + \alpha_t \langle \Delta_t(x_t), x - x_t \rangle + \beta_t \left[H(x) - H(x_{t+1}) \right]. \end{aligned}$$

We notice that, similarly to the standard stochastic mirror descent proof [3], the first two terms can be simplified by using the strong convexity of H and Young's inequality

$$\begin{aligned} \alpha_t \langle g_t, x_t - x_{t+1} \rangle - \gamma_t D^H(x_{t+1}, x_t) &\leq \alpha_t \langle g_t, x_t - x_{t+1} \rangle - \frac{\gamma_t}{2} \|x_{t+1} - x_t\|^2 \\ &\leq \frac{\alpha_t^2}{2\gamma_t} \|g_t\|_*^2. \end{aligned}$$

Summing the previous equation from $t = 1$ to $t = T$, for all $x \in \mathcal{X}$

$$\begin{aligned}
\sum_{t=1}^T \alpha_t [F(x_t) - F(x)] &\leq \gamma_1 D^H(x, x_1) - (\beta_T + \gamma_T) D^H(x, x_{T+1}) \\
&+ \sum_{t=1}^{T-1} \underbrace{(\gamma_{t+1} - \gamma_t - \beta_t)}_{\leq 0} D^H(x, x_{t+1}) + \sum_{t=1}^T \alpha_t \langle \Delta_t(x_t), x - x_t \rangle + \sum_{t=1}^T \frac{\alpha_t^2 \|g_t\|_*^2}{\gamma_t} \\
&+ \sum_{t=1}^T \beta_t (H(x) - H(x_{t+1})).
\end{aligned} \tag{3.1}$$

We decompose the regularizer term $H(x) - H(x_{t+1})$ in the following way

$$\begin{aligned}
H(x) - H(x_{t+1}) &= (H(\hat{x}_0) - H(x_{t+1})) + (H(x) - H(\hat{x}_0)) \\
&\leq \frac{F(x_{t+1}) + F(x)}{\Gamma},
\end{aligned}$$

where we have used the relative scale setting $|H(x) - H(\hat{x}_0)| \leq F(x)/\Gamma$ in the last line. Exceptionally for the iterate $t = T$, we use another decomposition

$$H(x) - H(x_{T+1}) \leq H(\hat{x}_0) - H(x_{T+1}) + \frac{F(x)}{\Gamma} \leq H(\hat{x}_0) - \min H + \frac{F(x)}{\Gamma}.$$

After introducing these decompositions in equation (3.1), we simplify and obtain

$$\begin{aligned}
\sum_{t=1}^T \alpha_t [F(x_t) - F(x)] &\leq \gamma_1 D^H(x, x_1) + \frac{F(x)}{\Gamma} \sum_{t=1}^T \beta_t + \sum_{t=1}^{T-1} \frac{\beta_t}{\Gamma} F(x_{t+1}) \\
&+ \beta_T (H(\hat{x}_0) - \min H) + \sum_{t=1}^T \alpha_t \langle \Delta_t(x_t), x - x_t \rangle + \sum_{t=1}^T \frac{\alpha_t^2 \|g_t\|_*^2}{\gamma_t}.
\end{aligned}$$

From now on, we fix $x = x_F := \arg \min F$ and consider the expectation

$$\begin{aligned}
\sum_{t=1}^T \alpha_t (\mathbb{E} F(x_t) - F_*) &\leq \gamma_1 D^H(x_F, x_1) + \frac{2F_*}{\Gamma} \sum_{t=1}^T \beta_t \\
&+ \beta_T \left(H(\hat{x}_0) - \min H - \frac{F_*}{\Gamma} \right) + \sum_{t=1}^{T-1} \frac{\beta_t}{\Gamma} (\mathbb{E} F(x_{t+1}) - F_*) + \sum_{t=1}^T \frac{\alpha_t^2 \mathbb{E} \|g_t\|_*^2}{\gamma_t}.
\end{aligned}$$

We also notice that $H(\hat{x}_0) - \min H \leq \frac{F(x_H)}{\Gamma}$ if we write $x_H = \arg \min_{x \in \mathcal{X}} H(x)$. We rearrange the previous equation in the following way

$$\begin{aligned}
& \sum_{t=1}^T \frac{\alpha_t}{2} (\mathbb{E}F(x_t) - F_\star) \\
& \leq \gamma_1 D^H(x_F, x_1) + \frac{2F_\star}{\Gamma} \sum_{t=1}^T \beta_t + \sum_{t=1}^T \frac{\alpha_t^2 \mathbb{E} \|g_t\|_*^2}{\gamma_t} + \beta_T \left(H(\hat{x}_0) - \min H - \frac{F_\star}{\Gamma} \right) \\
& + \sum_{t=1}^{T-1} \frac{\beta_{t-1}}{\Gamma} (\mathbb{E}F(x_t) - F_\star) - \sum_{t=1}^T \frac{\alpha_t}{2} (\mathbb{E}F(x_t) - F_\star) \\
& \leq \gamma_1 D^H(x_F, x_1) + \frac{2F_\star}{\Gamma} \sum_{t=1}^T \beta_t + \sum_{t=1}^T \frac{\alpha_t^2}{\gamma_t} (\mathcal{M} \mathbb{E}F(x_t)) + \beta_T \left(H(\hat{x}_0) - \min H - \frac{F_\star}{\Gamma} \right) \quad (3.2) \\
& + \sum_{t=1}^{T-1} \frac{\beta_{t-1}}{\Gamma} (\mathbb{E}F(x_t) - F_\star) - \sum_{t=1}^T \frac{\alpha_t}{2} (\mathbb{E}F(x_t) - F_\star) \\
& \leq \gamma_1 D^H(x_F, x_1) + \frac{2F_\star}{\Gamma} \sum_{t=1}^T \beta_t + \sum_{t=1}^T \frac{\alpha_t^2}{\gamma_t} (\mathcal{M} F_\star) + \beta_T \left(H(\hat{x}_0) - \min H - \frac{F_\star}{\Gamma} \right) \\
& + \sum_{t=1}^T \left(\frac{\mathcal{M} \alpha_t^2}{\gamma_t} - \frac{\alpha_t}{4} \right) (\mathbb{E}F(x_t) - F_\star) + \sum_{t=1}^{T-1} \left(\frac{\beta_{t-1}}{\Gamma} - \frac{\alpha_t}{4} \right) (\mathbb{E}F(x_t) - F_\star).
\end{aligned}$$

Now, for any $n \geq 0$, if we consider the polynomial step-size $\alpha_t \propto t^n, \beta_t \propto \mu t^n, \gamma_t \propto \mu t^{n+1}/n$, we notice the two terms in the last line will become negative when t grows. More precisely, we note that

$$\begin{aligned}
\frac{\mathcal{M} \alpha_t^2}{\gamma_t} \leq \frac{\alpha_t}{4} & \Leftrightarrow 4\mathcal{M} \leq \frac{\gamma_t}{\alpha_t} \propto \frac{\mu t}{n}, \\
\frac{\alpha_t}{4} \geq \frac{\beta_{t-1}}{\Gamma} & \Leftrightarrow t^n \geq \frac{4\mu}{\Gamma} (t-1)^n.
\end{aligned}$$

Since we supposed that the choice of μ satisfies the regime $2\mathcal{M}n/T \leq \mu \leq \frac{\Gamma}{2}$, we know there exists a transition time $T_0 \propto \lceil \mathcal{M}n/\mu \rceil$ where the last two terms become negative. Now if we introduce the transition time in (3.2), we obtain

$$\begin{aligned}
\mathbb{E} \left[\sum_{t=1}^T t^n (F(x_t) - F_\star) \right] & \lesssim \mu \left(\frac{1}{n} D^H(x_F, x_1) + \frac{F_\star}{\Gamma} \sum_{t=1}^{T-1} t^n \right) + \sum_{t=1}^T \frac{\mathcal{M} n t^{2n}}{\mu t^{n+1}} F_\star \\
& + \mu \left(H(\hat{x}_0) - H_\star - \frac{F_\star}{\Gamma} \right) T^n + \mathcal{L} \sum_{t=1}^{T_0} \frac{\mathcal{M} n t^{2n}}{\mu t^{n+1}}.
\end{aligned}$$

Note that

$$\sum_{t=1}^{T_0} \frac{\mathcal{M} n t^{2n}}{\mu t^{n+1}} \propto \frac{\mathcal{M} n}{\mu} \sum_{t=1}^{T_0} t^{n-1} \propto \frac{\mathcal{M}}{\mu} T_0^n.$$

We conclude by using the convexity of F with $x_T^{ag} = \frac{\sum \alpha_t x_t}{\sum \alpha_t}$

$$\begin{aligned} & \mathbb{E}[F(x_T^{ag})] - F_\star \\ &= O\left(\left(\frac{\mathcal{M}n}{\mu T} + \frac{\mu}{\Gamma}\right)F_\star + \frac{\mu(H(\hat{x}_0) - H_\star)}{T} + \mathcal{L}\left(\frac{\mathcal{M}n}{\mu T}\right)^{n+1} + \frac{\mu D^H(x_F, x_1)}{T^{n+1}}\right). \end{aligned}$$

□

Note that if \hat{x}_0 is known, the second term disappears by choosing H centered at \hat{x}_0 . We illustrate the previous result with concrete choices of (n, μ) in the next corollary.

Corollary 3.1. *Under the previous assumption of relative scale setting, we choose $(n, \mu) = (3, \frac{1}{\sqrt{T}})$. If T is large enough i.e. $T \geq \max\left(36\mathcal{M}^2, \frac{4}{\Gamma^2}\right)$ then our algorithm yields*

$$\mathbb{E}[F(x_T^{ag})] - F_\star = O\left(\left(\mathcal{M} + \frac{1}{\Gamma}\right)\frac{F_\star}{\sqrt{T}} + \frac{H(\hat{x}_0) - H_\star}{\sqrt{T}} + \frac{\mathcal{L}\mathcal{M}^4}{T^2} + \frac{\mu D^H(x_F, x_1)}{T^4}\right).$$

Proof. We apply the previous Theorem and verify that, for $\mu = \frac{1}{\sqrt{T}}$ and $n = 3$,

$$2\mathcal{M}n/T \leq \mu \leq \frac{\Gamma}{2} \quad \Leftrightarrow \quad \frac{6\mathcal{M}}{T} \leq \frac{1}{\sqrt{T}} \leq \frac{\Gamma}{2} \quad \Leftrightarrow \quad \frac{36\mathcal{M}}{T} \leq 1 \leq \frac{\Gamma^2 T}{4}$$

where we multiply by T in the last step. □

3.2. Accelerated Composite Mirror descent. For the smooth setting, we use the acceleration framework from [3] with step-size (2.1).

Algorithm 2 Accelerated Composite Mirror Descent

Require: Number of iterations $T \geq 0$, starting point $x_1 \in \mathcal{X}$, step-sizes $(\alpha_t, \beta_t, \gamma_t)_{t \geq 1}$, regularizer H

for $1 \leq t \leq T$ **do**

Compute $x_t^{md} = \frac{A_{t-1}}{A_t}x_t^{ag} + \frac{\alpha_t}{A_t}x_t$.

Compute $x_{t+1} = \arg \min_{x \in \mathcal{X}} \{\alpha_t \langle \nabla f_{\xi_t}(x_t^{md}), x \rangle + \beta_t H(x) + \gamma_t D^H(x, x_t)\}$.

Compute $x_{t+1}^{ag} = \frac{A_{t-1}}{A_t}x_t^{ag} + \frac{\alpha_t}{A_t}x_{t+1}$.

end for

Output x_{t+1}^{ag} .

As in the relative case, we show the convergence of the accelerated algorithm, by introducing a burn in time T_0 .

Theorem 3.2. *Assume that H is 1-strongly-convex with regard to $\|\cdot\|$ and that there exist $D, D_H, L, \sigma > 0$ such that for all $x, y \in \mathcal{X}$,*

$$\begin{aligned} D^F(x, y) &\leq \frac{L}{2}\|x - y\|^2, \\ \mathbb{E}_\xi \left[\|\nabla f_\xi(x) - \nabla F(x)\|_*^2 \right] &\leq \sigma^2, \\ \|x - y\|^2 &\leq D^2, \\ H(x_F) - H(x) &\leq D^2, \end{aligned}$$

where x_F is the minimum of F . Suppose Algorithm 2 runs under the oblivious step-size (2.1). For any $\mu > 0$, if $2\sqrt{L/\mu} \leq T$, then

$$\mathbb{E}[F(x_{T+1}^{ag})] - F_\star = O\left(\frac{\mu D^H(x_F, x_1)}{nT^{n+1}} + \mu D^2 + \frac{n\sigma^2}{\mu T} + \frac{LD^2}{T}\left(\frac{Ln}{\mu T^2}\right)^{n/2}\right)$$

Proof. Applying the proximal lemma to $u = x$, $y = x_t$, $u^\star = x_{t+1}$, $f(\cdot) = \alpha_t \langle \nabla f_{\xi_t}(x_t^{md}), \cdot \rangle + \beta_t H(\cdot)$, and $\nu(\cdot) = \gamma_t H(\cdot)$, we have, for all $x \in \mathcal{X}$,

$$\begin{aligned} & \alpha_t \langle \nabla f_{\xi_t}(x_t^{md}), x_{t+1} \rangle + \beta_t H(x_{t+1}) + \gamma_t D^H(x_{t+1}, x_t) + \beta_t D^H(x, x_{t+1}) \\ & \leq \alpha_t \langle \nabla f_{\xi_t}(x_t^{md}), x \rangle + \beta_t H(x) + \gamma_t [D^H(x, x_t) - D^H(x, x_{t+1})]. \end{aligned}$$

After rearranging, we have

$$\begin{aligned} & \alpha_t \langle \nabla F(x_t^{md}), x_{t+1} - x \rangle + \alpha_t \langle \Delta(x_t^{md}), x_{t+1} - x \rangle + \beta_t H(x_{t+1}) - \beta_t H(x) \\ & = \alpha_t \langle \nabla f_{\xi_t}(x_t^{md}), x_{t+1} - x \rangle + \beta_t H(x_{t+1}) - \beta_t H(x) \\ & \leq \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) - \gamma_t D^H(x_{t+1}, x_t), \end{aligned} \tag{3.3}$$

where the stochastic noise is defined as $\Delta(x_t^{md}) = \nabla f_{\xi_t}(x_t^{md}) - \nabla F(x_t^{md})$. Now we split the gradient term with $\nabla F(x_t^{md})$ in the following way

$$\begin{aligned} & \alpha_t \langle \nabla F(x_t^{md}), x_{t+1} - x \rangle \\ & = \alpha_t \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle + \alpha_t \langle \nabla F(x_t^{md}), x_t^{md} - x \rangle \\ & = \alpha_t \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle + \alpha_t \left(F(x_t^{md}) - F(x) + D^F(x, x_t^{md}) \right). \end{aligned}$$

The Bregman divergence $D^F(x, x_t^{md})$ is positive because F is convex, therefore we can reformulate (3.3) in the following way

$$\begin{aligned} & \alpha_t \left(F(x_t^{md}) - F(x) + \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle \right) + \beta_t H(x_{t+1}) - \beta_t H(x) \\ & \leq \alpha_t \langle \Delta(x_t^{md}), x - x_{t+1} \rangle + \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) - \gamma_t D^H(x_{t+1}, x_t). \end{aligned}$$

For the acceleration step, we recall that $x_{t+1}^{ag} = \frac{A_{t-1}}{A_t} x_t^{ag} + \frac{\alpha_t}{A_t} x_{t+1}$. From the smoothness of F , we know

$$\begin{aligned} F(x_{t+1}^{ag}) & \leq F(x_t^{md}) + \langle \nabla F(x_t^{md}), x_{t+1}^{ag} - x_t^{md} \rangle + \frac{L}{2} \|x_{t+1}^{ag} - x_t^{md}\|^2 \\ & = F(x_t^{md}) + \langle \nabla F(x_t^{md}), \frac{A_{t-1}}{A_t} x_t^{ag} + \frac{\alpha_t}{A_t} x_{t+1} - x_t^{md} \rangle + \frac{L}{2} \frac{\alpha_t^2}{A_t^2} \|x_{t+1} - x_t\|^2 \\ & = \frac{A_{t-1}}{A_t} [F(x_t^{md}) + \langle \nabla F(x_t^{md}), x_t^{ag} - x_t^{md} \rangle] \\ & \quad + \frac{\alpha_t}{A_t} [F(x_t^{md}) + \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle] + \frac{L}{2} \frac{\alpha_t^2}{A_t^2} \|x_{t+1} - x_t\|^2. \end{aligned}$$

By scaling every term by A_t , we know that, for all $x \in \mathcal{X}$,

$$\begin{aligned} & A_t \left(F(x_{t+1}^{ag}) - F(x) \right) \\ & \leq A_{t-1} \left(F(x_t^{md}) - F(x) + \langle \nabla F(x_t^{md}), x_t^{ag} - x_t^{md} \rangle \right) \\ & \quad + \alpha_t \left(F(x_t^{md}) - F(x) + \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle \right) + \frac{L}{2} \frac{\alpha_t^2}{A_t} \|x_{t+1} - x_t\|^2. \end{aligned}$$

Combining with previous equations, we have

$$\begin{aligned} & A_t \left(F(x_{t+1}^{ag}) - F(x) \right) + \beta_t \left(H(x_{t+1}) - H(x) \right) \\ & \leq A_{t-1} \left(F(x_t^{md}) - F(x) + \langle \nabla F(x_t^{md}), x_t^{ag} - x_t^{md} \rangle \right) + \frac{L}{2} \frac{\alpha_t^2}{A_t} \|x_{t+1} - x_t\|^2 \\ & \quad + \alpha_t \langle \Delta(x_t^{md}), x - x_{t+1} \rangle + \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) - \gamma_t D^H(x_{t+1}, x_t). \end{aligned}$$

We also notice that by the convexity of F

$$\langle \nabla F(x_t^{md}), x_t^{ag} - x_t^{md} \rangle \leq F(x_t^{ag}) - F(x_t^{md}),$$

which implies

$$\begin{aligned} & A_t \left(F(x_{t+1}^{ag}) - F(x) \right) + \beta_t \left(H(x_{t+1}) - H(x) \right) \\ & \leq A_{t-1} \left(F(x_t^{ag}) - F(x) \right) + \frac{L}{2} \frac{\alpha_t^2}{A_t} \|x_{t+1} - x_t\|^2 + \alpha_t \langle \Delta(x_t^{md}), x - x_t \rangle \\ & \quad + \alpha_t \langle \Delta(x_t^{md}), x_t - x_{t+1} \rangle + \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) - \gamma_t D^H(x_{t+1}, x_t). \end{aligned}$$

We use the strong convexity of H and split the term into two parts

$$\gamma_t D^H(x_{t+1}, x_t) \geq \frac{\gamma_t}{4} \|x_{t+1} - x_t\|^2 + \frac{\gamma_t}{4} \|x_{t+1} - x_t\|^2,$$

and we apply Young's inequality

$$\alpha_t \langle \Delta(x_t^{md}), x_t - x_{t+1} \rangle - \frac{\gamma_t}{4} \|x_{t+1} - x_t\|^2 \leq \frac{2\alpha_t^2}{\gamma_t} \|\Delta(x_t^{md})\|_*^2.$$

We use the other $\gamma_t/4$ to balance the $L\alpha_t^2/2A_t$ term. In the end, we obtain for all $x \in \mathcal{X}$

$$\begin{aligned} & A_t \left(F(x_{t+1}^{ag}) - F(x) \right) - A_{t-1} \left(F(x_t^{ag}) - F(x) \right) \\ & \leq \beta_t \left(H(x_{t+1}) - H(x) \right) + \left(\frac{L}{2} \frac{\alpha_t^2}{A_t} - \frac{\gamma_t}{4} \right) \|x_{t+1} - x_t\|^2 + \alpha_t \langle \Delta(x_t^{md}), x - x_t \rangle \\ & \quad + \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) + \frac{2\alpha_t^2}{\gamma_t} \|\Delta(x_t^{md})\|_*^2. \end{aligned}$$

Adding all terms from $t = 1$ to $t = T$, we obtain

$$\begin{aligned}
& A_T \left([F(x_{t+1}^{ag})] - F(x) \right) + \gamma_T D^H(x, x_{T+1}) \\
& \leq \gamma_1 D^H(x, x_1) + \sum_{t=1}^T \alpha_t \langle \Delta(x_t^{md}), x - x_t \rangle + \sum_{t=1}^T \beta_t \left(H(x) - H(x_{t+1}) \right) \\
& + \sum_{t=1}^T \frac{2\alpha_t^2}{\gamma_t} \|\Delta(x_t^{md})\|_*^2 + \sum_{t=1}^T \left(\frac{L}{2} \frac{\alpha_t^2}{A_t} - \frac{\gamma_t}{4} \right) \|x_{t+1} - x_t\|^2 \\
& + \sum_{t=1}^{T-1} (\gamma_{t+1} - \gamma_t - \beta_t) D^H(x, x_{t+1}).
\end{aligned}$$

We know that the conditional expectation on t for $\mathbb{E}[\alpha_t \langle \Delta(x_t^{md}), x - x_t \rangle] = 0$. By the tower law, it disappears when we consider the sum of the expectation. By fixing $x = x_F := \arg \min F$, we obtain

$$\begin{aligned}
& A_T \left(\mathbb{E} [F(x_{t+1}^{ag})] - F_\star \right) \\
& \leq \gamma_1 D^H(x_F, x_1) + \sum_{t=1}^T \beta_t \mathbb{E} \left(H(x_F) - H(x_{t+1}) \right) + \sum_{t=1}^T \frac{2\sigma^2 \alpha_t^2}{\gamma_t} \\
& + \sum_{t=1}^T \left(\frac{L}{2} \frac{\alpha_t^2}{A_t} - \frac{\gamma_t}{4} \right) \|x_{t+1} - x_t\|^2 + \sum_{t=1}^{T-1} (\gamma_{t+1} - \gamma_t - \beta_t) D^H(x_F, x_{t+1}).
\end{aligned}$$

Recall that for all $x, y \in \mathcal{X}$, $H(x_F) - H(x), \|x - y\| \leq D^2$. Now for any $n \geq 0$, if we consider the polynomial step-size $\alpha_t \propto t^n, \beta_t \propto \mu t^n, \gamma_t \propto \mu t^{n+1}/n$, we look at the transition time $T_0 := \arg \max_t \{2L \leq \frac{\gamma_t A_t}{\alpha_t^2}\}$. For $\mu > 0$,

$$2L \lesssim \frac{\gamma_t A_t}{\alpha_t^2} \Leftrightarrow 2L \lesssim \frac{\mu t^{2n+2}}{n^2 t^{2n}} \Leftrightarrow \frac{\sqrt{2L}n}{\sqrt{\mu}} \lesssim t.$$

Therefore $T_0 \propto n \sqrt{\frac{L}{\mu}}$, and we have

$$\begin{aligned}
& A_T \left(\mathbb{E} [F(x_{t+1}^{ag})] - F_\star \right) \\
& \leq \gamma_1 D^H(x_F, x_1) + D^2 \sum_{t=1}^T \beta_t + D^2 \sum_{t=1}^{T_0} \left(\frac{L}{2} \frac{\alpha_t^2}{A_t} - \frac{\gamma_t}{4} \right) \\
& + \sum_{t=1}^T \frac{2\sigma^2 \alpha_t^2}{\gamma_t} + \sum_{t=1}^{T-1} (\gamma_{t+1} - \gamma_t - \beta_t) D^H(x_F, x_{t+1}).
\end{aligned}$$

The remaining step is to simplify the transition time term. We note that

$$\sum_{t=1}^{T_0} \frac{\alpha_t^2}{A_t} \propto \sum_{t=1}^{T_0} \frac{nt^{2n}}{t^{n+1}} \propto n \sum_{t=1}^{T_0} t^{n-1} \propto T_0^n$$

By diving every term by A_T , we conclude that

$$\mathbb{E}[F(x_{T+1}^{ag})] - F_\star \lesssim \frac{\mu D^H(x_F, x_1)}{nT^{n+1}} + \mu D^2 + \frac{n\sigma^2}{\mu T} + \frac{LD^2}{T} \left(\frac{Ln}{\mu T^2} \right)^{n/2}.$$

□

We would like to underline that transition time analysis can co-exist with the original analysis [3]. We will illustrate the previous result by choosing (n, μ) .

Corollary 3.2. *Under the assumptions of Theorem 3.2, if we tune $(n, \mu) = \left(2, \frac{1}{\sqrt{T}}\right)$ (with $3L^{2/3} \leq T$), then*

$$\mathbb{E}[F(x_{T+1}^{ag})] - F_\star = O\left(\frac{D^H(x_F, x_1)}{T^{3.5}} + \frac{\sigma^2 + D^2}{\sqrt{T}} + \frac{LD^2}{T} \left(\frac{L}{T^{3/2}}\right)\right).$$

Proof. We apply the previous Theorem with $n = 2$ and $\mu = \frac{1}{\sqrt{T}}$, we note that

$$2\sqrt{L/\mu} \leq T \Leftrightarrow 4L\sqrt{T} \leq T^2 \Leftrightarrow 4L \leq T^{3/2}.$$

□

3.3. Composite Mirror descent for Relative smoothness. For the relative smoothness setting, we recall for each random sampling ξ , f_ξ is L relatively smooth with respect to H if for all $x, y \in \mathcal{X}$,

$$D^{f_\xi}(x, y) \leq L_\star D^H(x, y).$$

The concept of randomness with relative smoothness is more subtle than the usual way since the standard variance could be arbitrary large in the relative smooth case. We will consider the new variance definition from [6], we suppose that there exists η_0 and σ^2 such that, for all $x \in \mathcal{X}$ and $\eta < \eta_0$,

$$\mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] \leq 2\eta^2 \sigma^2.$$

For our composite mirror descent, we will also suppose that for all $\eta > 0$

$$\mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] \leq 2\eta^2 G^2.$$

The constant G could be much larger than σ , but we will see that its impact is limited in our convergence result.

Algorithm 3 Composite Mirror descent for Relative smoothness

Require: Number of iterations $T \geq 0$, starting point $x_1 \in \mathcal{X}$, step-sizes $(\alpha_t, \beta_t, \gamma_t)_{t \geq 1}$, regularizer H

for $1 \leq t \leq T$ **do**

$$x_{t+1} = \arg \min_{x \in \mathcal{X}} \{ \alpha_t \langle \nabla f_\xi(x_t), x \rangle + \beta_t H(x) + \gamma_t D^H(x, x_t) \}$$

end for

Output $x_t^{ag} := \frac{\sum_{t=1}^T \alpha_t x_t}{\sum_{t=1}^T \alpha_t}.$

Theorem 3.3. *For the bounded relative scale case, assume that there exists $\eta_0, \sigma, G, L_\star > 0$ such that, for all $x, y \in \mathcal{X}$,*

$$\begin{aligned} \forall \xi, \quad D^{f_\xi}(x, y) &\leq L_\star D^H(x, y), \\ \forall \eta < \eta_0, \quad \mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] &\leq 2\eta^2 \sigma^2, \\ \forall \eta, \quad \mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] &\leq 2\eta^2 G^2, \\ \exists \hat{x}_0 \in \mathcal{X}, \nabla H(\hat{x}_0) &= 0. \end{aligned}$$

Suppose Algorithm 1 runs under the oblivious step-size (2.1). For any $\mu > \max\left(\frac{L_\star}{T}, \frac{\eta_0}{T}\right)$,

$$\frac{\sum_{t=1}^T nt^n \mathbb{E}[D^F(x_\star, x_t)]}{T^{n+1}} = \mathcal{O} \left(\frac{\mu D^H(x_F, x_1)}{nT^{n+1}} + \mu D^2 + \frac{\sigma^2 n}{\mu T} + \frac{G^2}{n\mu T^{n+1}} \max \left\{ \frac{L^n}{\mu^n}, \frac{\eta_0^n}{\mu^n} \right\} \right).$$

Proof. Since we supposed that the mirror descent steps stay in the interior of the domain \mathcal{X} , we know that the gradients at iteration t is zero at x_{t+1}

$$\begin{aligned} x_{t+1} &= \arg \min_{x \in \mathcal{X}} \{ \alpha_t \langle \nabla f_\xi(x_t), x \rangle + \beta_t H(x) + \gamma_t D^H(x, x_t) \} \\ \implies \alpha_t \nabla f_\xi(x_t) + \beta_t \nabla H(x_{t+1}) + \gamma_t (\nabla H(x_{t+1}) - \nabla H(x_t)) &= 0 \\ \Leftrightarrow \alpha_t \nabla f_\xi(x_t) = \gamma_t [\nabla H(x_t) - \nabla H(x_{t+1})] - \beta_t \nabla H(x_{t+1}) \\ \Leftrightarrow (\beta_t + \gamma_t) \nabla H(x_{t+1}) &= \gamma_t \nabla H(x_t) - \alpha_t \nabla f_\xi(x_t). \end{aligned} \tag{3.4}$$

Now, we apply the proximal lemma 2.2 where $u = x$, $y = x_t$, $u^\star = x_{t+1}$, $f(\cdot) = \alpha_t \langle \nabla f_\xi(x_t), \cdot \rangle + \beta_t H(\cdot)$, and $\nu(\cdot) = \gamma_t H(\cdot)$. We obtain for all $x \in \mathcal{X}$

$$\begin{aligned} &\alpha_t \langle \nabla f_\xi(x_t), x_{t+1} - x \rangle + \beta_t (H(x_{t+1}) - H(x)) + \gamma_t D^H(x_{t+1}, x_t) + \beta_t D^H(x, x_{t+1}) \\ &= \gamma_t [D^H(x, x_t) - D^H(x, x_{t+1})]. \end{aligned}$$

Considering $x = x_\star = x_F$ leads us to

$$\begin{aligned} &\alpha_t (\langle \nabla f_\xi(x_t), x_{t+1} - x_\star \rangle) + \gamma_t D^H(x_{t+1}, x_t) \\ &= \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t) D^H(x, x_{t+1}) + \beta_t (H(x_\star) - H(x_{t+1})). \end{aligned}$$

After arranging different terms, we have

$$\begin{aligned} &\alpha_t \langle \nabla f_\xi(x_t), x_t - x_\star \rangle + \alpha_t \langle \nabla f_\xi(x_t), x_{t+1} - x_t \rangle + \gamma_t D^H(x_{t+1}, x_t) \\ &= \gamma_t D^H(x_\star, x_t) - (\beta_t + \gamma_t) D^H(x_\star, x_{t+1}) + \beta_t (H(x_\star) - H(x_{t+1})). \end{aligned} \tag{3.5}$$

We replace the term $\alpha_t \nabla f_\xi(x_t)$ in (3.5) with the one in (3.4). We obtain

$$\begin{aligned} &\alpha_t \langle \nabla f_\xi(x_t), x_t - x_\star \rangle + \gamma_t \langle \nabla H(x_t) - \nabla H(x_{t+1}), x_{t+1} - x_t \rangle + \gamma_t D^H(x_{t+1}, x_t) \\ &= \gamma_t D^H(x_\star, x_t) - (\beta_t + \gamma_t) D^H(x_\star, x_{t+1}) \\ &+ \beta_t (H(x_\star) - H(x_{t+1})) + \beta_t \langle \nabla H(x_{t+1}), x_{t+1} - x_t \rangle. \end{aligned}$$

We introduce the following simplifications

$$\begin{aligned}\langle \nabla H(x_{t+1}) - \nabla H(x_t), x_{t+1} - x_t \rangle &= D^H(x_t, x_{t+1}) + D^H(x_{t+1}, x_t) \\ \langle \nabla H(x_{t+1}), x_{t+1} - x_t \rangle &= H(x_{t+1}) - H(x_t) + D^H(x_t, x_{t+1}).\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}& \alpha_t \langle \nabla f_\xi(x_t), x_t - x_\star \rangle - \gamma_t \left(D^H(x_t, x_{t+1}) + D^H(x_{t+1}, x_t) \right) + \gamma_t D^H(x_{t+1}, x_t) \\ &= \gamma_t D^H(x_\star, x_t) - (\beta_t + \gamma_t) D^H(x_\star, x_{t+1}) + \beta_t \left(H(x_\star) - H(x_t) \right) + \beta_t D^H(x_t, x_{t+1}),\end{aligned}$$

which can be rearranged into

$$\begin{aligned}\alpha_t \langle \nabla f_\xi(x_t), x_t - x_\star \rangle &= \gamma_t D^H(x_\star, x_t) - (\beta_t + \gamma_t) D^H(x_\star, x_{t+1}) \\ &\quad + \beta_t \left(H(x_\star) - H(x_t) \right) + (\beta_t + \gamma_t) D^H(x_t, x_{t+1}).\end{aligned}$$

Since $\nabla F(x_\star) = 0$, by considering the conditional expectation at iteration t , we obtain

$$\begin{aligned}\alpha_t \left(D^F(x_t, x_\star) + D^F(x_\star, x_t) \right) &= \alpha_t \langle \nabla F(x_t) - \nabla F(x_\star), x_t - x_\star \rangle \\ &= \mathbb{E}_\xi \left[\alpha_t \langle \nabla f_\xi(x_t), x_t - x_\star \rangle \right] \\ &= \gamma_t D^H(x_\star, x_t) - (\beta_t + \gamma_t) \mathbb{E} D^H(x_\star, x_{t+1}) \\ &\quad + \beta_t \left(H(x_\star) - H(x_t) \right) + (\beta_t + \gamma_t) \mathbb{E} D^H(x_t, x_{t+1}).\end{aligned}$$

Our goal is to rewrite $\mathbb{E} D^H(x_t, x_{t+1})$. We use the decomposition of $\nabla H(x_{t+1})$ in (3.4)

$$\begin{aligned}D^H(x_t, x_{t+1}) &= D^{H_\star}(\nabla H(x_{t+1}), \nabla H(x_t)) \\ &= D^{H_\star} \left(\frac{\gamma_t}{\beta_t + \gamma_t} \nabla H(x_t) - \frac{\alpha_t}{\beta_t + \gamma_t} \nabla f_\xi(x_t), \nabla H(x_t) \right) \\ &= D^{H_\star} \left(\frac{\gamma_t}{\beta_t + \gamma_t} \left(\nabla H(x_t) - \frac{\alpha_t}{\gamma_t} \nabla f_\xi(x_t) \right) + \frac{\beta_t}{\beta_t + \gamma_t} \nabla H(\hat{x}_0), \nabla H(x_t) \right),\end{aligned}$$

where $\hat{x}_0 \in \mathcal{X}$ satisfies $\nabla H(\hat{x}_0) = 0$. Since $D^{H_\star}(\cdot, \nabla H(x_t))$ is a convex function,

$$\begin{aligned}D^H(x_t, x_{t+1}) &\leq \frac{\gamma_t}{\beta_t + \gamma_t} D^{H_\star} \left(\nabla H(x_t) - \frac{\alpha_t}{\gamma_t} \nabla f_\xi(x_t), \nabla H(x_t) \right) \\ &\quad + \frac{\beta_t}{\beta_t + \gamma_t} D^{H_\star}(\nabla H(\hat{x}_0), \nabla H(x_t)).\end{aligned}$$

We also notice that

$$\begin{aligned}D^{H_\star}(\nabla H(\hat{x}_0), \nabla H(x_t)) &= D^H(x_t, \hat{x}_0) = H(x_t) - H(\hat{x}_0) - \langle \nabla H(\hat{x}_0), x_t \rangle \\ &= H(x_t) - H(\hat{x}_0).\end{aligned}$$

After simplification, we obtain

$$\begin{aligned}& \alpha_t \left(D^F(x_t, x_\star) + D^F(x_\star, x_t) \right) \\ &\leq \gamma_t D^H(x_\star, x_t) - (\beta_t + \gamma_t) \mathbb{E} D^H(x_\star, x_{t+1}) + \beta_t \left(H(x_\star) - H(\hat{x}_0) \right) \\ &\quad + \gamma_t \mathbb{E} D^{H_\star} \left(\nabla H(x_t) - \frac{\alpha_t}{\gamma_t} \nabla f_\xi(x_t), \nabla H(x_t) \right).\end{aligned}$$

The last term corresponds to the variance that we are trying to bound. If $\frac{\alpha_t}{\gamma_t} \geq \min\left(\frac{1}{2L_*}, \frac{1}{2\eta_0}\right)$, we simply use the assumption the quadratic upper bound G of the stochastic noise

$$D^{H_*}\left(\nabla H(x_t) - \frac{\alpha_t}{\gamma_t}\nabla f_\xi(x_t), \nabla H(x_t)\right) \leq \frac{\alpha_t^2 G^2}{\gamma_t^2}.$$

In the other case, we decompose in the following way,

$$\begin{aligned} & D^{H_*}\left(\nabla H(x_t) - \frac{\alpha_t}{\gamma_t}\nabla f_\xi(x_t), \nabla H(x_t)\right) \\ & \leq \frac{1}{2}D^{H_*}\left(\nabla H(x_t) - \frac{2\alpha_t}{\gamma_t}\nabla f_\xi(x_*) , \nabla H(x_t)\right) \\ & \quad + \frac{1}{2}D^{H_*}\left(\nabla H(x_t) - \frac{2\alpha_t}{\gamma_t}\left(\nabla f_\xi(x_t) - \nabla f_\xi(x_*)\right), \nabla H(x_t)\right). \end{aligned}$$

The first term is related to the variance assumption, if $\frac{2\alpha_t}{\gamma_t} \leq \frac{1}{\eta_0}$,

$$\frac{\gamma_t}{2}\mathbb{E}D^{H_*}\left(\nabla H(x_t) - \frac{2\alpha_t}{\gamma_t}\nabla f_\xi(x_*), \nabla H(x_t)\right) \leq \frac{\gamma_t}{2} \times \left(\frac{2\alpha_t}{\gamma_t}\right)^2 \sigma^2 = \frac{2\sigma^2\alpha_t^2}{\gamma_t}.$$

The second term is related to the relative smoothness and the cocoercivity of f_ξ . By Lemma 2.3, if $\frac{2\alpha_t}{\gamma_t} \leq \frac{1}{L_*}$, then

$$\begin{aligned} \frac{\gamma_t}{2}\mathbb{E}D^{H_*}\left(\nabla H(x_t) - \frac{2\alpha_t}{\gamma_t}\left(\nabla f_\xi(x_t) - \nabla f_\xi(x_*)\right), \nabla H(x_t)\right) & \leq \frac{\gamma_t}{2} \times \frac{2\alpha_t}{\gamma_t}\mathbb{E}D^{f_\xi}(x_t, x_*) \\ & = \alpha_t D^F(x_t, x_*). \end{aligned}$$

Combining everything, we see

$$\begin{aligned} \alpha_t D^F(x_*, x_t) & \leq \gamma_t D^H(x, x_t) - (\beta_t + \gamma_t)\mathbb{E}D^H(x, x_{t+1}) + \beta_t\left(H(x_*) - H(\hat{x}_0)\right) \\ & \quad + \begin{cases} \frac{2\sigma^2\alpha_t^2}{\gamma_t} & \text{if } \frac{\alpha_t}{\gamma_t} \leq \min\left(\frac{1}{2L_*}, \frac{1}{2\eta_0}\right) \\ \frac{G^2\alpha_t^2}{\gamma_t} & \text{else.} \end{cases} \end{aligned}$$

Now we note $T_0 = \arg \max_t \left\{ \frac{2\alpha_t}{\gamma_t} \geq \min\left\{\frac{1}{L}, \frac{1}{\eta_0}\right\} \right\}$. By the telescopic sum

$$\sum_{t=1}^T \alpha_t D^F(x_*, x_t) = \gamma_1 D^H(x_*, x_t) + \sum_t \beta_t \left(H(x_*) - H(\hat{x}_0)\right) + \sum_{t=T_0}^T \frac{2\sigma^2\alpha_t^2}{\gamma_t} + \sum_{t=1}^{T_0} \frac{G^2\alpha_t^2}{\gamma_t},$$

we finish the proof by considering the polynomial step-size $\alpha_t \propto t^n$, $\beta_t \propto \mu t^n$, $\gamma_t \propto \mu t^{n+1}/n$ with $n \geq 0$. We notice that

$$\sum_{t=1}^{T_0} \frac{\alpha_t^2 G^2}{\gamma_t} = \frac{G^2}{\mu} \left(\sum_{t=1}^{T_0} t^{n-1} \right) = \mathcal{O}\left(\frac{G^2 T_0^n}{n\mu}\right) = \mathcal{O}\left(\frac{G^2}{n\mu} \max\left\{\frac{L^n}{\mu^n}, \frac{\eta_0^n}{\mu^n}\right\}\right).$$

□

To illustrate the previous theorem and to compare the current result to other existing results, we assume convexity and choose concrete values for (n, μ) .

Corollary 3.3. *Under previous assumption, if $D^F(x_\star, \cdot)$ is convex and $T \geq \max(L_\star^2, \eta_0^2)$, by tuning $(n, \mu) = (3, \frac{1}{\sqrt{T}})$, our algorithm yields*

$$\mathbb{E}[D^F(x_\star, x_{T+1}^{ag})] = O\left(\frac{D^H(x_F, x_1)}{T^{3.5}} + \frac{\sigma^2 + D^2}{\sqrt{T}} + \frac{G^2}{T^3}\right)$$

Proof. We apply the previous Theorem with $n = 3$ and $\mu = \frac{1}{\sqrt{T}}$. We apply Jensen's inequality and note that

$$\mu \geq \max\left(\frac{L_\star}{T}, \frac{\eta_0}{T}\right) \Leftrightarrow T \geq \max(L_\star^2, \eta_0^2)$$

□

4. GRID SEARCH

When the function value F can be evaluated, Nemirovski's grid method offers a parameter-independent algorithmic approach. The method operates by running K concurrent sessions over N iterations, with each session using a distinct value of the regularization parameter μ . The algorithm then selects the optimal solution from among these K candidate outputs.

The standard implementation sets the number of parallel sessions logarithmically in relation to the total iterations: $K := O(\log(T))$. However, when rough estimates of the problem parameters are available, a more informed choice of K may yield better performance. The algorithm presented below accommodates an arbitrary selection of K to allow for such flexibility.

Algorithm 4 Grid Search Algorithm

Require: Total iteration number $T \geq 0$, number of parallel session $K \geq 1$, Starting point $x_1 \in \mathcal{X}$, algorithm \mathcal{A} .

Find N such that $T/2 \leq KN < T$.

Consider initial start point: $x_1^0 = x_1$.

for $k \in [-\lfloor K/2 \rfloor + 1, \lfloor K/2 \rfloor - 1]$ **do**

 Run Algorithm \mathcal{A} with N iterations and regularizing coefficient $\mu = 2^k$, obtain x_{N+1}^k as output

end for

Compute the index $k_{out} = \arg \max_k F(x_{N+1}^k)$.

Return $x_{out} = x_{N+1}^{k_{out}}$.

The general idea of the parallelization algorithm is to use an exponential range of regularizer values μ to find the optimal one. Before giving the full version, we present first a simplified version of the Grid search lemma.

Lemma 4.1. *Consider an algorithm \mathcal{A} that takes as inputs the number of iterations N and a regularization parameter $\mu > 0$, and denote by x_N^μ the output of the algorithm. Assume that the algorithm \mathcal{A} satisfies, for all $N \geq 1, \mu > 0$,*

$$\mathbb{E}[F(x_N^\mu)] - F_\star \leq \mu A + \frac{B}{\mu N}$$

with $A, B \geq 0$ being positive constants. For simplicity, we suppose that $\log(T)$ is an integer. If

$$T \geq \max \left\{ \left(\frac{A}{B} \right)^{1/3}, \left(\frac{B}{A} \right)^{1/4} \right\}, \quad (4.1)$$

then running the Algorithm 4 with $K = 4 \log(T)$ ensures that, within T iterations, the output x_{out} satisfies

$$\mathbb{E}[F(x_{out})] - F_\star = O \left(\sqrt{\frac{AB \log(T)}{T}} \right).$$

Proof. Intuitively, the optimal value of μ is $\sqrt{\frac{B}{AN}}$. Therefore, our goal is to show that there exists $k \in [-\log(T), \log(T)]$ (logarithm base 2) such that

$$\frac{1}{2} \sqrt{\frac{B}{AN}} \leq \mu = 2^k \leq \sqrt{\frac{B}{AN}}. \quad (4.2)$$

To show the existence of such a k , we need to check the two extreme values $\mu = 2^{2 \log(T)}$ and $\mu = 2^{-2 \log(T)}$ (recall that $T \propto KN = 4N \log(T)$)

$$\text{Case } \mu = 2^{2 \log T}: \quad \mu \geq \frac{1}{2} \sqrt{\frac{B}{AN}} \Leftrightarrow T^2 \geq \frac{1}{2} \sqrt{\frac{B}{AN}} \Leftrightarrow T^5 \geq \log(T) \left(\frac{B}{A} \right) \Leftrightarrow T^4 \geq \frac{B}{A}.$$

$$\text{Case } \mu = 2^{-2 \log T}: \quad \mu \leq \sqrt{\frac{B}{AN}} \Leftrightarrow \frac{1}{T^2} \leq \sqrt{\frac{B}{AN}} \Leftrightarrow \frac{1}{T^4} \leq \frac{B}{AN} \Leftrightarrow \frac{A}{B} \leq T^3.$$

We notice that both conditions are satisfied when T is large enough in (4.1). Therefore, there exists k verifying (4.2). ($\mu = 2^k \approx \sqrt{\frac{B}{AN}}$) If we consider the output of the k -th run

$$\begin{aligned} \mathbb{E}[F(x_N^k)] - F_\star &\leq \mu A + \frac{B}{\mu N} = 2^k A + \frac{B}{2^k N} \\ &\leq \sqrt{\frac{AB}{N}} + 2 \sqrt{\frac{AB}{N}} = 3 \sqrt{\frac{AB}{N}}, \end{aligned}$$

and we conclude by recalling $T = 4N \log(T)$ and $F(x_{out}) \leq F(x_N^k)$,

$$\mathbb{E}[F(x_{out})] - F_\star \leq 3 \sqrt{\frac{4AB \log(T)}{T}} = 6 \sqrt{\frac{AB \log(T)}{T}}.$$

□

It is interesting to note that condition (4.1) scales differently than the initial hypothesis yielding $\mu \propto \frac{1}{\sqrt{T}}$

$$T \geq \max \{A, B\},$$

therefore the Grid search Algorithm could be very powerful in practice when we have a reliable way of comparing different values of F . Now, we present the Grid search lemma, where multiple terms could be involved.

Lemma 4.2. *Consider an algorithm \mathcal{A} that takes as inputs the number of iterations N and a regularization parameter $\mu > 0$, and denote by x_N^μ the output of the algorithm. Assume that algorithm \mathcal{A} satisfies, for all $N \geq 1, \mu > 0$,*

$$\mathbb{E}[F(x_N^\mu)] - F_\star \leq \mu A + \frac{B}{\mu^{\beta_1} N^{\alpha_1}} + \frac{C}{\mu^{\beta_2} N^{\alpha_2}}$$

with $A, B, C, \alpha_1, \alpha_2, \beta_1, \beta_2 \geq 0$ being positive constants. If the following condition is satisfied

$$\begin{aligned} K + \frac{2\alpha_1}{1+\beta_1} \log N &\geq \frac{2}{1+\beta_1} \log \left(\frac{2B}{A} \right) + 2, \\ K + \frac{2\alpha_2}{1+\beta_2} \log N &\geq \frac{2}{1+\beta_2} \log \left(\frac{2C}{A} \right) + 2, \\ K &\geq 2 \min \left(\frac{1}{1+\beta_1} \log \left(\frac{2AN^{\alpha_1}}{B} \right), \frac{1}{1+\beta_2} \log \left(\frac{2AN^{\alpha_2}}{C} \right) \right), \end{aligned} \quad (4.3)$$

then running Algorithm 4 ensures that, within T iterations, the output x_{out} satisfies

$$\mathbb{E}[F(x_{out})] - F_\star = O \left(\frac{A^{\frac{\beta_1}{1+\beta_1}} B^{\frac{1}{1+\beta_1}} K^{\frac{\alpha_1}{1+\beta_1}}}{T^{\frac{\alpha_1}{1+\beta_1}}} + \frac{A^{\frac{\beta_2}{1+\beta_2}} C^{\frac{1}{1+\beta_2}} K^{\frac{\alpha_2}{1+\beta_2}}}{T^{\frac{\alpha_2}{1+\beta_2}}} \right).$$

Proof. The main difference with Lemma 4.1 is that multiple terms involving μ are in the equation and therefore there are multiple critical values of μ depending on A, B, C, N .

In condition (4.3), the first two lines mean that, when μ is at its maximum value, μA is the biggest term. In fact,

$$\begin{aligned} \text{Case } \mu = 2^{\lfloor K/2 \rfloor - 1}: \quad \mu A &\geq \frac{B}{\mu^{\beta_1} N^{\alpha_1}} \Leftrightarrow \mu \geq \frac{B^{\frac{1}{1+\beta_1}}}{A^{\frac{1}{1+\beta_1}} N^{\frac{\alpha_1}{1+\beta_1}}} \\ &\Leftrightarrow 2^{\lfloor K/2 \rfloor - 1} \geq \frac{B^{\frac{1}{1+\beta_1}}}{A^{\frac{1}{1+\beta_1}} N^{\frac{\alpha_1}{1+\beta_1}}} \\ &\Leftrightarrow \lfloor K/2 \rfloor \geq \frac{1}{1+\beta_1} \log \left(\frac{2B}{AN^{\alpha_1}} \right) \\ &\Leftrightarrow K \geq \frac{2}{1+\beta_1} \log \left(\frac{2B}{AN^{\alpha_1}} \right) + 2 \\ &\Leftrightarrow K + \frac{2\alpha_1}{1+\beta_1} \log N \geq \frac{2}{1+\beta_1} \log \left(\frac{2B}{A} \right) + 2 \\ \mu A &\geq \frac{C}{\mu^{\beta_2} N^{\alpha_2}} \Leftrightarrow K + \frac{2\alpha_2}{1+\beta_2} \log N \geq \frac{2}{1+\beta_2} \log \left(\frac{2C}{A} \right) + 2. \end{aligned}$$

For the smallest value of μ , we require one of the two following conditions to be satisfied (we recall that $T/2 \leq KN \leq T$):

$$\begin{aligned} \text{Case } \mu = 2^{-\lfloor K/2 \rfloor + 1}: \quad \mu A \leq \frac{B}{\mu^{\beta_1} N^{\alpha_1}} &\Leftrightarrow 2^{-\lfloor K/2 \rfloor + 1} \leq \frac{B^{\frac{1}{1+\beta_1}}}{A^{\frac{1}{1+\beta_1}} N^{\frac{\alpha_1}{1+\beta_1}}} \\ &\Leftrightarrow \frac{1}{1+\beta_1} \log \left(\frac{2AN^{\alpha_1}}{B} \right) \leq K/2 \\ \mu A \leq \frac{C}{\mu^{\beta_2} N^{\alpha_2}} &\Leftrightarrow \frac{1}{1+\beta_2} \log \left(\frac{2AN^{\alpha_2}}{C} \right) \leq K/2. \end{aligned}$$

Since the dominant terms are different for the biggest and smallest values of k , we can consider the largest value of k where μA is not dominating the other two terms. In other words, we consider the largest k such that one of the two following conditions is satisfied

$$\begin{aligned} \frac{1}{2} \left(\frac{B}{AN^{\alpha_1}} \right)^{\frac{1}{1+\beta_1}} \leq \mu = 2^k &\leq \left(\frac{B}{AN^{\alpha_1}} \right)^{\frac{1}{1+\beta_1}}, \\ \frac{1}{2} \left(\frac{C}{AN^{\alpha_2}} \right)^{\frac{1}{1+\beta_2}} \leq \mu = 2^k &\leq \left(\frac{C}{AN^{\alpha_2}} \right)^{\frac{1}{1+\beta_2}}. \end{aligned} \quad (4.4)$$

Since it is the largest term not dominant, we know by replacing $\mu = 2^k$ by 2^{k+1} that μA will become dominant. Therefore, the other two terms with B and C cannot be much larger than μA . In other words, if one of the two previous conditions is satisfied with $\mu = 2^k$, then

$$\begin{aligned} 2^{k+1} &\geq \frac{B^{\frac{1}{1+\beta_1}}}{A^{\frac{1}{1+\beta_1}} N^{\frac{\alpha_1}{1+\beta_1}}} \implies 2^{1+\beta_1} \mu^{1+\beta_1} \geq \frac{B}{AN^{\alpha_1}} \implies 2^{1+\beta_1} \mu A \geq \frac{B}{\mu^{\beta_1} N^{\alpha_1}}, \\ 2^{k+1} &\geq \frac{C^{\frac{1}{1+\beta_2}}}{A^{\frac{1}{1+\beta_2}} N^{\frac{\alpha_2}{1+\beta_2}}} \implies 2^{1+\beta_2} \mu^{1+\beta_2} \geq \frac{C}{AN^{\alpha_2}} \implies 2^{1+\beta_2} \mu A \geq \frac{C}{\mu^{\beta_2} N^{\alpha_2}}. \end{aligned}$$

Therefore, we can consider the output of the k -th run

$$\begin{aligned} \mathbb{E}[F(x_N^k)] - F_\star &\leq \mu A + \frac{B}{\mu^{\beta_1} N^{\alpha_1}} + \frac{C}{\mu^{\beta_2} N^{\alpha_2}} \\ &\leq \mu A (1 + 2^{1+\beta_1} + 2^{1+\beta_2}). \end{aligned}$$

If the first line of (4.4) is satisfied, we have

$$\begin{aligned} \mathbb{E}[F(x_N^k)] - F_\star &\leq (1 + 2^{1+\beta_1} + 2^{1+\beta_2}) A^{\frac{\beta_1}{1+\beta_1}} B^{\frac{1}{1+\beta_1}} N^{-\frac{\alpha_1}{1+\beta_1}} \\ &\leq (1 + 2^{1+\beta_1} + 2^{1+\beta_2}) A^{\frac{\beta_1}{1+\beta_1}} B^{\frac{1}{1+\beta_1}} K^{\frac{\alpha_1}{1+\beta_1}} T^{-\frac{\alpha_1}{1+\beta_1}}. \end{aligned}$$

If the second line of (4.4) is satisfied, we have

$$\mathbb{E}[F(x_N^k)] - F_\star \leq (1 + 2^{1+\beta_1} + 2^{1+\beta_2}) A^{\frac{\beta_2}{1+\beta_2}} C^{\frac{1}{1+\beta_2}} K^{\frac{\alpha_2}{1+\beta_2}} T^{-\frac{\alpha_2}{1+\beta_2}}.$$

We conclude by merging the conditions and replacing the maximum of the two cases by their sum. \square

We note that the previous Lemma could work if we have more than three terms in the equation, the reasoning would be similar by distinguishing different cases. We also note that grid search could work even if the original algorithm \mathcal{A} does not have a guarantee

for all $\mu > 0$. The only important values of μ are the critical ones that we use in the previous proof.

For the rest of section, we will assume that is possible to evaluate the population function F in an efficient way and we will run grid search $K = 4\log(T)$, and show the derived convergence rates.

Optimization in relative scale We recall the initial result from Theorem 3.1

$$\mathbb{E}[F(x_T^{ag})] - F_\star = O\left(\mu\left(\frac{F_\star}{\Gamma} + \frac{H(\hat{x}_0) - H_\star}{T}\right) + \frac{\mathcal{M}nF_\star}{\mu T} + \mathcal{L}\left(\frac{\mathcal{M}n}{\mu T}\right)^{n+1} + \frac{\mu D^H(x_F, x_1)}{T^{n+1}}\right). \quad (4.5)$$

For the sake of simplicity, we would like to only choose the optimal value of μ to balance the coefficients in front of F_\star ($\frac{\mu}{\Gamma} \approx \frac{\mathcal{M}n}{\mu T}$). The final result might not be optimal for the other terms, but general conditions are easier to satisfy. We can balance the value of μ around $\mu = \sqrt{\frac{\mathcal{M}n\Gamma}{T}}$, and we note that this value of μ satisfies $2\mathcal{M}n/T \leq \mu \leq \frac{\Gamma}{2}$. The new corollary is as follows.

Corollary 4.1. *Assume that H is 1-strongly-convex with regard to $\|\cdot\|$ and there exist $\Gamma, \mathcal{M}, \mathcal{L} > 0$ such that, for all $x \in \mathcal{X}$,*

$$\begin{aligned} \Gamma(H(x) - \min H) &\leq F(x), \\ \mathbb{E}\left[\|\nabla f_\xi(x, \xi)\|_*^2\right] &\leq \mathcal{M}F(x), \\ F(x) - F_\star &\leq \mathcal{L}. \end{aligned}$$

Assume that T is large enough,

$$T \geq \max\left\{\left(\frac{1}{\Gamma\mathcal{M}n}\right)^{1/3}, \left(\Gamma\mathcal{M}n\right)^{1/4}\right\}.$$

Suppose Algorithm 1 runs under the oblivious step-size (2.1) with the Grid search Algorithm 4 and $K = 4\lceil\log T\rceil$. Then

$$\begin{aligned} \mathbb{E}[F(x_{out})] - F_\star &= O\left(\log^{1/2}(T)\sqrt{\frac{\mathcal{M}n}{\Gamma T}}F_\star + \log^{3/2}(T)\sqrt{\mathcal{M}n\Gamma}\frac{H(\hat{x}_0) - H_\star}{T\sqrt{T}}\right) \\ &\quad + O\left(\mathcal{L}\left(\frac{\mathcal{M}n\Gamma\log(T)}{T}\right)^{\frac{n+1}{2}} + \sqrt{\frac{\mathcal{M}n\Gamma D^H(x_F, x_1)}{T}}\left(\frac{\log(T)}{T}\right)^{n+1}\right). \end{aligned}$$

Proof. The proof strategy is similar to the one in Lemma 4.1. We are looking for k to balance $\frac{\mu}{\Gamma}$ with $\frac{\mathcal{M}n}{\mu T}$ while neglecting other terms. More precisely, we look for k such that

$$\frac{1}{2}\sqrt{\frac{\mathcal{M}n}{\Gamma T}} \leq \mu = 2^k \leq \sqrt{\frac{\mathcal{M}n}{\Gamma T}}$$

and we choose this value of μ for other terms such as $\frac{\mu D^H(x_F, x_1)}{T^{n+1}}$. \square

Optimization with smooth functions We recall the result from Theorem 3.2. If $2\sqrt{L/\mu} \leq T$, we know

$$\mathbb{E}[F(x_{T+1}^{ag})] - F_\star = O\left(\mu D^2 + \frac{n\sigma^2}{\mu T} + \frac{LD^2}{T}\left(\frac{Ln}{\mu T^2}\right)^{n/2} + \frac{\mu D^H(x_F, x_1)}{nT^{n+1}}\right).$$

In the smooth case, there are two important values of μ to consider. To optimize the choice of μ , we need to balance the values of μ around

$$\mu = \frac{\sigma}{D} \sqrt{\frac{n}{T}} \quad (\mu D \approx \frac{n\sigma^2}{\mu T})$$

and

$$\mu = \frac{Ln^{\frac{n}{n+2}}}{T^{\frac{2(n+1)}{n+2}}} \quad (\mu \approx \frac{L}{T} (\frac{Ln}{\mu T^2})^{n/2}).$$

Theorem 4.1. Assume that H is 1-strongly-convex with regard to $\|\cdot\|$ and that there exist $D, D_H, L, \sigma > 0$ such that, for all $x, y \in \mathcal{X}$,

$$\begin{aligned} D^F(x, y) &\leq \frac{L}{2} \|x - y\|^2, \\ \mathbb{E}_\xi \left[\|\nabla f_\xi(x) - \nabla F(x)\|_*^2 \right] &\leq \sigma^2, \\ \|x - y\|^2 &\leq D^2, \\ H(x_F) - H(x) &\leq D^2, \end{aligned}$$

where x_F is the minimum of F . Assume that T is large enough,

$$\begin{aligned} T &\geq \max \left\{ L^{\frac{2+n}{6+4n}} n^{\frac{n}{6+4n}}, \frac{n^{1/5} \sigma^{2/5}}{D^{\frac{2}{5}}} \right\}, \\ T &\geq \max \left\{ \frac{1}{L^{1+n/2} n^{n/2}}, \frac{D^2}{n\sigma^2} \right\}. \end{aligned}$$

Suppose Algorithm 2 runs under the oblivious step-size (2.1) with the Grid search Algorithm 4 and $K = 4\lceil \log T \rceil$. Then

$$\begin{aligned} \mathbb{E}[F(x_{out})] - F_* &= \mathcal{O} \left(\frac{LD^2 n^{\frac{n}{n+2}} \log^2(T)}{T^{\frac{2(n+1)}{n+2}}} + \frac{\sigma D \sqrt{n \log(T)}}{\sqrt{T}} \right) \\ &+ \mathcal{O} \left(\max \left\{ \frac{L \log^2(T)}{T^{\frac{2(n+1)}{n+2}}}, \frac{\sigma \log^{1/2}(T)}{D \sqrt{nT}} \right\} \frac{\log^{n+1}(T) D^H(x_F, x_1)}{T^{n+1}} \right). \end{aligned}$$

Proof. The proof strategy is similar to the one in Lemma 4.2. When $\mu = 2^{2\lceil \log T \rceil} \geq T^2$, we verify that

$$\begin{aligned} \mu D^2 \geq \frac{LD^2}{T} \left(\frac{Ln}{\mu T^2} \right)^{n/2} &\Leftrightarrow \mu^{2+n} T^{2+2n} \geq L^{2+n} n^n \Leftarrow T^{6+4n} \geq L^{2+n} n^n, \\ \mu D^2 \geq \frac{n\sigma^2}{\mu T} &\Leftrightarrow \mu^2 T \geq \frac{n\sigma^2}{D^2} \Leftarrow T^5 \geq \frac{n\sigma^2}{D^2}. \end{aligned}$$

When $\mu = 2^{-2\lceil \log T \rceil} \leq T^{-2}$, we verify that

$$\begin{aligned} \mu D^2 \leq \frac{LD^2}{T} \left(\frac{Ln}{\mu T^2} \right)^{n/2} &\Leftrightarrow \mu^{2+n} T^{2+2n} \leq L^{2+n} n^n \Leftarrow T^2 \geq \frac{1}{L^{2+n} n^n}, \\ \mu D^2 \leq \frac{n\sigma^2}{\mu T} &\Leftrightarrow \mu^2 T \leq \frac{n\sigma^2}{D^2} \Leftarrow T \geq \frac{D^2}{n\sigma^2}. \end{aligned}$$

Since we supposed that T is large enough to verify previous condition, now we choose the largest k such that one of the two following conditions is verified

$$\begin{aligned} \frac{1}{2} \frac{Ln^{\frac{n}{n+2}}}{T^{\frac{2(n+1)}{n+2}}} &\leq \mu = 2^k \leq \frac{Ln^{\frac{n}{n+2}}}{T^{\frac{2(n+1)}{n+2}}}, \\ \frac{1}{2} \frac{\sigma}{D} \sqrt{\frac{n}{T}} &\leq \mu = 2^k \leq \frac{\sigma}{D} \sqrt{\frac{n}{T}}. \end{aligned}$$

If the first line is verified, since $\mu \geq \frac{4L}{T^2}$, we can apply the result from Theorem 3.2,

$$\mathbb{E}[F(x_N^k)] - F_\star = \mathcal{O} \left(\frac{LD^2 n^{\frac{n}{n+2}}}{T^{\frac{2(n+1)}{n+2}}} + \frac{LD^H(x_F, x_1)}{T^{n+1+\frac{2(n+1)}{n+2}}} \right).$$

If the second line is verified, we can deduce that :

$$\mu \geq \frac{1}{2} \frac{\sigma}{D} \sqrt{\frac{n}{T}} \geq \frac{Ln^{\frac{n}{n+2}}}{T^{\frac{2(n+1)}{n+2}}} \geq \frac{4L}{T^2}.$$

Therefore,

$$\mathbb{E}[F(x_N^k)] - F_\star = \mathcal{O} \left(\frac{\sigma D \sqrt{n}}{\sqrt{T}} + \frac{\sigma D^H(x_F, x_1)}{D \sqrt{n} T^{n+3/2}} \right).$$

We conclude the proof by taking the two bounds and adding the poly-logarithmic terms. \square

Optimization with relative smooth functions We recall the initial result from Theorem 3.1, if $\mu \geq \max \left\{ \frac{L_\star}{T}, \frac{\eta_0}{T} \right\}$

$$\frac{\sum_{t=1}^T nt^n \mathbb{E}[D^F(x_\star, x_t)]}{T^{n+1}} = \mathcal{O} \left(\frac{\mu D^H(x_F, x_1)}{n T^{n+1}} + \mu D^2 + \frac{\sigma^2 n}{\mu T} + \frac{G^2}{n \mu T^{n+1}} \max \left\{ \frac{L^n}{\mu^n}, \frac{\eta_0^n}{\mu^n} \right\} \right). \quad (4.6)$$

For the sake of simplicity, we balance the value of $\mu = \frac{\sigma}{D} \sqrt{\frac{n}{T}}$ only to equalize two terms $\mu D^2 \approx \frac{\sigma^2 n}{\mu T}$. This choice of μ is not optimal for other terms, but the conditions are easier to be satisfied. The new corollary becomes

Corollary 4.2. *Assume that H is strictly-convex, $D^F(x_\star, \cdot)$ is convex, and there exist $\Gamma, \mathcal{M}, \mathcal{L} > 0$ such that, for all $x \in \mathcal{X}$,*

$$\begin{aligned} D^F(x, y) &\leq L_\star D^H(x, y), \\ \forall \eta < \eta_0, \quad \mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] &\leq 2\eta^2 \sigma^2, \\ \forall \eta, \quad \mathbb{E}_\xi \left[D^{H_\star} \left(\nabla H(x) - 2\eta \nabla f_\xi(x_F), \nabla H(x) \right) \right] &\leq 2\eta^2 G^2, \\ H(x_F) - \min H &\leq D^2. \end{aligned}$$

Assume that T is large enough,

$$T \geq \max \left\{ \left(\frac{D^2}{\sigma^2 n} \right)^{1/3}, \left(\frac{\sigma^2 n}{D^2} \right)^{1/4}, \frac{4L_\star^2 D^2}{\sigma^2 n}, \frac{4\eta_0^2 D^2}{\sigma^2 n} \right\}$$

Suppose Algorithm 1 runs under the oblivious step-size (2.1) with the Grid search Algorithm 4 and $K = 4\lceil \log T \rceil$. Then

$$\begin{aligned} \mathbb{E}[F(x_{out})] - F_\star &= O\left(\frac{\sigma D\sqrt{n} \log^{1/2}(T)}{\sqrt{T}} + \frac{\sigma \log^{n+1}(T)}{D\sqrt{n}T^{n+1/2}} D^H(x_F, x_1)\right) \\ &\quad + O\left(\frac{G^2 D \log^{2n}(T)}{\sigma n \sqrt{n} T^{n+1/2}} \max\left\{\frac{L\sigma}{D\sqrt{nT}}, \frac{\eta_0\sigma}{D\sqrt{nT}}\right\}^n\right). \end{aligned}$$

Proof. The proof strategy is similar to the one in Lemma 4.1. We are looking for k such that

$$\frac{1}{2} \frac{\sigma}{D} \sqrt{\frac{n}{T}} \leq \mu = 2^k \leq \frac{\sigma}{D} \sqrt{\frac{n}{T}}.$$

The extra condition is to verify that when T is large, we have

$$\frac{1}{2} \frac{\sigma}{D} \sqrt{\frac{n}{T}} \geq \max\left\{\frac{L_\star}{T}, \frac{\eta_0}{T}\right\} \Leftrightarrow \sqrt{T} \geq \left\{\frac{2L_\star D}{\sigma\sqrt{n}}, \frac{2\eta_0 D}{\sigma\sqrt{n}}\right\}.$$

□

5. NUMERICAL EXPERIMENTS

For simulation, we are interested in minimizing the maximum eigenvalue function over a hypercube. We can apply both of our methods here, as the feasible set is bounded and the objective function is bounded over this feasible set as well

$$\begin{aligned} \min_{x \in \mathbb{R}^{d \times d}} \quad & \lambda_{\max}(x) \\ \text{s.t.} \quad & \|x - A\|_\infty \leq \rho. \end{aligned}$$

In our simulation, to generate A , we will first generate random diagonal matrices then add Gaussian noise. When the Gaussian noise level is low, we refer to it as the sparse case. We mainly reproduce the setting from [4], more details can be found in Appendix A. We plot performance versus problem dimension and, having some flexibility on the degree of polynomials we are using in $\alpha_t \propto t^n, \gamma_t \propto t^{n+1}/n$ for the oblivious step-sizes, we study how that degree affects algorithm performance in practice.

5.1. Smoothed oracle. In the first part of the experiments, we run simulations using the smoothed stochastic oracle suggested in [4]. This reference provides an estimation of the smoothness constant L , but the authors mention that it could be overestimating the parameter in practice. As a benchmark, we have run the method in [8] using prior knowledge on L , and the Levy method [14] using prior knowledge on D .

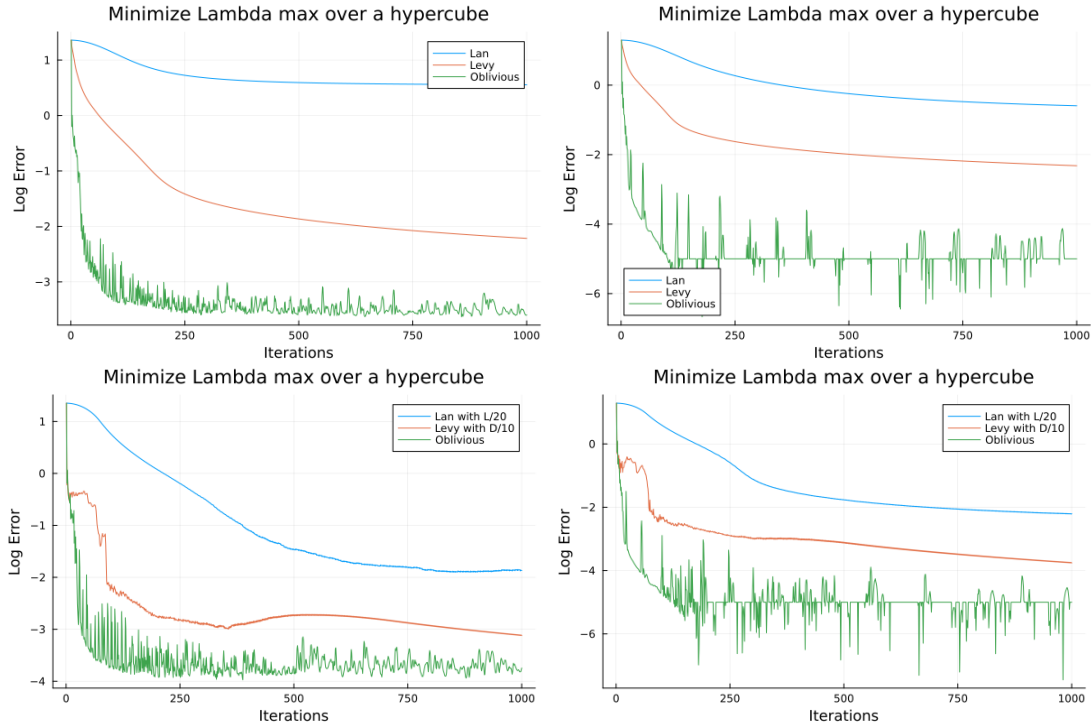


FIGURE 1. Performance on synthetic data, using stochastic smoothing in normal/sparse cases. Upper figures show performance under theoretical step-sizes, bottom figures use hyper-tuned step-sizes.

The first results in Fig. 5.1 show that the current over-estimation of L and D slows down significantly both Lan and Levy's algorithms. In order to give an advantage to these methods, we also investigate the use of hyper-parameter tuning. We observe that the sparse case is well handled by our algorithms, whereas the Levy adaptive algorithm still keeps conservative step sizes.

Dimension	Lan	Levy	Oblivious1	Oblivious2	Oblivious3
$d=50$	> 1000	125	22	29	39
$d=100$	> 1000	529	45	57	70
$d=150$	> 1000	679	106	163	244

TABLE 2. Number of iterations required to reach the precision $1e-2$ with stochastic smoothing.

We notice that, at least in high dimensions, these algorithms' performance is dramatically improved by hyper-parameter tuning (see Fig. 5.1). In particular, our method still appears to be substantially faster than the alternatives.

5.2. Power method oracle. We reproduce these results using the power method oracle, with mirror descent in relative scale [22] replacing Lan algorithm. The global phenomenon is similar as before, with L_* and D over-estimated by the theory and we need to use hyper-tuned step-sizes for Levy and relative scale algorithms to obtain good performance.

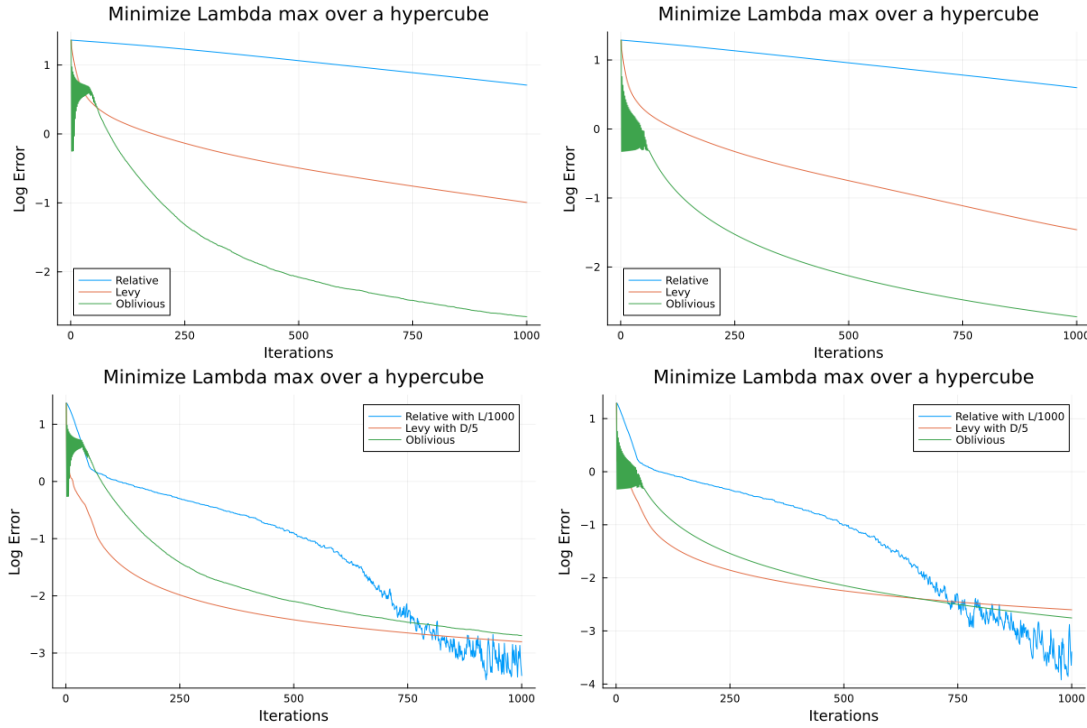


FIGURE 2. Performance of algorithms on synthetic data, using power method in the normal case and in the sparse case. The upper figures show performances when we use the parameters L_* and D provided by the theory, the bottom one are using hyper-tuned step-sizes.

However there are two major differences with the smooth case. (the objective function is not smooth anymore) First, we notice that our oblivious algorithm does not converge directly to the minimum as in the smooth case. Instead, our algorithm's first iterations are unstable and converges after a transition time, as our theory predicts. The other point we would like to underline is that power method oracles cost less to compute in practice, compared to the stochastic smoothing oracles, but the accuracy is also worse. More discussions and simulations results can be found in the supplementary material. [Appendix A](#)

REFERENCES

- [1] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* 2 (2009), 183–202.
- [2] A. Cutkosky, Anytime online-to-batch conversions, optimism, and acceleration, arXiv 1903.00974, 2019.
- [3] A. d’Aspremont, C. Guzmán, C. Lezane, Optimal algorithms for stochastic complementary composite minimization, arXiv 2211.01758, 2022.
- [4] A. d’Aspremont, N. El Karoui, A stochastic smoothing algorithm for semidefinite programming, arXiv 1204.0665, 2014.
- [5] J. Diakonikolas, C. Guzmán, Complementary composite minimization, small gradients in general norms, and applications, arXiv 2101.11041, 2023.
- [6] R.A. Dragomir, M. Even, H. Hendrikx, Fast stochastic bregman gradient methods: Sharp analysis and variance reduction, In: *International Conference on Machine Learning*, 2021.

- [7] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.* 12 (2011), 2121–2159.
- [8] S. Ghadimi, G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework, *SIAM J. Optim.* 22 (2012), 1469–1492.
- [9] S. Ghadimi, G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms, *SIAM J. Optim.* 23 (2013), 2061–2089.
- [10] A. Jacobsen, A. Cutkosky, Parameter-free mirror descent, arXiv 2203.00444, 2022.
- [11] P. Joulani, A. Raj, A. Gyorgy, C. Szepesvari, A simpler approach to accelerated optimization: iterative averaging meets optimism, In: Hal Daumé III and A. Singh (ed.), *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, pp. 4984–4993, PMLR, 2020.
- [12] A. Juditsky, Y. Nesterov, Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization, *Stochastic Systems* 4 (2014), 44–80.
- [13] G. Lan, An optimal method for stochastic composite optimization, *Math. Program.* 133 (2012), 1–33.
- [14] K.Y. Levy, A. Yurtsever, V. Cevher, Online adaptive methods, universality and acceleration, arXiv 1809.02864, 2018.
- [15] H. Lu, R.M. Freund, Y. Nesterov, Relatively smooth convex optimization by first order methods, and applications, *SIAM J. Optim.* 28 (2016), 333–354.
- [16] A. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, *SIAM J. Optim.* 15 (2004), 229–251.
- [17] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.* 19 (2009), 1574–1609.
- [18] A.S. Nemirovsky, D.B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, 1983.
- [19] Y. Nesterov, Gradient methods for minimizing composite functions, *Math. Program.* 140 (2013), 125–161.
- [20] Y. Nesterov, How to make the gradients small, *Optima.* 88 (2012), 10–11.
- [21] Y. Nesterov, Gradient methods for minimizing composite functions, *Math. Program.* 140 (2013), 125–161.
- [22] Y. Nesterov, A. Rodomanov, Randomized minimization of eigenvalue functions, arXiv 2301.08352, 2023.
- [23] B. Polyak, *Introduction to Optimization*, Optimization Software, Inc., New York 1987.
- [24] K. Scheinberg, D. Goldfarb, X. Bai, Fast first-order methods for composite convex optimization with backtracking, *Found. Comput. Math.* 14 (2014), 389–417.
- [25] A. Yurtsever, J.A. Tropp, O. Fercoq, M. Udell, V. Cevher, Scalable semi-definite programming, *SIAM J. Math. Data Sci.* 3 (2021), 171–200.

APPENDIX A. SIMULATIONS

We recall the simulation problem that we are solving for the stochastic smoothing problem.

$$\begin{aligned} \min_{x \in \mathbb{R}^{d \times d}} \quad & \lambda_{\max}(x) \\ \text{s.t.} \quad & \|x - A\|_{\infty} \leq \rho. \end{aligned}$$

Note that, when using the power method oracle, we are actually solving an approximation of $\lambda_{\max}(x)^2$.

Problem setting To generate A , we chose a diagonal matrix C with coefficients that decreases exponentially $C_{i,i} = \exp(-i)$, then we add Gaussian noise to each coefficients $A_{i,j} = C_{i,j} + \mathcal{N}(0, \sigma^2)$ ($\sigma = 0.2$ for the normal case and $\sigma = 0.05$ for the sparse case). We normalise and project to make A symmetric and maximum coefficient equal to 1, and we chose $\rho = \max(\text{diag}(A))/2$. For the stochastic smoothing oracle, we chose the target precision as $\epsilon = 1e - 2$. For the power method oracle, we use order $p = 21$.

Hyper-tuned step-sizes For Lan, Levy and relative scale algorithm, we have run simulations with fewer iterations and lower dimension ($T = 200, d = 50$). We discovered that by diminishing given constants L , D or L_* , all these algorithms require less iteration to reach $\epsilon = 1e - 2$ precision. So for the rest of experiments, we have chosen some hyper-tuned coefficients $L/50$, $D/10$ or $L_*/50$ (with associated hyper-tuned step-sizes) and we run these algorithms in higher dimension.

Power method oracle in higher dimension The idea is to repeat the same experience as for the stochastic smoothed case, as the power method oracle costs less than the stochastic smoothing oracle, we have enlarged the number of iterations from $T = 1000$ to $T = 5000$. This will affect our oblivious algorithm since we choose $\mu = \sqrt{T}$.

Dimension	Relative	Levy	Oblivious1	Oblivious2	Oblivious3
$d=50$	1591	342	881	644	641
$d=100$	> 5000	959	1172	713	629
$d=150$	> 5000	2207	2659	1127	767

TABLE 3. Number of iterations required to reach the precision 1e-2 with power method oracle for different algorithms.

On the previous table, we see that our oblivious algorithm is still efficient for problems in higher dimension. We also notice a major difference with the smooth case, in fact the degree of polynomial of oblivious algorithm plays a great role here, the oblivious algorithm using $\alpha_t \propto t^3$ require less iterations than the oblivious algorithm using $\alpha_t \propto t$. We think a good choice of the degree could be interesting in those situations.

Execution time and stability of the results From a practical point of view, it is interesting to collect the computation time for these algorithms and see how the complexity evolves in higher dimension. It would also be interesting to know if the experience are easily reproducible, so we have repeated the simulations 10 times and put the results into the following table with the empirical mean and confidence interval.

Dimension $d = 50$	Lan	Levy	Oblivious
Mean	> 6	0.68	0.14
Confidence interval	NA	[0.51, 1.03]	[0.12, 0.16]
Dimension $d = 100$	Lan	Levy	Oblivious
Mean	> 18	9.98	0.8
Confidence interval	NA	[7.73, 12.6]	[0.75, 0.88]
Dimension $d = 150$	Lan	Levy	Oblivious
Mean	> 33	31.2	5.58
Confidence interval	NA	[19.6, > 37]	[4.9, 6.25]

TABLE 4. Empirical measures about computation time (in seconds) required to reach the precision $1e-2$ with stochastic smoothing.

Dimension $d = 50$	Relative	Levy	Oblivious3
Mean	1.86	0.39	0.79
Confidence interval	[1.80, 1.92]	[0.28, 0.5]	[0.77, 0.81]
Dimension $d = 100$	Relative	Levy	Oblivious3
Mean	> 18	4.9	2.34
Confidence interval	NA	[3.63, 6.9]	[2.25, 2.4]
Dimension $d = 150$	Relative	Levy	Oblivious3
Mean	> 26	24.35	4.0
Confidence interval	NA	[20.5, 26.9]	[3.5, 4.74]

TABLE 5. Empirical measures about computation time (in seconds) required to reach the precision $1e-2$ with power method.

When the dimension increases, we notice that computational costs increase at non-linear speed. The reason is that the algorithm requires more computational time for each iteration and it requires more iterations in higher dimension to reach the precision $1e-2$. We also observe that our oblivious algorithm outperforms the other two algorithms in higher dimension.