2018

On acceleration with noise-corrupted gradients

On Acceleration with Noise-Corrupted Gradients

Michael B. Cohen [1] **Jelena Diakonikolas** [2] **Lorenzo Orecchia** [2]

Abstract

Accelerated algorithms have broad applications in large-scale optimization, due to their generality and fast convergence. However, their stability in the practical setting of noise-corrupted gradient oracles is not well-understood. This paper provides two main technical contributions: (i) a new accelerated method AGD+ that generalizes Nesterov's AGD and improves on the recent method AXGD (Diakonikolas & Orecchia, 2018), and (ii) a theoretical study of accelerated algorithms under noisy and inexact gradient oracles, which is supported by numerical experiments. This study leverages the simplicity of AGD+ and its analysis to clarify the interaction between noise and acceleration and to suggest modifications to the algorithm that reduce the mean and variance of the error incurred due to the gradient noise.

1. Introduction

First-order methods for convex optimization play a fundamental role in the solution of modern large-scale computational problems, encompassing applications in machine learning (Bubeck, 2014), scientific computing (Spielman & Teng, 2004; Kelner et al., 2013) and combinatorial optimization (Sherman, 2017; Ene & Nguyen, 2016). A central object of study in this area is the notion of *acceleration* – an algorithmic technique that can be deployed when minimizing a *smooth* convex function $f(\cdot)$ via queries to a first-order oracle (a *blackbox* that on input $\mathbf{x} \in \mathcal{X}$, returns the vector $\nabla f(\mathbf{x})$ in constant time). In this setting, a function $f(\cdot)$ is L-smooth if it is differentiable and its gradient is L-Lipschitz continuous w.r.t to a pair of dual norms $\|\cdot\|$, $\|\cdot\|_*$, i.e.:

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X},\ \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L \cdot \|\mathbf{x} - \mathbf{y}\|. \quad (1.1)$$

Acceleration is interesting because it yields faster algorithms than classical steepest-descent algorithms, often matching or closely approximating known information-theoretic lower bounds on the number of necessary queries to the oracle. In the simplest smooth setting, the optimal accelerated algorithm, Accelerated Gradient Descent (Nesterov, 1983), achieves an error that scales as $O(1/k^2)$, where k is the number of oracle queries. This should be compared to the convergence of steepest-descent methods, which attempt to locally minimize the first-order approximation to the function and only yield $O(1/k)$-convergence (Ben-Tal & Nemirovski, 2001; Nesterov, 2013). Many of the workhorses of optimization, such as conjugate gradient and FISTA (Beck & Teboulle, 2009), are instantiations of accelerated algorithms.

Because of its generality, acceleration still proves an active topic of research. In particular, two weaknesses in the classical presentation of accelerated methods have recently attracted attention of scholars and practitioners alike: 1) the complexity and lack of underlying intuition in the convergence analysis of accelerated methods, and 2) the apparent lack of robustness to perturbations of the gradient oracle displayed by accelerated methods when compared to their non-accelerated counterparts.

Recently, some of the mystery of acceleration has faded, as different works have provided natural interpretations and alternative proofs for accelerated methods (Allen-Zhu & Orecchia, 2017; Krichene et al., 2015; Wibisono et al., 2016; Bubeck et al., 2015; Lessard et al., 2016; Hu & Lessard, 2017; Diakonikolas & Orecchia, 2017). Of particular interest to our work is the framework of (Diakonikolas & Orecchia, 2017), which completely derives accelerated algorithms from the Euler discretization of a continuous dynamics that minimizes a natural notion of duality gap.

In terms of robustness, it has long been observed empirically that a naïve application of accelerated algorithms to inexact oracles often leads to error accumulation, even in the setting of random perturbations, while standard steepest descent algorithms do not suffer from this problem (Hardt, 2014). From a theoretical point of view, a number of papers have introduced oracle models that account for inexact gradient information. For example, (d'Aspremont, 2008) proposed a restricted model of perturbations to the gradient that preserves the possibility of acceleration. More

[*]Equal contribution [1]Department of EECS, Massachusetts Institute of Technology, Cambridge, MA, USA [2]Department of Computer Science, Boston University, Boston, MA, USA. Correspondence to: Jelena Diakonikolas <jelenad@bu.edu>, Lorenzo Orecchia <orecchia@bu.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

recently, (Devolder et al., 2014) proposed a more general framework that allows for larger perturbations and seems to capture the error accumulation and instability observed in practice for accelerated methods. In these works, the inexact oracle outputs an arbitrary deterministic perturbation of the true gradient oracle. In particular, (Devolder et al., 2014) shows that such perturbations can be adversarially chosen to encode non-smooth problems.

For stochastic perturbations, (Lan, 2012; Ghadimi & Lan, 2012; 2013) considered an additive-noise model, under which (Lan, 2012; Ghadimi & Lan, 2012) obtained an optimal convergence bound for the accelerated algorithm AC-SA in the smooth, non-strongly convex setting, but sub-optimal for the smooth, strongly-convex case.[1] Further, (Ghadimi & Lan, 2013) improved the convergence bound in the setting of *constrained* smooth and strongly convex minimization to the optimal one, by coupling AC-SA algorithm from (Ghadimi & Lan, 2012) with a domain-shrinking procedure. More recently, (Jain et al., 2018) completely closed this gap for the case of linear regression. Additionally, (Dvurechensky & Gasnikov, 2016) unified the deterministic model (Devolder et al., 2014), the stochastic model (Ghadimi & Lan, 2012), and the associated results. These references are the most closely related to our work.

Our contributions We study the issue of robustness of accelerated methods in three steps. First, we propose a novel, simple, generic accelerated algorithm AGD+ following the framework of (Diakonikolas & Orecchia, 2017). This algorithm has a simple interpretation and analysis, and generalizes other known accelerated algorithms.

Second, we leverage the simplicity of the analysis of AGD+ to characterize its behavior on different models of inexact oracles. Our analysis recovers the results for the deterministic oracle models of (d'Aspremont, 2008) and (Devolder et al., 2014). More importantly, we consider the more general model of noise-corrupted gradient oracle, in which the true gradient $\nabla f(\mathbf{x})$ is corrupted by additive noise $\boldsymbol{\eta}$:

$$\widetilde{\nabla} f(\mathbf{x}) = \nabla f(\mathbf{x}) + \boldsymbol{\eta}, \tag{1.2}$$

where the perturbation $\boldsymbol{\eta}$ may be a random variable. Such a model captures the setting of stochastic methods, in which the gradient is only estimated from a subset of its components (Lan, 2012; Ghadimi & Lan, 2012; 2013; Atchade et al., 2014; Krichene & Bartlett, 2017; Jain et al., 2018), the setting of differentially private empirical risk minimization, in which Gaussian noise is intentionally added to the gradient to protect the privacy of the data (Bassily et al., 2014), and the setting of engineering systems in which the gradient is estimated from noisy measurements (Birand et al., 2013).

Our algorithm AGD+ is closely related to AC-SA from (Lan, 2012) and can in fact be seen as a "lazy" (dual averaging) counterpart of AC-SA. After this paper had been submitted, Gasnikov and Nesterov independently proposed a universal method for stochastic composite optimization (Gasnikov & Nesterov, 2018). While their algorithm is defined recursively and does not explicitly account for the iterative construction of a dual solution, a simple unwinding of the recursion shows that it is identical to AGD+. However, the fact that AGD+ is obtained and analyzed through the use of the approximate duality gap technique (Diakonikolas & Orecchia, 2017) allows us to streamline the analysis and obtain various bounds for both deterministic and stochastic models of noise. Further, in the setting of smooth and strongly convex minimization, our analysis leads to a tighter convergence bound for a single-stage algorithm (without domain-shrinking) than previously obtained in (Ghadimi & Lan, 2012; 2013) (see Appendix B for a precise statement).

There are other models of noise that are not considered here. For example, we do not consider the model with both multiplicative and additive error in the gradient oracle (Hu et al., 2017). Further, stochastic methods with variance reduction (see, e.g., (Schmidt et al., 2017; Allen-Zhu, 2017) and references therein) lead to a particularly structured gradient noise variance (e.g., Lemma 3.4 in (Allen-Zhu, 2017)) that is not explored in this work. Nevertheless, we believe that our analysis is general enough to be extended to these settings as well, which is deferred to the future version of this paper.

Our results reveal an interesting discrepancy between noise tolerance in the settings of constrained and unconstrained smooth minimization. Namely, in the setting of constrained optimization, the error due to noise does not accumulate and is proportional to the diameter of the feasible region and the expected norm of the noise. In the setting of unconstrained optimization, the bound on the error incurred due to the noise accumulates, as observed empirically by (Hardt, 2014). However, our analysis also suggests a simple restart and slow down semi-heuristic for stabilizing the noise-incurred error, which allows taking advantage of both the acceleration and the noise stability under stochastic noise.

In the case of smooth and strongly convex minimization (Appendix B), the error due to noise does not accumulate even if the region is unconstrained, as long as the noise is zero-mean, independent, and has bounded variance.[2] Further, using smaller step sizes than in the standard accelerated version of the method, the error due to noise decreases at rate $1/k$ (compare this to the $1/\sqrt{k}$ rate for smooth non-strongly

[1] In particular, the deterministic term in the convergence bound in (Ghadimi & Lan, 2012) decreases as $O(1/k^2)$ instead of the optimal $O(1 - 1/\sqrt{\kappa})^k$ convergence, where κ is the objective function's condition number.

[2] Obtaining similar bounds for a slightly more general model that relaxes independence (similar to (Lan, 2012; Ghadimi & Lan, 2012)) is also possible; see Appendix C.2.

convex functions). This means that strong convexity of a function implies higher robustness to noise.

Finally, we verify the predictions and insights from our analysis of AGD+ by performing numerical experiments comparing AGD+ to other accelerated and non-accelerated methods on noise-corrupted gradient oracles. A noteworthy outcome of these experiments is the following: when a natural generic restart & slow-down semi-heuristic is applied, the accelerated algorithm AXGD (Diakonikolas & Orecchia, 2017) and the algorithm AGD+ presented in this paper seem to outperform Nesterov's AGD both in expectation and in variance in the presence of large noise. Further, we note that compared to AXGD, AGD+ reduces the oracle complexity (the number of queried gradients) by a factor of two.

2. Notation and Preliminaries

We assume that we are given a continuously differentiable convex function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set. Hence:

$$\forall \mathbf{y}, \mathbf{x} \in \mathcal{X} : \; f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle , \quad (2.1)$$

where $\nabla f(\cdot)$ denotes the gradient of $f(\cdot)$.

Given oracle access to (possibly noise-corrupted) gradients of $f(\cdot)$, we are interested in minimizing $f(\cdot)$. We denote by $\mathbf{x}_* \in \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ any (fixed) minimizer of $f(\cdot)$.

We assume that there is an arbitrary (but fixed) norm $\|\cdot\|$ associated with the space, and all the statements about function properties are stated with respect to that norm. We also define the dual norm $\|\cdot\|_*$ in the standard way: $\|\mathbf{z}\|_* = \sup\{\langle \mathbf{z}, \mathbf{x}\rangle : \|\mathbf{x}\| = 1\}$. The following definitions will be useful in our analysis.

Definition 2.1. A function $f : \mathcal{X} \to \mathbb{R}$ is L-smooth on $\mathcal{X}$ with respect to a norm $\|\cdot\|$, if for all $\mathbf{x}, \hat{\mathbf{x}} \in \mathcal{X}$: $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{L}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$. This is equivalent to Equation (1.1).

A gradient step is defined in a standard way as $\mathrm{Grad}(\mathbf{x}) = \arg\min_{\hat{\mathbf{x}}\in\mathcal{X}}\{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{L}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2\}$.

Definition 2.2. A function $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex on $\mathcal{X}$ with respect to a norm $\|\cdot\|$, if for all $\mathbf{x}, \hat{\mathbf{x}} \in \mathcal{X}$: $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{\mu}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$.

Definition 2.3. (Convex Conjugate) Function ψ^* is the convex conjugate of $\psi : \mathcal{X} \to \mathbb{R}$, if $\psi^*(\mathbf{z}) = \max_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$, $\forall \mathbf{z} \in \mathbb{R}$.

We assume that there is a strongly-convex differentiable function $\psi : \mathcal{X} \to \mathbb{R}$ such that $\max_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$ is easily solvable, possibly in a closed form. Notice that this problem defines the convex conjugate of $\psi(\cdot)$, i.e., $\psi^*(\mathbf{z}) = \max_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$. The following fact is a simple corollary of Danskin's Theorem[3].

Fact 2.4. *Let $\psi : \mathcal{X} \to \mathbb{R}$ be differentiable and strongly-convex. Then: $\nabla\psi^*(\mathbf{z}) = \arg\max_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$.*

Fact 2.5. *If $\psi(\cdot)$ is μ-strongly convex w.r.t. a norm $\|\cdot\|$ for $\mu > 0$, then $\psi^*(\cdot)$ is $\frac{1}{\mu}$-smooth w.r.t. the norm $\|\cdot\|_*$.*

Definition 2.6. (Bregman Divergence) $D_\psi(\mathbf{x}, \hat{\mathbf{x}}) \stackrel{\text{def}}{=} \psi(\mathbf{x}) - \psi(\hat{\mathbf{x}}) - \langle \nabla\psi(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}}\rangle$, for $\mathbf{x} \in \mathcal{X}, \hat{\mathbf{x}} \in \mathcal{X}^o$, where $\mathcal{X}^o$ denotes the set of all points from $\mathcal{X}$ for which $\psi(\cdot)$ admits a (sub)gradient.

The Bregman divergence $D_\psi(\mathbf{x}, \mathbf{y})$ captures the difference between $\psi(\mathbf{x})$ and its first order approximation at $\mathbf{y}$. Notice that, for a differentiable ψ, we have: $\nabla_{\mathbf{x}} D_\psi(\mathbf{x}, \mathbf{y}) = \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{y})$. The Bregman divergence $D_\psi(\mathbf{x}, \mathbf{y})$ as a function $g_{\mathbf{y}}(\mathbf{x})$ is convex. Its Bregman divergence is itself, i.e., $D_{g_{\mathbf{y}}}(\mathbf{v}, \mathbf{u}) = D_\psi(\mathbf{v}, \mathbf{u})$.

3. Improved Accelerated Method

In this section, we focus on the setting of smooth minimization. The case of smooth and strongly convex minimization is treated in Appendix B. To design AGD+, we define an approximate duality gap, similar to (Diakonikolas & Orecchia, 2017; 2018), but allowing for an inexact gradient oracle according to Eq. (1.2). The construction is based on maintaining three points at each iteration k: $\mathbf{x}_k$ is the point at which the gradient is queried, while $(\mathbf{y}_k, \mathbf{z}_k)$ is the current primal-dual solution pair at the end of iteration k. For this setup, the dual solution $\mathbf{z}_k$ is a conic combination of the negative gradients seen so far, taken at an initial dual point $z_0 = \nabla\psi(\mathbf{x}_0)$,where $\mathbf{x}_0$ is an arbitrary initial primal solution, i.e.,

$$\mathbf{z}_k = -\sum_{i=1}^{k} a_i \widetilde{\nabla} f(\mathbf{x}_i) + \mathbf{z}_0. \quad (3.1)$$

where the sequence $a_k > 0$, $A_k = \sum_{i=1}^k a_i$ will be specified later. By convention, $A_0 = 0$.

3.1. Approximate Duality Gap

The choice of sequences above immediately implies upper and lower bounds on the optimum at each iteration k. The upper bound is simply chosen as $U_k = f(\mathbf{y}_k)$. For the lower bound, by convexity of $f(\cdot)$ (see Eq. (2.1)):

$$f(\mathbf{x}_*) \geq \frac{\sum_{i=1}^k a_i f(\mathbf{x}_i) + \sum_{i=1}^k a_i \langle \nabla f(\mathbf{x}_i), \mathbf{x}_* - \mathbf{x}_i\rangle}{A_k}.$$

To relate the lower bound to the output of the inexact oracle, it is useful to express the gradients $\nabla f(\mathbf{x}_i)$ as $\nabla f(\mathbf{x}_i) = \widetilde{\nabla} f(\mathbf{x}_i) - \boldsymbol{\eta}_i$. Adding and subtracting $\frac{1}{A_k} D_\psi(\mathbf{x}_*, \mathbf{x}_0)$ in the

[3] See, e.g., Proposition 4.15 in (Bertsekas et al., 2003).

last equation, we have:

$$f(\mathbf{x}_*) \geq \frac{\sum_{i=1}^k a_i f(\mathbf{x}_i) + \sum_{i=1}^k a_i \left\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{x}_* - \mathbf{x}_i \right\rangle}{A_k} + \frac{-\sum_{i=1}^k a_i \langle \boldsymbol{\eta}_i, \mathbf{x}_* - \mathbf{x}_i \rangle + D_\psi(\mathbf{x}_*, \mathbf{x}_0) - D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k}.$$

Finally, we can replace $\mathbf{x}_*$ by a minimization over $\mathcal{X}$ to obtain our final lower bound:

$$\frac{\sum_{i=1}^k a_i f(\mathbf{x}_i) - \sum_{i=1}^k a_i \langle \boldsymbol{\eta}_i, \mathbf{x}_* - \mathbf{x}_i \rangle - D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + \frac{\min_{\mathbf{u} \in \mathcal{X}} \left\{ \sum_{i=1}^k a_i \left\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i \right\rangle + D_\psi(\mathbf{u}, \mathbf{x}_0) \right\}}{A_k} \stackrel{\text{def}}{=} L_k.$$

Applying Fact 2.4 and the definition of $\mathbf{z}_k$ from (3.1), we have the following characterization of the last term of L_k.

Proposition 3.1. *Let $\mathbf{z}_k$ be defined as in (AGD+). Then:*

$$\nabla \psi^*(\mathbf{z}_k) = \arg\min_{\mathbf{u} \in \mathcal{X}} \left\{ \sum_{i=1}^k a_i \left\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i \right\rangle + D_\psi(\mathbf{u}, \mathbf{x}_0) \right\}.$$

The approximate duality gap is simply defined as $G_k = U_k - L_k$. Observe that, by construction of U_k and L_k, $f(\mathbf{y}_k) - f(\mathbf{x}_*) \leq G_k$. Hence, to prove the convergence of the algorithm, it suffices to bound G_k. To do so, we will track the evolution of the quantity $A_k G_k$, i.e., we will bound[4] $E_k = A_k G_k - A_{k-1} G_{k-1}$, so that

$$G_k = \frac{A_1}{A_k} G_1 + \frac{\sum_{i=2}^k E_i}{A_k}.$$

3.2. The AGD+ Algorithm

The steps of AGD+ are defined as follows:

$$\begin{aligned} \mathbf{x}_k &= \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \nabla \psi^*(\mathbf{z}_{k-1}), \\ \mathbf{z}_k &= \mathbf{z}_{k-1} - a_k \widetilde{\nabla} f(\mathbf{x}_k), \qquad \text{(AGD+)} \\ \mathbf{y}_k &= \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \nabla \psi^*(\mathbf{z}_k), \end{aligned}$$

To seed AGD+, we let $\mathbf{x}_1 = \mathbf{x}_0$, $\mathbf{y}_1 = \mathbf{v}_1 = \nabla \psi^*(\mathbf{z}_1)$.

Related Algorithms Compared to Nesterov's AGD, AGD+ differs in the sequence $\mathbf{y}_k$: AGD sets $\mathbf{y}_k = \mathrm{Grad}(\mathbf{x}_k)$. The two algorithms are equivalent when $\psi(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_0\|^2$, $\frac{a_k^2}{A_k} = \frac{\mu}{L}$, and $\mathcal{X} = \mathbb{R}^n$, but in general they produce different sequences of points. Thus, AGD+ can be seen as a generalization of AGD. Compared to a more recent accelerated method AXGD of (Diakonikolas & Orecchia, 2018), AGD+ differs in sequences $\mathbf{y}_k$ and $\mathbf{z}_k$. In particular, in AXGD, $\mathbf{z}_k = \mathbf{z}_{k-1} - a_k \widetilde{\nabla} f(\mathbf{y}_k)$, while $\mathbf{y}_k = \frac{A_{k-1}}{A_k} \mathbf{y}_{k-1} + \frac{a_k}{A_k} \nabla \psi^*(\mathbf{z}_k - a_k \widetilde{\nabla} f(\mathbf{x}_k))$. As AXGD uses the gradients of $f(\cdot)$ at both sequences $\mathbf{x}_k$ and $\mathbf{y}_k$ to define $\mathbf{x}_k$ and $\mathbf{y}_k$, it is more wasteful: its oracle complexity is twice as high as that of AGD and AGD+. Most closely related to AGD+ is the AC-SA algorithm (Lan, 2012); namely, for some step sizes, AGD+ can be seen as a "lazy" (dual averaging) version of AC-SA. The relationship between μAGD+ (see Appendix B) and AC-SA for smooth and strongly convex minimization (Ghadimi & Lan, 2012) is not immediately clear, due to the different parameter choices.

3.3. Convergence Analysis for AGD+

To simplify the notation, from now on we denote:

$$\mathbf{v}_k \stackrel{\text{def}}{=} \nabla \psi^*(\mathbf{z}_k).$$

We can now bound the change $E_k = A_k G_k - A_{k-1} G_{k-1}$ by decomposing it into two terms: $E_k \leq E_k^e + E_k^\eta$, where the latter term is due to the inexact nature of the gradient oracle. The following lemma allows us to bound these terms. Its proof can be found in Appendix A.

Lemma 3.2. *Let $E_k^\eta = \langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k \rangle$ and $E_k^e = A_k(f(\mathbf{y}_k) - f(\mathbf{x}_k)) - A_k \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \rangle - D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1})$. Then $E_k \leq E_k^\eta + E_k^e$.*

The last piece that is needed for the analysis is the bound on the initial gap G_1, obtained in the following proposition.

Proposition 3.3. *$A_1 G_1 \leq D_\psi(\mathbf{x}_*, \mathbf{x}_0) + E_1^\eta + E_1^e$, where E_1^η is defined as in Lemma 3.2 and $E_1^e = A_1(f(\mathbf{y}_1) - f(\mathbf{x}_1) - \langle \nabla f(\mathbf{x}_1), \mathbf{v}_1 - \mathbf{x}_1 \rangle) - D_\psi(\mathbf{v}_1, \mathbf{x}_0)$.*

The proof is a straightforward application of the previously introduced definitions.

3.4. Convergence of AGD+ with Exact Oracle

To prove the convergence of the method in the noiseless case, in this section we assume that $\boldsymbol{\eta}_k = \mathbf{0}$, and, consequently, $E_k^\eta = 0$. Hence, to obtain a convergence bound for AGD+, we only need to bound E_k^e.

Theorem 3.4. *Let $f : \mathcal{X} \to \mathbb{R}$ be an L-smooth function and let $\mathbf{x}_0 \in \mathcal{X}$ be an arbitrary initial point. If sequences $\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k$ evolve according to (AGD+) for some μ-strongly convex function $\psi(\cdot)$, $\boldsymbol{\eta}_k = \mathbf{0}$, and $\frac{a_k^2}{A_k} \leq \frac{\mu}{L}$, then $\forall k \geq 1$:*

$$f(\mathbf{y}_k) - f(\mathbf{x}_*) \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k}.$$

[4] From (Diakonikolas & Orecchia, 2017) it can be derived that $A_k G_k$ is a Lyapunov function for the continuous dynamic underlying AGD+, i.e., E_k is the discretization error at iteration k.

Proof. By smoothness of $f(\cdot)$, $f(\mathbf{y}_k) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \rangle \leq \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_k\|^2$. Hence:

$$E_k = E_k^e \leq A_k \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_k\|^2 - D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}).$$

From (AGD+), $\mathbf{y}_k - \mathbf{x}_k = \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{v}_{k-1})$. As $D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) \geq \frac{\mu}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2$, it follows that:

$$E_k \leq \frac{1}{2}\Big(\frac{{a_k}^2 L}{A_k} - \mu\Big)\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 \leq 0,$$

as $\frac{{a_k}^2}{A_k} \leq \frac{\mu}{L}$ by the theorem assumptions. Thus: $G_k \leq \frac{A_1}{A_k} G_1$ and it remains to bound $A_1 G_1$, which is just:

$$\begin{aligned} A_1 G_1 =& A_1(f(\mathbf{y}_1) - f(\mathbf{x}_1) - \langle \nabla f(\mathbf{x}_1), \mathbf{v}_1 - \mathbf{x}_1 \rangle) \\ & - D_\psi(\mathbf{v}_1, \mathbf{x}_0) + D_\psi(\mathbf{x}_*, \mathbf{x}_0) \leq D_\psi(\mathbf{x}_*, \mathbf{x}_0), \end{aligned}$$

as $\mathbf{y}_1 = \mathbf{v}_1$ and $\mathbf{x}_1 = \mathbf{x}_0$. □

Observe that for $a_k = \frac{\mu}{L} \cdot \frac{k+1}{2}$ we recover the standard $1/k^2$ convergence rate of accelerated methods.

3.5. Convergence of AGD+ with Inexact Oracle

In this subsection, we focus on bounding the error E_k^η that is accrued due to the additive noise $\boldsymbol{\eta}_k$. Additional results (including other models of noise) can be found in Appendix C. As $E_k^\eta = a_k \langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k \rangle$ by Lemma 3.2, we have:

Proposition 3.5. *Let $f : \mathcal{X} \to \mathbb{R}$ be an L-smooth function and let $\mathbf{x}_0 \in \mathcal{X}$ be an arbitrary initial point. If sequences $\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k$ evolve according to (AGD+) for some μ-strongly convex function $\psi(\cdot)$, where $\boldsymbol{\eta}_k$'s are independent, $R_{\mathbf{x}_*} = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}_*\|$, and $\frac{{a_k}^2}{A_k} \leq \frac{\mu}{L}$, then $\forall k \geq 1$:*

$$\mathbb{E}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right] \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + R_{\mathbf{x}_*} \frac{\sum_{i=1}^k a_i \mathbb{E}[\|\boldsymbol{\eta}_i\|_*]}{A_k},$$

$$\mathrm{Var}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right] \leq {R_{\mathbf{x}_*}}^2 \frac{\sum_{i=1}^k {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}{{A_k}^2}.$$

Proof. From Theorem 3.4 and Lemma 3.2:

$$A_k G_k \leq D_\psi(\mathbf{x}_*, \mathbf{x}_0) + \sum_{i=1}^k a_i \langle \boldsymbol{\eta}_i, \mathbf{x}_* - \mathbf{v}_i \rangle. \tag{3.2}$$

The bound on the expectation follows by applying $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|\|\mathbf{y}\|_*$ (by the duality of norms), linearity of expectation, and $A_k = \sum_{i=1}^k a_i$. The bound on the variance follows by, in addition, using the standard facts that $\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y]$ and $\mathrm{Var}[X] \leq \mathbb{E}[X^2]$, where a, b are constants and X, Y are independent random variables. □

Remark 3.6. Observe that if $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*] \leq M$, $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] \leq \sigma^2$, $\forall i$, then $\mathbb{E}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right] \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + R_{\mathbf{x}_*} M$ and $\mathrm{Var}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right] \leq (R_{\mathbf{x}_*}\sigma)^2 \frac{\sum_{i=1}^k {a_i}^2}{{A_k}^2}$. The same bound on $\mathbb{E}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right]$ as in Prop. 3.5 (and the special case stated here) holds *even if $\boldsymbol{\eta}_k$'s are not independent*.

The bound from Proposition 3.5 is mainly useful when $R_{\mathbf{x}_*}$ is bounded, which is the case when, e.g., the diameter of $\mathcal{X}$ is bounded. For the case of unconstrained optimization (i.e., when $\mathcal{X} = \mathbb{R}^n$), the bound from Proposition 3.5 is uninformative. Hence, we derive another bound that is independent of $R_{\mathbf{x}_*}$, but it requires that the noise samples $\boldsymbol{\eta}_k$ are both zero-mean and independent.[5]

Lemma 3.7. *Let $f : \mathcal{X} \to \mathbb{R}$ be an L-smooth function and let $\mathbf{x}_0 \in \mathcal{X}$ be an arbitrary initial point. If sequences $\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k$ evolve according to (AGD+) for some μ-strongly convex function $\psi(\cdot)$, where $\boldsymbol{\eta}_k$'s are zero-mean independent random variables and $\frac{{a_k}^2}{A_k} \leq \frac{\mu}{L}$, then $\forall k \geq 1$:*

$$\mathbb{E}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right] \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + \frac{\sum_{i=1}^k {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}{\mu A_k}.$$

Proof. Let $\hat{\mathbf{v}}_k = \nabla\psi^*(\mathbf{z}_{k-1} - a_k \nabla f(\mathbf{x}_k)) = \nabla\psi^*(\mathbf{z}_k + a_k\boldsymbol{\eta}_k)$. Recall that $E_k^\eta = a_k \langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k \rangle$. Adding and subtracting $\hat{\mathbf{v}}_k$:

$$E_k^\eta = a_k \langle \boldsymbol{\eta}_k, \mathbf{x}_* - \hat{\mathbf{v}}_k \rangle + a_k \langle \boldsymbol{\eta}_k, \hat{\mathbf{v}}_k - \mathbf{v}_k \rangle.$$

As $\hat{\mathbf{v}}_k$ is independent of $\boldsymbol{\eta}_k$ and $\mathbb{E}[\boldsymbol{\eta}_k] = \mathbf{0}$, $\mathbb{E}\left[\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \hat{\mathbf{v}}_k \rangle\right] = 0$. On the other hand, as $\mathbf{v}_k = \nabla\psi^*(\mathbf{z}_k)$ and $\psi^*(\cdot)$ is $\frac{1}{\mu}$-smooth by Fact 2.5, by the duality of norms: $\langle \boldsymbol{\eta}_k, \hat{\mathbf{v}}_k - \mathbf{v}_k \rangle \leq \frac{a_k}{\mu}\|\boldsymbol{\eta}_k\|_*^2$. The rest of the proof is by Theorem 3.4 and Lemma 3.2. □

Remark 3.8. It is possible to relax the assumption that $\boldsymbol{\eta}_k$'s are independent. In fact, for Lemma 3.7 to apply, it suffices that, conditioned on the natural filtration $\mathcal{F}_{k-1}$ (all the information about the noise up to the beginning of iteration k), $\boldsymbol{\eta}_k$ is independent of $\hat{\mathbf{v}}_k$. More details are provided in Appendix C.2.

Lemma 3.7 suggests that for unconstrained smooth minimization the sequence a_k that leads to accelerated methods aggregates noise, as for accelerated methods $a_k \sim k$, $A_k \sim k^2$. However, if we were to resort to a slower, uniform sequence (and slower $1/k$ convergence rate), then the noise would average out, as we would have constant a_k's and $A_k \sim k$. Even more, if $a_k \sim 1/\sqrt{k}$, then the error due to noise would decrease at rate $\log(k)/\sqrt{k}$. This is confirmed by our numerical experiments and matches the experience of practitioners, as discussed by (Hardt, 2014).

The lemma also shows that error accumulation can be avoided if we postulate that the magnitude of the noise vanishes with the number of iterations. This can be achieved

[5] The assumption that $\boldsymbol{\eta}_k$'s are independent can be relaxed – see Remark 3.8 below and Appendix C.2.

if the estimates of the gradient improve over iterations (Atchade et al., 2014). For example, if we have $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] = O\left(\frac{1}{a_i}\right) = O\left(\frac{1}{i}\right)$, the noise-error term averages out, making the total error due to noise bounded. If $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] = O\left(\frac{1}{i^2}\right)$, the noise-error term vanishes at rate $k/A_k = 1/k$. Finally, if $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] = O\left(\frac{1}{i^3}\right)$, the noise-error term vanishes at rate $\log(k)/k^2$, essentially recovering accelerated convergence.

Observe that we could not get a bound on variance that is independent of $R_{\mathbf{x}_*}$, as the variance (unlike the expectation) of $\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \hat{\mathbf{v}}_k \rangle$ is not zero. Instead, since we upper-bound the expectation of $f(\mathbf{y}_k) - f(\mathbf{x}_*)$ by a non-negative quantity (and $f(\mathbf{y}_k) - f(\mathbf{x}_*)$ is always non-negative as $\mathbf{x}_*$ is the minimizer of $f(\cdot)$), we can apply Markov's Inequality to obtain a concentration bound on $f(\mathbf{y}_k) - f(\mathbf{x}_*)$. Finally, the step sizes a_k can be chosen so as to balance the deterministic error and the error due to noise in the convergence bound. This leads to the following corollary.

Corollary 3.9. *Let $f : \mathcal{X} \to \mathbb{R}$ be an L-smooth function and let $\mathbf{x}_0 \in \mathcal{X}$ be an arbitrary initial point. If sequences $\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k$ evolve according to (AGD+) for some μ-strongly convex function $\psi(\cdot)$, where $\boldsymbol{\eta}_k$'s are zero-mean independent random variables, $K \geq 1$ is an arbitrary (but fixed) number of iterations of AGD+, $\gamma = \mu / \max\{L, \sqrt{\sum_{i=1}^K {b_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}\}$, $a_i = \gamma b_i$, and $\frac{{a_i}^2}{A_i} \leq \gamma$, $\forall i$, then:*

$$\mathbb{E}\left[f(\mathbf{y}_K) - f(\mathbf{x}_*)\right] \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_K} + \frac{\sqrt{\sum_{i=1}^K {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}}{A_K}.$$

In particular, if $b_i = \frac{i+1}{2}$ and, in addition, $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] \leq \sigma^2$, $\forall i$, then:

$$\mathbb{E}\left[f(\mathbf{y}_K) - f(\mathbf{x}_*)\right] \leq \frac{4LD_\psi(\mathbf{x}_*, \mathbf{x}_0)}{\mu K(K+3)} + O\left(\frac{\sigma(\mu + D_\psi(\mathbf{x}_*, \mathbf{x}_0))}{\mu\sqrt{K}}\right).$$

If $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] \leq \sigma^2$, $\forall i$, but the value of σ is unknown, then setting $b_i = \frac{i+1}{2}$ and $\gamma = \mu / \max\{L, \sqrt{\sum_{i=1}^K {b_i}^2}\}$ gives:

$$\mathbb{E}\left[f(\mathbf{y}_K) - f(\mathbf{x}_*)\right] \leq O\left(\frac{\sigma^2}{\sqrt{K}}\right) + \max\left\{\frac{4LD_\psi(\mathbf{x}_*, \mathbf{x}_0)}{\mu K(K+3)}, O\left(\frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{\mu\sqrt{K}}\right)\right\}.$$

Proof. By the choice of parameters,

$$\mu = \gamma \max\left\{L, \sqrt{\sum_{i=1}^K {b_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}\right\} \geq \gamma \sqrt{\sum_{i=1}^K {b_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]} = \sqrt{\sum_{i=1}^K {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]},$$

which bounds the stochastic term. The rest of the proof follows by plugging in particular choices of parameters and using that $\gamma L \leq \mu \leq \gamma L + \sqrt{\sum_{i=1}^K {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}$. □

Remark 3.10. Observe that the optimal choice of γ to balance the terms from Lemma 3.7 would be $\gamma = \mu / \max\left\{L, \sqrt{\frac{\sum_{i=1}^K {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}{D_\phi(\mathbf{x}_*, \mathbf{x}_0)}}\right\}$. When $\mathcal{X}$ is unconstrained, it is not always possible to estimate (an upper bound on) $D_\phi(\mathbf{x}_*, \mathbf{x}_0)$, which is why we made the particular choice of μ in Corollary 3.9. However, when the diameter $\Omega_\mathcal{X}$ of $\mathcal{X}$ is bounded, we can choose $\gamma = \mu / \max\left\{L, \sqrt{\frac{\sum_{i=1}^K {a_i}^2 \mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2]}{\Omega_\mathcal{X}}}\right\}$, leading to the same (optimal) asymptotic bound as in (Lan, 2012). Finally, for bounded-diameter region, (Ghadimi & Lan, 2012) provides a step-size policy that can relax the assumption that K is fixed in advance. We expect that a similar policy should also apply in the case of AGD+. The details are omitted.

4. Noise-Error Reduction

We now discuss how the results of Section 3 can be used to control the error of AGD+ that is incurred due to the gradient oracle noise. First, we discuss how to prevent error accumulation from Lemma 3.7, which is incurred when running a vanilla version of AGD+. The main idea is to take advantage of acceleration until the noise accumulation starts dominating the convergence, and then switch to a slower sequence $\{a_k\}$ for which the error averages out and the algorithm further reduces the mean. Finally, we show how, through another algorithm restart and slow down, the sequence of updates can be made convergent (i.e., the mean error is further reduced at a rate $\sim 1/\sqrt{k}$).

Observe that the result from Corollary 3.9 already gives a convergent sequence of updates. However, the choice of parameters in Corollary 3.9 is fixed and tailored to the global problem properties and worst-case effect of the additive noise. Instead, the strategy of incrementally slowing down the algorithm can take advantage of the more local, fine-grained properties of the objective function, as confirmed by the experiments in Section 5 and in Appendix D.

4.1. Mean-Error Stabilization

To take advantage of acceleration at the initial stage and then stabilize the mean error due to noise, we propose the following RESTART+SLOWDOWN semi-heuristics:

> RESTART+SLOWDOWN: If $\|\mathbf{z}_k\|_2^2 \leq \sum_{i=1}^k {a_i}^2 \mathbb{E}\left[\|\eta_i\|_2^2\right]$, restart the algorithm taking $\mathbf{y}_k$ as the initial point and slow down the sequence $\{a_i\}$ to $a_i = \frac{\mu}{L}$, $\forall i \geq 1$.

The only "heuristic" part of RESTART+SLOWDOWN is deciding when to switch to the slower sequence, as, due to Lemma 3.7, slower sequence is guaranteed to lead to a better bound on the approximation error due to noise in the case of unconstrained minimization. Further, switching to a slower, linearly growing sequence A_k is guaranteed to further reduce the error mean, as discussed in the next subsection.

The intuition behind RESTART+SLOWDOWN criterion is restarting when "the signal is drowning in noise". In particular, $\mathbf{z}_k$ (the weighted sum of the noisy negative gradients) is the only gradient information used in defining all steps of AGD+ and we can interpret it as the "signal" that is used to guide the algorithm updates. When the gradients are corrupted by noise, $\mathbf{z}_k = -\sum_{i=1}^k a_i \nabla f(\mathbf{x}_i) - \sum_{i=1}^k a_i \boldsymbol{\eta}_i$. As the noise is assumed to be independent, the expected energy of the signal-plus-noise is equal to the sum of the energy of the signal and the expected energy of the noise:

$$\|\mathbf{z}_k\|_2^2 = \mathbb{E}\left[\left\|\sum_{i=1}^k a_i \nabla f(\mathbf{x}_i)\right\|_2^2\right] + \mathbb{E}\left[\left\|\sum_{i=1}^k a_i \boldsymbol{\eta}_i\right\|_2^2\right]$$
$$= \left\|\sum_{i=1}^k a_i \nabla f(\mathbf{x}_i)\right\|_2^2 + \sum_{i=1}^k {a_i}^2 \mathbb{E}\left[\|\eta_i\|_2^2\right].$$

Hence, when the criterion of RESTART+SLOWDOWN is satisfied, the energy component due to noise dominates the energy component of the signal in $\mathbf{z}_k$.

For constrained minimization with a small diameter, RESTART+SLOWDOWN cannot reduce the theoretical mean of the error due to noise (unless the bound from Lemma 3.7 dominates the bound from Proposition 3.5), as the noise term averages out regardless of the sequence $\{a_i\}$ (Proposition 3.5). Nevertheless, a slower, uniform sequence $\{a_i\}$ has lower variance than the accelerated sequence, and can be beneficial in the settings where the accelerated sequence produces high error variance.

4.2. Further Mean-Error Reduction

Quadratically-growing sequence $\{A_i\}$ (or linearly growing sequence $\{a_i\}$) is the fastest-growing sequence which guarantees that $A_k G_k$ is non-increasing in the case of smooth minimization with exact gradients. When we switch to a slower sequence $\{A_i\}$ by invoking RESTART+SLOWDOWN, this creates more slack in making $A_k G_k$ non-increasing in the presence of gradient noise. Hence, RESTART+SLOWDOWN reduces the mean error and keeps it bounded. However, with RESTART+SLOWDOWN alone, the mean error cannot converge to zero. To ensure that the error is converging to zero, we can perform an additional RESTART+SLOWDOWN (RESTART+SLOWDOWN-2), which uses the same criterion for restart, but slows down the sequence a_k to $a_k \sim 1/\sqrt{k}$, as follows.

RESTART+SLOWDOWN-2: If $\|\mathbf{z}_k\|_2^2 \leq \sum_{i=1}^k {a_i}^2 \mathbb{E}\left[\|\eta_i\|_2^2\right]$, restart the algorithm taking $\mathbf{y}_k$ as the initial point and slow down the sequence $\{a_i\}$ to $a_i = \mu/(L\sqrt{i}), \forall i \geq 1$.

Thus, we have the following Corollary (of Lemma 3.7):

Corollary 4.1. *Let $f : \mathcal{X} \to \mathbb{R}$ be an L-smooth function and let $\mathbf{x}_0 \in \mathcal{X}$ be an arbitrary initial point. If sequences $\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k$ evolve according to (AGD+) for some μ-strongly convex function $\psi(\cdot)$, where $\boldsymbol{\eta}_k$'s are zero-mean i.i.d. random variables and $a_k = \frac{\mu}{L\sqrt{k}}$, then $\forall k \geq 1$:*

$$\mathbb{E}\left[f(\mathbf{y}_k) - f(\mathbf{x}_*)\right] \leq \frac{L D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{\mu\sqrt{k}} + \frac{\log(k+1)\mathbb{E}[\|\boldsymbol{\eta}_1\|_*^2]}{L\sqrt{k}}.$$

Finally, observe that the factor of $\log(k+1)$ in the bound from Corollary 4.1 can be removed if the number of steps K is fixed in advance and a_k's are set to $a_k = \frac{\mu}{L\sqrt{K}}$.

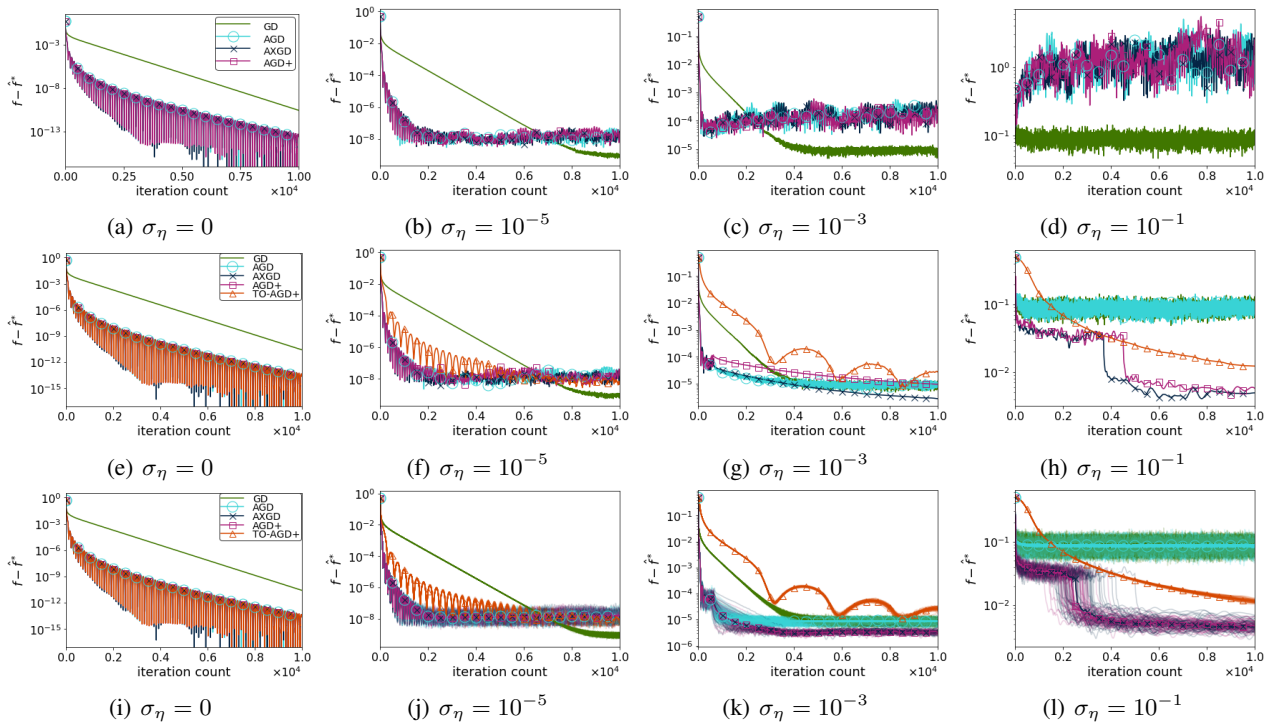
5. Numerical Experiments

To illustrate the results, we consider two main problems: a hard instance for smooth minimization (see, e.g., (Nesterov, 2013)) and regression problems on Epileptic Seizure Recognition Dataset (Andrzejak et al., 2001) obtained from the UCI Machine Learning Repository (Lichman, 2013). Most results can be found in Appendix D. In all the experiments, we used standard Python libraries to solve the considered problems to high accuracy. The resulting function value is denoted by $\hat{f}^*$ in the figures. In all the problems, we used $\psi(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|_2^2$ as the regularizer. For constrained problems, we implemented projected gradient descent as the "GD" algorithm.

In the graphs, TO-AGD+ denotes the "theoretically optimal" version of AGD+; namely, it corresponds to AGD+ with step sizes chosen according to Corollary 3.9 and Remark 3.10. In all the experiments, we compare the different accelerated algorithms (AGD+, AGD, AXGD) and the non-accelerated GD under i.i.d. additive gradient noise $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\eta I)$.

"Hard" Instance for Smooth Minimization To understand the worst-case performance of AGD+, we first compare it to Nesterov's AGD and (Diakonikolas & Orecchia, 2018)'s AXGD. The instance is an unconstrained minimization problem, where $f(\mathbf{x}) = \frac{1}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$, $\mathbf{A}$ is the graph Laplacian of a cycle[6], $b_1 = -b_n = 1$ and vector $\mathbf{b}$ is zero elsewhere. The initial point $\mathbf{x}_0$ is an all-zeros vector. The dimension of the problem is $n = 100$.

[6]Namely, the difference of a tridiagonal square matrix $\mathbf{C}$ with 1's on the main diagonal and -1's on the remaining diagonals, and matrix $\mathbf{B}$, which is zero everywhere except for $B_{1n} = B_{n1} = 1$.



(a) $\sigma_\eta = 0$ (b) $\sigma_\eta = 10^{-5}$ (c) $\sigma_\eta = 10^{-3}$ (d) $\sigma_\eta = 10^{-1}$

(e) $\sigma_\eta = 0$ (f) $\sigma_\eta = 10^{-5}$ (g) $\sigma_\eta = 10^{-3}$ (h) $\sigma_\eta = 10^{-1}$

(i) $\sigma_\eta = 0$ (j) $\sigma_\eta = 10^{-5}$ (k) $\sigma_\eta = 10^{-3}$ (l) $\sigma_\eta = 10^{-1}$

Figure 1: Performance of gradient descent (GD) and accelerated algorithms (AGD, AXGD, AGD+) on a hard instance for unconstrained smooth minimization for $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \sigma_\eta I)$ over $\mathbb{R}^n$: (a)-(d) sample run; (e)-(h) sample run with both RESTART+SLOWDOWN and RESTART+SLOWDOWN-2 implemented on all accelerated algorithms, except TO-AGD+; (i)-(l) 50 runs and the median with same algorithms as previous row.

The performance of AGD+, AGD, and AXGD together with the performance of the slower, unaccelerated GD on the described worst-case instance is shown in Fig. 1(a)-1(d), for the exact gradient oracle (Fig. 1(a)) and noise-corrupted gradient oracle with i.i.d. $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\eta I)$ (Fig. 1(b)-1(d)). We repeated the same experiments when the parameters for AGD+ are chosen according to Corollary 3.9 (denoted as TO-AGD+) and when both RESTART+SLOWDOWN and RESTART+SLOWDOWN-2 are employed on all other accelerated algorithms. (Fig. 1(e)-1(l)).

Without restart and slow-down, all accelerated algorithms perform similarly. In particular, as the noise standard deviation σ_η is increased, the mean and the variance of the approximation error of all accelerated algorithms increases and the noise appears to be accumulating (see, e.g., Fig. 1(d)). On the other hand, GD generally converges to an approximation error with lower mean and variance, at the expense of converging at a slower $1/k$ rate.

When restart and slow-down are used, in the noiseless case (Fig. 1(e) and 1(i)), there is no difference compared to the vanilla case (Fig. 1(a)), which is what we want – there is no need to slow down the accelerated algorithms unless their performance is compromised by noise. In the low-noise scenario (Fig. 1(f), 1(j)), RESTART+SLOWDOWN does not change the performance of the algorithms in a noticeable way, although, in that case, the performance degradation due to noise is low. As the noise becomes higher (Fig. 1(g), 1(k), 1(f), 1(l)), restart and slow-down noticeably stabilizes all accelerated algorithms, reducing both their mean and their variance. Further, restart and slow-down generally outperforms the "theoretically optimal" AGD+ (TO-AGD+), which, as stated before, is a "lazy" version of AC-SA (Lan, 2012; Ghadimi & Lan, 2012) and is thus equivalent to it in the setting of unconstrained smooth minimization. Additional results are provided in Appendix D.

6. Conclusion

This paper presents a new accelerated algorithm together with the analysis of its associated error bounds in the cases when the gradient oracle is corrupted by additive noise. Moreover, motivated by the analytical results, we also provide simple semi-heuristics that restart and slow down the accelerated algorithms to reduce their error mean and variance. Our numerical experiments corroborate the analytical results. There are several interesting directions for future work that merit further investigation. For example, restart & slow-down approaches that do not require the explicit knowledge of the noise variance would be interesting for applications in engineered systems where gradients are estimated from noise-corrupted measurements.

Acknowledgements

Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing. It was partially supported by NSF grant #CCF-1718342, by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant #CCF-1740425 and by DHS-ALERT subaward 505035-78050. JD and LO would like to thank Guanghui Lan, Pavel Dvurechensky and the anonymous reviewers for useful comments.

The algorithm AGD+ for the noiseless case is due to Michael B. Cohen, who termed it a "proper extension of Nesterov's method" and shared it with JD during the Fall 2017 semester at the Simons Institute for the Theory of Computing. JD and LO dedicate this paper to the memory of Michael's brilliance and scholarship.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proc. ACM STOC'17*, 2017.

Allen-Zhu, Z. and Orecchia, L. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proc. ITCS'17*, 2017.

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Phys. Rev. E*, 64(6):061907, 2001.

Atchade, Y. F., Fort, G., and Moulines, E. On stochastic proximal gradient algorithms. *arXiv preprint arXiv:1402.2365*, 2014.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proc. IEEE FOCS'14*, 2014.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

Ben-Tal, A. and Nemirovski, A. *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications*. MPS-SIAM Series on Optimization. SIAM, 2001.

Bertsekas, D. P., Nedic, A., and Ozdaglar, A. E. *Convex Analysis and Optimization*. Athena Scientific, 2003.

Birand, B., Wang, H., Bergman, K., and Zussman, G. Measurements-based power control-a cross-layered framework. In *Proc. OSA OFC/NFOEC'13*, 2013.

Bubeck, S. *Theory of Convex Optimization for Machine Learning*. 2014. arXiv preprint, arXiv:1405.4980v1.

Bubeck, S., Lee, Y. T., and Singh, M. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint, arXiv:1506.08187*, 2015.

d'Aspremont, A. Smooth optimization with approximate gradient. *SIAM J. Optimiz.*, 19(3):1171–1183, 2008.

Devolder, O., Glineur, F., and Nesterov, Y. First-order methods of smooth convex optimization with inexact oracle. *Math. Prog.*, 146(1-2):37–75, 2014.

Diakonikolas, J. and Orecchia, L. The approximate duality gap technique: A unified theory of first-order methods. *arXiv preprint, arXiv:1712.02485*, 2017.

Diakonikolas, J. and Orecchia, L. Accelerated extra-gradient descent: A novel, accelerated first-order method. In *Proc. ITCS'18*, 2018.

Dvurechensky, P. and Gasnikov, A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *J. Optim. Theory Appl.*, 171(1):121–145, 2016.

Ene, A. and Nguyen, H. L. Constrained submodular maximization: Beyond 1/e. In *Proc. IEEE FOCS'16*, 2016.

Gasnikov, A. V. and Nesterov, Y. E. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58:48–64, 2018. doi: 10.1134/S0965542518010050.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optimiz.*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM J. Optimiz.*, 23(4):2061–2089, 2013.

Hardt, M. Robustness versus acceleration. http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration.html, 2014.

Hu, B. and Lessard, L. Control interpretations for first-order optimization methods. In *Proc. IEEE ACC'17*, 2017.

Hu, B., Seiler, P., and Lessard, L. Analysis of approximate stochastic gradient using quadratic constraints and sequential semidefinite programs. *arXiv preprint arXiv:1711.00987*, 2017.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. Accelerating stochastic gradient descent, 2018.

Kelner, J. A., Orecchia, L., Sidford, A., and Zhu, Z. A. A simple, combinatorial algorithm for solving SDD systems in nearly-linear time. In *Proc. ACM STOC'13*, 2013.

Krichene, W. and Bartlett, P. Acceleration and averaging in stochastic descent dynamics. In *Proc. NIPS'17*, 2017.

Krichene, W., Bayen, A., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. In *Proc. NIPS'15*, 2015.

Lan, G. An optimal method for stochastic composite optimization. *Math. Prog.*, 133(1-2):365–397, 2012.

Lessard, L., Recht, B., and Packard, A. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM J. Optimiz.*, 26(1):57–95, 2016.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Nemirovskii, A. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady AN SSSR (translated as Soviet Mathematics Doklady)*, volume 269, pp. 543–547, 1983.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Math. Prog.*, 162(1-2):83–112, 2017.

Sherman, J. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *Proc. ACM STOC'17*, 2017.

Spielman, D. A. and Teng, S.-H. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. ACM STOC'04*, 2004.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. In *Proc. Natl. Acad. Sci. U.S.A.*, 2016.