ESTIMATE SEQUENCE FOR CONVERGENCE OF ACCELERATED GRADIENT ITERATES MINIMIZING COSTS INVOLVING NON-EUCLIDEAN NORMS

*Bingcong Li** *Mario Coutiño*[†] *Georgios B. Giannakis**

* University of Minnesota - Twin Cities, Minneapolis, MN, USA
† Delft University of Technology, Delft, The Netherlands

ABSTRACT

A task of paramount importance in several applications deals with minimizing a convex cost function $f(\mathbf{x})$ with Lipschitz continuous gradient via accelerated gradient methods (AGMs). The focus of the present contribution is on the so-termed estimate sequence (ES) analysis tool for establishing convergence of AGM iterates. A generalized ES is introduced to support Lipschitz continuous gradient on *any* norm, a valuable advancement in practical settings involving optimization of non-Euclidean norms. Traditionally, ES consists of a sequence of quadratic functions that serve as surrogate functions of $f(\mathbf{x})$. However, such quadratic functions preclude the possibility of supporting Lipschitz continuous gradient defined for non-Euclidean norms. This much needed generalization to non-Euclidean norms is accomplished here through a *simple* yet nontrivial modification of the standard ES. The novel analysis provides useful insights on how acceleration is achieved along with interpretability of the involved parameters in ES. Finally, numerical tests demonstrate the convergence benefits of taking non-Euclidean norms into account.

***Index Terms*—** Nesterov's accelerated gradient method, estimate sequences, gradient descent, optimization

1. INTRODUCTION

In this work, we focus on solving the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \tag{1}$$

where f is a convex function with Lipschitz continuous gradient; d is the dimension of the decision vector $\mathbf{x}$. Throughout, $\mathbf{x}^*$ denotes the optimal solution of (1), and it is assumed that $f(\mathbf{x}^*) > -\infty$.

One of the standard solvers of (1) is through gradient descent (GD), which iteratively updates the decion vector as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k)$$

where k is the iteration index, and η_k is the step size. It is well known that GD converges with rate $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}(1/k)$. As the lower bound of first-order methods for convex problems is $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}(1/k^2)$, clearly GD is not optimal in terms of convergence rate [1].

To accelerate GD, Nesterov proposed an accelerated gradient method (AGM), which entails the following iterative updates

$$\mathbf{x}_{k+1} = \mathbf{y}_k - \alpha_k \nabla f(\mathbf{y}_k) \tag{2a}$$
$$\mathbf{y}_{k+1} = (1-\eta_k)\mathbf{x}_{k+1} + \eta_k \mathbf{x}_k \tag{2b}$$

where α_k and η_k are carefully designed step sizes [1, 2]. It has been established that AGM convergence matches the lower bound of first-order methods; that is, $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}(1/k^2)$. Thanks to its fast convergence, AGM and its variants, including FISTA [3] and variance-reduced AGM for finite-sum costs [4, 5, 6], are useful for several statistical signal processing applications; see e.g., [7, 8, 9].

Even though the fastest convergence rate is guaranteed, understanding the machinery behind AGM turns out to be challenging because most existing analyses do not provide intuition as clear as those for analyzing GD. In this work, the estimate sequence (ES) analysis tool that was first proposed in [1], with the goal of unveiling the mysteries behind it.

As formalized next, ES "estimates" f using a sequence of surrogate functions.

Definition 1. *(Estimate sequence) A tuple* $\left(\{\Phi_k(\mathbf{x})\}_{k=0}^{\infty}, \{\lambda_k\}_{k=0}^{\infty}\right)$ *is an ES of* $f(\mathbf{x})$*, if* $\lim_{k\to\infty}\lambda_k = 0$*, and for any* $\mathbf{x}\in\mathbb{R}^d$ *we have*

$$\Phi_k(\mathbf{x}) \le (1-\lambda_k)f(\mathbf{x}) + \lambda_k\Phi_0(\mathbf{x}).$$

As the choice of $\{\Phi_k(\mathbf{x})\}$ and $\{\lambda_k\}$ will become clear later, AGM iterations (2) can be derived from ES [1]. Though the intuition behind ES is still unclear, ES is a powerful tool that has been adopted for analyzing different algorithms [4, 5, 6, 10]. In this work, we will argue that ES "estimates" f in a two-way manner: i) through the progress made per iteration using (2); and ii) through the distance of $f(\mathbf{x}_{k+1})$ is from $f(\mathbf{x}^*)$. In addition, although the importance of smoothness defined on non-Euclidean norm is widely recognized [1, 2, 11, 12], existing analyses with ES only deal with Lipschitz continuous gradient defined on ℓ_2-norm. This prompted us to generalize ES to support smoothness on any norm.

Our detailed contributions are summarized below.

- **c1)** ES is generalized to support a Lipschitz continuous gradient defined on *any* norm.
- **c2)** In-depth explanation of acceleration is provided, along with its reflection on ES.
- **c3)** It is shown empirically that considering $\|\cdot\|_{\mathbf{Q}}$ with a *simple* but *carefully* designed $\mathbf{Q}$ can significantly improve convergence over that of standard AGM.

This research is supported in part by NSF 1508993, 1711471, 1901134 and the ASPIRE project (project 14926 within the STW OTP programme), financed by the Netherlands Organization for Scientific Research (NWO). Mario Coutino is partially supported by CONACYT. Emails:{lixx5599, georgios}@umn.edu; m.a.coutinominguez@tudelft.nl

2. PRELIMINARIES

Assumptions and definitions are introduced in this section.The importance of non-Euclidean norms in optimization is also mentioned. Throughout, the dual norm of a given norm $\|\cdot\|$ is given by $\|\cdot\|_*$.

Assumption 1. *(Convexity.) Function $f:\mathbb{R}^d\to\mathbb{R}$ is convex; that is, $f(\mathbf{y})-f(\mathbf{x})\geq\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle,\forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.*

Assumption 2. *(Gradient Lipschitz.) Function $f:\mathbb{R}^d\to\mathbb{R}$ has L-Lipchitz gradient with respect to (wrt) some norm $\|\cdot\|$; that is, $\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|_*\leq L\|\mathbf{x}-\mathbf{y}\|,\forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.*

Assumptions 1 and 2 hold throughout. For convenience, the notion of Lipschitz-continuous gradient and smoothness will be used interchangeably despite their slight difference. When considering the ℓ_2-norm, Assumption 2 reduces to the standard one $\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|_2\leq L\|\mathbf{x}-\mathbf{y}\|_2$. The consequence of Assumption 2 is the so-termed descent lemma that demands [11, Appendix B.1]

$$f(\mathbf{y})-f(\mathbf{x})\leq\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2. \tag{3}$$

As a simple example that relies on (3) to illustrate the importance of non-Euclidean norms, consider an f having L_1 and L_2 Lipchitz continuous gradient wrt the ℓ_1- and ℓ_2-norm, respectively. Plugging L_1 and L_2 in (3), and using that $\|\mathbf{x}\|_1\leq\sqrt{d}\|\mathbf{x}\|_2$, we deduce that $L_2\approx dL_1$. Since L_1 and L_2 influence the convergence rate of first-order methods, this suggests supporting smoothness wrt ℓ_1-norm is helpful to speedup converge.

To handle non-Euclidean norms, one would rely on the Bregman divergence [2, 11, 12].

Definition 2. *(Bregman divergence). If $R(\cdot)$ is 1-strongly convex wrt some norm $\|\cdot\|$, that is $R(\mathbf{y})\geq R(\mathbf{x})+\langle\nabla R(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2,\forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, the Bregman divergence wrt R is*

$$\mathcal{D}_R(\mathbf{y},\mathbf{x})=R(\mathbf{y})-R(\mathbf{x})-\langle\nabla R(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle.$$

Function $R(\cdot)$ is sometimes termed distance generating function (DGF). To illustrate the Bregman divergence, consider $R(\mathbf{x})=\frac{1}{2}\|\mathbf{x}\|_2^2$ that is 1-strongly convex wrt the ℓ_2-norm, and has Bregman divergence given by $\mathcal{D}_R(\mathbf{x},\mathbf{y})=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$. Another example involves the negative entropy as DGF, for which $R(\mathbf{x})=\sum_{i=1}^d x_i\ln x_i$ is known to be 1-strongly convex wrt the ℓ_1-norm, and the Bregman divergence is $\mathcal{D}_R(\mathbf{x},\mathbf{y})=\sum_{i=1}^d x_i\ln\frac{x_i}{y_i}-\sum_{i=1}^d(x_i-y_i)$, and it is also known as generalized KL divergence.

3. GENERALIZED ESTIMATE SEQUENCE

In this section, we broaden the scope of ES and lay out the *generic framework* of AGM with support of non-Euclidean norms. To this end, we construct a sequence of surrogate functions of f using $\mu_0>0$, $\{\mathbf{y}_k\}$, $\{\delta_k\in(0,1)\}$ (that will be specified later), as

$$\Phi_0(\mathbf{x})=\Phi_0^*+\mu_0\mathcal{D}_R(\mathbf{x},\mathbf{x}_0) \tag{4a}$$

$$\begin{aligned}\Phi_{k+1}(\mathbf{x})=&(1-\delta_k)\Phi_k(\mathbf{x}) \\ &+\delta_k\Big[f(\mathbf{y}_k)+\big\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\big\rangle\Big],\forall\,k\geq 0.\end{aligned} \tag{4b}$$

Our first result asserts that (4) with proper $\{\lambda_k\}$ is an ES for f.

Lemma 1. *If $\lambda_0=1$ and $\lambda_k=\lambda_{k-1}(1-\delta_{k-1})$, then the tuple $\big(\{\Phi_k(\mathbf{x})\}_{k=0}^\infty,\{\lambda_k\}_{k=0}^\infty\big)$ is an estimate sequence of $f(\mathbf{x})$.*

Proof. We show this by induction. For $\lambda_0=1$, it holds that $\Phi_0(\mathbf{x})=(1-\lambda_0)f(\mathbf{x})+\lambda_0\Phi_0(\mathbf{x})$. Suppose that $\Phi_k(\mathbf{x})\leq(1-\lambda_k)f(\mathbf{x}_k)+\lambda_k\Phi_0(\mathbf{x})$ is true for some k. We then have

$$\begin{aligned}\Phi_{k+1}(\mathbf{x})&=(1-\delta_k)\Phi_k(\mathbf{x})+\delta_k\Big[f(\mathbf{y}_k)+\big\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\big\rangle\Big]\\&\overset{(a)}{\leq}(1-\delta_k)\Phi_k(\mathbf{x})+\delta_k f(\mathbf{x})\\&\leq(1-\delta_k)\Big[(1-\lambda_k)f(\mathbf{x})+\lambda_k\Phi_0(\mathbf{x})\Big]+\delta_k f(\mathbf{x})\\&=(1-\lambda_{k+1})f(\mathbf{x})+\lambda_{k+1}\Phi_0(\mathbf{x})\end{aligned}$$

where (a) is due to the convexity of f; and the last equation follows from the definition of λ_{k+1}. Since $\lim_{k\to\infty}\lambda_k=0$, the tuple $\big(\{\Phi_k(\mathbf{x})\}_{k=0}^\infty,\{\lambda_k\}_{k=0}^\infty\big)$ satisfies the definition of an ES. □

We term $\{\Phi_k(\mathbf{x})\}$ in (4) and the corresponding $\{\lambda_k\}$ as *generalized ES*. If $R(\mathbf{x})=\frac{1}{2}\|\mathbf{x}\|_2^2$, the surrogate functions in (4) boil down to the standard ones in [1]. The key difference of (4) will be discussed later, but for now the ensuing result will demonstrate why ES is useful for analyzing AGM.

Proposition 1. *If for a sequence $\{\mathbf{x}_k\}$ it holds that $f(\mathbf{x}_k)\leq\min_{\mathbf{x}}\Phi_k(\mathbf{x})$, we have*

$$f(\mathbf{x}_k)\leq\lambda_k\big(\Phi_0(\mathbf{x}^*)-f(\mathbf{x}^*)\big),\forall\,k.$$

Proof. If $f(\mathbf{x}_k)\leq\min_{\mathbf{x}}\Phi_k(\mathbf{x})$ holds, then we have

$$f(\mathbf{x}_k)\leq\min_{\mathbf{x}}\Phi_k(\mathbf{x})\leq\Phi_k(\mathbf{x}^*)\leq(1-\lambda_k)f(\mathbf{x}^*)+\lambda_k\Phi_0(\mathbf{x}^*)$$

where in the last inequality we used Definition 1. Subtracting $f(\mathbf{x}^*)$ from both sides, we arrive at

$$f(\mathbf{x}_k)-f(\mathbf{x}^*)\leq\lambda_k\big(\Phi_0(\mathbf{x}^*)-f(\mathbf{x}^*)\big)$$

which completes the proof. □

Proposition 1 illustrates that the generalized ES is helpful to find a sequence $\{\mathbf{x}_k\}$ that is converging to $\mathbf{x}^*$. One can see that $\lambda_k=\prod_{\tau=0}^{k-1}(1-\delta_\tau)$ in Proposition 1 characterizes the convergence rate of $\{\mathbf{x}_k\}$. On the other hand, although the surrogate functions $\{\Phi_k(\mathbf{x})\}$ do not appear in Proposition 1 directly, they pose requirements on $\{\mathbf{x}_k\}$; that is, $\{\mathbf{x}_k\}$ should be chosen to satisfy $f(\mathbf{x}_k)\leq\min_{\mathbf{x}}\Phi_k(\mathbf{x})$.

The rest of this section will deal with the construction of sequences $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$, so that $f(\mathbf{x}_k)\leq\min_{\mathbf{x}}\Phi_k(\mathbf{x})$ is guaranteed for all k. To this end, we need to take a closer look at the surrogate functions $\{\Phi_k(\mathbf{x})\}$ in (4).

Lemma 2. *The functions $\Phi_k(\mathbf{x})$ in (4) can be rewritten as $\Phi_k(\mathbf{x})=\Phi_k^*+\mu_k\mathcal{D}_R(\mathbf{x},\mathbf{v}_k)$, where $\Phi_k^*=\min_{\mathbf{x}}\Phi_k(\mathbf{x})$, and $\Phi_k(\mathbf{v}_k)=\Phi_k^*$. Furthermore, we have*

$$\mu_{k+1}=(1-\delta_k)\mu_k \tag{5a}$$

$$\mathbf{v}_{k+1}=\arg\min_{\mathbf{v}}\Big\langle\frac{\delta_k}{\mu_{k+1}}\nabla f(\mathbf{y}_k),\mathbf{v}-\mathbf{v}_k\Big\rangle+\mathcal{D}_R(\mathbf{v},\mathbf{v}_k) \tag{5b}$$

$$\begin{aligned}\Phi_{k+1}^*=&(1-\delta_k)\Phi_k^*+\delta_k f(\mathbf{y}_k)+\mu_{k+1}\mathcal{D}_R(\mathbf{v}_{k+1},\mathbf{v}_k)\\&-\delta_k\big\langle\nabla f(\mathbf{y}_k),\mathbf{y}_k-\mathbf{v}_{k+1}\big\rangle.\end{aligned} \tag{5c}$$

Proof. See supplemental material online at [13]. □

Algorithm 1 AGM

```
Initialize: $\mathbf{x}_0$, $\{\delta_k\}$, and $\{\mu_k\}$
$\mathbf{v}_0 = \mathbf{x}_0$
for $k = 0, 1, \ldots, K-1$ do
    $\mathbf{y}_k = \delta_k \mathbf{v}_k + (1-\delta_k)\mathbf{x}_k$
    $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \langle \nabla f(\mathbf{y}_k), \mathbf{x}-\mathbf{y}_k \rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}_k\|^2$
    $\mathbf{v}_{k+1} = \arg\min_{\mathbf{v}} \left\langle \frac{\delta_k}{\mu_{k+1}} \nabla f(\mathbf{y}_k), \mathbf{v}-\mathbf{v}_k \right\rangle + \mathcal{D}_R(\mathbf{v}, \mathbf{v}_k)$
end for
Return: $\mathbf{x}_K$
```

With the alternative expressions of $\Phi_k(\mathbf{x})$, Lemma 2 relates $\mathbf{v}_{k+1}$ with $\mathbf{v}_k$ (Φ^*_{k+1} and Φ^*_k). In addition, Lemma 2 shows the key difference of our generalized ES with the standard one in [1]. As $R(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ in standard ES, simple calculation shows that $\Phi_k(\mathbf{x})$ is exactly μ_k-strongly convex wrt ℓ_2-norm (in fact Φ_k is a quadratic function). However, when considering a general $R(\mathbf{x})$, we have $\mathcal{D}_R(\mathbf{x}, \mathbf{y}) \geq \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$. This means that though $\Phi_k(\mathbf{x})$ is strongly convex wrt $\|\cdot\|$, the parameter μ_k is always an underestimate of its strongly convexity parameter.

Based on Lemma 2, the following lemma guides the choice of $\mathbf{y}_k$ and $\mathbf{x}_k$ to ensure $f(\mathbf{x}_k) \leq \Phi^*_k$, which is the requirement in Proposition 1 for establishing the convergence of $\mathbf{x}_k$.

Lemma 3. *Choose* $\Phi^*_0 = f(\mathbf{x}_0)$, $\mathbf{y}_k = \delta_k \mathbf{v}_k + (1-\delta_k)\mathbf{x}_k$, *and* $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \langle \nabla f(\mathbf{y}_k), \mathbf{x}-\mathbf{y}_k \rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}_k\|^2$. *If* $L\delta_k^2 \leq \mu_{k+1}$ *is satisfied, it holds that* $f(\mathbf{x}_k) \leq \Phi^*_k, \forall k \geq 0$.

Proof. See supplemental material online at [13]. □

With the choices of $\{\mathbf{x}_k\}$, $\{\mathbf{y}_k\}$, and $\{\mathbf{v}_k\}$ in Lemmas 2 and 3, we summarize the AGM with support to non-Euclidean norms in Alg. 1. For non-Euclidean norms induced by a positive definite matrix, the closed-form updates for $\mathbf{x}_{k+1}$ and $\mathbf{v}_{k+1}$ will be discussed in Section 4.2.

Next, we establish the convergence rate of Alg. 1.

Theorem 1. *Choosing* $\mu_0 = 2L$, $\delta_k = \frac{2}{k+3}$, *Alg.1 guarantees that*

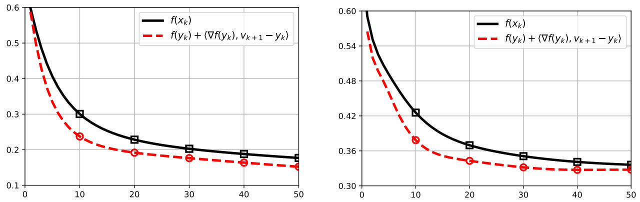
$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}\Big(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*) + L\mathcal{D}_R(\mathbf{x}^*, \mathbf{x}_0)}{k^2}\Big), \forall k.$$

Proof. By the choice of parameters, one can verify that $L\delta_k^2 \leq \mu_{k+1}$ holds. And the choices of $\{\mathbf{x}_k\}$, $\{\mathbf{y}_k\}$, and $\{\mathbf{v}_k\}$ guarantee $f(\mathbf{x}_k) \leq \Phi^*_k$ as shown in Lemma 3. Therefore, we can directly apply Proposition 1 to obtain

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^*) &\leq \lambda_k \big(\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big) \\ &= \frac{2\Big[f(\mathbf{x}_0) - f(\mathbf{x}^*) + 2L\mathcal{D}_R(\mathbf{x}^*, \mathbf{x}_0)\Big]}{(k+1)(k+2)} \end{aligned}$$

which completes the proof. □

Theorem 1 suggests that AGM has a lower bound matching convergence rate $\mathcal{O}(1/k^2)$. Note also that Alg. 1 recovers the so-termed "linear coupling" [11], which is believed to be very different from AGM. However, our generalized ES suggests that linear coupling is a natural consequence of Nesterov's acceleration technique. The only minor difference is that the analysis in [11] supports the choice $\delta_k = \frac{2}{k+2}$, while ours selects[1] $\delta_k = \frac{2}{k+3}$. Although different, both choices exhibit an $\mathcal{O}(1/k)$ behavior.



(a) dataset *w1a*

(b) dataset *a3a*

Fig. 1. Validation of the intuitive explanation of acceleration.

4. INTUITION AND NON-EUCLIDEAN NORMS

In this section, we examine Alg. 1 from a "linear coupling" [11] point of view to better understand the generalized ES. Subsequently, we present a case study to illustrate the merits of considering non-Euclidean norms together with numerical tests.

4.1. Redux of ES through the "linear coupling" lens

In "linear coupling" [11], the gradient descent and mirror descent are coupled to effect acceleration. We first rewrite the updates of AGM using the notation in [11]. The variable $\mathbf{x}_{k+1}$ is obtained via a generalized GD, that is

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathtt{Grad}(\mathbf{y}_k) \\ &:= \arg\min_{\mathbf{x}} \langle \nabla f(\mathbf{y}_k), \mathbf{x}-\mathbf{y}_k \rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}_k\|^2 \end{aligned} \tag{6}$$

while $\mathbf{v}_{k+1}$ is obtained by mirror descent (MD)

$$\begin{aligned} \mathbf{v}_{k+1} &= \mathtt{Mirr}\Big(\mathbf{v}_k, \frac{\delta_k}{\mu_{k+1}} \nabla f(\mathbf{y}_k)\Big) \\ &:= \arg\min_{\mathbf{v}} \Big\langle \frac{\delta_k}{\mu_{k+1}} \nabla f(\mathbf{y}_k), \mathbf{v}-\mathbf{v}_k \Big\rangle + \mathcal{D}_R(\mathbf{v}, \mathbf{v}_k) \\ &= \arg\min_{\mathbf{v}} \langle \nabla f(\mathbf{y}_k), \mathbf{v}-\mathbf{v}_k \rangle + \frac{\mu_{k+1}}{\delta_k} \mathcal{D}_R(\mathbf{v}, \mathbf{v}_k). \end{aligned} \tag{7}$$

The consequence of finding $\mathbf{x}_{k+1}$ using (6) is $f(\mathbf{x}_{k+1}) - f(\mathbf{y}_k) \leq -\frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|_*^2$ as shown in the proof of Lemma 3. This inequality reveals how much progress is made per iteration when moving from $\mathbf{y}_k$ to $\mathbf{x}_{k+1}$.

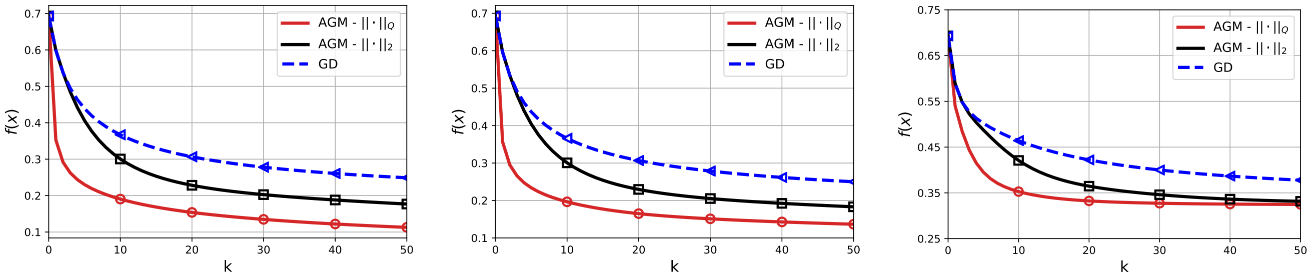
On the other hand, the MD step is used to estimate the optimality gap of current iterates. To see this, recall that for any $\mathbf{u} \in \mathbb{R}^d$ convexity implies that the following inequality holds

$$\begin{aligned} f(\mathbf{u}) &\geq f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{u}-\mathbf{y}_k \rangle \\ &= f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{u}-\mathbf{v}_k \rangle + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_k-\mathbf{y}_k \rangle. \end{aligned} \tag{8}$$

Since $f(\mathbf{u}) \geq f(\mathbf{x}^*), \forall \mathbf{u}$, it is natural to use (8) to obtain an estimate of $f(\mathbf{x}^*)$. Since the RHS of (8) is linear in $\mathbf{u}$, one would instead minimize the regularized version of the RHS of (8) as in (7) to obtain a worst-case estimate of $f(\mathbf{x}^*)$. Hence, obtaining $\mathbf{v}_{k+1}$ amounts to finding an approximation of the optimality gap via (8). The role of $\{\mathbf{v}_k\}$ in the generalized ES is thus unveiled: *it helps to construct the optimality gap*. This intuition is validated by the numerical tests in Fig. 1, where the RHS of (8) is always less than $f(\mathbf{x}_k)$ as an estimate of $f(\mathbf{x}^*)$.

In a nutshell, both GD and MD effect acceleration: using GD for descent; while consulting MD for estimating the optimality gap.

[1] Theoretically, $\delta_k = \frac{2}{k+3}$ also works for linear coupling.



(a) *w1a*

(b) *w7a*

(c) *a9a*

Fig. 2. Tests AGM with $\|\cdot\|_{\mathbf{Q}}$ on different datasets.

4.2. Case study: quadratic norm

In this subsection, we consider smoothness w.r.t. the quadratic norm, $\|\cdot\|_{\mathbf{Q}}$, where $\mathbf{Q}\in\mathbb{S}^d_{++}$ is a positive definite matrix. In this case, it is natural to choose $R(\mathbf{x})=\frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$ with $\mathcal{D}_R(\mathbf{x},\mathbf{y})=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbf{Q}}$. The updates on $\mathbf{x}_{k+1}$ and $\mathbf{v}_{k+1}$ (Lines 5 and 6 in Alg. 1) can thus be rewritten in closed form as

$$\mathbf{x}_{k+1}=\mathbf{y}_k-\frac{1}{L}\mathbf{Q}^{-1}\nabla f(\mathbf{y}_k) \tag{9a}$$

$$\mathbf{v}_{k+1}=\mathbf{v}_k-\frac{\delta_k}{\mu_{k+1}}\mathbf{Q}^{-1}\nabla f(\mathbf{y}_k). \tag{9b}$$

Despite the closed-form update, the main massage here is that a properly designed $\mathbf{Q}$ can be helpful to speedup convergence. Intuitively, choosing $\mathbf{Q}$ as an approximation of the Hessian can be helpful. However, since AGM is a first-order method, one wants to find $\mathbf{Q}$ using first-order information only.

Inspired by the well known AdaGrad [14, 15], which has similar updates as (9a), we propose to obtain $\mathbf{Q}$ using a few gradients as AdaGrad does. Specifically, setting $\mathbf{z}_0=\mathbf{x}_0$, and performing t steps of GD on $\mathbf{z}_k$, i.e., $\mathbf{z}_{k+1}=\mathbf{z}_k-\frac{1}{L_2}\nabla f(\mathbf{z}_k)$, where L_2 is the smoothness parameter w.r.t. ℓ_2-norm, we can then choose $\mathbf{Q}$ as

$$\mathbf{Q}=c\cdot\mathtt{diag}\left(\sqrt{\frac{1}{t}\sum_{k=0}^{t-1}\big(\nabla f(\mathbf{z}_k)\big)^2+\epsilon\mathbf{1}}\right) \tag{10}$$

where $(\cdot)^2$ and $\sqrt{\cdot}$ denote element-wise square and square-root operations, respectively; $\mathtt{diag}(\boldsymbol{\theta})$ denotes a diagonal matrix whose diagonal entries are given by the vector $\boldsymbol{\theta}$; $\epsilon>0$ is a small offset to guarantee the positive definiteness of $\mathbf{Q}$; and $c>0$ is a tunable scaler. One can view $\mathbf{Q}$ as an estimated Hessian using first-order information. As for the choice of t, in practice we have found in our experiments that a small number ($t\approx 3$) performs well. Hence, using (10) to find $\mathbf{Q}$ does not incur a major computational overhead.

5. NUMERICAL TESTS

In this section, we illustrate our theoretical findings in the classical binary classification task using logistic regression, and the proposed construction for the matrix $\mathbf{Q}$ [cf. (10)]. The loss function is

$$f(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-b_i\langle\mathbf{a}_i,\mathbf{x}\rangle\big)\Big)$$

where $\mathbf{a}_i$ and b_i are the feature and label of datum i, respectively; and n is the total number of data. We choose standard GD and Nesterov's standard acceleration approach (AGM with l_2-norm) as benchmarks. For the implementation of AGM with $\|\cdot\|_{\mathbf{Q}}$, we consider $\mathbf{Q}$ specified by (10) with $\epsilon=10^{-4}$ and $c=10$.

Table 1. Parameters of datasets used, where d is the dimensionality of feature vectors, n the number of data, and "density" refers to the percentage of non-zero entries among all feature vectors.

dataset	d	n	density
w1a	300	2477	3.82%
w7a	300	24,692	3.89%
a9a	122	32,561	11.37%

We run tests on datasets *w1a*, *w7a*, and *a9a*[2], whose detailed descriptions are listed in Tab. 1. Performance of the considered algorithms is depicted in Fig. 2. The proposed AGM with $\|\cdot\|_{\mathbf{Q}}$ significantly outperforms the original AGM with $\|\cdot\|_2$. For example, on dataset *w1a*, the proposed method uses around 10 iterations to achieve $f(\mathbf{x}_k)=0.2$, while standard AGM requires 30 iterations.

Notice that the convergence improvement achieved by using quadratic norm in AGM over the standard AGM is more pronounced when sparsity is present (see Fig. 2 (b) and (c)). As $\mathbf{Q}$ is obtained in the spirit of AdaGrad, such "sparsity preference" behavior is consistent with the observation made in [15], where AdaGrad also performs better on sparse data.

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this work, the estimate sequence (ES) analysis tool was extended to support smoothness defined on any norm. In-depth explanation of how acceleration is achieved, and the meaning of $\{\mathbf{v}_k\}$ in ES were provided. Our theoretical findings led to an efficient method, where $\|\cdot\|_{\mathbf{Q}}$ is taken advantage of to improve the performance of the standard AGM. Numerical tests corroborated that the novel algorithm markedly outperforms standard AGM.

Investigating generalized ES on strongly convex problems is a challenging future research topic because μ_k is an underestimate of the strongly convex surrogate $\Phi_k(\mathbf{x})$.

[2] Online available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`

7. REFERENCES

[1] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2004, vol. 87.

[2] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[4] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," in *Proc. Advances in Neural Info. Process. Syst.*, Montreal, Canada, 2014, pp. 1574–1582.

[5] H. Lin, J. Mairal, and Z. Harchaoui, "A universal catalyst for first-order optimization," in *Proc. Advances in Neural Info. Process. Syst.*, Montreal, Canada, 2015, pp. 3384–3392.

[6] A. Kulunchakov and J. Mairal, "Estimate sequences for variance-reduced stochastic composite optimization," in *Proc. Intl. Conf. on Machine Learning*, 2019.

[7] A. P. Liavas, G. Kostoulas, G. Lourakis, K. Huang, and N. D. Sidiropoulos, "Nesterov-based alternating optimization for nonnegative tensor factorization: Algorithm and parallel implementation," *IEEE Trans. Signal Processing*, vol. 66, no. 4, pp. 944–953, 2017.

[8] R. Gu and A. Dogandžić, "Projected Nesterov's proximal-gradient algorithm for sparse signal recovery," *IEEE Trans. Signal Processing*, vol. 65, no. 13, pp. 3510–3525, 2017.

[9] T. Ramachandran, M. H. Nazari, S. Grijalva, and M. Egerstedt, "Overcoming communication delays in distributed frequency regulation," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2965–2973, 2015.

[10] B. Li, L. Wang, and G. B. Giannakis, "Almost tune-free variance reduction," *arXiv preprint arXiv:1908.09345*, 2019.

[11] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *arXiv preprint arXiv:1407.1537*, 2014.

[12] A. S. Nemirovsky and D. B. Yudin, "Problem complexity and method efficiency in optimization." 1983.

[13] B. Li, M. Coutino, and G. B. Giannakis, "Revisit of estimate sequence for accelerated gradient methods." [Online]. Available: https://www.dropbox.com/s/85nus46df6xk1pz/Acc.pdf?dl=0

[14] H. B. McMahan and M. Streeter, "Adaptive bound optimization for online convex optimization," *arXiv preprint arXiv:1002.4908*, 2010.

[15] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.