J. Nonlinear Var. Anal. 10 (2026), No. 3, pp. 523-553
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.10.2026.3.03

NESTEROV'S ACCELERATION AT THE LIMIT: FIRST-ORDER SCHEMES

ENDRIT DOSTI[1], SERGIY A. VOROBYOV[1,*], THEMISTOKLIS CHARALAMBOUS[2,3]

[1]*Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[2]*Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus*
[3]*Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland*

Abstract. One of the major early-career contributions to numerical optimization by Yurii Nesterov is the development of Nesterov's acceleration and the corresponding Fast Gradient Method (FGM). Accelerated first-order methods are very important in large-scale optimization and have applications in different fields of engineering and science. Such methods are devised in the context of the estimating sequences framework, which was also developed by Nesterov, but much later than FGM, and exhibit desirable properties such as fast convergence rate and low per-iteration complexity. In this paper, we devise new generalized estimating sequences with an objective of pushing acceleration to its limit and show how they can be used to construct accelerated first-order methods. We start our summary by considering the case of minimizing smooth convex objective functions. For this class of problems, we present a class of generalized estimating sequences, constructed by exploiting the history of the estimating functions that are obtained during the minimization process. Using these generalized estimating sequences, we devise an accelerated gradient method and prove that it converges to a tolerated neighborhood of the optimal solution faster than FGM and other first-order methods. We then consider a more general class of optimization problems, namely composite objectives. For this class of problems, we introduce the class of composite estimating sequences, which are obtained by making use of the gradient mapping framework and a tight lower bound on the function that should be minimized. Using these composite estimating sequences, we devise a composite objective accelerated multi-step estimating sequence technique and prove its accelerated convergence rate. Last, embedding the memory term coming from the previous iterates into the composite estimating sequences, we obtain the generalized composite estimating sequences. Using these estimating sequences, we construct another accelerated gradient method and prove its accelerated convergence rate.

Keywords. Convergence rate; Estimating sequences; First-order accelerated optimization algorithms; Non-smooth optimization; Nesterov's acceleration.

1. INTRODUCTION

One of the major contributions to numerical optimization by Yurii Nesterov is the development of what is nowadays known as Nesterov's acceleration and the corresponding Nesterov

*Corresponding author.
E-mail address: ext-endrit.dosti@aalto.fi (E. Dosti), sergiy.vorobyov@aalto.fi (S.A. Vorobyov), charalambous.themistoklis@ucy.ac.cy (T. Charalambous).

Received 2 December 2025; Accepted 8 January 2026; Published online 1 April 2026.

©2026 Journal of Nonlinear and Variational Analysis

accelerated gradient algorithm better known as Fast Gradient Method (FGM). Although in recent years Nesterov is more concerned with accelerated higher-order schemes, accelerated first-order methods gained exceptionally high importance in large-scale optimization and machine learning with dramatically extended applications in different fields of engineering and science. Such methods are devised in the context of the estimating sequences framework, which was also developed by Nesterov, but much later than FGM itself. This paper summarizes our original recent work on generalizing the estimating sequences with the objective of finding the limit of acceleration for the corresponding first-order schemes considering the black-box framework.

All optimization problems that are encountered in science and engineering can be either non-convex or convex [1, 2, 3]. Despite the recent advances in optimization theory, finding and certifying their global solutions remain challenging [4]. On the other hand, the class of convex optimization problems has gathered significant attention in the research community [1, 4, 5]. Different from the case of non-convex problems, for the class of convex problems it is possible to find and certify the globally optimal solutions (or an arbitrarily tight approximation of them) [1]. Such problems arise often in the context of applications in many fields such as data analysis, signal processing, information theory and wireless communications [6, 7, 8, 9]. A myriad of such problems can be solved exactly. Nevertheless, in the context of modern engineering applications which are enabled by big data, we are more interested in finding approximate solutions, which can be computed efficiently [10].

Depending on the size of the underlying datasets, it is natural to seek to optimize the trade-off between a high per-iteration complexity and convergence speed. Methods that exhibit a higher per-iteration complexity, such as Newton and/or quasi-Newton type methods (e.g., Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [1]) also exhibit fast convergence. However, as the size of the datasets grows large, it becomes necessary to seek to devise methods that exhibit a low per-iteration complexity, and as fast as possible convergence. One of the most popular tools used to solve the large-scale optimization problems, are gradient-based methods designed to be agnostic to the problem formulation, i.e., considering the black-box framework. At each iteration, these methods query a black-box oracle to obtain relevant insight about the function that is being minimized [11]. To build efficient gradient-based methods, the following aspects need to be considered: 1. Such methods need to converge to a neighborhood of the optimal solution; 2. The number of first-order oracle calls, together with additional computations, need to be minimized [4, 5]. The performance bounds for different black-box gradient-based methods for different types of convex problems have been thoroughly investigated and established in [8, 9, 11, 12].

In this paper, we consider the problem of devising accelerated methods for solving smooth and non-smooth convex optimization problems. Considering only the problem of devising efficient gradient-based methods for solving convex optimization problems with smooth objective, one of the most celebrated results is the development of the FGM [13]. Based on the framework devised in [11], FGM is referred to as an optimal method, i.e., the method minimizes the calls of a first-order oracle while exhibiting a convergence rate $\mathcal{O}(1/k^2)$, where k is the iteration counter. On a framework level, one of the most significant advances was the development of the estimating sequences framework, initially introduced in [14] and later refined in [15, 16]. Using this framework, further variants of FGM were constructed for solving optimization problems which have smooth and strongly convex cost functions [14], [16, Constant Step Scheme I].

These variants of FGM require at most $\sqrt{\kappa}\left(\ln \varepsilon^{-1} + \mathcal{O}(1)\right)$ iterations to converge to a point x with $f(x) - f^* \leq \varepsilon$, where $\kappa = L/\mu$ with L and μ denoting the Lipschitz constant and strong convexity parameter, respectively, while $\varepsilon > 0$ is the required tolerance and f^* stand for the optimal value of the objective function $f(x)$.

Despite the consideration that the complexity bounds reached by FGM-type methods are only proportional to the fundamental performance bounds introduced in [11], FGM and its different variants have always been regarded in the literature to be optimal methods. Interestingly, these methods started gathering more attention only after the publication of the seminal work on smoothing techniques [17], wherein Nesterov approximated a non-smooth convex cost function by another smooth convex cost function. FGM was then used to minimize the approximated function. The authors in [18] further extended the work by devising new interior gradient algorithms which also exhibit an accelerated convergence rate. In another line of work, detailed in [19, 20, 21], several researchers have studied the problem of robustness of FGM-type methods with respect to the usage of inaccurate gradients of the objective function in the minimization process.

More recently, in addition to the estimating sequences framework, other frameworks that can be used to accelerate gradient-based methods have been developed. In the line of work presented in [22, 23, 24], existing links between the integration of ordinary differential equations (ODE) and optimization have been considered in the context of devising a different perspective on acceleration of first-order methods with an objective to devise also an *intuitive* interpretation of Nesterov's acceleration. More specifically, in [23] the authors derived a second-order ODE which is the limit of FGM when the step size goes to zero. In [22], the authors showed that different accelerated gradient methods can be reformulated as constant parameter second-order ODEs. Moreover, they have shown the equivalence between the stability of such systems and the accelerated convergence rate, which was also found in [4, 25] by setting a relationship between estimating function used for defining estimating sequences in accelerated optimization methods and Lyapunov function used for studying stability. Then, accelerated schemes such as FGM can be intuitively interpreted as looking ahead and measuring a curvature of an objective function in an extrapolated point, rather than in a point given by a current update. To ensure stability/convergence, the extrapolated point has to be selected based on the corresponding Lyapunov function/estimating function. Last, the authors of [24] have demonstrated that different variants of FGM can be viewed as instances of a structured approach to transition from the continuous-time curves created by the Bregman Lagrangian to accelerated algorithms.

In another line of work, the authors of [26] have shown that it is possible to devise different variants of FGM by making use of the linear coupling between mirror and gradient descent. Yet another line of work has been introduced in [27, 28]. Specifically, the authors of [27] have developed an accelerated gradient method by extending the results originated for the ellipsoid method. The resulting method called Geometric Descent Method (GDM) is more efficient than FGM, however, it has the drawback that it requires an exact line search to ensure accelerated convergence. The links between the GDM of [27] and the strongly-convex variants of FGM have been later established in [28]. Another line of work [29] has used principles of robust control theory to derive convergence rate results for accelerated gradient methods. The authors of [30] have used the analysis presented in [29] to construct a more efficient method, which they name as Triple Momentum Method (TMM). TMM is more efficient than FGM, in the

sense that it exhibits a faster convergence rate, however, it has the drawback that it is defined only for strongly convex objective functions. Even for this class of problems, when the value of the condition number is large, TMM exhibits slower convergence rate than FGM (for more details see [31, Figure 1]).

Another interesting line of work has been introduced in [32]. Therein, the authors have casted a semidefinite program (SDP) which is used to model the improvement of the worst accuracy that a black-box numerical method can exhibit. Later, the authors of [33] have analyzed the tightness of the worst-case accuracies that the SDP yields. These results have paved path to the development of new classes of optimal methods for minimizing smooth and non-strongly convex cost functions [34, 35]. Using the framework introduced in [32], the authors of [31] have developed an optimal method for solving smooth and strongly convex optimization problems. The method proposed therein reaches the complexity bounds established in [11]; however, it has several drawbacks. First, it is difficult to extend the framework to broader and more practical optimization setups, such as non-smooth optimization, stochastic optimization, etc. Second, the results demonstrated for the method are achieved by assuming that parameters relevant to the objective function (e.g., μ, L) are known. The sensitivity and robustness of the method to the inexact values of these parameters in the context of practical deployments requires further analysis and evaluation.

Different from all the other frameworks which have been used to develop accelerated gradient-based algorithms, estimating sequences have been consistently used to develop numerical methods that exhibit a competitive performance in a myriad of applications and optimization setups. In the context of applications, a myriad of novel results have been presented in [36, 37, 38, 39]. Specifically, in [36] the authors have devised an accelerated gradient method used for minimizing a smooth loss function regularized by the trace norm of the matrix variable. In [37], the authors have developed efficient distributed methods and shown that their results match the existing results for FGM, with the additional cost coming from the communication constrains. Moreover, the authors of [38] have considered the coupling of FGM-type of acceleration, multi-consensus and gradient tracking to devise algorithms that achieve optimal computation complexity and near-optimal communication complexity. Last, in [39] the authors have developed an efficient variant of FGM by using the principle of differential quantization.

Estimating sequence-based approaches have also been successfully extended to other optimization setups. Many interesting results have been established in the context of stochastic optimization [40, 41, 42]. In [40], the authors have developed a stochastic accelerated gradient method for solving regularized risk minimization problems. An accelerated stochastic approximation algorithm based on FGM has been presented in [41]. A new class of stochastic estimating sequences has been presented in [42]. These stochastic estimating sequences have been then used to devise efficient and robust stochastic methods. The development of non-Euclidean methods has also been widely investigated recently [43, 44]. The new estimating functions introduced in [44] have been used to devise a novel bound on the nonlinear metric distortion to devise a Riemannian version of FGM. The method proposed therein exhibits accelerated convergence rate for finding the optimal solution of geodesically convex problems, which are smooth and strongly convex. In [44], the authors have presented new estimating functions, which have been then used to devise the first global accelerated gradient method for Riemannian manifolds. Another relevant setup to which the estimating sequences framework

has been successfully extended is the design of higher-order methods [15, 45, 46], which is the main theme of Nesterov's current research.

In [15], the author has presented a unified framework which can be used for studying estimating sequences methods, and shown how to use the framework to develop accelerated algorithms. An accelerated version of the Newton method has been presented in [45]. Moreover, accelerated high-order proximal methods developed using the inexact oracle framework have been presented in [46]. Another setup wherein the acceleration effect obtained by utilizing the estimating sequences framework becomes relevant is related to non-convex problems [47, 48]. The generalization of FGM to non-convex setups has been introduced in [47]. Moreover, for nonconvex functions with Lipschitz continuous first and second derivatives, a Hessian-free accelerated gradient method has been developed in [48].

Estimating sequences can also be considered to devise efficient methods to solve constrained optimization problems. The fundamentals behind such extensions have been introduced in [16, Chapters 2.2.4 - 2.2.5]. The key idea behind such extension lies in exploring the coupling of the estimating sequences framework together with the gradient mapping framework [11]. A similar approach can also be used for solving problems with convex composite objective functions. Extensions of these frameworks to solving such problems have been introduced in [49, 50, 51]. In [49], Nesterov also has introduced a new class of estimating sequences and used them to derive an accelerated gradient method called Accelerated Multi-step Gradient Scheme (AMGS). Together with AMGS, Nesterov has also introduced a backtracking strategy which is used to estimate the value of the Lipschitz constant. In the same work, he has also presented an efficient technique for approximating the strong convexity parameter of the cost function. The main drawback of AMGS comes due to its high per-iteration complexity because for each iteration it needs two projection-like operations. This issue has been mitigated with the development of Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [50]. The method exhibits an accelerated rate of convergence and a lower per-iteration complexity. Despite the attractive properties, FISTA does not convergence as fast as AMGS when considered in practical deployments [25, 52]. Another class of composite estimating sequences has been introduced in [51]. Different from the estimating sequences presented in [49], the composite estimating sequences are used to devise accelerated gradient-based schemes which require one projection-like operation per iteration. Moreover, the method constructed therein, converges faster than both AMGS and FISTA when tested on practical problems and real-world datasets.

In summary, many gradient-based methods have already been studied in the literature in the context of different applications and optimization setups. Independently of the framework used for designing the methods, in order for them to be optimal in the sense of [11] when considering smooth convex optimization, the following are important: 1. a method achieves an accelerated convergence rate; 2. estimated number of iterations is proportional to the complexity bounds given in [11]. In the context of composite objectives with non-smooth term, it is desirable for the resulting methods to exhibit an accelerated convergence rate. A unified framework that can be used for developing gradient-based algorithms is presented in [53]. In [33], the authors have derived the exact worst-case bounds for the variant of FGM presented in [13]. As we discussed earlier, the variant of FGM built using the estimating sequences framework is presented in [14, 16]. In [16], the author have argued that one of the most relevant considerations for designing optimal methods relates to parsing global topological information about the cost function. The

collection of such information is enabled by the estimating sequences. They consist of the sequences $\{\lambda_k\}_k$ and $\{\phi_k(x)\}_k$, which enable the computation of the rate of convergence for the iterates and accumulation of information around them. Considering the popularity of estimating sequence methods, one can easily conclude that such an intuition is correct.

A major challenge with the estimating sequences framework arises because the estimating functions are not unique. Finding a structure for estimating functions that always results in the most efficient (both when considering the theoretical bounds and practical performance) methods that can be devised for the corresponding problem classes remains an open question. As we have already discussed, different FGM variants, e.g., the ones presented in [16, Constant Step Scheme I], [13, 49], etc., are built using different estimating functions. Nevertheless, they are all very efficient and enjoy the accelerated convergence rate properties. Despite the different structures for the estimating functions, all variants of FGM share the commonality that the parameters in iteration $k+1$, are updated by considering the values of the parameters in iteration k.

1.1. **Objectives.** Considering the plethora of frameworks for devising accelerated first-order methods, together with the possibilities to construct more efficient estimating functions, it is natural to ask: "Is it possible to construct more efficient methods by changing the structure of the estimating functions?". This is a central question positively answered in this paper for both minimizing smooth and convex cost functions and minimizing composite objective functions with a non-smooth term. This paper is a summary of our currently published and unpublished works considering different problem setups. The main contributions collected in this paper are summarized in the next Subsection.

1.2. **Summarized contributions.**

- We formalize the generalized estimating sequences framework and provide links between the momentum terms used to construct the proposed generalized estimating sequences with the heavy-ball momentum. Moreover, therein we establish the convergence results of our proposed method and prove that it converges faster than FGM [51].
- We prove new results and implications of using our proposed composite estimating sequences. Furthermore, we formalize and establish the convergence of our proposed composite objective multistep estimating sequence technique (COMET). We show that COMET requires only one projection-like operation per iteration and is more efficient than existing numerical methods for minimizing functions with composite structure [54].
- We show how to further extend the generalized estimating sequences framework for minimizing convex and composite cost functions. We embed the heavy-ball type of momentum into the composite estimating sequences introduced in the previously listed contribution. We use the new estimating functions to construct another numerical method and demonstrate its efficiency in solving practical problems with real-world datasets [55].

1.3. **Organization of the paper.** The remainder of the paper is organized as follows. Section 2 introduces the generalized estimating sequences framework for smooth functions. In the same Section, we also present the associated method and establish its convergence. Sections 3 and 4 focus on further extending the framework to composite objectives. Section 3 uses a tight bound

for the cost function and the gradient mapping framework to construct composite estimating sequences and the corresponding method for minimizing convex and composite cost functions. Section 4 further extends the work and introduces the generalized composite estimating sequences. Using these estimating sequences, we build yet another algorithm and establish its accelerated rate. Numerical example demonstrating the performance of the proposed methods using real-dataset is given in Section 5. Section 6 presents our final remarks of the work and highlights several remaining open problems.

2. Generalizing the Estimating Sequences

Many accelerated gradient-based algorithms have been devised based on the estimating sequences framework [13, 15, 16]. In this section, we begin by looking at the simplest case of smooth and strongly convex objective functions and highlight the main findings of [51]. We have already discussed that the existing variants of FGM that are built using different estimating sequences share the commonality that updates at iteration $k+1$ are obtained by considering *only* the updates at iteration k. Considering the existing results on the heavy-ball method [56], we formulate the first two research questions addressed in this paper: 1. Can we construct estimating sequences which also consider information coming from the past iterates? 2. How does this impact the resulting optimization method?

In the sequel, we present our answers to the aforementioned questions[1]. The main contributions are summarized as follows:

- We introduce new estimating functions, whose values are dependent on the history of iterates[2].
- We revisit the lemmas and theorems derived in the context of the classical estimating sequences framework and introduce new approaches to establish our findings. We also highlight the intuition behind the selection of the estimating sequences and design of the corresponding methods.
- For black-box optimization, we introduce a novel type of heavy-ball momentum, and show how to couple it with the estimating sequences framework. Different from the framework presented in [56], wherein the heavy-ball momentum stabilizes the iterates, our newly introduced momentum stabilizes the estimating functions themselves.
- We introduce a new gradient-based algorithm which allows for embedding our newly introduced momentum term into the classical FGM. We also show how FGM can be derived by discarding the additional memory terms.
- We improve upon the existing convergence results for FGM. We establish the optimality of our proposed method, and prove that the bound on the number of iterations becomes $\sqrt{\frac{L}{2\mu}}\left(\ln\left(\frac{\mu R_0^2}{2\varepsilon}\right)+\ln(5)\right)$, where $R_0=||x_0-x^*||$, x_0 is the initial point, x^* is the optimal point, $\varepsilon \leq \frac{\mu}{2}R_0^2$, and $||\cdot||$ stands for the Euclidean norm of a vector. This results in an improvement over the bound for FGM by more than $1/\sqrt{2}$.

[1]Note that the detailed derivations used to establish the Lemmas and Theorems presented in the paper are provided as part of [51, 54, 55].

[2]The proposed framework allows for embedding any form of information that can accelerate the convergence of iterates.

- The newly introduced convergence results allow for setting $\gamma_0 = 0$, which is the initial value of γ_k, where γ_k is the radius of a ball around x_k, and it will be formally defined later. In [51], we also show numerically that this result alone enables a faster convergence than FGM. We note that such result is an extension of the existing analysis for FGM, wherein the convergence was established only for $\gamma_0 \in [\mu, 3L+\mu]$. Moreover, it enables the robustness of the initialization of our method to the inexact estimate of μ.

2.1. **Proposed method.** Let us begin by considering the following problem

$$\underset{x\in\mathscr{R}^n}{\text{minimize}}\, f(x), \tag{2.1}$$

where $f : \mathscr{R}^n \to \mathscr{R}$ has strong convexity parameter μ and Lipschitz continuous gradient L, defined by a deterministic black-box oracle.

First, let $\mathscr{I} = \{x \,|\, x_0 + \mathbf{span}\{\nabla f(x_0), \dots \nabla f(x_{k-1}), \cdots\}\} \subset \mathscr{R}^n$ for $k = 0, 1, 2, \dots, t$, where t is the current iteration and $\mathbf{span}\{\cdot\}$ is the linear space formed by all the gradients of the objective function at all iterations until the current iteration t. The process of designing the minimization sequence $\{x_k\}_k$ can be understood as a reduction of the search space over iterations until convergence to the optimal point. Next, we highlight the following definition.

Definition 2.1. The sequences $\{\Phi_k\}_k$ and $\{\lambda_k\}_k$, $\lambda_k \geq 0$, are called generalized estimating sequences of the function $f(x)$, if there exists a sequence of bounded for the corresponding values of $x \in \mathscr{I}, \forall k$ functions $\{\psi_k\}_k$, $\lambda_k \to 0$, and we have

$$\Phi_k(x) \leq \lambda_k \Phi_0(x) + (1-\lambda_k)\left(f(x) - \psi_k(x)\right). \tag{2.2}$$

Using $\psi_k(x)$ in (2.2) allows for including more information on the cost function that can enable faster convergence. Let us now show how to use the generalized estimating sequences to measure the rate of convergence for the iterates formed during the minimization process.

Lemma 2.1. *If for some sequence of points $\{x_k\}_k$ we have*

$$f(x_k) \leq \Phi_k^* \triangleq \min_{x\in\mathscr{I}} \Phi_k(x), \tag{2.3}$$

then

$$f(x_k) - f(x^*) \leq \lambda_k \left[\Phi_0(x^*) - f(x^*)\right] - (1-\lambda_k)\,\psi_k(x^*), \tag{2.4}$$

where $x^* = \arg\min_{x\in\mathscr{R}^n} f(x)$.

Let us next proceed to presenting our proposed definitions for the terms that comprise the generalized estimating sequences.

Lemma 2.2. *Assume that there exist sequences $\{\alpha_k\}_k$, where $\alpha_k \in (0,1)$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\{y_k\}_k$, where $y_k \in \mathscr{R}^n$, and a sequence of functions $\{\psi_k\}_k$, with a finite upper bound value Ψ over $x \in \mathscr{I}, \forall k$, such that $\psi_k(x) \geq 0$, $\forall k$. Let $\psi_0(x) = 0$ and $\lambda_0 = 1$. Then, the sequences $\{\Phi_k\}_k$ and $\{\lambda_k\}_k$, which are defined recursively as*

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k, \tag{2.5}$$

$$\begin{aligned}\Phi_{k+1}(x) = {} & (1-\alpha_k)(\Phi_k(x) + \psi_k(x)) - \psi_{k+1}(x) - \Psi + \alpha_k \psi_k(x) \\ & + \alpha_k \left(f(y_k) + \nabla f(y_k)^T (x - y_k) + \frac{\mu}{2}||x-y_k||^2\right),\end{aligned} \tag{2.6}$$

are generalized estimating sequences. Here, $(\cdot)^T$ denotes the transposition operation.

Recall that the structure for $\{\Phi_k(x)\}_k$ has not been presented yet. As discussed in [16], accelerated methods need to exploit some of the topological features of the cost function. Such observation can be validated based on existing results on second-order methods. Considering Newton's method, as shown in [1, Fig. 9.19], making use of the information available in the Hessian enables the construction of ellipsoids around each iterate. Such ellipsoids facilitate corrections of the selected descent direction. For gradient-based methods, which do not have access to Hessian-related information, we can devise balls around each x_k, without "discriminating" the different search directions. Mathematically, this is modeled by using isotropic functions, which scan the ball around each x_k with radius γ_k. The resulting Hessian then becomes $\nabla^2\phi_k(x) = \gamma_k I$. The estimating function is

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}||x - v_k||^2, \ \forall k, \tag{2.7}$$

and has minimum value ϕ_k^*, radius $\gamma_k \in \mathscr{R}^+$ and is centered around $v_k \in \mathscr{R}^n$. Similar structure as (2.7) is also used for constructing FGM [16]. Different from (2.7), we let

$$\Phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}||x - v_k||^2 - \psi_k(x), \ \forall k. \tag{2.8}$$

The added term $\psi_k(x)$ in (2.8) can be set as the following memory term, if no other additional information about the objective function is available,

$$\psi_k(x) \triangleq \sum_{i=0}^{k-1} \beta_{i,k} \frac{\gamma_i}{2}||x - v_i||^2, \ \forall k. \tag{2.9}$$

A simple example for $\beta_{i,k}$ is

$$\beta_{i,k} = \begin{cases} \min\left(1, \frac{\mu}{\gamma_{k-1}}\right), & \text{if } i = k-1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

In this way, by considering the black-box setting, wherein prior knowledge of the structure of the objective function is not available, we allow our newly introduced scanning functions to "self-regulate" and encompass information that was already available from the earlier iterates. Selecting $\beta_{i,k}$ according to (2.10) ensures that (2.9) remains finite since only the estimating function in iteration $k-1$ is used.

Let us now present the recursive relations for ϕ_k^*, γ_k, and v_k.

Lemma 2.3. *Assume that the coefficients $\beta_{i,k}$ are selected according to (2.10), and let $\Phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}||x - v_0||^2$. Then, the process defined in Lemma 2.2 preserves the quadratic canonical structure of the scanning function introduced in (2.7). Moreover, the sequences $\{\gamma_k\}_k$, $\{v_k\}_k$*

and $\{\phi_k^*\}_k$ *can be computed as*

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\left(\mu + \sum_{i=0}^{k-1}\beta_{i,k}\gamma_i\right), \tag{2.11}$$

$$v_{k+1} = \frac{1}{\gamma_{k+1}}\left((1-\alpha_k)\gamma_k v_k + \mu\alpha_k\left(y_k - \frac{1}{\mu}\nabla f(y_k) + \sum_{i=0}^{k-1}\frac{\beta_{i,k}\gamma_i}{\mu}v_i\right)\right), \tag{2.12}$$

$$\begin{aligned}\phi_{k+1}^* &= \alpha_k f(y_k) + (1-\alpha_k)\phi_k^* + \frac{\alpha_k\gamma_k(1-\alpha_k)(\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i)}{2\gamma_{k+1}}||y_k - v_k||^2 \\ &+ \frac{\alpha_k^3}{\gamma_{k+1}}\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i||v_i - y_k||\,||\nabla f(y_k)|| - \frac{\alpha_k^2||\nabla f(y_k)||^2}{2\gamma_{k+1}} \\ &+ \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left((v_k-y_k)^T\nabla f(y_k) + \sum_{i=0}^{k-1}\beta_{i,k}\gamma_i||y_k - v_i||||y_k - v_k||\right) \\ &+ (1-\alpha_k)\frac{\gamma_k}{2}||x_{\Phi_k}^* - v_k||^2 + \alpha_k\sum_{i=0}^{k-1}\frac{\beta_{i,k}\gamma_i}{2}||y_k - v_i||^2 \\ &+ \frac{(1-\alpha_k)\,\alpha_k^2}{\gamma_{k+1}}\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i(v_i-y_k)^T\nabla f(y_k) + \sum_{i=0}^{k-1}\beta_{i,k}\frac{\gamma_i}{2}||x_{\Phi_k}^* - v_i||^2.\end{aligned} \tag{2.13}$$

We will choose $\{x_k\}_k$, $\{y_k\}_k$ and $\{v_k\}_k$ to ensure that $f(x_k) \le \Phi_k^*$, $\forall k$. For iteration k, suppose that $\phi_k^* \ge f(x_k)$. At iteration $k+1$, by relaxing (2.13) and making some algebraic manipulations, we reach

$$\begin{aligned}\phi_{k+1}^* \ge f(y_k) &+ (1-\alpha_k)\nabla f(y_k)^T(x_k - y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}||\nabla f(y_k)||^2 \\ &+ \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}(v_k-y_k)^T\nabla f(y_k) + (1-\alpha_k)\frac{\alpha_k^2}{\gamma_{k+1}}\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i(v_i-y_k)^T\nabla f(y_k).\end{aligned} \tag{2.14}$$

The necessary conditions of Lemma 2.1 are fulfilled if $\phi_{k+1}^* \ge f(x_{k+1})$. Thus, we further relax the lower bound by making use of

$$f(y_k) - \frac{1}{2L}||\nabla f(y_k)||^2 \ge f(x_{k+1}). \tag{2.15}$$

To ensure that (2.15) is satisfied, it suffices to take a gradient step for y_k [16, Theorem 2.1.5]. This allows for computing α_k as

$$\alpha_k = \sqrt{\frac{\gamma_{k+1}}{L}}. \tag{2.16}$$

Considering (2.11), we can write

$$\alpha_k = \frac{\left(\mu + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i - \gamma_k\right)}{2L} + \frac{\sqrt{\left(\mu + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i - \gamma_k\right)^2 + 4L\gamma_k}}{2L}. \tag{2.17}$$

Substituting the expression for α_k presented in (2.17), we can revise (2.14) as

$$\phi^*_{k+1} \geq f(x_{k+1}) + (1-\alpha_k)\nabla f(y_k)^T\Big((x_k - y_k) + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + \frac{\alpha_k^2}{\gamma_{k+1}}\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i(v_i - y_k)\Big). \tag{2.18}$$

The terms of $\{y_k\}_k$ can be acquired from

$$x_k - y_k + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + \frac{\alpha_k^2}{\gamma_{k+1}}\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i(v_i - y_k) = 0.$$

This yields

$$y_k = \frac{\gamma_{k+1}x_k + \alpha_k\gamma_k v_k + \alpha_k^2\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i v_i}{\gamma_{k+1} + \alpha_k\gamma_k + \alpha_k^2\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i}. \tag{2.19}$$

The complete procedure is presented in Algorithm 1.

Algorithm 1 Proposed Method

1: **Input** $x_0 \in \mathscr{R}^n$, set $\gamma_0 \in [0,\mu)\cup[2\mu, 3L+\mu]$ and $v_0 = x_0$.
2: **while** stopping criterion is not meet **do**
3: $\quad \alpha_k \leftarrow \frac{(\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i-\gamma_k)+\sqrt{(\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i-\gamma_k)^2+4L\gamma_k}}{2L}$
4: $\quad \gamma_{k+1} \leftarrow (1-\alpha_k)\gamma_k + \alpha_k\left(\mu + \sum_{i=0}^{k-1}\beta_{i,k}\gamma_i\right)$
5: $\quad y_k \leftarrow \frac{\gamma_{k+1}x_k+\alpha_k\gamma_k v_k+\alpha_k^2\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i v_i}{\gamma_{k+1}+\alpha_k\gamma_k+\alpha_k^2\sum_{i=0}^{k-1}\beta_{i,k}\gamma_i}$
6: $\quad x_{k+1} \leftarrow y_k - \frac{1}{L}\nabla f(y_k)$
7: $\quad v_{k+1} \leftarrow \frac{1}{\gamma_{k+1}}\left((1-\alpha_k)\gamma_k v_k + \mu\alpha_k\left(y_k - \frac{1}{\mu}\nabla f(y_k) + \sum_{i=0}^{k-1}\frac{\beta_{i,k}\gamma_i}{\mu}v_i\right)\right)$
8: **end while**
9: **Output** x_{k+1}

Let us next compare Algorithm 1 with [16, (2.2.19)]. First, observe that the relations for computing α_k and γ_k are different due to the different estimating functions. A similar observation can be made by looking at the recurrent relation for computing y_k, $\forall k$. An important difference is the range of values for which γ_0 can be selected. The existing convergence results for FGM are limited to the range $\gamma_0 \in [\mu, 3L+\mu]$. Our algorithm converges for a larger range of γ_0. The extension of the convergence results to cover also the case where $\gamma_0 = 0$ enables robust initialization of our method that relaxes the need to know inexact/accurate estimate of μ. Note that estimating the exact/accurate value for μ would require additional computations. Moreover, the additional terms coming from using $\{\psi_k\}_k$ appear as multipliers of α_k^2. They are also present in the update of v_{k+1}. Last, observe that FGM can be derived by letting $\beta_{i,k} = 0$, $\forall i,k$.

2.2. **Bounds on convergence rate.** Let us now present the key convergence results for Algorithm 1. First, we show that the convergence of the iterates obtained during the minimization process is dependent on both $\{\lambda_k\}_k$ and $\{\psi_k\}_k$.

Theorem 2.1. *If we let $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$, Algorithm 1 generates a sequence of points $\{x_k\}_k$ such that*

$$f(x_k) - f^* \leq \lambda_k \left[f(x_0) - f(x^*) + \frac{\gamma_0}{2}||x_0 - x^*||^2 \right] - (1-\lambda_k)\psi_k(x^*). \tag{2.20}$$

From Lemma 2.2, we have that $\{\lambda_k\}_k \to 0$ when $k \to \infty$. The estimate of the rate of convergence for $\{\lambda_k\}_k$ is given in the following Lemma.

Lemma 2.4. *For all $k \geq 0$, Algorithm 1 guarantees that*

$$\lambda_k \leq \frac{2\mu}{L\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i}{L}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i}{L}}} \right)^2} \leq \frac{2\mu}{\left(\mu + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)(k+1)^2}. \tag{2.21}$$

Last, we demonstrate that Algorithm 1 is optimal.

Theorem 2.2. *In Algorithm 1, let $\mu > 0$. Then, the scheme generates a sequence of points such that*

$$f_k - f^* \leq \frac{\mu||x_0 - x^*||^2}{\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i}{L}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i}{L}}} \right)^2} - (1-\lambda_k)\psi_k(x^*). \tag{2.22}$$

where $f_k = f(x_k)$ and $f^* = f(x^*)$. This means that the method is optimal when the accuracy $\varepsilon \leq \frac{\mu}{2}||x_0 - x^*||^2$.

For the class of smooth strongly convex functions considered in this Section, FGM reaches the following bound on the number of iterations [16, (2.2.17)]

$$k_{FGM} \geq \sqrt{\frac{L}{\mu}} \left(\ln\left(\frac{\mu R_0^2}{2\varepsilon} \right) + \ln(23/3) \right). \tag{2.23}$$

On the other hand, if we select $\beta_{i,k}$ according to (2.10), the proposed method reaches the following bound on the number of iterations

$$k_{Proposed} \geq \sqrt{\frac{L}{\mu + \min(\gamma_{k-1}, \mu)}} \left(\ln\left(\frac{\mu R_0^2}{2\varepsilon} \right) + \ln(5) \right). \tag{2.24}$$

Observe that the bound presented in (2.24) is impacted by the rate of increase for the terms in $\{\gamma_k\}_k$. As we also demonstrate in [51], the terms in thesequence $\{\gamma_k\}_k$ grow exponentially in k, and converge to 2μ. Thus, the bound to the required number of iterations converges to

$$k_{Proposed} \to \sqrt{\frac{L}{2\mu}} \left(\ln\left(\frac{\mu R_0^2}{2\varepsilon} \right) + \ln(5) \right). \tag{2.25}$$

Comparing the convergence results presented in (2.25) to the existing lower bound for FGM given in (2.23), we highlight the improvement by at least a factor of $1/\sqrt{2}$. This improvement of the constant in the linear convergence rate is shown to be even more significant by experiments for solving practical problem with real-world data in [51].

3. Extending the Existing Estimating Sequence Framework to Composite Objectives

In this section, we focus on a broader class of convex problems, which are expressed as the sum of a smooth convex function together with a non-smooth convex function. In the sequel, we present the main findings of [54]. For this class of problems, several estimating sequences methods have been introduced in, e.g., [25, 49, 52]. Links between methods that were not originally devised by using the estimating sequences framework, such as FISTA, with estimating sequences methods have been also presented in [52]. Despite these methods being devised using different frameworks, they all share in common the accelerated rate of convergence. Nevertheless, when comparing their performance in solving practical problems with real-world data, we have observed that they exhibit different convergence properties. Moreover, comparing the original FGM with FISTA and AMGS for minimizing smooth convex functions, we have observed that FGM is more efficient. Thus, it becomes relevant to extend the estimating sequences framework used for devising FGM to the setup of composite objectives.

In the sequel, we introduce our proposed composite estimating sequences and show how to construct a composite objective estimating sequence technique that exhibits an accelerated rate of convergence. The main contributions are summarized as follows:

- We present new estimating sequences that are useful for devising numerical methods for minimizing the broader class of composite functions.
- We introduce new composite estimating functions, devised by coupling the gradient mapping framework introduced in [11] together with a tight bound on the composite cost function.
- Different from the functions used in [16], our proposed composite estimating functions exploit a tight bound on the composite cost function, together with its subgradients. This allows for developing accelerated gradient-based methods applicable to more general optimization problems.
- Based on the composite estimating sequences, we develop the Composite Objective Multi-step Estimating-sequence Techniques (COMET). The proposed scheme, is equipped with an efficient step-size adaption strategy. Different from AMGS, COMET requires only one projection-like operation per iteration.
- We prove that COMET exhibits an accelerate rate despite the inexact knowledge of the Lipschitz constant. Moreover, through computational experiments reported in [54], we highlight the robustness of COMET to the inexact knowledge of μ.

3.1. **Preliminaries.**

The problems of interest have the following structure

$$\underset{x\in\mathscr{R}^n}{\text{minimize}} \quad F(x)=f(x)+\tau g(x), \quad \tau>0, \tag{3.1}$$

Transferring the strong convexity parameter of $g(x)$ inside $F(x)$ yields

$$F(x)=\left(f(x)+\frac{\tau\mu_g}{2}||x-x_0||^2\right)+\tau\left(g(x)-\frac{\mu_g}{2}||x-x_0||^2\right)=\hat{f}(x)+\tau\hat{g}(x). \tag{3.2}$$

Considering the above-mentioned strong convexity transfer, we can write $L_{\hat{f}}=L_f+\tau\mu_g$ and $\mu_{\hat{f}}=\mu_f+\tau\mu_g$. Moreover, we observe that $\mu_{\hat{g}}=0$.

Recall that for $\hat{f}(x)$ we can write

$$\hat{f}(x) \leq \hat{f}(y) + \nabla \hat{f}(y)^T (x-y) + \frac{L_{\hat{f}}}{2}||y-x||^2, \tag{3.3}$$

$$\hat{f}(x) \geq \hat{f}(y) + \nabla \hat{f}(y)^T (x-y) + \frac{\mu_{\hat{f}}}{2}||y-x||^2. \tag{3.4}$$

In a similar manner, by definition of the subgradient of a function, we can write

$$\hat{g}(x) \geq \hat{g}(y) + s(y)^T (x-y), \tag{3.5}$$

where $s(y)$ denotes a subgradient of $\hat{g}(y)$. Furthermore, consider

$$m_L(y;x) \triangleq \hat{f}(y) + \nabla \hat{f}(y)^T (x-y) + \frac{L}{2}||x-y||^2 + \tau \hat{g}(x), \tag{3.6}$$

where $L \geq L_{\hat{f}}$. Substituting (3.3) in (3.6), yields

$$m_L(y;x) \geq F(x), \forall x, y \in \mathscr{R}^n. \tag{3.7}$$

Next, let us introduce the composite gradient mapping

$$T_L(y) \triangleq \arg\min_{x \in \mathscr{R}^n} m_L(y;x). \tag{3.8}$$

Morever, we define the composite reduced gradient

$$r_L(y) \triangleq L\left(y - T_L(y)\right). \tag{3.9}$$

Considering the special case $\tau = 0$, (3.2) results in $\hat{f}(x) = f(x)$. Observe that this would be the case wherein $m_L(y;x)$ would be differentiable in its variables. Applying the optimality condition for (3.8), we can write $\nabla m_L(y;x) = 0$. Replacing (3.6) in (3.8), and evaluating the first order condition, yields $T_L(y) = y - \frac{\nabla \hat{f}(y)}{L}$. Replacing such result in (3.9), we obtain $r_L(y) = \nabla F(y) = \nabla f(y)$. Considering the case when $\tau \neq 0$, based on the optimality criteria for (3.8), we can write

$$\begin{aligned} \partial m_L(y;T_L(y))^T (x - T_L(y)) \geq 0, \\ \left(\nabla \hat{f}(y) + L(T_L(y) - y) + \tau s_L(y)\right)^T (x - T_L(y)) \geq 0, \end{aligned} \tag{3.10}$$

where $\partial m_L(y;T_L(y))$ is the subdifferential of $m_L(y;T_L(y))$ and $s_L(y) \in \partial \hat{g}(T_L(y))$ is a subgradient of $T_L(y)$. Letting the first factor of (3.10) be equal to 0 and utilizing (3.9) results in

$$r_L(y) = L(y - T_L(y)) = \nabla \hat{f}(y) + \tau s_L(y). \tag{3.11}$$

The next theorem introduces a tighter lower bound than the one given in (3.4) for $F(x)$.

Theorem 3.1. *Let $F(x)$ be a composition of an $L_{\hat{f}}$-smooth and $\mu_{\hat{f}}$-strongly convex function $\hat{f}(x)$, and a simple convex function $\hat{g}(x)$, as given in (3.2). For $L \geq L_{\hat{f}}$, and $x, y \in \mathscr{R}^n$ we have*

$$F(x) \geq \hat{f}(T_L(y)) + \tau \hat{g}(T_L(y)) + r_L(y)^T (x-y) + \frac{\mu_{\hat{f}}}{2}||x-y||^2 + \frac{1}{2L}||r_L(y)||^2. \tag{3.12}$$

3.2. **Algorithm.** Similar to the previous Section, let us define the following.

Definition 3.1. The sequences $\{\phi_k\}_k$ and $\{\lambda_k\}_k$, $\lambda_k \geq 0$ are called composite estimating sequences of the function $F(\cdot)$ defined in (3.2), if $\lambda_k \to 0$ as $k \to \infty$, and $\forall x \in \mathscr{R}^n$, $\forall k \geq 0$, we have

$$\phi_k(x) \leq \lambda_k \phi_0(x) + (1-\lambda_k)F(x). \tag{3.13}$$

Observe that the proposed composite estimating sequences can estimate the rate of convergence of $\{x_k\}_k$. This is captured in the sequel.

Lemma 3.1. *If, for some sequence of points* $\{x_k\}_k$,

$$F(x_k) \leq \phi_k^* \triangleq \min_{x \in \mathscr{R}^n} \phi_k(x), \tag{3.14}$$

then

$$F(x_k) - F(x^*) \leq \lambda_k \left[\phi_0(x^*) - F(x^*)\right], \tag{3.15}$$

where $x^* = \arg\min_{x \in \mathscr{R}^n} F(x)$.

The terms comprising the composite estimating sequences are computed recursively as shown below.

Lemma 3.2. *Let* $\{\alpha_k\}_k$ *be a sequence such that* $\sum_{k=0}^{\infty} \alpha_k = \infty$ *with* $\alpha_k \in (0,1)$ $\forall k$, *and* $\{y_k\}_{k=0}^{\infty}$ *be an arbitrary sequence. Furthermore, let* $\lambda_0 = 1$ *and assume that the estimates* L_k, $\forall k$, *of the Lipschitz constant* $L_{\hat{f}}$ *are selected in a way that inequality* (3.3) *is satisfied for all the iterates* x_k *and* y_k. *Then, the sequences* $\{\phi_k\}_k$ *and* $\{\lambda_k\}_k$, *which are defined recursively as*

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k, \tag{3.16}$$

$$\begin{aligned}\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x) + \alpha_k F\left(T_{L_k}(y_k)\right) + \alpha_k \frac{1}{2L_k}||r_{L_k}(y_k)||^2 \\ + \alpha_k \left(r_{L_k}(y_k)^T (x-y_k) + \frac{\mu_f}{2}||x-y_k||^2\right),\end{aligned} \tag{3.17}$$

are composite estimating sequences.

Let us now compare our findings presented in Definition 3.1, Lemma 3.1 and Lemma 3.2 with the results obtained in [16, Definition 2.2.1, Lemma 2.2.1, Lemma 2.2.2]. If the objective function were differentiable, the proposed Definition 3.1 and Lemma 3.1 would reduce to the baseline results introduced for FGM, which are obtained under the assumption of smooth objective function. From this viewpoint, our proposed framework extends results introduced in [16] to a broader setup. Second, based on the results proved for Lemma 3.1, the rate of convergence of $\{x_k\}_k$ would be characterized by the rate at which $\lambda_k \to 0$. Third, (3.17) highlights the effect of using the tighter bound introduced in Theorem 3.1. Last, the objective function given in (3.17) is now computed based on the composite gradient mapping. Unlike the case of FGM, the proposed composite estimating functions exploit the subgradients of the non-smooth cost function to build $\{\phi_k\}_k$.

The terms of the sequence $\{\phi_k\}_k$ can be computed as follows

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}||x-v_k||^2, \ \forall k = 1,2,\ldots. \tag{3.18}$$

We highlight that there could be choices for $\phi_k(x)$, which can lead to different algorithms (see, e.g., [43, 44]). We can now proceed to presenting the recursive relations for the terms $\{\gamma_k\}_k$, $\{v_k\}_k$, and $\{\phi_k^*\}_k$.

Lemma 3.3. *Let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}||x - v_0||^2$, where $\gamma_0 \in \mathscr{R}^+$ and $v_0 \in \mathscr{R}^n$. Then, the process defined in Lemma 3.2 preserves the canonical form of the function presented in (3.18), where the sequences $\{\gamma_k\}_k$, $\{v_k\}_k$, and $\{\phi_k^*\}_k$ can be computed recursively as follows*

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu_{\hat{f}}, \tag{3.19}$$

$$v_{k+1} = \frac{1}{\gamma_{k+1}}\Big((1-\alpha_k)\gamma_k v_k + \alpha_k\left(\mu_{\hat{f}} y_k - L_k(y_k - T_{L_k}(y_k))\right)\Big), \tag{3.20}$$

$$\begin{aligned}\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k\left(F(T_{L_k}(y_k)) + \frac{1}{2L_k}||r_{L_k}(y_k)||^2\right) - \frac{L_k^2\alpha_k^2}{2\gamma_{k+1}}||y_k - T_{L_k}(y_k)||^2 \\ + \frac{\mu_{\hat{f}}\alpha_k\gamma_k(1-\alpha_k)}{2\gamma_{k+1}}||y_k - v_k||^2 + \frac{L_k\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}}(y_k - v_k)^T(y_k - T_{L_k}(x_k)).\end{aligned} \tag{3.21}$$

Different from the results given in [16], our proposed framework also allows for the line search adaptation[3]. To enable faster convergence to the optimal solution, it is desirable to choose the smallest value L_k for which the inequality (3.3) holds, wherein $L_{\hat{f}} = L_k$ is satisfied $\forall k = 0, 1, \ldots$. Then, it is desirable to control the increase of its value throughout the minimization process. Such approach would enforce the algorithm to perform "larger steps towards x^*" during the first iterations. In the later iterations, i.e., when x_k is closer to x^*, having large L_k would prevent the method from overshooting past x^*. Unfortunately, such approach relies on the assumption that $L_{\hat{f}}$ is perfectly known. This makes it unsuitable for practical setups. Instead, we choose a line search strategy which enables: 1) Robustness of the algorithm with respect to the selection of L_0, and 2) Dynamic changes of the values of L_k, $\forall k = 0, 1, \ldots$. Our proposed line search strategy utilizes $\eta_u > 1$, which increases the value of L_k and $\eta_d \in (0, 1)$, which decreases the value of L_k. As we have shown in [54], the impact of the additional backtracks is minimal. This has also been observed in [25, 52]. The proposed algorithm for solving problems with composite objectives is given in Algorithm 2.

In line 3 of Algorithm 2, we use $K_{\max}$ to denote the maximum number of iterations. Its value can be chosen to optimize the trade-off between the required accuracy and computations/processing time. Comparing Algorithm 2 to FGM (CSS I in [16]), we can observe that $\{\alpha_k\}_k$ and $\{\gamma_k\}_k$ share similar recursive structures. The update of y_k is different. In Algorithm 2, y_k is computed independently of the value of $\mu_{\hat{f}}$. The update rule for the iterates x_k is also different. Because of the structure of the cost function, the next iterate is obtained through a proximal gradient step. The assumption on the simplicity of the non-smooth term $g(x)$ ensures that the proximal term can be computed with complexity $\mathscr{O}(n)$ [57]. The update rule for v_k is also different. In Algorithm 2, we can observe the effect of using the composite reduced gradient.

3.3. **Bounds on the convergence rate.** Our proposed convergence analysis establishes the converge of Algorithm 2 for a wider selection of γ_0. This is different from the existing results presented in [16, Lemma 2.2.4], wherein convergence is established only for $\gamma_0 \in [\mu_{\hat{f}}; 3L_{\hat{f}} + \mu_{\hat{f}}]$.

[3]Many backtracking line search strategies have already been presented in the literature (see [49, 50]).

Algorithm 2 Proposed Method

Input $x_0 \in \mathscr{R}^n$, $L_0 > 0$, $\mu_{\hat{f}}$, $\gamma_0 \in [0, 3L_0 + \mu_{\hat{f}}]$,
$\eta_u > 1$ and $\eta_d \in (0,1)$.
Set $k = 0$, $i = 0$ and $v_0 = x_0$.
while $k \leq K_{\max}$ **do**
  $\hat{L}_i \leftarrow \eta_d L_k$
  while True **do**
    $\hat{\alpha}_i \leftarrow \frac{(\mu_{\hat{f}} - \gamma_k) + \sqrt{(\mu_{\hat{f}} - \gamma_k)^2 + 4\hat{L}_i \gamma_k}}{2\hat{L}_i}$
    $\hat{\gamma}_{i+1} \leftarrow (1 - \hat{\alpha}_i)\gamma_k + \hat{\alpha}_i \mu_{\hat{f}}$
    $\hat{y}_i \leftarrow \frac{\hat{\gamma}_{i+1} x_k + \hat{\alpha}_i \gamma_k v_k}{\hat{\gamma}_{i+1} + \hat{\alpha}_i \gamma_k}$
    $\hat{x}_{i+1} \leftarrow \text{prox}_{\frac{1}{\hat{L}_i}\hat{g}}\left(\hat{y}_i - \frac{1}{\hat{L}_i}\nabla f(\hat{y}_i)\right)$
    $\hat{v}_{i+1} \leftarrow \frac{1}{\hat{\gamma}_{i+1}}\left((1 - \hat{\alpha}_i)\gamma_k v_k + \hat{\alpha}_i\left(\mu_{\hat{f}}\hat{y}_i - \hat{L}_i(\hat{y}_i - \hat{x}_{i+1})\right)\right)$
    if $F(\hat{x}_{i+1}) \leq m_{\hat{L}_i}(\hat{y}_i, \hat{x}_{i+1})$ **then**
      Break from loop
    else
      $\hat{L}_{i+1} \leftarrow \eta_u \hat{L}_i$
    end if
    $i \leftarrow i + 1$
  end while
  $L_{k+1} \leftarrow \hat{L}_i$, $x_{k+1} \leftarrow \hat{x}_i$, $\alpha_k \leftarrow \hat{\alpha}_{i-1}$,
  $y_k \leftarrow \hat{y}_{i-1}$, $i \leftarrow 0$, $k \leftarrow k + 1$
end while
Output x_k

Choosing $\gamma_0 = 0$, also provides the robustness to initialization of Algorithm 2 with respect to the imperfect knowledge of $\mu_{\hat{f}}$.

First, we demonstrate that the convergence rate of the minimization process of Algorithm 2 is characterized by the rate that $\lambda_k \to 0$.

Theorem 3.2. *If we let* $\lambda_0 = 1$ *and* $\lambda_k = \prod_{i=0}^{k-1}(1 - \alpha_i)$, *Algorithm 2 generates a sequence of points* $\{x_k\}_{k=0}^{\infty}$ *such that*

$$F(x_k) - F(x^*) \leq \lambda_k \left[F(x_0) - F(x^*) + \frac{\gamma_0}{2}||x_0 - x^*||^2\right]. \tag{3.22}$$

Since $\lambda_k \to 0$, Theorem 3.2 suffices to conclude that the iterates generated by Algorithm 2 converge to x^*.

We now estimate the rate at which $\lambda_k \to 0$.

Lemma 3.4. *For all* $k \geq 0$, *Algorithm 2 guarantees that*

(1) If $\gamma_0 \in [0, \mu_{\hat{f}})$, *then*

$$\lambda_k \leq \frac{2\mu_{\hat{f}}}{L_k\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}}{L_k}}}\right)^2} \leq \frac{2}{(k+1)^2}. \tag{3.23}$$

(2) If $\gamma_0 \in [\mu_{\hat{f}}, 3L_0 + \mu_{\hat{f}}]$, *then*

$$\lambda_k \leq \frac{4\mu_{\hat{f}}}{(\gamma_0 - \mu_{\hat{f}})\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}}{L_k}}}\right)^2} \leq \frac{4L_k}{(\gamma_0 - \mu_{\hat{f}})(k+1)^2}. \tag{3.24}$$

Contrasting the results in Lemma 3.4 with their counterpart, i.e., [16, Lemma 2.2.4], we have the following main differences. First, we prove the convergence of the iterates also in the absence of the exact knowledge of the Lipschitz constant. Moreover, we prove the convergence of the minimization process for a wider range of γ_0. Such a finding is important for several reasons. First, Algorithm 2 enjoys a faster theoretical (and practical as shown in [54]) convergence rate when $\gamma_0 = 0$. Second, setting $\gamma_0 = 0$ provides robustness to the inexact knowledge of $\mu_{\hat{f}}$.

The following lemma yields an upper bound on the distance $F(x_0) - F(x^*)$.

Lemma 3.5. *Let* $F(x)$ *be a convex function with composite structure as shown in* (2.1). *Moreover, let* $T_L(y)$ *and* $r_L(y)$ *be computed as given in* (3.8) *and* (3.11), *respectively. Then, for any starting point* x_0 *in the domain of* $F(x)$, *we have*

$$F(x_0) - F(x^*) \leq \frac{L_0}{2}||x_0 - x^*||^2. \tag{3.25}$$

Combining Lemmas 3.4 and 3.5 with Theorem 3.2, yields the following convergence rate for Algorithm 2.

Theorem 3.3. *Algorithm 2 generates a sequence of points such that*

(1) If $\gamma_0 \in [0, \mu_{\hat{f}})$, *then*

$$F(x_k) - F(x^*) \leq \frac{\mu_{\hat{f}}(L_0 + \gamma_0)||x_0 - x^*||^2}{L_k\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}}{L_k}}}\right)^2}. \tag{3.26}$$

(2) If $\gamma_0 \in [\mu_{\hat{f}}, 3L_0 + \mu_{\hat{f}}]$, *then*

$$F(x_k) - F(x^*) \leq \frac{2\mu_{\hat{f}}(L_0 + \gamma_0)||x_0 - x^*||^2}{(\gamma_0 - \mu_{\hat{f}})\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu}{L_k}}}\right)^2}. \tag{3.27}$$

Based on Theorem 3.3, we can observe that Algorithm 2 converges over a wider interval than its counterpart devised for the class of smooth and strongly convex functions. Initializing $\gamma_0 = 0$ guarantees the fastest convergence of the method. Such a result is relevant in the context of practical deployments since $\mu_{\hat{f}}$ and $L_{\hat{f}}$ are not known in practice and their values need to

be estimated based on the available data, which is typically computationally expensive. The convergence rate of the iterates is also dependent on the value of L_0. Based on (3.26) and (3.27), we can see that choosing small values for L_0 enables faster convergence of Algorithm 2.

4. Generalizing the Estimating Sequences Framework for Problems with Composite Objectives

In this section, we further extend the results presented in Sections 2 and 3. To clarify the path of developments more, we have proposed estimating sequences constructions that extend in different directions. In Section 2, we proposed a new class of generalized estimating sequences that support the embedding of a heavy-ball type of momentum into the classical estimating sequences. Based on the framework introduced in Section 2, we established that it is possible to devise a scheme that enjoys a provably faster convergence rate than FGM. In Section 3, we proposed a new class of estimating sequences that can be used for solving optimization problems with composite objectives. Therein, we showed that our proposed black-box method also enjoys the same acceleration as the existing benchmarks among black-box methods, i.e., AMGS and FISTA, however it is more efficient in terms of required number of iterations than them. The remaining question of interest relates to exploring the coupling of the frameworks introduced in Sections 2 and 3.

In the sequel, we present the final class of estimating sequences that we devise here, which we name *generalized composite estimating sequences*, and show that they enable the construction of a class of very efficient accelerated algorithms. The main contributions are summarized as follows:

- We introduce a new structure for the estimating functions, which we call the generalized composite estimating functions. The proposed estimating functions are constructed by making use of the generalized estimating sequences, which contain a heavy-ball type of momentum embedded into them, together with the gradient mapping technique [11]. Similar to Section 3, we use a tighter global lower bound on the objective function than the one obtained from the Taylor series expansion of a convex function, and which it traditionally used.
- We use the proposed generalized composite estimating sequences to devise a new class of accelerated gradient methods, which are also equipped with an efficient backtracking line-search technique. Similarly to the algorithm introduced in Section 3, and different from AMGS, the method proposed in this section also requires one projection-like operation per iteration.
- Independently from the knowledge of the true value of the Lipschitz constant, we prove that our proposed method enjoys the accelerated convergence rate, which makes the method robust to initial iterations in practice.
- We also show that the initialization of our proposed method can be made robust to the imperfect knowledge of the strong convexity parameter. This reduces the computational burden of computing a tight estimate of the strong convexity parameter.

4.1. **Algorithm.**

Let us now present the last structure of estimating sequences that we report in this paper.

Definition 4.1. The sequences $\{\Phi_k\}_k$ and $\{\lambda_k\}_k$, $\lambda_k \geq 0$, are called generalized composite estimating sequences of the function $F(\cdot)$ defined in (3.2), if there exists a sequence of bounded for the corresponding values of $x \in \mathscr{I}, \forall k$ functions $\{\psi_k\}_k$, $\lambda_k \to 0$, and $\forall x \in \mathscr{I}$, $\forall k \geq 0$ we have

$$\Phi_k(x) \leq \lambda_k \Phi_0(x) + (1-\lambda_k)\left(F(x) - \psi_k(x)\right). \tag{4.1}$$

Similar to other estimating sequences, we can use the generalized composite estimating sequences to characterize the convergence rate of the minimization process.

Lemma 4.1. *If, for some sequence* $\{x_k\}_k$,

$$F(x_k) \leq \Phi_k^* \triangleq \min_{x \in \mathscr{Q}} \Phi_k(x), \tag{4.2}$$

then

$$F(x_k) - F(x^*) \leq \lambda_k \left[\Phi_0(x^*) - F(x^*)\right] - (1-\lambda_k)\psi_k(x^*), \tag{4.3}$$

where $x^* = \arg\min_{x \in \mathscr{Q}} F(x)$.

To construct an optimization scheme, we will need the following recurrent expressions of the estimating functions.

Lemma 4.2. *Let* $\{\alpha_k\}_k$ *with* $\alpha_k \in (0,1)$ $\forall k$ *be a sequence such that* $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\{\psi_k\}_k$ *with a finite upper bound value* Ψ *over* $x \in \mathscr{I}, \forall k$, *such that* $\{\psi_k\}_k \geq 0$. *Let also* $\{y_k\}_k$ *be an arbitrary sequence. Furthermore, let* $\psi_0(x) = 0$, $\lambda_0 = 1$ *and assume that the estimates* L_k, $\forall k = 0, 1, \ldots$, *of the Lipschitz constant* $L_{\hat{f}}$ *are selected so that inequality* (3.3) *is satisfied for all the iterates* x_k *and* y_k, $\forall k = 0, 1, \ldots$. *Then, the sequences* $\{\Phi_k\}_k$ *and* $\{\lambda_k\}_k$, *which are defined recursively as*

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k, \tag{4.4}$$

$$\begin{aligned}\Phi_{k+1}(x) = {} & (1-\alpha_k)\left(\Phi_k(x) + \psi_k(x)\right) - \psi_{k+1}(x) - \Psi + \alpha_k \left(\frac{\mu_{\hat{f}}}{2}||x - y_k||^2\right) \\ & + \alpha_k \left(F\left(T_{L_k}(y_k) + r_{L_k}(y_k)^T (x - y_k)\right) + \psi_k(x) + \frac{1}{2L_k}||r_{L_k}(y_k)||^2\right),\end{aligned} \tag{4.5}$$

are generalized composite estimating sequences.

Observe that the estimating sequences used for devising FGM in [16, Lemma 2.2.4] are obtained as the special case of our generalized composite estimating sequences when $\tau = 0$ and $\psi_k(x) = 0, \forall k = 0, 1, \ldots$. Similarly, the generalized estimating sequences devised in Section 2 is the special case of the proposed generalized composite estimating sequences obtained when $\tau = 0$. Last, the composite estimating sequences presented in Section 3 correspond to the special case obtained when $\{\psi_k\}_k = \{0\}_k$. In this sense, the generalized composite estimating sequences presented in this section encompass different variants of estimating sequences presented in the literature and earlier in this paper.

Considering $\gamma_k \in \mathscr{R}^+$, $v_k \in \mathscr{R}^n$, $\forall k = 0, 1, \ldots$, let us choose $\{\Phi_k\}_k$ as (2.7), $\{\psi_k(x)\}_k$ as (2.9), and $\beta_{i,k}$ as (2.10). Based on these selections, the minimal value of the estimating function introduced in (2.7) is

$$\Phi_k^* = \min_x \Phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}||x_{\Phi_k}^* - v_k||^2 - \sum_{i=1}^{k-1} \frac{\beta_{i,k}\gamma_i}{2}||x_{\Phi_k}^* - v_i||^2, \tag{4.6}$$

where $x_{\Phi_k}^* = \arg\min_x \Phi_k(x)$. The recursive relations for the parameters of $\{\phi_k\}_k$ are presented in the following Lemma.

Lemma 4.3. *Let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}||x - v_0||^2$, where $\gamma_0 \in \mathscr{R}^+$ and $v_0 \in \mathscr{R}^n$. Then, the process defined in Lemma 4.2 preserves the canonical form of the function $\{\Phi_k(x)\}_k$ presented in (2.7), where the sequences $\{\gamma_k\}_k$, $\{v_k\}_k$ and $\{\phi_k^*\}_k$ can be computed as follows*

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\left(\mu_{\hat{f}} + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right), \tag{4.7}$$

$$v_{k+1} = \frac{1}{\gamma_{k+1}}\left((1-\alpha_k)\gamma_k v_k + \alpha_k\left(\mu_{\hat{f}} y_k + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i - L(y_k - T_{L_k}(y_k))\right)\right), \tag{4.8}$$

$$\begin{aligned}\phi_{k+1}^* = {} & (1-\alpha_k)\phi_k^* + \alpha_k\left(F(T_{L_k}(y_k)) + \frac{1}{2L_k}||r_{L_k}(y_k)||^2 + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||y_k - v_i||^2\right)\\ & - \frac{L_k^2\alpha_k^2}{2\gamma_{k+1}}||y_k - T_{L_k}(y_k)||^2 + \frac{\alpha_k\gamma_k(1-\alpha_k)\left(\mu_{\hat{f}} + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{2\gamma_{k+1}}||y_k - v_k||^2\\ & + \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}||x_{\Phi_k}^* - v_k||^2 + \sum_{i=1}^{k}\frac{\beta_{i,k+1}\gamma_i}{2}||x_{\Phi_{k+1}}^* - v_i||^2\\ & + \frac{\alpha_k^2(1-\alpha_k)}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i(v_i - y_k)^T r_{L_k}(y_k) + \frac{\alpha_k^3}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||v_i - y_k||\,||r_{L_k}(y_k)||\\ & + \frac{\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}}\left((v_k - y_k)^T r_{L_k}(y_k) + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||y_k - v_i||\,||y_k - v_k||\right).\end{aligned} \tag{4.9}$$

Comparing the results presented in Lemma 4.3 to corresponding results in [16, Lemma 2.2.3], we can highlight that the recursive relations obtained for computing the terms $\{v_k\}_k$ and $\{\phi_k^*\}_k$ now reflect the usage of a new lower bound on the function that is being minimized. The impact of using the proposed reduced composite gradient is also visible. Moreover, observe that the computation of the terms $\{\gamma_k\}_k$, $\{v_k\}_k$, and $\{\phi_k^*\}_k$ highlight the presence of the heavy-ball type of momentum term that was used to construct them. Comparing the above obtained results to the the results reported in Section 2, we can observe the presence of the subgradients of the objective function together with the multistep nature of our newly obtained method. Last, different from the results highlighted in Section 3, we can observe the additional terms coming from the newly introduced heavy-ball type of momentum.

Similar to the previous sections, we will devise our proposed method by using an induction-based argument. Suppose that at step k we have

$$\Phi_k^* \overset{(4.6)}{=} \phi_k^* + \frac{\gamma_k}{2}||x_{\Phi_k}^* - v_k||^2 - \sum_{i=1}^{k-1}\frac{\beta_{i,k}\gamma_i}{2}||x_{\Phi_k}^* - v_i||^2 \geq F(x_k). \tag{4.10}$$

We need to prove that $\Phi^*_{k+1} \geq F(x_{k+1})$. Using (4.10) and (3.9) in (4.9), we obtain

$$\begin{aligned}
\phi^*_{k+1} \geq\ & (1-\alpha_k)F(x_k)+\alpha_k\left(F\left(T_{L_k}(y_k)\right)+\frac{1}{2L_k}||r_{L_k}(y_k)||^2+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||y_k-v_i||^2\right)\\
&-\frac{L_k^2\alpha_k^2}{2\gamma_{k+1}}||y_k-T_{L_k}(y_k)||^2+\frac{\alpha_k\gamma_k(1-\alpha_k)\left(\mu_{\hat{f}}+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{2\gamma_{k+1}}||y_k-v_k||^2\\
&+\frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}||x^*_{\Phi_k}-v_k||^2+\sum_{i=1}^{k}\frac{\beta_{i,k+1}\gamma_i}{2}||x^*_{\Phi_{k+1}}-v_k||^2\\
&+\frac{\alpha_k^2(1-\alpha_k)}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i(v_i-y_k)^T r_{L_k}(y_k)+\frac{\alpha_k^3}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||v_i-y_k||||r_{L_k}(y_k)||\\
&+\frac{\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}}\left((v_k-y_k)^T r_{L_k}(y_k)+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||y_k-v_i||||y_k-v_k||\right). \qquad (4.11)
\end{aligned}$$

Substituting (3.12) into (4.11), we have

$$\begin{aligned}
\phi^*_{k+1} \geq\ & (1-\alpha_k)\left(F(T_{L_k}(y_k))+r_{L_k}(y_k)^T(x_k-y_k)+\frac{\mu}{2}||x_k-y_k||^2+\frac{1}{2L_k}||r_{L_k}(y_k)||^2\right)\\
&+\alpha_k\left(F\left(T_{L_k}(y_k)\right)+\frac{1}{2L_k}||r_{L_k}(y_k)||^2+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||y_k-v_i||^2\right)\\
&-\frac{L_k^2\alpha_k^2}{2\gamma_{k+1}}||y_k-T_{L_k}(y_k)||^2+\frac{\alpha_k\gamma_k(1-\alpha_k)\left(\mu_{\hat{f}}+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{2\gamma_{k+1}}||y_k-v_k||^2\\
&+\frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}||x^*_{\Phi_k}-v_k||^2+\sum_{i=1}^{k}\frac{\beta_{i,k+1}\gamma_i}{2}||x^*_{\Phi_{k+1}}-v_k||^2\\
&+\frac{\alpha_k^2(1-\alpha_k)}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i(v_i-y_k)^T r_{L_k}(y_k)+\frac{\alpha_k^3}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||v_i-y_k||||r_{L_k}(y_k)||\\
&+\frac{\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}}\left((v_k-y_k)^T r_{L_k}(y_k)+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i||y_k-v_i||||y_k-v_k||\right). \qquad (4.12)
\end{aligned}$$

Making some manipulations in (4.12), we reach

$$\begin{aligned}
\phi^*_{k+1} \geq\ & F(T_{L_k}(y_k))+(1-\alpha_k)r_{L_k}(y_k)^T(x_k-y_k)+\sum_{i=1}^{k}\frac{\beta_{i,k+1}\gamma_i}{2}||x^*_{\Phi_{k+1}}-v_i||^2\\
&+\left(\frac{1}{2L_k}-\frac{\alpha_k^2}{2\gamma_{k+1}}\right)||r_{L_k}(y_k)||^2+\frac{\alpha_k^2(1-\alpha_k)}{\gamma_{k+1}}\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i(v_i-y_k)^T r_{L_k}(y_k)\\
&+\frac{\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}}(v_k-y_k)^T r_{L_k}(y_k). \qquad (4.13)
\end{aligned}$$

Next, we add $\frac{\gamma_{k+1}}{2}||x^*_{\Phi_{k+1}} - v_{k+1}||^2$ to the right-hand side of (4.13) and move $\sum_{i=1}^{k}\frac{\beta_{i,k+1}\gamma_i}{2}||x^*_{\Phi_{k+1}} - v_i||^2$ to the left-hand side of (4.13). This results in

$$\Phi^*_{k+1} \geq F(T_{L_k}(y_k)) + (1-\alpha_k) r_{L_k}(y_k)^T (x_k - y_k) + \left(\frac{1}{2L_k} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right) ||r_{L_k}(y_k)||^2 + \frac{\alpha_k^2(1-\alpha_k)}{\gamma_{k+1}} \sum_{i=1}^{k-1} \beta_{i,k}\gamma_i (v_i - y_k)^T r_{L_k}(y_k) + \frac{\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}} (v_k - y_k)^T r_{L_k}(y_k). \tag{4.14}$$

We can simplify (4.14) by letting

$$\alpha_k = \sqrt{\frac{\gamma_{k+1}}{L_k}}. \tag{4.15}$$

Plugging (4.7) into (4.15), yields the following recurrent relation for α_k

$$\alpha_k = \frac{\mu_{\hat{f}} + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i - \gamma_k + \sqrt{\left(\mu_{\hat{f}} + \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i - \gamma_k\right)^2 + 4L_k\gamma_k}}{2L_k}. \tag{4.16}$$

Thus, we can rewrite (4.14) as

$$\Phi^*_{k+1} \geq F(T_{L_k}(y_k)) + (1-\alpha_k) r_{L_k}(y_k)^T (x_k - y_k) + \frac{\alpha_k^2(1-\alpha_k)}{\gamma_{k+1}} \sum_{i=1}^{k-1} \beta_{i,k}\gamma_i (v_i - y_k)^T r_L(y_k) + \frac{\alpha_k\gamma_k(1-\alpha_k)}{\gamma_{k+1}} (v_k - y_k)^T r_L(y_k).$$

Letting

$$x_k - y_k + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + \frac{\alpha_k^2}{\gamma_{k+1}} \sum_{i=1}^{k-1} \beta_{i,k}\gamma_i (v_i - y_k) = 0, \tag{4.17}$$

results in the following recurrent relation for y_k:

$$y_k = \frac{\gamma_{k+1}x_k + \alpha_k\gamma_k v_k + \alpha_k^2 \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i v_i}{\gamma_{k+1} + \alpha_k\gamma_k + \alpha_k^2 \sum_{i=1}^{k-1}\beta_{i,k}\gamma_i}. \tag{4.18}$$

Finally, to establish $\Phi^*_{k+1} \geq F(x_{k+1})$, we can now simply let $x_{k+1} = T_{L_k}(y_k)$. Our proposed method is summarized in Algorithm 3.

Comparing Algorithm 3 to FGM, we highlight based on lines 6 and 7 in Algorithm 3, that the updates for α_k and γ_k are now computed differently from the corresponding updates in FGM. For Algorithm 3, their values exhibit dependency on the heavy-ball type of momentum term that is utilized in building the estimating sequences. The update for y_k is also computed differently. Furthermore, the value is not dependent on $\mu_{\hat{f}}$. Another significant difference is the update for x_k, which is now obtained through a proximal gradient step. The last difference can be observed from the update of v_k, whose value now reflects our selected subgradient. Further, comparing between Algorithm 3 and Algorithm 1, we highlight the differences coming due to the usage of the proposed subgradient of the objective function and due to the multistep structure of our proposed generalized composite estimating sequences. Last, comparing between Algorithm 3 and Algorithm 2, we can see that the biggest differences arise from the usage of the heavy-ball type of momentum term.

Algorithm 3 Proposed Method

Input $x_0 \in \mathscr{R}^n$, $L_0 > 0$, $\mu_{\hat{f}}$, $\gamma_0 \in [0, \mu_{\hat{f}}) \cup [2\mu_{\hat{f}}, 3L_0 + \mu_{\hat{f}}]$, $\eta_u > 1$ and $\eta_d \in]0,1[$.
Set $k = 0$, $i = 0$ and $v_0 = x_0$.
while $k \leq K_{\max}$ **do**
  $\hat{L}_i \leftarrow \eta_d L_k$
  while True **do**
    $\hat{\alpha}_i \leftarrow \frac{\mu_{\hat{f}} + \sum_{i=1}^{k-1} \beta_{i,k}\hat{\gamma}_i - \gamma_k + \sqrt{\left(\mu_{\hat{f}} + \sum_{i=1}^{k-1} \beta_{i,k}\hat{\gamma}_i - \gamma_k\right)^2 + 4\hat{L}_i\gamma_k}}{2\hat{L}_i}$
    $\hat{\gamma}_{i+1} \leftarrow (1-\hat{\alpha}_i)\gamma_k + \hat{\alpha}_i\left(\mu_{\hat{f}} + \sum_{i=1}^{k-1} \beta_{i,k}\hat{\gamma}_i\right)$
    $\hat{y}_i \leftarrow \frac{\hat{\gamma}_{i+1}x_k + \hat{\alpha}_i\gamma_k v_k + \hat{\alpha}_i^2 \sum_{i=1}^{k-1} \beta_{i,k}\hat{\gamma}_i v_i}{\hat{\gamma}_{i+1} + \hat{\alpha}_i\gamma_k + \hat{\alpha}_i^2 \sum_{i=1}^{k-1} \beta_{i,k}\hat{\gamma}_i}$
    $\hat{x}_{i+1} \leftarrow \text{prox}_{\frac{1}{\hat{L}_i}\hat{g}}\left(\hat{y}_i - \frac{1}{\hat{L}_i}\nabla f(\hat{y}_i)\right)$
    $\hat{v}_{i+1} \leftarrow \frac{1}{\hat{\gamma}_{i+1}}\left((1-\hat{\alpha}_i)\gamma_k v_k + \hat{\alpha}_i\left(\mu_{\hat{f}}\hat{y}_i + \sum_{i=1}^{k-1} \beta_{i,k}\hat{\gamma}_i - \hat{L}_i(\hat{y}_i - \hat{x}_{i+1})\right)\right)$
    if $F(\hat{x}_{i+1}) \leq m_{\hat{L}_i}(\hat{y}_i, \hat{x}_{i+1})$ **then**
      Break from loop
    else
      $\hat{L}_{i+1} \leftarrow \eta_u \hat{L}_i$
    end if
    $i \leftarrow i+1$
  end while
  $L_{k+1} \leftarrow \hat{L}_i$, $x_{k+1} \leftarrow \hat{x}_i$, $\alpha_k \leftarrow \hat{\alpha}_{i-1}$, $y_k \leftarrow \hat{y}_{i-1}$, $\gamma_{k+1} \leftarrow \hat{\gamma}_i$, $v_{k+1} \leftarrow \hat{v}_i$, $i \leftarrow 0$, $k \leftarrow k+1$
end while
Output x_k

Observe that the recursive relations obtained for our method presented in Algorithm 3 reduce to the ones obtained for FGM when $\tau = 0$ and $\psi_k(x) = 0, \forall k = 0, 1, \ldots$. Further, observe that Algorithm 3 reduces to Algorithm 1 when $\tau = 0$. Last, observe that Algorithm 3 reduces to Algorithm 2 when $\psi_k(x) = 0, \forall k = 0, 1, \ldots$. In this sense, Algorithm 3 is a generalization of all the aforementioned algorithms.

4.2. **Convergence Analysis.** Based on Lemma 4.1, we can deduce that the convergence rate of the minimization process is dependent on the sequences $\{\lambda_k\}_k$ and $\{\psi_k\}_k$. This is clarified in the following Theorem.

Theorem 4.1. *If we let* $\lambda_0 = 1$ *and* $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$, *Algorithm 3 generates a sequence of points* $\{x_k\}_k$ *such that*

$$F(x_k) - F(x^*) \leq \lambda_k\left(F(x_0) - F(x^*) + \frac{\gamma_0}{2}||x_0 - x^*||^2\right) - (1-\lambda_k)\psi_k(x). \tag{4.19}$$

The rate of convergence for $\{\lambda_k\}_k$ is characterized in the sequel.

Lemma 4.4. *For all* $k \geq 0$, *Algorithm 3 guarantees that*

(1) If $\gamma_0 \in [0, \mu_{\hat{f}})$, *then*

$$\lambda_k \leq \frac{2\mu_{\hat{f}}}{L_k\left(e^{\frac{k+1}{2}\sqrt{\frac{\left(\mu_{\hat{f}}+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\left(\mu_{\hat{f}}+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{L_k}}}\right)^2} \leq \frac{2}{(k+1)^2}. \tag{4.20}$$

(2) If $\gamma_0 \in [2\mu_{\hat{f}}, 3L_0 + \mu_{\hat{f}}]$, *then*

$$\begin{aligned}\lambda_k &\leq \frac{4\mu_{\hat{f}}}{(\gamma_0-\mu_{\hat{f}})\left(e^{\frac{k+1}{2}\sqrt{\frac{\left(\mu_{\hat{f}}+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\left(\mu_{\hat{f}}+\sum_{i=1}^{k-1}\beta_{i,k}\gamma_i\right)}{L_k}}}\right)^2} \\ &\leq \frac{4L_k}{(\gamma_0-\mu_{\hat{f}})(k+1)^2}.\end{aligned} \tag{4.21}$$

Comparing to [16, Lemma 2.2.4], the results in Lemma 4.4 highlight the following benefits: 1. Algorithm 3 converges also when the exact value of $L_{\hat{f}}$ is not known; 2. Algorithm 3 converges for a wider range of γ_0. Such finding is important as it suggests that the initialization of Algorithm 3 is robust to the inexact knowledge of $\mu_{\hat{f}}$.

Moreover, comparing the results of Lemma 4.4 to that of Lemma 3.4, we can also observe the structural similarity, which is the consequence of the fact that the same estimating function was used for constructing the composite estimating sequences in Section 3 and the generalized composite estimating sequences extended by the term $\psi_k(x)$ (see (2.2)) in this section. The fundamental difference of using the generalized composite estimating sequences for constructing the accelerated optimization scheme here appears in the power of exponents in the denominators, specifically, the addition of $\sum_{j=1}^{k-1}\beta_{j,k}\gamma_j$ in the results of Lemma 4.4 as compared to that of in Lemma 3.4. This additional term $\sum_{j=1}^{k-1}\beta_{j,k}\gamma_j$ defines how much improvement to the constant of the convergence rate we have for Algorithm 3 compared to Algorithm 2 when the generalized composite estimating sequences are extended by the term $\psi_k(x)$.

To establish the accelerated convergence rate of Algorithm 3, it suffices to combine Lemma 4.4 and Theorem 3.1 with Theorem 4.1 to come to the following finalizing theorem.

Theorem 4.2. *Algorithm 3 generates a sequence of points such that*

(1) If $\gamma_0 \in [0, \mu_{\hat{f}})$, *then*

$$F(x_k) - F(x^*) \leq \frac{\mu_{\hat{f}}(L_0+\gamma_0)||x_0-x^*||^2}{L_k\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}+\sum_{j=1}^{k-1}\beta_{j,k}\gamma_j}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}+\sum_{j=1}^{k-1}\beta_{j,k}\gamma_j}{L_k}}}\right)^2}. \tag{4.22}$$

(2) If $\gamma_0 \in [2\mu_{\hat{f}}, 3L_0 + \mu_{\hat{f}}]$, *then*

$$F(x_k) - F(x^*) \leq \frac{2\mu_{\hat{f}}(L_0+\gamma_0)||x_0 - x^*||^2}{(\gamma_0 - \mu_{\hat{f}})\left(e^{\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}+\Sigma_{j=1}^{k-1}\beta_{j,k}\gamma_j}{L_k}}} - e^{-\frac{k+1}{2}\sqrt{\frac{\mu_{\hat{f}}+\Sigma_{j=1}^{k-1}\beta_{j,k}\gamma_j}{L_k}}}\right)^2}. \tag{4.23}$$

Note that we provide here a measure of the convergence rate for Algorithm 3 in the challenging framework of the unknown Lipschitz constant with an account of the memory of the algorithm through the memory term in our generalized estimating sequences. The above results, however, are applicable also to the case when the Lipschitz constant is known/estimated and fixed. In this sense, our convergence results generalize the existing convergence results typically derived for known Lipschitz constant. With backtracking for L_k, the slope of the convergence curve is expected to be steeper than for fixed L, especially if L is overestimated, because backtracking helps to improve the condition number estimate at each iteration. This is in line with the other first-order algorithms extended with the backtracking procedure such as, for example, Algorithm 20 in [4] (see Corollary 4.23 there) and Algorithm 2 in [25].

5. Real-data Based Example

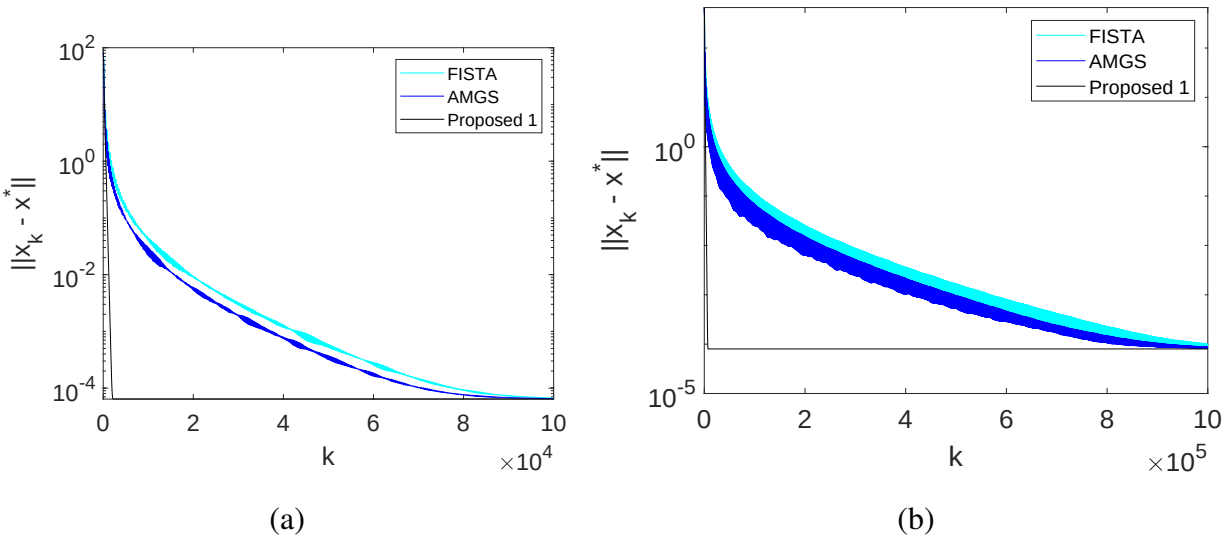
The detailed numerical studies can be found in out papers [51, 54, 55]. Here we demonstrate just one example of applying the resulting most general Algorithm 3 to real-data processing problem. Specifically, we test the performance of Algorithm 3 in minimizing the following regularized logistic regression problem

$$\underset{x\in\mathscr{R}^n}{\text{minimize}} \quad \frac{1}{m}\sum_{i=1}^{m}\log\left(1-\mathrm{e}^{-b_i x^T a_i}\right) + \frac{\tau_1}{2}\|x\|^2 + \tau_2\|x\|_1, \tag{5.1}$$

where $\|\cdot\|_1$ denotes the l_1-norm of a vector argument, and τ_1 and τ_2 are some weights assign to the corresponding regularization terms. We consider datasets namely "rcv1.binary", for which $A_{\text{"rcv1.binary"}} = [a_1, a_2, \cdots, a_i, \cdots] \in \mathscr{R}^{1000\times 2000}$ [58].

We find x^* by using CVX [59]. We choose the terms $\beta_{j,k} = \min\left(1, \frac{\mu}{\gamma_{k-1}}\right)$, for $j = k-1$ for Algorithm 3. Depending on the selection of the terms γ_0, we consider Algorithm 3 and set $\gamma_0 = 0$.We compare the performance of Algorithm 3 (called as "Proposed") against those of AMGS and FISTA. To estimate the value of the Lipschitz constant for AMGS and FISTA, we make use of the line-search strategies introduced in the corresponding papers [49, 50]. We select the point x_0 at random and use it as a starting point for all the algorithms that are compared. Moreover, the convergence of FISTA is significantly affected by the selection of L_0, which happens because the line-search strategy proposed for FISTA does not allow for decreasing the estimate of the Lipschitz constant. Since in this paper the goal is to devise more efficient black-box algorithms, we assume that the true value of $L_{\hat{f}}$ is known. For the dataset "rcv1.binary", we have $L_{\text{"rcv1.binary"}} = 1.13$. Regarding the strong convexity parameter, we select its value to be the same as the l_2 regularizer term in (5.1), which are selected to be $\tau_1 = \tau_2 \in \{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$.

Our findings are depicted in Fig. 1. From the figure, we can clearly see that Algorithm 3 significantly outperforms the selected benchmarks in minimizing the regularized logistic loss function.



(a) (b)

FIGURE 1. Performance evaluation of our proposed method and the selected benchmarks on real data. We consider the logistic objective function and elastic net regularizer. (a) Evaluating the distance to x^* for "rcv1.binary" dataset, $\tau_1 = \tau_2 = 10^{-4}$. (b) Evaluating the distance to x^* for "rcv1.binary" dataset, $\tau_1 = \tau_2 = 10^{-5}$.

6. CONCLUSION AND DISCUSSION

We have presented several accelerated first-order estimating sequence schemes that can be used for minimizing different classes of convex functions. In Section 2, we have established the generalized estimating sequences framework, and have shown that it enables to develop the algorithm with provably faster convergence rate than FGM. In Section 3, we have presented the class of composite estimating sequences and have shown that they can be used to devise efficient accelerated methods for minimizing convex function with composite structure. Then, in Section 4, we have introduced the generalized composite estimating sequences, which encompass all the previously introduced classes of estimating sequences. These estimating sequences have also been used to define an accelerated gradient-based method, which is more efficient than the existing benchmarks. Based on our convergence results, we can also conclude that for non-strongly convex problems our proposed methods retains the $\mathcal{O}(1/k^2)$ convergence rate. However, for arbitrarily small values of the strong convexity parameter, our proposed methods exhibit an accelerated linear convergence rate. Moreover, different from classical FGM-type of methods, the initialization of our proposed methods can be made robust to the imperfect knowledge of the strong convexity parameter. Moreover, for the methods presented in Sections 3 and 4, we have also introduced an efficient backtracking line search strategy.

Finally, we introduce several open problems that arise based on our newly introduced framework.

- Several open questions relate to the selection of the structure for the terms $\{\psi_k(x)\}_k$ and the choice of the coefficients $\beta_{i,k}$. Obtaining more efficient constructions for these terms can be used to devise more efficient first-order methods. It would also be of interest to evaluate their impact in designing methods that are optimal in decreasing the

norm of the gradient for the case of smooth objective functions. Devising such methods is particularly relevant in the context of nonconvex optimization [47, 60, 61], which aim to find stationary points of the objective function.

- Finding alternative constructions for $\psi_k(x)$ would also be of interest, both in the context of black-box optimization and beyond. A related concept is introduced in [62], wherein the authors develop the notions of relative smoothness and relative strong convexity. Considering twice differentiable functions, the relative smoothness and strong convexity parameters are influenced by the weighted difference of the Hessians of the objective function with a differentiable and convex reference function [62, Proposition 1.1]. Here we used a similar approach in establishing our estimating functions, with the main difference being that our proposed construction for $\psi_k(x)$ is dynamically changing over iterations. From the perspective of the framework introduced in [62], our selection of the coefficients $\beta_{i,k}$ suggests that the relative strong convexity parameter between $f(x)$ and $\psi_k(x)$ is not unique. As a matter of fact, it is contained in an interval which diminishes as the value of k increases, and as $k \to \infty$, it is restrained in $[0, 1]$. Thus, it is desirable to assess the co-existence aspects of these frameworks.
- In practice, the performance of FGM-type methods can be improved by restarting them. Several restarting conditions have been presented in the literature [63, 64]. It is of interest to assess if similar conditions can be devised for our proposed algorithm and measure the improvements in their performance. Here, we avoided any heuristic approaches such as restarting for further improving the efficiency of our proposed algorithms. Nevertheless, it would be beneficial to devise restarting conditions applicable to our proposed methods.
- It would be also relevant to extend our proposed framework to broader optimization setups, such as nonconvex, stochastic and distributed optimization. Several extensions of the estimating sequences framework used for devising FGM have already been presented in the literature. Considering the gains observed for the foundational setups, it would be of interest to extend our proposed estimating sequences to such optimization setups.

REFERENCES

[1] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[2] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.
[3] K. R. Kesari and J. Honorio, First order methods take exponential time to converge to global minimizers of non-convex functions, In: Proc. IEEE International Symposium on Information Theory, pp. 2322-2327, Melbourne, 2021.
[4] A. d'Aspremont, D. Scieur, and A. Taylor, Acceleration Methods. Foundations and Trends in Optimization, Now Publisher, 2021.
[5] A. Beck, First-order Methods in Optimization, vol. 25, SIAM, 2017.
[6] M. S. Ibrahim, A. Konar and N. D. Sidiropoulos, Fast algorithms for joint multicast beamforming and antenna selection in massive MIMO, IEEE Trans. Signal Process. 68 (2020), 1897–1909
[7] R. Gu and A. Dogandžić, Projected Nesterov's proximal-gradient algorithm for sparse signal recovery, IEEE Trans. Signal Process. 65 (2017), 3510–3525.
[8] M. Raginsky and A. Rakhlin, Information-based complexity, feedback and dynamics in convex programming, IEEE Trans. Info. Theory, 57 (2011), 7036-7056.

[9] A. Agarwal, P. L. Bartlett, P. Ravikumar and M. J. Wainwright, Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization, IEEE Trans. Info. Theory, 58 (2012), 3235-3249.
[10] Y. Nesterov, Subgradient methods for huge-scale optimization problems, Math. Program. 146 (2014), 275–297.
[11] A. Nemirovsky and D. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley, 1983.
[12] A. Pensia, V. Jong and P. L. Loh, Generalization error bounds for noisy, iterative algorithms, In: Proc. IEEE Inter. Sympos. Information Theory, pp. 546-550 Colorado, 2018.
[13] Y. Nesterov, A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, Doklady AN USSR, 269 (1983), 543–547.
[14] Y. Nesterov, On an approach to the construction of optimal methods of minimization of smooth convex functions, Ekonomika i Mateaticheskie Metody, 24 (1988), 509–517.
[15] M. Baes, Estimate sequence methods: Extensions and approximations, Institute for Operations Research, ETH, Zürich, 2020.
[16] Y. Nesterov, Lectures on Convex Optimization, vol. 137, Springer, 2018.
[17] Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program. 103 (2005), 127–152.
[18] A. Auslender and M. Teboulle, Interior Gradient and Proximal Methods for Convex and Conic Optimization, SIAM J. Optim. 16 (2006), 697–725.
[19] A. d'Aspremont, Smooth optimization with approximate gradient, SIAM J. Optim. 19 (2008), 1171–1183.
[20] O. Devolder, F. Glineur and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, Math. Program. 146 (2014), 37–75.
[21] M. Schmidt, N. L. Roux and F. R. Bach, Convergence rates of inexact proximal-gradient methods for convex optimization, In: Proc. 25th Annual Conference on Neural Information Processing Systems, pp. 1458–1466, Granada, 2011.
[22] N. Flammarion and F. Bach, From averaging to acceleration, there is only a step-size, In: Proc. Conference on Learning Theory, pp. 658–695, Paris, 2015.
[23] W. Su, S. Boyd and E. J. Candès, A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights, J. Mach. Learn. Res. 17 (2016), 1–43.
[24] A. Wibisono, A. C. Wilson and M. I. Jordan, A variational perspective on accelerated methods in optimization, Proc. Nat. Acad. Sci. 113 (2016), E7351–E7358.
[25] M. I. Florea and S. A. Vorobyov, An accelerated composite gradient method for large-scale composite objective problems, IEEE Trans. Signal Process. 67 (2019), 444-–459.
[26] Z. Allen-Zhu and L. Orecchia, Linear coupling: An ultimate unification of gradient and mirror descent, arXiv: 1407.1537, 2016.
[27] S. Bubeck, Y. T. Lee and M. Singh, A geometric alternative to Nesterov's accelerated gradient descent, arXiv: 1506.08187, 2015.
[28] D. Drusvyatskiy, M. Fazel and S. Roy, An optimal first order method based on optimal quadratic averaging, SIAM J. Optim. 28 (2018), 251–271.
[29] L. Lessard, B. Recht and A. Packard, Analysis and design of optimization algorithms via integral quadratic constraints, SIAM J. Optim. 26 (2016), 57–95.
[30] B. Van Scoy, R. A. Freeman and K. M. Lynch, The fastest known globally convergent first-order method for minimizing strongly convex functions, IEEE Control Sys. Lett. 2 (2018), 49-54.
[31] A. Taylor and Y. Drori, An optimal gradient method for smooth strongly convex minimization, Math. Program. 199 (2023), 557–594.
[32] Y. Drori and M. Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach, Math. Program. 145 (2014), 451–482.
[33] A. B. Taylor, J. M. Hendrickx and F. Glineur, Smooth strongly convex interpolation and exact worst-case performance of first-order methods, Math. Program. 161 (2014), 307—345.
[34] D. Kim and J. A. Fessler, Optimized first-order methods for smooth convex minimization, Math. Program. 159 (2016), 81–107.
[35] Y. Drori, The exact information-based complexity of smooth convex minimization, J. Complexity, 39 (2016), 1—16.

[36] S. Ji and J. Ye, An accelerated gradient method for trace norm minimization, In: Proc of the 26th Annual International Conference on Machine Learning, pp. 457–464, Montreal, 2009.
[37] C. A. Uribe, S. Lee, A. Gasnikov and A. Nedić, A dual approach for optimal algorithms in distributed optimization over networks, In: Proc. Information Theory and Applications Workshop, pp. 1–37, California, 2020.
[38] H. Ye, L. Luo, Z. Zhou and T. Zhang, Multi-consensus decentralized accelerated gradient descent, J. Mach. Learn. Res. 24 (2023), 1–50.
[39] C. Lin, V. Kostina and B. Hassibi, Differentially quantized gradient methods, EEE Trans. Info. Theory, 68 (2022), 6078–6097.
[40] C. Hu, W. Pan and J. Kwok, Accelerated gradient methods for stochastic optimization and online learning, In: Proc. 22nd Annual Conference on Neural Information Processing Systems, pp. 781–789, Vancouver, 2009.
[41] G. Lan, An optimal method for stochastic composite optimization, Math. Program. 133 (2016), 365–397.
[42] A. Kulunchakov and J. Mairal, Estimate Sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise, J. Mach. Learn. Res. 21 (2020), 1–52.
[43] H. Zhang and S. Sra, An estimate sequence for geodesically convex optimization, In: Proc. Conference on Learning Theory, pp. 1703–1723, Stockholm, 2018.
[44] K. Ahn and S. Sra, From Nesterov's estimate sequence to Riemannian acceleration, In: Proc. Conference on Learning Theory, pp. 88–118, Graz, 2020.
[45] Y. Nesterov, Accelerating the cubic regularization of Newton's method on convex problems, Math. Program. 112 (2008), 159–181.
[46] Y. Nesterov, Inexact high-order proximal-point methods with auxiliary search procedure, SIAM J. Optim. 31 (2021), 2807–2828.
[47] S. Ghadimi and G. Lan, Accelerated gradient methods for nonconvex nonlinear and stochastic programming, Math. Program. 156 (2016), 59–99.
[48] Y. Carmon, J. C. Duchi, O. Hinder and A. Sidford, Accelerated methods for nonconvex optimization, SIAM J. Optim. 28 (2018), 1751–1772.
[49] Y. Nesterov, Gradient methods for minimizing composite objective function, Math. Program. 140 (2013), 125–161.
[50] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009), 183–202.
[51] E. Dosti, S. A. Vorobyov, and T. Charalambous, A new class of composite objective multi-step estimating sequence techniques, Signal Process. 206 (2023), 108889.
[52] M. I. Florea and S. A. Vorobyov, A generalized accelerated composite gradient method: Uniting Nesterov's fast gradient method and FISTA, IEEE Trans. Signal Process. 68 (2020), 3033–3048.
[53] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, submitted to SIAM Journal on Optimization, 2008.
[54] E. Dosti, S. A. Vorobyov, and T. Charalambous, Embedding a heavy-ball type of momentum into the estimating sequences, Signal Process. 233 (2025), 109865.
[55] E. Dosti, S. A. Vorobyov, and T. Charalambous, Estimating sequences with memory for minimizing convex non-smooth composite functions, arxiv.org/pdf/2510.02965, 2025.
[56] B. T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Comput. Math. Math. Phys. 4 (1964), 1–17.
[57] N. Parikh, S. Boyd, Proximal Algorithms, Foundations and Trends in Optimization 1 (2014), 127–239.
[58] C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2 (2011), 1–27.
[59] M. Grant, S. Boyd and Y. Ye, CVX: Matlab software for disciplined convex programming (web page and software), 2009.
[60] Y. Carmon, J. C. Duchi, O. Hinder and A. Sidford, Lower bounds for finding stationary points II: First-order methods, Math. Program. 185 (2021), 315–355.
[61] J. Liang and R. D. C. Monteiro, An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems, SIAM J. Optim. 31 (2021), 217–243.

[62] H. Lu, R. M. Freund and Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, SIAM J. Optim. 28 (2021), 333–354.
[63] B. O'Donoghue and E. J. Candès, Adaptive restart for accelerated gradient schemes, Found. Comput. Math. 15 (2015), 715–732.
[64] P. Giselsson and S. Boyd, Monotonicity and restart in fast gradient methods, In: Proc. 53rd IEEE Conference on Decision and Control, pp. 5058—5063, Los Anngeles, 2015.