New Analysis of Linear Convergence of Gradient-type Methods via Unifying Error Bound Conditions

Hui Zhang*

June 23, 2017

Abstract

This paper reveals that a common and central role, played in many error bound (EB) conditions and a variety of gradient-type methods, is a residual measure operator. On one hand, by linking this operator with other optimality measures, we define a group of abstract EB conditions, and then analyze the interplay between them; On the other hand, by using this operator as an ascent direction, we propose an abstract gradient-type method, and then figure out EB conditions that are necessary and sufficient for its linear convergence. Both of these two points of view are refreshing and useful. The former provides a unified framework that not only allows us to find new connections between many existing EB conditions, but also paves a way to construct new EB conditions. The latter allows us to claim the weakest conditions guaranteeing linear convergence for a number of fundamental algorithms, including the gradient method, the proximal point algorithm, and the forward-backward splitting algorithm. In addition, we show linear convergence for the proximal alternating linearized minimization algorithm under a group of equivalent EB conditions, which are strictly weaker than the traditional strongly convex condition. Moreover, by defining a new EB condition, we show Q-linear convergence of the Nesterov's accelerated forward-backward algorithm without strong convexity. Finally, we verify EB conditions for a class of dual objective functions.

Keywords. residual measure operator, gradient descent, linear convergence, error bound condition, proximal point algorithm, forward-backward splitting algorithm, proximal alternating linearized minimization, Nesterov's acceleration, dual objective function

AMS subject classifications. 90C25, 90C60, 65K10, 49M29

1 Introduction

It is well-known that the standard assumption for proving linear convergence of gradient-type methods is strong convexity [40]. In practice, however, strong convexity is too stringent. Moreover, various gradient-type methods for solving convex optimization problems have exhibited linear convergence in numerical experiments even when strong convexity is absent; see for example [23, 30, 56]. Thereby, one would wonder whether such a phenomenon can be explained theoretically, and whether there exist weaker alternatives to strong convexity that retain fast rates.

*College of Science, National University of Defense Technology, Changsha, Hunan, 410073, P.R.China. Corresponding author. Email: h.zhang1984@163.com

A very powerful idea to address these questions is to connect error bound (EB) conditions with the convergence rate estimation of gradient-type methods. This idea has a long history dating back to 1963, Polyak introduced an EB inequality as a sufficient condition for gradient descent to attain linear convergence [42]. In the same year, a wide class of inequalities, which include Polyak's as a special case, were introduced by Łojasiewicz [35]. In the recent manuscript [26], the EB condition of Polyak-Łojasiewicz's type was further developed for linear convergence of gradient and proximal gradient methods. The second type of EB conditions is due to Hoffman, who proposed an EB inequality for systems of linear inequalities [24] in 1952. Along this line, Luo and Tseng in the early 90's contributed several aspects for connecting EB conditions of Hoffman's type with convergence analysis of descent methods [37]. Recently, global versions of EB conditions of Hoffman's type attracts a lot of attentions [49, 54, 65]. The third type of EB conditions is the quadratic growth condition (also called zero-order EB condition in [10]), which might go back to the work [66]. It was recently rediscovered in the special case of convex functions, and widely used to derive linear convergence for many gradient-type methods as well [34, 22, 39]. In particular, after this work was submitted to review, we obtained a group of linear convergence results for the proximal incremental aggregated gradient method under the quadratic growth condition [58].

Moreover, there recently emerges a surge of interests in developing new EB conditions guaranteeing (global) linear convergence for various gradient-type methods. For example, the authors of [30, 64, 61] proposed a restricted secant inequality (RSI), and developed the restricted strongly convex (RSC) property for analyzing linear convergence of (dual) gradient descent methods and Nesterov's restart accelerated methods; the authors of [39] proposed several relaxations of strong convexity that are sufficient for obtaining linear convergence for (projected and accelerated) gradient-type methods.

Another line of recent works is to find connections between existing EB conditions. For example, the authors of [19] discussed the relationship between the quadratic growth condition and the EB condition of Hoffman's type (also called Luo-Tseng's type in [31]); Parallel to and partially influenced by the work [19], the author of [60] established several new types of equivalence between the RSC property, the quadratic growth condition, and the EB condition of Hoffman's type; the authors of [10] showed the equivalence between the zero-order EB condition and the Kurdyka-Łojasiewicz inequality. We note that works [39] and [26] also discussed the relationships among many of these EB conditions.

Based on these two lines of recent developments, two natural questions arise. The first one is whether there is a unified framework for defining different EB conditions and analyzing the connection between them. The second one is whether these sufficient conditions guaranteeing linear convergence for gradient-type methods are also necessary. To answer these two questions, we will rely on a vital observation: a common and key role, played in many EB conditions and a variety of gradient-type methods, is a residual measure operator. This observation immediately leads us to the following discoveries:

1. By linking the residual measure operator with other optimality measures, we define a group of abstract EB conditions. Then, we comprehensively analyze the interplay between them by means of technique developed in [10]. The definition of abstract EB conditions not only unifies many existing EB conditions, but also helps us to construct new ones. The interplay between the abstract EB conditions allows us find new connections between many existing EB conditions.

2. By viewing the residual measure operator as an ascent direction, we propose an abstract

gradient-type method, and then figure out EB conditions that are necessary and sufficient for its linear convergence. The latter allows us to claim the weakest (or say, necessary and sufficient) conditions guaranteeing linear convergence for a number of fundamental algorithms, including the gradient method (applied to possibly nonconvex optimization), the proximal point algorithm, and the forward-backward splitting algorithm. The sufficiency of these EB conditions for linear convergence has been widely known. In contrast, less attention has focused on the discussion of necessity.

In addition, we also make the following contributions, separatively from aspects of block coordinate gradient descent, Nesterov's acceleration, and verifying EB conditions:

3. We show linear convergence for the proximal alternating linearized minimization (PALM) algorithm under a group of equivalent EB conditions. It has been recently shown [48, 25, 32] that PALM achieves sublinear convergence for convex problems and linear convergence for strongly convex problems. We in this study show its linear convergence under strictly weaker conditions than strong convexity.

4. By defining a new EB condition, we obtain Q-linear convergence of the Nesterov's accelerated forward-backward algorithm, which generalizes the Q-linear convergence of the Nesterov's accelerated gradient method, recently independently discoverd in [28] and [55]. The new EB condition in some special cases can be viewed as a strictly weaker relaxation of strong convexity. In such sense, we show Q-linear convergence of the Nesterov's accelerated method without strong convexity. Our proof idea is partially inspired by [5] but might be of interest in its own right.

5. We provide a new proof to show that a class of dual objective functions satisfy EB conditions, under slightly weaker assumptions, again by means of technique developed in [10]. The authors of [30] gave the first proof for a special case of this class of functions, and the author of [47] gave the first general proof by contradiction.

The paper is organized as follows. In Section 2, we present the basic notation and some elementary preliminaries. In Section 3, we analyze necessary and sufficient conditions guaranteeing linear convergence of gradient descent. In Section 4, we define a group of abstract EB conditions, and analyze the interplay between them. In Section 5, we define an abstract gradient-type method, and figure out EB conditions that are necessary and sufficient guaranteeing its linear convergence. In Section 6, we study linear convergence of the PALM algorithm. In Section 7, we study linear convergence of the Nesterov's accelerated forward-backward algorithm. In Section 8, we verify EB conditions for a class of dual objective functions. Finally, in Section 9, we give a short summary of this paper, along with some discussion for future work.

2 Notation and preliminaries

Throughout the paper, $\mathbb{R}^n$ will denote an n-dimensional Euclidean space associated with inner-product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. For any nonempty $Q \subset \mathbb{R}^n$, we define the distant function by $d(x, Q) := \inf_{y \in Q} \|x - y\|$. For a nonempty set $Q \subset \mathbb{R}^n$, we define the indicator function of Q by

$$\delta_Q(x) := \begin{cases} 0, & \text{if } x \in Q; \\ +\infty, & \text{otherwise.} \end{cases}$$

We say that f is gradient-Lipschitz-continuous with modulus $L > 0$ if

$$\forall x, y \in \mathbb{R}^n, \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

and f is strongly convex with modulus $\mu > 0$ if for any $\alpha \in [0, 1]$,

$$\forall x, y \in \mathbb{R}^n, \quad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{1}{2}\mu\alpha(1-\alpha)\|x - y\|^2,$$

or if (when it is differentiable)

$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2.$$

We will consider the following classes of functions.

- $\mathcal{F}^1(\mathbb{R}^n)$: the class of continuously differentiable convex functions from $\mathbb{R}^n \to \mathbb{R}$;
- $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$: the class of gradient-Lipschitz-continuous convex functions from $\mathbb{R}^n \to \mathbb{R}$ with Lipschitz modulus L;
- $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$: the class of gradient-Lipschitz-continuous and strongly convex functions from $\mathbb{R}^n \to \mathbb{R}$ with Lipschitz modulus L and strongly convex modulus μ;
- $\Gamma(\mathbb{R}^n)$: the class of proper and lower semicontinuous functions from $\mathbb{R}^n \to (-\infty, +\infty]$;
- $\Gamma_0(\mathbb{R}^n)$: the class of proper and lower semicontinuous convex functions from $\mathbb{R}^n \to (-\infty, +\infty]$.

Obviously, we have the following inclusions:

$$\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n) \subseteq \mathcal{F}_L^{1,1}(\mathbb{R}^n) \subseteq \mathcal{F}^1(\mathbb{R}^n), \quad \Gamma_0(\mathbb{R}^n) \subseteq \Gamma(\mathbb{R}^n).$$

It is convenient to denote by $\operatorname{Arg\,min} f$ the set of optimal solutions of minimizing f over $\mathbb{R}^n$, and to use "$\arg\min f$", if the solution is unique, to stand for the unique solution. If $\operatorname{Arg\,min} f$ is nonempty, we let $\min f$ present the minimum of f over $\mathbb{R}^n$.

The notation of subgradient plays a central role in (non)convex optimization.

Definition 1 (subgradients, [45])**.** *Let $f \in \Gamma(\mathbb{R}^n)$. Its domain is defined by*

$$\mathrm{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

(a) For a given $x \in \mathrm{dom} f$, the regular subgradient of f at x, written $\hat{\partial} f(x)$, is the set of all vectors $u \in \mathbb{R}^n$ which satisfy

$$\lim_{y \neq x} \inf_{y \to x} \frac{f(y) - f(x) - \langle u, y - x\rangle}{\|y - x\|} \geq 0.$$

When $x \notin \mathrm{dom} f$, we set $\hat{\partial} f(x) = \emptyset$.

(b) The (general) subgradient, of f at $x \in \mathbb{R}^n$, written $\partial f(x)$, is defined through the following closure process

$$\partial f(x) := \{u \in \mathbb{R}^n : \exists x^k \to x, f(x^k) \to f(x) \text{ and } u^k \in \hat{\partial} f(x^k) \to u \text{ as } k \to \infty\}.$$

(c) *If we further assume that f is convex, then the subgradient of f at $x \in \mathrm{dom} f$ can also be defined by*

$$\partial f(x) := \{v \in \mathbb{R}^n : f(z) \geq f(x) + \langle v, z - x\rangle, \quad \forall\ z \in \mathbb{R}^n\}.$$

It should be noted that for each $x \in \mathrm{dom} f$, $\partial f(x)$ is closed (see Theorem 8.6 in [45]). Moreover, if $f \in \Gamma_0(\mathbb{R}^n)$, then for each $x \in \mathrm{dom} f$, $\partial f(x)$ is a nonempty closed and convex set. In the later case, we denote by $\partial^0 f(x)$ the unique least-norm element of $\partial f(x)$. Points whose subgradient contains 0 are called critical points. The set of critical points of f is denoted by $\mathbf{crit} f$. If $f \in \Gamma_0(\mathbb{R}^n)$, then $\mathbf{crit} f = \mathrm{Arg}\min f$.

Let $f \in \Gamma_0(\mathbb{R}^n)$; its Fenchel conjugate function $f^* : \mathbb{R}^n \to (-\infty, +\infty]$ is defined by

$$f^*(x) := \sup_{y \in \mathbb{R}^n} \{\langle y, x\rangle - f(y)\},$$

and the proximal mapping operator by

$$\mathbf{prox}_{\lambda f}(x) := \arg\min_{y \in \mathbb{R}^n} \{f(y) + \frac{1}{2\lambda}\|y - x\|^2\}.$$

For each $x \in \overline{\mathrm{dom} f}$, there is a unique absolutely continuous curve $\chi_x : [0, \infty) \to \mathbb{R}^n$ such that $\chi_x(0) = x$ and for almost every $t > 0$,

$$\dot{\chi}_x(t) \in -\partial f(\chi_x(t)).$$

We say that $\Omega \subset \mathbb{R}^n$ is ∂f-invariant if

$$(\forall x \in \Omega \cap \mathrm{dom}\ \partial f)(\text{for a.e.}, t > 0) \ \ \chi_x(t) \in \Omega.$$

This concept was proposed in [12] and recently used in [21]. There are several types of Ω being ∂f-invariant; see Example 7.2 in [21] and Section IV.4 in [12]. In Sections 5 and 8, we will use the fact that sublevel $X_r := \{x : f(x) \leq r\}$ is always ∂f-invariant for any function $f \in \Gamma_0(\mathbb{R}^n)$.

At last, we present some variational analysis tools. Let $\mathcal{T}$, $\mathcal{E}$, and $\mathcal{E}_i$, $i = 1, 2$ be finite-dimensional Euclidean spaces. The closed ball around $x \in \mathcal{E}$ with radius $r > 0$ is denoted by $\mathbb{B}_{\mathcal{E}}(x, r) := \{y \in \mathcal{E} : \|x - y\| \leq r\}$. The unit ball is denoted by $\mathbb{B}_{\mathcal{E}}$ for simplicity, and the open unit ball around the original in $\mathcal{E}$ is by $\mathbb{B}^o_{\mathcal{E}}$. A muti-function $S : \mathcal{E}_1 \rightrightarrows \mathcal{E}_2$ is a mapping assigning each point in $\mathcal{E}_1$ to a subset of $\mathcal{E}_2$. The graph of S is defined by

$$\mathtt{gph}(S) := \{(u, v) \in \mathcal{E}_1 \times \mathcal{E}_2 : v \in S(u)\}.$$

The inverse map $S^{-1} : \mathcal{E}_2 \rightrightarrows \mathcal{E}_1$ is defined by setting

$$S^{-1}(v) := \{u \in \mathcal{E}_1 : v \in S(u)\}.$$

Calmness and metric subregularity have been considered in various contexts and under various names. Here, we follow the terminology of Dontchev and Rockafellar [16].

Definition 2 ([16], Chapter 3H)**.** *(a) A muti-function $S : \mathcal{E}_1 \rightrightarrows \mathcal{E}_2$ is said to be calm with constant $\kappa > 0$ around $\bar{u} \in \mathcal{E}_1$ for $\bar{v} \in \mathcal{E}_2$ if $(\bar{u}, \bar{v}) \in \mathtt{gph}(S)$ and there exist constants $\epsilon, \delta > 0$ such that*

$$S(u) \cap \mathbb{B}_{\mathcal{E}_2}(\bar{v}, \epsilon) \subseteq S(\bar{u}) + \kappa \cdot \|u - \bar{u}\|_2 \mathbb{B}_{\mathcal{E}_2}, \quad \forall u \in \mathbb{B}_{\mathcal{E}_1}(\bar{u}, \delta), \tag{1}$$

or equivalently,

$$S(u) \cap \mathbb{B}_{\mathcal{E}_2}(\bar{v}, \epsilon) \subseteq S(\bar{u}) + \kappa \cdot \|u - \bar{u}\|_2 \mathbb{B}_{\mathcal{E}_2}, \quad \forall u \in \mathcal{E}_1. \tag{2}$$

(b) A muti-function $S:\mathcal{E}_1 \rightrightarrows \mathcal{E}_2$ is said to be metrically sub-regular with constant $\kappa > 0$ around $\bar{u} \in \mathcal{E}_1$ for $\bar{v} \in \mathcal{E}_2$ if $(\bar{u},\bar{v}) \in \mathtt{gph}(S)$ and there exists a constant $\delta > 0$ such that

$$d(u, S^{-1}(\bar{v})) \le \kappa \cdot d(\bar{v}, S(u)), \quad \forall u \in \mathbb{B}_{\mathcal{E}_1}(\bar{u},\delta). \tag{3}$$

Note that the calmness defined above is weaker than the locally upper Lipschitz-continuous property [43]:

$$S(u) \subseteq S(\bar{u}) + \kappa \cdot \|u-\bar{u}\|_2 \mathbb{B}_{\mathcal{E}_2}, \quad \forall u \in \mathbb{B}_{\mathcal{E}_1}(\bar{u},\delta), \tag{4}$$

which requires the muti-functions S to be calm around $\bar{u} \in \mathcal{E}_1$ with constant $\kappa > 0$ for any $\bar{v} \in \mathcal{E}_2$. Recently, the locally upper Lipschitz-continuous property (4) was employed in [47] as a main assumption for verifying EB conditions of a class of dual objective functions.

3 The gradient descent: a necessary and sufficient condition for linear convergence

In this section, we first figure out the weakest condition that ensures gradient descent to converge linearly, and then we show that a number of existing linear convergence results can be recovered in a unified and transparent manner. This is a "warm-up" section for the forthcoming abstract theory in Sections 4 and 5.

Now, we start by considering the following unconstrained optimization problem

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} f(x),$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is a differentiable function achieving its minimum $\min f$ so that $\operatorname{Arg\,min} f \neq \emptyset$. Note that $\operatorname{Arg\,min} f$ is closed since f is differentiable. For any $x \in \mathbb{R}^n$, the set of its projection points onto $\operatorname{Arg\,min} f$, denoted by $\mathcal{Y}_f(x)$, is nonempty. Let $\{x_k\}_{k\ge 0}$ be generated by the gradient descent method

$$x_{k+1} = x_k - h \cdot \nabla f(x_k), \;\; k \ge 0, \tag{5}$$

where $h > 0$ is the step size. Observe that $d(x_k, \operatorname{Arg\,min} f)$ measures how close x_k is to $\operatorname{Arg\,min} f$, and the ratio of $d(x_{k+1}, \operatorname{Arg\,min} f)$ to $d(x_k, \operatorname{Arg\,min} f)$ measures how fast x_k converges to $\operatorname{Arg\,min} f$. Now, we analyze the ratio of $d(x_{k+1}, \operatorname{Arg\,min} f)$ to $d(x_k, \operatorname{Arg\,min} f)$ as follows

$$\begin{aligned} d^2(x_{k+1}, \operatorname{Arg\,min} f) &= \|x_{k+1} - x'_{k+1}\|^2 \le \|x_{k+1} - x'_k\|^2 \\ &= \|x_k - h\cdot\nabla f(x_k) - x'_k\|^2 \\ &= d^2(x_k, \operatorname{Arg\,min} f) - 2h\langle \nabla f(x_k), x_k - x'_k\rangle + h^2\|\nabla f(x_k)\|^2, \end{aligned}$$

where $x'_{k+1} \in \mathcal{Y}_f(x_{k+1})$ and $x'_k \in \mathcal{Y}_f(x_k)$. To ensure gradient descent to converge linearly in the following sense:

$$d^2(x_{k+1}, \operatorname{Arg\,min} f) \le \tau \cdot d^2(x_k, \operatorname{Arg\,min} f), \;\; k \ge 0. \tag{6}$$

it suffices to require that for $k \ge 0, x'_k \in \mathcal{Y}_f(x_k)$,

$$d^2(x_k, \operatorname{Arg\,min} f) - 2h\langle \nabla f(x_k), x_k - x'_k\rangle + h^2\|\nabla f(x_k)\|^2 \le \tau \cdot d^2(x_k, \operatorname{Arg\,min} f),$$

i.e.,

$$\inf_{u\in\mathcal{Y}_f(x^k)} \langle \nabla f(x_k), x_k - u\rangle \ge \frac{1-\tau}{2h} d^2(x_k, \operatorname{Arg\,min} f) + \frac{h}{2}\|\nabla f(x_k)\|^2, \;\; k \ge 0. \tag{7}$$

It turns out that this sufficient condition is also necessary when the objective function f belongs to $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and the step size h lies in some interval.

Proposition 1. *Let f be a differentiable function achieving its minimum* $\min f$ *so that* $\operatorname{Arg\,min} f \neq \emptyset$*, and let $h > 0$ and $\tau \in (0, 1)$.*

(i) If the condition (7) *holds, then the sequence $\{x_k\}_{k\geq 0}$ generated by the gradient descent method* (5) *must converge linearly in the sense of* (6)*.*

(ii) Let $f(x) \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$. If the sequence $\{x_k\}$ generated by the gradient descent method (5) *with $0 < h \leq \frac{1-\sqrt{\tau}}{L}$ converges linearly as* (6)*, then the condition* (7) *must hold.*

Proof. The proof of sufficiency part has been done. We now show the necessity part. Pick $u_{k+1} \in \mathcal{Y}_f(x_{k+1})$ to derive that

$$\begin{aligned} d(x_k, \operatorname{Arg\,min} f) &\leq \|x_k - u_{k+1}\| \leq \|x_{k+1} - u_{k+1}\| + \|x_{k+1} - x_k\| \\ &= d(x_{k+1}, \operatorname{Arg\,min} f) + h\|\nabla f(x_k)\|, \ k \geq 0. \end{aligned} \tag{8}$$

Combine (8) and the fact of linear convergence

$$d(x_{k+1}, \operatorname{Arg\,min} f) \leq \sqrt{\tau} \cdot d(x_k, \operatorname{Arg\,min} f), \ k \geq 0$$

to obtain

$$(1 - \sqrt{\tau}) d(x_k, \operatorname{Arg\,min} f) \leq h\|\nabla f(x_k)\|, \ k \geq 0. \tag{9}$$

According to Theorem 2.1.5 in [40], we know that $f(x) \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ implies

$$\langle \nabla f(x_k), x_k - v_k \rangle \geq \frac{1}{L}\|\nabla f(x_k)\|^2, \ v_k \in \mathcal{Y}_f(x_k), \ k \geq 0.$$

By letting $\alpha + \beta \leq 1$ and $\alpha, \beta > 0$, we have that for any $v_k \in \mathcal{Y}_f(x_k)$,

$$\begin{aligned} \langle \nabla f(x_k), x_k - v_k \rangle &\geq \frac{\alpha}{L}\|\nabla f(x_k)\|^2 + \frac{\beta}{L}\|\nabla f(x_k)\|^2 \\ &\geq \frac{\alpha}{L}\|\nabla f(x_k)\|^2 + \frac{\beta(1-\sqrt{\tau})^2}{Lh^2} d(x_k, \operatorname{Arg\,min} f)^2, \ k \geq 0, \end{aligned}$$

where the last inequality follows by (9). Thus, by letting $\frac{\alpha}{L} = \frac{h}{2}$ and $\frac{\beta(1-\sqrt{\tau})^2}{Lh^2} = \frac{1-\tau}{2h}$, we get the condition (7). At last, we need

$$\alpha + \beta = \frac{Lh}{2} + \frac{Lh(1-\tau)}{2(1-\sqrt{\tau})^2} = \frac{hL}{1-\sqrt{\tau}} \leq 1,$$

which forces $h \leq \frac{1-\sqrt{\tau}}{L}$. This completes the proof. □

The condition (7) means that if the steepest descent direction $-\nabla f(x)$ is well correlated to any desired descent directions $u - x$, where $u \in \mathcal{Y}_f(x)$, then a linear convergence rate of the gradient descent method can be ensured. Conversely, when $f(x) \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and if the gradient descent converges linearly and the step size lies in the interval $(0, \frac{1-\sqrt{\tau}}{L}]$, then $-\nabla f(x)$ must be well correlated to $u - x$. Now, we list some direct applications of this basic observation.

In our first illustrating example, we consider functions in $S_{\mu,L}^{1,1}(\mathbb{R}^n)$. First, we introduce an important property about this type of functions.

Lemma 1 ([40])**.** *If* $f \in S^{1,1}_{\mu,L}(\mathbb{R}^n)$*, then we have*

$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\mu L}{\mu + L}\|x-y\|^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Let x^* be the unique minimizer of $f \in S^{1,1}_{\mu,L}(\mathbb{R}^n)$; then $\operatorname{Arg\,min} f = \{x^*\}$. Using the inequality above with $x = x_k, y = x^*$ and noting that $\nabla f(x^*) = 0$ and $\|x_k - x^*\| = d(x_k, \operatorname{Arg\,min} f)$, we obtain

$$\langle \nabla f(x_k), x_k - x^*\rangle \geq \frac{\mu L}{\mu+L} d^2(x_k, \operatorname{Arg\,min} f) + \frac{1}{\mu+L}\|\nabla f(x_k)\|^2, \ k \geq 0.$$

To guarantee the condition (7), we only need

$$\frac{\mu L}{\mu+L} \geq \frac{1-\tau}{2h} \quad \text{and} \quad \frac{1}{\mu+L} \geq \frac{h}{2},$$

which implies that

$$\frac{(1-\tau)(\mu+L)}{2\mu L} \leq h \leq \frac{2}{\mu+L}, \quad \tau \geq \tau_0 := (\frac{L-\mu}{L+\mu})^2.$$

The optimal linear convergence rate τ_0 can be obtained by setting $h = \frac{2}{\mu+L}$. This gives the corresponding result in Nesterov's book; see Theorem 2.1.15 in [40].

In our second illustrating example, we consider RSC functions [64, 61]. The following property can be viewed as a convex combination of the restricted strong convexity and the gradient-Lipschitz-continuous property; see Lemma 3 in [61].

Lemma 2 ([61])**.** *If* $f \in \mathcal{F}^{1,1}_L(\mathbb{R}^n)$ *and* f *is RSC with* $0 < \nu < L$*, then for every* $\theta \in [0,1]$ *it holds:*

$$\forall x \in \mathbb{R}^n, \quad \langle \nabla f(x), x - x'\rangle \geq \frac{\theta}{L}\|\nabla f(x)\|^2 + (1-\theta)\nu d^2(x, \operatorname{Arg\,min} f),$$

where x' *is the unique projection point of* x *onto* $\operatorname{Arg\,min} f$ *since* $\operatorname{Arg\,min} f$ *is a nonempty closed convex set.*

Similarly, to guarantee the condition (7) , we only need

$$(1-\theta)\nu \geq \frac{1-\tau}{2h} \quad \text{and} \quad \frac{\theta}{L} \geq \frac{h}{2},$$

which implies that

$$\frac{1-\tau}{2(1-\theta)\nu} \leq h \leq \frac{2\theta}{L}, \quad \tau \geq 1 - \frac{4\theta(1-\theta)\nu}{L} \geq 1 - \frac{\nu}{L}.$$

The optimal linear convergence rate $1 - \frac{\nu}{L}$ can be obtained at $\theta = \frac{1}{2}$ and $h = \frac{1}{L}$. This gives the corresponding result in [61]. The argument here is much simpler than that previously employed to derive the same result; see the proof of Theorem 2 in [61].

The last example to be illustrated is a nonconvex minimization. The following definition can be viewed as a local version of Lemma 2. Therefore, it is not difficult to predict a local linear convergence under such property.

Definition 3 (Regularity Condition, [14])**.** *Let $\mathcal{N}$ be a neighborhood of* $\operatorname{Arg\,min} f$ *and let* $\alpha, \beta > 0$*. We say that f satisfies the regularity condition if*

$$\forall x \in \mathcal{N}, \quad \inf_{u \in \mathcal{Y}_f(x)} \langle \nabla f(x), x - u \rangle \geq \frac{1}{\alpha} d^2(x, \operatorname{Arg\,min} f) + \frac{1}{\beta} \|\nabla f(x)\|^2.$$

Again, to guarantee the condition (7) locally, we only need

$$\frac{1}{\alpha} \geq \frac{1-\tau}{2h} \quad \text{and} \quad \frac{1}{\beta} \geq \frac{h}{2},$$

which implies that

$$\frac{(1-\tau)\alpha}{2} \leq h \leq \frac{2}{\beta}, \quad \tau \geq \tau_0 := (1 - \frac{4}{\alpha\beta}).$$

The optimal linear convergence rate τ_0 can be obtained by setting $h = \frac{2}{\beta}$ and assuming $\alpha\beta > 4$. The latter must hold if the regularity condition holds; see the argument below Lemma 7.10 in [14]. Therefore, we obtain the corresponding result in [14]. Regularity condition provably holds for nonconvex optimization problems that appear in phase retrieve and low-rank matrix recover; interested readers can refer to [14] and [53] for details.

Observe that the right-hand side of (7) has two terms. In order to better analyze such condition, we decompose it into two parts:

$$\inf_{u \in \mathcal{Y}_f(x^k)} \langle \nabla f(x_k), x_k - u \rangle \geq \theta_1 \cdot d^2(x_k, \operatorname{Arg\,min} f),$$
$$\inf_{u \in \mathcal{Y}_f(x^k)} \langle \nabla f(x_k), x_k - u \rangle \geq \theta_2 \cdot \|\nabla f(x_k)\|^2,$$

where $\theta_i, i = 1, 2$ are some positive parameters. This idea of separating the right-hand side of (7) partially inspires us to consider new and abstract error bound conditions, which are the main context of next section.

4 Abstract EB conditions: definition and interplay

This section is divided into two parts. In the first part, we define a group of EB conditions in a unified and abstract way. In the second part, we discuss some interplay between them, along with new connections between many existing EB conditions.

4.1 Definition of abstract EB conditions

The concept of residual measure operator, given by the following definition, will play a key role in the forthcoming theory.

Definition 4. *Let $\varphi \in \Gamma(\mathbb{R}^n)$. We say that $G_\varphi : \mathbb{R}^n \to \mathbb{R}^n$ is a residual measure operator related to φ, if it satisfies*

$$\{x \in \mathbb{R}^n : G_\varphi(x) = 0\} = \mathbf{crit}\varphi.$$

Especially, if we further assume that φ is convex, the above condition can be written as

$$\{x \in \mathbb{R}^n : G_\varphi(x) = 0\} = \operatorname{Arg\,min} \varphi.$$

Now, we define a group of abstract EB conditions.

Definition 5. *Let $\varphi \in \Gamma(\mathbb{R}^n)$ be such that it achieves its minimum* $\min \varphi$ *and that its critical point set* $\mathbf{crit}\varphi$ *is nonempty and closed. Let $\Omega \subset \mathbb{R}^n$ and let G_φ be a residual measure operator related to φ. Define the projection operator $\mathcal{P}_\varphi : \mathbb{R}^n \to \mathbb{R}^n$ onto* $\mathbf{crit}\varphi$ *by:*

$$\mathcal{P}_\varphi(x) := \operatorname{Arg\,min}_{u \in \mathbf{crit}\varphi} \|x - u\|.$$

We call $d(x, \mathbf{crit}\varphi)$ point value error, $\varphi(x) - \min \varphi$ objective value error, $\|G_\varphi(x)\|$ residual value error, and $\inf_{x_p \in \mathcal{P}_\varphi(x)} \langle G_\varphi(x), x - x_p \rangle$ least correlated error. With these optimality measures, we say that

1. *φ satisfies residual-point values EB condition with operator G_φ and constant $\kappa > 0$ on Ω, abbreviated $(G_\varphi, \kappa, \Omega)$-(res-EB) condition, if:*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \|G_\varphi(x)\| \geq \kappa \cdot d(x, \mathbf{crit}\varphi); \tag{res-EB}$$

2. *φ satisfies correlated-point values EB condition with operator G_φ and constant $\nu > 0$ on Ω, abbreviated (G_φ, ν, Ω)-(cor-EB) condition, if:*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \inf_{x_p \in \mathcal{P}_\varphi(x)} \langle G_\varphi(x), x - x_p \rangle \geq \nu \cdot d^2(x, \mathbf{crit}\varphi); \tag{cor-EB}$$

3. *φ satisfies objective-point values EB condition with constant $\alpha > 0$ on Ω, abbreviated $(\varphi, \alpha, \Omega)$-(obj-EB) condition, if:*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \varphi(x) - \min \varphi \geq \frac{\alpha}{2} \cdot d^2(x, \mathbf{crit}\varphi); \tag{obj-EB}$$

4. *φ satisfies residual-objective values EB condition with operator G_φ and constant $\eta > 0$ on Ω, abbreviated $(G_\varphi, \eta, \Omega)$-(res-obj-EB) condition, if:*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \|G_\varphi(x)\| \geq \eta \cdot \sqrt{\varphi(x) - \min \varphi}; \tag{res-obj-EB}$$

5. *φ satisfies correlated-residual values EB condition with operator G_φ and constant $\beta > 0$ on Ω, abbreviated $(G_\varphi, \beta, \Omega)$-(cor-res-EB) condition, if:*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \inf_{x_p \in \mathcal{P}_\varphi(x)} \langle G_\varphi(x), x - x_p \rangle \geq \beta \cdot \|G_\varphi(x)\|^2; \tag{cor-res-EB}$$

6. *φ satisfies correlated-objective values EB condition with operator G_φ and constant $\omega > 0$ on Ω, abbreviated $(G_\varphi, \omega, \Omega)$-(cor-obj-EB) condition, if:*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \inf_{x_p \in \mathcal{P}_\varphi(x)} \langle G_\varphi(x), x - x_p \rangle \geq \omega \cdot (\varphi(x) - \min \varphi). \tag{cor-obj-EB}$$

We will refer to these EB conditions as global if $\Omega = \mathbb{R}^n$. For global EB conditions, we will omit Ω for simplicity.

In order to gain some intuition of the abstract EB conditions, we point out their correspondences to existing notions: (res-EB) corresponds to the EB condition of Hoffman's type, (res-obj-EB) to the Polyak-Łojasiewicz's type, (obj-EB) to the quadratic growth condition, (cor-EB) to the RSI's type, and (cor-obj-EB) to the subgradient inequality of convex function. The (cor-res-EB) condition, which will be used in Section 5, is a relaxation of the following property:

$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla\varphi(x) - \nabla\varphi(y), x - y\rangle \geq \frac{1}{L}\|\nabla\varphi(x) - \nabla\varphi(y)\|^2,$$

which is equivalent to $\varphi \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$; see Theorem 2.1.5 in [40].

In our early manuscript [59], we only roughly gave global EB conditions in definition 5. By incorporating the referee's comments, we present the current version, which is much more accurate than the previous one.

4.2 Interplay between the EB conditions

We first show the interplay between the abstract EB conditions. The proof of equivalence will rely heavily on a technical result developed in [10].

Theorem 1. *Let $\varphi \in \Gamma(\mathbb{R}^n)$ be such that it achieves its minimum $\min\varphi$ and that $\mathbf{crit}\varphi$ is nonempty and closed. Let $\Omega \subset \mathbb{R}^n$ and let G_φ be a residual measure operator related to φ. Assume that the $(G_\varphi, \omega, \Omega)$-(cor-obj-EB) condition holds. Then, we have the following implications*

$$\text{(obj-EB)} \Rightarrow \text{(cor-EB)} \Rightarrow \text{(res-EB)} \Rightarrow \text{(res-obj-EB)}.$$

One can respectively take $\nu = \frac{\alpha\omega}{2}, \kappa = \nu, \eta = \sqrt{\kappa\omega}$. If we further assume that $\varphi \in \Gamma_0(\mathbb{R}^n)$, Ω is $\partial\varphi$-invariant, and G_φ satisfies

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \|G_\varphi(x)\| \leq \inf_{g\in\partial\varphi(x)} \|g\|. \tag{10}$$

Then, we have the following equivalent relationship

$$\text{(obj-EB)} \Leftrightarrow \text{(cor-EB)} \Leftrightarrow \text{(res-EB)} \Leftrightarrow \text{(res-obj-EB)}.$$

For (res-obj-EB) $\Rightarrow$ (obj-EB), *one can take $\alpha = \frac{1}{2}\eta^2$.*

Proof. We prove this theorem by showing the following implications

$$\text{(obj-EB)} \Rightarrow \text{(cor-EB)} \Rightarrow \text{(res-EB)} \Rightarrow \text{(res-obj-EB)} \Rightarrow \text{(obj-EB)}.$$

Firstly, the implication of (obj-EB) $\Rightarrow$ (cor-EB) follows from that

$$\inf_{x_p\in\mathcal{P}_\varphi(x)} \langle G_\varphi(x), x - x_p\rangle \geq \omega \cdot (\varphi(x) - \min\varphi) \geq \frac{\alpha\omega}{2} \cdot d^2(x, \mathbf{crit}\varphi),$$

where the left inequality is (cor-obj-EB) and the right one is (obj-EB).

Secondly, the implication of (cor-EB) $\Rightarrow$ (res-EB) follows from a direct application of the Cauchy-Schwartz inequality to (cor-EB).

Thirdly, we show (res-EB) $\Rightarrow$ (res-obj-EB). By (cor-obj-EB) and (res-EB), we derive that for $\forall\, x \in \Omega \cap \mathrm{dom}\varphi$,

$$\begin{aligned}\omega \cdot (\varphi(x) - \min \varphi) &\leq \inf_{x_p \in \mathcal{P}_\varphi(x)} \langle G_\varphi(x), x - x_p \rangle \\ &\leq \inf_{x_p \in \mathcal{P}_\varphi(x)} \|G_\varphi(x)\| \|x - x_p\| = \|G_\varphi(x)\| \cdot d(x, \mathbf{crit}\varphi) \\ &\leq \kappa^{-1} \|G_\varphi(x)\|^2.\end{aligned}$$

Thus, it holds that $\forall\, x \in \Omega \cap \mathrm{dom}\varphi,\quad \|G_\varphi(x)\| \geq \sqrt{\kappa\omega} \cdot \sqrt{\varphi(x) - \min \varphi}$, which is just (res-obj-EB).

At last, we show (res-obj-EB) $\Rightarrow$ (obj-EB). The following is based on an argument used for proving Theorem 27 in [10]. For the sake of completeness, we reproduce that proof in our particular case. First of all, take $x \in \Omega \cap \mathrm{dom}\varphi$ and recall that we have additionally assumed $\mathbf{crit}\varphi = \mathrm{Arg}\min \varphi$. Without loss of generality, we assume that $\min \varphi = 0$ and $x \notin \mathrm{Arg}\min \varphi$. According to the result about subgradient curves due to Brézis [12] and Bruck [13] and recently appeared in [10], we can find the unique absolutely continuous curve $\chi_x : [0, +\infty) \to \mathbb{R}^n$ such that $\chi_x(0) = x$ and

$$\dot{\chi}_x(t) \in -\partial\varphi(\chi_x(t))$$

for almost every $t > 0$. Moreover, $\chi_x(t)$ converges to some point in $\mathrm{Arg}\min \varphi$ as $t \to +\infty$ and the function $t \mapsto \varphi(\chi_x(t))$ is nonincreasing and

$$\lim_{t \to +\infty} \varphi(\chi_x(t)) = \min \varphi = 0.$$

By the $\partial\varphi$-invariant property of Ω, we have $\chi_x(t) \in \Omega$ and hence $\chi_x(t) \in \Omega \cap \mathrm{dom}\varphi$ due to the nonincreasing of $\varphi(\chi_x(t))$. Let

$$T := \inf\{t \in [0, +\infty) : \varphi(\chi_x(t)) = 0\}.$$

We claim that $T > 0$. Otherwise, $T = 0$ and then, by the lower semicontinuous property of φ, we can derive that

$$\varphi(x) = \varphi(\chi_x(0)) \leq \liminf_{t \to 0^+} \varphi(\chi_x(t)) = 0.$$

This contradicts $x \notin \mathrm{Arg}\min \varphi$. Now, combining (10) and (res-obj-EB), we derive that

$$\frac{\|\dot{\chi}_x(t)\|}{\sqrt{\varphi(\chi_x(t))}} \geq \frac{\inf_{g \in \partial\varphi(\chi_x(t))} \|g\|}{\sqrt{\varphi(\chi_x(t))}} \geq \frac{\|G_\varphi(\chi_x(t))\|}{\sqrt{\varphi(\chi_x(t))}} \geq \eta, \quad \forall t \in [0, T).$$

Observe that for $p, q \in [0, T)$,

$$\begin{aligned}&\sqrt{\varphi(\chi_x(p))} - \sqrt{\varphi(\chi_x(q))} = \int_q^p \frac{d\sqrt{\varphi(\chi_x(t))}}{dt} dt \\ =&\frac{1}{2} \int_p^q (\varphi(\chi_x(p)))^{-\frac{1}{2}} \|\dot{\chi}_x(t)\|^2 dt = \frac{1}{2} \int_p^q \frac{\|\dot{\chi}_x(t)\|}{\sqrt{\varphi(\chi_x(t))}} \|\dot{\chi}_x(t)\| dt \\ \geq&\frac{1}{2} \int_p^q \eta \|\dot{\chi}_x(t)\| dt = \frac{\eta}{2} \cdot \mathrm{length}(\chi_x(t), p, q) \geq \frac{\eta}{2} \cdot \|\chi_x(p) - \chi_x(q)\|,\end{aligned}$$

where $\text{length}(\chi_x(t), p, q)$ stands for the length of subgradient curve from p to q. By letting $p = 0$ and $q \to +\infty$ if $T = +\infty$ and $q \to T$ if $T < +\infty$, we obtain

$$\sqrt{\varphi(\chi_x(0))} = \sqrt{\varphi(x)} \geq \frac{\eta}{2} \cdot \|x - \hat{x}\|$$

for some $\hat{x} \in \operatorname{Arg\,min} \varphi$. Therefore, for $\forall\ x \in \Omega \cap \mathrm{dom}\varphi$ we always have

$$\varphi(x) - \min \varphi \geq \frac{\eta^2}{4} \cdot \|x - \hat{x}\|^2 \geq \frac{\eta^2}{4} \cdot d^2(x, \operatorname{Arg\,min} \varphi) = \frac{\eta^2}{4} \cdot d^2(x, \mathbf{crit}\varphi),$$

which implies that (obj-EB) with $\alpha = \frac{\eta^2}{2}$ holds. This completes the proof. □

As a direct consequence, we have the following corollary.

Corollary 1. *Let $\varphi \in \Gamma_0(\mathbb{R}^n)$ be such that its achieves its minimum $\min \varphi$ so that $\operatorname{Arg\,min} \varphi \neq \emptyset$. Let $\Omega \subset \mathbb{R}^n$ be $\partial\varphi$-invariant, and $G^i_\varphi, i = 1, 2$ be two different residual measure operators related to the same function φ. We assume that $G^i_\varphi, i = 1, 2$ satisfy*

$$\forall\ x \in \Omega \cap \mathrm{dom}\varphi, \quad \|G^i_\varphi(x)\| \leq \inf_{g \in \partial\varphi(x)} \|g\|, \tag{11}$$

and $(G^i_\varphi, \omega, \Omega)$-(cor-obj-EB) conditions hold. Then, we have

$$\begin{gathered}(G^1_\varphi, \kappa, \Omega)\text{-(res-EB)} \Leftrightarrow (G^1_\varphi, \nu, \Omega)\text{-(cor-EB)} \Leftrightarrow (G^1_\varphi, \eta, \Omega)\text{-(res-obj-EB)} \\ \Leftrightarrow (\varphi, \alpha, \Omega)\text{-(obj-EB)} \Leftrightarrow \\ (G^2_\varphi, \kappa, \Omega)\text{-(res-EB)} \Leftrightarrow (G^2_\varphi, \nu, \Omega)\text{-(cor-EB)} \Leftrightarrow (G^2_\varphi, \eta, \Omega)\text{-(res-obj-EB)}.\end{gathered}$$

Now, we list some cases where the equivalence between the EB conditions indeed holds.

Corollary 2. *The EB conditions* (cor-EB), (res-EB), (obj-EB), *and* (res-obj-EB) *are equivalent under each of the following situations:*

case 1: *$\varphi \in \mathcal{F}^1(\mathbb{R}^n)$ achieves its minimum $\min \varphi$, Ω is $\nabla\varphi$-invariant, and $G_\varphi = \nabla\varphi$;*

case 2: *$\varphi \in \Gamma_0(\mathbb{R}^n)$ achieves its minimum $\min \varphi$, Ω is $\partial\varphi$-invariant, and $G_\varphi = \partial^0\varphi$;*

case 3: *$\varphi = f + g$, where $f \in \mathcal{F}^{1,1}_L(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$, achieves its minimum $\min \varphi$, Ω is $\partial\varphi$-invariant, and $G_\varphi(x) = \mathcal{R}_t(x)$, where*

$$\forall\ x \in \mathrm{dom}\varphi, \quad \mathcal{R}_t(x) := t^{-1}\left(x - \mathbf{prox}_{tg}(x - t\nabla f(x))\right),$$

with some $t \in (0, \frac{1}{L}]$. In addition, we assume that there exists a positive constant $\epsilon \leq \frac{2}{t}$ such that

$$\|G_\varphi(x)\|^2 \geq \epsilon(\varphi(x) - \varphi(x^+)), \tag{12}$$

where $x^+ = x - t \cdot G_\varphi(x)$.

Proof. First of all, $\mathbf{crit}\varphi$ is nonempty since $\mathbf{crit}\varphi = \operatorname{Arg\,min} \varphi \neq \emptyset$, and is closed since φ a proper and lower semicontinuous function, in all the listed cases. Secondly, by optimality conditions, one can easily verify that G_φ in all the listed cases are residual measure operators. We only need to verify the remained assumptions in Theorem 1.

For both cases 1 and 2, the convexity of φ implies the (cor-obj-EB) condition with $\omega = 1$.

In case 1, the assumption (10) holds obviously because of $\partial\varphi(x) = \{\nabla\varphi(x)\}$.

In case 2, the assumption (10) follows from the definition of $\partial^0\varphi(x)$.

Now, let us consider the case 3. Since $f(x) \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$, we have that $G_\varphi(x)$ satisfies the standard result

$$\forall x, y \in \mathbb{R}^n, \quad \varphi(x^+) \leq \varphi(y) + \langle G_\varphi(x), x - y\rangle - \frac{t}{2}\|G_\varphi(x)\|^2;$$

see e.g. Lemma 2.3 in [8] or Lemma 2 in the very recent work [4]. Since φ also belongs to $\Gamma_0(\mathbb{R}^n)$, we can conclude that $\operatorname{Arg\,min}\varphi$ is a nonempty closed convex set. Thus, by the projection theorem, there exists a unique projection point of x onto $\operatorname{Arg\,min}\varphi$, denoted by x_p. Using the inequality above with $y = x_p$ and the assumption (12), we derive that

$$\begin{aligned}
\langle G_\varphi(x), x - x_p\rangle &\geq \varphi(x^+) - \min\varphi + \frac{t}{2}\|G_\varphi(x)\|^2 \\
&\geq \varphi(x^+) - \min\varphi + \frac{t\epsilon}{2}(\varphi(x) - \varphi(x^+)) \\
&= \frac{t\epsilon}{2}(\varphi(x) - \min\varphi) + (1 - \frac{t\epsilon}{2})(\varphi(x^+) - \min\varphi) \\
&\geq \frac{t\epsilon}{2}(\varphi(x) - \min\varphi),
\end{aligned}$$

from which the $(G_\varphi, \omega, \Omega)$-(cor-obj-EB) condition with $\omega = \frac{t\epsilon}{2}$ follows. The assumption (10) in this case was established in Theorem 3.5 in [19] and Lemma 4.1 in [31]. This completes the proof. □

When this work was under review, we note that the authors of [26] independently also recently obtained the equivalent relationship between the EB conditions (cor-EB), (res-EB), (obj-EB), and (res-obj-EB) for functions in $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and we also note that the authors of [21] independently recently obtained the equivalent relationship between the EB conditions (res-EB), (obj-EB), and (res-obj-EB) for functions in $\Gamma_0(\mathbb{R}^n)$. The former is merely limited to $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and the latter mainly focuses on $\Gamma_0(\mathbb{R}^n)$ and does not consider (cor-EB).

Observe that the condition (12) is implied by the (res-obj-EB) condition since

$$\forall x \in \mathbb{R}^n, \quad \|G_\varphi(x)\|^2 \geq \eta^2(\varphi(x) - \min\varphi) \geq \eta^2(\varphi(x) - \varphi(x^+)).$$

And also, note that $\varphi = f + g \in \Gamma_0(\mathbb{R}^n)$ if $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$. With a little efforts, we can get the following result.

Corollary 3. *Let $\varphi = f + g$ with $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$ achieve its minimum $\min\varphi$, and let $\Omega \subset \mathbb{R}^n$ be $\partial\varphi$-invariant and $t \in (0, \frac{1}{L}]$. If the $(\mathcal{R}_t, \eta, \Omega)$-(res-obj-EB) condition holds, then we have the following equivalent relationship:*

$$\begin{gathered}
(\partial^0\varphi, \kappa, \Omega)\text{-(res-EB)} \Leftrightarrow (\partial^0\varphi, \nu, \Omega)\text{-(cor-EB)} \Leftrightarrow (\partial^0\varphi, \eta, \Omega)\text{-(res-obj-EB)} \\
\Leftrightarrow (\varphi, \alpha, \Omega)\text{-(obj-EB)} \Leftrightarrow \\
(\mathcal{R}_t, \kappa, \Omega)\text{-(res-EB)} \Leftrightarrow (\mathcal{R}_t, \nu, \Omega)\text{-(cor-EB)} \Leftrightarrow (\mathcal{R}_t, \eta, \Omega)\text{-(res-obj-EB)},
\end{gathered}$$

and each of the equivalent conditions holds.

Based on the relationship established in Theorem 2 in [60], that is $(\varphi, \alpha, \Omega)$-(obj-EB) $\Leftrightarrow$ $(\mathcal{R}_t, \kappa, \Omega)$-(res-EB)$\Leftrightarrow$ $(\mathcal{R}_t, \nu, \Omega)$-(cor-EB), and together with the case 2 of Corollary 2, we still have the following result even if we do not take the $(\mathcal{R}_t, \eta, \Omega)$-(res-obj-EB) condition as an assumption.

Corollary 4. *Let $\varphi = f + g$ with $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$ achieve its minimum* $\min \varphi$*, and let $\Omega \subset \mathbb{R}^n$ be $\partial\varphi$-invariant and $t \in (0, \frac{1}{L}]$. Then, we have*

$$(\partial^0\varphi, \kappa, \Omega)\text{-(res-EB)} \Leftrightarrow (\partial^0\varphi, \nu, \Omega)\text{-(cor-EB)} \Leftrightarrow (\partial^0\varphi, \eta, \Omega)\text{-(res-obj-EB)}$$
$$\Leftrightarrow (\varphi, \alpha, \Omega)\text{-(obj-EB)} \Leftrightarrow (\mathcal{R}_t, \kappa, \Omega)\text{-(res-EB)} \Leftrightarrow (\mathcal{R}_t, \nu, \Omega)\text{-(cor-EB)}.$$

In all corollaries above, parameters involved in different EB conditions can be set explicitly as Theorem 1, but we omit the details here.

5 An abstract gradient-type method: linear convergence and applications

In this section, we define an abstract gradient-type method by viewing residual measure operator as an ascent direction, and then figure out a necessary and sufficient condition for linear convergence based on the abstract EB conditions defined before. The following main result generalizes Proposition 1.

Theorem 2. *Let $\varphi \in \Gamma(\mathbb{R}^n)$ be such that it achieves its minimum $\min \varphi$ and that $\mathbf{crit}\varphi$ is nonempty closed. Let $\Omega \subset \mathbb{R}^n$ and let G_φ be a residual measure operator related to φ. Suppose that φ satisfies the $(G_\varphi, \beta, \Omega)$-(cor-res-EB) condition. Define the abstract gradient-type method by*

$$x_{k+1} = x_k - h \cdot G_\varphi(x_k), \;\; k \geq 0,$$

with step size $h > 0$ and arbitrary initial point $x_0 \in \Omega$. Assume that $x_k \in \Omega, k \geq 0$. Let $\tau, \theta \in (0, 1)$.

(i) If φ satisfies the (G_φ, ν, Ω)-(cor-EB) condition with $\nu < \frac{1}{\beta}$ and the following inequalities hold

$$\frac{1-\tau}{2\theta\nu} \leq h \leq 2(1-\theta)\beta, \quad \tau \geq 1 - 4\theta(1-\theta)\beta\nu, \tag{13}$$

then the abstract gradient-type method converges linearly in the sense that

$$d^2(x_{k+1}, \mathbf{crit}\varphi) \leq \tau \cdot d^2(x_k, \mathbf{crit}\varphi), \quad k \geq 0. \tag{14}$$

The optimal rate $\tau_0 := 1 - \beta\nu$ is obtained at $h = \beta$ and $\theta = \frac{1}{2}$.

(ii) Conversely, if the abstract gradient-type method converges linearly in the sense of (14), then φ satisfies the (G_φ, ν, Ω)-(cor-EB) condition with $\nu = \frac{\beta(1-\sqrt{\tau})^2}{h^2}$.

Proof. First, we repeat the argument before (6) to obtain that for $v_k \in \mathcal{P}_\varphi(x_k)$,

$$d^2(x_{k+1}, \mathbf{crit}\varphi) \leq d^2(x_k, \mathbf{crit}\varphi) - 2h\langle G_\varphi(x_k), x_k - v_k\rangle + h^2\|G_\varphi(x_k)\|^2, \;\; k \geq 0.$$

Take $\theta \in (0, 1)$ and then use a convex combination of the (cor-res-EB) and (cor-EB) conditions at $x = x_k$ to obtain

$$\inf_{v_k \in \mathcal{P}_\varphi(x_k)} \langle G_\varphi(x_k), x_k - v_k\rangle \geq \theta\nu \cdot d^2(x_k, \mathbf{crit}\varphi) + (1-\theta)\beta \cdot \|G_\varphi(x_k)\|^2, \;\; k \geq 0.$$

Therefore, we can derive that

$$\begin{aligned} d^2(x_{k+1}, \mathbf{crit}\varphi) &\leq (1-2\theta\nu h)d^2(x_k, \mathbf{crit}\varphi) + (h^2 - 2h(1-\theta)\beta)\|G_\varphi(x_k)\|^2 \\ &\leq \tau \cdot d^2(x_k, \mathbf{crit}\varphi),\ k \geq 0, \end{aligned}$$

where the second inequality follows from the condition (13) on the step size. Obviously, the optimal linear convergence rate $\tau_0 = 1 - \beta\nu$ can be obtained at $h = \beta, \theta = \frac{1}{2}$.

Conversely, pick $u_{k+1} \in \mathcal{P}_\varphi(x_{k+1})$ to derive that

$$\begin{aligned} d(x_k, \mathbf{crit}\varphi) &\leq \|x_k - u_{k+1}\| \leq \|x_{k+1} - u_{k+1}\| + \|x_{k+1} - x_k\| \\ &= d(x_{k+1}, \mathbf{crit}\varphi) + h\|G_\varphi(x_k)\|,\ k \geq 0. \end{aligned} \tag{15}$$

Combine (15) and the fact of linear convergence

$$d(x_{k+1}, \mathbf{crit}\varphi) \leq \sqrt{\tau} \cdot d(x_k, \mathbf{crit}\varphi),\ k \geq 0$$

to obtain

$$(1-\sqrt{\tau})^2 d^2(x_k, \mathbf{crit}\varphi) \leq h^2\|G_\varphi(x_k)\|^2,\ k \geq 0.$$

Thus, together with the (cor-res-EB) condition, we can derive that

$$\inf_{v_k \in \mathcal{P}_\varphi(x_k)} \langle G_\varphi(x_k), x_k - v_k\rangle \geq \beta\|G_\varphi(x_k)\|^2 \geq \frac{\beta(1-\sqrt{\tau})^2}{h^2} d^2(x_k, \mathbf{crit}\varphi),\ k \geq 0.$$

Observe that the starting point $x_0 \in \Omega$ can be arbitrary. Therefore, the (cor-EB) condition with $\nu = \frac{\beta(1-\sqrt{\tau})^2}{h^2}$ holds. This completes the proof. □

With Theorem 2 in hand, we now claim the necessary and sufficient EB conditions guaranteeing linear convergence for the gradient method, the proximal point algorithm, and the forward-backward splitting algorithm. These conditions, previously known to be sufficient for linear convergence, are actually necessary. We start by the gradient method, applied to possibly nonconvex optimization.

Corollary 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a gradient-Lipschitz-continuous function with modulus $L > 0$ and let $\Omega \subset \mathbb{R}^n$. Assume that f achieves its minimum $\min f$ and $\mathbf{crit} f = \operatorname{Arg\,min} f \neq \emptyset$. Let $\{x_k\}_{k\geq 0}$ be generated by the gradient descent method (5) with $h = \frac{1}{L}$ and assume that $x_k \in \Omega, k \geq 0$.*

(i) If f satisfies the $(\nabla f, \nu, \Omega)$-(cor-EB) condition with $\nu < L$, then the gradient descent method (5) with $h = \frac{1}{L}$ converges linearly in the sense that

$$f(x_{k+1}) - \min f \leq \left(1 - (\frac{\nu}{L})^2\right)(f(x_k) - \min f),\ k \geq 0. \tag{16}$$

(ii) If we further assume that f is convex, then the gradient descent method (5) with $h = \frac{1}{L}$ attains the following linear convergence:

$$d^2(x_{k+1}, \operatorname{Arg\,min} f) \leq (1 - \frac{\nu}{L}) \cdot d^2(x_k, \operatorname{Arg\,min} f),\ k \geq 0. \tag{17}$$

(iii) *Conversely, if f is convex and if starting from arbitrary initial point $x_0 \in \Omega$, the gradient descent method* (5) *with $h = \frac{1}{L}$ converges linearly like* (17) *but replacing $1 - \frac{\nu}{L}$ with τ, then f satisfies the* $(\nabla f, \nu, \Omega)$-(cor-EB) *condition with $\nu = L(1-\sqrt{\tau})^2$.*

Proof. We first show (16) by modifying the argument due to Polyak [42] and recently highlighted in [27, 26]. The gradient-Lipschitz-continuous of f implies

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2, \ \forall \ x, y \in \mathbb{R}^n. \tag{18}$$

Using this inequality with $y = x_{k+1}$ and $x = x_k$ and together with the update rule of gradient descent, we get

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2, \ k \geq 0. \tag{19}$$

Using again the inequality (18) with $y = x_k$ and $x = u_k \in \mathcal{P}_f(x_k)$, and noting that $u_k \in \mathbf{crit} f = \operatorname{Arg\,min} f$ and hence $f(u_k) = \min f$ and $\nabla f(u_k) = 0$, we have

$$f(x_k) - \min f \leq \frac{L}{2} d^2(x_k, \mathbf{crit} f), \ k \geq 0. \tag{20}$$

Applying the Cauchy-Schwartz inequality to the $(\nabla f, \nu, \Omega)$-(cor-EB) condition, we obtain

$$\forall x \in \Omega \cap \operatorname{dom} f, \quad \|\nabla f(x)\| \geq \nu \cdot d(x, \mathbf{crit} f).$$

Thus, combining the inequalities (19) and (20), we have that

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2 \leq -\frac{\nu^2}{L^2}(f(x_k) - \min f), \ k \geq 0,$$

from which (16) follows.

Now, with the additional convex assumption of f, we have $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, which is equivalent to the following condition

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2, \quad x, y \in \mathbb{R}^n;$$

see Theorem 2.1.5 [40]. Using this inequality with $y \in \mathcal{P}_f(x)$, we obtain

$$\inf_{y \in \mathcal{P}_f(x)} \langle \nabla f(x), x - y\rangle \geq \frac{1}{L}\|\nabla f(x)\|^2, \ x \in \mathbb{R}^n,$$

which is just the $(\nabla f, \beta, \Omega)$-(cor-res-EB) condition with $\beta = \frac{1}{L}$. Therefore, the remained results follow from Theorem 2. This completes the proof. □

Remark 1. *In Example 2 in [61], we constructed a one-dimensional nonconvex function, that satisfies all the conditions in Corollary 5 that ensure* (16). *In this sense,* (16) *is one of the few general results for global linear convegence on non-convex problems. We note that a similar phenomenon was observed by the authors of [26] under the Polyak-Lojasiewicz condition.*

Before we discuss the linear convergence of the proximal point algorithm (PPA), we introduce the following result.

Lemma 3 ([7, 46]). *Let $f \in \Gamma_0(\mathbb{R}^n)$ and $\lambda > 0$. Let the Moreau-Yosida regularization of f be defined by*

$$f_\lambda(x) := \min_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{1}{2\lambda}\|x-u\|^2 \right\}.$$

Then,

- *f_λ is real-valued, convex, and continuously differentiable and can be formulated as*

$$f_\lambda(x) = f(\mathbf{prox}_{\lambda f}(x)) + \frac{1}{2\lambda}\|x - \mathbf{prox}_{\lambda f}(x)\|^2;$$

- *Its gradient*

$$\nabla f_\lambda(x) = \lambda^{-1}(x - \mathbf{prox}_{\lambda f}(x))$$

  is λ^{-1}-Lipschitz continuous.

- $\operatorname{Arg\,min} f_\lambda = \operatorname{Arg\,min} f$ *and* $\min f = \min f_\lambda$.

Now, we are ready to present the result of linear convergence for PPA.

Corollary 6. *Let $f \in \Gamma_0(\mathbb{R}^n)$ achieve its minimum $\min f$, $\Omega \subset \mathbb{R}^n$, and $\lambda > 0$. The PPA can be defined by*

$$x_{k+1} = \mathbf{prox}_{\lambda f}(x_k) = x_k - \lambda \cdot \nabla f_\lambda(x_k),\ \ k \geq 0.$$

Assume that $x_k \in \Omega, k \geq 0$.

(i) If f satisfies the (f, α, Ω)-(obj-EB) condition, then f_λ satisfies the $(\nabla f_\lambda, \nu, \Omega)$-(cor-EB) condition with $\nu = \min\{\frac{\alpha}{4}, \frac{1}{4\lambda}\}$, and hence the PPA converges linearly in the sense that

$$d^2(x_{k+1}, \operatorname{Arg\,min} f) \leq \left(1 - \min\{\frac{\alpha\lambda}{4}, \frac{1}{4}\}\right) \cdot d^2(x_k, \operatorname{Arg\,min} f),\ \ k \geq 0. \tag{21}$$

(ii) Conversely, if starting from arbitrary initial point $x_0 \in \Omega$ the PPA converges linearly like (21) but replacing the rate $1 - \min\{\frac{\alpha\lambda}{4}, \frac{1}{4}\}$ with a constant $\tau \in (0,1)$, then f satisfies the (f, α, Ω)-(obj-EB) condition with $\alpha = \frac{(1-\sqrt{\tau})^2}{2\lambda}$.

Proof. First of all, we remark that

$$\mathbf{crit} f = \operatorname{Arg\,min} f = \operatorname{Arg\,min} f_\lambda = \mathbf{crit} f_\lambda. \tag{22}$$

From Lemma 3, we have $f_\lambda \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ with $L = \lambda^{-1}$ and hence the $(\nabla f_\lambda, \beta, \Omega)$-(cor-res-EB) condition with $\beta = \lambda$ holds. Now, we first prove that the (f, α, Ω)-(obj-EB) condition implies the (f_λ, c, Ω)-(obj-EB) condition with $c = \min\{\frac{\alpha}{2}, \frac{1}{2\lambda}\}$. Indeed, letting $v = \mathbf{prox}_{\lambda f}(x)$ and $v' \in \mathcal{P}_f(v)$, we derive that

$$\begin{aligned}
f_\lambda(x) - \min f_\lambda &= f(\mathbf{prox}_{\lambda f}(x)) + \frac{1}{2\lambda}\|x - \mathbf{prox}_{\lambda f}(x)\|^2 - \min f \\
&\geq \frac{\alpha}{2} d^2(\mathbf{prox}_{\lambda f}(x), \mathbf{crit} f) + \frac{1}{2\lambda}\|x - \mathbf{prox}_{\lambda f}(x)\|^2 \\
&= \frac{\alpha}{2}\|v - v'\|^2 + \frac{1}{2\lambda}\|x - v\|^2 \geq c \cdot (\|v - v'\|^2 + \|x - v\|^2) \\
&\geq \frac{c}{2}(\|v - v'\| + \|x - v\|)^2 \geq \frac{c}{2}\|x - v'\|^2 \geq \frac{c}{2} d^2(x, \mathbf{crit} f_\lambda),
\end{aligned}$$

where the last inequality follows by $v' \in \mathcal{P}_f(v) \subset \mathbf{crit} f = \mathbf{crit} f_\lambda$. From case 1 of Corollary 2, the (f_λ, c, Ω)-(obj-EB) condition implies the $(\nabla f_\lambda, \nu, \Omega)$-(cor-EB) condition with $\nu = \min\{\frac{\alpha}{4}, \frac{1}{4\lambda}\}$. Therefore, (21) follows from Theorem 2 and the fact (22).

Now, we turn to the necessity part. Invoking Theorem 2 again to conclude that f_λ satisfies the $(\nabla f_\lambda, \nu, \Omega)$-(cor-EB) condition with $\nu = \frac{(1-\sqrt{\tau})^2}{\lambda}$, that is

$$\forall\, x \in \Omega \cap \mathrm{dom} f_\lambda, \quad \inf_{x_p \in \mathcal{P}_{f_\lambda}(x)} \langle \nabla f_\lambda(x), x - x_p \rangle \geq \nu \cdot d^2(x, \mathbf{crit} f_\lambda). \tag{23}$$

Together with the fact of $\mathbf{crit} f = \mathbf{crit} f_\lambda$, we can get

$$\forall\, x \in \Omega \cap \mathrm{dom} f_\lambda, \quad \|\nabla f_\lambda(x)\| \geq \nu \cdot d(x, \mathbf{crit} f). \tag{24}$$

On the other hand, using the definition of $v = \mathbf{prox}_{\lambda f}(x)$, which implies $\frac{1}{\lambda}(x - v) \in \partial f(v)$, and the convexity of f, we obtain that

$$\forall\, x \in \mathrm{dom} f, \forall\, g \in \partial f(x), \quad \langle \frac{1}{\lambda}(x - v) - g, v - x \rangle \geq 0, \tag{25}$$

which further implies that

$$\forall\, x \in \mathrm{dom} f, \quad \inf_{g \in \partial f(x)} \|g\| \geq \frac{1}{\lambda}\|x - v\| = \|\nabla f_\lambda(x)\|. \tag{26}$$

Thus, combining (24) and (26) and noting that $\mathrm{dom} f \subset \mathrm{dom} f_\lambda$, we obtain

$$\forall\, x \in \Omega \cap \mathrm{dom} f, \quad \|\partial^0 f(x)\| = \inf_{g \in \partial f(x)} \|g\| \geq \nu \cdot d(x, \mathbf{crit} f). \tag{27}$$

This is just the $(\partial^0 f, \kappa, \Omega)$-(res-EB) condition with $\kappa = \nu$. Therefore, the (f, α, Ω)-(obj-EB) condition with $\alpha = \frac{(1-\sqrt{\tau})^2}{2\lambda}$ holds by case 2 of Corollary 2. □

Remark 2. *Linear convergence of PPA was previously provided based on different EB conditions, such as the Łojasiewicz inequality (corresponding to* (res-obj-EB)*) in [2, 3], the quadratic growth condition (corresponding to* (obj-EB)*) in Proposition 6.5.2 in [9], and the EB condition of Hoffman's type (corresponding to* (res-EB)*) in Theorem 2.1 in [38]. Our novelty here mainly lies in the necessity part, i.e., conclusion (ii).*

Finally, we discuss linear convergence for the forward-backward splitting (FBS) algorithm. Recall that $\mathcal{R}_{1/L}(x) = L\left(x - \mathbf{prox}_{tg}(x - \frac{1}{L}\nabla f(x))\right)$.

Corollary 7. *Let $\varphi = f + g$, where $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$, achieve its minimum $\min \varphi$, and let $\Omega \subset \mathbb{R}^n$. Let the sequence $\{x_k\}_{k \geq 0}$ be generated by FBS, that is*

$$x_{k+1} = \mathbf{prox}_{\frac{1}{L}g}(x_k - \frac{1}{L}\nabla f(x_k)) = x_k - \frac{1}{L} \cdot \mathcal{R}_{1/L}(x_k), \ k \geq 0.$$

Assume that $x_k \in \Omega, k \geq 0$.

(i) If φ satisfies the $(\mathcal{R}_{1/L}, \nu, \Omega)$-(cor-EB) condition with $\nu < 2L$, then FBS converges linearly in the sense that

$$\varphi(x_{k+1}) - \min \varphi \leq (1 - \frac{\nu}{2L})(\varphi(x_k) - \min \varphi), \ k \geq 0. \tag{28}$$

and

$$d^2(x_{k+1}, \mathrm{Arg}\min \varphi) \leq (1 - \frac{\nu}{2L}) \cdot d^2(x_k, \mathrm{Arg}\min \varphi), \ k \geq 0. \tag{29}$$

(ii) Conversely, if starting from arbitrary initial point $x_0 \in \Omega$, FBS converges linearly like (29) *but replacing $1 - \frac{\nu}{2L}$ with τ, then φ satisfies the* $(\mathcal{R}_{1/L}, \nu, \Omega)$-(cor-EB) *condition with $\nu = \frac{L}{2}(1 - \sqrt{\tau})^2$.*

Proof. We rely on the following standard result (see again Lemma 2.3 in [8]):

$$\forall x, y \in \mathbb{R}^n, \quad \langle \mathcal{R}_{1/L}(y), y - x \rangle \geq \varphi(\mathbf{prox}_{\frac{1}{L}g}(y - \frac{1}{L}\nabla f(y))) - \varphi(x) + \frac{1}{2L}\|\mathcal{R}_{1/L}(y)\|^2. \tag{30}$$

Using successively this result at $x = y = x_k$, and then at $y = x_k, x = u_k \in \mathcal{P}_\varphi(x_k)$, together with the fact of $x_{k+1} = \mathbf{prox}_{\frac{1}{L}g}(x_k - \frac{1}{L}\nabla f(x_k))$, we obtain

$$\varphi(x_{k+1}) - \varphi(x_k) \leq -\frac{1}{2L}\|\mathcal{R}_{1/L}(x_k)\|^2, \ k \geq 0, \tag{31}$$

and

$$\varphi(x_{k+1}) - \min \varphi + \frac{1}{2L}\|\mathcal{R}_{1/L}(x_k)\|^2 \leq \langle \mathcal{R}_{1/L}(x_k), x_k - u_k \rangle, \ k \geq 0.$$

Applying the Cauchy-Schwartz inequality to the $(\mathcal{R}_{1/L}, \nu, \Omega)$-(cor-EB) condition, we obtain

$$\forall \ x \in \Omega \cap \mathrm{dom}\varphi, \quad \|\mathcal{R}_{1/L}(x)\| \geq \nu \cdot d(x, \mathbf{crit}\varphi),$$

from which the following inequality follows

$$\langle \mathcal{R}_{1/L}(x_k), x_k - u_k \rangle \leq \frac{1}{\nu}\|\mathcal{R}_{1/L}(x_k)\|^2, \ k \geq 0.$$

Thus, we obtain

$$\varphi(x_{k+1}) - \min \varphi \leq (\frac{1}{\nu} - \frac{1}{2L})\|\mathcal{R}_{1/L}(x_k)\|^2, \ k \geq 0. \tag{32}$$

Combining (31) and (32), we get

$$\varphi(x_{k+1}) - \varphi(x_k) \leq -\frac{1}{2L}\left(\frac{1}{\nu} - \frac{1}{2L}\right)^{-1}(\varphi(x_{k+1}) - \min \varphi), \ k \geq 0,$$

from which the announced result (28) follows.

Now, using the standard result (30) with $x = y_p \in \mathcal{P}_\varphi(y)$ to yield

$$\langle \mathcal{R}_{1/L}(y), y - y_p \rangle \geq \varphi(\mathbf{prox}_{\frac{1}{L}g}(y - \frac{1}{L}\nabla f(y))) - \varphi(y_p) + \frac{1}{2L}\|\mathcal{R}_{1/L}(y)\|^2, \tag{33}$$

and noting that

$$\varphi(\mathbf{prox}_{\frac{1}{L}g}(y - \frac{1}{L}\nabla f(y))) - \varphi(y_p) = \varphi(\mathbf{prox}_{\frac{1}{L}g}(y - \frac{1}{L}\nabla f(y))) - \min \varphi \geq 0,$$

we obtain

$$\forall y \in \mathbb{R}^n, \quad \langle \mathcal{R}_{1/L}(y), y - y_p \rangle \geq \frac{1}{2L}\|\mathcal{R}_{1/L}(y)\|^2.$$

Thus, φ satisfies the $(\mathcal{R}_{1/L}, \beta, \Omega)$-(cor-res-EB) condition with $\beta = \frac{1}{2L}$. Therefore, the remained results follow from Theorem 2 and the fact of $\mathbf{crit}\varphi = \mathrm{Arg}\min \varphi$. □

Remark 3. *The results* (28) *and* (29) *were essentially shown in [19] and [60] respectively, with different methods. Our novelty here lies in conclusion (ii), which was independently also recently observed by the authors in [21].*

We remark that a common assumption employed in Corollaries 5-7 is $x_k \in \Omega, k \geq 0$. Note that each sublevel $X_r = \{x \in \mathbb{R}^n : f(x) \leq r\}$ for any function $f \in \Gamma_0(\mathbb{R}^n)$ is ∂f-invariant, and Observe that $x_k \in X_{\varphi(x_0)}, k \geq 0$ or $x_k \in X_{f(x_0)}, k \geq 0$ for the algorithms studied in Corollaries 5-7. Thereby, if $\Omega = X_{f(x_0)}$ or $\Omega = X_{\varphi(x_0)}$, then the common assumption of $x_k \in \Omega, k \geq 0$, holds trivially; for more details please refer to section 4.2 in [21].

6 Linear convergence of the PALM algorithm

The PALM algorithm was recently introduced by the authors of [11] for a class of composition optimization problem in the general non-convex and non-smooth setting. The authors developed a convergence analysis framework relying on the Kurdyka-Łojasiewicz (KL) inequality and proved that PALM converges globally to a critical point for problem with semi-algebraic data. A global non-asymptotic sublinear rate of convergence of PALM for convex problems was obtained independently in [48] and [25]. Very recently, a globally linear convergence of PALM for strongly convex problems was obtained in [32]. Note that PALM is called block coordinate proximal gradient algorithm in [25] and cyclic block coordinate descent-type method in [32]. In this section, we show linear convergence of PALM under EB conditions, which are strictly weaker than strong convexity.

The following is our main result in this section.

Theorem 3. *Consider the following composite convex nonsmooth minimization problem*

$$\operatorname*{minimize}_{x \in \mathbb{R}^d} \varphi(x) := f(x_1, \cdots, x_p) + \sum_{j=1}^{p} g_j(x_j), \tag{34}$$

where $\mathbb{R}^d \ni x = (x_1, \cdots, x_p)$ *with the* j*-th block* $x_j \in \mathbb{R}^{d_j}$, *and* $d = \sum_{j=1}^{p} d_j$. *Set* $g(x) := \sum_{j=1}^{p} g_j(x_j)$ *so that* $\mathrm{dom} g = \Pi_{j=1}^{p} \mathrm{dom} g_j$. *With these notations, the objective function of* (34) *reads as* $\varphi = f + g$. *Assume that*

- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^d)$, $g_j \in \Gamma_0(\mathbb{R}^{d_j}), j = 1, \cdots, p$, *and* $\Omega \subset \mathrm{dom} g$;
- $f(x_{1:(j-1)}, x_j, x_{(j+1):p}) \in \mathcal{F}_{L_j}^{1,1}(\mathbb{R}^{d_j})$ *for all* $x_{1:(j-1)}$ *and* $x_{(j+1):p}$, $j = 1, \cdots ; p$;
- $\varphi = f + g$ *is such that it achieves its minimum* $\min \varphi$;
- φ *satisfies the* $(\partial^0 \varphi, \eta, \Omega)$-(res-obj-EB) *condition (or its equivalent conditions from case 2 of Corollary 2), which is strictly weaker than strong convexity.*

Here, $L_j, j = 1, \cdots, p$ *and* L *are positive constants, and* $x_{1:k} := (x_1, x_2, \cdots, x_k)$. *Denote* $x_{1:(j-1)}^{(t+1)} := (x_1^{(t+1)}, \cdots, x_{j-1}^{(t+1)})$, $x_{(j+1):p}^{(t)} := (x_{j+1}^{(t)}, \cdots, x_p^{(t)})$, $\psi_j^{(t)}(x_j) := f(x_{1:(j-1)}^{(t+1)}, x_j, x_{(j+1):p}^{(t)})$, *and* $\varphi_j^{(t)}(x_j) := \psi_j^{(t)}(x_j) + g_j(x_j)$. *Start with given initial points* $\{x_j^{(0)}\}_{j=1}^{p}$. *PALM generates* $\{x_j^{(t+1)}\}_{j=1}^{p}$ *via solving a collection of subproblems*

$$x_j^{(t+1)} = \arg\min_{x_j} \left\{ \langle x_j - x_j^{(t)}, \nabla \psi_j^{(t)}(x_j^{(t)}) \rangle + \frac{L_j}{2} \|x_j - x_j^{(t)}\|^2 + g_j(x_j) \right\}, \quad j = 1, \cdots, p, \; t \geq 0.$$

Assume that $x^{(t)} \in \Omega, t \geq 0$. Then, PALM converges linearly in the sense that

$$\varphi(x^{(t+1)}) - \min \varphi \leq \left(\frac{\eta^2 L_{\min}}{4pL^2 + 4L_{\max}^2} + 1 \right)^{-1} (\varphi(x^{(t)}) - \min \varphi), \quad t \geq 0,$$

where $L_{\min} = \min_j L_j$ and $L_{\max} = \max_j L_j$.

Proof. We divide the proof into three steps.

Step 1. We prove that

$$\varphi(x^{(t)}) - \varphi(x^{(t+1)}) \geq \frac{L_{\min}}{2} \|x^{(t)} - x^{(t+1)}\|^2, \ t \geq 0. \tag{35}$$

Let $G_j^{(t)} = L_j(x_j^{(t)} - x_j^{(t+1)})$. By the definition of $x_j^{(t+1)}$ and Lemma 2.3 in [8], we get

$$\varphi_j^{(t)}(x_j^{(t)}) - \varphi_j^{(t)}(x_j^{(t+1)}) \geq \frac{1}{2L_j} \|G_j^{(t)}\|^2 = \frac{L_j^2}{2L_j} \|x_j^{(t)} - x_j^{(t+1)}\|^2 = \frac{L_j}{2} \|x_j^{(t)} - x_j^{(t+1)}\|^2.$$

In addition, note that

$$\sum_{j=1}^{p} \varphi_j^{(t)}(x_j^{(t)}) = \sum_{j=1}^{p} \left(f(x_{1:(j-1)}^{(t+1)}, x_{j:p}^{(t)}) + g_j(x_j^{(t)}) \right)$$

and

$$\sum_{j=1}^{p} \varphi_j^{(t)}(x_j^{(t+1)}) = \sum_{j=1}^{p} \left(f(x_{1:j}^{(t+1)}, x_{(j+1):p}^{(t)}) + g_j(x_j^{(t+1)}) \right).$$

Thus, we derive that for $t \geq 0$,

$$\varphi(x^{(t)}) - \varphi(x^{(t+1)}) = \sum_{j=1}^{p} \varphi_j^{(t)}(x_j^{(t)}) - \sum_{j=1}^{p} \varphi_j^{(t)}(x_j^{(t+1)}) \geq \sum_{j=1}^{p} \frac{L_j}{2} \|x_j^{(t)} - x_j^{(t+1)}\|^2,$$

from which (35) follows.

Step 2. The $(\partial^0 \varphi, \eta, \Omega)$-(res-obj-EB) condition at $x = x^{(t+1)}$ reads as

$$\varphi(x^{(t+1)}) - \min \varphi \leq \frac{\|\partial^0 \varphi(x^{(t+1)})\|^2}{\eta^2}.$$

At the $(t+1)$-th iteration, there exists $\xi_j^{(t+1)} \in \partial g_j(x_j^{(t+1)})$ satisfying the optimality condition:

$$\nabla_j f(x_{1:(j-1)}^{(t+1)}, x_j^{(t)}, x_{(j+1):p}^{(t)}) + L_j(x_j^{(t+1)} - x_j^{(t)}) + \xi_j^{(t+1)} = 0.$$

Here and below, we denote the partial gradient $\nabla_{x_j} f(x)$ by $\nabla_j f(x)$ for notational simplicity. Let $\xi^{(t+1)} = (\xi_1^{(t+1)}, \cdots, \xi_p^{(t+1)})$. Then,

$$\nabla f(x^{(t+1)}) + \xi^{(t+1)} \in \partial \varphi(x^{(t+1)})$$

and hence

$$\varphi(x^{(t+1)}) - \min \varphi \leq \frac{\|\partial^0 \varphi(x^{(t+1)})\|^2}{\eta^2} \leq \frac{\|\nabla f(x^{(t+1)}) + \xi^{(t+1)}\|^2}{\eta^2}.$$

Using the optimality condition and the fact of $f(x) \in \mathcal{F}_L^{1,1}(\mathbb{R}^d)$, we derive that

$$\begin{aligned}
\|\nabla f(x^{(t+1)}) + \xi^{(t+1)}\|^2 =& \sum_{j=1}^{p} \|\nabla_j f(x^{(t+1)}) - \nabla_j f(x^{(t+1)}_{1:(j-1)}, x^{(t)}_j, x^{(t)}_{(j+1):p}) - L_j(x^{(t+1)}_j - x^{(t)}_j)\|^2 \\
\leq& \sum_{j=1}^{p} 2\|\nabla_j f(x^{(t+1)}) - \nabla_j f(x^{(t+1)}_{1:(j-1)}, x^{(t)}_j, x^{(t)}_{(j+1):p})\|^2 + \sum_{j=1}^{p} 2L_j^2\|x^{(t+1)}_j - x^{(t)}_j\|^2 \\
\leq& \sum_{j=1}^{p} 2\|\nabla f(x^{(t+1)}) - \nabla f(x^{(t+1)}_{1:(j-1)}, x^{(t)}_j, x^{(t)}_{(j+1):p})\|^2 + \sum_{j=1}^{p} 2L_j^2\|x^{(t+1)}_j - x^{(t)}_j\|^2 \\
\leq& \sum_{j=1}^{p} 2L^2\|x^{(t+1)}_{j:p} - x^{(t)}_{j:p}\|^2 + \sum_{j=1}^{p} 2L_j^2\|x^{(t+1)}_j - x^{(t)}_j\|^2 \\
\leq& (2pL^2 + 2L_{\max}^2)\|x^{(t+1)} - x^{(t)}\|^2.
\end{aligned}$$

Therefore, we obtain

$$\varphi(x^{(t+1)}) - \min \varphi \leq \frac{(2pL^2 + 2L_{\max}^2)}{\eta^2}\|x^{(t+1)} - x^{(t)}\|^2. \tag{36}$$

Step 3. Combining (35) and (36), we derive that

$$\begin{aligned}
\varphi(x^{(t)}) - \min \varphi =& \left(\varphi(x^{(t)}) - \varphi(x^{(t+1)})\right) + \left(\varphi(x^{(t+1)}) - \min \varphi\right) \\
\geq& \frac{L_{\min}}{2}\|x^{(t)} - x^{(t+1)}\|^2 + \left(\varphi(x^{(t+1)}) - \min \varphi\right) \\
\geq& \left(\frac{\eta^2 L_{\min}}{4pL^2 + 4L_{\max}^2} + 1\right)\left(\varphi(x^{(t+1)}) - \min \varphi\right),
\end{aligned}$$

from which the claimed result follows. This completes the proof. □

On one hand, the $(\varphi, \alpha, \Omega)$-(obj-EB) condition is obviously weaker than strong convexity. On the other hand, we can easily construct functions that satisfies (obj-EB) but fail to be strong convexity. For example, the composition $f(Ax)$, where $f(\cdot)$ is a strongly convex and A is rank deficient, is such a function. This explains why we say that the $(\partial^0\varphi, \eta, \Omega)$-(res-obj-EB) condition, which is equivalent to the $(\varphi, \alpha, \Omega)$-(obj-EB) condition, is strictly weaker than strong convexity

We note that the authors of [6] very recently showed that the regularized Jacobi algorithm-a type of cyclic block coordinate descent method-achieves a linear convergence rate under similar conditions to that of Theorem 3.

7 Linear convergence of Nesterov's accelerated forward-backward algorithm

This section is divided into two parts. In the first part, we first introduce a composition optimization problem, and then we give a new EB condition. In the second part, we introduce the Nesterov's accelerated forward-backward algorithm and show its Q-linear convergence.

7.1 Problem formulation and a new EB condition

Given a nonnegative real sequence $\{r_k\}_{k\geq 0}$. Following the terminology from [41], we say that r_k converges:

- Q-linearly if there exists a constant $\tau \in (0,1)$ such that $\forall k \geq 0$, $r_{k+1} \leq \tau \cdot r_k$,
- R-linearly if there exists a sequence $\{s_k\}_{k\geq 0}$ Q-linearly converging to zero such that $\forall k \geq 0$, $r_k \leq s_k$.

It is well-known that the Nesterov's accelerated gradient method with the following form

$$\begin{cases} y_k &= x_k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k), \end{cases} \tag{37}$$

converges R-linearly for minimizing $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ in the sense that $\{f(x_k) - \min f\}_{k\geq 0}$ converges R-linearly. Very recently, the following Q-linear convergence was independently discovered in [28] and [55] by quite different methods:

$$f(x_{k+1}) - \min f + \frac{\mu}{2}\|w_{k+1} - x^*\|^2 \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(f(x_k) - \min f + \frac{\mu}{2}\|w_k - x^*\|^2\right), \ \forall k \geq 0, \tag{38}$$

where $w_k = (1 + \sqrt{\frac{L}{\mu}})y_k - \sqrt{\frac{L}{\mu}}x_k$. In Nesterov's book [45], via replacing gradient with gradient mapping, the accelerated scheme (37) was successfully extended to solve the following minimization problems:

$$\operatorname*{minimize}_{x\in Q} f(x), \tag{39}$$

and

$$\operatorname*{minimize}_{x\in Q} f(x) := \max_{1\leq i\leq m} f_i(x), \tag{40}$$

where $f, f_i \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n), i = 1, \cdots, m$ and Q is a nonempty closed convex set. Similarly, the accelerated scheme (37) can also be successfully extended to solve

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} \varphi(x) := f(x) + g(x), \tag{41}$$

where $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$. The extended Nesterov's accelerated methods have been proved to achieve R-linear convergence. A natural question arises: Whether there exists Q-linear convergence for the Nesterov's accelerated method applied to problems (39)-(41) as well. In order to study problems (39)-(41) in a unified way, we consider the following composite optimization problem:

$$\operatorname*{minimize}_{x} \varphi(x) := f(e(x)) + g(x). \tag{42}$$

This is a very powerful expression covering many optimization problems, including problems (39)-(41), as its special cases; see [19, 18]. Now, we introduce a new EB condition, commonly satisfied by problems (39)-(41). Our forthcoming argument, will heavily rely on this condition.

Definition 6. *Let $\varphi := f \circ e + g$ be such that $f : \mathbb{R}^m \to \mathbb{R}$ is a closed convex functions, $g \in \Gamma_0(\mathbb{R}^n)$, and $e : \mathbb{R}^n \to \mathbb{R}^m$ is a smooth mapping with its Jacobian given by $\nabla e(x)$. Let $L > 0$ and define*

$$\ell(x; y) := g(x) + f(e(y) + \nabla e(y)(x - y)) + \frac{L}{2}\|x - y\|^2,$$

and

$$p(y) := \arg\min_{x \in \mathbb{R}^n} \ell(x; y),$$
$$G(y) := L(y - p(y)).$$

We say that φ satisfies the composition EB condition with positive constants μ, L obeying $\mu < L$ if

$$\forall x, y \in \mathbb{R}^n, \quad \langle G(y), y - x\rangle \geq \varphi(p(y)) - \varphi(x) + \frac{1}{2L}\|G(y)\|^2 + \frac{\mu}{2}\|x - y\|^2. \tag{43}$$

Let us give several comments on this definition.

Remark 4. *1. Both $p(y)$ and $G(y)$ are well defined due to the strong convexity of $\ell(x; \cdot)$ for any $x \in \mathbb{R}^n$. Moreover, operator G is a residual measure operator related to φ. In fact, observe that the optimality conditions for the proximal subproblem $\mathrm{Arg}\min_{x \in \mathbb{R}^n} \ell(x; y)$ reads as*

$$G(y) \in \partial g(p(y)) + \nabla e(y)^T \partial f(e(y) + \nabla e(y)(p(y) - y)),$$

which implies $y \in \mathbf{crit}\varphi$ if $G(y) = 0$. On the other hand, by the definition of $p(y)$ and using the convexity of g and f, we derive that

$$\begin{aligned}
\varphi(y) &= \ell(y; y) \geq \ell(p(y); y) \\
&= g(p(y)) + f(e(y) + \nabla e(y)(p(y) - y)) + \frac{L}{2}\|p(y) - y\|^2 \\
&\geq (g(y) + \langle z, p(y) - y\rangle) + (f(e(y)) + \langle w, \nabla e(y)(p(y) - y)\rangle) + \frac{L}{2}\|p(y) - y\|^2 \\
&= \varphi(y) + \langle z + \nabla e(y)^T w, p(y) - y\rangle + \frac{L}{2}\|p(y) - y\|^2,
\end{aligned} \tag{44}$$

where $z \in \partial g(y)$ and $w \in \partial f(e(y))$, and hence $z + \nabla e(y)^T w \in \partial\varphi(y)$. Thus, if $0 \in \partial\varphi(y)$, then we can take some $z \in \partial g(y)$ and $w \in \partial f(e(y))$ such that $z + \nabla e(y)^T w = 0$. Hence, the inequality (44) implies that $G(y) = 0$ if $y \in \mathbf{crit}\varphi$. Therefore, we have $\{x \in \mathbb{R}^n : G(y) = 0\} = \mathbf{crit}\varphi$, i.e., G is a residual measure operator related to φ.

2. The composition EB condition can be viewed as a relaxation of strong convexity to some degree. This perspective is in the same spirit of work [39]. Indeed, in case of $m = 1$, $g(x) \equiv 0$, $f(t) \equiv t, t \in \mathbb{R}$, and $e \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, (43) reads as

$$\forall x, y \in \mathbb{R}^n, \quad e(x) \geq \left(e(y - \frac{1}{L}\nabla e(y)) + \frac{1}{2L}\|\nabla e(y)\|^2\right) + \langle \nabla e(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2. \tag{45}$$

On the other hand, $e \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ implies that

$$\forall x, y \in \mathbb{R}^n, \quad e(y) \geq e(y - \frac{1}{L}\nabla e(y)) + \frac{1}{2L}\|\nabla e(y)\|^2.$$

Therefore, (45) *is a relaxation of strong convexity in the following form:*

$$\forall x, y \in \mathbb{R}^n, \quad e(x) \geq e(y) + \langle \nabla e(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2.$$

In the case of $f \circ e(x) \equiv 0$ *and* $g \in \Gamma(\mathbb{R}^n)$, (43) *reads as*

$$\forall x, y \in \mathbb{R}^n, \quad g(x) \geq g_\lambda(y) + \langle \nabla g_\lambda(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2, \tag{46}$$

where $\lambda = \frac{1}{L}$. *Recall that* g_λ *is the Moreau-Yosida regularization of* g *and note that* $g(x) \geq g_\lambda(x)$. *We can see that* (46) *is a relaxation of strong convexity of* g_λ.

3. *Although we have shown that* (43) *can be viewed as a relaxation of strong convexity, it is still a very strong property. Now, we construct an example to show that even strongly convex property of* f *is not enough to ensure* (43) *to hold. This example is obtained by setting* $n = m = 2$, $x = (x_1, x_2)^T$, $e(x) = (x_1, x_1)^T$, $f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$, $g(x) \equiv 0$*; then* $\varphi(x) = f \circ e(x) = x_1^2$. *It is obviously to see that* f *is strongly convex. Let us show that in this special case* (43) *fails to hold. Actually, after some simple calculus, we can get*

$$p(y) = \begin{pmatrix} \frac{L}{L+2}y_1 \\ y_2 \end{pmatrix}, \qquad G(y) = \begin{pmatrix} \frac{2L}{L+2}y_1 \\ 0 \end{pmatrix},$$

and therefore (43) *reads as*

$$\begin{aligned}\frac{L}{L+2}y_1(y_1 - x_1) \geq &(\frac{L}{L+2}y_1)^2 - x_1^2 + \frac{2L}{(L+2)^2}y_1^2 \\ &+ \frac{\mu}{2}(x_1 - y_1)^2 + \frac{\mu}{2}(x_2 - y_2)^2, \quad \forall x_i, y_i \in \mathbb{R}, i = 1, 2.\end{aligned}$$

But, if we take $x_1 = y_1 \equiv 0$, *then it should have*

$$0 \geq \frac{\mu}{2}(x_2 - y_2)^2, \quad \forall x_2, y_2 \in \mathbb{R}.$$

Obviously, this is impossible for any positive constant μ.

4. *Let* $A \in \mathbb{R}^{m\times n}$ *with* $m < n$ *be a given matrix and* $b \in \mathbb{R}^m$ *be a given vector. A well-known fact in the community of EB is that the quadratic function* $\frac{1}{2}\|Ax - b\|^2$ *is not strongly convex but satisfies EB conditions. Unfortunately, this function fails to satisfy* (45). *We show this point by contradiction. It is enough to consider* $e(x) = \frac{1}{2}x^T aa^T x$ *with* $\|a\|^2 = L$. *In this case,* (45) *reads as*

$$\frac{1}{2}(a^T x - a^T y)^2 \geq \frac{\mu}{2}\|x - y\|^2, \ \forall x, y \in \mathbb{R}^n.$$

Let $h \neq 0$ *be an orthogonal vector of* a. *Now, take* $y - x = \lambda h, \lambda \in \mathbb{R}$. *Then, we have*

$$0 \geq \frac{\mu}{2}\lambda^2\|h\|^2, \ \forall \lambda \in \mathbb{R},$$

which is impossible for any positive constant μ*; otherwise, the right-hand side can be arbitrarily large.*

5. *In order to show that* (46) *can be strictly weaker than strong convexity, we now construct a one-dimensional example that satisfies* (46) *but fails to be strongly convex. Define the shrinkage operator by* $\mathcal{S}(t) := \mathrm{sign}(t) \cdot \max\{|t| - 1, 0\}$ *and the projection operator by* $[x]_I^+ := \arg\min_{y \in I} \|x - y\|$*, where* I *is some closed interval. Now, we take* $\lambda = 1$*,* $I = [-2, 2]$*, and* $g(x) = |x| + \delta_I(x)$*. Obviously, such* $g(x)$ *is convex but not strongly convex. Using formula (14) in [62] and Lemma 3, we have*

$$g_\lambda(x) = |[\mathcal{S}(x)]_I^+| + \frac{1}{2}(x - [\mathcal{S}(x)]_I^+)^2.$$

Here, g_λ *is the Moreau-Yosida regularization of* g*. Denote* $\ell_{g_\lambda}(x; y) := g_\lambda(y) + \langle \nabla g_\lambda(y), x - y \rangle$*. We have the following expression:*

$$\ell_{g_\lambda}(x; y) = \begin{cases} (y+2)x - \frac{1}{2}y^2 + 4, & y \le -3, \\ -x - \frac{1}{2}, & -3 \le y \le -1, \\ yx - \frac{1}{2}y^2, & -1 \le y \le 1, \\ x - \frac{1}{2}, & 1 \le y \le 3, \\ (y-2)x - \frac{1}{2}y^2 + 4, & y \ge 3 \end{cases} \tag{47}$$

Then, one can verify case by case that for any $\mu \in (0, \frac{1}{9}]$*,* (46) *always holds. For example, in the case of* $y \le -3$*, we only need to verify that*

$$|x| \ge (y+2)x - \frac{1}{2}y^2 + 4 + \frac{\mu}{2}(x - y)^2, \quad x \in [-2, 2],$$

i.e., $\frac{1-\mu}{2}(x - y)^2 \ge \frac{1}{2}x^2 + 2x - |x| + 4, \quad x \in [-2, 2]$*. Thus, it is sufficient to require that*

$$\mu \le 1 - \max_{y \le -3, |x| \le 2} \frac{x^2 + 4x - 2|x| + 8}{(x - y)^2}.$$

After some simple calculus, we have $\mu \in (0, \frac{1}{9}]$*. The other cases can be similarly verified; we omit the details here. This example shows that the composition EB condition indeed holds for some non-strongly convex functions.*

Now, we explain why we say that the condition (43) is commonly satisfied by problems (39)-(41).

Remark 5. *(i) The minimization problem* (42) *with* $m = 1$*,* $e(x) \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$*,* $f(t) \equiv t, t \in \mathbb{R}$*,* $g(x) = \delta_Q(x)$*, and* Q *being nonempty closed convex, corresponds to problem* (39)*. The condition* (43) *holds in this setting; see Theorem 2.2.7 in [40].*

(ii) The minimization problem (42) *with* $f(y) = \max_{1 \le i \le m}\{y_i\}$*,* $f_i(x) \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$*,* $e(x) = (f_1(x), f_2(x), \cdots, f_m(x))$*,* $g(x) = \delta_Q(x)$*, and* Q *being nonempty closed convex, corresponds to problem* (40)*. The condition* (43) *holds in this setting; see Corollary 2.3.2 in [40].*

(iii) The minimization problem (42) *with* $m = 1$*,* $e(x) \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$*,* $f(t) \equiv t, t \in \mathbb{R}$*, and* $g(x) \in \Gamma_0(\mathbb{R}^n)$*, corresponds to problem* (41)*. The condition* (43) *holds in this setting; see the inequality (4.36) in [15].*

In general, we have to admit that it is difficulty to verify the composition EB condition, which therefore deserves further study in the future.

7.2 Q-linear convergence of the Nesterov's acceleration

In this part, we show Q-linear convergence of the Nesterov's acceleration under the composition EB condition, which is more general than strong convexity. First, in light of the Nesterov accelerated scheme (2.2.11) in [40], the Nesterov's accelerated forward-backward algorithm for solving the problem (42) reads as: choosing $x_{-1} = x_0 \in \mathbb{R}^n$, for $k \geq 0$,

$$\begin{cases} y_k &= x_k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L}G(y_k). \end{cases}$$

Let

$$\alpha = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}},\ \beta = \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}},\ \gamma = \frac{1}{2L}(1+\sqrt{\frac{L}{\mu}}).$$

Let

$$\Phi_k(x^*;\tau) := \varphi(x_k) - \min\varphi + \tau\cdot\|z_k - x^*\|^2,\ k \geq 0,$$

where $x^* \in \operatorname{Arg\,min}\varphi$ (assumed to be nonempty) and

$$z_k = \frac{1}{2}(1+\sqrt{\frac{L}{\mu}})y_k + \frac{1}{2}(1-\sqrt{\frac{L}{\mu}})x_k,\ k \geq 0.$$

Now, we are ready to present the main result in this section. The proof idea behind is partially inspired by the argument in [5] but might be of interest in its own right.

Theorem 4. *Let $\varphi := f\circ e+g$ be such that $f:\mathbb{R}^m\to\mathbb{R}$ is a closed convex functions, $g\in\Gamma_0(\mathbb{R}^n)$, and $e:\mathbb{R}^n\to\mathbb{R}^m$ is a smooth mapping with its Jacobian given by $\nabla e(x)$. Let φ satisfy the composition EB condition with positive constants μ, L obeying $\mu < L$. Assume that φ achieves its minimum $\min\varphi$ so that $\operatorname{Arg\,min}\varphi \neq \emptyset$. Then, there exist a unique vector x^* such that $\operatorname{Arg\,min}\varphi = \{x^*\}$, and the Nesterov's accelerated forward-backward method converges Q-linearly in the sense that there exists a positive constant $\theta_0 < 1$ such that for any $\theta\in[\theta_0, 1)$ it holds*

$$\Phi_{k+1}(x^*;\tau) \leq \rho\cdot\Phi_k(x^*;\tau),\ k\geq 0, \tag{48}$$

where $\rho = \max\{\alpha,\theta\} < 1$ and $\tau = \frac{\theta\beta}{2\rho\gamma}$. Especially, by taking $\theta = \max\{\theta_0,\alpha\}$, we have

$$\Phi_{k+1}(x^*;\frac{2L\mu}{(\sqrt{L}+\sqrt{\mu})^2}) \leq \max\{\theta_0,\alpha\}\cdot\Phi_k(x^*;\frac{2L\mu}{(\sqrt{L}+\sqrt{\mu})^2}),\ k\geq 0. \tag{49}$$

Proof. We first show the uniqueness of optimal solution x^* of φ. In fact, by statement (i) in Remark 4 and the fact of $\operatorname{Arg\,min}\varphi\subset \mathbf{crit}\varphi$, we have that $G(x^*)=0$ and $p(x^*)=x^*$, and hence (43) at $y=x^*$ reads as

$$\varphi(x) - \min\varphi \geq \frac{\mu}{2}\|x-x^*\|^2,\quad \forall x\in\mathbb{R}^n,$$

which clearly implies that $\operatorname{Arg\,min}\varphi = \{x^*\}$.

Now, we analyze rates of linear convergence. Using successively (43) at $x=x_k$ and $y=y_k$, and then at $y=y_k$ and $x=x^*$, together with the fact of $x_{k+1}=p(y_k)$, we obtain

$$\varphi(x_{k+1}) \leq \varphi(x_k) + \langle G(y_k), y_k - x_k\rangle - \frac{1}{2L}\|G(y_k)\|^2 - \frac{\mu}{2}\|x_k - y_k\|^2$$

and

$$\varphi(x_{k+1}) \leq \varphi(x^*) + \langle G(y_k), y_k - x^* \rangle - \frac{1}{2L}\|G(y_k)\|^2 - \frac{\mu}{2}\|x^* - y_k\|^2.$$

Multiplying the first inequality by α and the second one by β, and then adding the two resulting inequalities, we obtain

$$\begin{aligned}\varphi(x_{k+1}) \leq &\alpha\varphi(x_k) + \beta\varphi(x^*) + \langle G(y_k), \alpha(y_k - x_k) + \beta(y_k - x^*) \rangle \\ &- \frac{1}{2L}\|G(y_k)\|^2 - \frac{\mu\alpha}{2}\|x_k - y_k\|^2 - \frac{\mu\beta}{2}\|x^* - y_k\|^2.\end{aligned}$$

In order to estimate the right-hand side of the inequality above, we first write down:

$$\alpha(y_k - x_k) + \beta(y_k - x^*) = \beta(z_k - x^*). \tag{50}$$

Secondly, using the expression of $y_{k+1} = x_{k+1} + \alpha(x_{k+1} - x_k)$, we get

$$z_{k+1} = \frac{1}{2}(1 + \sqrt{\frac{L}{\mu}})x_{k+1} + \frac{1}{2}(1 - \sqrt{\frac{L}{\mu}})x_k. \tag{51}$$

Then, substitute $x_{k+1} = y_k - \frac{1}{L}G(y_k)$ into formula (51) to obtain

$$z_{k+1} - x^* = z_k - x^* - \gamma \cdot G(y_k). \tag{52}$$

Using equality (52), we derive that

$$\begin{aligned}\langle G(y_k), z_k - x^* \rangle &= \frac{1}{\gamma}\langle z_k - x^* - (z_{k+1} - x^*), z_k - x^* \rangle \\ &= \frac{1}{\gamma}\|z_k - x^*\|^2 - \frac{1}{\gamma}\langle z_{k+1} - x^*, z_k - x^* \rangle \\ &= \frac{1}{\gamma}\|z_k - x^*\|^2 - \frac{1}{\gamma}\langle z_{k+1} - x^*, z_{k+1} - x^* + \gamma \cdot G(y_k) \rangle \\ &= \frac{1}{\gamma}\|z_k - x^*\|^2 - \frac{1}{\gamma}\|z_{k+1} - x^*\|^2 - \langle z_{k+1} - x^*, G(y_k) \rangle \\ &= \frac{1}{\gamma}\|z_k - x^*\|^2 - \frac{1}{\gamma}\|z_{k+1} - x^*\|^2 - \langle G(y_k), z_k - x^* \rangle + \gamma\|G(y_k)\|^2.\end{aligned}$$

Thus, we have

$$\langle G(y_k), z_k - x^* \rangle = \frac{1}{2\gamma}(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2) + \frac{\gamma}{2}\|G(y_k)\|^2. \tag{53}$$

Combining formula (53) and formula (50), we derive that

$$\begin{aligned}\varphi(x_{k+1}) \leq &\alpha\varphi(x_k) + \beta\varphi(x^*) + \frac{\beta}{2\gamma}(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2) \\ &+ (\frac{\beta\gamma}{2} - \frac{1}{2L})\|G(y_k)\|^2 - \frac{\mu\alpha}{2}\|x_k - y_k\|^2 - \frac{\mu\beta}{2}\|x^* - y_k\|^2 \\ =&\alpha\varphi(x_k) + \beta\varphi(x^*) + \frac{\beta}{2\gamma}(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2) - \frac{\mu\alpha}{2}\|x_k - y_k\|^2 - \frac{\mu\beta}{2}\|x^* - y_k\|^2,\end{aligned}$$

where the term $\|G(y_k)\|^2$ is eliminated since $\frac{\beta\gamma}{2} = \frac{1}{2L}$. Note that (50) can be written as

$$z_k - x^* = (y_k - x^*) + \frac{1}{2}(\sqrt{\frac{L}{\mu}} - 1)(y_k - x_k),$$

with which we further derive that

$$\begin{aligned}\|z_k - x^*\|^2 \leq & 2\|x^* - y_k\|^2 + \frac{1}{2}(\sqrt{\frac{L}{\mu}} - 1)^2\|y_k - x_k\|^2 \\ & \leq \max\left\{2, \frac{1}{2}(\sqrt{\frac{L}{\mu}} - 1)^2\right\}(\|x^* - y_k\|^2 + \|y_k - x_k\|^2).\end{aligned}$$

Denote $\eta_1 := \min\left\{\frac{\mu\alpha}{2}, \frac{\mu\beta}{2}\right\}$ and $\eta_2 := \max\left\{2, \frac{1}{2}(\sqrt{\frac{L}{\mu}} - 1)^2\right\}$. Then, we have

$$\begin{aligned}\varphi(x_{k+1}) &\leq \alpha\varphi(x_k) + \beta\varphi(x^*) + \frac{\beta}{2\gamma}(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2) - \eta_1(\|x^* - y_k\|^2 + \|y_k - x_k\|^2) \\ &\leq \alpha\varphi(x_k) + \beta\varphi(x^*) + \frac{\beta}{2\gamma}(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2) - \frac{\eta_1}{\eta_2}\|z_k - x^*\|^2.\end{aligned}$$

Rearrange the terms to obtain

$$\varphi(x_{k+1}) - \varphi(x^*) + \frac{\beta}{2\gamma}\|z_{k+1} - x^*\|^2 \leq \alpha(\varphi(x_k) - \varphi(x^*)) + (\frac{\beta}{2\gamma} - \frac{\eta_1}{\eta_2})\|z_k - x^*\|^2.$$

Thus, there exists a positive constant $\theta_0 < 1$ such that for any $\theta \in [\theta_0, 1)$ it holds

$$\varphi(x_{k+1}) - \varphi(x^*) + \frac{\beta}{2\gamma}\|z_{k+1} - x^*\|^2 \leq \alpha(\varphi(x_k) - \varphi(x^*)) + \frac{\theta\beta}{2\gamma}\|z_k - x^*\|^2.$$

Since $\rho = \max\{\alpha, \theta\}$, we have that $\rho < 1$ and $\frac{\theta}{\rho} \leq 1$. Thus, we obtain

$$\begin{aligned}\varphi(x_{k+1}) - \varphi(x^*) + \frac{\theta\beta}{2\rho\gamma}\|z_{k+1} - x^*\|^2 &\leq \alpha(\varphi(x_k) - \varphi(x^*)) + \frac{\theta\beta}{2\gamma}\|z_k - x^*\|^2 \\ &\leq \rho\left(\varphi(x_k) - \varphi(x^*)) + \frac{\theta\beta}{2\rho\gamma}\|z_k - x^*\|^2\right),\end{aligned}$$

i.e., $\Phi_{k+1}(x^*; \tau) \leq \rho \cdot \Phi_k(x^*; \tau)$ with $\tau = \frac{\theta\beta}{2\rho\gamma}$. This is just the announced result (48).

It remains to show (49). In fact, if $\theta = \max\{\theta_0, \alpha\}$, then $\rho = \max\{\alpha, \theta\} = \max\{\theta_0, \alpha\} = \theta$ and hence

$$\tau = \frac{\theta\beta}{2\rho\gamma} = \frac{\beta}{2\gamma} = \frac{2L\mu}{(\sqrt{L} + \sqrt{\mu})^2}.$$

This completes the proof. □

Remark 6. *It should be noted that we here only show the existence of rates of linear convergence for the Nesterov's accelerated forward-backward method. But, we are not clear whether one can derive an exact rate of linear convergence as* $1 - \sqrt{\frac{\mu}{L}}$ *as obtained for the Nesterov's accelerated gradient method.*

8 A class of dual functions satisfying EB conditions

Verifying EB conditions for functions with certain structure is a difficulty topic. In this section, we consider a class of dual objective functions, that have achieved many interesting applications in signal processing and compressive sensing [63, 30]. We first describe the problem, along with some direct conclusions.

Proposition 2. *Consider the linearly constrained optimization problem*

$$\underset{y\in\mathbb{R}^m}{\text{minimize}}\, g(y), \quad \textit{subject to} \quad Ay = b, \tag{P}$$

where $g : \mathbb{R}^m \to \mathbb{R}$ is a real-valued closed and strongly convex function with modulus $c > 0$, $A \in \mathbb{R}^{n\times m}$ is a given matrix, and $b \in R(A)$ is a given vector. Here, $R(A)$ stands for the range of A. The dual problem is

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x) := g^*(A^T x) - \langle b, x\rangle. \tag{D}$$

Then, we have that

- *the primal problem* (P) *has a unique optimal solution $\bar{y}$,*
- *the dual objective function f belongs to $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ with $L = \frac{\|A\|^2}{c}$, and*
- *the set of optimal solutions of the dual problem,*

  $$\operatorname{Arg\,min} f := \{x \in \mathbb{R}^n : A\nabla g^*(A^T x) = b\},$$

  is a nonempty convex closed set, and can be characterized by $\{x \in \mathbb{R}^n : A^T x \in \partial g(\bar{y})\}$ or equivalently by $\{x \in \mathbb{R}^n : \nabla g^*(A^T x) = \bar{y}\}$.

Proof. The first two statements are standard results which can be found in textbooks on convex analysis and no proof will be given here. Now, we prove the third statement. First, let the Lagrangian function be given by $L(y, x) = g(y) - \langle Ay - b, x\rangle$. By the assumption of $b \in R(A)$ and the finiteness of the optimal value of the primal problem, according to Proposition 5.3.3 in [9], for any $\bar{x} \in \operatorname{Arg\,min} f$ we have that $\bar{y} \in \operatorname{Arg\,min} L(y, \bar{x})$. Hence, $A^T\bar{x} \in \partial g(\bar{y})$ or equivalently $\nabla g^*(A^T\bar{x}) = \bar{y}$ due to $(\partial g)^{-1} = \nabla g^*$, which holds by Corollary 23.5.1 in [44]. This implies that $\operatorname{Arg\,min} f \subseteq \{x \in \mathbb{R}^n : \nabla g^*(A^T x) = \bar{y}\}$. The inverse inclusion is obvious since $A\bar{y} = b$. Thereby,

$$\operatorname{Arg\,min} f = \{x \in \mathbb{R}^n : \nabla g^*(A^T x) = \bar{y}\} = \{x \in \mathbb{R}^n : A^T x \in \partial g(\bar{y})\}.$$

This completes the proof. □

Now, we state the main result of this section.

Theorem 5. *Use the same setting as Proposition 2. Denote $X_r := \{x \in \mathbb{R}^n : f(x) \le \min f + r\}$, and $V_r := A^T X_r$. If the following assumptions hold:*

(a) ∂g is calm around $\bar{y}$ for any $\bar{z} \in V_0$,

(b) the collection $\{\partial g(\bar{y}), R(A^T)\}$ is linearly regular with constant $\gamma > 0$, that is

$$d(A^T x, \partial g(\bar{y})) \ge \gamma \cdot d(A^T x, \partial g(\bar{y}) \cap R(A^T)), \quad \forall x \in \mathbb{R}^n,$$

then we have that

(i) *there exist positive constants* r_0, τ *such that the* (f, τ, X_{r_0})-(obj-EB) *condition holds, that is*

$$f(x) - \min f \geq \frac{\tau}{2} \cdot d^2(x, \mathbf{crit} f), \quad \forall x \in X_{r_0}. \tag{54}$$

Specifically, if ∂g *is calm with constant* $\kappa > 0$ *around* $\bar{y}$ *for any* $\bar{z} \in V_0$*, then* (54) *holds for all* $\tau \in (0, \kappa^{-1})$.

(ii) *For any sublevel set* X_r*, pick* $r_1 \in (0, r_0)$ *and let* $c_r := \sqrt{\frac{r_1}{r}}$ *and*

$$\rho_r := \begin{cases} c_r, & \text{when } r \geq r_0, \\ 1, & \text{when } r \leq r_0. \end{cases}$$

Then, the $(\nabla f, \nu, X_r)$-(cor-EB) *condition with* $\nu = \frac{\tau \rho_r^2}{8}$ *holds.*

Proof. For simplicity, denote $\mathbb{R}^m$ by $\mathcal{E}$. The proof is divided into four steps.

Step 1. First, we prove that $V_0 = A^T X_0$ is compact. In fact, we have shown that $X_0 = \{x \in \mathbb{R}^n : A^T x \in \partial g(\bar{y})\}$ in Proposition 2. Hence, $V_0 = A^T X_0 \subseteq \partial g(\bar{y})$. Since g is a real-valued convex function, $\partial g(\bar{y})$ must be nonempty and bounded according to Theorem 23.4 in [44] or Proposition 5.4.2 in [9]. Therefore, $V_0 = A^T X_0$ is bounded and hence compact because of the closedness of X_0. Since ∂g is calm at $\bar{y}$ for any $\bar{z} \in V_0$ and $V_0 \subseteq \partial g(\bar{y})$ is compact, by Proposition 2 in [65] we can conclude that there exist constants $\kappa, \epsilon > 0$ such that

$$\partial g(y) \cap (V_0 + \epsilon \mathbb{B}_{\mathcal{E}}) \subseteq \partial g(\bar{y}) + \kappa \cdot \|y - \bar{y}\|_2 \mathbb{B}_{\mathcal{E}}, \quad \forall y \in \mathcal{E}. \tag{55}$$

Let $r > 0$ be small enough such that

$$A^T X_r = V_r \subseteq V_0 + \epsilon \mathbb{B}_{\mathcal{E}}.$$

Pick $x \in X_r$ and let $y = \nabla g^*(A^T x)$. Then, $A^T x \in \partial g(y)$ due to $\partial g = (\nabla g^*)^{-1}$ and hence $A^T x \in \partial g(y) \cap (V_0 + \epsilon \mathbb{B}_{\mathcal{E}})$. By the inclusion (55), we obtain

$$d(A^T x, \partial g(\bar{y})) \leq \kappa \|y - \bar{y}\|_2 = \kappa \cdot d(\bar{y}, \nabla g^*(A^T x)), \quad \forall x \in X_r. \tag{56}$$

Step 2. Let $z = A^T x$ and note that $\partial g = (\nabla g^*)^{-1}$. The inequality (56) can be written as

$$d(z, (\nabla g^*)^{-1}(\bar{y})) \leq \kappa \cdot d(\bar{y}, \nabla g^*(z)), \quad \forall z \in V_r.$$

This implies that ∇g^* is always metrically subregular at each $\bar{z} \in V_0$ for $\bar{y}$ because that V_r is a neighborhood for every $\bar{z} \in V_0$. Thus, by Theorem 3.1 in [20], for each $\bar{z} \in V_0$ there exists a neighborhood $\bar{z} + \epsilon(\bar{z}) \mathbb{B}_{\mathcal{E}}$ and a positive constant $\alpha(\bar{z})$ such that

$$g^*(z) \geq g^*(\bar{z}) - \langle \bar{y}, \bar{z} - z \rangle + \frac{\alpha(\bar{z})}{2} \cdot d^2(z, (\nabla g^*)^{-1}(\bar{y})), \quad \forall z \in \mathcal{E} \text{ with } \|z - \bar{z}\|_2 \leq \epsilon(\bar{z}), \tag{57}$$

where the constant $\alpha(\bar{z})$ can be chosen arbitrarily in $(0, \kappa^{-1})$. Note that $\{\bar{z} + \epsilon(\bar{z}) \mathbb{B}_{\mathcal{E}}^o\}_{\bar{z} \in V_0}$ forms an open cover of the compact set V_0. Hence, by the Heine-Borel theorem, there exist K points (where $K \geq 1$ is finite) $\bar{z}_1, \cdots, \bar{z}_K \in V_0$ such that

$$V_0 \subseteq U := \bigcup_{i=1}^{K} (\bar{z}_i + \epsilon(\bar{z}_i) \mathbb{B}_{\mathcal{E}}^o).$$

Let $\alpha = \min\{\alpha(\bar{z}_1), \cdots, \alpha(\bar{z}_K)\}$, which can be chosen arbitrarily in $(0, \kappa^{-1})$, and note that $\min f = g^*(\bar{z}) - \langle \bar{y}, \bar{z}\rangle,\ \forall \bar{z} \in V_0$. By the relationship (57), we thus get

$$g^*(z) - \langle \bar{y}, z\rangle \geq \min f + \frac{\alpha}{2} \cdot d^2(z, (\nabla g^*)^{-1}(\bar{y})), \quad \forall z \in U.$$

Finally, letting r be small enough such that $V_r \subseteq U$ and using again the fact of $(\nabla g^*)^{-1} = \partial g$, we obtain

$$g^*(z) - \langle \bar{y}, z\rangle \geq \min f + \frac{\alpha}{2} \cdot d^2(z, \partial g(\bar{y})), \quad \forall z \in V_r,$$

or equivalently,

$$f(x) - \min f \geq \frac{\alpha}{2} \cdot d^2(A^T x, \partial g(\bar{y})), \quad \forall x \in X_r.$$

Step 3. Using the linear regular property of $\{\partial g(\bar{y}), R(A^T)\}$, we derive that

$$\begin{aligned} d(A^T x, \partial g(\bar{y})) \geq \gamma \cdot d(A^T x, \partial g(\bar{y}) \cap R(A^T)) &= \min_{A^T u \in \partial g(\bar{y})} \|A^T x - A^T u\| \\ &= \min_{u \in X_0} \|A^T x - A^T u\| = \|A^T x - A^T \hat{x}\|, \end{aligned}$$

where such $\hat{x} \in X_0$ exists since X_0 is a nonempty closed set. Now, we follow the argument in [47] to finish the proof of (i). Denote the null space of A^T by $N(A^T)$ and the minimal positive singular value of A by $\sigma(A)$. Note that $\operatorname{Arg\,min} f + N(A^T) \subseteq \operatorname{Arg\,min} f$. We derive that

$$d(x, \operatorname{Arg\,min} f) \leq \|x - (\hat{x} + \mathcal{P}_{N(A^T)}(x - \hat{x}))\| \leq \frac{1}{\sigma(A)} \|A^T x - A^T \hat{x}\| \leq \frac{d(A^T x, \partial g(\bar{y}))}{\sigma(A)},$$

where $\mathcal{P}_{N(A^T)}$ stands for the orthogonal projection operator onto $N(A^T)$. Note that $\operatorname{Arg\,min} f = \mathbf{crit} f$. Thereby, the (obj-EB) condition follows with $\tau = \alpha \cdot \sigma^2(A)$.

Step 4. Let us prove (ii). Without loss of generality, we assume that $\min f = 0$ and $r \geq r_0$. Since for any $r > 0$ the sublevel set X_r is ∇f-invariant, using (54) and together with the equivalence established in Corollary 2, we can conclude that f satisfies the $(\nabla f, \eta, X_{r_0})$-(res-obj-EB) conditions with $\eta = \sqrt{\frac{\tau}{2}}$, that is

$$\forall\, x \in X_{r_0}, \quad \|\nabla f(x)\| \geq \eta \cdot \sqrt{f(x)}. \tag{58}$$

Let $\varphi(t) := 2\eta^{-1} t^{\frac{1}{2}}$. Then, the property (58) can be written as

$$\forall\, x \in X_{r_0}, \quad \|\nabla f(x)\| \varphi'(f(x)) \geq 1. \tag{59}$$

By applying Proposition 30 in [10], a globalization result for KL inequalities, to (59), we have that for the given $r_1 \in (0, r_0)$, the function given by

$$\phi(t) := \begin{cases} \varphi(t), & \text{when } \ t \leq r_1, \\ \varphi(r_1) + (t - r_1)\varphi'(r_1), & \text{when } \ t \geq r_1, \end{cases}$$

is desingularising for f on all of $\mathbb{R}^n$ and hence it holds

$$\forall\, x \in X_r, \quad \|\nabla f(x)\| \phi'(f(x)) \geq 1. \tag{60}$$

Thereby, we can get

$$\|\nabla f(x)\| \geq \eta \sqrt{r_1}, \quad \forall x \in X_r \cap X_{r_1}^c,$$

where $X_{r_1}^c$ is the complementary set of X_{r_1}. By the definition of c_r, we can further obtain

$$\|\nabla f(x)\| \geq \eta c_r \sqrt{r} \geq \eta c_r \sqrt{f(x)}, \quad \forall x \in X_r \cap X_{r_1}^c.$$

Finally, noting that $c_r < 1$ and together with (58), we have

$$\|\nabla f(x)\| \geq \eta c_r \sqrt{f(x)}, \quad \forall x \in X_r,$$

which is just the $(\nabla f, \eta c_r, X_r)$-(res-obj-EB) condition for $r \geq r_0$, and hence the $(\nabla f, \eta \rho_r, X_r)$-(res-obj-EB) condition holds for $r > 0$. Thus, the $(\nabla f, \nu, X_r)$-(cor-EB) condition follows from Corollary 2. The parameter ν can be established by using the relevant formulas in Theorem 1. This completes the proof. □

Remark 7. *By directly invoking Corollary 4.3 in [1], we can derive* (57) *with the constant satisfying* $\alpha(\bar{z}) \in (0, \frac{1}{4\kappa})$*, which is slightly worse than that of* $\alpha(\bar{z}) \in (0, \kappa^{-1})$*.*

Remark 8. *The author of [47], with slightly different assumptions, proved by contradiction that the dual objective function* $f(x) = g^*(A^T x) - \langle b, x \rangle$ *satisfies the* $(\nabla f, \nu, X_r)$-(cor-EB) *condition. While the author of [47] requires that* ∂g *is calm around* $\bar{y}$ *for any* $\bar{z} \in \mathbb{R}^m$*, i.e. the locally upper Lipschitz-continuous property* (4)*, we only require that* ∂g *is calm around* $\bar{y}$ *for any* $\bar{z} \in V_0$*, and additionally assume that g is real-valued closed, which holds trivially for all the cases listed in Example 2.10. in [47]. Moreover, our proof is by means of the KL inequality globalization technique developed in [10], and hence quite different from that of [47].*

Remark 9. *Verifying EB conditions for more general functions with the form* $f(x) := h(Ax) + l(x)$ *was studied recently in [19, 65, 31]. Specialized to the dual objective function* $f(x) = g^*(A^T x) - \langle b, x \rangle$*, existing theory usually requires* g^* *to be strictly or strongly convex; see e.g. Corollary 4.3 in [19] and Assumption 1 in [65]. In contrast, our study, following the research line of work [47], relies on exploiting the prime-dual structure, and hence quite different from that in [19, 65, 31].*

9 Discussion

In this paper, we provide a new perspective for studying EB conditions and analyzing linear convergence of gradient-type methods. Under our theoretical framework, a group of new technical results are discovered. Especially, some EB conditions, previously known to be sufficient for linear convergence, are also necessary; and the Nesterov's forward-backward algorithm, previously known to be R-linearly convergent, are also Q-linearly convergent. Finally, we complete this paper with the following possible future works:

1. We have defined a group of abstract EB conditions of "square type". But we do not know whether the idea behind can be extended to that of general types by introducing so-called desingularizing functions [10], so that the other EB conditions discussed in [21] can be included in a more general framework.

2. Although we have shown sufficient conditions guaranteeing linear convergence for PALM and Nesterov's accelerated forward-backward algorithms, it is still unclear whether they are necessary. The very recent work [36] might shed light on this topic.

3. Verifying EB conditions with *high probability* for non-convex functions has proven to be a very powerful approach for non-convex optimization; see e.g. [14, 53, 33]. Thus, seeking or verifying new classes of non-convex functions, satisfying EB condition with high-probility, deserves future study.

4. What are the optimal rates of linear convergence (or say, exact worst-case convergence rates) for gradient-type methods under general EB conditions? The method of performance estimation, originally proposed in [17] and further developed in [29, 52, 51], might be useful for this topic.

5. The ordinary differential equation (ODE) approaches are recently used to study (accelerated) gradient-type methods [50, 55]. Except one paper [57], existing analyses only consider general convex and strongly convex conditions, and do not work on general EB conditions. We wonder whether the EB condition presented in this paper can be embedded in the ODE approaches to study linear convergence for gradient-type methods. This is also interesting for future work.

Acknowledgements

I am grateful to the anonymous referees, the associate editor, and the coeditor Prof. Adrian S. Lewis for many useful comments, which allowed me to significantly improve the original presentation. I would like to thank Prof. Zaiwen Wen for his invitation and hospitality during my visit to BeiJing International Center for Mathematical Research, and to thank Prof. Dmitriy Drusvyatskiy for a careful reading of an early draft of this manuscript, and for valuable comments and suggestions. I also thank Profs. Chao Ding, Bin Dong, Lei Guo, Yongjin Liu, Deren Han, Mark Schmidt, Anthony Man-Cho So, and Wotao Yin for their time and many helpful discussions with me. Further thanks due to my cousin Boya Ouyang who helped me with my English writing, and to PhD students Ke Guo, Wei Peng, Ziyang Yuan, Xiaoya Zhang, who looked over the manuscript and corrected several typos. This work is supported by the National Science Foundation of China (No.11501569 and No.61571008).

References

[1] F. J. A. Artacho and M. H. Geoffroy. Metric subregularity of the convex subdifferential in banach spaces. *J. Nonlinear Convex A.*, 15(15):35–47, 2015.

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program. Ser. B*, 116:5–16, 2009.

[3] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Math. Program. Ser. A*, 137(1-2):91–129, 2013.

[4] H. Attouch and A. Cabot. Convergence rates of inertial forward-backward algorithms. <hal-01453170>, 2017.

[5] H. Attouch and J. Peypouquet. The rate of convergence of nesterov's accelerated forward-backard method is actually $o(k^{-2})$. *SIAM J. Optim.*, 26(3):1824–1834, 2016.

[6] G. Banjac, K. Margellos, and P. J. Goulart. On the convergence of a regularized jacobi algorithm for convex optimization. manuscript, 2016.

[7] H. H. Bauschke and P. L. Combettes. Convex analysis and monotone operator theory in hilbert spaces. *Springer Berlin*, 2011.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.*, 17(4):183–202, 2009.

[9] D. P. Bertsekas. *Convex optimization theory.* Athena Scienfitic and Tsinghua University Press, 2011.

[10] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program. Ser. A*, doi:10.1007/s10107-016-1091-6, 2016.

[11] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program., Ser. A*, 146(1-2):459–494, 2014.

[12] H. Brézis. Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de hilbert. *North-Holland Mathematics studies 5*, 1973.

[13] R. E. Bruck. Asymptotic convergence of nonlinear contraction semigroups in hilbert space. *J. Funct. Anal.*, 18(1):15–26, 1975.

[14] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[15] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numer.*, 25:161–319, 2016.

[16] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings.* Springer, 2009.

[17] Y. Drori and M. Teboulle. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Math. Program., Ser. A*, 145(1-2):451–482, 2014.

[18] D. Drusvyatskiy and C. Kempton. An accelerated algorithm for minimizaing convex compositions. *arXiv:1605.00125v1[math.OC] 30 Apr 2016.*

[19] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv:1602.06661v1 [math.OC] 22 Feb 2016.*

[20] D. Drusvyatskiy, B. S. Mordukhovich, and T. T. A. Nghia. Second-order growth, tilt stability, and metric regularity of the subdifferential. *J. Convex Anal.*, 21(4):1165–1192, 2014.

[21] G. Garrigos, L. Rosasco, and S. Villa. Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. *arXiv:1703.09477v2 [math.OC] 29 Mar 2017.*

[22] P. Gong and J. Ye. Linear convergence of variance-reduced projected stochastic gradient without strong convexity. *arXiv:1406.1102v2 [cs.NA] 10 Jul 2015.*

[23] E. T. Hale, W. Yin, and Y. Zhang. Fixed point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130, 2008.

[24] A. Hoffman. On approximate solutions of systems of linear inequalities. *J.Res.Natl. Bureau Standards*, 49:263–265, 1952.

[25] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo. Iteration complexity analysis of block coordinate descent methods. *Math. Program., Ser. A*, 163:85–114, 2017.

[26] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of proximal-gradient methods under the Polyak-Łojasiewicz condition. *arXiv:1608.04636v1 [cs.LG] 16 Aug 2016.*, 2016.

[27] H. Karimi and M. Schmidt. Linear convergence of proximal-gradient methods under the Polyak-Łojasiewicz condition. *NIPS OPT*, 2015.

[28] S. Karimi and S. Vavasis. A unified convergence bound for conjugate gradient and accelerated gradient. *arXiv:1605.003200v1[math.OC] 11 May 2016.*

[29] D. Kim and J. Fessler. Optimized first-order methods for smooth convex minimization. *Math. Program., Ser. A*, 159(1):81–107, 2016.

[30] M. J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM J. Imaging Sci.*, 6(2):1059–1091, 2013.

[31] G. Li and T. K. Pong. Calculus of the exponent of kurdyka-lojasiewicz inequality and its applications to linear convergence of first-order methods. *aeXiv:1602.02915v1 [math.OC] 9 Feb 2016.*

[32] X. Li, T. Zhao, R. Arora, H. Liu, and M. Hong. An improved convergence analysis of cyclic block coordinate descent-type methods for strongly convex minimization. *arXiv:1607.02793v1[math.OC] 10 Jul 2016.*

[33] H. Liu, M.-C. Yue, and A. M.-C. So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *manuscript, 2016.*

[34] J. Liu and S. J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM J. Optim.*, 30:351–376, 2015.

[35] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Colloques internationaux du CNRS*, 117:87–89, 1963.

[36] D. R. Luke, N. H. Thao, and M. Teboulle. Necessary conditions for linear convergence of picard iterations and application to alternating projections. *arXiv:1704.08926v1 [math.OC] 28 Apr 2017.*

[37] Z. Q. Luo and P. Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. In *SIAM J. Control Optim*, pages 408–425, 1992.

[38] F. Luque. Asymptotic convergence analysis of the proximal point algorithm. *SIAM J. Optim.*, 22(2):277–293, 1984.

[39] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv:1504.06298v2 [math.OC] 23 Apr 2015.*

[40] Y. Nesterov. *Introductory lectures on convex optimization: A basic course.* Kluwer Academic Publishers, 2004.

[41] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer-Verlag, New York, 1997.

[42] B. T. Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[43] S. M. Robinson. Some continuity properties of polyhedral multifunctions. *Math. Program. Study*, 14:206–214, 1981.

[44] R. T. Rockafellar. *Convex analysis.* Princeton university press, 1970.

[45] R. T. Rockafellar and R. J. B. Wets. *Variational analysis.* Springer, 1998.

[46] A. P. Ruszczyński. *Nonlinear optimization*, volume 13. Princeton university press, 2006.

[47] F. Schöpfer. Linear convergence of descent methods for the unconstrained minimization of restricted strongly convex functions. *SIAM J. Optim.*, 26(3):1883–1911, 2016.

[48] R. Shefi and M. Teboulle. On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. *EURO J Comput. Optim.*, 4:27–46, 2016.

[49] M. C. So. Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. *arXiv:1309.0113v1 [math.OC] 31 Aug 2013.*

[50] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling nesterovs accelerated gradient method: theory and insights. *J. Mach. Learn. Res.*, 17:1–13, 2016.

[51] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *arXiv:1705.04398v1 [math.OC] 11 May 2017.*

[52] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Math. Program., Ser. A*, 161(1-2):307–345, 2017.

[53] S. Tu, R. Boczar, M. Simchowitz, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv:1507.03566v2[math.OC] 5 Feb 2016.*

[54] P. W. Wang and C. J. Lin. Iteration complexity of feasible descent methods for convex optimization. *J. Mach. Learn. Res.*, 15(2):1523–1548, 2014.

[55] A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv:1611.02635v1 [math.OC] 8 Nov 2016.*

[56] L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM J. Optim.*, 23(2):1062–1091, 2013.

[57] L. Yang, R. Arora, V. Braverman, and T. Zhao. The physical systems behind optimization algorithms. *arXiv:1612.02803v1 [cs.LG] 8 Dec 2016.*

[58] H. Zhang. Linear convergence of the proximal incremental aggregated gradient method under quadratic growth condition. *arXiv:1702.08166v1 [math.OC] 27 Feb 2017.*

[59] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *arXiv:1606.00269v3 [math.OC] 17 Aug 2016.*

[60] H. Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. *Optim. Lett.*, 11(4):817–833, 2017.

[61] H. Zhang and L. Cheng. Restricted strong convexity and its applications to convergence analysis of gradient-type methods. *Optim. Lett.*, 9(5):961–979, 2015.

[62] H. Zhang and L. Cheng. Projected shrinkage algorithm for box-constrained ℓ_1-minimization. *Optim. Lett.*, 11(1):55–70, 2017.

[63] H. Zhang, L. Cheng, and W. Yin. A dual algorithm for a class of augmented convex signal recovery models. *Commun. Math. Sci.*, 13(1):103–112, 2015.

[64] H. Zhang and W. T. Yin. Gradient methods for convex minimization: better rates under weaker conditions. Technical report, CAM Report 13-17, UCLA, 2013.

[65] Z. Zhou and M. C. So. A unified approach to error bounds for structured convex optimization problems. *Math. Program., Ser. A*, DOI 10.1007/s10107-016-1100-9, 2017.

[66] T. Zolezzi. On equiwellset minimum problems. *Appl. Math. Optim*, 4:203–223, 1978.