Safeguarding adaptive methods: global convergence of Barzilai-Borwein and other stepsize choices

Ou, Hongjia
Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University

Themelis, Andreas
Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University

出版情報：2024 10th International Conference on Control, Decision and Information Technologies (CoDIT), pp.2802-2807, 2024-10-18. Institute of Electrical and Electronics Engineers (IEEE)
バージョン：

Safeguarding adaptive methods: global convergence of Barzilai-Borwein and other stepsize choices

Hongjia Ou and Andreas Themelis

***Abstract*— Leveraging on recent advancements on adaptive methods for convex minimization problems, this paper provides a linesearch-free proximal gradient framework for globalizing the convergence of popular stepsize choices such as Barzilai-Borwein and one-dimensional Anderson acceleration. This framework can cope with problems in which the gradient of the differentiable function is merely locally Hölder continuous. Our analysis not only encompasses but also refines existing results upon which it builds. The theory is corroborated by numerical evidence that showcases the synergetic interplay between fast stepsize selections and adaptive methods.**

I. Introduction

Convex nonsmooth optimization problems are encountered in various engineering applications such as image denoising [4], signal processing and digital communication [16], machine learning [7], and control [15], to name a few. Many such problems can be cast in composite form as

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper and lower semicontinuous (lsc). A textbook algorithm in this setting is the proximal gradient method that involves updates of the form

$$x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k)), \tag{1}$$

where

$$\operatorname{prox}_g(x) := \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{g(w) + \tfrac{1}{2}\|w - x\|^2\right\} \tag{2}$$

is the *proximal mapping* of g, available in closed form for many practical applications [3].

In (1), the choice of the *stepsize* parameters γ_{k+1} plays a pivotal role in dictating the efficiency of the algorithm. Traditional constant stepsizes require the gradient of the function f to satisfy global Lipschitz continuity [5, Prop. 1.2.3], and stepsizes that are either too large or too small can lead to slow convergence or failure thereof. Alternatively, the utilization of backtracking linesearch for stepsize determination [6, §9.3] provides an online tuning that works under less restrictive assumptions, but incurs computational overhead due to the inner subroutines for assessing each stepsize.

A revolutionary concept was conceived in [17], which proposes an adaptive method for automatically tuning the stepsizes without backtracking routines to address scenarios of convex smooth optimization, that is, when $g = 0$ in (1), yet under mere *local* (as opposed to *global*) Lipschitz gradient continuity. This pioneering work has been attracting considerable attention in the last couple of years, leading to proximal (and primal-dual) variants [13], [18], [12].

The present paper continues this trend, inspired by two recent interesting developments: [20] which reduces the assumptions on ∇f to *local Hölder continuity*, and [23] which incorporates Barzilai-Borwein stepsizes [2] to enhance convergence speed. The Barzilai-Borwein can be classified as a one-dimensional quasi-Newton method, and its convergence without backtracking routines has only been established for smooth and strongly convex quadratic problems [21], [10], [14]. A first attempt beyond the quadratic case was advanced in [8] with the stabilized BB method, which provides a suitable dampening of the BB stepsizes ensuring convergence for strongly convex and smooth problems. However, convergence is established only for (undefined) small enough choices of a parameter, precluding an off-the-shelf use in practice.

The above-mentioned [23] obviates this problem by designing a tailored algorithm, adaPBB, that carefully integrates the adaptive ideas of [17] and successive works. The method is an alternative way of "stabilizing" BB stepsizes, promoting larger stepsizes (on average) which are associated with faster convergence (see the next section).

The present paper patterns the same rationale and proposes a simpler "stabilizing" approach, not necessarily tied to the choice of BB stepsizes, that builds upon the recent findings of [20] to apply to a much broader class of problems. Our working assumptions on problem (P) are the following:

Assumption I. *The following hold in problem* (P)*:*

A1 $f : \mathbb{R}^n \to \mathbb{R}$ *is convex and has locally Hölder-continuous gradient of order* $\nu \in (0, 1]$ *(see §III).*

A2 $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper, lsc, and convex.*

A3 *A solution exists:* $\operatorname{arg\,min}\varphi \neq \emptyset$.

Although $\nu > 0$ is ultimately required in the analysis, in some results that are valid in the limiting case $\nu = 0$ we will explicitly mention the extended range $\nu \in [0, 1]$. As formally elaborated in [20], the case $\nu = 0$ amounts to f being merely convex and real-valued, and the notation ∇f is used to indicate any *subgradient* selection $\nabla f(x) \in \partial f(x)$.

A. Contributions

As a means to globalize in a linesearch-free fashion a virtually arbitrary variety of stepsize choices, we employ adaptive methods as *"safeguarding"* mechanisms to automatically

Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University, 744 Motooka, Nishi-ku 819-0395, Fukuoka, Japan.
Email: ou.honjia.069@s.kyushu-u.ac.jp, andreas.themelis@ees.kyushu-u.ac.jp

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI grants JP21K17710 and JP24K20737.

dampen, when necessary, overshooting stepsizes that may cause divergence. Our focus is on adaPG$^{q,\frac{q}{2}}$ developed in [12], as it constitutes an umbrella framework enabling us to encompass many other adaptive methods at once.

To embrace the generality of Assumption I, we provide a "convergence recipe" that identifies which are the stepsize choices that can be *safeguarded*, including, but not limited to, Barzilai-Borwein [2] and Anderson acceleration–like choices [1], [11]. Even when specialized to the plain adaPG$^{q,\frac{q}{2}}$ algorithm, this general view results in tighter constants for the convergence rate compared to the tailored analysis of [20].

Finally, we provide a list of stepsize choices that comply with our framework, and provide numerical evidence in support of their employment.

B. Paper organization

In §II we provide a brief overview on adaptive methods and build on the motivating examples that led to this work; the section provides a gradual walkthrough to the idea and the rationale of *"safeguarding"*. The technical core of the paper is in §III, where the general convergence recipe for a class of adaptive methods is presented and validated with formal proofs. Some stepsize choices compliant with the recipe are listed in §IV, and §V presents numerical experiments followed by some closing remarks.

II. Adaptive methods as safeguards

In contrast to the employment of constant parameters based on global (worst-case) Lipschitz moduli, adaptive methods generate stepsizes at every iteration from *local* estimates based on *past* information, thereby completely waiving any need of inner subroutines. At iteration k, the new stepsize γ_{k+1} is retrieved based on the following quantities:

$$\ell_k := \frac{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1}\rangle}{\|x^k - x^{k-1}\|^2}, \tag{3a}$$

$$L_k := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}, \tag{3b}$$

and/or

$$c_k := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2}{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1}\rangle}. \tag{3c}$$

When ∇f is *locally* Lipschitz, [12] shows that the update

$$\gamma_{k+1} := \min\left\{\gamma_k\sqrt{\tfrac{1}{q}+\tfrac{\gamma_k}{\gamma_{k-1}}}, \frac{\gamma_k}{\sqrt{2\left[\gamma_k^2 L_k^2 - (2-q)\gamma_k\ell_k + 1 - q\right]_+}}\right\} \tag{4}$$

for some $q \in [1,2]$ guarantees, for convex problems, convergence of (1) to a solution ($[\cdot]_+ := \max\{\cdot, 0\}$). The remarkable performance in practice owes to two factors: the ability to adapt to the *local* geometry of f, and the consequent employment of much larger stepsizes than textbook constant "worst-case" choices would yield. It is worth noting that the analysis of [12] and related references still holds if equality in (4) is replaced by a "$\leq$", as long as the stepsizes remain bounded away from zero. This fact was used to demonstrate that the adaptive schemes [17], [13], [18] proposed in previous works could be covered by the same theory, despite the update in (4) may be larger. Nevertheless, as the sublinear worst-case rate

$$\min_{k\leq K}\varphi(x^k) - \min\varphi \leq \frac{C}{\sum_{k=1}^{K+1}\gamma_k}$$

for some $C > 0$ confirms, see Fact 2*(3)*, larger stepsizes are typically associated with faster convergence, suggesting that selecting "=" in (4) would be preferable.

A. Motivating examples: Barzilai-Borwein stepsizes

A first motivation for this work is the observation that the above argument may be shortsighted. To gain an intuition, consider $q = 1$ in (4) and assume that ∇f is globally L_f-Lipschitz continuous for simplicity; as shown in [12], a lower bound $\gamma_k \geq \frac{1}{\sqrt{2}L_f}$ can easily be derived. This estimate can in fact be tightened by considering the cases in which the stepsize increases or not: using $\ell_k \leq L_k \leq c_k \leq L_f$ and $\ell_k c_k = L_k^2$ [13, Lem. 2.1] it is easy to see that

$$\gamma_{k+1} \geq \gamma_k \quad \vee \quad \gamma_k \geq \tfrac{1+\sqrt{3}}{2c_k}$$

holds for every k. By loosening the update into, say,

$$\gamma_{k+1} = \min\left\{\gamma_k\sqrt{1+\tfrac{\gamma_k}{\gamma_{k-1}}}, \frac{\gamma_k}{\sqrt{2\left[\gamma_k^2 L_k^2 - \gamma_k\ell_k\right]_+}}, \tfrac{1}{\ell_k}\right\}, \tag{5a}$$

a similar comparison yields

$$\gamma_{k+1} \geq \gamma_k \quad \vee \quad \gamma_k \geq \tfrac{3}{2c_k} \quad \vee \quad \gamma_k \geq \gamma_{k+1} = \tfrac{1}{\ell_k}.$$

Similarly, reducing the last term as

$$\gamma_{k+1} = \min\left\{\gamma_k\sqrt{1+\tfrac{\gamma_k}{\gamma_{k-1}}}, \frac{\gamma_k}{\sqrt{2\left[\gamma_k^2 L_k^2 - \gamma_k\ell_k\right]_+}}, \tfrac{1}{c_k}\right\} \tag{5b}$$

benefits the second case, with new lower bounds given by

$$\gamma_{k+1} \geq \gamma_k \quad \vee \quad \gamma_k \geq \tfrac{2}{c_k} \quad \vee \quad \gamma_k \geq \gamma_{k+1} = \tfrac{1}{c_k}.$$

Interestingly, (5) amount to *dampened* versions of the celebrated *long* and *short Barzilai-Borwein* (BB) stepsizes $1/\ell_k$ and $1/c_k$ [2]. It is easy to see that through the updates (1) and (5) the generated stepsize sequences are bounded away from zero; then, since γ_{k+1} in both is dominated by the right-hand side of (4) (with $q = 1$), the analysis of [12] directly applies, resulting in a simple update rule to make BB stepsizes globally convergent, for possibly nonsmooth convex problems.

This is the sense in which we say that adaPG$^{q,\frac{q}{2}}$ acts as a *"safeguard"* for BB stepsizes, namely in providing a suitable dampening to trigger global convergence. These claims will be substantiated with a formal proof addressing a richer choice of stepsizes spanning beyond long and short BB.

B. The locally Hölder case

As a matter of fact, the (locally) Lipschitz differentiable case is a no-brainer consequence of [12, Thm. 1.1]. Nevertheless, as the numerical evidence demonstrates, such a simple change in the update rule can have a strong impact. A deeper theoretical contribution in this paper is provided by the *locally Hölder* analysis which builds upon the recent developments of [20]. In this setting, the theory of [20] cannot be invoked, since the proofs therein are tightly linked to the update rule (4), complicated by possibly vanishing stepsizes. The proposed solution is a convergence recipe prescribing two conditions: (i) do not overshoot the update (4)

of adaPG$^{q,\frac{q}{2}}$, and (ii) control the stepsize from below based on a past fixed-point residual. Surprisingly, not only does this solution prove that adaPG$^{q,\frac{q}{2}}$ can *safeguard* even in the locally Hölder regime, but it also improves the results in [20] for adaPG$^{q,\frac{q}{2}}$ itself by providing tighter rate constants.

III. Convergence of adaptive methods revisited

In what follows we adopt the same notation of [13]:

$$\rho_k := \tfrac{\gamma_k}{\gamma_{k-1}}, \quad P_k := \varphi(x^k) - \min\varphi, \quad P_k^{\min} := \min_{i\le k} P_i. \tag{6}$$

Moreover, following [20], for $\nu \in [0,1]$ we introduce

$$\lambda_{k,\nu} := \frac{\gamma_k}{\|x^k - x^{k-1}\|^{1-\nu}} \tag{7}$$

acting as a *scaled* stepsize, and the local Hölder estimates

$$\ell_{k,\nu} := \ell_k\|x^k - x^{k-1}\|^{1-\nu} \text{ and } L_{k,\nu} := L_k\|x^k - x^{k-1}\|^{1-\nu}. \tag{8}$$

When ∇f is locally ν-Hölder continuous, for any compact convex set $\Omega \subset \mathbb{R}^n$ there exists a ν-Hölder modulus $L_{\Omega,\nu} > 0$ for ∇f on Ω, namely such that

$$\|\nabla f(x) - \nabla f(y)\| \le L_{\Omega,\nu}\|x - y\|^\nu \quad \forall x, y \in \Omega.$$

Then, as long as $(x^k)_{k\in\mathbb{N}}$ remains in a bounded and convex set Ω, one has that [20, Eq. (13)]

$$\ell_{k,\nu} \le L_{k,\nu} \le L_{\Omega,\nu}. \tag{9}$$

A. A convergence recipe

We now list the fundamental ingredients of the convergence analysis. After stating some preliminary lemmas, we will prove that the given items are all is needed for ensuring convergence of proximal gradient iterations in the generality of Assumption I. In the following §IV we will translate these requirements into conditions on stepsize oracles that can be *safeguarded*, followed by some notable examples.

Convergence recipe
Let a sequence $(x^k)_{k\in\mathbb{N}}$ be generated by proximal gradient iterations (1) with stepsizes $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$, and let ℓ_k and L_k be as in (3). We say that $(x^k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ satisfy Property P1 (resp. P2/P3) if there exist $\nu \in (0,1]$, $q \in [1,2]$ and $\lambda_{\min,\nu} > 0$ such that, for all $k \in \mathbb{N}$:

- P1 $1 + q\rho_k - q\rho_{k+1}^2 \ge 0$
- P2 $\frac{1}{2} - \rho_{k+1}^2\big[\gamma_k^2 L_k^2 - \gamma_k\ell_k(2-q) + 1 - q\big] \ge 0$
- P3 either $\gamma_{k+1} \ge \gamma_k$ or there exists and index $j_k \le k$ such that $\min\{\gamma_{j_k}, \gamma_{k+1}\} \ge \lambda_{\min,\nu}\|x^{j_k} - x^{j_k-1}\|^{1-\nu}$.

Properties P1 and P2 dictate that the stepsize should not overshoot the one in (4): adaPG$^{q,\frac{q}{2}}$ shall act as *safeguard*. As stated in Fact 2, these two alone already ensure boundedness of the generated iterates x^k. The remaining Property P3 demands a bound from below: for $\nu = 1$, it is tantamount to requiring $(\gamma_k)_{k\in\mathbb{N}}$ bounded away from zero; more generally, it involves a bound on a *scaled* stepsize $\lambda_{i,\nu}$, cf. (7), akin to that observed in [20, Lem. 3.6]; see Lemma 3. The turning point lies in allowing full flexibility for the index $j_k \le k$ in Property P3, enabling us to safeguard stepsizes that depend on a window of past iterations, as will be showcased in §IV.

B. Preliminary results on adaptive methods

Fact 1 ([20, Lem. 3.3]). *Let f and g be convex, $q > 0$ and $x^\star \in \arg\min\varphi$. Relative to $(x^k)_{k\in\mathbb{N}}$ generated by proximal gradient iterations (1) with stepsizes $\gamma_k > 0$ $\forall k \in \mathbb{N}$, let*

$$\mathcal{U}_k^q(x^\star) := \tfrac{1}{2}\|x^k - x^\star\|^2 + \tfrac{1}{2}\|x^k - x^{k-1}\|^2 + \gamma_k(1 + q\rho_k)P_{k-1}. \tag{10}$$

Then, for every $k \ge 1$ and with ℓ_k and L_k as in (3),

$$\begin{aligned}\mathcal{U}_{k+1}^q(x^\star) \le \mathcal{U}_k^q(x^\star) &- \gamma_k\big(1 + q\rho_k - q\rho_{k+1}^2\big)P_{k-1} \\ &- \Big\{\tfrac{1}{2} - \rho_{k+1}^2\big[\gamma_k^2 L_k^2 - \gamma_k\ell_k(2-q) \\ &\quad + 1 - q\big]\Big\}\|x^k - x^{k-1}\|^2.\end{aligned} \tag{11}$$

Fact 2 ([20, Lem. 3.4]). *Let Assumption I hold with $\nu \in [0,1]$, $x^\star \in \arg\min\varphi$, and $(x^k)_{k\in\mathbb{N}}$ be generated by proximal gradient iterations (1) with stepsizes $\gamma_k > 0$ $\forall k \in \mathbb{N}$. If $(x^k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ satisfy Properties P1 and P2, then:*

1. *$(\mathcal{U}_k^q(x^\star))_{k\in\mathbb{N}}$ decreases and converges to a finite value.*
2. *The sequence $(x^k)_{k\in\mathbb{N}}$ is bounded and admits at most one optimal limit point.*
3. *$P_K^{\min} \le \mathcal{U}_1^q(x^\star)/\big(\sum_{k=1}^{K+1}\gamma_k\big)$ for any $K \ge 1$.*

Proof. The proof is identical to that of [20, Lem. 3.4], after observing that Properties P1 and P2 together amount to

$$\gamma_{k+1} \le \gamma_k \min\left\{\sqrt{\tfrac{1}{q} + \tfrac{\gamma_k}{\gamma_{k-1}}}, \frac{1}{\sqrt{2\big[\gamma_k^2 L_k^2 - (2-q)\gamma_k\ell_k + 1 - q\big]_+}}\right\}.$$

□

Lemma 3 (compliance of adaPG$^{q,\frac{q}{2}}$). *Let Assumption I hold with $\nu \in [0,1]$. Then, $(x^k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ generated by adaPG$^{q,\frac{q}{2}}$ ((1) and (4)) satisfy Properties P1 to P3 with*

$$j_k \in \{k-1, k\} \quad \text{and} \quad \lambda_{\min,\nu} = \tfrac{1}{\sqrt{2q}L_{\Omega,\nu}},$$

where $L_{\Omega,\nu}$ is a ν-Hölder modulus for ∇f on a compact and convex set Ω that contains all the iterates x^k.

Proof. Compliance with Properties P1 and P2, as well as the existence of such an Ω and the consequent finiteness of $L_{\Omega,\nu}$, follows from Fact 2. Recall in particular that $\ell_{k,\nu} \le L_{k,\nu} \le L_{\Omega,\nu}$ holds for every k, cf. (9). Suppose that $\gamma_{k+1} < \gamma_k$ at some iteration k, and let $K_i \subseteq \mathbb{N}$ denote the set of iterates k for which the i-th element in the minimum of (4) is active, $i = 1, 2$. We have two cases:

♢ If $k \in K_2$, then

$$\begin{aligned}\gamma_k > \gamma_{k+1} &= \frac{\gamma_k}{\sqrt{2\big[\gamma_k^2 L_k^2 - (2-q)\gamma_k\ell_k - (q-1)\big]}} \\ &\ge \tfrac{1}{\sqrt{2}L_k} = \frac{\|x^k - x^{k-1}\|^{1-\nu}}{\sqrt{2}L_{k,\nu}} \ge \frac{\|x^k - x^{k-1}\|^{1-\nu}}{\sqrt{2}L_{\Omega,\nu}}.\end{aligned} \tag{12}$$

Noticing that $\frac{1}{\sqrt{2}L_{\Omega,\nu}} \ge \lambda_{\min,\nu}$ (since $q \ge 1$), apparently the claimed Property P3 holds with $j_k = k$.

♢ If $k \in K_1$ (necessarily $q > 1$), then $1 > \rho_{k+1}^2 = \frac{1}{q} + \rho_k$, which yields that $\rho_k < 1 - \frac{1}{q} < 1$ and that consequently $\gamma_{k+1} < \gamma_k < \gamma_{k-1}$. Necessarily $k - 1 \in K_2$, for otherwise, denoting $t = q^{-1} \ge \frac{1}{2}$, it would hold that $\rho_{k+1} = \sqrt{t + \rho_k} = \sqrt{t + \sqrt{t + \rho_{k-1}}} \ge \sqrt{1/2 + 1/\sqrt{2}} > 1$. Thus,

$$\gamma_{k-1} > \gamma_{k+1} = \gamma_k\sqrt{\tfrac{1}{q} + \rho_k} \ge \tfrac{1}{\sqrt{q}}\gamma_k \overset{(12)}{>} \frac{\|x^{k-1} - x^{k-2}\|^{1-\nu}}{\sqrt{2q}L_{\Omega,\nu}}.$$

In this case, the same conclusion holds with $j_k = k-1$. □

C. Convergence analysis

We now present the main result of this paper, which generalizes and improves upon the analysis of [20]. To minimize overlapping arguments, proofs of intermediate claims that are verbatim identical are deferred to the reference.

Theorem 4. *Let Assumption 1 hold (with $\nu > 0$), and let $(x^k)_{k\in\mathbb{N}}$ be generated by proximal gradient iterations (1). If all Properties P1 to P3 hold for some $q \in [1,2]$ and $\lambda_{\min,\nu} > 0$, then $(x^k)_{k\in\mathbb{N}}$ converges to some $x^\star \in \arg\min\varphi$ with*

$$P_K^{\min} \leq \max\left\{\frac{\mathcal{U}_1^q(x^\star)}{\gamma_0(K+1)}, \frac{2^{\frac{1-\nu}{2}}\mathcal{U}_1^q(x^\star)^{\frac{1+\nu}{2}}(1+\lambda_{\min,\nu}L_{\Omega,\nu})^{1-\nu}}{\lambda_{\min,\nu}(K+1)^\nu}\right\}$$

for every $K \geq 1$, where $\mathcal{U}^q(x^\star)$ is as in (10) and $L_{\Omega,\nu}$ is a ν-Hölder modulus for ∇f on a convex and compact set Ω that contains all the iterates x^k.

Proof. The assumptions of Fact 2 are met, and therefore all the claims therein hold. In particular, the existence of a ν-Hölder modulus $L_{\Omega,\nu}$ for ∇f on a convex and compact set Ω that contains all the iterates x^k is guaranteed. Properties P1 and P2 ensure that the coefficients of P_{k-1} and $\|x^k - x^{k-1}\|^2$ in (11) are negative. A telescoping argument then yields that

$$\sum_{k\geq 1}\gamma_k\left(1 + q\rho_k - q\rho_{k+1}^2\right)P_{k-1} < \infty. \tag{13}$$

We next proceed by intermediate steps. In what follows, we denote $K^\uparrow := \{k \in \mathbb{N} \mid \gamma_{k+1} \geq \gamma_k\}$ and $K^\downarrow := \mathbb{N} \setminus K^\uparrow$.

♢ *Claim 4(a):* $\inf_{k\in\mathbb{N}} P_k = 0$, *and* $(x^k)_{k\in\mathbb{N}}$ *admits a (unique) optimal limit point.*

Boundedness of $(x^k)_{k\in\mathbb{N}}$ and at most uniqueness of the optimal limit point follow from Fact 2*(2)*. By lower semicontinuity of φ, in order to show existence it suffices to prove that $\inf_{k\in\mathbb{N}} P_k = 0$. If $\gamma_k \not\to 0$, then $\sum_k \gamma_k = \infty$ and we know from Fact 2*(3)* that $P_k \to 0$. Suppose instead that $\gamma_k \to 0$, so that in particular $K^\downarrow$ must be infinite. If $K^\uparrow$ is finite, then $1 + q\rho_k - q\rho_{k+1}^2 \not\to 0$, for otherwise $(\gamma_k)_{k\in\mathbb{N}}$ would be eventually linearly increasing. If $K^\uparrow$ is infinite, then so is $\tilde{K} := \left\{k \in K^\downarrow \mid k-1 \in K^\uparrow\right\}$ with $\rho_k \geq 1$ and $\rho_{k+1} < 1$ for all $k \in \tilde{K}$; in particular, $1 + q\rho_k - q\rho_{k+1}^2 \geq 1$ for all $k \in \tilde{K}$. Also in this case we conclude that $0 \leq 1 + q\rho_k - q\rho_{k+1}^2 \not\to 0$ as $K^\downarrow \ni k \to \infty$, where the inequality is ensured by Property P1. Then, (regardless of whether $K^\uparrow$ is finite or not) there exists $\varepsilon > 0$ together with an infinite set $\hat{K} \subseteq K^\downarrow$ such that $1 + q\rho_k - q\rho_{k+1}^2 \geq \varepsilon$ holds for all $k \in \hat{K}$. Since $\hat{K} \subseteq K^\downarrow$, by virtue of Property P3 we then have that $1 + q\rho_k - q\rho_{k+1}^2 \geq \varepsilon$ and $\gamma_{k+1} \geq \lambda_{\min,\nu}\|x^{j_k} - x^{j_k-1}\|^{1-\nu}$ hold $\forall k \in \hat{K}$, where $j_k \leq k$ for every $k \in \hat{K}$. By combining with (13) we obtain

$$\infty > \sum_{k\in\hat{K}}\gamma_{k+1}P_k \geq \lambda_{\min,\nu}\sum_{k\in\hat{K}}\|x^{j_k} - x^{j_k-1}\|^{1-\nu}P_k.$$

Therefore, either $\liminf_{\hat{K}\ni k\to\infty}\|x^{j_k} - x^{j_k-1}\|^{1-\nu} = 0$ or $\liminf_{\hat{K}\ni k\to\infty} P_k = 0$. In the latter case, the claim is proven. In the former case, necessarily ($\nu < 1$ and) the sequence $(j_k)_{k\in\hat{K}}$ contains infinitely many indices (for otherwise $(\|x^{j_k} - x^{j_k-1}\|)_{k\in\hat{K}}$ would alternate between a finite number of nonzero elements). Up to extracting, we have that $\lim_{\hat{K}\ni k\to\infty}\|x^{j_k} - x^{j_k-1}\| = 0$. For any $x^\star \in \arg\min\varphi$ we thus have (recall that $P_k = \varphi(x^k) - \varphi(x^\star)$)

$$\begin{aligned}
P_{j_k} &\leq \langle x^{j_k} - x^\star, \tfrac{x^{j_k-1} - x^{j_k}}{\gamma_{j_k}} - \left(\nabla f(x^{j_k-1}) - \nabla f(x^{j_k})\right)\rangle \\
&\leq \|x^{j_k} - x^\star\|\Big(\tfrac{1}{\gamma_{j_k}}\|x^{j_k-1} - x^{j_k}\| \\
&\qquad + \|\nabla f(x^{j_k-1}) - \nabla f(x^{j_k})\|\Big) \qquad (14)\\
&\leq \|x^{j_k} - x^\star\|\|x^{j_k-1} - x^{j_k}\|^\nu \tfrac{1+\lambda_{\min,\nu}L_{\Omega,\nu}}{\lambda_{\min,\nu}} \quad \forall k \in \hat{K}.
\end{aligned}$$

Since $\nu > 0$, by taking the limit as $\hat{K} \ni k \to \infty$ we obtain that $\lim_{\hat{K}\ni k\to\infty} P_{j_k} = 0$.

♢ *Claim 4(b):* $(x^k)_{k\in\mathbb{N}}$ *converges to a solution.*

Having shown the previous claim and since $\sup_{k\in\mathbb{N}}\rho_k < \infty$ (by Property P1), the proof is the same as in [20, Thm. 3.8].

To simplify the notation we let $D := \sqrt{2\mathcal{U}_1^q(x^\star)}$.

♢ *Claim 4(c):* $P_k^{\min} \leq D\frac{1+\lambda_{k,\nu}L_{\Omega,\nu}}{\lambda_{k,\nu}}\|x^k - x^{k-1}\|^\nu$ *holds for any $k \in \mathbb{N}$.*

See [20, Claim 3.9(b)].

♢ *Claim 4(d): For any $k \in \mathbb{N}$ it holds that*

$$\gamma_{k+1} \geq \begin{cases} \lambda_{\min,\nu}^{\frac{1}{\nu}}\left(\frac{P_k^{\min}}{D(1+\lambda_{\min,\nu}L_{\Omega,\nu})}\right)^{\frac{1-\nu}{\nu}} & \text{if } k \geq \min K^\downarrow \\ \gamma_0 & \text{otherwise.}\end{cases}$$

Suppose first that $k \in K^\downarrow$ (in particular, $k \geq \min K^\downarrow$). Since $\gamma_{k+1} < \gamma_k$, it follows from Property P3 that

$$\gamma_{k+1} \geq \min\{\gamma_{j_k}, \gamma_{k+1}\} \geq \lambda_{\min,\nu}\|x^{j_k} - x^{j_k-1}\|^{1-\nu}$$

holds for some $j_k \leq k$, and in particular that $\lambda_{j_k,\nu} \geq \lambda_{\min,\nu}$. Noticing that $P_{j_k}^{\min} \geq P_k^{\min}$ since $j_k \leq k$, the sought inequality follows by using the lower bound

$$\|x^{j_k} - x^{j_k-1}\| \geq \left(\frac{\lambda_{j_k,\nu}P_{j_k}^{\min}}{D(1+\lambda_{j_k,\nu}L_{\Omega,\nu})}\right)^{\frac{1}{\nu}} \geq \left(\frac{\lambda_{\min,\nu}P_k^{\min}}{D(1+\lambda_{\min,\nu}L_{\Omega,\nu})}\right)^{\frac{1}{\nu}}$$

obtained from Claim 4(c) raised to the power $\frac{1}{\nu}$.

Next, suppose that $k \in K^\uparrow$. Let

$$K_{<k}^\downarrow := K^\downarrow \cap \{0, 1, \ldots, k-1\}$$

denote the (possibly empty) set of all iteration indices up to $k-1$ in which the next stepsize is strictly smaller.

If $K_{<k}^\downarrow = \emptyset$, then $\gamma_k \geq \gamma_{k-1} \geq \cdots \geq \gamma_0$.

Suppose instead that $K_{<k}^\downarrow \neq \emptyset$, and let i_k denote its largest element:

$$i_k := \max K_{<k}^\downarrow = \max\{i < k \mid \gamma_{i+1} < \gamma_i\} < k.$$

Then,

$$\begin{aligned}
\gamma_{k+1} \geq \gamma_{i_k+1} &\geq \lambda_{\min,\nu}^{\frac{1}{\nu}}\left(\tfrac{1}{D(1+\lambda_{\min,\nu}L_{\Omega,\nu})}P_{i_k}^{\min}\right)^{\frac{1-\nu}{\nu}} \\
&\geq \lambda_{\min,\nu}^{\frac{1}{\nu}}\left(\tfrac{1}{D(1+\lambda_{\min,\nu}L_{\Omega,\nu})}P_k^{\min}\right)^{\frac{1-\nu}{\nu}}
\end{aligned}$$

where the second inequality follows from the fact that $i_k \in K^\downarrow$, and the last one from the fact that $(P_k^{\min})_{k\in\mathbb{N}}$ is decreasing (note that $k > i_k$).

The sum of stepsizes can then be lower bounded by

$$\begin{aligned}
\sum_{k=1}^{K+1}\gamma_k &\geq \sum_{k=0}^{K}\min\left\{\gamma_0, \lambda_{\min,\nu}^{\frac{1}{\nu}}\left(\frac{P_k^{\min}}{D(1+\lambda_{\min,\nu}L_{\Omega,\nu})}\right)^{\frac{1-\nu}{\nu}}\right\} \\
&\geq (K+1)\min\left\{\gamma_0, \lambda_{\min,\nu}^{\frac{1}{\nu}}\left(\frac{P_K^{\min}}{D(1+\lambda_{\min,\nu}L_{\Omega,\nu})}\right)^{\frac{1-\nu}{\nu}}\right\},
\end{aligned}$$

where the last inequality uses the fact that $P_K^{\min} \le P_k^{\min}$ for every $k \le K$. Therefore, in light of Fact 2*(3)* we have

$$\mathcal{U}_1^q(x^\star) \ge (K+1)\min\left\{\gamma_0 P_K^{\min},\ \frac{(\lambda_{\min,\nu} P_K^{\min})^{1/\nu}}{(D(1+\lambda_{\min,\nu}L_{\Omega,\nu}))^{\frac{1-\nu}{\nu}}}\right\}.$$

From $D := \sqrt{2\mathcal{U}_1^q(x^\star)}$, the claimed bound follows. □

Neglecting the first term for simplicity, from Lemma 3 we obtain the sublinear rate

$$P_K^{\min} \le (1+\sqrt{2q})^{1-\nu}\frac{\sqrt{2}\sqrt{q}^\nu \mathcal{U}_1^q(x^\star)^{\frac{1+\nu}{2}} L_{\Omega,\nu}}{(K+1)^\nu} \tag{15}$$

for plain adaPG$^{q,\frac{q}{2}}$, improving the estimate

$$P_K^{\min} \le (1+\sqrt{2}\rho_{\max})^{1-\nu}\frac{\sqrt{2}\sqrt{q}^\nu \mathcal{U}_1^q(x^\star)^{\frac{1+\nu}{2}} L_{\Omega,\nu}}{(K+1)^\nu}$$

with $\rho_{\max} := \frac{1+\sqrt{1+\frac{4}{q}}}{2}$ of [20, Thm. 3.9].

IV. Choice of stepsizes

In this section we translate the convergence recipe into sufficient conditions for "fast" stepsize choices $\gamma_{k+1}^{\mathsf{fast}}$ to be safeguardable by adaPG$^{q,\frac{q}{2}}$. To this end, we restrict our scrutiny to stepsizes oracles of the form

$$\gamma_{k+1}^{\mathsf{fast}} = \Gamma^{\mathsf{fast}}(x^{k-m},\dots,x^k)$$

for some $m \ge 1$ and $\Gamma^{\mathsf{fast}} : (\mathbb{R}^n)^{m+1} \to \mathbb{R}_{++}$. The safeguard framework is elementary: take the minimum between the desired stepsize $\gamma_{k+1}^{\mathsf{fast}}$ and the *safe* one of adaPG$^{q,\frac{q}{2}}$:

Algorithm 1 Safeguard for stepsizes $\Gamma^{\mathsf{fast}}:(\mathbb{R}^n)^{m+1}\to\mathbb{R}_{++}$

REQUIRE $q \in [1,2]$, $x^0,\dots,x^m \in \mathbb{R}^n$, $\gamma_{m-1},\gamma_m > 0$

REPEAT FOR $k = m, m+1,\dots$ until convergence

$\gamma_{k+1}^{\mathsf{safe}} = \min\left\{\gamma_k\sqrt{\frac{1}{q}+\frac{\gamma_k}{\gamma_{k-1}}},\ \frac{\gamma_k}{\sqrt{2\left[\gamma_k^2L_k^2-(2-q)\gamma_k\ell_k+1-q\right]_+}}\right\}$

$\gamma_{k+1}^{\mathsf{fast}} = \Gamma^{\mathsf{fast}}(x^{k-m},\dots,x^k)$

$\gamma_{k+1} = \min\left\{\gamma_{k+1}^{\mathsf{safe}},\gamma_{k+1}^{\mathsf{fast}}\right\}$

$x^{k+1} = \mathrm{prox}_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k))$

Lemma 5 (convergence of Algorithm 1)**.** *Let Assumption I hold for some $\nu \in (0,1]$, and let $\Gamma^{\mathsf{fast}} : (\mathbb{R}^n)^{m+1} \to \mathbb{R}_{++}$ for some $m \ge 1$. Suppose that for any compact $\Omega \subset \mathbb{R}^n$ there exists $\lambda_{\Omega,\nu} > 0$ with the property that whenever $z^0,\dots,z^m \in \Omega$ there exists $i \in \{1,\dots,m\}$ such that $\Gamma^{\mathsf{fast}}(z^0,\dots,z^m) \ge \lambda_{\Omega,\nu}\|z^i - z^{i-1}\|^{1-\nu}$. Then, denoting $\rho_{\max} = \frac{1+\sqrt{1+\frac{4}{q}}}{2}$, the iterates generated by Algorithm 1 comply with the convergence recipe with $\lambda_{\min,\nu} = \min\left\{\frac{1}{\sqrt{2q}L_{\Omega,\nu}}, \frac{\lambda_{\Omega,\nu}}{\rho_{\max}^{m-1}}\right\}$, and in particular Theorem 4 holds true.*

Proof. Properties P1 and P2 follow from $\gamma_{k+1} \le \gamma_{k+1}^{\mathsf{safe}}$. Fact 2*(2)* then ensures boundedness of $(x^k)_{k\in\mathbb{N}}$, whence the existence of a $\lambda_{\Omega,\nu} > 0$ as in the statement. Thus, for every k there exists $j_k \in \{k-m+1,\dots,k\}$ such that $\gamma_{k+1}^{\mathsf{fast}} \ge \lambda_{\Omega,\nu}\|x^{j_k} - x^{j_k-1}\|^{1-\nu}$. If $\gamma_{k+1} < \gamma_{k+1}^{\mathsf{fast}}$, Property P3 with $\lambda_{\min,\nu} = \frac{1}{\sqrt{2q}L_{\Omega,\nu}}$ follows from Lemma 3. Otherwise,

$$\gamma_{j_k} \ge \frac{\gamma_k}{\rho_{\max}^{j_k-1}} \ge \frac{\gamma_k}{\rho_{\max}^{m-1}} \ge \frac{\gamma_{k+1}}{\rho_{\max}^{m-1}} \ge \frac{\lambda_{\Omega,\nu}}{\rho_{\max}^{m-1}}\|x^{j_k} - x^{j_k-1}\|^{1-\nu},$$

with first inequality owing to Property P1. □

We now provide some examples of stepsizes that well fit in the safeguarding framework of Algorithm 1. As will be evident from the experiment in §V, we strongly advocate for the choice of the Anderson acceleration–like one described in §IV-D, which dramatically boosts convergence speed in practice. To simplify the notation, in what follows we denote $s^k := x^k - x^{k-1}$ and $y^k := \nabla f(x^k) - \nabla f(x)$.

A. Long and short Barzilai-Borwein

When f satisfies Assumption I.A1 for some $\nu \in (0,1]$, these choices respectively correspond to

$$\gamma_{k+1}^{\mathsf{BB^{long}}} = \frac{1}{\ell_k} \quad\text{and}\quad \gamma_{k+1}^{\mathsf{BB^{short}}} = \frac{1}{c_k^\nu L_k^{1-\nu}} = \frac{1}{\sqrt{c_k^{1+\nu}\ell_k^{1-\nu}}}. \tag{16}$$

It is immediate to verify that the long one complies with Lemma 5, having $\ell_k = \ell_{k,\nu}\|x^k - x^{k-1}\|^{1-\nu} \le L_{\Omega,\nu}\|x^k - x^{k-1}\|^{1-\nu}$, see (8) and (9). For the short one, the involvement of a geometric average with the long one when $\nu < 1$ is necessary to ensure compliance with Lemma 5, having $\frac{1}{c_k} = \frac{\langle y^k,s^k\rangle}{\|y^k\|^2} \propto \|y^k\|^{\frac{1-\nu}{\nu}}$; see [20, Fact 2.2.3] for the details.

B. Martinez' rule for long and short BB

BB$^{\mathsf{long}}$ and BB$^{\mathsf{short}}$ are one-dimensional equivalents of Broyden's "good" and "bad" quasi-Newton method, namely $\gamma_{k+1}^{\mathsf{BB^{long}}} = \operatorname{argmin}_{\gamma\in\mathbb{R}}\|\frac{s^k}{\gamma} - y^k\|^2$ and $\gamma_{k+1}^{\mathsf{BB^{short}}} = \operatorname{argmin}_{\gamma\in\mathbb{R}}\|\gamma y^k - s^k\|^2$ are chosen to minimize the secant and inverse secant quasi-Newton approximation [19]. As suggested in [19, §3.1], we may choose between the two based on the rule

$$\gamma_{k+1}^{\mathsf{Martinez}} = \begin{cases}\gamma_{k+1}^{\mathsf{BB^{long}}} & \text{if } \gamma_k > \frac{\langle s^k,s^{k-1}\rangle}{\langle y^k,y^{k-1}\rangle}\\ \gamma_{k+1}^{\mathsf{BB^{short}}} & \text{otherwise,}\end{cases} \tag{17}$$

which selects the one for which the inverse secant error on the previous pair is minimized.

C. Least normalized secant error for long and short BB

We can take both direct and inverse errors into account. Relative to the previous pair (s^{k-1},y^{k-1}), we may opt for BB$^{\mathsf{long}}$ when it has lower inverse secant error, and BB$^{\mathsf{short}}$ if this has lower secant error. If neither holds, we compare the respective scaled residuals, thereby selecting BB$^{\mathsf{long}}$ when

$$\|s^k\|^{-1}\|s^k - \gamma_{k+1}^{\mathsf{BB^{long}}}y^k\| \le \|y^k\|^{-1}\|y^k - s^k/\gamma_{k+1}^{\mathsf{BB^{short}}}\| \tag{18}$$

or else BB$^{\mathsf{short}}$. The proposed rule boils down to

$$\gamma_{k+1}^{\mathsf{LNSE}} = \begin{cases}\gamma_{k+1}^{\mathsf{BB^{long}}} & \text{if } \gamma_{k+1}^{\mathsf{BB^{long}}} + \gamma_{k+1}^{\mathsf{BB^{short}}} \le 2\gamma_k^{\mathsf{BB^{short}}}\\ \gamma_{k+1}^{\mathsf{BB^{short}}} & \text{else if } \frac{1}{\gamma_{k+1}^{\mathsf{BB^{long}}}} + \frac{1}{\gamma_{k+1}^{\mathsf{BB^{short}}}} \ge \frac{2}{\gamma_k^{\mathsf{BB^{long}}}}\\ \gamma_{k+1}^{\mathsf{BB^{long}}} & \text{else if (18)}\\ \gamma_{k+1}^{\mathsf{BB^{short}}} & \text{otherwise.}\end{cases} \tag{19}$$

D. Anderson acceleration

Keep the latest $m \ge 1$ pairs s^i and y^i, and set

$$\gamma_{k+1}^{\mathsf{AA}_m} = \frac{\sum_{i=k-m+1}^k\langle s^i,y^i\rangle}{\sum_{i=k-m+1}^k\|y^i\|^2} = \frac{\sum_{i=k-m+1}^k\|y^i\|^2\frac{1}{c_i}}{\sum_{i=k-m+1}^k\|y^i\|^2}. \tag{20}$$

Apparently, this is a weighted average (with weights $\|y^i\|^2$) of the most recent BB$^{\mathsf{short}}$ stepsizes, hence Lemma 5 is verified when Assumption I holds with $\nu = 1$. When $\nu < 1$,

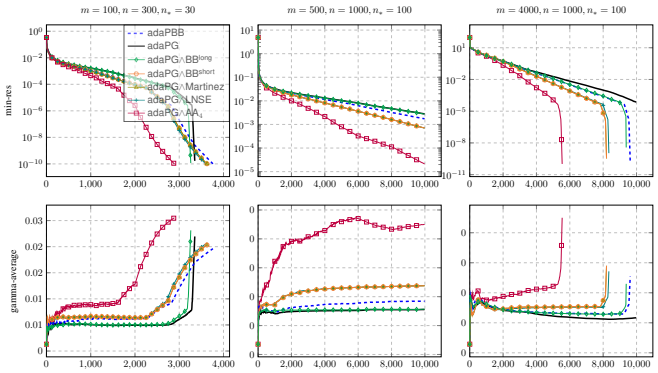

Fig. 1. Random lasso problem with ℓ_1-regularization parameter $\lambda = 0.1$. $n_\star$ represents the number of nonzero elements in the solution.

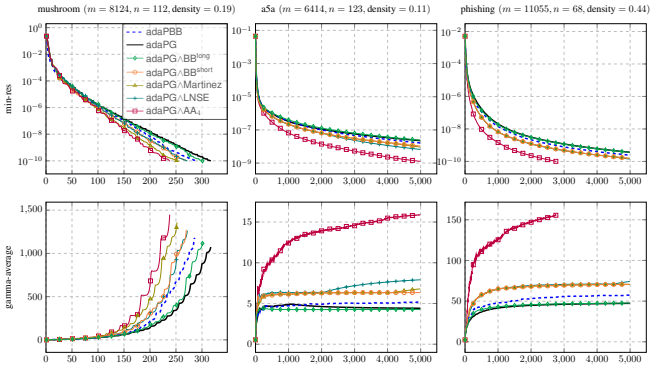

Fig. 2. Regularized logistic regression (m and n are the number of samples and features). The ℓ_1-regularization parameter λ is set as in [23, §6.1].

consistently with the discussion in §IV-A, a suitable geometric averaging is needed. The name Anderson acceleration [1] is evocative of its multi-secant (inverse) update interpretation, see [11] and [22, §3.3.2], as it is easily seen to equal $\gamma_{k+1}^{\mathsf{AA}} = \arg\min_{\gamma\in\mathbb{R}} \sum_{i=k-m+1}^{k} \|\gamma y^i - s^i\|^2$.

V. Numerical results

We conducted a series of simulations to test the effectiveness of adaPG$^{q,\frac{q}{2}}$ safeguarding the five stepsizes listed in §IV. The safeguarding of adaPG$^{q,\frac{q}{2}}$ is represented in the legend plots with "adaPG$^{q,\frac{q}{2}}\wedge$". Our tests are based on the Julia code provided in [13] on problems from the LIBSVM dataset [9]. In all instances of adaPG$^{q,\frac{q}{2}}$ we use $q = 1.2$, consistently with its good performance reported in [12], while for Anderson acceleration (20) a memory $m = 4$ is used. The resulting methods are tested against adaPG$^{q,\frac{q}{2}}$ (4) and adaPBB [23, Alg. 3].

We refer to [13, §4] for a detailed account of the experiments, which are here reproduced with minimal variations. In each figure, top plots report the best-so-far residual

$$r_k = \|\tfrac{x^k - x^{k-1}}{\gamma_k} - (\nabla f(x^k) - \nabla f(x^{k-1}))\|,$$

and bottom ones the stepsize cumulative average $\frac{1}{k}\sum_{i=1}^{k}\gamma_i$. On the x-axis we report the number of gradient evaluations, which also corresponds to the iteration count in all methods.

All experiments confirm that quasi-Newton–type ideas can yield considerable improvements within adaptive methods. The convergence speed is evidently correlated with the magnitude of the stepsizes, as confirmed by Fact 2*(3)*. The winner

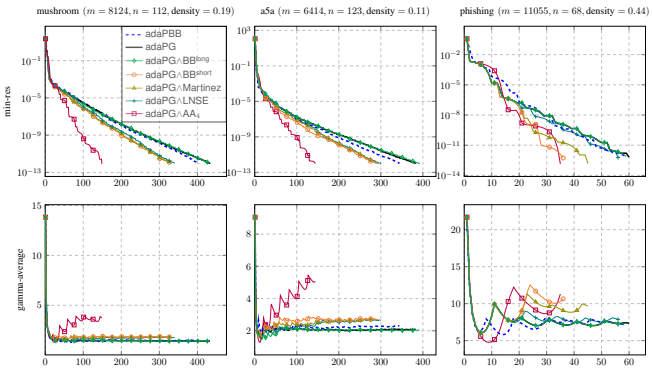

Fig. 3. Cubic regularization problem with Hessian and gradient generated from the logistic loss problem evaluated at zero on the mushroom and phishing datasets. The cubic regularization parameter is set as $M = 0.01$.

turns out to be Anderson acceleration (20) ($m = 4$), consistently outperforming all other choices.

References

[1] D.G. Anderson. Iterative procedures for nonlinear integral equations. *J. ACM*, 12(4):547–560, 1965.

[2] J. Barzilai and J.M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, jan 1988.

[3] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[4] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, 18(11):2419–2434, 2009.

[5] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] S. Bubeck. Theory of convex optimization for machine learning. *arXiv:1405.4980*, 2014.

[8] O. Burdakov, Y. Dai, and N. Huang. Stabilized Barzilai-Borwein method. *J. Comput. Math.*, 37(6):916–936, 2019.

[9] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)*, 2(3):1–27, 2011.

[10] Y. Dai and L. Liao. R-linear convergence of the Barzilai and Borwein gradient method. *IMA J. Numer. Anal.*, 22(1):1–10, 2002.

[11] H. Fang and Y. Saad. Two classes of multisecant methods for nonlinear acceleration. *Numer. Linear Algebra Appl.*, 16(3):197–221, 2009.

[12] P. Latafat, A. Themelis, and P. Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. *arXiv:2311.18431*, 2023.

[13] P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *arXiv:2301.04431*, 2023.

[14] D. Li and R. Sun. On a faster R-linear convergence rate of the Barzilai-Borwein method. *arXiv:2101.00205*, 2021.

[15] Y. Li, X. Tang, X. Lin, L. Grzesiak, and X. Hu. The role and application of convex modeling and optimization in electrified vehicles. *Renewable Sustainable Energy Rev.*, 153:111796, 2022.

[16] Z. Luo. Applications of convex optimization in signal processing and digital communication. *Math. Program.*, 97(1):177–207, 2003.

[17] Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *Proc. 37th Int. Conf. Mach. Learn.*, volume 119, pages 6702–6712, 2020.

[18] Y. Malitsky and K. Mishchenko. Adaptive proximal gradient method for convex optimization. *arXiv:2308.02261*, 2023.

[19] J.M. Martínez. Practical quasi-Newton methods for solving nonlinear systems. *J. of Comp. Appl. Math.*, 124(1-2):97–121, 2000.

[20] K.A. Oikonomidis, E. Laude, P. Latafat, A. Themelis, and P. Patrinos. Adaptive proximal gradient methods are universal without approximation. *arXiv:2402.06271*, 2024.

[21] M. Raydan. On the Barzilai and Borwein choice of steplength for the gradient method. *IMA J. Numer. Anal.*, 13(3):321–326, jul 1993.

[22] A. Themelis, L. Stella, and P. Patrinos. Douglas-Rachford splitting and ADMM for nonconvex optimization: Accelerated and Newton-type algorithms. *Comput. Optim. Appl.*, 82:395–440, 2022.

[23] D. Zhou, S. Ma, and J. Yang. AdaBB: Adaptive Barzilai-Borwein method for convex optimization. *arXiv:2401.08024*, 2024.