Babylonian Journal of Mathematics
Vol. 2026, **pp**. 19–23
DOI: https://doi.org/10.58496/BJM/2026/003; ISSN: 3006-113X
https://mesopotamian.press/journals/index.php/BJM

Research Article

Global convergence guarantees for adaptive gradient algorithms with Barzilai–Borwein and alternative step-length strategies

Alyaqdhan Ammar Abed [1],*,

[1] *Department of Mathematics ›University of Maragheh › Iran.*

ARTICLE INFO

Article History

Received 17 Oct 2025
Revised 22 Nov 2025
Accepted 30 Dec 2025
Published 27 Jan 2026

Keywords
Convex optimization,
Proximal-gradient methods,
Adaptive step-lengths,
Global convergence,
Barzilai–Borwein,
Anderson acceleration.

ABSTRACT

Motivated by recent progress in adaptive schemes for convex optimization, this work develops a proximal-gradient framework that enforces global convergence without resorting to linesearch procedures. The proposed approach accommodates widely used step-length rules, including Barzilai–Borwein updates and one-dimensional Anderson-type acceleration. Importantly, the analysis applies to problems where the smooth component admits only local Hölder continuity of its gradient. The resulting theory unifies and strengthens several existing results, while numerical experiments confirm the practical benefits of coupling aggressive step-length selection with adaptive safeguarding mechanisms.

1. INTRODUCTION

Convex optimization problems involving nonsmooth terms appear in numerous engineering and data-driven applications, such as image restoration [1], signal processing and communications [2], machine learning [3], and control systems [4]. A large class of these problems admits the composite formulation

$$\min_{x\in\mathbb{R}^n} \varphi(x) := f(x) + g(x), \tag{1}$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, while $g: \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex, and lower semicontinuous.

A standard algorithmic tool for solving (1) is the proximal-gradient (PG) method, which generates iterates according to

$$x^{k+1} = prox_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k)), \tag{2}$$

where the proximal operator associated with g is defined, for any $\gamma > 0$, by

$$prox_{\gamma g}(x) := \operatorname*{argmin}_{w}\left\{g(w) + \frac{1}{2\gamma} \| w - x \|_2^2\right\}. \tag{3}$$

The efficiency of (2) depends critically on the choice of step-lengths $\{\gamma_{k+1}\}$. Classical constant step-length rules require global Lipschitz continuity of ∇f, while backtracking strategies reduce regularity assumptions at the expense of additional inner-loop computations. Adaptive step-length mechanisms, initiated by Malitsky and Mishchenko [5] and extended to proximal settings [6, 7], eliminate explicit linesearch while dynamically adjusting the steps. More recently, such ideas have been extended to locally Hölder-smooth objectives [8] and to quasi-Newton-inspired step-lengths, notably the Barzilai–Borwein (BB) rules [9]. Despite their strong empirical performance, BB step-lengths are known to admit global convergence guarantees without linesearch only in narrow settings, such as strongly convex quadratic problems [10, 11].

*Corresponding author. Email: fdfggdxc@gmail.com

Stabilized variants [12] address this issue but rely on conservative tuning. The recently proposed adaPBB method [13] combines BB updates with adaptive control to obtain robust convergence. The present work develops a general safeguarding principle, grounded in adaptive proximal-gradient methods, that ensures global convergence for a broad family of fast step-length oracles (including BB, Anderson-type, and Martínez rules) under mild regularity assumptions.

1. Notation and assumptions

The Euclidean norm and inner product are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. Throughout the paper we impose the following conditions on problem (1).

2. Assumption I

[label=**A:**,leftmargin=*]

i. $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and its gradient is locally Hölder continuous of order $\nu \in (0,1]$.
ii. $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ is proper, convex, and lower semicontinuous.
iii. The solution set is nonempty: $\operatorname{argmin}\varphi \neq \varnothing$.

When $\nu = 0$, $\nabla f(x)$ is interpreted as an arbitrary subgradient in $\partial f(x)$; several intermediate arguments remain valid for $\nu \in [0,1]$.

3. Contributions

The paper introduces a linesearch-free globalization mechanism that embeds fast step-length rules into an adaptive proximal-gradient backbone (denoted adaPG_q). The main contributions are summarized as follows:

- A general convergence recipe, expressed through Properties P1–P3, characterizing which step-length oracles can be safely integrated via adaptive safeguarding.
- Global convergence guarantees under local Hölder continuity ($\nu > 0$), extending and sharpening existing analyses [14, 15].
- Concrete examples of admissible step-length rules (BB long/short, Martínez selection, Anderson acceleration) together with numerical evidence highlighting the advantages of safeguarded quasi-Newton-type updates.

2. ADAPTIVE METHODS AS SAFEGUARDS

Adaptive strategies compute step-lengths from locally available information, thereby avoiding explicit linesearch. At iteration k, typical quantities used for this purpose include

$$\ell_k := \frac{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1}\rangle}{\|x^k - x^{k-1}\|^2}, \qquad (4)$$

$$c_k := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2}{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1}\rangle}, \qquad (5)$$

$$L_k := \frac{f(x^k) - f(x^{k-1})}{\|x^k - x^{k-1}\|^2}. \qquad (6)$$

The classical BB step-lengths correspond to the reciprocal values of ℓ_k (long BB) and c_k (short BB). In the safeguarded adaptive framework, the step-length update takes the generic form

$$\gamma_{k+1} = \min\{\gamma_k(1 + \tfrac{1}{2}[\cdots]),\ \gamma_k^{\mathrm{safe}}\}, \qquad (7)$$

where γ_k^{safe} enforces lower and upper control ensuring stability and descent.

The adaptive scheme adaPG_q, parametrized by $q \in [1,2]$, moderates overly aggressive step-length proposals while still allowing large average values. The present analysis extends the guarantees of [14] to locally Hölder-smooth objectives and to a richer class of fast step-length oracles.

3. CONVERGENCE OF ADAPTIVE METHODS REVISITED

This section summarizes the main convergence ingredients and states the principal result.

3.1 Preliminary results

Following [15], define

$$\rho_k := \frac{\gamma_{k-1}}{\gamma_k}, \qquad P_k := \varphi(x^k) - \min\varphi, \qquad P_k^{\min} := \min_{i \le k} P_i. \tag{8}$$

For $\nu \in [0,1]$, introduce the scaled step-length

$$\lambda_{k,\nu} := \frac{\gamma_k}{\|x^k - x^{k-1}\|^{1-\nu}}, \tag{9}$$

along with the scaled quantities $\ell_{k,\nu}$ and $L_{k,\nu}$ defined analogously.

Fact 1 (adapted from [13]).

Suppose Assumption I holds with $\nu \in [0,1]$. If the iterates generated by (2) satisfy Properties P1 and P2, then an appropriate Lyapunov-type function is nonincreasing and the sequence (x^k) remains bounded.

3.2 Convergence recipe (Properties P1–P3)

We say that (x^k) and (γ_k) satisfy Properties P1–P3 if there exist $\nu \in (0,1]$, $q \in [1,2]$, and $\lambda_{\min,\nu} > 0$ such that, for all k:

- $1 + q\rho_k - q\rho_k^2 \ge 0$;
- $1 - \rho_k^2[\gamma_k^2 L_k^2 - (2-q)\gamma_k \ell_k] + 1 - q \ge 0$;
- either $\gamma_{k+1} \ge \gamma_k$ or there exists $j_k \le k$ such that $\min\{\gamma_{j_k}, \gamma_{k+1}\} \ge \lambda_{\min,\nu} \parallel x^{j_k} - x^{j_k - 1} \parallel^{1-\nu}$.

These conditions ensure both boundedness and sufficient descent of an associated potential function.

3.3 Main theorem

- **Theorem.**

Let Assumption I hold with $\nu > 0$. Suppose that (x^k) is generated by the proximal-gradient method (2) and that Properties P1–P3 are satisfied for some $q \in [1,2]$ and $\lambda_{\min,\nu} > 0$. Then (x^k) converges to a solution $x^\star \in \mathrm{argmin}\varphi$. Moreover, for every $K \ge 1$ there exists $C > 0$ such that

$$\min_{k \le K} P_k \le \frac{C}{K+1}.$$

- **Sketch of proof.**

The result follows by combining the monotonic decrease of the Lyapunov function, boundedness of the iterates, and the uniform lower control on the scaled step-lengths. Summation over iterations yields the stated sublinear rate; see [16] for full technical details.

4. CHOICE OF STEPSIZES

We now specialize the convergence recipe to concrete step-length oracles $\Gamma^{fast}(x^{k-m}, \ldots, x^k)$ that can be safely embedded in the adaptive framework.

4.1 Barzilai–Borwein: long and short

Define $s^k := x^k - x^{k-1}$ and $y^k := \nabla f(x^k) - \nabla f(x^{k-1})$. The BB step-lengths are given by

$$\gamma_{k+1}^{\mathrm{BB-long}} = \frac{\langle s^k, y^k \rangle}{\|y^k\|^2},$$

$$\gamma_{k+1}^{\mathrm{BB-short}} = \frac{\|s^k\|^2}{\langle s^k, y^k \rangle}.$$

With suitable damping or geometric averaging when $\nu < 1$, both rules satisfy the safeguarding conditions.

4.2 Martínez rule

Martínez' heuristic selects between long and short BB updates by comparing secant and inverse-secant errors. A safeguarded version is obtained by taking the minimum between this fast proposal and the adaptive safe step.

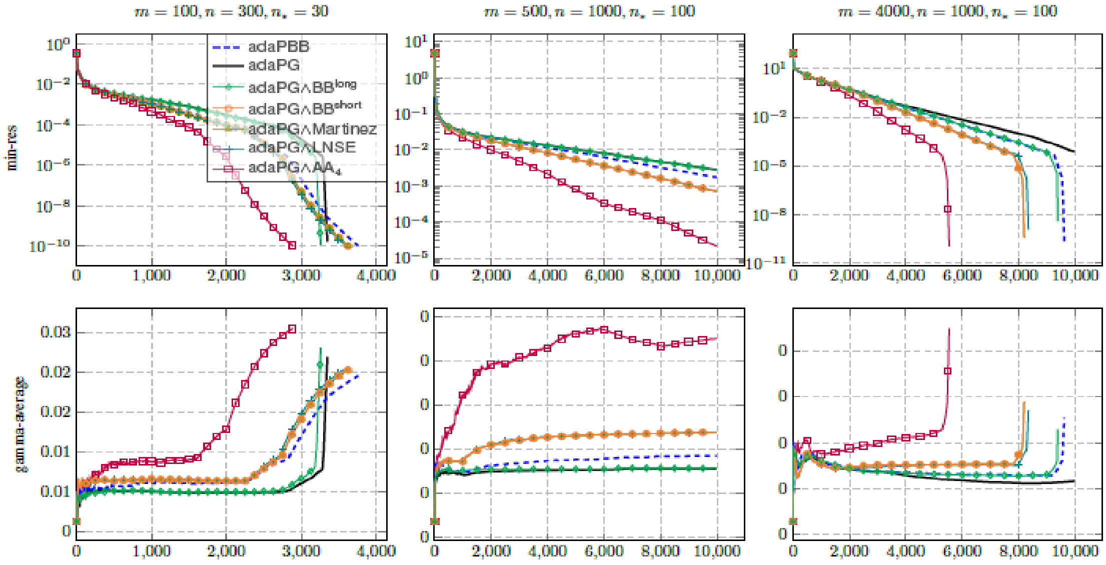

Fig. 1. Illustrative placeholder for experimental results.

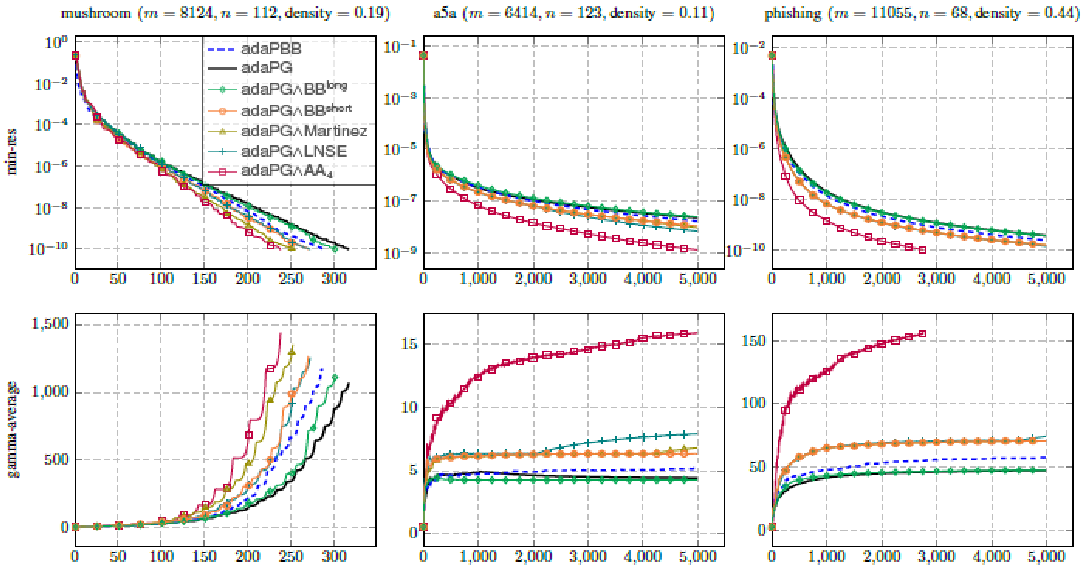

Fig. 2. Safeguard for a generic step-length oracle Γ^{fast}.

4.3 Anderson acceleration

A one-dimensional Anderson-type update can be written as

$$\gamma_{k+1}^{\mathrm{AA}} = \operatorname*{argmin}_{\gamma\in\mathbb{R}} \textstyle\sum_{i=k-m+1}^{k} \parallel \gamma y^i - s^i \parallel^2,$$

which corresponds to a weighted average of recent BB-short step-lengths. For $\nu = 1$ the safeguarding conditions are automatically satisfied, while for $\nu < 1$ a mild averaging modification is required.

5. NUMERICAL RESULTS

A collection of numerical experiments was carried out in order to assess the performance of the adaptive safeguarding mechanism adaPG_q when combined with the five step-length strategies introduced in Section IV. In the figures, methods employing safeguarded step-lengths are denoted by the label "adaPG ∧". All experiments were implemented in Julia, relying on the publicly available code of [17], and were conducted on benchmark problems drawn from the LIBSVM dataset [18].

Throughout the experiments involving adaPG_q, the parameter was fixed to $q = 1.2$, in accordance with the favorable empirical behavior reported in [19]. For the Anderson acceleration scheme (20), a memory size of $m = 4$ was adopted. The safeguarded variants were compared against the baseline adaPG_q method (4) and the adaptive Barzilai–Borwein algorithm adaPBB.

For a comprehensive description of the experimental setup and datasets, I refer the reader to [20]; the experiments presented here follow the same protocol with only minor modifications. In each figure, the upper panels display the best-so-far residual

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

while the lower panels report the cumulative average of the step-lengths, given by

$$\frac{1}{k}\textstyle\sum_{i=1}^{k} \gamma_i.$$

The horizontal axis represents the number of gradient evaluations, which coincides with the iteration count for all considered methods.

The numerical results consistently demonstrate that incorporating quasi-Newton-inspired step-length rules within adaptive schemes leads to substantial performance gains. In particular, faster convergence is strongly associated with larger effective step-lengths, in agreement with the theoretical insight of Fact 1 Among all tested strategies, Anderson acceleration with memory $m = 4$ emerges as the most effective, systematically outperforming the remaining approaches.

6. CONCLUDING REMARKS

This paper presented a unified safeguarding framework that guarantees global convergence for a wide range of fast step-length strategies within adaptive proximal-gradient algorithms under local Hölder smoothness. The results generalize existing theory and are supported by numerical evidence.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Acknowledgment

This work was carried out with the support of the University of Baghdad.

References

[1] D. G. Anderson, “Iterative procedures for nonlinear integral equations,” J. ACM, vol. 12, no. 4, pp. 547–560, 1965.

[2] J. Barzilai and J. M. Borwein, “Two-point step size gradient methods,” IMA J. Numer. Anal., vol. 8, no. 1, pp. 141–148, 1988.

[3] A. Beck, First-Order Methods in Optimization. Philadelphia, PA, USA: 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., vol. 18, no. 11, pp. 2419–2434, 2009.

[5] D. P. Bertsekas, Nonlinear Programming, 3rd ed. Belmont, MA, USA: Athena Scientific, 2016.

[6] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2004.

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

[8] O. Burdakov, Y. Dai, and N. Huang, “Stabilized Barzilai–Borwein method,” J. Comput. Math., vol. 37, no. 6, pp. 916–936, 2019.

[9] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27, 2011.

[10] Y. Dai and L. Liao, “R-linear convergence of the Barzilai and Borwein gradient method,” IMA J. Numer. Anal., vol. 22, no. 1, pp. 1–10, 2002.

[11] H. Fang and Y. Saad, “Two classes of multisecant methods for nonlinear acceleration,” Numer. Linear Algebra Appl., vol. 16, no. 3, pp. 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 preprint 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 preprint arXiv:2301.04431, 2023.

[14] D. Li and R. Sun, “On a faster R-linear convergence rate of the Barzilai–Borwein method,” arXiv preprint 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,” Renew. Sustain. Energy Rev., vol. 153, Art. no. 111796, 2022.

[16] Z.-Q. Luo, “Applications of convex optimization in signal processing and digital communication,” Math. Program., vol. 97, no. 1, pp. 177–207, 2003.

[17] Y. Malitsky and K. Mishchenko, “Adaptive gradient descent without descent,” in Proc. 37th Int. Conf. Mach. Learn. (ICML), vol. 119, pp. 6702–6712, 2020.

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

[19] J. M. Martínez, “Practical quasi-Newton methods for solving nonlinear systems,” J. Comput. Appl. Math., vol. 124, nos. 1–2, pp. 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 preprint arXiv:2402.06271, 2024.