Mathematical Programming
https://doi.org/10.1007/s10107-024-02143-7

FULL LENGTH PAPER

Series A

Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient

Puya Latafat[1] · **Andreas Themelis**[2] · **Lorenzo Stella**[3] · **Panagiotis Patrinos**[1]

Received: 13 January 2023 / Accepted: 10 September 2024
© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2024

Abstract

Backtracking linesearch is the de facto approach for minimizing continuously differentiable functions with locally Lipschitz gradient. In recent years, it has been shown that in the convex setting it is possible to avoid linesearch altogether, and to allow the stepsize to adapt based on a local smoothness estimate without any backtracks or evaluations of the function value. In this work we propose an adaptive proximal gradient method, adaPGM, that uses novel estimates of the local smoothness modulus which leads to less conservative stepsize updates and that can additionally cope with nonsmooth terms. This idea is extended to the primal-dual setting where an adaptive three-term primal-dual algorithm, adaPDM, is proposed which can be viewed as an extension of the PDHG method. Moreover, in this setting the "essentially" fully adaptive variant adaPDM$^+$ is proposed that avoids evaluating the linear operator norm by invoking a backtracking procedure, that, remarkably, does not require extra gradient evaluations. Numerical simulations demonstrate the effectiveness of the proposed algorithms compared to the state of the art.

Lorenzo Stella: Work done prior to joining Amazon.

✉ Puya Latafat
puya.latafat@kuleuven.be

Andreas Themelis
andreas.themelis@ees.kyushu-u.ac.jp

Lorenzo Stella
lorenzostella@gmail.com

Panagiotis Patrinos
panos.patrinos@esat.kuleuven.be

1 Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

2 Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

3 Amazon Web Services, Berlin, Germany

Published online: 28 October 2024

Keywords Convex minimization · Proximal gradient method · Primal-dual algorithms · Locally Lipschitz gradient · Linesearch-free adaptive stepsizes

Mathematics Subject Classification 65K05 · 90C06 · 90C25 · 90C30 · 90C47

1 Introduction

Backtracking linesearch is one of the most successful ideas in smooth optimization. It is well known that gradient descent with linesearch converges under mild differentiability assumptions [8, §1.2]. Even under Lipschitz gradient continuity such techniques can lead to significant speedups compared to using a constant stepsize dictated by a global Lipschitz modulus due to their ability to adapt to the local geometry of the problem. In this work we explore an alternative approach which can cope with nonsmooth formulations, does not require any backtracking procedures or function value evaluations, and yet only requires *local* Lipschitz gradient continuity of the differentiable term. The key property that allows for this improvement is the assumption of convexity.

Our work is inspired by [47] where an adaptive gradient method $x^{k+1} = x^k - \gamma_{k+1}\nabla f(x^k)$ is studied with stepsizes updated by the rule

$$\gamma_{k+1} = \min\left\{\gamma_k\sqrt{1+\tfrac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{2L_k}\right\}, \tag{1.1}$$

where

$$L_k := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|} \tag{1.2}$$

is a local Lipschitz estimate of ∇f. The idea of using an estimate for the Lipschitz modulus has also been explored in the setting of variational inequalities [10, 12, 64, 69, 70], but often at the cost of enforcing the stepsize sequence to be nonincreasing, which can lead to slow convergence. Allowing the stepsize to increase is a crucial feature of (1.1) which our proposed methods maintain. It is worth noting that in [46] another adaptive scheme, aGRAAL, was proposed for hemivariational inequalities which also allows for increasing stepsizes (see also [1]). Recently, in the setting of the gradient method for Lipschitz smooth minimization, [31] advanced an interesting choice of stepsizes according to predefined cyclic patterns. A similar idea appears in [2] that adopts nonrepeating *fractal-like* patterns. These methods provably yield improved worst-case rates over the standard gradient method but are bound to globally Lipschitz smooth problems. Moreover, they incorporate predetermined stepsize sequences agnostic to the local geometry of the cost function. Interestingly, adaPGM, the adaptive scheme presented in this paper, as well as the one in [47], automatically leads to sequences of large stepsizes that exhibit a seemingly cyclic behavior, see Fig. 4 and the discussion in Sect. 4.3. We also mention recent works [51, 63] whose adaptive rules are designed to guarantee worst-case rates, and [3] that exploits a continuous-time viewpoint to develop adaptive algorithms.

In [47] it was observed that the line of proof therein does not provide any route for generalization to the composite proximal setting. Additionally to showing that this is in fact possible, in this work we will actually provide larger stepsizes. As better detailed in the discussion before Sect. 2.2, the improvement is partially due to tighter estimates of the geometry of f in which both Lipschitz and (inverse) cocoercivity estimates are taken into account, namely

$$\ell_k := \frac{\langle \nabla f(x^{k-1}) - \nabla f(x^k), x^{k-1} - x^k \rangle}{\|x^{k-1} - x^k\|^2} \tag{1.3a}$$

and

$$c_k := \frac{\|\nabla f(x^{k-1}) - \nabla f(x^k)\|^2}{\langle \nabla f(x^{k-1}) - \nabla f(x^k), x^{k-1} - x^k \rangle} \tag{1.3b}$$

for a pair of points $x^{k-1}, x^k \in \mathbb{R}^n$. Noting that the denominator of ℓ_k or c_k is zero iff $\nabla f(x^k) - \nabla f(x^{k-1}) = 0$ (the latter owing to the Baillon-Haddad theorem [4, Cor. 10]; see Lemma A.2 for the local version needed here), we stick to the convention $\frac{0}{0} = 0$ so that both ℓ_k and c_k are (well-defined, positive) real numbers. Throughout, we shall also adhere to $\frac{1}{0} = \infty$. Note that ℓ_k and c_k are the inverse of the Barzilai-Borwein stepsize choices [5], which have been considered in the setting of gradient descent [21, 56] but whose convergence results are limited to the quadratic setting (see also [14, 62] for extensions). Our proposed adaptive proximal gradient scheme adaPGM (Algorithm 2.1) combines the two estimates and involves the update rule

$$\gamma_{k+1} = \min\left\{ \gamma_k \sqrt{1 + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \frac{\gamma_k}{2\sqrt{[\gamma_k \ell_k (\gamma_k c_k - 1)]_+}} \right\} \tag{1.4}$$

on the stepsize. Note that whenever $\gamma_k c_k \leq 1$ the update reduces to $\gamma_{k+1} = \gamma_k \sqrt{1 + \frac{\gamma_k}{\gamma_{k-1}}}$, effectively strictly increasing the stepsize. Regardless, this update is easily seen to be less conservative than (1.1), the one prescribed in [47, Alg. 1]; see Remark 2.4 for the details. We also point out the recent follow-up work [48] of [47], subsequent to the preprint version of our manuscript, that also considers the proximal gradient setting (with a different stepsize update) and improves the second term in (1.1) by a factor of $\sqrt{2}$ in the smooth case.

In the second part of the paper this idea is extended to the primal-dual setting to address more general problems of the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \varphi(x) := f(x) + g(x) + h(Ax), \tag{1.5}$$

where A is a linear mapping, g and h are (possibly nonsmooth) extended-real-valued convex functions, and f is a convex function typically assumed to have Lipschitz continuous gradient (this is relaxed to local Lipschitz continuity here, cf. Assumption II).

In the past decade primal-dual algorithms have gained a lot of popularity in areas ranging from machine learning and signal processing to control [18, 34, 35, 37, 38, 61]. Their popularity is primarily due to their ability to achieve *full splitting* on composite problems of the form (1.5). Moreover, inherent properties of first-order operations facilitate block-coordinate and distributed variants, see for instance [9, 27, 41–43] and the references therein.

There is a large body of literature on primal-dual algorithms; see, e.g., [11, 15, 20, 26, 38, 68]. Despite employing different techniques in their convergence analysis, the majority of existing methods rely on establishing a Fejér-type inequality. In fact, most can be viewed as intelligent applications of a monotone splitting technique such as forward-backward, Douglas-Rachford, and forward-backward-forward splittings for solving the associated primal-dual optimality conditions; see, e.g., [11, 19, 20, 33, 66]. More recently, the introduction of new splitting techniques such as AFBA [39, 40], NOFOB [28], forward-Douglas-Rachford-forward [59], forward-backward-half forward [13], forward-reflected-backward [50], has led to new primal-dual algorithms. We remark also that when $A = \mathrm{id}$ in (1.5) one can directly solve the problem without any lifting by using the three-term splitting [22]. There exists also an adaptive/linesearch variant of this algorithm (see, e.g., [54]) which however requires potentially costly extra gradient evaluations during the backtracking procedure.

Unlike the above described correspondence with splitting techniques, our proposed method cannot be viewed as an instance of any splitting technique for solving *general* monotone inclusions, in that it relies heavily on the knowledge that the operators involved are subdifferentials of convex functions. Although the proposed idea is extendable to other primal-dual methods such as those in [39], our focus here is on an adaptive variant of PDHG [15, 20, 66]. An interesting algorithm in this line of work was proposed in [65] which however cannot handle the third nonsmooth term g in (1.5). In adaPDM (Algorithm 3.1) we provide a different stepsize rule that not only can handle (1.5) but also inherits the same idea of using tighter estimates c_k, ℓ_k as in the case of the proximal gradient method.

A second consideration for primal-dual methods is that in the usual (nonadaptive) setting the primal-dual stepsizes γ, σ should typically satisfy a condition of the form $\gamma\sigma\|A\|^2 \leq 1 - \frac{\gamma L_f}{2}$ (see for instance [20, Thm. 3.1] and [39, Prop. 5.1]). In practical applications the norm of the linear operator may be costly to compute and can lead to smaller stepsizes, and thus slower convergence. Recently, in the setting where $f \equiv 0$, a linesearch procedure was proposed in [49] for selecting the stepsizes based on an estimate of the norm of the linear operator. A linesearch extension of [46] in the primal-dual setting was also proposed in [17]. Instead, the linesearch procedure that we propose naturally integrates our adaptive primal-dual algorithm adaPDM$^+$ (Algorithm 3.2) to handle the more general problem (1.5) without any extra gradient evaluations during the backtracks.

1.1 Contributions

The main contributions of the paper are summarized below.

1. We propose a nonmonotone adaptive stepsize rule for the proximal gradient method that departs from the usual linesearch technique. In contrast to backtracking linesearch, the new approach eliminates the need for backtracks or function value evaluations altogether. More importantly, the proposed algorithm does not require any parameter tuning and can quickly recover from a bad stepsize initialization. This is achieved by adapting the stepsize to the local geometry of the smooth function, combining local estimates of cocoercivity and Lipschitz moduli of the differentiable term along the last two iterates. Compared to [47], even when restricted to the case of gradient descent, the proposed approach allows for less restrictive stepsizes. Through this observation, convergence of the aforementioned work in the proximal case follows immediately as a by-product of our analysis.
2. This idea is extended to the primal-dual setting where an adaptive three-term splitting for composite minimization problems is developed. The proposed algorithm can be viewed as an extension of the Condat-Vũ algorithm [20, 66], which in turn is an extension of the PDHG algorithm [15].
3. As a final contribution, an "essentially" fully adaptive variant of the primal-dual method is presented. This is meant in the sense that it no longer requires evaluating the norm of the linear operator A, and is thus "fully" adaptive, but only "essentially" so, for all this comes at the expense of performing a backtracking to potentially correct the given (local) estimates. Remarkably nevertheless, the proposed linesearch does not require any extra gradient evaluations and can thus be implemented efficiently.

1.2 Organization

We conclude this section by introducing the adopted notation. The proposed adaptive proximal gradient method adaPGM is formally studied in Sect. 2. The underlying idea is then extended to the primal-dual setting in Sect. 3, where adaPDM is presented that can handle one additional nonsmooth term composed with a linear operator. The issue of estimating the norm of the linear operator is resolved through the introduction of a linesearch procedure in Sect. 3.2. The convergence results for both variants of the primal-dual algorithm are presented in a unified fashion in Sect. 3.3 with some of the proofs deferred to Appendix A. Numerical simulations for the proposed algorithms are presented in Sect. 4, together with a commentary on some empirical observations. Section 5 concludes the paper.

1.3 Notation

The set of real and extended-real numbers are $\mathbb{R} := (-\infty, \infty)$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, while the positive and strictly positive reals are $\mathbb{R}_+ := [0, \infty)$ and $\mathbb{R}_{++} := (0, \infty)$. We use the notation $[x]_+ = \max\{0, x\}$. With id we indicate the identity function defined on a suitable space. We denote by $\langle \cdot\,, \cdot \rangle$ and $\|\cdot\|$ the standard Euclidean inner product and the induced norm. Given a set $\mathcal{D} \subseteq \mathbb{R}^n$, with $\operatorname{int} \mathcal{D}$, $\operatorname{relint} \mathcal{D}$ and

bdry $\mathcal{D}$ we respectively denote its interior, relative interior, and boundary, and for a sequence $(x^k)_{k\in\mathbb{N}}$ we write $(x^k)_{k\in\mathbb{N}} \subseteq \mathcal{D}$ to indicate that $x^k \in \mathcal{D}$ for all $k \in \mathbb{N}$. The diameter of $\mathcal{D}$ is $\operatorname{diam}\mathcal{D} := \sup\{\|x-y\| \mid x, y \in \mathcal{D}\}$, while its indicator function is $\iota_{\mathcal{D}}$, namely $\iota_{\mathcal{D}}(x) = 0$ if $x \in \mathcal{D}$ and ∞ otherwise. Given two sets $\mathcal{D}, \mathcal{D}' \subseteq \mathbb{R}^n$, the notation $\mathcal{D} + \mathcal{D}' := \left\{x + x' \mid x \in \mathcal{D},\ x' \in \mathcal{D}'\right\}$ indicates their Minkowski sum.

The notation $\overline{\mathrm{B}}(\bar{x}; r) := \{x \mid \|x - \bar{x}\| \leq r\}$ indicates the closed ball centered at $\bar{x}$ and with radius r. The *domain* and *epigraph* of an extended-real-valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ are the sets $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$ and $\operatorname{epi} h := \{(x, c) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \leq c\}$. Function h is said to be *proper* if $\operatorname{dom} h \neq \emptyset$, and *lower semicontinuous (lsc)* if $\operatorname{epi} h$ is a closed subset of $\mathbb{R}^{n+1}$. We say that h is *level bounded* if its c-sublevel set $\operatorname{lev}_{\leq c} h := \{x \in \mathbb{R}^n \mid h(x) \leq c\}$ is bounded for all $c \in \mathbb{R}$. The *conjugate* of h is defined by $h^*(y) := \sup_{x\in\mathbb{R}^n}\{\langle y, x\rangle - h(x)\}$.

We say that a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ has *locally Lipschitz continuous gradient* if for every convex and compact set $\mathcal{V} \subset \mathbb{R}^n$ there exists $L_{h,\mathcal{V}} > 0$ such that $\|\nabla h(x) - \nabla h(x')\| \leq L_{h,\mathcal{V}}\|x - x'\|$ holds for all $x, x' \in \mathcal{V}$.[1]

2 Adaptive proximal gradient method

The proximal gradient method (PGM) is the natural extension of gradient descent for constrained and nonsmooth problems. It addresses nonsmooth minimization problems by splitting them into the sum of two terms as follows:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \varphi(x) := f(x) + g(x). \tag{2.1}$$

Throughout this section the following underlying assumptions are imposed on problem (2.1).

Assumption I (Requirements for the proximal gradient setting) The following hold in problem (2.1):

A1 $f : \mathbb{R}^n \to \mathbb{R}$ is convex and has locally Lipschitz continuous gradient.

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

A3 A solution exists: $\arg\min \varphi \neq \emptyset$.

In addition to the gradient of the differentiable term, the fundamental oracle of PGM is the *proximal mapping* [7, Def. 6.1] defined as

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

where $\tau > 0$ is a given stepsize. In the convex setting the proximal map is single valued and in fact Lipschitz continuous. It is well known that for many applications of interest such as constrained or regularized problems the nonsmooth term admits

[1] By virtue of [58, Thm. 9.2] this condition is equivalent to *strict continuity* of ∇h at every point, in the sense of [58, Def. 9.1].

Algorithm 2.1 Adaptive proximal gradient method (adaPGM)

REQUIRE starting point $x^{-1} \in \mathbb{R}^n$ and stepsizes $\gamma_0 \geq \gamma_{-1} > 0$

INITIALIZE $x^0 = \operatorname{prox}_{\gamma_0 g}(x^{-1} - \gamma_0 \nabla f(x^{-1}))$

REPEAT FOR $k = 0, 1, \ldots$ until convergence

2.1.1: With ℓ_k and c_k as in (1.3), define the stepsize as // $[r]_+ := \max\{r, 0\}$

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{1 + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{2\sqrt{[\gamma_k \ell_k (\gamma_k c_k - 1)]_+}}\right\}$$

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

closed form proximal operator (e.g., projection on sets, shrinkage operator, etc.). The most common variant of PGM involves a constant stepsize that is upper bounded by $2/L_f$, where L_f is the global Lipschitz constant of ∇f. A common strategy in practice is to estimate such modulus via backtracking linesearch.

2.1 Algorithmic overview

Diverging from the linesearch technique, we propose adaPGM that adaptively selects the stepsizes based on the estimates c_k, ℓ_k as in (1.3). Note that an adaptive gradient method with $g \equiv 0$ was proposed in [47] with a different update rule, see (1.1), where it was observed that the adopted line of proof *does not seem to provide any route for generalization* to account for a nonsmooth term. This appears to be fundamentally due to the fact that the analysis therein revolves around the fact that the difference of consecutive iterates is a multiple of the gradient. In contrast, we circumvent this by combining the subgradient inequality for the nonsmooth term g at *three* different pairs of points, namely $(x^\star, x^{k+1})$, (x^{k+1}, x^k), and (x^{k-1}, x^k). In addition, the combined use of the quantities ℓ_k and c_k as in (1.3) allows for estimating, along with that of the gradient, the local Lipschitz constant of the forward operator $\mathrm{id} - \gamma_k \nabla f$. This appears to be fundamental for recovering the update rule (1.1) of [47], and in fact leads to the less conservative update rule of adaPGM.

2.1.1 Initialization and practical considerations

AdaPGM has the ability to recover from small stepsizes which may be due to bad initialization, or stumbling upon steep/ill-conditioned regions, by linearly increasing the stepsize (by a factor of at least $\sqrt{2}$) until a value proportional to the inverse of a local Lipschitz estimate is attained; see the proof of Theorem 2.3*(ii)* for the details. On the other extreme, too large an initial stepsize is corrected in one iteration thanks to the second term in (1.4). Nevertheless, this event can also result in the very first algorithmic step being pushed far away without control.

To mitigate such scenarios at initialization, γ_0 can be refined by running offline proximal gradient updates; starting from the initial point x_{-1}, γ_0 can be updated by the inverse of either one of the quantities in (1.3) or (1.2) evaluated between x^{-1} and the

prox-grad point. If the updated stepsize is orders of magnitude smaller than the original one, the same procedure may be repeated an additional time. This procedure is helpful even for problems with globally Lipschitz gradient continuity. Once a reasonable γ_0 is obtained, we suggest selecting γ_{-1} small enough such that $\gamma_0\sqrt{1+\gamma_0/\gamma_{-1}} \geq \frac{1}{2L_0}$, ensuring that $\gamma_1 \geq \frac{1}{2L_0}$. It is important to note that this choice of γ_{-1} doesn't affect the convergence results of Theorem 2.3. In fact, with this initialization Theorem 2.3*(ii)* holds with $k_0 = 1$.

2.2 Preliminary lemmas

Throughout, we will make use of the following shorthand for the forward operator with stepsize γ_k:

$$H_k := \mathrm{id} - \gamma_k \nabla f.$$

The subgradient characterization of the proximal step implies that (see (2.4))

$$\tfrac{1}{\gamma_k}\big(H_k(x^{k-1}) - H_k(x^k)\big) \in \partial\varphi(x^k). \tag{2.2}$$

This quantity plays an important role in our analysis. As we are about to show, the combined adoption of the estimates ℓ_k and c_k provides an estimate of the Lipschitz modulus of not only ∇f but also the forward operator H_k.

Lemma 2.1 *Suppose that Assumption I holds, and let $x^{k-1}, x^k \in \mathbb{R}^n$. Then, with L_k, ℓ_k, and c_k as in (1.2) and (1.3) the following hold:*

(i) $\|\nabla f(x^{k-1}) - \nabla f(x^k)\|^2 = c_k\ell_k\|x^{k-1} - x^k\|^2$, *that is,* $c_k\ell_k = L_k^2$.

(ii) $\|H_k(x^{k-1}) - H_k(x^k)\|^2 = \big(1 - \gamma_k\ell_k(2-\gamma_k c_k)\big)\|x^{k-1} - x^k\|^2$.

(iii) $\ell_k \leq L_k \leq L_{f,\mathcal{V}}$ *and* $L_k \leq c_k$, *where $L_{f,\mathcal{V}}$ is a Lipschitz modulus for ∇f on a compact convex set $\mathcal{V}$ containing x^{k-1} and x^k.*

Proof The first assertion is of trivial verification, and similarly the third one follows from the Cauchy-Schwarz inequality and the Baillon-Haddad theorem [4, Cor. 10], see also [6, Cor. 18.17]. To conclude, observe that

$$\begin{aligned}\|H_k(x^{k-1}) - H_k(x^k)\|^2 = \|x^{k-1} - x^k\|^2 + \gamma_k^2\|\nabla f(x^{k-1}) - \nabla f(x^k)\|^2 \\ - 2\gamma_k\langle \nabla f(x^{k-1}) - \nabla f(x^k), x^{k-1} - x^k\rangle,\end{aligned}$$

from which assertion *(ii)* follows. □

Our convergence analysis will rely on first establishing boundedness of the generated sequence, thereby entailing the existence of a Lipschitz constant $L_{f,\mathcal{V}} > 0$ for ∇f on a bounded convex set $\mathcal{V}$ that contains the iterates. It will then follow from Lemma 2.1*(iii)* that both L_k and ℓ_k are upper bounded by this quantity, which in turn will be used to show that this modulus provides a lower bound for the stepsize separating it from zero. Before that, we show how the combined use of ℓ_k and c_k can be employed to estimate the progress of the iterates generated by PGM with arbitrary

stepsizes, not necessarily dictated by the update rule of adaPGM. To simplify the presentation, we introduce the following notation:

$$\rho_k := \tfrac{\gamma_k}{\gamma_{k-1}} \quad \text{and} \quad P_k := \varphi(x^k) - \min \varphi. \tag{2.3}$$

Lemma 2.2 *Suppose Assumption I holds, and consider a sequence $(x^k)_{k\in\mathbb{N}}$ generated by PGM iterations $x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k))$. Then*

$$\begin{aligned}&\tfrac{1}{2}\|x^{k+1} - x^\star\|^2 + \gamma_{k+1}(1+\rho_{k+1})P_k + \tfrac{1}{4}\|x^k - x^{k+1}\|^2 \\ &\le \tfrac{1}{2}\|x^k - x^\star\|^2 + \rho_{k+1}\gamma_{k+1}P_{k-1} - \rho_{k+1}^2\gamma_k\ell_k(1-\gamma_k c_k)\|x^{k-1} - x^k\|^2\end{aligned}$$

holds for any $k \ge 1$ and $x^\star \in \arg\min\varphi$, where ℓ_k and c_k are as in (1.3), and ρ_k and P_k as in (2.3).

Proof The subgradient characterization

$$\tfrac{H_{k+1}(x^k) - x^{k+1}}{\gamma_{k+1}} = \tfrac{x^k - x^{k+1}}{\gamma_{k+1}} - \nabla f(x^k) \in \partial g(x^{k+1}) \tag{2.4}$$

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

$$\begin{aligned}0 &\le g(x^\star) - g(x^{k+1}) + \langle \nabla f(x^k), x^\star - x^{k+1}\rangle - \tfrac{1}{\gamma_{k+1}}\langle x^k - x^{k+1}, x^\star - x^{k+1}\rangle \\ &= g(x^\star) - g(x^{k+1}) + \underbrace{\langle \nabla f(x^k), x^\star - x^{k+1}\rangle}_{\text{(A)}} + \tfrac{1}{2\gamma_{k+1}}\|x^k - x^\star\|^2 \\ &\quad - \tfrac{1}{2\gamma_{k+1}}\|x^{k+1} - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x^k - x^{k+1}\|^2\end{aligned}$$

holds for any solution $x^\star$. We next proceed to upper bound the term (A) as

$$\begin{aligned}\text{(A)} &= \langle \nabla f(x^k), x^\star - x^k\rangle + \langle \nabla f(x^k), x^k - x^{k+1}\rangle \\ &= \langle \nabla f(x^k), x^\star - x^k\rangle + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - x^k, x^{k+1} - x^k\rangle \\ &\qquad + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle \\ &\overset{(2.4)}{\le} f(x^\star) - f(x^k) + g(x^{k+1}) - g(x^k) \\ &\qquad + \underbrace{\tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle}_{\text{(B)}}.\end{aligned}$$

We bound the term (B) by Young's inequality with parameter ε_{k+1} as

$$\begin{aligned}\text{(B)} &\le \tfrac{\varepsilon_{k+1}}{2\gamma_k}\|x^k - x^{k+1}\|^2 + \tfrac{1}{2\varepsilon_{k+1}\gamma_k}\|H_k(x^{k-1}) - H_k(x^k)\|^2 \\ &\overset{2.1(ii)}{=} \tfrac{\varepsilon_{k+1}}{2\gamma_k}\|x^k - x^{k+1}\|^2 + \tfrac{1-\gamma_k\ell_k(2-\gamma_k c_k)}{2\varepsilon_{k+1}\gamma_k}\|x^{k-1} - x^k\|^2.\end{aligned} \tag{2.5}$$

Let $\varphi_\star := \min \varphi$. The three inequalities combined give

$$0 \leq \varphi_\star - \varphi(x^k) + \tfrac{1}{2\gamma_{k+1}}\|x^k - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x^{k+1} - x^\star\|^2 + \left(\tfrac{\varepsilon_{k+1}}{2\gamma_k} - \tfrac{1}{2\gamma_{k+1}}\right)\|x^k - x^{k+1}\|^2 + \tfrac{1-\gamma_k\ell_k(2-\gamma_k c_k)}{2\varepsilon_{k+1}\gamma_k}\|x^{k-1} - x^k\|^2.$$

Using again the subgradient (2.4) (since $\partial\varphi = \nabla f + \partial g$) one has

$$v^k := \tfrac{x^{k-1}-x^k}{\gamma_k} - (\nabla f(x^{k-1}) - \nabla f(x^k)) \in \partial\varphi(x^k), \tag{2.6}$$

hence, for any $\vartheta_{k+1} \geq 0$,

$$\begin{aligned} 0 &\leq \vartheta_{k+1}\Big(\varphi(x^{k-1}) - \varphi(x^k) - \langle v^k, x^{k-1} - x^k\rangle\Big) \\ &= \vartheta_{k+1}\Big(\varphi(x^{k-1}) - \varphi(x^k) - \tfrac{1}{\gamma_k}\|x^k - x^{k-1}\|^2 + \langle \nabla f(x^{k-1}) - \nabla f(x^k), x^{k-1} - x^k\rangle\Big) \\ &= \vartheta_{k+1}\Big(\varphi(x^{k-1}) - \varphi(x^k) - \tfrac{1-\gamma_k\ell_k}{\gamma_k}\|x^k - x^{k-1}\|^2\Big). \end{aligned} \tag{2.7}$$

By summing the last two inequalities, multiplying by γ_{k+1}, and observing that

$$\varphi_\star - \varphi(x^k) + \vartheta_{k+1}(\varphi(x^{k-1}) - \varphi(x^k)) = \vartheta_{k+1}P_{k-1} - (1+\vartheta_{k+1})P_k,$$

we obtain

$$\begin{aligned} &\tfrac{1}{2}\|x^{k+1} - x^\star\|^2 + \gamma_{k+1}(1+\vartheta_{k+1})P_k + \tfrac{1-\varepsilon_{k+1}\rho_{k+1}}{2}\|x^k - x^{k+1}\|^2 \\ &\quad\leq \tfrac{1}{2}\|x^k - x^\star\|^2 + \vartheta_{k+1}\gamma_{k+1}P_{k-1} \\ &\qquad + \rho_{k+1}\Big(\tfrac{1-\gamma_k\ell_k(2-\gamma_k c_k)}{2\varepsilon_{k+1}} - \vartheta_{k+1}(1-\gamma_k\ell_k)\Big)\|x^{k-1} - x^k\|^2. \end{aligned} \tag{2.8}$$

Selecting $\vartheta_{k+1} = \rho_{k+1}$ and $\varepsilon_{k+1} = 1/2\rho_{k+1}$ results in the claimed inequality. □

As detailed in the proof, the inequality in Lemma 2.2 is a special case of the more general (2.8) obtained by setting $\vartheta_k = \rho_k$ and $\varepsilon_{k+1} = 1/2\rho_{k+1}$ for any k. As we are about to see in the following Theorem 2.3, these choices strike a nice balance between simplifying inequality (2.8) and enabling large stepsizes: the result is a rather simple update rule for the stepsize that works very well in practice. It would be tempting to explore if more sophisticated tunings of these parameters could lead to further improvements, an aspect that we believe is worth investigating in the future, see Remark 2.6.

2.3 Convergence results

While adaPGM can be seen as a special case of the more general primal-dual adaPDM, the convergence results of adaPGM is obtained under fewer restrictions (cf. Section 3.1.3). For this reason, we provide a dedicated proof for the adaptive proximal gradient algorithm.

Theorem 2.3 *Suppose that Assumption I holds, and consider the iterates generated by adaPGM. Then, for any $x^\star \in \arg\min \varphi$, with ρ_k and P_k as in (2.3) and with $\mathcal{U}_k(x^\star)$ defined as*

$$\mathcal{U}_k(x^\star) := \tfrac{1}{2}\|x^k - x^\star\|^2 + \tfrac{1}{4}\|x^k - x^{k-1}\|^2 + \gamma_k(1+\rho_k)P_{k-1},$$

the following hold:

(i) For all $k \geq 1$, $\mathcal{U}_{k+1}(x^\star) \leq \mathcal{U}_k(x^\star) - \overbrace{\left(\tfrac{1}{4} - \rho_{k+1}^2\gamma_k\ell_k(\gamma_k c_k - 1)\right)}^{\geq 0}\|x^{k-1} - x^k\|^2$ $-\gamma_k\overbrace{\left(1 + \rho_k - \rho_{k+1}^2\right)}^{\geq 0}P_{k-1}$.

(ii) The sequence $(x^k)_{k\in\mathbb{N}}$ is bounded and $\gamma_k > \frac{1}{2L_{f,\mathcal{V}}}$ holds for every $k \geq k_0 := \left\lfloor 2\log_2 \frac{1}{2\gamma_0 L_{f,\mathcal{V}}} \right\rfloor_+$, where $L_{f,\mathcal{V}}$ is a Lipschitz modulus for ∇f on a compact convex set $\mathcal{V}$ containing $(x^k)_{k\in\mathbb{N}}$.

(iii) The sequence $(x^k)_{k\in\mathbb{N}}$ converges to a solution. The claim remains true if the stepsizes are chosen in such a way that (1.4) holds with "$\leq$", as long as $(\gamma_k)_{k\in\mathbb{N}}$ is bounded away from zero.

(iv) The following rate holds

$$\min_{k=0,1,\ldots,K}(\varphi(x^k) - \min\varphi) \leq \frac{\mathcal{U}_1(x^\star)}{\sum_{k=1}^{K+1}\gamma_k},$$

which by assertion (ii) implies a sublinear $\mathcal{O}(1/K)$ convergence rate on the best-so-far cost.

Proof

♠ *(i)* Expressed in terms of $\mathcal{U}_k(x^\star)$, the inequality in Lemma 2.2 simplifies to the one in the statement. The update rule for γ_k implies that the coefficients indicated in the statement are greater than or equal to zero, and the claim follows.

♠ *(ii)* The proven inequality implies that $(\mathcal{U}_k(x^\star))_{k\in\mathbb{N}}$ converges for any $x^\star \in \arg\min\varphi$. Taking into account the nonnegativity of the last two terms in the definition of $\mathcal{U}_k(x^\star)$, it follows that $(x^k)_{k\in\mathbb{N}}$ is bounded. As such, $\mathcal{V}$ and $L_{f,\mathcal{V}}$ as in the statement exist. Observe that

$$\frac{\gamma_k}{2\sqrt{[\gamma_k\ell_k(\gamma_k c_k - 1)]_+}} = \frac{1}{2}\sqrt{\frac{1}{\ell_k c_k}\frac{\gamma_k c_k}{[\gamma_k c_k - 1]_+}} \geq \frac{1}{2\sqrt{\ell_k c_k}} = \frac{1}{2L_k}, \tag{2.9}$$

where the last identity owes to Lemma 2.1*(i)*, and therefore

$$\gamma_{k+1} \geq \min\left\{\gamma_k\sqrt{1+\rho_k}, \tfrac{1}{2L_k}\right\} \tag{2.10}$$

holds for every k. Suppose that $\gamma_k \leq 1/2L_{f,\mathcal{V}}$ for $k = 0, \ldots, K$. Because of (2.10) and Lemma 2.1*(iii)*, this implies that $\gamma_{k+1} = \gamma_k\sqrt{1+\rho_k}$ for $k = 0, \ldots, K-1$. Then,

for these k one has that $\rho_{k+1}^2 = 1 + \rho_k > 1$, which inductively gives $\rho_k^2 \geq 2$ for all $k = 1, \ldots, K$ (since $\rho_0 \geq 1$). We then have

$$\gamma_K^2 = \rho_K^2 \gamma_{K-1}^2 \geq 2\gamma_{K-1}^2 \geq \cdots \geq 2^K \gamma_0^2, \tag{2.11}$$

showing that in at most k_0 iterations stepsize exceeds $1/2L_{f,\mathcal{V}}$. From this point, a trivial induction argument using (2.10) reveals that $\gamma_k \geq 1/2L_{f,\mathcal{V}}$ for all $k \geq k_0$.

♠ *(iii)* We begin by observing that, as is apparent from its proof, assertion *(i)* remains valid if the identity in (1.4) is replaced by an inequality "$\leq$". Either way, a telescoping argument yields that

$$\gamma_k(1 + \rho_k - \rho_{k+1}^2)P_{k-1} \to 0 \quad \text{as} \quad k \to \infty. \tag{2.12}$$

We proceed by intermediate claims.

Claim 2.3.1 $\liminf_{k\to\infty} P_k = 0$*, and in particular the sequence* $(x^k)_{k\in\mathbb{N}}$ *admits a limit point* $x_\infty \in \arg\min \varphi$.
Since $(x^k)_{k\in\mathbb{N}}$ is bounded and φ is lsc, it suffices to show that $\liminf_{k\to\infty} P_k = 0$. If $\limsup_{k\to\infty}(1 + \rho_k - \rho_{k+1}^2) > 0$, then (2.12) yields the claim, since γ_k is bounded away from zero. Alternatively, since $1+\rho_k-\rho_{k+1}^2 \geq 0$, necessarily $1+\rho_k-\rho_{k+1}^2 \to 0$, from which it easily follows that $\liminf_{k\to\infty} \rho_k > 1$, and that therefore $\gamma_k \to \infty$; in this case, because of the inequalities $\gamma_k(1 + \rho_k)P_{k-1} \leq \mathcal{U}_k \leq \mathcal{U}_1$, this directly proves that $P_k \to 0$.

Claim 2.3.2 x_∞ *is the only limit point that belongs to* $\arg\min \varphi$.
Suppose that $x'_\infty \in \arg\min \varphi$ is a limit point. Observe that

$$\langle x^k, x_\infty - x'_\infty\rangle = \mathcal{U}_k(x'_\infty) - \mathcal{U}_k(x_\infty) + \tfrac{1}{2}\|x_\infty\|^2 - \tfrac{1}{2}\|x'_\infty\|^2,$$

and since $(\mathcal{U}_k(x^\star))_{k\in\mathbb{N}}$ is convergent for all $x^\star \in \arg\min \varphi$, then so is $(\langle x^k, x_\infty - x'_\infty\rangle)_{k\in\mathbb{N}}$. Passing to the limit along the two converging subsequences thus yields $\langle x_\infty, x_\infty - x'_\infty\rangle = \langle x'_\infty, x_\infty - x'_\infty\rangle$, which after rearranging results in $\|x_\infty - x'_\infty\|^2 = 0$, establishing the claim.

Claim 2.3.3 $(x^k)_{k\in\mathbb{N}}$ *converges to a point in* $\arg\min \varphi$.
Let x_∞ be the (unique) optimal limit point of $(x^k)_{k\in\mathbb{N}}$ as in Claim 2.3.2, and let

$$U := \lim_{k\to\infty} \mathcal{U}_k(x_\infty), \tag{2.13}$$

which exists since the sequence is monotonically decreasing by assertion *(i)*. Because of the inequality $\frac{1}{2}\|x^k - x_\infty\|^2 \leq \mathcal{U}_k(x_\infty)$, it suffices to show that $U = 0$. Observe that, for every k,

$$\rho_k \leq \rho_{\max} \quad \text{where} \quad \rho_{\max} := \max\left\{\tfrac{\gamma_0}{\gamma_{-1}}, \tfrac{1}{2}(1 + \sqrt{5})\right\} > 1, \tag{2.14}$$

as it can be verified by induction. We consider two mutually exclusive scenarios.

$\diamond$ Suppose that $(\gamma_k)_{k\in\mathbb{N}}$ is bounded, and consider a subsequence $(x^k)_{k\in K}$ such that $\lim_{K\ni k\to\infty} P_k = 0$, which exists and converges to x_∞ by virtue of the previous claims. Extract $K' \subseteq K$ such that $(\gamma_{k+1})_{k\in K'} \to \gamma$ for some $\gamma > 0$. Then, $x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k)) \to \operatorname{prox}_{\gamma g}(x_\infty - \gamma\nabla f(x_\infty)) = x_\infty$ as $K' \ni k \to \infty$ as well. This yields the sought identity

$$\begin{aligned} U &= \lim_{k\to\infty} \mathcal{U}_k(x_\infty) = \lim_{K'\ni k\to\infty} \mathcal{U}_{k+1}(x_\infty) \\ &= \lim_{K'\ni k\to\infty} \left(\tfrac{1}{2}\|x^{k+1} - x_\infty\|^2 + \tfrac{1}{2}\|x^{k+1} - x^k\|^2 + \gamma_{k+1}(1+\rho_{k+1})P_k \right) \\ &= 0, \end{aligned}$$

where the vanishing of the last term owes to boundedness of $(\gamma_k)_{k\in\mathbb{N}}$ and $(\rho_k)_{k\in\mathbb{N}}$.
$\diamond$ Suppose instead that $(\gamma_k)_{k\in\mathbb{N}}$ is unbounded, and let $(\gamma_k)_{k\in K'}$ be an arbitrary subsequence such that $\gamma_k \to \infty$ as $K' \ni k \to \infty$. From the inequality $\gamma_k P_{k-1} \le \mathcal{U}_k(x_\infty) \le \mathcal{U}_1(x_\infty) < \infty$, we conclude that

$$P_{k-1} \to 0 \ \text{ and } \ x^{k-1} \to x_\infty \text{ as } K' \ni k \to \infty \tag{2.15}$$

(recall that φ is lsc, $(x^k)_{k\in\mathbb{N}}$ is bounded, and x_∞ is the *unique* optimal limit point). Then, since by (2.14) $\gamma_{k-1} \ge \gamma_k/\rho_{\max} \to \infty$ as $K' \ni k \to \infty$, by the same argument also $x^{k-2} \to x_\infty$ as $K' \ni k \to \infty$. With U as in (2.13), we thus have

$$U = \lim_{k\to\infty} \mathcal{U}_k(x_\infty) = \lim_{K'\ni k\to\infty} \mathcal{U}_{k-1}(x_\infty) = \lim_{K'\ni k\to\infty} \gamma_{k-1}(1+\rho_{k-1})P_{k-2} \tag{2.16}$$

holding for *any* set of indices $K' \subseteq \mathbb{N}$ along which the stepsizes are divergent.
We now construct a specific subsequence $K := \{k_0, k_1, \ldots\}$ as follows: start with $k_0 = 1$, and for $i \ge 0$ let

$$k_{i+1} = \min\left\{k \ge k_i \mid \gamma_k \ge \rho_{\max}\gamma_{k_i}\right\}. \tag{2.17}$$

Then, $(\gamma_{k_i})_{i\in\mathbb{N}} \to \infty$, which as argued after (2.15) implies that both $(x^{k_i-1})_{i\in\mathbb{N}}$ and $(x^{k_i-2})_{i\in\mathbb{N}}$ converge to x_∞. Observe that $\rho_{k_i} > 1$ holds for all $i \in \mathbb{N}$ by minimality of k_i. Notice further that

$$\rho_{k_i-1} \ge \rho_{\max}^{-1} \quad \forall i \in \mathbb{N}. \tag{2.18}$$

To see why, suppose on the contrary that $\rho_{k_i-1} < \rho_{\max}^{-1}$; then in particular $\rho_{k_i-1} < 1$, thereby ensuring by the previous observation that $k_i - 1 \notin K$, and thus $k_{i-1} \le k_i - 2$. This leads to the contradiction

$$\rho_{\max}\gamma_{k_{i-1}} \overset{(2.17)}{\le} \gamma_{k_i} \overset{(2.14)}{\le} \rho_{\max}\gamma_{k_i-1} = \rho_{\max}\rho_{k_i-1}\gamma_{k_i-2} < \gamma_{k_i-2} < \rho_{\max}\gamma_{k_{i-1}},$$

where the last inequality follows either in case $k_i - 2 = k_{i-1}$ (since $\rho_{\max} > 1$) or when $k_{i-1} < k_i - 2 < k_i$ (from minimality in the definition of k_i). This shows (2.18).

Next, by the stepsize update at step 2.1.1,

$$\begin{aligned}\tfrac{1}{\rho_{\max}}\gamma_{k_i-2}L_{k_i-2} \overset{(2.18)}{\leq} \gamma_{k_i-1}L_{k_i-2} &\leq \frac{\gamma_{k_i-2}L_{k_i-2}}{2\sqrt{\left[\gamma_{k_i-2}^2\ell_{k_i-2}c_{k_i-2}-\gamma_{k_i-2}\ell_{k_i-2}\right]_+}}\\ &\leq \frac{\gamma_{k_i-2}L_{k_i-2}}{2\sqrt{\left[\gamma_{k_i-2}^2L_{k_i-2}^2-\gamma_{k_i-2}L_{k_i-2}\right]_+}}\\ &= \frac{1}{2\sqrt{\left[1-\frac{1}{\gamma_{k_i-2}L_{k_i-2}}\right]_+}},\end{aligned}$$

where the third inequality uses the fact that $L_{k_i-2}^2 = \ell_{k_i-2}c_{k_i-2}$, and $\ell_{k_i-2} \leq L_{k_i-2}$ by Lemma 2.1. This implies that $(\gamma_{k_i-2}L_{k_i-2})_{i\in\mathbb{N}}$ must be bounded. The subgradient characterization in (2.6) and the Cauchy-Schwarz inequality then yield

$$\begin{aligned}&\gamma_{k_i-1}P_{k_i-2}\\ &\leq \rho_{k_i-1}\langle x^{k_i-2}-x_\infty, x^{k_i-2}-x^{k_i-3}-\gamma_{k_i-2}(\nabla f(x^{k_i-2})-\nabla f(x^{k_i-3}))\rangle\\ &\leq \rho_{\max}\underbrace{(1+\gamma_{k_i-2}L_{k_i-2})}_{\text{bounded}}\underbrace{\|x^{k_i-2}-x_\infty\|\|x^{k_i-2}-x^{k_i-3}\|}_{\to 0} \to 0 \quad \text{as } i\to\infty.\end{aligned}$$

Combined with the fact that $\rho_{k_i-1} \leq \rho_{\max}$, it follows from (2.16) that $U = 0$.

♠ *(iv)* A telescoping argument in the descent inequality of assertion *(i)* yields $\mathcal{U}_{K+1} + \sum_{k=1}^{K}\gamma_k\left(1+\rho_k-\rho_{k+1}^2\right)P_{k-1} \leq \mathcal{U}_1$ for $K \geq 1$. By further lower bounding $\mathcal{U}_{K+1}$ we obtain

$$\gamma_{K+1}(1+\rho_{K+1})P_K + \sum_{k=1}^{K}\gamma_k\left(1+\rho_k-\rho_{k+1}^2\right)P_{k-1} \leq \mathcal{U}_1.$$

Considering the best-so-far quantity $\min_{k=0,1,\ldots,K} P_k$, it follows that

$$\begin{aligned}\min_{k=0,1,\ldots,K} P_k &\leq \frac{\mathcal{U}_1}{\gamma_{K+1}(1+\rho_{K+1})+\sum_{k=1}^{K}\gamma_k(1+\rho_k-\rho_{k+1}^2)}\\ &= \frac{\mathcal{U}_1}{\gamma_{K+1}(1+\rho_{K+1})+\sum_{k=1}^{K}\gamma_k+\sum_{k=1}^{K}(\gamma_k\rho_k-\gamma_{k+1}\rho_{k+1})}\\ &= \frac{\mathcal{U}_1}{\gamma_1\rho_1+\sum_{k=1}^{K+1}\gamma_k} \leq \frac{\mathcal{U}_1}{(K+1)\gamma_{\min}},\end{aligned} \tag{2.19}$$

where $\gamma_{\min} := \min\left\{\gamma_0, \frac{1}{2L_{f,\mathcal{V}}}\right\}$ is a lower bound on the stepsize, as demonstrated in the proof of assertion *(ii)*. □

We remark that if the suggested initialization in Sect. 2.1.1 is used, the initial stepsize satisfies $\gamma_0 \geq 1/L_{f,\mathcal{V}}$ and consequently (2.19) holds with $\gamma_{\min} = 1/2L_{f,\mathcal{V}}$. The rate for

the best-so-far cost thus simplifies to $\min_{k=0,\dots,K} P_k \leq \frac{2\mathcal{U}_1 L_{f,\mathcal{V}}}{K+1}$ in this case. More importantly, as a consequence of Theorem 2.3 it can be easily seen that [47, Alg. 1] can in fact cope with nonsmooth problems of the form (2.1). This fact can be deduced from the inequality (2.10) in the proof of Theorem 2.3*(iii)*, as formalized next.

Remark 2.4 (Comparison with [47, Alg. 1]) With γ_{k+1} as in step 2.1.1 and L_k as in (1.2), as shown in (2.10) one has that

$$\gamma_{k+1} \geq \min\left\{\gamma_k\sqrt{1+\rho_k},\ \tfrac{1}{2L_k}\right\} \tag{2.20}$$

holds for every k. Note that the right-hand side corresponds to the stepsize update of [47, Alg. 1]. As a consequence, in addition to accommodating proximal terms, adaPGM also comes with a less conservative stepsize update rule.

Inequality (2.20) cannot be reiterated inductively, and in particular there is no guarantee that, iteration-wise, the *entire* sequence of stepsizes produced by adaPGM is larger than that generated by [47, Alg. 1], even if the algorithms are started with same initial conditions. This is nevertheless enough to infer convergence of [47, Alg. 1] applied to composite minimization problems as in (2.1).

Corollary 2.5 (Proximal extension of [47, Alg. 1]) *Suppose that Assumption I holds. Then, PGM iterations* $x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k))$ *with stepsize rule* (1.2) *converge to a solution of* (2.1).

Proof The validity of Theorem 2.3*(iii)* guarantees that the generated sequence remains bounded. As also observed in [47, Thm. 1], this guarantees that $\gamma_k \geq \frac{1}{2L_{f,\mathcal{V}}}$ (up to possibly excluding initial iterates), where $L_{f,\mathcal{V}}$ is a Lipschitz constant of ∇f on a compact convex set $\mathcal{V}$ that contains all the iterates. The proof follows by invoking Theorem 2.3*(iii)* in light of (2.20). □

Remark 2.6 (Alternative stepsize choices) The update for γ_k in adaPGM is designed to ensure descent on the Lyapunov function $\mathcal{U}_k$, cf. Theorem 2.3*(i)*. As commented after the proof of Lemma 2.2, choices of the parameter ϑ_k appearing in (2.8) other than $\vartheta_k = \rho_k$ lead to different update rules for the stepsize. For instance, retaining $\varepsilon_k = 1/2\rho_k$ but setting $\vartheta_k = \pi\rho_k$ for some $\pi > 0$ results in[2]

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\tfrac{1}{\pi} + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{2\sqrt{\left[\gamma_k\ell_k(\gamma_k c_k - 2 + \pi) + 1 - \pi\right]_+}}\right\}, \tag{2.21}$$

bringing about a trade-off between improving either term at the expense of the other. A similar concept is already pursued in [48], which advances the update

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\tfrac{2}{3} + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{\sqrt{\left[2\gamma_k^2 L_k^2 - 1\right]_+}}\right\}$$

[2] The update (2.21) is obtained by expressing (2.8) in terms of $\mathcal{U}_k^\pi(x^\star) := \frac{1}{2}\|x^k - x^\star\|^2 + \frac{1}{4}\|x^k - x^{k-1}\|^2 + \gamma_k(1+\pi\rho_k)P_{k-1}$, and enforcing descent as in Theorem 2.3*(i)*.

(unrelated to (2.21), and obtained through different arguments) remarkably improving the worst-case rate coefficient, see [48, Thm. 3 and §3.2].

3 Adaptive three-term primal-dual methods

In this section the idea of adaptively estimating the local geometry of f will be extended to composite problems of the form (1.5), which we rewrite here for the reader's convenience

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \varphi(x) := f(x) + g(x) + h(Ax).$$

Although the analysis in this setting is inevitably more complicated, the key idea of using the estimates c_k, ℓ_k in (1.3) remains the same. We will first propose an adaptive algorithm under the assumption that the norm of the linear operator A is known. This assumption will then be lifted through a certain linesearch procedure.

Problem (1.5) is studied under the following assumptions.

Assumption II (Requirements for the primal-dual setting) The following hold in problem (1.5):

A1 $f : \mathbb{R}^n \to \mathbb{R}$ is convex and has locally Lipschitz continuous gradient.

A2 $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper convex and lsc, and $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear mapping.

A3 A solution exists: $\arg\min \varphi \neq \emptyset$.

A4 The problem is strictly feasible: there exists $x \in \operatorname{relint} \operatorname{dom} g$ such that $Ax \in \operatorname{relint} \operatorname{dom} h$.

A well-established approach for addressing (1.5) is to lift it into the primal-dual space and solve the associated convex-concave saddle point problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \underset{y\in\mathbb{R}^m}{\text{maximize}}\ \mathcal{L}(x, y) := f(x) + g(x) + \langle Ax, y\rangle - h^*(y).$$

This lifting is the key to splitting the composed term $h \circ A$. Moreover, in doing so the primal and the dual solutions can be obtained simultaneously. A pair $z^\star = (x^\star, y^\star)$ will be referred to as a primal-dual solution if the following primal-dual optimality condition holds

$$0 \in -Ax^\star + \partial h^*(y^\star), \quad 0 \in A^\top y^\star + \nabla f(x^\star) + \partial g(x^\star). \tag{3.1}$$

The set of all such pairs will be denoted by $\mathcal{S}_\star$. Under the constraint qualification of Assumption II.A4, the set of solutions for the dual problem is nonempty, and thus so is $\mathcal{S}_\star$, and the duality gap is zero, see [57, Cor. 31.2.1] and [6, Thm. 19.1]. Moreover, the pair $(x^\star, y^\star)$ is a primal-dual solution if and only if $x^\star$ is a primal and $y^\star$ is a dual solution. Since (1.5) is a convex problem, the primal-dual solution pairs are equivalently characterized by the saddle point inequality

$$\mathcal{L}(x^\star, y) \le \mathcal{L}(x^\star, y^\star) \le \mathcal{L}(x, y^\star) \quad \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^m. \tag{3.2}$$

In our analysis we will measure deviation from $\mathcal{L}(x^\star, y^\star)$ along the primal and dual sequences using shorthand notations $P_k = P_k(x^\star, y^\star)$ and $Q_k = Q_k(x^\star, y^\star)$ defined as

$$P_k := \mathcal{L}(x^k, y^\star) - \mathcal{L}(x^\star, y^\star) = (f+g)(x^k) - (f+g)(x^\star) + \langle x^k - x^\star, A^\top y^\star \rangle \tag{3.3a}$$

and

$$Q_k := \mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^k) = h^*(y^k) - h^*(y^\star) + \langle Ax^\star, y^\star - y^k \rangle, \tag{3.3b}$$

which are both positive due to the saddle point inequality (3.2).

3.1 Algorithmic overview

The proposed algorithm is presented in adaPDM (Algorithm 3.1). It can be viewed as an adaptive variant of the algorithm proposed in [20, 66], which itself is an extension of the PDHG method [15]. In comparison to the aforementioned algorithms with constant stepsizes, here a varying and potentially increasing stepsize rule is proposed based on the estimates c_k, ℓ_k. Moreover, in PDHG the dual update combines two consequent primal updates as $2Ax^k - Ax^{k-1}$ (in our notation). In [39, Alg. 3] it was shown that many primal-dual algorithms can be unified by modifying the dual update to use terms of the form $\theta Ax^k + (1-\theta)Ax^{k-1}$ followed by a correction step. While depending on the application this can lead to parallel implementations and potentially larger stepsizes compared to PDHG (with $\theta = 2$), the improvement in speed is limited by the use of global estimates. Differently from the aforementioned works, in adaPDM the mixing constant θ is selected adaptively as $1 + \gamma_{k+1}/\gamma_k$. When viewed as an extension of the proximal gradient method adaPGM, this idea appears natural. In fact, this was proposed in [65] as a primal-dual extension of [47] where superior convergence rate compared to constant stepsize variants was observed. AdaPDM uses a different stepsize update rule and inherits the tighter estimates in (1.3) while permitting for the third nonsmooth term g. Notice also that it is not symmetric with respect to the primal and dual variables and a different algorithm can be obtained by applying it to the dual problem.

3.1.1 Initialization and practical considerations

AdaPDM involves a few parameters, some of which require tuning in practice. The constant ϵ is required to be strictly positive due to theoretical reasons, and in practice can be selected very close to zero (see Remark 3.1 and the discussion thereafter). The constant ν affects the second and third arguments of the min operator in step 3.1.2; too large a ν results in the second term limiting the stepsizes, while a value close to $1+\epsilon$ can potentially result in the third term becoming too small during the next iterate. In practice, we suggest a value between 1.1 and 1.5 for this constant. Possibly a more critical parameter in play is $t > 0$, which denotes the ratio between the primal

and the dual stepsizes. It is well known that this parameter can have a big impact on the performance of primal-dual methods. Even for the Condat-Vũ algorithm, where a simple stepsize condition $\gamma\sigma\|A\|^2 \leq 1 - \frac{\gamma L_f}{2}$ is available, this parameter has to be tuned in general. Having an algorithm that adaptively selects the parameter t is an open research question. Finally, the initial primal stepsize can be chosen equal to the minimum between $1/2\nu t\|A\|$ and a suitable estimate (obtained, e.g., as in Sect. 2.1.1 for adaPGM).

Algorithm 3.1 Adaptive primal-dual method (adaPDM)

REQUIRE primal/dual (square inverse) stepsize ratio $t > 0$
stepsize parameters $\epsilon > 0$ and $\nu > 1 + \epsilon$ (e.g., $\epsilon = 10^{-6}$, $\nu = 1.2$)
initial primal-dual pair $(x^{-1}, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$ and stepsizes $\gamma_{-1} \leq \gamma_0 \leq 1/2\nu t\|A\|$

INITIALIZE $x^0 = \operatorname{prox}_{\gamma_0 g}\left(x^{-1} - \gamma_0 \nabla f(x^{-1}) - \gamma_0 A^\top y^0\right)$

REPEAT FOR $k = 0, 1, \ldots$ until convergence

3.1.1: Set $\delta_k = \gamma_k \ell_k(\gamma_k c_k - 1)$ and $\xi_k = t^2\gamma_k^2\|A\|^2$, where ℓ_k and c_k are as in (1.3)

3.1.2: Define the stepsizes as

$$\begin{cases} \gamma_{k+1} = \min\left\{ \gamma_k\sqrt{1 + \frac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{2\nu t\|A\|},\ \gamma_k\sqrt{\frac{1-4\xi_k(1+\epsilon)^2}{2(1+\epsilon)\left(\sqrt{\delta_k^2 + \xi_k(1-4\xi_k(1+\epsilon)^2)} + \delta_k\right)}} \right\} \\ \sigma_{k+1} = t^2\gamma_{k+1} \end{cases}$$

3.1.3: $y^{k+1} = \operatorname{prox}_{\sigma_{k+1}h^*}\left(y^k + \sigma_{k+1}\left(\left(1 + \frac{\gamma_{k+1}}{\gamma_k}\right)Ax^k - \frac{\gamma_{k+1}}{\gamma_k}Ax^{k-1}\right)\right)$

3.1.4: $x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}\left(x^k - \gamma_{k+1}\nabla f(x^k) - \gamma_{k+1}A^\top y^{k+1}\right)$

3.1.2 Termination criteria

The primal and dual proximal updates lead to the following inclusions:

$$\begin{aligned} v_1^{k+1} &:= \tfrac{1}{\sigma_{k+1}}(y^k - y^{k+1}) + \tfrac{\gamma_{k+1}}{\gamma_k}\left(Ax^k - Ax^{k-1}\right) + \left(Ax^k - Ax^{k+1}\right) \\ &\in \partial h^*(y^{k+1}) - Ax^{k+1} \qquad (3.4a) \\ v_2^{k+1} &:= \tfrac{1}{\gamma_{k+1}}(x^k - x^{k+1}) + \nabla f(x^{k+1}) - \nabla f(x^k) \\ &\in \partial(g + f)(x^{k+1}) + A^\top y^{k+1}. \qquad (3.4b) \end{aligned}$$

The sequence $(v^k)_{k\in\mathbb{N}} = (v_1^k, v_2^k)_{k\in\mathbb{N}}$ is uniquely determined by the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, y^k)_{k\in\mathbb{N}}$ and can be computed efficiently without the need for additional gradient evaluations or matrix–vector products. Note that $v^k \in Tz^k$, where $T(x, y) := \left(\partial h^*(y) - Ax, \partial g(x) + \nabla f(x) + A^\top y\right)$ denotes the operator associated with the primal-dual optimality conditions (3.1). Therefore, the quantity

$$\|v^k\| \geq \operatorname{dist}(0, Tz^k) \qquad (3.5)$$

serves as a measure of optimality and can be used as a termination criterion.

3.1.3 Comparison to the proximal gradient method

When $h \equiv 0$ (and $A = 0$), problem (1.5) reduces to problem (2.1) and adaPDM with $\epsilon = 0$ reduces to adaPGM. In fact, in this case apparently $y^k \equiv 0$ and $x^{k+1} = \operatorname{prox}_{\gamma_{k+1} g}(x^k - \gamma_{k+1} \nabla f(x^k))$, and since $\xi_k \equiv 0$ the stepsize update on γ_k reduces to

$$\begin{aligned} \gamma_{k+1} &= \gamma_k \min\left\{\sqrt{1 + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \tfrac{1}{\sqrt{2\left(|\delta_k| + \delta_k\right)}}\right\} \\ &= \gamma_k \min\left\{\sqrt{1 + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \tfrac{1}{2\sqrt{[\delta_k]_+}}\right\}, \end{aligned}$$

which is precisely the stepsize update rule in adaPGM.

Remark 3.1 Although there is an apparent relation between P_k as in (3.3a) and the one in Lemma 2.2, while for adaPGM it was sufficient to show that $\liminf_{k\to\infty} P_k = 0$ to infer existence of optimal limit points, in the primal-dual setting an argument through the cost function cannot be used. In fact, although $\liminf_{k\to\infty} P_k = 0$ and $\lim_{k\to\infty} Q_k = 0$ do hold even when $\epsilon = 0$ (cf. Theorem 3.3*(iii)*), and thus limit points $(\hat{x}, \hat{y})$ of $(x^k, y^k)_{k\in\mathbb{N}}$ generated by either one of Algorithms 3.1 and 3.2 exist that satisfy

$$\mathcal{L}(\hat{x}, y^\star) = \mathcal{L}(x^\star, \hat{y}) = \mathcal{L}(x^\star, y^\star) \quad \forall (x^\star, y^\star) \in \mathcal{S}_\star, \tag{3.6}$$

we can only guarantee convergence to solutions by establishing sufficient descent in terms of the residual $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|$ (enforced in the algorithms through the introduction of the parameter $\epsilon > 0$).

The fact that primal optimality of $\hat{x}$ and/or dual optimality of $\hat{y}$ cannot be inferred from (3.6) can be demonstrated with a simple counterexample.[3]

Example 3.2 Consider the Lagrangian $\mathcal{L}(x, y) = x(1 + y)$ of the problem

$$\operatorname*{minimize}_{x\in\mathbb{R}} \ x \quad \text{subject to } x = 0.$$

The unique saddle point of $\mathcal{L}$ is $(x^\star, y^\star) = (0, -1)$, this being the only primal-dual solution of the problem. Nevertheless, any $x \in \mathbb{R}$ minimizes $\mathcal{L}(x, y^\star) \equiv 0$ and similarly $\mathcal{L}(x^\star, y) \equiv 0$ is maximized by any $y \in \mathbb{R}$; equivalently, P_k and Q_k as in (3.3) are identically zero independently of x^k and y^k.

Example 3.2 demonstrates that P_k and Q_k cannot, in general, be employed as optimality measures. Sufficient conditions involve the minimizer of $\mathcal{L}(\,\cdot\,, y^\star)$ and/or the maximizer of $\mathcal{L}(x^\star, \,\cdot\,)$ being unique, which can be guaranteed under strict convexity assumptions on the primal and/or dual formulations. Whether adaPDM can accommodate $\epsilon = 0$ in the generality of Assumption II remains an open question and is left for future work.

[3] Example taken from https://math.stackexchange.com/a/3039783/53739.

3.2 Linesearch variant without linear operator norm

The employment of the quantities ℓ_k and c_k in the adaptive stepsize strategies of adaPGM and adaPDM enables much tighter estimates of the local geometry of the problem, as opposed to adopting preset constants such as global Lipschitz moduli (when available). By the same principle, the norm of the linear operator A only offers a worst-case bound of the kind $\|A^\top \cdot\| \leq \|A\|\|\cdot\|$, which can potentially be very loose on specific instances. To completely remove any dependency on global quantities, in this subsection we introduce adaPDM$^+$ (Algorithm 3.2), a *fully* adaptive primal-dual algorithm that replaces also the norm of the linear operator A with local estimates. As was also the case for the other two algorithms, adaPDM$^+$ follows the convention of indexing with k all the variables that depend on quantities defined up to iteration k. This choice highlights the nested dependency of γ_{k+1} and η_{k+1} at step 3.2.4, which appears to be solvable only by means of a linesearch procedure (see also the discussion before Theorem 3.3). In account of this, we dub the method *essentially* adaptive. The employment of a linesearch within an adaptive scheme is also pursued in [44], where an extension of adaPGM is developed in the context of simple bilevel optimization, a setting that encompasses the primal-dual problem setting (1.5) as a special case. Backtracking linesearch has also been employed in combination with PDHG in various forms. In [29, 30] it is used to potentially increase the speed of convergence by balancing the primal and dual residuals. In [49] an adaptive linesearch algorithm is presented for PDHG and the idea is extended to the composite form (1.5), see [49, Alg. 4]. In addition to a different stepsize update rule, a major difference here is that the backtracks involved do not require evaluations of the gradient ∇f. We also remark that adaPDM$^+$ provides a practical way of initializing the stepsize, and that it reduces to adaPDM if η_k is taken as $\|A\|$ for all k.

3.2.1 Practical considerations and termination criteria

The same initialization and parameter values described for adaPDM can be used in adaPDM$^+$ with the difference that the initial estimate η_0 replaces $\|A\|$. The estimate η_0 can be computed, for instance, by evaluating the Frobenius norm of the linear operator as an underestimation for it. Subsequent values of η_{k+1} can be initialized based on the previously accepted value η_k; consistently with what suggested in step 3.2.2, a multiple in the range $[0.9, 1)$ is recommended. Regarding the termination criterion, the same strategy discussed in Sect. 3.1 for adaPDM can be employed.

3.3 Convergence results

The convergence analysis will once again revolve around showing descent on a suitable merit function, this time defined as $\mathcal{U}_k = \mathcal{U}_k(x^\star, y^\star)$ given by

$$\mathcal{U}_k := \tfrac{1}{2}\|x^k - x^\star\|^2 + \tfrac{1-4\xi_k(1+\epsilon)}{4}\|x^k - x^{k-1}\|^2 + \tfrac{1}{2t^2}\|y^k - y^\star\|^2 + \gamma_k(1+\rho_k)P_{k-1}, \quad (3.7)$$

Algorithm 3.2 Adaptive primal-dual method with linesearch (adaPDM$^+$)

REQUIRE primal/dual (square inverse) stepsize ratio $t > 0$
stepsize parameters $\epsilon > 0$ and $\nu > 1 + \epsilon$ (e.g., $\epsilon = 10^{-6}$, $\nu = 1.2$)
initial primal-dual pair $(x^{-1}, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$, estimate $\eta_0 > 0$ of $\|A\|$,
stepsizes $0 < \gamma_{-1} \leq \gamma_0 \leq 1/2t\nu\eta_0$, backtracking parameter $r > 1$ (e.g., $r = 2$)

INITIALIZE $x^0 = \operatorname{prox}_{\gamma_0 g}\left(x^{-1} - \gamma_0 \nabla f(x^{-1}) - \gamma_0 A^\top y^0\right)$

REPEAT FOR $k = 0, 1, \ldots$ until convergence

3.2.1: Set $\delta_k = \gamma_k \ell_k (\gamma_k c_k - 1)$ and $\bar{\xi}_k = t^2 \gamma_k^2 \eta_k^2 (1+\epsilon)^2$, with ℓ_k and c_k as in (1.3)

3.2.2: Choose an estimate $0 < \eta_{k+1} \leq \eta_k$ of $\|A\|$ (e.g., $\eta_{k+1} = 0.95\eta_k$)

3.2.3: WHILE TRUE DO // (linesearch loop)

3.2.4: Define the stepsizes as

$$\begin{cases} \gamma_{k+1} = \min\left\{\gamma_k\sqrt{1+\frac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{2\nu t\eta_{k+1}},\ \gamma_k\sqrt{\frac{1-4\bar{\xi}_k}{2(1+\epsilon)\left(\sqrt{\delta_k^2+(t\eta_{k+1}\gamma_k)^2(1-4\bar{\xi}_k)}+\delta_k\right)}}\right\} \\ \sigma_{k+1} = t^2\gamma_{k+1} \end{cases}$$

3.2.5: $y^{k+1} = \operatorname{prox}_{\sigma_{k+1}h^*}\left(y^k + \sigma_{k+1}\left(\left(1+\frac{\gamma_{k+1}}{\gamma_k}\right)Ax^k - \frac{\gamma_{k+1}}{\gamma_k}Ax^{k-1}\right)\right)$

3.2.6: IF $\eta_{k+1} \geq \frac{\|A^\top(y^{k+1}-y^k)\|}{\|y^{k+1}-y^k\|}$ THEN BREAK, ELSE $\eta_{k+1} \leftarrow r\eta_{k+1}$

3.2.7: $x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}\left(x^k - \gamma_{k+1}\nabla f(x^k) - \gamma_{k+1}A^\top y^{k+1}\right)$

where $(x^\star, y^\star) \in \mathcal{S}_\star$ is any primal-dual optimal pair. The next theorem allows us to study the convergence of Algorithms 3.1 and 3.2 under a unified analysis. It should be noted that, unless $\|A\|$ is known and η_{k+1} is chosen greater to or equal than that quantity (as it happens in adaPDM), the following theorem does not furnish an implementable stepsize update rule, owing to the implicit dependency between the stepsize γ_{k+1} and the norm estimate η_{k+1}. The linesearch prescribed by adaPDM$^+$ circumvents this issue.

Theorem 3.3 *Suppose that Assumption II holds, and let $t > 0$, $\epsilon \geq 0$, and $\nu > 1 + \epsilon$ be fixed. Consider a sequence $(x^k, y^k)_{k\geq 1}$ generated by*

$$\begin{cases} y^{k+1} = \operatorname{prox}_{\sigma_{k+1}h^*}\left(y^k + \sigma_{k+1}\left(\left(1+\frac{\gamma_{k+1}}{\gamma_k}\right)Ax^k - \frac{\gamma_{k+1}}{\gamma_k}Ax^{k-1}\right)\right) \\ x^{k+1} = \operatorname{prox}_{\gamma_{k+1}g}\left(x^k - \gamma_{k+1}\nabla f(x^k) - \gamma_{k+1}A^\top y^{k+1}\right), \end{cases} \tag{3.8}$$

starting from a triplet $(x^{-1}, x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ and with initial primal stepsizes $\gamma_0 \geq \gamma_{-1} > 0$. Denote $\xi_k := t^2\eta_k^2\gamma_k^2$ and $\delta_k := \gamma_k\ell_k(\gamma_k c_k - 1)$ with ℓ_k and c_k as in (1.3), $\eta_0 \leq \frac{1}{2\nu t\gamma_0}$ and η_{k+1} any such that $\frac{\|A^\top(y^k-y^{k+1})\|}{\|y^k-y^{k+1}\|} \leq \eta_{k+1} \leq \eta_{\max}$ for some $\eta_{\max} < \infty$, $k \in \mathbb{N}$. Suppose that the sequences of stepsizes comply with the rules

$$\gamma_{k+1} = \min\left\{\gamma_k\sqrt{1+\frac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{2\nu t\eta_{k+1}},\ \gamma_k\sqrt{\frac{1-4\xi_k(1+\epsilon)^2}{2(1+\epsilon)\left(\sqrt{\delta_k^2+(t\eta_{k+1}\gamma_k)^2(1-4\xi_k(1+\epsilon)^2)}+\delta_k\right)}}\right\}$$

and $\sigma_k = t^2\gamma_k$, $k \in \mathbb{N}$. *Then, for any primal-dual solution* $(x^\star, y^\star)$ *of* (1.5) *and with* $\mathcal{U}_k$ *as in* (3.7), *the following hold:*

(i) $\mathcal{U}_{k+1} \leq \mathcal{U}_k - \underbrace{\frac{\epsilon}{2t^2(1+\epsilon)}}_{\geq 0}\|y^k - y^{k-1}\|^2 - \underbrace{\frac{1-4\xi_k-4\rho_{k+1}^2(\delta_k+\xi_{k+1})}{4}}_{\geq \epsilon/4(1+\epsilon)}\|x^k - x^{k+1}\|^2 - \gamma_k(1+\rho_k-\rho_{k+1}^2)P_{k-1} - \gamma_{k+1}Q_{k+1}.$

(ii) The sequence $(x^k, y^k)_{k\in\mathbb{N}}$ *is bounded, and* $\gamma_k \geq \hat{\gamma} > 0$ *for all* $k \geq 1$ *(see* (A.7) *for the value of* $\hat{\gamma}$*).*

(iii) $\liminf_{k\to\infty} P_k = \lim_{k\to\infty} Q_k = 0$.

(iv) If $\epsilon > 0$, *the sequence* $(x^k, y^k)_{k\in\mathbb{N}}$ *converges to a primal-dual solution.*

We also remark that, thanks to the descent inequality in Theorem 3.3*(i)*, the same telescoping arguments as in [15] can be used to show an $\mathcal{O}(1/k)$ convergence rate in terms of the partial primal-dual gap function introduced in [15]. The sequential convergence results stated next for adaPDM follow from the more general Theorem 3.3*(iv)*, specialized to $\eta_k \equiv \eta_{\max} = \|A\|$. For adaPDM$^+$ the assertions follow from Theorem 3.3 this time with $\eta_{\max} = \max\{\eta_0, r\|A\|\}$, as this furnishes an upper bound on $(\eta_k)_{k\in\mathbb{N}}$ (this fact simply follows by observing that whenever $\eta_{k+1} \geq \|A\|$ one has that the backtracking procedure terminates at step 3.2.6, and that η_{k+1} is otherwise increased by a factor r at any failed attempt).

Theorem 3.4 *(convergence of Algorithms 3.1 and 3.2) Under Assumption II, all the assertions in Theorems 3.3(i) to 3.3(iv) remain valid for both of the sequences generated by Algorithms 3.1 and 3.2.*

4 Numerical simulations

In this section the performance of the proposed algorithms is evaluated through a series of simulations on standard problems on both synthetic data as well as datasets from the LIBSVM library [16]. All the algorithms are implemented in the Julia programming language and are available online (Table 1).[4]

The backtracking procedure in PGM-ls is meant in the sense of [7, §10.4.2], (see [60, LS1] and [23, Alg. 3] for the locally Lipschitz smooth case), without enforcing monotonic decrease on the stepsize sequence. Specifically, the initial guess for γ_{k+1} is warm-started as $r\gamma_k$, where γ_k is the accepted value in the previous iteration and $r \geq 1$ is a scaling factor. In each simulation the best outcome for PGM-ls among the choices $r \in \{1, 1.5, 2\}$ is reported.

We also considered two linesearch variants of Nesterov's accelerated method. The first variant is described in [52, Eq. 4.9] and allows the stepsize to increase (by warm starting the backtracks with a multiple of the last accepted stepsize). The second variant is described in [7, §10.7] and initializes the backtracking procedure with the last accepted stepsize. We tested the first algorithm with scaling factors of 1.5 and 2. Both variants performed worse than the constant stepsize regime, and we have

[4] https://github.com/pylat/adaptive-proximal-algorithms.

Table 1 Algorithms compared against in the numerical simulations (when applicable)

PGM	Proximal gradient with constant stepsize $1/L_f$
PGM-ls	Proximal gradient with backtracking
Nesterov	Nesterov's acceleration with constant stepsize $1/L_f$ [7, §10.7]
aGRAAL	The golden ratio algorithm [46]
PDHG	The algorithm of [15]
CV	The algorithm of Condat and Vũ, proposed in [20, 66]
MP-ls	The linesearch method of Malitsky and Pock [49, Alg. 4]
adaPGM	Algorithm 2.1
adaPGM-MM	Proximal extension of Malitsky and Mishchenko [47, Alg. 1][a]
adaPDM	Algorithm 3.1
adaPDM$^+$	Algorithm 3.2

[a] See Corollary 2.5

therefore omitted them from the plots. The degradation in performance is likely due to the extra cost of evaluating the backtracking condition.

4.1 Adaptive proximal gradient

We compare the performance of adaPGM (Algorithm 2.1) on three practical problems that can be cast as (2.1). In the figures, the distance of the cost from the minimum is plotted against the number of calls to linear operations, which in all problems correspond to the most costly operation. This accounts for all gradient evaluations, and all additional cost evaluations that linesearch methods incur.

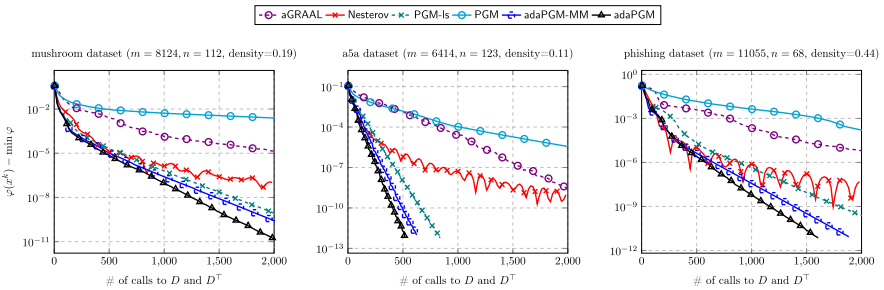
In order to have a fair comparison between different methods, for globally Lipschitz-smooth problems we initialized the stepsize for all the methods with $1/L_f$ where L_f denotes the Lipschitz constant of ∇f. For cubic regularization we perturbed the initial point with a random vector and used the inverse of the estimate as in (1.2) using these two points. We note that $r = 1.5$ usually worked best in backtracking linesearch variants; for other parameters we used standard choices. For aGRAAL, we set the algorithm parameters as suggested in [46, §5]. The inital point x^0 was set to zero for all algorithms. The results of all the simulations are reported in Fig. 1.

4.1.1 Logistic regression

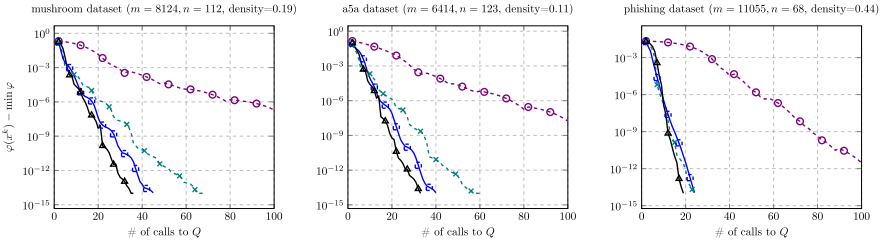
We consider an ℓ_1-regularized logistic regression problem

$$\underset{x\in\mathbb{R}^{n+1}}{\text{minimize}} \; -\tfrac{1}{m}\sum_{i=1}^{m}\Big[y_i \log\big(s_i(x)\big) + (1-y_i)\log\big(1-s_i(x)\big)\Big] + \lambda\|x\|_1,$$

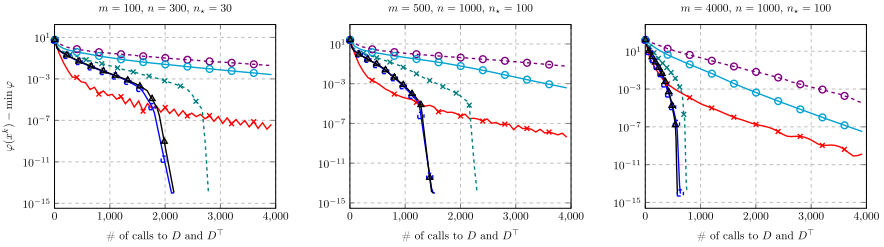
where $\lambda > 0$ is the regularization parameter, m, n are the number of samples and features, $s_i(x) = (1+\exp(-d_i^\top x))^{-1}$ are logistic sigmoid functions with $d_i \in \mathbb{R}^{n+1}$ denoting the i-th data sample (up to absorbing the bias terms), and $y_i \in \{0, 1\}$ is the associated label. In Fig. 1a the algorithms are compared in terms of the total number



(a) Logistic regression problem of Section 4.1.1



(b) Cubic regularization problem of Section 4.1.2



(c) Lasso problem of Section 4.1.3

Fig. 1 Simulations for problem (2.1)

of calls to D and its adjoint, where D denotes the data matrix constructed by stacking the vectors d_i. In all the simulations $\lambda = 0.01$ was used.

4.1.2 Cubic regularization

The subproblem in the cubic Newton method [53] involves solving

$$\underset{x \in \mathbb{R}^{n+1}}{\text{minimize}} \ \tfrac{1}{2}\langle x, Qx\rangle + \langle x, q\rangle + \tfrac{M}{6}\|x\|^3,$$

where $Q \in \mathbb{R}^{(n+1)\times(n+1)}$, $q \in \mathbb{R}^{n+1}$, and $M > 0$ is some regularization parameter. In the simulations, the Hessian Q and the gradient q are generated for the logistic loss problem evaluated at zero on the mushroom, a5a, and phishing datasets. Moreover, $M = 1$ is used in the plots, though we remark that the behavior of the algorithms for different values is similar.

4.1.3 Regularized least squares

Consider the lasso problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}\|Dx-b\|^2+\|x\|_1,$$

where matrix $D\in\mathbb{R}^{m\times n}$ and vector $b\in\mathbb{R}^n$ are generated based on the procedure described in [52, §6], and $n_\star$ denotes the number of nonzero elements of the solution. In all simulations, the parameter controlling the magnitude of the primal solution was set equal to one ($\rho=1$ in the notation of [52, §6]), but the general behavior of the algorithms is similar for alternative values.

4.2 Adaptive primal-dual algorithms

We now compare the performance of adaPDM (Algorithm 3.1) and its operator norm–free extension adaPDM⁺ (Algorithm 3.2) on problems fitting into formulation (1.5).

Optimality criterion

The optimality criterion defined in Sect. 3.1.2 will be used in all the simulations for comparison between algorithms. It consists of computing the norm of the sequence v^k uniquely determined by the generated sequence for which $v^k\in Tz^k$ holds. This quantity is provided in (3.4) for adaPDM and adaPDM⁺, as well as the Condat-Vũ and PDHG algorithms (both corresponding to updates in steps 3.1.3 and 3.1.4 with constant stepsize). In the case of MP-ls [49, Alg. 4], it is similarly obtained according to the optimality conditions of the proximal updates at steps 1 and 2.a therein (the extra evaluations needed are not counted in the plots).

Stepsize selection

The stepsize parameters for adaPDM and adaPDM⁺ were configured as follows: $\epsilon=10^{-8}$, $\nu=1.2$, and $\gamma_0=\gamma_{-1}=1/(2\nu t\eta_0)$. In the presented simulations, for comparison purposes we used $\eta_0=\|A\|$, noting that adaPDM⁺ displays low sensitivity to the initial choice of η_0. Consenquently, there is little need for fine-tuning this parameter and any reasonable estimate results in a very similar trajectory.

We observed that the stepsize ratio t requires some tuning. In our preliminary simulations for Algorithms 3.1 and 3.2, on the problems presented in here, we ran a grid search for $t\in[0.01,100]$ and observed that often $t=1$ performed well enough on the presented examples. For the Condat-Vũ and PDHG algorithms, we used the heuristic stepsize selection rule suggested in [36, Eq. (3.53)]. In the case of MP-ls, in [49, §5] it was suggested to set t equal to the ratio obtained by tuning the constant stepsize regime of the Condat-Vũ algorithm. In our experiments, in addition to this heuristic the performance of MP-ls was fine-tuned through a grid search in the range $t\in[0.01,100]$.

4.2.1 Dual support vector machine problem

The support vector machine (SVM) problem consists of solving

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ C\sum_{n=1}^{N}\max\left\{0, 1-a_i(d_i^\top x+x_0)\right\}+\tfrac{1}{2}\|x\|^2,$$

where $(d_i, a_i) \in \mathbb{R}^n \times \mathbb{R}$ represents the i-th data pair, and $C > 0$ is some positive constant. Computationally, a popular approach is to instead consider the dual SVM problem [32, §12.2.1]

$$\begin{aligned}&\underset{\alpha_1,\ldots,\alpha_N}{\text{minimize}}\ \tfrac{1}{2}\|\sum_{i=1}^{N}\alpha_i a_i d_i\|^2-\sum_{i=1}^{N}\alpha_i\\ &\text{subject to } 0\le\alpha_i\le C,\quad i=1,\ldots,N\\ &\qquad\sum_{i=1}^{N}\alpha_i a_i=0.\end{aligned}\tag{4.1}$$

In the simulations we used $C \in \{1, 0.1\}$. The problem is cast in the standard form (1.5) by letting f represent the quadratic cost, g the indicator of the box constraints, $h = \iota_{\{0\}}$, and $A = [a_1, \ldots, a_N] \in \mathbb{R}^{1\times N}$. Notice that the dual vectors y^k are scalars in this case, and in particular one has $\|A^\top(y^{k+1} - y^k)\| = \|A\|\|y^{k+1} - y^k\|$ for any k, where $\|A\|$ is a vector norm which is negligible to compute. Apparently, in this case (up to dicarding inital iterates) the linesearch variant adaPDM$^+$ produces the same iterates of adaPDM with no computational advantage, and is thus omitted from the plots. The results of this simulation are reported in Fig. 2.

4.2.2 Least absolute deviation regression and square-root lasso

As a final application we consider regularized regression, consisting of

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \|Dx-b\|_p+\lambda\|x\|_1.\tag{4.2}$$

Whenever $p = 2$, this problem is referred to as square-root lasso, and for $p = 1$ it is known as the least absolute deviation (LAD) regression. Both problems are cast as (1.5) by letting $f \equiv 0$, $g = \|\cdot\|_1$, $h = \|\cdot - b\|_p$, and $A = D$. For the regression data $D \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, three different datasets from the LIBSVM library were used. The results are reported in Fig. 3. It is apparent that the linesearch variant adaPDM$^+$ (as well as MP-ls, for similar reasons) excels by adaptively employing tighter bounds of the linear operator norm along one-dimensional subspaces. We ran the simulations for $\lambda \in \{0.1, 1, 10\}$, but only report them for $\lambda = 10$ remarking that the behavior is similar for the other values.

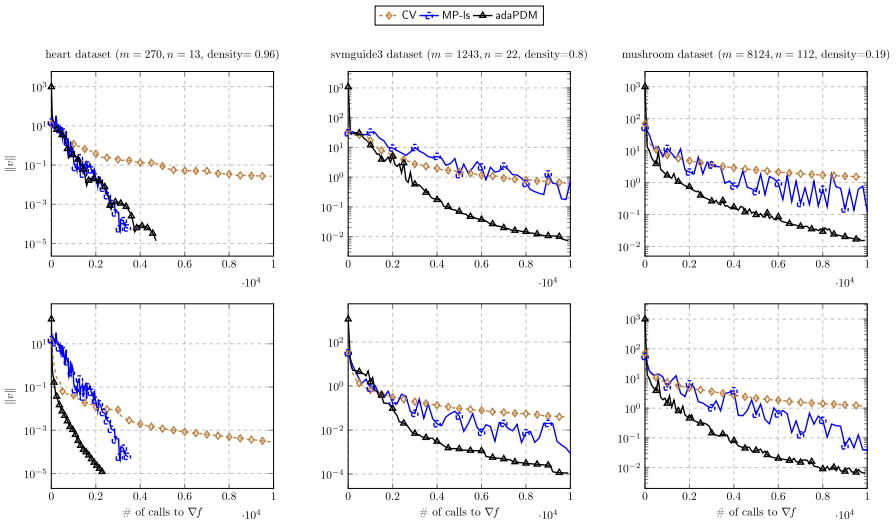


Fig. 2 Simulations for the dual SVM problem (4.1). First row: $C = 1$, second row: $C = 0.1$. The x-axis reports gradient evaluations, which is the most expensive operation (since $A \in \mathbb{R}^{1\times N}$, calls to A and $A^\top$ are negligible). As explained, in this case adaPDM$^+$ is indistinguishable from adaPDM and thus omitted. AdaPDM and MP-ls are tuned for best performance by a grid search for $t \in [0.01, 100]$

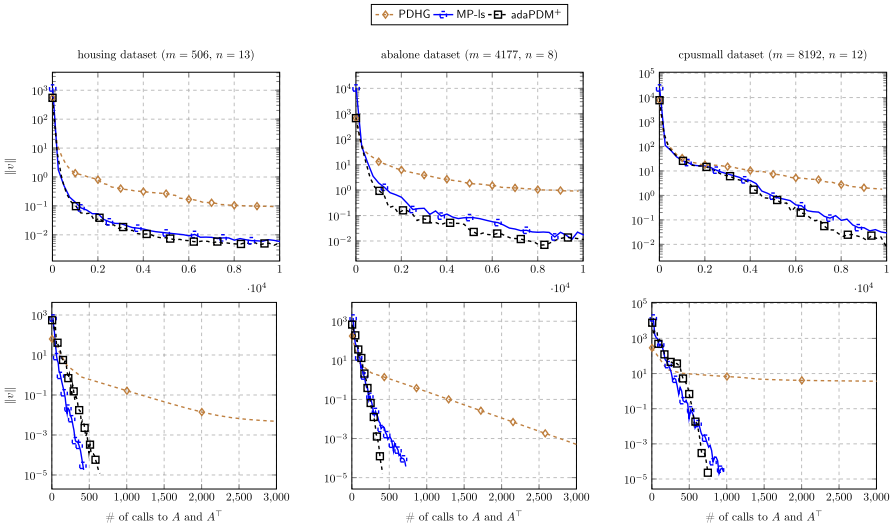


Fig. 3 Simulations for problem (4.2) of Sect. 4.2.2 with $\lambda = 10$. First row: regularized least-absolute deviation ($p = 1$); second row: square-root lasso ($p = 2$). AdaPDM$^+$ and MP-ls are tuned for best performance by a grid search for $t \in [0.01, 100]$. Having $f = 0$, adaPDM reduces to PDHG with worse (constant) stepsizes and is thus omitted from the comparisons

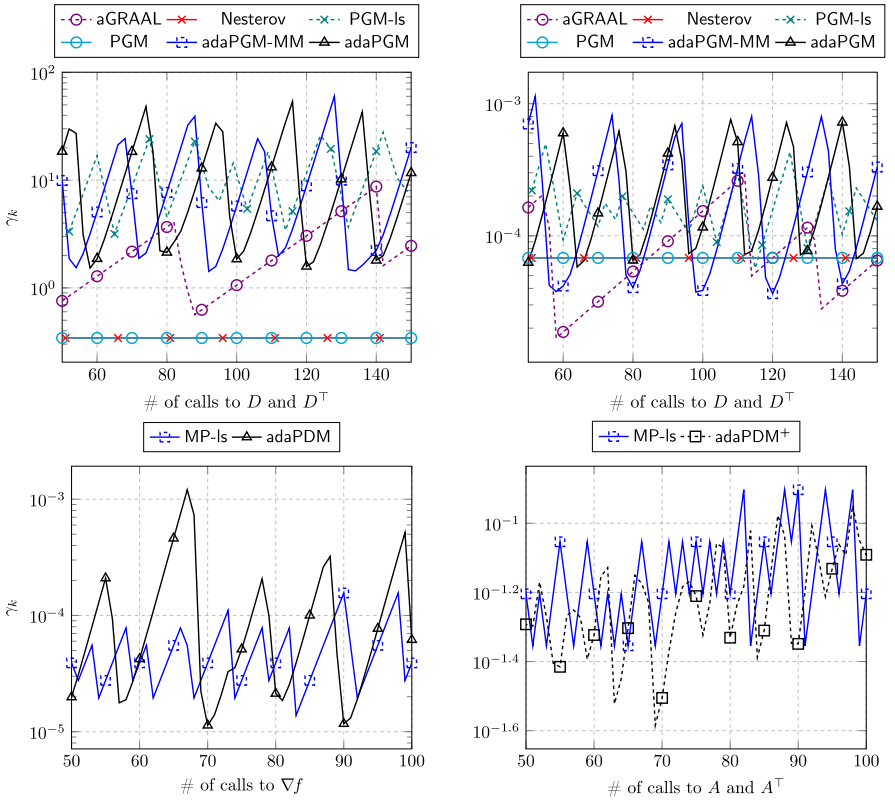


Fig. 4 Demonstrative plot of stepsize magnitudes in windows of 50 and 100 iterations extracted from the simulations in this section. First row, left: logistic regression, mushroom dataset; right: lasso problem with $m = 500$, $n = 1000$, $n_\star = 100$. As commented after Remark 2.4, despite the fact that the stepsize update rule of adaPGM is less conservative than that of adaPGM-MM, the comparison does not carry over iterationwise. Second row, left: dual SVM, mushroom dataset, $C = 0.1$; right: square-root lasso (housing dataset, $\lambda = 10$). Primal-dual algorithms are compared on problems where the ratio $t^2 = \sigma_k/\gamma_k$ coincides, so that the plots are representative also for dual stepsizes σ_k

4.3 Observations about the stepsize sequence

Theorem 2.3*(ii)* certifies that the stepsizes γ_k produced by adaPGM do not drop below $1/2L_{f,\mathcal{V}}$, where $L_{f,\mathcal{V}}$ is a Lipschitz modulus for ∇f on a convex and bounded set $\mathcal{V}$ that contains all the iterates. A refined analysis can nevertheless reveal some interesting additional details that, informally, confirm the pattern exhibited by the sequence $(\gamma_k)_{k\in\mathbb{N}}$ in all our experiments.

Whenever a stepsize is below a certain threshold, a first phase is triggered where the stepsize strictly increases until the threshold is surmounted. As ensured by Lemma A.2, c_k is upper bounded by a local Lipschitz modulus $L_{f,\overline{\mathcal{V}}}$ for ∇f on some enlargement $\overline{\mathcal{V}}$ of the set $\mathcal{V}$ containing the iterates. Because of the update rule at step 2.1.1, if $\gamma_k \leq 1/L_{f,\overline{\mathcal{V}}}$ (hence $\gamma_k c_k \leq 1$) then $\gamma_{k+1} > \gamma_k$ (in fact it increases at least linearly, cf. (2.11)). This is followed by a second phase of unspecified length, during which

the stepsizes remain strictly greater than $1/L_{f,\overline{\nu}}$. Should a stepsize again drop below it (yet surely not below $1/2L_{f,\mathcal{V}}$), the cycle restarts from phase one. Such an oscillatory behavior is confirmed in all our simulations (see Fig. 4 for particular instances) and could also explain the adoption of much larger stepsizes compared to the standard constant stepsize regime. Whether more rigorous arguments could be made in support of this empirical evidence is an interesting open question left for future work.

5 Conclusions

In this paper we studied convex composite problems involving the sum of a locally Lipschitz differentiable term and two nonsmooth terms, one of which composed with a linear operator. The primal-dual algorithm adaPDM was proposed that updates the stepsizes adaptively using already available information without any backtracking procedures. The stepsize update is based on a novel rule that combines estimates of local cocoercivity and Lipschitz continuity of the gradient of the differentiable function. When the linear operator is zero we obtain adaPGM that not only extends the adaptive gradient descent algorithm of [47], but also involves a less conservative stepsize update rule. More generally, we further waive the computation of the norm of the linear operator (required by the stepsize update rule) through a linesearch procedure. The resulting algorithm, adaPDM$^+$, can be implemented efficiently given that the backtracks do not require additional gradient evaluations.

Future research directions involve extending the presented line of proof to the setting of variational inequalities in the framework of algorithms such as the extragradient method and other projection-based splittings such as those in [28, 39]. Other potential directions include nonmonotone extensions for problems satisfying the so-called weak Minty assumptions [25, 55], as well as block-coordinate and stochastic variants. It would also be interesting to explore extensions of the ideas presented in this work to other types of adaptive settings such as those in [24, 45, 67, 71].

A Convergence analysis for Sect. 3

We begin by establishing an inequality for the iterates in (3.8) that would eventually under proper stepsize choices lead to a descent inequality.

Lemma A.1 *Let $(\gamma_k)_{k\in\mathbb{N}}$, $(\sigma_k)_{k\in\mathbb{N}}$ be sequences of strictly positive scalars, and starting from a triplet $(x^{-1}, x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ let $(x^k, y^k)_{k\in\mathbb{N}}$ be recursively defined by (3.8). Then, for any $\varepsilon_k, \tau_k, \mu_k > 0$ and $\eta_k \geq \frac{\|A^\top(y^k - y^{k-1})\|}{\|y^k - y^{k-1}\|}$ (e.g., $\eta_k = \|A\|$) it holds that*

$$
\begin{aligned}
0 \le\ & \tfrac{1}{2}\|x^k - x^\star\|^2 - \tfrac{1}{2}\|x^{k+1} - x^\star\|^2 + \tfrac{\varepsilon_{k+1}\rho_{k+1}+\mu_{k+1}\eta_{k+1}\gamma_{k+1}-1}{2}\|x^k - x^{k+1}\|^2 \\
& + \rho_{k+1}\Big(\tfrac{1-\alpha_k(2-\beta_k)}{2\varepsilon_{k+1}} + \tfrac{\tau_{k+1}\eta_{k+1}\gamma_{k+1}}{2} - \rho_{k+1}(1-\alpha_k)\Big)\|x^{k-1} - x^k\|^2 \\
& + \tfrac{\gamma_{k+1}}{2\sigma_{k+1}}\|y^k - y^\star\|^2 - \Big(\tfrac{\gamma_{k+1}}{2\sigma_{k+1}} - \tfrac{\eta_{k+1}\gamma_{k+1}}{2}\Big(\tfrac{1}{\mu_{k+1}} + \tfrac{\rho_{k+1}}{\tau_{k+1}}\Big)\Big)\|y^{k+1} - y^k\|^2 \\
& - \tfrac{\gamma_{k+1}}{2\sigma_{k+1}}\|y^{k+1} - y^\star\|^2 \\
& + \gamma_{k+1}\rho_{k+1}P_{k-1} - \gamma_{k+1}(1+\rho_{k+1})P_k - \gamma_{k+1}Q_{k+1},
\end{aligned}
$$

where $\rho_{k+1} := \frac{\gamma_{k+1}}{\gamma_k}$, $\alpha_k := \gamma_k \ell_k$, $\beta_k := \gamma_k c_k$, *and* P_k, Q_k *are as in* (3.3).

Proof Let $\bar{y}^{k+1} := y^k + \sigma_{k+1}\frac{\gamma_{k+1}}{\gamma_k}A(x^k - x^{k-1})$, so that $y^{k+1} = \operatorname{prox}_{\sigma_{k+1}h^*}(\bar{y}^{k+1} + \sigma_{k+1}Ax^k)$. The characterization of y^k and x^{k+1} as in the respective updates then reads

$$
\tfrac{\bar{y}^{k+1}-y^{k+1}}{\sigma_{k+1}} + Ax^k \in \partial h^*(y^{k+1}) \tag{A.1a}
$$

and

$$
\tfrac{H_{k+1}(x^k)-x^{k+1}}{\gamma_{k+1}} - A^\top y^{k+1} = \tfrac{x^k-x^{k+1}}{\gamma_{k+1}} - \nabla f(x^k) - A^\top y^{k+1} \in \partial g(x^{k+1}). \tag{A.1b}
$$

In particular, for any solution pair $(x^\star, y^\star) \in \mathcal{S}_\star$

$$
\begin{aligned}
0 \overset{(A.1b)}{\le}\ & g(x^\star) - g(x^{k+1}) + \langle A^\top y^{k+1}, x^\star - x^{k+1}\rangle + \overbrace{\langle \nabla f(x^k), x^\star - x^{k+1}\rangle}^{(A')} \\
& + \tfrac{1}{2\gamma_{k+1}}\|x^k - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x^{k+1} - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x^k - x^{k+1}\|^2
\end{aligned}
$$

and

$$
\begin{aligned}
0 \overset{(A.1a)}{\le}\ & h^*(y^\star) - h^*(y^{k+1}) - \langle Ax^k, y^\star - y^{k+1}\rangle - \tfrac{1}{\sigma_{k+1}}\langle \bar{y}^{k+1} - y^{k+1}, y^\star - y^{k+1}\rangle \\
=\ & h^*(y^\star) - h^*(y^{k+1}) - \langle Au^{k+1}, y^\star - y^{k+1}\rangle \\
& + \tfrac{1}{2\sigma_{k+1}}\|y^k - y^\star\|^2 - \tfrac{1}{2\sigma_{k+1}}\|y^{k+1} - y^k\|^2 - \tfrac{1}{2\sigma_{k+1}}\|y^{k+1} - y^\star\|^2,
\end{aligned}
$$

where $u^{k+1} := (1+\rho_{k+1})x^k - \rho_{k+1}x^{k-1}$. We next proceed to upper bound the term (A′) as

$$
\begin{aligned}
(A') &= \langle \nabla f(x^k), x^\star - x^k\rangle + \langle \nabla f(x^k), x^k - x^{k+1}\rangle \\
&= \langle \nabla f(x^k), x^\star - x^k\rangle + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - x^k, x^{k+1} - x^k\rangle \\
&\quad + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle \\
&\overset{(A.1b)}{\le} f(x^\star) - f(x^k) + g(x^{k+1}) - g(x^k) + \langle A^\top y^k, x^{k+1} - x^k\rangle \\
&\quad + \underbrace{\tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle}_{(B)}.
\end{aligned}
$$

Let $\varphi = f + g$. We bound the term (B) as done in (2.5) and combine the three inequalities to obtain

$$\begin{aligned}0 \leq\ & (f+g)(x^\star) - (f+g)(x^k) + \tfrac{1}{2\gamma_{k+1}}\|x^k - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x^{k+1} - x^\star\|^2 \\ & + \left(\tfrac{\varepsilon_{k+1}}{2\gamma_k} - \tfrac{1}{2\gamma_{k+1}}\right)\|x^k - x^{k+1}\|^2 + \tfrac{1-\gamma_k\ell_k(2-\gamma_k c_k)}{2\varepsilon_{k+1}\gamma_k}\|x^{k-1} - x^k\|^2 \\ & + h^*(y^\star) - h^*(y^{k+1}) + \langle A^\top y^k, x^{k+1} - x^k\rangle + \langle A^\top y^{k+1}, x^\star - x^{k+1}\rangle \\ & - \langle Au^{k+1}, y^\star - y^{k+1}\rangle \\ & + \tfrac{1}{2\sigma_{k+1}}\|y^k - y^\star\|^2 - \tfrac{1}{2\sigma_{k+1}}\|y^{k+1} - y^k\|^2 - \tfrac{1}{2\sigma_{k+1}}\|y^{k+1} - y^\star\|^2.\end{aligned}$$

Using again the subgradient (A.1), one has

$$v^k := \tfrac{x^{k-1}-x^k}{\gamma_k} - (\nabla f(x^{k-1}) - \nabla f(x^k)) - A^\top y^k \in \partial(f+g)(x^k), \tag{A.2}$$

hence

$$\begin{aligned}0 &\leq \tfrac{\gamma_{k+1}}{\gamma_k}\Big((f+g)(x^{k-1}) - (f+g)(x^k) - \langle v^k, x^{k-1} - x^k\rangle\Big) \\ &= \tfrac{\gamma_{k+1}}{\gamma_k}\Big((f+g)(x^{k-1}) - (f+g)(x^k)\Big) - \tfrac{\gamma_{k+1}}{\gamma_k}\tfrac{1-\gamma_k\ell_k}{\gamma_k}\|x^k - x^{k-1}\|^2 \\ &\quad + \tfrac{\gamma_{k+1}}{\gamma_k}\langle A^\top y^k, x^{k-1} - x^k\rangle.\end{aligned} \tag{A.3}$$

Sum the last two inequalities and use the identity $\rho_{k+1}(x^k - x^{k-1}) = u^{k+1} - x^k$ to obtain

$$\begin{aligned}0 \leq\ & \tfrac{1}{2\gamma_{k+1}}\|x^k - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x^{k+1} - x^\star\|^2 + \langle Ax^\star, y^{k+1}\rangle - \langle Au^{k+1}, y^\star\rangle \\ & + \left(\tfrac{\varepsilon_{k+1}}{2\gamma_k} - \tfrac{1}{2\gamma_{k+1}}\right)\|x^k - x^{k+1}\|^2 + \left(\tfrac{1-\gamma_k\ell_k(2-\gamma_k c_k)}{2\varepsilon_{k+1}\gamma_k} - \tfrac{\gamma_{k+1}}{\gamma_k}\tfrac{1-\gamma_k\ell_k}{\gamma_k}\right)\|x^{k-1} - x^k\|^2 \\ & + \tfrac{1}{2\sigma_{k+1}}\|y^k - y^\star\|^2 - \tfrac{1}{2\sigma_{k+1}}\|y^{k+1} - y^k\|^2 - \tfrac{1}{2\sigma_{k+1}}\|y^{k+1} - y^\star\|^2 \\ & + \Big(h^*(y^\star) - h^*(y^{k+1})\Big) \\ & + \underbrace{\tfrac{\gamma_{k+1}}{\gamma_k}\Big((f+g)(x^{k-1}) - (f+g)(x^k)\Big) - \Big((f+g)(x^k) - (f+g)(x^\star)\Big)}_{\text{(C)}} \\ & + \underbrace{\langle A^\top(y^{k+1} - y^k), u^{k+1} - x^{k+1}\rangle}_{\text{(D)}}.\end{aligned}$$

To conclude, observe that

$$\text{(C)} = \rho_{k+1}P_{k-1} - (1+\rho_{k+1})P_k + \langle u^{k+1} - x^\star, A^\top y^\star\rangle$$

and

$$\begin{aligned}(\mathrm{D}) &= \tfrac{\gamma_{k+1}}{\gamma_k}\langle x^k - x^{k-1}, A^\top(y^{k+1}-y^k)\rangle + \langle x^k - x^{k+1}, A^\top(y^{k+1}-y^k)\rangle \\ &\le \tfrac{\gamma_{k+1}}{\gamma_k}\tfrac{\tau_{k+1}\eta_{k+1}}{2}\|x^{k-1}-x^k\|^2 + \tfrac{\mu_{k+1}\eta_{k+1}}{2}\|x^{k+1}-x^k\|^2 \\ &\quad + \tfrac{\eta_{k+1}}{2}\Big(\tfrac{1}{\mu_{k+1}} + \tfrac{\gamma_{k+1}}{\gamma_k\tau_{k+1}}\Big)\|y^{k+1}-y^k\|^2, \end{aligned} \tag{A.4}$$

where $\mu_k, \tau_k > 0$ are parameters related to Young's inequality, so that the proof follows from the identity $\langle Ax^\star, y^{k+1} - y^\star\rangle = \mathcal{L}(x^\star, y^{k+1}) - \mathcal{L}(x^\star, y^\star)$. □

In addition to the upper estimate $L_{f,\mathcal{V}}$ for L_k and ℓ_k in Lemma 2.1*(iii)*, the proof of Theorem 3.3 will exploit an upper bound for c_k which is obtained in the next lemma. The proof is a slight refinement of known cocoercivity results in the globally Lipschitz setting [7, Thm. 5.8], and is provided for completeness to highlight the need of the enlarged set $\overline{\mathcal{V}}$

Lemma A.2 *Suppose that Assumption* IA1 *holds. Then, for every $x, y \in \mathbb{R}^n$ it holds that*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2L_{f,\overline{\mathcal{V}}}}\|\nabla f(x) - \nabla f(y)\|^2,$$

where $L_{f,\overline{\mathcal{V}}}$ is a Lipschitz modulus for ∇f on $\overline{\mathcal{V}} := \mathcal{V} + \overline{\mathrm{B}}(0; \mathrm{diam}(\mathcal{V}))$, and $\mathcal{V} \subseteq \mathbb{R}^n$ is a bounded and convex set that contains x and y. In particular,

$$\frac{\|\nabla f(x) - \nabla f(y)\|^2}{\langle \nabla f(x) - \nabla f(y), x - y\rangle} \le L_{f,\overline{\mathcal{V}}}.$$

Proof We henceforth fix $x, y \in \mathcal{V} \subseteq \overline{\mathcal{V}}$. Since ∇f is Lipschitz continuous on $\overline{\mathcal{V}}$ with modulus $L_{f,\overline{\mathcal{V}}}$, it follows from the descent inequality [8, Prop. A.24] that

$$f(z) \le f(y) + \langle \nabla f(y), z - y\rangle + \tfrac{L_{f,\overline{\mathcal{V}}}}{2}\|z - y\|^2 \quad \forall z \in \overline{\mathcal{V}}. \tag{A.5}$$

Let $l_x(y) := f(y) - f(x) - \langle \nabla f(x), y - x\rangle$, and note that l_x is a convex function with $\nabla l_x(y) = \nabla f(y) - \nabla f(x)$. For any $z \in \overline{\mathcal{V}}$, we have

$$\begin{aligned} l_x(z) &= f(z) - f(x) - \langle \nabla f(x), z - x\rangle \\ &\overset{(A.5)}{\le} f(y) + \langle \nabla f(y), z - y\rangle + \tfrac{L_{f,\overline{\mathcal{V}}}}{2}\|z-y\|^2 - f(x) - \langle \nabla f(x), z - x\rangle \\ &= f(y) - f(x) - \langle \nabla f(x), y - x\rangle + \langle \nabla f(y) - \nabla f(x), z - y\rangle + \tfrac{L_{f,\overline{\mathcal{V}}}}{2}\|z-y\|^2 \\ &= l_x(y) + \langle \nabla l_x(y), z - y\rangle + \tfrac{L_{f,\overline{\mathcal{V}}}}{2}\|z-y\|^2. \end{aligned}$$

Noticing that $\nabla l_x(x) = 0$, it follows from convexity of l_x that x is its global minimizer. Hence that $\min l_x = l_x(x) = 0$. Let us denote $v := \frac{1}{\|\nabla l_x(y)\|}\nabla l_x(y)$ and set $z = y - \frac{\|\nabla l_x(y)\|}{L_{f,\overline{\mathcal{V}}}}v$. Note that

$$\|z - y\| = \tfrac{\|\nabla l_x(y)\|}{L_{f,\overline{\mathcal{V}}}} = \tfrac{\|\nabla f(y) - \nabla f(x)\|}{L_{f,\overline{\mathcal{V}}}} \le \|y - x\|,$$

and in particular $z \in \mathcal{V} + \overline{\mathrm{B}}(y; \|y - x\|) \subseteq \overline{\mathcal{V}}$. From the previous inequality we get

$$\begin{aligned} 0 = \min l_x = l_x(x) &\leq l_x\Big(y - \tfrac{\|\nabla l_x(y)\|}{L_{f,\overline{\mathcal{V}}}} v\Big) \\ &\overset{(A.5)}{\leq} l_x(y) - \tfrac{\|\nabla l_x(y)\|}{L_{f,\overline{\mathcal{V}}}} \langle \nabla l_x(y), v\rangle + \tfrac{L_{f,\overline{\mathcal{V}}}}{2} \tfrac{1}{L_{f,\overline{\mathcal{V}}}^2} \|\nabla l_x(y)\|^2 \\ &= f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \tfrac{1}{2L_{f,\overline{\mathcal{V}}}} \|\nabla f(x) - \nabla f(y)\|^2. \end{aligned}$$

□

Proof of Theorem 3.3

♠ *(i)* Denoting

$$\tilde{\mathcal{U}}_k := \tfrac{1}{2}\|x^k - x^\star\|^2 + \tfrac{1-\varepsilon_k\rho_k-\mu_k\eta_k\gamma_k}{2}\|x^k - x^{k-1}\|^2 + \tfrac{\gamma_k}{2\sigma_k}\|y^k - y^\star\|^2 + \gamma_k(1+\rho_k)P_{k-1},$$

the inequality in Lemma A.1 can be expressed as

$$\begin{aligned} \tilde{\mathcal{U}}_{k+1} \leq \tilde{\mathcal{U}}_k &- \Big(\tfrac{\gamma_k}{2\sigma_k} - \tfrac{\gamma_{k+1}}{2\sigma_{k+1}}\Big)\|y^k - y^\star\|^2 - \gamma_k(1 + \rho_k - \rho_{k+1}^2)P_{k-1} \\ &- \Big(\tfrac{1-\varepsilon_k\rho_k-\mu_k\eta_k\gamma_k}{2} + \rho_{k+1}\Big(\rho_{k+1}(1-\alpha_k) - \tfrac{1-\alpha_k(2-\beta_k)}{2\varepsilon_{k+1}} - \tfrac{\tau_{k+1}\eta_{k+1}\gamma_{k+1}}{2}\Big)\Big)\|x^{k-1} - x^k\|^2 \\ &- \Big(\tfrac{\gamma_{k+1}}{2\sigma_{k+1}} - \tfrac{\eta_{k+1}\gamma_{k+1}}{2}\Big(\tfrac{1}{\mu_{k+1}} + \tfrac{\rho_{k+1}}{\tau_{k+1}}\Big)\Big)\|y^{k+1} - y^k\|^2 - \gamma_{k+1}\big(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{k+1})\big). \end{aligned}$$

Since $\sigma_k = t^2\gamma_k$ and $\xi_k = t^2\gamma_k^2\eta_k^2$, by selecting $\varepsilon_k := \frac{1}{2\rho_k}$ and $\mu_k := 2(1+\epsilon)t\xi_k^{1/2}$ one has that $\tilde{\mathcal{U}}_k = \mathcal{U}_k$. Note also that the coefficient of $\|x^k - x^{k-1}\|$ in $\mathcal{U}_k$ is strictly positive since the stepsize update ensures $\xi_k \leq 1/4\nu^2$. Moreover, with this choice the above inequality becomes

$$\begin{aligned} \mathcal{U}_{k+1} \leq \mathcal{U}_k &- \gamma_k(1 + \rho_k - \rho_{k+1}^2)P_{k-1} - \gamma_{k+1}\big(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{k+1})\big) \\ &- \left(\tfrac{1-4\xi_k(1+\epsilon)}{4} + \rho_{k+1}\left(\rho_{k+1}\alpha_k(1-\beta_k) - \tfrac{\tau_{k+1}\xi_{k+1}^{1/2}}{2t}\right)\right)\|x^{k-1} - x^k\|^2 \\ &- \tfrac{1}{2t}\left(\tfrac{1+2\epsilon}{2t(1+\epsilon)} - \tfrac{\xi_{k+1}^{1/2}\rho_{k+1}}{\tau_{k+1}}\right)\|y^{k+1} - y^k\|^2. \end{aligned}$$

We now set $\tau_{k+1} := \rho_{k+1}\mu_{k+1} = 2(1+\epsilon)t\xi_{k+1}^{1/2}\rho_{k+1}$ so the inequality overall simplifies to the one of the statement

$$\begin{aligned} \mathcal{U}_{k+1} \leq \mathcal{U}_k &- \gamma_k(1 + \rho_k - \rho_{k+1}^2)P_{k-1} - \gamma_{k+1}\big(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{k+1})\big) \\ &\underbrace{- \Big(\tfrac{1-4\xi_k(1+\epsilon)}{4} + \rho_{k+1}(\rho_{k+1}\alpha_k(1-\beta_k) - \rho_{k+1}\xi_{k+1}(1+\epsilon))\Big)}_{\geq \frac{\epsilon}{4(1+\epsilon)}}\|x^{k-1} - x^k\|^2 \\ &- \tfrac{\epsilon}{2t^2(1+\epsilon)}\|y^{k+1} - y^k\|^2, \end{aligned}$$

up to ensuring that the coefficient of P_{k-1} is positive and that the inequality for the coefficient of $\|x^k - x^{k-1}\|^2$ holds. The former is of trivial verification, having

$\rho_{k+1} \leq \sqrt{1+\rho_k}$. It thus remains to show that

$$\tfrac{1}{4} - \tfrac{\epsilon}{4(1+\epsilon)} - \xi_k(1+\epsilon) - \rho_{k+1}^2(\delta_k + \xi_{k+1}(1+\epsilon)) \geq 0$$

holds for every k. By using the fact that $\xi_{k+1} = (t\eta_{k+1}\gamma_{k+1})^2 = (t\eta_{k+1}\gamma_k)^2\rho_{k+1}^2$, this reduces to the second-order inequality (in ρ_{k+1}^2)

$$(t\eta_{k+1}\gamma_k)^2(1+\epsilon)\rho_{k+1}^4 + \delta_k\rho_{k+1}^2 - \left[\tfrac{1}{4(1+\epsilon)} - \xi_k(1+\epsilon)\right] \leq 0.$$

Note that the bound $\gamma_k \leq \frac{1}{2vt\eta_k}$ implies $\frac{1}{4(1+\epsilon)} - \xi_k(1+\epsilon) \geq \frac{v^2-(1+\epsilon)^2}{4v^2(1+\epsilon)} > 0$, thus ensuring that the inequality always admits solutions for small enough ρ_{k+1}^2. Namely, letting $\bar{\xi}_k := \xi_k(1+\epsilon)^2$

$$\begin{aligned}\rho_{k+1}^2 &\leq \frac{-\delta_k + \sqrt{\delta_k^2 + (t\eta_{k+1}\gamma_k)^2(1-4\bar{\xi}_k)}}{2(1+\epsilon)(t\eta_{k+1}\gamma_k)^2} \\ &= \frac{1-4\bar{\xi}_k}{2(1+\epsilon)\left(\delta_k + \sqrt{\delta_k^2 + (t\eta_{k+1}\gamma_k)^2(1-4\bar{\xi}_k)}\right)},\end{aligned}$$

which is indeed guaranteed by one of the bounds on $\gamma_{k+1} = \gamma_k\rho_{k+1}$. Note that the second expression removes the singularity in case $\eta_{k+1} = 0$.

♠ *(ii)* Boundedness of the sequence follows from the fact that $\frac{1}{2}\|x^k - x^\star\|^2 + \frac{1}{2t^2}\|y^k - y^\star\|^2 \leq \mathcal{U}_k \leq \mathcal{U}_1$, where the first inequality follows by definition of $\mathcal{U}_k$, cf. (3.7), and the second one from assertion *(i)*. In particular, there exists a convex and compact set $\mathcal{V} \subseteq \mathbb{R}^n$ that contains $(x^k)_{k\in\mathbb{N}}$. Up to considering a suitable enlargement $\overline{\mathcal{V}}$ as in Lemma A.2, $\ell_k \leq c_k \leq L_{f,\overline{\mathcal{V}}}$ holds for every k by Lemma 2.1*(iii)* and Lemma A.2, where $L_{f,\overline{\mathcal{V}}} > 0$ is a Lipschitz modulus for ∇f on $\overline{\mathcal{V}}$. To prove the lower bound on the stepsize, we will show that whenever $\gamma_{k+1} < \gamma_k$ occurs, then necessarily γ_{k+1} is greater than some constant $\hat{\gamma}$ as in the statement. The proof will then follow from a trivial inductive argument. Suppose that $\gamma_{k+1} < \gamma_k$. If $\gamma_{k+1} = \frac{1}{2vt\eta_{k+1}}$, then clearly $\gamma_{k+1} \geq \frac{1}{2vt\eta_{\max}}$. Otherwise, necessarily

$$\begin{aligned}\gamma_k^2 > \gamma_{k+1}^2 &= \frac{\gamma_k^2(1-4\bar{\xi}_k)}{2(1+\epsilon)\left(\sqrt{\delta_k^2 + (t\eta_{k+1}\gamma_k)^2(1-4\bar{\xi}_k)} + \delta_k\right)} \\ &\geq \frac{\gamma_k^2\bar{\nu}}{2(1+\epsilon)\left(\sqrt{\delta_k^2 + (t\eta_{\max}\gamma_k)^2\bar{\nu}} + \delta_k\right)},\end{aligned} \tag{A.6}$$

where $\bar{\nu} := \frac{\nu^2-(1+\epsilon)^2}{\nu^2}$, and the second inequality uses the fact that $1 - 4\bar{\xi}_k \geq \bar{\nu} > 0$ together with the fact that $(0, \infty) \ni x \mapsto \frac{x}{\sqrt{b^2+a^2x}+b}$ is increasing for any value of $a, b \in \mathbb{R}$. We now distinguish two cases:

$\diamond$ Case 1. $\delta_k \leq 0$. By comparing the outermost terms of the chain of inequalities in (A.6) we obtain that

$$\sqrt{\delta_k^2 + (t\eta_{\max}\gamma_k)^2\bar{\nu}} + \delta_k > \tfrac{\bar{\nu}}{2(1+\epsilon)} \quad \Leftrightarrow \quad \delta_k > \tfrac{\bar{\nu}}{4(1+\epsilon)} - (1+\epsilon)(t\eta_{\max}\gamma_k)^2,$$

and in particular $\gamma_k > \frac{\sqrt{\bar{\nu}}}{2(1+\epsilon)t\eta_{\max}}$. Then, since $x \mapsto \frac{1}{\sqrt{x^2+b^2}+x}$ is decreasing for any value of $b \in \mathbb{R}$, by setting δ_k equal to 0 in (A.6) we obtain

$$\gamma_{k+1}^2 \geq \gamma_k \frac{\sqrt{\bar{\nu}}}{2(1+\epsilon)t\eta_{\max}} \geq \frac{\bar{\nu}}{(2(1+\epsilon)t\eta_{\max})^2}.$$

$\diamond$ Case 2. $\delta_k > 0$ or, equivalently, $\gamma_k c_k > 1$. Denoting $\alpha := \gamma_k L_{f,\overline{\nu}}$, one has that $\alpha \geq \gamma_k \ell_k$ and $\alpha \geq \gamma_k c_k > 1$, hence that $\delta_k \leq \alpha(\alpha-1)$. Arguing as in the previous case, this time by setting $\delta_k \leftarrow \alpha(\alpha-1)$ in (A.6), yields

$$\begin{aligned}
\gamma_{k+1}^2 L_{f,\overline{\nu}}^2 &\geq \frac{\alpha^2\bar{\nu}}{2(1+\epsilon)\left(\sqrt{\alpha^2(\alpha-1)^2 + (t\eta_{\max}\gamma_k)^2\bar{\nu}} + \alpha(\alpha-1)\right)} \\
&= \frac{\alpha\bar{\nu}}{2(1+\epsilon)\left(\sqrt{(\alpha-1)^2 + (t\eta_{\max}/L_{f,\overline{\nu}})^2\bar{\nu}} + \alpha - 1\right)} \\
&\geq \min\left\{\tfrac{L_{f,\overline{\nu}}\sqrt{\bar{\nu}}}{2(1+\epsilon)t\eta_{\max}},\ \tfrac{\bar{\nu}}{4(1+\epsilon)}\right\},
\end{aligned}$$

where the last inequality owes to the fact that $[1, \infty) \ni x \mapsto \frac{x}{\sqrt{(x-1)^2+b^2}+x-1}$ attains the infimum at either 1 or ∞ for any $b \in \mathbb{R}$.

Putting all the cases together yields

$$\gamma_k \geq \hat{\gamma} := \min\left\{\gamma_0,\ \tfrac{\sqrt[4]{\bar{\nu}}}{\sqrt{2(1+\epsilon)t\eta_{\max}L_{f,\overline{\nu}}}},\ \tfrac{\sqrt{\bar{\nu}}}{2\sqrt{(1+\epsilon)}L_{f,\overline{\nu}}},\ \tfrac{1}{2\nu t\eta_{\max}},\ \tfrac{\sqrt{\bar{\nu}}}{2(1+\epsilon)t\eta_{\max}}\right\} > 0 \tag{A.7}$$

(where we remind that $\bar{\nu} = \frac{\nu^2-(1+\epsilon)^2}{\nu^2}$), establishing the claim.

♠ *(iii)* That $Q_k \to 0$ as $k \to \infty$ follows from a telescoping argument in assertion *(i)*, since $(\gamma_k)_{k\in\mathbb{N}}$ is bounded away from zero by assertion *(ii)*. As to $(P_k)_{k\in\mathbb{N}}$, if $\limsup_{k\to\infty}(1 + \rho_k - \rho_{k+1}^2) > 0$, then the same telescoping argument yields the claim. Alternatively, by the stepsize update rule one has that $1 + \rho_k - \rho_{k+1}^2 \geq 0$ and therefore $1 + \rho_k - \rho_{k+1}^2 \to 0$, from which it easily follows that $\liminf_{k\to\infty} \rho_k > 1$ and that therefore $\gamma_k \to \infty$. Consequently, since $\gamma_k(1 + \rho_k)P_{k-1} \leq \mathcal{U}_k \leq \mathcal{U}_0$, this directly proves that $P_k \to 0$.

♠ *(iv)* Let $(\hat{x}, \hat{y})$ denote a limit point of $(x^k, y^k)_{k\in\mathbb{N}}$. Telescoping the inequality in assertion *(i)* yields that both $\|x^{k+1} - x^k\|$ and $\|y^{k-1} - y^k\|$ vanish. By the optimality condition for step 3.1.4 we have

$$\tfrac{1}{\gamma_{k+1}}(x^k - x^{k+1}) \in \nabla f(x^k) + A^\top y^{k+1} + \partial g(x^{k+1}). \tag{A.8a}$$

Passing to the limit along the subsequence converging to $(\hat{x}, \hat{y})$, using the fact that $(\gamma_k)_{k\in\mathbb{N}}$ is bounded away from zero by assertion *(ii)* and outer semicontinuity of ∂g yield $0 \in \nabla f(\hat{x}) + \partial g(\hat{x}) + A^\top \hat{y}$. Similarly, for the dual variable,

$$\tfrac{1}{\sigma_{k+1}}(y^k - y^{k+1}) + \tfrac{\gamma_{k+1}}{\gamma_k} A(x^k - x^{k-1}) \in \partial h^*(y^{k+1}) - Ax^k. \tag{A.8b}$$

A trivial induction argument reveals that $\frac{\gamma_{k+1}}{\gamma_k}$ is upper bounded by $\max\left\{\frac{\gamma_0}{\gamma_{-1}}, \frac{1}{2}(1+\sqrt{5})\right\}$. Therefore, passing to the limit along the same subsequence and recalling that $\sigma_{k+1} = t^2\gamma_{k+1}$ is bounded away from zero yield $0 \in \partial h^*(\hat{y}) - A\hat{x}$. Along with the previous inclusion, primal-dual optimality of the limit pair $(\hat{x}, \hat{y})$ is established. Therefore, any limit point of $(x^k, y^k)_{k\in\mathbb{N}}$ is a primal-dual optimal pair. Suppose that (x^∞, y^∞) and $(\hat{x}^\infty, \hat{y}^\infty)$ are two primal-dual optimal limit points of $(x^k, y^k)_{k\in\mathbb{N}}$. Define $v^k = (x^k, \frac{1}{t}y^k)$, and consistently $v^\infty = (x^\infty, \frac{1}{t}y^\infty)$, $\hat{v}^\infty = (\hat{x}^\infty, \frac{1}{t}\hat{y}^\infty)$. Observe that

$$\langle v^k, v^\infty - \hat{v}^\infty\rangle = \mathcal{U}_k(\hat{x}^\infty, \hat{y}^\infty) - \mathcal{U}_k(x^\infty, y^\infty) + \tfrac{1}{2}\|v^\infty\|^2 - \tfrac{1}{2}\|\hat{v}^\infty\|^2,$$

and since $(\mathcal{U}_k(x^\star, y^\star))_{k\in\mathbb{N}}$ is convergent for all primal-dual optimal pairs $(x^\star, y^\star)$, then so is $(\langle v^k, v^\infty - \hat{v}^\infty\rangle)_{k\in\mathbb{N}}$. Passing to the limit along the two converging subsequences thus yields $\langle v^\infty, v^\infty - \hat{v}^\infty\rangle = \langle \hat{v}^\infty, v^\infty - \hat{v}^\infty\rangle$, which after rearranging results in $\|v^\infty - \hat{v}^\infty\|^2 = 0$ establishing uniqueness of the optimal limit point. □

Funding This work was supported by: the Research Foundation Flanders (FWO) postdoctoral grant 12Y7622N and research projects G081222N, G033822N, and G0A0920N; Research Council KU Leuven C1 project No. C14/24/103; European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 953348; Japan Society for the Promotion of Science (JSPS) KAKENHI grants JP21K17710 and JP24K20737.

Data availability All data generated or analysed during this study are included in this manuscript.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

1. Alacaoglu, A., Böhm, A., Malitsky, Y.: Beyond the golden ratio for variational inequality algorithms. J. Mach. Learn. Res. **24**(172), 1–33 (2023)

2. Altschuler, J.M., Parrilo, P.A.: Acceleration by stepsize hedging II: silver stepsize schedule for smooth convex optimization. arXiv preprint arXiv:2309.16530 (2023)
3. Attouch, H., Boţ, R.I., Nguyen, D.K.: Fast convex optimization via closed-loop time scaling of gradient dynamics. arXiv preprint arXiv:2301.00701 (2023)
4. Baillon, J.B., Haddad, G.: Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. Israel J. Math. **26**(2), 137–150 (1977)
5. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**(1), 141–148 (1988)
6. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics, Springer (2017)
7. Beck, A.: First-Order Methods in Optimization. SIAM, Philadelphia, PA (2017)
8. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific (2016)
9. Bianchi, P., Hachem, W.: A primal-dual algorithm for distributed optimization. In: IEEE 53rd Annual Conference on Decision and Control (CDC), pp. 4240–4245 (2014)
10. Böhm, A.: Solving nonconvex-nonconcave min-max problems exhibiting weak Minty solutions. arXiv preprint arXiv:2201.12247 (2022)
11. Boţ, R.I., Hendrich, C.: A Douglas-Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. SIAM J. Optim. **23**(4), 2541–2565 (2013)
12. Boţ, R.I., Sedlmayer, M., Vuong, P.T.: A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs. J. Mach. Learn. Res. **24**, 1–37 (2023)
13. Briceño-Arias, L.M., Davis, D.: Forward-backward-half forward algorithm for solving monotone inclusions. SIAM J. Optim. **28**(4), 2839–2871 (2018)
14. Burdakov, O., Dai, Y.H., Huang, N.: Stabilized barzilai-borwein method. J. Comput. Math. **37**(6), 916–936 (2019)
15. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis **40**(1), 120–145 (2011)
16. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transact. Intell. Syst. Technol. (TIST) **2**, 1–27 (2011)
17. Chang, X.K., Yang, J., Zhang, H.: Golden ratio primal-dual algorithm with linesearch. SIAM J. Optim. **32**(3), 1584–1613 (2022)
18. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer New York (2011)
19. Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Valued Var. Anal. **20**(2), 307–330 (2012)
20. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
21. Dai, Y.H., Fletcher, R.: Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. Numer. Math. **100**(1), 21–47 (2005)
22. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-Valued Var. Anal. **25**(4), 829–858 (2017)
23. De Marchi, A., Themelis, A.: Proximal gradient algorithms under local Lipschitz gradient continuity: a convergence and robustness analysis of PANOC. J. Optim. Theory Appl. **194**, 771–794 (2022)
24. Defazio, A., Zhou, B., Xiao, L.: Grad-GradaGrad? A Non-monotone Adaptive Stochastic Gradient Method. arXiv preprint arXiv:2206.06900 (2022)
25. Diakonikolas, J., Daskalakis, C., Jordan, M.: Efficient methods for structured nonconvex-nonconcave min-max optimization. In: International Conference on Artificial Intelligence and Statistics, pp. 2746–2754. PMLR (2021)
26. Drori, Y., Sabach, S., Teboulle, M.: A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. Oper. Res. Lett. **43**(2), 209–214 (2015)
27. Fercoq, O., Bianchi, P.: A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. SIAM J. Optim. **29**(1), 100–134 (2019)
28. Giselsson, P.: Nonlinear forward-backward splitting with projection correction. SIAM J. Optim. **31**(3), 2199–2226 (2021)
29. Goldstein, T., Li, M., Yuan, X.: Adaptive primal-dual splitting methods for statistical learning and image processing. Adv. Neural Inform. Process. Syst. **28** (2015)

30. Goldstein, T., Li, M., Yuan, X., Esser, E., Baraniuk, R.: Adaptive primal-dual hybrid gradient methods for saddle-point problems. arXiv preprint arXiv:1305.0546 (2013)
31. Grimmer, B., Shu, K., Wang, A.L.: Accelerated gradient descent via long steps. arXiv preprint arXiv:2309.09961 (2023)
32. Hastie, T., Friedman, J., Tibshirani, R.: The Elements of Statistical Learning. Springer, New York (2001)
33. He, B., Yuan, X.: Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. SIAM J. Imag. Sci. **5**(1), 119–149 (2012)
34. Jezierska, A., Chouzenoux, E., Pesquet, J.C., Talbot, H.: A primal-dual proximal splitting approach for restoring data corrupted with Poisson-Gaussian noise. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1085–1088. IEEE (2012)
35. Komodakis, N., Pesquet, J.C.: Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems. IEEE Signal Process. Mag. **32**(6), 31–54 (2015)
36. Latafat, P.: Distributed proximal algorithms for large-scale structured optimization. Ph.D. thesis, KU Leuven (2020)
37. Latafat, P., Bemporad, A., Patrinos, P.: Plug and play distributed model predictive control with dynamic coupling: a randomized primal-dual proximal algorithm. In: European Control Conference (ECC), pp. 1160–1165 (2018)
38. Latafat, P., Freris, N.M., Patrinos, P.: A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. IEEE Trans. Autom. Control **64**(10), 4050–4065 (2019)
39. Latafat, P., Patrinos, P.: Asymmetric forward-backward-adjoint splitting for solving monotone inclusions involving three operators. Comput. Optim. Appl. **68**(1), 57–93 (2017)
40. Latafat, P., Patrinos, P.: Primal-dual proximal algorithms for structured convex optimization: A unifying framework. In: P. Giselsson, A. Rantzer (eds.) Large-Scale and Distributed Optimization, pp. 97–120. Springer International Publishing (2018)
41. Latafat, P., Stella, L., Patrinos, P.: New primal-dual proximal algorithm for distributed optimization. In: 55th IEEE Conference on Decision and Control (CDC), pp. 1959–1964 (2016)
42. Latafat, P., Themelis, A., Ahookhosh, M., Patrinos, P.: Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity. SIAM J. Optim. **32**(3), 2230–2262 (2022)
43. Latafat, P., Themelis, A., Patrinos, P.: Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems. Math. Program. **193**(1), 195–224 (2022)
44. Latafat, P., Themelis, A., Villa, S., Patrinos, P.: On the convergence of proximal gradient methods for convex simple bilevel optimization. arXiv preprint arXiv:2305.03559 (2023)
45. Li, X., Orabona, F.: On the convergence of stochastic gradient descent with adaptive stepsizes. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 983–992. PMLR (2019)
46. Malitsky, Y.: Golden ratio algorithms for variational inequalities. Math. Program. **184**(1), 383–410 (2020)
47. Malitsky, Y., Mishchenko, K.: Adaptive gradient descent without descent. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 6702–6712. PMLR (2020)
48. Malitsky, Y., Mishchenko, K.: Adaptive proximal gradient method for convex optimization. arXiv preprint arXiv:2308.02261 (2023)
49. Malitsky, Y., Pock, T.: A first-order primal-dual algorithm with linesearch. SIAM J. Optim. **28**(1), 411–432 (2018)
50. Malitsky, Y., Tam, M.K.: A forward-backward splitting method for monotone inclusions without cocoercivity. SIAM J. Optim. **30**(2), 1451–1472 (2020)
51. Marumo, N., Takeda, A.: Parameter-free accelerated gradient descent for nonconvex minimization (2022)
52. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
53. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
54. Pedregosa, F., Gidel, G.: Adaptive three operator splitting. In: International Conference on Machine Learning, pp. 4085–4094. PMLR (2018)
55. Pethick, T., Latafat, P., Patrinos, P., Fercoq, O., Cevher, V.: Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In: International Conference on Learning Representations (2022)

56. Raydan, M.: On the Barzilai and Borwein choice of steplength for the gradient method. IMA J. Numer. Anal. **13**(3), 321–326 (1993)
57. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970)
58. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer (2009)
59. Ryu, E.K., Vũ, B.C.: Finding the forward-Douglas-Rachford-forward method. J. Optim. Theory Appl. **184**(3), 858–876 (2020)
60. Salzo, S.: The variable metric forward-backward splitting algorithm under mild differentiability assumptions. SIAM J. Optim. **27**(4), 2153–2181 (2017)
61. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. MIT Press (2012)
62. Tan, C., Ma, S., Dai, Y.H., Qian, Y.: Barzilai-Borwein step size for stochastic gradient descent. Adv. Neural Inform. Process. Syst. **29** (2016)
63. Teboulle, M., Vaisbourd, Y.: An elementary approach to tight worst case complexity analysis of gradient based methods. Mathematical Programming pp. 1–34 (2022)
64. Thong, D.V., Van Hieu, D., Rassias, T.M.: Self adaptive inertial subgradient extragradient algorithms for solving pseudomonotone variational inequality problems. Optim. Lett. **14**(1), 115–144 (2020)
65. Vladarean, M.L., Malitsky, Y., Cevher, V.: A first-order primal-dual method with adaptivity to local smoothness. Adv. Neural. Inf. Process. Syst. **34**, 6171–6182 (2021)
66. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**(3), 667–681 (2013)
67. Ward, R., Wu, X., Bottou, L.: AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In: K. Chaudhuri, R. Salakhutdinov (eds.) Proceedings of the 36th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 97, pp. 6677–6686. PMLR (2019)
68. Yan, M.: A new primal-dual algorithm for minimizing the sum of three functions with a linear operator. J. Sci. Comput. **76**(3), 1698–1717 (2018)
69. Yang, J.: Self-adaptive inertial subgradient extragradient algorithm for solving pseudomonotone variational inequalities. Appl. Anal. **100**(5), 1067–1078 (2021)
70. Yang, J., Liu, H.: A modified projected gradient method for monotone variational inequalities. J. Optim. Theory Appl. **179**(1), 197–211 (2018)
71. Yurtsever, A., Gu, A., Sra, S.: Three operator splitting with subgradients, stochastic gradients, and adaptive learning rates. Adv. Neural. Inf. Process. Syst. **34**, 19743–19756 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.