Proceedings of Machine Learning Research vol 242:197–208, 2024

On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms

Puya Latafat PUYA.LATAFAT@KULEUVEN.BE
Department of Electrical Engineering (ESAT-STADIUS)
KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Andreas Themelis ANDREAS.THEMELIS@EES.KYUSHU-U.AC.JP
Faculty of Information Science and Electrical Engineering (ISEE)
Kyushu University, 744 Motooka, Nishi-ku 819-0395, Fukuoka, Japan

Panagiotis Patrinos PANOS.PATRINOS@ESAT.KULEUVEN.BE
Department of Electrical Engineering (ESAT-STADIUS)
KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Abstract

Building upon recent works on linesearch-free adaptive proximal gradient methods, this paper proposes AdaPG^{q,r}, a framework that unifies and extends existing results by providing larger stepsize policies and improved lower bounds. Different choices of the parameters q and r are discussed and the efficacy of the resulting methods is demonstrated through numerical simulations. In an attempt to better understand the underlying theory, its convergence is established in a more general setting that allows for time-varying parameters. Finally, an adaptive alternating minimization algorithm is presented by exploring the dual setting. This algorithm not only incorporates additional adaptivity, but also expands its applicability beyond standard strongly convex settings.

Keywords: Convex minimization, proximal gradient method, alternating minimization algorithm, locally Lipschitz gradient, linesearch-free adaptive stepsizes

1. Introduction

The proximal gradient (PG) method is the natural extension of gradient descent for constrained and nonsmooth problems. It addresses nonsmooth minimization problems by splitting them as

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

where f is here assumed *locally* Lipschitz differentiable, and g possibly nonsmooth but with an easy-to-compute proximal mapping, while both being convex (see Assumption 2.1 for details). It has long been known that performance of first-order methods can be drastically improved by an appropriate stepsize selection as evident in the success of linesearch based approaches.

Substantial effort has been devoted to developing adaptive methods. Most notably, in the context of stochastic (sub)gradient descent, numerous adaptive methods have been proposed starting with Duchi et al. (2011). We only point the reader to few recent works in this area Li and Orabona (2019); Ward et al. (2019); Yurtsever et al. (2021); Ene et al. (2021); Defazio et al. (2022); Ivgi et al. (2023). However, although applicable to a more general setting, such approaches tend to suffer from diminishing stepsizes, which can hinder their performance.

Closer to our setting are recent works Grimmer et al. (2023); Altschuler and Parrilo (2023) which consider smooth optimization problems, and propose predefined stepsize patterns. These methods obtain accelerated worst-case rates under global Lipschitz continuity assumptions. We also mention the recent work Li and Lan (2023) in the constrained smooth setting which, while also being bound to a global Lipschitz continuity assumption, uses an adaptive estimate for the Lipschitz modulus and achieves an *accelerated* worst-case rate.

In this paper, we extend recent results pioneered in Malitsky and Mishchenko (2020) and later further developed in Latafat et al. (2023b); Malitsky and Mishchenko (2023), where novel (self-) adaptive schemes are developed. We provide a unified analysis that bridges together and improves upon all these works by enabling larger stepsizes and, in some cases, providing tighter lower bounds. Adaptivity refers to the fact that, in contrast to linesearch methods that employ a *look forward* approach based on trial and error to ensure a sufficient descent in the cost, we *look backward* to yield stepsizes only based on past information. Specifically, we estimate the Lipschitz modulus of ∇f at consecutive iterates $x^{k-1}, x^k \in \mathbb{R}^n$ generated by the algorithm using the quantities

$$\ell_k := \frac{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1} \rangle}{\|x^k - x^{k-1}\|^2} \quad \text{and} \quad L_k := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}. \tag{1.1}$$

Throughout, we stick to the convention $\frac{0}{0} = 0$ so that ℓ_k and L_k are well-defined, positive real numbers. In addition, we adhere to $\frac{1}{0} = \infty$. Note also that

$$\ell_k \leq L_k \leq L_{f,\mathcal{V}} \tag{1.2}$$

holds whenever $L_{f,\mathcal{V}}$ is a Lipschitz modulus for ∇f on a convex set $\mathcal{V}$ containing x^{k-1} and x^k. Despite the mere dependence of these quantities on the previous iterates, they provide a sufficiently refined estimate of the local geometry of f. In fact, a carefully designed stepsize update rule not only ensures that the stepsize sequence is separated from zero, but also that a sufficient descent-type inequality can be indirectly ensured between (x^{k+1}, x^k) and (x^k, x^{k-1}) without any backtracks.

The ultimate deliverable of this manuscript is the general adaptive framework outlined in Algorithm 2.1. A special case of it is here condensed into a two-parameter simplified algorithm.

AdaPG^{q,r} Fix $x^{-1} \in \mathbb{R}^n$ and $\gamma_0 = \gamma_{-1} > 0$. With ℓ_k and L_k as in (1.1), starting from $x^0 = \operatorname{prox}_{\gamma_0 g}(x^{-1} - \gamma_0 \nabla f(x^{-1}))$, iterate for $k = 0, 1, \ldots$

$$\gamma_{k+1} = \gamma_k \min\left\{ \sqrt{\tfrac{1}{q} + \tfrac{\gamma_k}{\gamma_{k-1}}}, \sqrt{\tfrac{1-\frac{r}{q}}{\left[\gamma_k^2 L_k^2 + 2\gamma_k \ell_k (r-1) - (2r-1)\right]_+}} \right\} \tag{1.3a}$$

$$x^{k+1} = \operatorname{prox}_{\gamma_{k+1} g}(x^k - \gamma_{k+1} \nabla f(x^k)) \tag{1.3b}$$

Theorem 1.1 *Under Assumption 2.1, for any $q > r \geq \frac{1}{2}$ the sequence $(x^k)_{k\in\mathbb{N}}$ generated by AdaPG^{q,r} converges to some $x^\star \in \arg\min \varphi$. If in addition $q \leq \frac{1}{2}(3 + \sqrt{5})$, then*

$$\gamma_k \geq \gamma_{\min} := \sqrt{\tfrac{1-\frac{r}{q}}{\max\{1,q\}}} \tfrac{1}{L_{f,\mathcal{V}}} \quad \textit{holds for all } k \geq 2\left\lceil \log_{1+\frac{1}{q}}\left(\tfrac{1}{\gamma_0 L_{f,\mathcal{V}}}\right)\right\rceil_+,$$

where $L_{f,\mathcal{V}}$ is a Lipschitz modulus for ∇f on a convex and compact set $\mathcal{V}$ that contains $(x^k)_{k\in\mathbb{N}}$. Moreover, $\min_{k\leq K}(\varphi(x^k) - \min\varphi) \leq \frac{\mathcal{U}_1(x^\star)}{\sum_{k=1}^{K+1}\gamma_k}$ holds for every $K \geq 1$, where $\mathcal{U}_1(x^\star)$ is as in (2.4).

The above worst-case sublinear rate depends on the aggregate of the stepsize sequence, providing a partial explanation for the fast convergence of the algorithm observed in practice. AdaPG^{q,r}

and Theorem 1.1 are particular instances of the general framework provided in Section 2, see Remark 2.2 and Theorem 2.7 for the details. Specific choices of the parameters q, r nevertheless allow AdaPG^{q,r} to embrace and extend existing algorithms:

- $r = \frac{1}{2}$ and $q = 1$. Then, $\gamma_{k+1} = \gamma_k \min\left\{\sqrt{1+\frac{\gamma_k}{\gamma_{k-1}}}, \frac{1}{\sqrt{2\left[\gamma_k^2 L_k^2 - \gamma_k \ell_k\right]_+}}\right\}$ coincides with the update in (Latafat et al., 2023b, Alg. 2.1) with second term improved by a $\sqrt{2}$ factor.
- Owing to the relation $\gamma_k^2 L_k^2 - \gamma_k \ell_k \leq \gamma_k^2 L_k^2$, the case above is also a proximal extension of (Malitsky and Mishchenko, 2023, Alg. 1) which considers $\gamma_{k+1} = \min\left\{\gamma_k\sqrt{1+\frac{\gamma_k}{\gamma_{k-1}}}, \frac{1}{\sqrt{2}L_k}\right\}$ when $g = 0$, and which in turn is also a strict improvement over the previous work Malitsky and Mishchenko (2020).
- $r = \frac{3}{4}$ and $q = \frac{3}{2}$. Then, $\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\frac{2}{3}+\frac{\gamma_k}{\gamma_{k-1}}}, \frac{1}{\sqrt{[2\gamma_k^2 L_k^2 - 1 - \gamma_k \ell_k]_+}}\right\}$ recovers the update rule (Malitsky and Mishchenko, 2023, Alg. 2) (in fact tighter because of the extra $-\gamma_k \ell_k$ term).

The interplay between the parameters can then be understood by noting that $\sqrt{\frac{1}{q}+\frac{\gamma_k}{\gamma_{k-1}}}$ allows the algorithm to recover from a potentially small stepsize, which can only decrease for a controlled number of iterations and will then rapidly enter a phase where it increases linearly until a certain threshold is reached, see the proof of Theorem 2.7. A smaller q allows for a more aggressive recovery, but comes at the cost of more conservative second term. As for r, values in the range $[1/2, 1]$, such as in the combinations reported in Table 1, work well in practice.

$1 - r/q$	q	r	$\gamma_{\min} L_{f,\mathcal{V}}$
$1/4$	$10/9$	$5/6$	$3/2\sqrt{1/10} \approx 0.47$
$2/5$	$8/5$	$24/25$	$1/2$
$1/2$	$5/3$	$5/6$	$\sqrt{3/10} \approx 0.55$
$1/2$	$3/2$	$3/4$	$1/\sqrt{3} \approx 0.57$
$1/2$	1	$1/2$	$1/\sqrt{2} \approx 0.71$
$3/5$	$5/2$	1	$\sqrt{6}/5 \approx 0.49$

Table 1. Suggested options for q and r in AdaPG^{q,r}. Green cells strike a nice balance between aggressive increases and large lower bounds ($\gamma_{\min}$) for the stepsize sequence, while the orange cell yields the largest theoretical lower bound. Here, $L_{f,\mathcal{V}}$ is a local Lipschitz modulus for ∇f as in Theorem 1.1.

As a final contribution, an adaptive variant of the *alternating minimization algorithm* (AMA) of Tseng (1991) is proposed that addresses composite problems of the form

$$\operatorname*{minimize}_{x \in \mathbb{R}^n} \psi_1(x) + \psi_2(Ax). \tag{CP}$$

AMA is particularly interesting in settings where ψ_1 is either nonsmooth or its gradient is computationally demanding. Its convergence was established in Tseng (1991) by framing it as the dual form of the splitting method introduced in Gabay (1983), and acceleration techniques have also been adapted to this setting Goldstein et al. (2014). In contrast to existing methods, ours not only incorporates an adaptive stepsize mechanism but also relaxes the strong convexity assumption to mere *local* strong convexity, see Assumption 3.1 for details. Due to space limitations, some proofs are deferred to the preprint version Latafat et al. (2023a).

2. A general framework for adaptive proximal gradient methods

In this section we consider plain proximal gradient iterations of the form

$$x^{k+1} = \operatorname{prox}_{\gamma_{k+1} g}\bigl(x^k - \gamma_{k+1}\nabla f(x^k)\bigr), \tag{2.1}$$

where $(\gamma_k)_{k\in\mathbb{N}}$ is a sequence of strictly positive stepsize parameters. The main oracles of the method are gradient and proximal maps (see (Beck, 2017, §6) for examples of *proximable* functions). Whenever g is convex, for any $\gamma > 0$ it is well known that $\operatorname{prox}_{\gamma g}$ is *firmly nonexpansive* (Bauschke and Combettes, 2017, §4.1 and Prop. 12.28), a property stronger than Lipschitz continuity. We here show that even when the stepsizes are time-varying as in (2.1) a similar property still holds for the iterates therein. This fact is a refinement of (Malitsky and Mishchenko, 2023, Lem. 12) that follows after an application of Cauchy-Schwarz and that will be used in our main descent inequality.

Lemma 2.1 (FNE-like inequality) *Suppose that g is convex and that f is differentiable. Then, for any $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$ and with $H_k \coloneqq \mathrm{id} - \gamma_k\nabla f$, proximal gradient iterates (2.1) satisfy*

$$\|x^{k+1} - x^k\|^2 \le \rho_{k+1}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle \le \rho_{k+1}^2\|H_k(x^{k-1}) - H_k(x^k)\|^2. \tag{2.2}$$

Throughout, we study problem (P) under the following assumptions.

Assumption 2.1 (Requirements for problem (P))

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 There exists $x^\star \in \operatorname{arg\,min} f + g$.

The main adaptive framework, involving two time-varying parameters q_k, ξ_k, is given in Algorithm 2.1. The shorthand notation $\mathbb{1}_{\gamma_k\ell_k\ge 1}$ equals 1 if $\gamma_k\ell_k \ge 1$ and 0 otherwise, while for $t \in \mathbb{R}$ we denote $[t]_+ \coloneqq \max\{0, t\}$.

Algorithm 2.1 General adaptive proximal gradient framework

REQUIRE starting point $x^{-1} \in \mathbb{R}^n$, stepsizes $\gamma_0 = \gamma_{-1} > 0$
parameters $\frac{1}{2} < q_{\min} \le q_{\max}$, $0 < \xi_{\min} \le 2q_{\min} - 1$,

INITIALIZE $x^0 = \operatorname{prox}_{\gamma_0 g}(x^{-1} - \gamma_0\nabla f(x^{-1}))$, $\rho_0 = 1$, $q_0 \in [q_{\min}, q_{\max}]$, $\xi_0 \ge \xi_{\min}$

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

2.1.1: Let ℓ_k and L_k be as in (1.1), and choose q_{k+1}, ξ_{k+1} such that

$$\xi_{k+1} \ge \xi_{\min}, \quad r_{k+1} \coloneqq \tfrac{q_{k+1}}{1+\xi_{k+1}} \ge \tfrac{1}{2}, \quad q_{\min} \le q_{k+1} \le \min\{q_{\max}, q_k + \mathbb{1}_{\gamma_k\ell_k\ge 1}\}$$

2.1.2: $\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\tfrac{1+q_k\rho_k}{q_{k+1}}},\ \sqrt{\tfrac{r_{k+1}}{q_{k+1}}\tfrac{\xi_k}{\left[\gamma_k^2L_k^2 + 2\gamma_k\ell_k(r_{k+1}-1) - (2r_{k+1}-1)\right]_+}}\right\}$

2.1.3: Set $\rho_{k+1} = \frac{\gamma_{k+1}}{\gamma_k}$ and update $x^{k+1} = \operatorname{prox}_{\gamma_{k+1} g}(x^k - \gamma_{k+1}\nabla f(x^k))$

Remark 2.2 (relation to AdaPG^{q,r}). Whenever $q_k \equiv q_{\min} = q_{\max} =: q$ and $\xi_k \equiv \xi_{\min} =: \xi$, the conditions in Algorithm 2.1 reduce to $q > \frac{1}{2}$, $r = \frac{q}{\xi+1} \geq \frac{1}{2}$, and $\xi = \frac{q}{r} - 1 > 0$; equivalently, $q > r \geq \frac{1}{2}$ as in Theorem 1.1.

In what follows, for $x \in \operatorname{dom}\varphi$ we adopt the notation

$$P_k(x) \coloneqq \varphi(x^k) - \varphi(x). \tag{2.3}$$

Our convergence analysis revolves around showing that under appropriate stepsize update and parameter selection the function

$$\mathcal{U}_k(x) \coloneqq \tfrac{1}{2}\|x^k - x\|^2 + \gamma_k(1 + q_k\rho_k)P_{k-1}(x) + \tfrac{\xi_k}{2}\|x^k - x^{k-1}\|^2, \tag{2.4}$$

monotonically decreases along the iterates for all $x \in \arg\min\varphi$. The main inequality, outlined in Theorem 2.3, extends the one in (Latafat et al., 2023b, Eq. (2.8)) by blending it with Lemma 2.1. This combination is achieved by adding and subtracting a multiple of the residual scaled by a newly added parameter ξ_k. The proof is otherwise adapted from that of (Latafat et al., 2023b, Lem. 2.2), and is included in full detail in the preprint version.

Theorem 2.3 (main PG inequality) *Consider a sequence $(x^k)_{k\in\mathbb{N}}$ generated by PG iterations (2.1) under Assumption 2.1, and denote $\rho_{k+1} \coloneqq \frac{\gamma_{k+1}}{\gamma_k}$. Then, for any $x \in \operatorname{dom}\varphi$, $q_k, \xi_k \geq 0$ and $\nu_k > 0$, $k \in \mathbb{N}$,*

$$\begin{aligned}\mathcal{U}_{k+1}(x) \leq \mathcal{U}_k(x) - \gamma_k(1 + q_k\rho_k - q_{k+1}\rho_{k+1}^2)P_{k-1}(x) - \tfrac{1}{2}\|x^k - x^{k-1}\|^2\Big\{1 + \xi_k - \tfrac{1}{\nu_k} \\ - \rho_{k+1}^2(\nu_{k+1} + \xi_{k+1})\Big[\gamma_k^2 L_k^2 + 2\gamma_k\ell_k\big(\tfrac{q_{k+1}}{\nu_{k+1}+\xi_{k+1}} - 1\big) - \big(2\tfrac{q_{k+1}}{\nu_{k+1}+\xi_{k+1}} - 1\big)\Big]\Big\},\end{aligned} \tag{2.5}$$

where $\mathcal{U}_k(x)$ is as in (2.4). In particular, with $\nu_k \equiv 1$, if $\varphi(x) \leq \inf_{k\in\mathbb{N}}\varphi(x^k)$ (for instance, if $x \in \arg\min\varphi$), and

$$0 < \rho_{k+1}^2 \leq \min\left\{\frac{1+q_k\rho_k}{q_{k+1}}, \frac{\xi_k}{(1+\xi_{k+1})\Big[\gamma_k^2L_k^2 + 2\gamma_k\ell_k\Big(\frac{q_{k+1}}{1+\xi_{k+1}} - 1\Big) + \Big(1 - 2\frac{q_{k+1}}{1+\xi_{k+1}}\Big)\Big]_+}\right\} \tag{2.6}$$

(with $\xi_k > 0$) holds for every k, then the coefficients of $P_{k-1}(x)$ and $\|x^k - x^{k-1}\|^2$ in (2.5) are negative, $\mathcal{U}_{k+1}(x) \leq \mathcal{U}_k(x)$ and thus $(\mathcal{U}_k(x))_{k\in\mathbb{N}}$ converges and $(x^k)_{k\in\mathbb{N}}$ is bounded.

Consistently with what was first observed in Malitsky and Mishchenko (2020), inequality (2.6) confirms that stepsizes should both not grow too fast and be controlled by the local curvature of f. We next show that, under a technical condition on q_k, all that remains to do is ensuring that the stepsizes do not vanish, which is precisely the reason behind the restrictions on the parameters q_k and ξ_k prescribed in Algorithm 2.1, as Theorem 2.7 will ultimately demonstrate. The technical condition turns out to be a controlled growth of q_k, needed to guarantee that a sequence $\varrho_{k+1} \approx \sqrt{\frac{1+q_k\varrho_k}{q_{k+1}}}$ will eventually stay above 1.

Theorem 2.4 (convergence of PG with nonvanishing stepsizes) *Consider the iterates generated by (2.1) under Assumption 2.1, with $\gamma_{k+1} = \gamma_k \rho_{k+1}$ complying with (2.6). If $q_{k+1} \leq 1 + q_k$ holds for every k and $\inf_{k\in\mathbb{N}} \gamma_k > 0$, then:*

(i) The (bounded) sequence $(x^k)_{k\in\mathbb{N}}$ has exactly one optimal accumulation point.

(ii) If, in addition, $(q_k)_{k\in\mathbb{N}}$ and $(\xi_k)_{k\in\mathbb{N}}$ are chosen bounded and bounded away from zero, then the entire sequence $(x^k)_{k\in\mathbb{N}}$ converges to a solution $x^\star \in \arg\min \varphi$, and $\mathcal{U}_k(x^\star) \searrow 0$.

We now turn to the last piece of the puzzle, namely enforcing a strictly positive lower bound on the stepsizes. The following elementary lemma provides the key insight to achieve this.

Lemma 2.5 *Let f be convex and differentiable, and consider the iterates generated by Algorithm 2.1. Then, for every $k \in \mathbb{N}$ such that $\gamma_k \ell_k < 1$ it holds that*

$$\gamma_{k+1} \geq \min\left\{\gamma_k \sqrt{\tfrac{1}{q_{\max}} + \rho_k},\ \sqrt{\tfrac{\xi_{\min} r_{\min}}{q_{\max}}}\,\tfrac{1}{L_k}\right\}. \tag{2.7}$$

Proof We start by observing that the assumptions on f guarantee that $0 \leq \ell_k \leq L_k$. The (squared) second term in the minimum of step 2.1.2 can be lower bounded as follows

$$\begin{aligned}\tfrac{r_{k+1}}{q_{k+1}} \tfrac{\xi_k}{\left[\gamma_k^2 L_k^2 + 2\gamma_k \ell_k (r_{k+1}-1) + (1-2r_{k+1})\right]_+} &\geq \tfrac{\xi_{\min}}{q_{\max}} \tfrac{r_{\min}}{\left[\gamma_k^2 L_k^2 + 2\gamma_k \ell_k (r_{\min}-1) + (1-2r_{\min})\right]_+} \\ &= \tfrac{\xi_{\min}}{q_{\max}} \tfrac{r_{\min}}{\left[\gamma_k^2 L_k^2 - \gamma_k \ell_k + (\gamma_k \ell_k - 1)(2r_{\min}-1)\right]_+} \geq \tfrac{\xi_{\min} r_{\min}}{q_{\max}\gamma_k^2 L_k^2},\end{aligned} \tag{2.8}$$

where the first inequality follows from the fact that the left-hand side is increasing with respect to r_{k+1}, and the second inequality follows since $r_{\min} \geq 1/2$. In turn, the claimed inequality (2.7) follows from the fact that $q_{k+1} \leq q_k \leq q_{\max}$ whenever $\gamma_k \ell_k < 1$, see step 2.1.1. ■

This lemma already hints at a potential lower bound for the stepsize, since boundedness of the sequence $(x^k)_{k\in\mathbb{N}}$ ensures lower boundedness of the second term in the minimum. As for the first term, as long as q_k is upper bounded, the stepsize can only decrease for a controlled number of iterations. This arguments will be formally completed in the proof of Theorem 2.7, where the following notation will be instrumental.

Definition 2.6 *Let $\varepsilon > 0$. With $\varrho_1 = \sqrt{\varepsilon}$ and $\varrho_{t+1} = \sqrt{\varepsilon + \varrho_t}$ for $t \geq 1$, we denote*

$$t_\varepsilon \coloneqq \max\{t \in \mathbb{N} \mid \varrho_1, \ldots, \varrho_t < 1\} \quad \textit{and} \quad \mathrm{m}(\varepsilon) \coloneqq \textstyle\prod_{t=1}^{t_\varepsilon} \varrho_t.$$

Notice that $\mathrm{m}(\varepsilon) \leq 1$ and equality holds iff $\varepsilon \geq 1$ (equivalently, iff $t_\varepsilon = 0$). For $\varepsilon \in (0,1)$, t_ε is a well-defined strictly positive integer, owing to the monotonic increase of ϱ_t and its convergence to the positive root of the equation $\varrho^2 - \varrho - \varepsilon = 0$. In particular, $\mathrm{m}(\varepsilon) \leq \varrho_1 = \sqrt{\varepsilon}$ and identity holds if $t_\varepsilon = 1$, that is, $\sqrt{\varepsilon + \sqrt{\varepsilon}} \geq 1$ (and $\varepsilon < 1$). This leads to a partially explicit expression

$$\begin{cases} t_\varepsilon = 1 & \text{and } \ \mathrm{m}(\varepsilon) = \sqrt{\min\{1, \varepsilon\}} & \text{if } \varepsilon \geq \tfrac{3-\sqrt{5}}{2} \approx 0.382 \\ 1 < t_\varepsilon \leq \left\lceil \tfrac{1}{\varepsilon(2-\varepsilon)} \right\rceil & \text{and } \ \sqrt{\varepsilon^{t_\varepsilon}} < \mathrm{m}(\varepsilon) < \sqrt{\varepsilon} & \text{otherwise.} \end{cases} \tag{2.9}$$

The bound on t_ε in the second case is obtained by observing that

$$1 > \varrho_{t_\varepsilon} = \varrho_{t_\varepsilon} - \varrho_{t_\varepsilon}^2 + \varepsilon + \varrho_{t_\varepsilon - 1} = \cdots = \sum_{t=1}^{t_\varepsilon} (\varrho_t - \varrho_t^2) + (t_\varepsilon - 1)\varepsilon \geq (t_\varepsilon - 1)\varepsilon(1-\varepsilon) + (t_\varepsilon - 1)\varepsilon,$$

where we used the fact that $\varepsilon \leq \varrho_t \leq 1 - \varepsilon$ and thus $\varrho_t - \varrho_t^2 \geq \varepsilon(1-\varepsilon)$ for $t = 1, \ldots, t_\varepsilon - 1$. We also remark that the lower bound in the simplified setting of Theorem 1.1 pertains to the case when $t_\varepsilon = 1$, since $\varepsilon = \frac{1}{q}$ falls under the first case above.

Theorem 2.7 (convergence of Algorithm 2.1) *Under Assumption 2.1, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 2.1 converges to a solution $x^\star \in \arg\min \varphi$ and $(\mathcal{U}_k(x^\star))_{k\in\mathbb{N}} \searrow 0$. Moreover, there exists $k_0 \leq 2\lceil \log_{1+\frac{1}{q_{\max}}}\big(\frac{1}{\gamma_0 L_{f,\mathcal{V}}}\big)\rceil_+$ such that*

$$\gamma_k \geq \gamma_{\min} := \mathrm{m}(1/q_{\max})\sqrt{\tfrac{\xi_{\min} r_{\min}}{q_{\max}}}\tfrac{1}{L_{f,\mathcal{V}}} \quad \forall k \geq k_0,$$

where $r_{\min} := \inf_{k\in\mathbb{N}} r_k \geq \frac{1}{2}$, $\mathrm{m}(\,\cdot\,)$ is as in Definition 2.6 (see also (2.9)), and $L_{f,\mathcal{V}}$ is a Lipschitz modulus for ∇f on a compact convex set $\mathcal{V}$ that contains $(x^k)_{k\in\mathbb{N}}$.

Proof The conditions prescribed in step 2.1.1 entail that the requirements of Theorem 2.4*(ii)* are met, so that the proof reduces to showing the claimed lower bound on $(\gamma_k)_{k\in\mathbb{N}}$. Boundedness of the sequence $(x^k)_{k\in\mathbb{N}}$ established in Theorem 2.3 ensures the existence of $L_{f,\mathcal{V}} > 0$ as in the statement. In particular, recall that $\ell_k \leq L_k \leq L_{f,\mathcal{V}}$ holds for all $k \in \mathbb{N}$, cf. (1.2). Lemma 2.5 then yields that

$$\gamma_k \ell_k < 1 \quad \Rightarrow \quad \gamma_{k+1} \geq \min\left\{\gamma_k\sqrt{\tfrac{1}{q_{\max}} + \rho_k},\ \sqrt{\tfrac{\xi_{\min} r_{\min}}{q_{\max}}}\tfrac{1}{L_{f,\mathcal{V}}}\right\}. \tag{2.10}$$

We first show that $\gamma_{k_0} L_{f,\mathcal{V}} \geq \sqrt{\frac{\xi_{\min} r_{\min}}{q_{\max}}}$ holds for some $k_0 \geq 0$ upper bounded as in the statement. To this end, suppose that $\gamma_k L_{f,\mathcal{V}} < \sqrt{\frac{\xi_{\min} r_{\min}}{q_{\max}}}$ for $k = 0, 1, \ldots, K$. The bounds $\xi_k \geq \xi_{\min}$ and $q_k \leq q_{\max}$ enforced in step 2.1.1 imply that $\frac{1}{2} \leq r_{\min} \leq r_k \leq \frac{q_{\max}}{1+\xi_{\min}}$ for any k. In particular, $\frac{\xi_{\min} r_{\min}}{q_{\max}} \leq \frac{\xi_{\min} r_k}{q_{\max}} \leq \frac{\xi_{\min}}{1+\xi_{\min}} < 1$ holds for every k. Then, $\gamma_k \ell_k \leq \gamma_k L_{f,\mathcal{V}} < \sqrt{\frac{\xi_{\min} r_{\min}}{q_{\max}}} < 1$, and (2.10) hold true for all such k, leading to $\gamma_{k+1} \geq \gamma_k \sqrt{1/q_{\max} + \rho_k}$ for $k = 0, \ldots, K-1$. Since $\rho_0 \geq 1$, it follows that $\rho_{k+1} = \gamma_{k+1}/\gamma_k \geq \sqrt{1/q_{\max} + 1}$ for $k = 0, \ldots, K-1$. Thus,

$$1 > \tfrac{a r_{\min}}{q_{\max}} > (\gamma_K L_{f,\mathcal{V}})^2 \geq \big(1 + \tfrac{1}{q_{\max}}\big)(\gamma_{K-1} L_{f,\mathcal{V}})^2 \geq \cdots \geq \big(1 + \tfrac{1}{q_{\max}}\big)^K (\gamma_0 L_{f,\mathcal{V}})^2,$$

from which the existence of k_0 bounded as in the statement follows.

Let $k \geq k_0$ be an index such that $\gamma_k L_{f,\mathcal{V}} \geq \sqrt{\frac{\xi_{\min} r_{\min}}{q_{\max}}}$, and suppose that $\gamma_{k+t} L_{f,\mathcal{V}} < \sqrt{\frac{\xi_{\min} r_{\min}}{q_{\max}}}$ for $t = 1, \ldots, T$. As before, the inequalities in (2.10) hold true for all such iterates, leading to

$$\rho_{k+t} \geq \sqrt{\tfrac{1}{q_{\max}} + \rho_{k+t-1}}, \quad t = 1, \ldots, T+1, \quad \text{and in particular} \quad \rho_{k+1} \geq \sqrt{\tfrac{1}{q_{\max}}}.$$

It then follows from the definition of $\mathrm{m}(\varepsilon)$ and t_ε as in Definition 2.6 with $\varepsilon = \frac{1}{q_{\max}}$ that $\gamma_{k+t} = \gamma_{k+t-1}\rho_{k+t}$ can only decrease for at most $t \leq t_\varepsilon$ iterations (that is, $T \leq t_\varepsilon$), at the end of which

$$\gamma_{k+t} = \big(\textstyle\prod_{i=1}^t \rho_{k+i}\big)\gamma_k \geq \mathrm{m}(\tfrac{1}{q_{\max}})\gamma_k \geq \mathrm{m}(\tfrac{1}{q_{\max}})\sqrt{\tfrac{\xi_{\min} r_{\min}}{q_{\max}}}\tfrac{1}{L_{f,\mathcal{V}}} \overset{\text{(def)}}{=} \gamma_{\min},$$

and then increases linearly up to when it is again larger than $\sqrt{\frac{\xi_{\min} r_{\min}}{q_{\max}}}\frac{1}{L_{f,\mathcal{V}}}$, proving that $\gamma_k \geq \gamma_{\min}$ holds for all $k \geq k_0$. ■

3. A class of adaptive alternating minimization algorithms

Leveraging an interpretation of AMA as the dual of the proximal gradient method, an adaptive variant is developed for solving (CP) under the following assumptions.

Assumption 3.1 (requirements for problem (CP))

*A1** $\psi_1 : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper, closed, locally strongly convex, and 1-coercive;*

*A2** $\psi_2 : \mathbb{R}^m \to \overline{\mathbb{R}}$ *is proper, convex and closed;*

*A3** $A \in \mathbb{R}^{m\times n}$ *and there exists* $x \in \operatorname{relint}\operatorname{dom}\psi_1$ *such that* $Ax \in \operatorname{relint}\psi_2$.

Under these requiremets, problem (CP) admits a unique solution $x^\star$, and by virtue of (Rockafellar, 1970, Thm.s 23.8, 23.9, and Cor. 31.2.1) also its dual

$$\operatorname*{minimize}_{y\in\mathbb{R}^m} \psi_1^*(-A^\top y) + \psi_2^*(y) \tag{D}$$

has solutions $y^\star$ characterized by $y^\star \in \partial\psi_2(Ax^\star)$ and $-A^\top y^\star \in \partial\psi_1(x^\star)$, and strong duality holds. In fact, Assumption 3.1.A1* ensures that the conjugate ψ_1^* is a (real-valued) locally Lipschitz differentiable function (Goebel and Rockafellar, 2008, Thm. 4.1). We note that the weaker notion of local strong monotonicity of $\partial\psi_1$ relative to its graph would suffice, and that this minor departure from the reference is used for simplicity of exposition. Problem (D) can then be addressed with proximal gradient iterations $y^+ = \operatorname{prox}_{\gamma g}(y - \gamma\nabla f(y))$ analized in the previous section, with $f \coloneqq \psi_1^*(-A^\top y)$ and $g \coloneqq \psi_2^*$. In terms of primal variables x and z, these iterations result in the alternating minimization algorithm. We here reproduce the simple textbook steps. First, observe that

$$x = \nabla\psi_1^*(-A^\top y) \quad\Leftrightarrow\quad -A^\top y \in \partial\psi_1(x) \quad\Leftrightarrow\quad 0 \in A^\top y + \partial\psi_1(x) = \partial[\langle A\,\cdot\,, y\rangle + \psi_1](x).$$

Hence, by strict convexity, $x = \arg\min\{\psi_1 + \langle A\,\cdot\,, y\rangle\}$. By the Moreau decomposition,

$$\operatorname{prox}_{\gamma g}(y - \gamma\nabla f(y)) = \operatorname{prox}_{\gamma\psi_2^*}(y + \gamma Ax) = y + \gamma Ax - \gamma\operatorname{prox}_{\psi_2/\gamma}(\gamma^{-1}y + Ax).$$

AMA iterations thus generate a sequence $(y^k)_{k\in\mathbb{N}}$ given in (3.1), where

$$\mathscr{L}_\gamma(x,z,y) \coloneqq \psi_1(x) + \psi_2(z) + \langle y, Ax - z\rangle + \tfrac{\gamma}{2}\|Ax - z\|^2$$

is the γ-augmented Lagrangian associated to (CP).

AdaAMA^{q,r} Fix $y^{-1} \in \mathbb{R}^m$ and $\gamma_0 = \gamma_{-1} > 0$. With ℓ_k and L_k as in (3.2), starting from

$$\begin{cases} x^{-1} = \arg\min_{x\in\mathbb{R}^n}\{\psi_1(x) + \langle y^{-1}, Ax\rangle\} \\ z^0 = \arg\min_{z\in\mathbb{R}^m} \mathscr{L}_{\gamma_0}(x^{-1}, z, y^{-1}) \\ y^0 = y^{-1} + \gamma_0(Ax^{-1} - z^0), \end{cases}$$

iterate for $k = 0, 1, \ldots$

$$x^k = \operatorname*{arg\,min}_{x\in\mathbb{R}^n}\{\psi_1(x) + \langle y^k, Ax\rangle\} \qquad \Big(= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \mathscr{L}_0(x, z^k, y^k)\Big) \tag{3.1a}$$

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\tfrac{1}{q} + \tfrac{\gamma_k}{\gamma_{k-1}}},\ \sqrt{\tfrac{1-\frac{r}{q}}{\left[(1-2r)+\gamma_k^2L_k^2+2\gamma_k\ell_k(r-1)\right]_+}}\right\} \tag{3.1b}$$

$$z^{k+1} = \operatorname{prox}_{\psi_2/\gamma_{k+1}}(\gamma_{k+1}^{-1}y^k + Ax^k) \qquad \Big(= \operatorname*{arg\,min}_{z\in\mathbb{R}^m} \mathscr{L}_{\gamma_{k+1}}(x^k, z, y^k)\Big) \tag{3.1c}$$

$$y^{k+1} = y^k + \gamma_{k+1}(Ax^k - z^{k+1}) \tag{3.1d}$$

The chosen iteration indexing reflects the dependency on the stepsize γ_k: x^k depends on y^k but *not* on γ_{k+1}, whereas z^{k+1} does depend on it. Moreover, this convention is consistent with the relation $\nabla f(y^k) = -Ax^k$. Local Lipschitz estimates of ∇f as in (1.1) are thus expressed as

$$\ell_k = -\frac{\langle Ax^k - Ax^{k-1}, y^k - y^{k-1}\rangle}{\|y^k - y^{k-1}\|^2} \quad \text{and} \quad L_k = \frac{\|Ax^k - Ax^{k-1}\|^2}{\|y^k - y^{k-1}\|^2}. \tag{3.2}$$

Being dually equivalent algorithms, convergence of AdaAMA^{q,r} is deduced from that of AdaPG^{q,r}.

Theorem 3.1 *Under Assumption 3.1, for any $q > r \geq \frac{1}{2}$ the sequence $(x^k)_{k\in\mathbb{N}}$ generated by AdaAMA^{q,r} converges to the (unique) primal solution of (CP), and $(y^k)_{k\in\mathbb{N}}$ to a solution of the dual problem (D).*

4. Numerical simulations

Performance of AdaPG^{q,r} with five different parameter choices from Table 1 is reported through a series of experiments on (i) logistic regression, (ii) cubic regularization for logistic loss, (iii) regularized least squares. The two former simulations use three standard datasets from the LIBSVM library Chang and Lin (2011), while for Lasso synthetic data is generated based on (Nesterov, 2013, §6); for further details the reader is referred to (Latafat et al., 2023b, §4.1) where the same problem setup is used. When applicable, the following algorithms are included in the comparisons.[1]

PG-ls^b	Proximal gradient method with nonmonotone backtracking
Nesterov	Nesterov's acceleration with constant stepsize $1/L_f$ (Beck, 2017, §10.7)
adaPG	(Latafat et al., 2023b, Alg. 2.1)
adaPG-MM	Proximal extension of (Malitsky and Mishchenko, 2020, Alg. 1)

The backtracking procedure in PG-ls^b is meant in the sense of (Beck, 2017, §10.4.2), (see also (Salzo, 2017, LS1) and (De Marchi and Themelis, 2022, Alg. 3) for the locally Lipschitz smooth case), without enforcing monotonic decrease on the stepsize sequence. To improve performance, the initial guess for γ_{k+1} is warm-started as $b\gamma_k$, where γ_k is the accepted value in the previous iteration and $b \geq 1$ is a backtracking factor. For each simulation we tested all values of $b \in \{1,\ 1.1,\ 1.3,\ 1.5,\ 2\}$ and only reported the best outcome.

5. Conclusions

This paper proposed a general framework for a class of adaptive proximal gradient methods, demonstrating its capacity to extend and tighten existing results when restricting to certain parameter choices. Moreover, application of the developed method was explored in the dual setting which led to a class of novel adaptive alternating minimization algorithms.

Future research directions include extensions to nonconvex problems, variational inequalities, and simple bilevel optimization expanding upon Malitsky (2020) and Latafat et al. (2023c). It would also be interesting to investigate the effectiveness of time-varying parameters in our framework for further improving performance and worst-case convergence rate guarantees.

1. https://github.com/pylat/adaptive-proximal-algorithms-extended-experiments

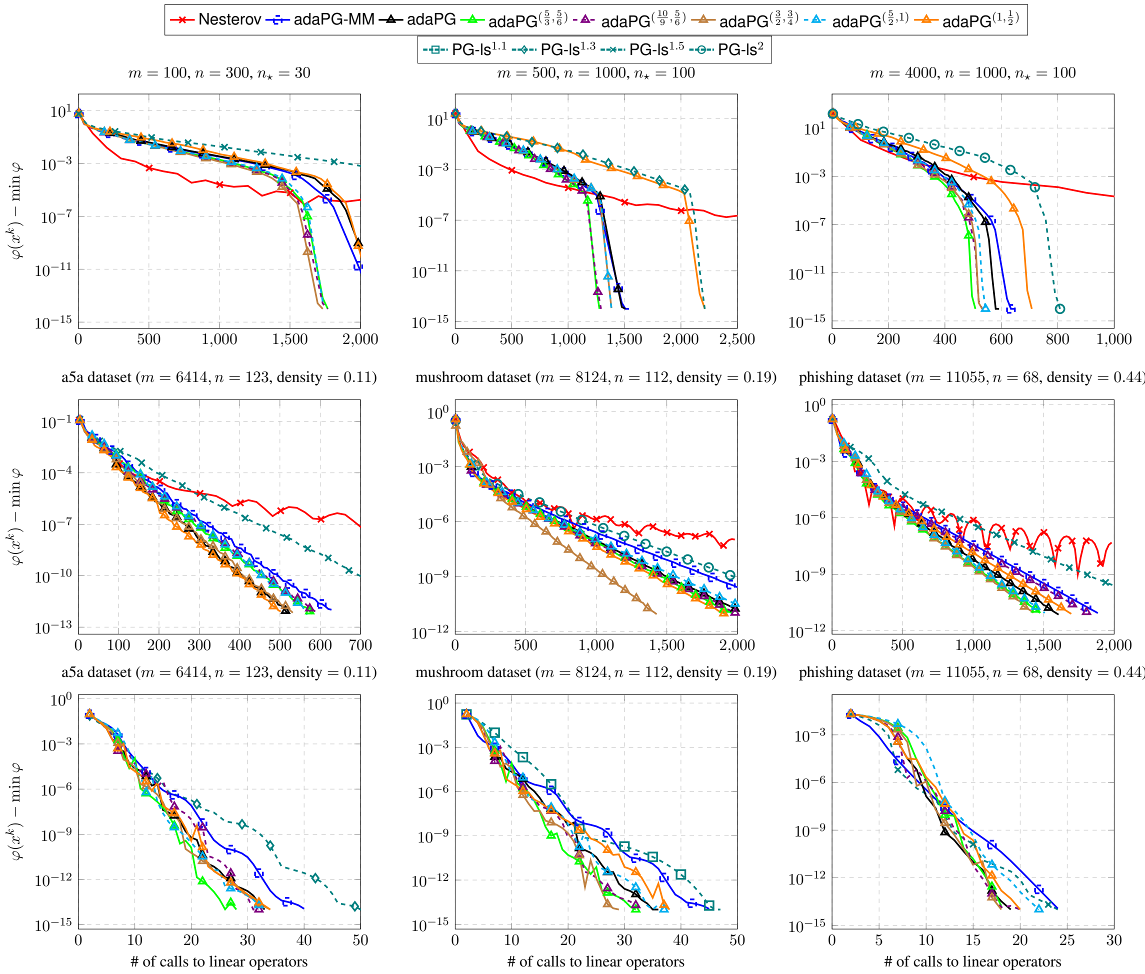

Figure 1: First row: regularized least squares, second row: ℓ_1-regularized logistic regression, third row: cubic regularization with Hessian generated for the logistic loss problem evaluated at zero. For the linesearch method $\mathsf{PG\text{-}ls}^b$, in each simulation only the best outcome for $b \in \{1,\ 1.1,\ 1.3,\ 1.5,\ 2\}$ is reported.

Acknowledgments

Work supported by: the Research Foundation Flanders (FWO) postdoctoral grant 12Y7622N and research projects G081222N, G033822N, and G0A0920N; 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.

References

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging II: Silver stepsize schedule for smooth convex optimization. *arXiv:2309.16530*, 2023.

Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics. Springer, 2017. ISBN 978-3-319-48310-8.

Amir Beck. *First-Order Methods in Optimization*. SIAM, Philadelphia, PA, 2017.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:1–27, 2011.

Alberto De Marchi and Andreas Themelis. Proximal gradient algorithms under local Lipschitz gradient continuity: A convergence and robustness analysis of PANOC. *Journal of Optimization Theory and Applications*, 194:771–794, 2022.

Aaron Defazio, Baoyu Zhou, and Lin Xiao. Grad-GradaGrad? A non-monotone adaptive stochastic gradient method. *arXiv:2206.06900*, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Alina Ene, Huy L Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7314–7321, 2021.

Daniel Gabay. Chapter IX Applications of the method of multipliers to variational inequalities. In Michel Fortin and Roland Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and Its Applications*, pages 299–331. Elsevier, 1983.

Rafal Goebel and R Tyrrell Rockafellar. Local strong convexity and local Lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263, 2008.

Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

Benjamin Grimmer, Kevin Shu, and Alex L Wang. Accelerated gradient descent via long steps. *arXiv:2309.09961*, 2023.

Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. *arXiv:2302.12022*, 2023.

Puya Latafat, Andreas Themelis, and Panagiotis Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. *arXiv:2311.18431*, 2023a.

Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *arXiv:2301.04431*, 2023b.

Puya Latafat, Andreas Themelis, Silvia Villa, and Panagiotis Patrinos. On the convergence of proximal gradient methods for convex simple bilevel optimization. *arXiv:2305.03559*, 2023c.

Tianjiao Li and Guanghui Lan. A simple uniformly optimal method without line search for convex optimization. *arXiv:2310.10082*, 2023.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6702–6712. PMLR, 13- 2020.

Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *arXiv:2308.02261*, 2023.

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, aug 2013.

Ralph T. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

Saverio Salzo. The variable metric forward-backward splitting algorithm under mild differentiability assumptions. *SIAM Journal on Optimization*, 27(4):2153–2181, 2017.

Paul Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR, 09- 2019.

Alp Yurtsever, Alex Gu, and Suvrit Sra. Three operator splitting with subgradients, stochastic gradients, and adaptive learning rates. *Advances in Neural Information Processing Systems*, 34: 19743–19756, 2021.