Adaptive Proximal Gradient Methods Are Universal Without Approximation

Oikonomidis, Konstantinos
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

Laude, Emanuel
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

Latafat, Puya
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

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

他

出版情報：Proceedings of Machine Learning Research. 235, pp.38663-38682, 2024-05-02.
Proceedings of Machine Learning Research(PMLR)
バージョン：

Adaptive Proximal Gradient Methods Are Universal Without Approximation

Konstantinos Oikonomidis [1] **Emanuel Laude** [1] **Puya Latafat** [1] **Andreas Themelis** [2] **Panagiotis Patrinos** [1]

Abstract

We show that adaptive proximal gradient methods for convex problems are not restricted to traditional Lipschitzian assumptions. Our analysis reveals that a class of linesearch-free methods is still convergent under mere local Hölder gradient continuity, covering in particular continuously differentiable semi-algebraic functions. To mitigate the lack of local Lipschitz continuity, popular approaches revolve around ε-oracles and/or linesearch procedures. In contrast, we exploit plain Hölder inequalities not entailing any approximation, all while retaining the linesearch-free nature of adaptive schemes. Furthermore, we prove full sequence convergence without prior knowledge of local Hölder constants nor of the order of Hölder continuity. Numerical experiments make comparisons with baseline methods on diverse tasks from machine learning covering both the locally and the globally Hölder setting.

1. Introduction

We consider composite minimization problems of the form

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and has *locally* Hölder continuous gradient of (possibly *unknown*) order $\nu \in (0, 1]$, and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, and convex with easy to compute proximal mapping.

The proximal gradient method is the de facto splitting technique for solving the composite problem (P). Under global Lipschitz continuity of ∇f, convergence results and complexity bounds are well established. Nevertheless, there exist many applications where such an assumption is not met. Among these are mixtures of maximum likelihood models (Grimmer, 2023), classification, robust regression (Yang & Lin, 2018; Forsythe, 1972), compressive sensing (Chartrand & Yin, 2008) and p-Laplacian problems on graphs (Hafiene et al., 2018), or the subproblems in the power augmented Lagrangian method (Luque, 1984; Oikonomidis et al., 2023; Laude & Patrinos, 2023).

Although often linesearch methods are still applicable under this setting, their additional backtracking procedures can be quite costly in practice. In response to this, this work investigates linesearch-free adaptive methods under mere *local Hölder continuity* of ∇f, covering in particular all continuously differentiable semi-algebraic functions.[1]

In the Hölder differentiable setting, a notable approach was introduced in the seminal works (Nesterov, 2015; Devolder et al., 2014) that rely on the notion of ε-oracles (Devolder et al., 2014, Def. 1). The main idea there is to approximate the Hölder smooth term with the squared Euclidean norm, resulting in an approximate descent lemma (Nesterov, 2015, Lem. 2) that can be leveraged for a linesearch procedure. More specifically, given $\eta \in (0, 1)$, some $\gamma_0 > 0$, and an accuracy threshold $\varepsilon > 0$, *Nesterov's universal primal gradient* (NUPG) method (Nesterov, 2015) consists of computing

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

where $\gamma_{k+1} = 2\gamma_k\eta^{m_k}$ and $m_k \in \mathbb{N}$ is the smallest such that

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \tfrac{1}{2\gamma_{k+1}}\|x^{k+1} - x^k\|^2 + \tfrac{1}{2}\varepsilon. \tag{1b}$$

It is clear that ε is a parameter of the algorithm, a fact which is further illustrated by the convergence rate

$$\varphi(x^k) - \varphi(x^\star) \le \tfrac{\varepsilon}{2} + \tfrac{1}{\varepsilon}^{\frac{1-\nu}{1+\nu}} \frac{2L_\nu^{\frac{2}{1+\nu}}\|x^0 - x^\star\|^2}{k+1}, \tag{1c}$$

where L_ν is a modulus of Hölder continuity of the gradient: the coefficient of the $1/(k+1)$ term becomes arbitrarily large as higher accuracy is demanded $\varepsilon \to 0$. This approach

[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 819-0395, Fukuoka, Japan. Correspondence to: Konstantinos Oikonomidis <konstantinos.oikonomidis@kuleuven.be>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]This is a consequence of the Łojasiewicz inequality: for a continuous semi-algebraic function $H : \mathbb{R}^n \to \mathbb{R}^n$, consider $f(x, y) = \|x - y\|$ and $g(x, y) = \|H(x) - H(y)\|$ in (Bochnak et al., 1998, Cor. 2.6.7).

nevertheless allows handling Hölder smooth problems in the same manner as Lipschitz smooth ones and thus implementing classical improvements such as acceleration (Nesterov, 2015; Kamzolov et al., 2021; Ghadimi et al., 2019). Moreover, its implementability has led to algorithms that go beyond the classical forward-backward splitting, such as primal-dual methods (Yurtsever et al., 2015; Nesterov et al., 2021) and even variational inequalities (Stonyakin et al., 2021).

In the context of classical majorization-minimization, when the order and the modulo of smoothness are known, the Hölder smoothness inequality itself has been used to generate descent without the aforementioned approximation. Akin to Lipschitz continuity, Hölder continuity of ∇f with constant H translates into a descent lemma inequality which, after addition of $g(x)$, yields the upper bound to the cost φ

$$\varphi(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + g(x) + \tfrac{1}{(1+\nu)\lambda_{k+1,\nu}} \|x - x^k\|^{1+\nu}, \tag{2}$$

for any $\lambda_{k+1,\nu} < 1/H$, cf. Fact 2.2.2. This was considered in (Bredies, 2008; Guan & Song, 2021) in a general Banach space setup as well as in (Yashtini, 2016; Bolte et al., 2023) for smooth but possibly nonconvex problems. With x^{k+1} denoting the minimizer of the above majorization model, the first-order optimality condition

$$0 \in \partial g(x^{k+1}) + \nabla f(x^k) + \tfrac{1}{\lambda_{k+1,\nu}} \|x^{k+1} - x^k\|^{-(1-\nu)}(x^{k+1} - x^k) \tag{3}$$

reveals that the resulting iterations are essentially an Euclidean proximal gradient method for a particular stepsize $\gamma_{k+1} := \lambda_{k+1,\nu}\|x^{k+1} - x^k\|^{1-\nu}$ that bears an implicit dependence on the future iterate x^{k+1}.

Such a majorize-minimize paradigm adopts $\lambda_{k+1,\nu}$ as an *explicit* stepsize parameter, and is thus tied to (the knowledge of) the order ν of Hölder differentiability. Instead, we directly derive conditions on the "Euclidean" stepsize γ_{k+1} and rather regard $\lambda_{k+1,\nu}$ as an *implicit* parameter, crucial to the convergence analysis yet absent in the algorithm. (In contrast to $\lambda_{k+1,\nu}$, the absence of the subscript ν in γ_{k+1} emphasizes the independence of the Hölder exponent on the Euclidean stepsize; this notational convention will be adopted throughout.)

We also remark that the implicit nature of the inclusion (3) is ubiquitous in algorithms that involve proximal terms of the form $\frac{1}{p}\|x - x^k\|^p$ for $p > 1$, such as the cubic Newton and tensor methods (Nesterov, 2021a; Doikov et al., 2024; Cartis et al., 2011) or the high-order proximal point algorithm (Luque, 1984; Nesterov, 2021b; Oikonomidis et al., 2023; Laude & Patrinos, 2023). Notice further that the Hölder proximal gradient update for non-Euclidean norms as described above differs from performing a "scaled" gradient step followed by a higher-order proximal point step. Instead, that corresponds to the anisotropic proximal gradient method (Laude & Patrinos, 2022) for choosing $\phi(x) = \frac{H}{1+\nu}\|x\|_{1+\nu}^{1+\nu}$.

Our contribution

Our approach departs from and improves upon existing works in the following aspects.

- Through a novel analysis of adaPG^{q,r} (Latafat et al., 2023a), we demonstrate that the class of linesearch-free adaptive methods advanced in (Malitsky & Mishchenko, 2020; 2023; Latafat et al., 2023a;b) are convergent even in the *locally* Hölder differentiable setting, covering in particular the class of semi-algebraic C^1 functions.
- Our approach bridges the gap between two fundamental approaches to minimizing Hölder-smooth functions: it is both *exact* as in (Bredies, 2008), in the sense that it does not involve nor depend on any predefined accuracy, and *universal* akin to the approach in (Nesterov, 2015), for it does not depend on (nor require the knowledge of) problem data such as the Hölder exponent ν.
- We establish sequential convergence (as opposed to subsequential or approximate cost convergence) with an exact rate

$$\min_{k \leq K} \varphi(x^k) - \min \varphi \leq O(\tfrac{1}{(K+1)^\nu}). \tag{4}$$

Differently from existing analyses that rely on a global lower bound on the stepsizes to infer convergence and an $O(1/(K+1))$ rate in the case $\nu = 1$, we identify a scaling of the stepsizes and a lower bound thereof that enables us to tackle the general $\nu \in (0, 1)$ regime.
- In numerical simulations we show that adaPG$^{q,\frac{q}{2}}$ performs well on a collection of locally and globally Hölder smooth problems, such as classification with Hölder-smooth SVMs and a p-norm version of Lasso. We show that our method performs consistenly better than Nesterov's universal primal gradient method (Nesterov, 2015) and in many cases better than its fast variant (Nesterov, 2015), as well as the recently proposed auto-conditioned fast gradient method (Li & Lan, 2023).

2. Universal, adaptive, without approximation

We consider standard (Euclidean) proximal gradient steps

$$x^{k+1} = \operatorname{prox}_{\gamma_{k+1} g}(x^k - \gamma_{k+1}\nabla f(x^k)) \tag{5}$$

for solving (P) under the following assumptions.

Assumption 2.1. The following hold in problem (P):

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

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

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

We build upon a series of adaptive algorithms, starting with a pioneering gradient method in (Malitsky & Mishchenko, 2020) and the follow-up studies (Latafat et al., 2023a;b; Malitsky & Mishchenko, 2023; Zhou et al., 2024) which contribute with proximal extensions, larger stepsizes, and tighter convergence rate estimates. While standard results of proximal algorithms guarantee a descent along the iterates in terms of the cost, distance to solutions, and fixed-point residual individually, the key idea behind this class of methods is to eliminate linesearch procedures by *implicitly* ensuring a descent on a (time-varying) combination of the three (see Lemma 3.3). This was achieved under local Lipschitz continuity of ∇f, by exploiting local Lipschitz estimates at consecutive iterates $x^{k-1}, x^k \in \mathbb{R}^n$ generated by the algorithm such as

$$\ell_k := \frac{\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\|x^k - x^{k-1}\|^2} \tag{6a}$$

and

$$L_k := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|} \tag{6b}$$

(with the convention $\frac{0}{0} = 0$).

Under the assumption considered in the aforementioned references of local Lipschitz continuity of ∇f, that is, with $\nu = 1$, the estimates ℓ_k and L_k as in (6) remain bounded whenever x^k and x^{k-1} range in a bounded set. Although this is no more the case in the setting investigated here, for ℓ_k and L_k may diverge as x^k and x^{k-1} get arbitrarily close, we show that these *Lipschitz* estimates can still be employed even if ∇f is merely locally *Hölder* continuous.

Specifically, we will consider the following stepsize update rule

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\frac{1}{q} + \frac{\gamma_k}{\gamma_{k-1}}}, \sqrt{\frac{1}{2[\gamma_k^2 L_k^2 - (2-q)\gamma_k \ell_k - (q-1)]_+}}\right\} \tag{7}$$

for some $q \in [1, 2]$, where $[\,\cdot\,]_+ := \max\{\,\cdot\,, 0\}$. This correponds to the update rule of the algorithm adaPG^{q,r} of (Latafat et al., 2023a) specialized to the choice $r = q/2$ for the second parameter. This restriction is nevertheless general enough to recover the update rules of (Malitsky & Mishchenko, 2020, Alg. 1), (Latafat et al., 2023b, Alg. 2.1), and (Malitsky & Mishchenko, 2023, Alg. 1) ($q = 1$), as well as the one of (Malitsky & Mishchenko, 2023, Alg. 3) ($q = 3/2$), see (Latafat et al., 2023b, Rem. 2.4). In particular, our analysis in the generality of $q \in [1, 2]$ demonstrates that *all these adaptive algorithms are convergent in the locally Hölder (convex) setting*.

A crucial challenge in the locally Hölder setting is the lack of a positive uniform lower bound for the stepsize sequence $(\gamma_k)_{k\in\mathbb{N}}$ generated by (7). To mitigate this, we factorize γ_k as

$$\gamma_k = \lambda_{k,\nu} \|x^k - x^{k-1}\|^{1-\nu}, \tag{8}$$

introducing scaled stepsizes $\lambda_{k,\nu}$ as suggested by the upper bound minimization procedure (3). This allows us to normalize the Hölder inequalities into Lipschitz-like ones, see Lemma 2.3. (Throughout, the subscript ν shall be used for quantities with dependence on ν for clarity of exposition.) Our analysis relies on showing that this scaled stepsize is lower bounded whenever the second term in (7) is active (see Lemma 3.6). We emphasize that this quantity, while crucial in our convergence analysis, does not appear in the algorithm, which only uses the estimates (6) and the update rule (7), neither of which depend on (the knowledge of) the local Hölder order ν.

2.1. Hölder continuity estimates

In this section, we set up some basic facts about Hölder continuity of ∇f that will be essential in our analysis. We again emphasize that our convergence analysis makes mere use of *existence* of $\nu \in (0, 1]$ (see Assumption 2.1.A1), but the algorithm is independent of (the knowledge of) this exponent, cf. adaPG$^{q,\frac{q}{2}}$ (Algorithm 1).

Local Hölder continuity of ∇f of order ν (which we shall refer to as local ν-Hölder continuity of ∇f for brevity) amounts to the existence, for every convex and bounded set $\Omega \subset \mathbb{R}^n$, of a constant $L_{\Omega,\nu} > 0$ such that

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

Note that both norms are the standard Euclidean 2-norms. We remark that the limiting case $\nu = 0$ amounts to *sub*-gradients of f being bounded on bounded sets, that is, to f being merely convex and real valued with no differentiability requirements. Although not covered by our convergence results in Section 3.2, the preliminary lemmas collected in Section 3.1 still remain valid for any real-valued convex f. To clearly emphasize this fact, whenever applicable we shall henceforth specify "(possibly with $\nu = 0$)" when invoking Assumption 2.1; in this case, the notation ∇f shall indicate any *sub*gradient map $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ with $\nabla f(x) \in \partial f(x)$, and

$$L_{\Omega,0} \leq 2 \operatorname{lip}_\Omega f \tag{9}$$

is bounded by twice the Lipschitz modulus for f on Ω (Rockafellar, 1970, Thm. 24.7).

Throughout, we will make use of the following inequalities, which reduce to well-known Lipschitz and cocoerciv-

ity properties of ∇f when $\nu = 1$. The proof of the second assertion can be found in (Yashtini, 2016, Lem. 1); our cocoercivity-like claims are a slight refinement of known global and/or scalar versions, see e.g., (Bauschke & Combettes, 2017, Cor. 18.14) and (Ying & Zhou, 2017, Prop. 1).

Fact 2.2 (Hölder-smoothness inequalities)**.** *Suppose that $h : \mathbb{R}^n \to \mathbb{R}$ is convex and ∇h is ν-Hölder continuous with modulus $L_{E,\nu} > 0$ on a convex set $E \subseteq \mathbb{R}^n$ for some $\nu \in [0,1]$. Then, for every $x, y \in E$ the following hold:*

1. $\langle x - y, \nabla h(x) - \nabla h(y)\rangle \leq L_{E,\nu}\|x - y\|^{1+\nu}$
2. $h(y) - h(x) - \langle \nabla h(x), y - x\rangle \leq \frac{L_{E,\nu}}{1+\nu}\|x - y\|^{1+\nu}$.

If $\nu \neq 0$ and ∇h is ν-Hölder continuous on an enlarged set $\overline{E} := E + \overline{\mathrm{B}}(0; \operatorname{diam} E)$ with modulus $L_{\overline{E},\nu}$, then the following local cocoercivity-type estimates also hold:

3. $\frac{2\nu}{1+\nu}\frac{1}{L_{\overline{E},\nu}^{1/\nu}}\|\nabla h(x) - \nabla h(y)\|^{\frac{1+\nu}{\nu}} \leq \langle x - y, \nabla h(x) - \nabla h(y)\rangle$
4. $\frac{\nu}{1+\nu}\frac{1}{L_{\overline{E},\nu}^{1/\nu}}\|\nabla h(x) - \nabla h(y)\|^{\frac{1+\nu}{\nu}} \leq h(y) - h(x) - \langle \nabla h(x), y - x\rangle$.

Based on the inequalities in Fact 2.2, given a sequence $(x^k)_{k\in\mathbb{N}}$ we define local estimates of Hölder continuity of ∇f with $\nu \in [0,1]$ as follows:

$$\ell_{k,\nu} := \frac{\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\|x^k - x^{k-1}\|^{1+\nu}} \tag{11a}$$

and

$$L_{k,\nu} := \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|^{\nu}}. \tag{11b}$$

Let us draw some comments on these quantities. Considering the scaled stepsize $\lambda_{k,\nu}$ given in (8), it is of immediate verification that

$$\lambda_{k,\nu}L_{k,\nu} = \gamma_k L_k, \quad \lambda_{k,\nu}\ell_{k,\nu} = \gamma_k \ell_k. \tag{12}$$

Moreover, observe that

$$\ell_{k,\nu} \leq L_{k,\nu} \leq L_{\Omega,\nu} \tag{13}$$

holds whenever $L_{\Omega,\nu}$ is a ν-Hölder modulus for ∇f on a compact convex set Ω that contains both x^{k-1} and x^k, the first inequality following from a simple application of Cauchy-Schwartz. We also remark that defining $\ell_{k,\nu}$ and $L_{k,\nu}$ as above in place of a cocoercivity-like estimate

$$c_{k,\nu} := \left(\frac{2\nu}{1+\nu}\frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|^{1+\frac{1}{\nu}}}{\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}\right)^{\nu}$$

causes no loss of generality, since each one among $c_{k,\nu}$, $\ell_{k,\nu}$ and $L_{k,\nu}$ can be derived based on the other two. The use of $\ell_{k,\nu}$ and $L_{k,\nu}$ provides nevertheless a simpler and more straightforward Hölder estimate, contrary to $c_{k,\nu}$ which instead involves counterintuitive powers and coefficients, as well as a potentially looser Hölder modulus, and fails to cover the limiting case $\nu = 0$, cf. Fact 2.2.3.

Let us denote the forward operator by

$$H_k := \mathrm{id} - \gamma_k \nabla f. \tag{14}$$

The subgradient characterization of the proximal gradient update (5) then reads

$$H_k(x^{k-1}) - H_k(x^k) \in \gamma_k \partial\varphi(x^k). \tag{15}$$

As in (Latafat et al., 2023b), the combined use of $\ell_{k,\nu}$ and $L_{k,\nu}$ yields a local Hölder modulus for the forward operator, though in this work it will be convenient to express it with respect to the scaled stepsize $\lambda_{k,\nu}$ as in (8).

Lemma 2.3. *Let $\ell_{k,\nu}$ and $L_{k,\nu}$ be as in (11) for some $x^{k-1}, x^k \in \mathbb{R}^n$, and let H_k be as in (14) for some $\gamma_k > 0$. Then, for any $\nu \in [0,1]$ and with $\lambda_{k,\nu}$ as in (8) it holds that*

$$\begin{aligned} M_k^2 := \frac{\|H_k(x^k) - H_k(x^{k-1})\|^2}{\|x^k - x^{k-1}\|^2} &= 1 + \lambda_{k,\nu}^2 L_{k,\nu}^2 - 2\lambda_{k,\nu}\ell_{k,\nu} \\ &= 1 + \gamma_k^2 L_k^2 - 2\gamma_k \ell_k. \end{aligned}$$

Being M_k a *Lipschitz* estimate, unless ∇f is locally Lipschitz continuous there is no guarantee that M_k is bounded for pairs x^{k-1}, x^k ranging in a compact set. The ν-Hölder estimate of the forward operator H_k, which instead *is* guaranteed to be bounded on bounded sets (for bounded $(\gamma_k)_{k\in\mathbb{N}}$), is given by

$$\begin{aligned} M_{k,\nu} := M_k\|x^k - x^{k-1}\|^{1-\nu} = \frac{\|H_k(x^k) - H_k(x^{k-1})\|}{\|x^k - x^{k-1}\|^{\nu}} \\ \leq (1 + \lambda_{k,\nu}L_{k,\nu})\|x^k - x^{k-1}\|^{1-\nu}, \end{aligned} \tag{16}$$

where the inequality follows from the triangle inequality and the definition of $L_{k,\nu}$, cf. (11b).

3. AdaPG$^{q,\frac{q}{2}}$ revisited

In this section adaPG$^{q,\frac{q}{2}}$ is presented in Algorithm 1 for solving composite problems (P). The main oracles of the algorithm are plain proximal and gradient evaluations. We refer to (Beck, 2017, §6) for examples of functions with easy to evaluate proximal maps.

AdaPG$^{q,\frac{q}{2}}$ incorporates the simple stepsize update rule (10) with a parameter $q \in [1,2]$ that strikes a balance between speed of recovery from small values (e.g., due to steep or ill-conditioned regions), and magnitude of the stepsize dictated by the second term. If $q = 1$, whenever $\gamma_k^2 L_k^2 \leq \gamma_k \ell_k$, the second term reduces to $1/0 = \infty$, and $\gamma_{k+1} = \gamma_k\sqrt{1 + \gamma_k/\gamma_{k-1}}$ strictly increases. On the other end of the spectrum, if $q = 2$ the update reduces to

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

Algorithm 1 AdaPG$^{q,\frac{q}{2}}$: A universal adaptive proximal gradient method

Require: starting point $x^{-1} \in \mathbb{R}^n$, stepsizes $\gamma_0 \geq \gamma_{-1} > 0$, parameter $q \in [1, 2]$

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

Repeat for $k = 0, 1, \ldots$ until convergence

1: With ℓ_k, L_k given in (6), define the stepsize as (with notation $[z]_+ = \max\{z, 0\}$ and convention $\frac{1}{0} = \infty$)

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{\frac{1}{q} + \frac{\gamma_k}{\gamma_{k-1}}}, \frac{1}{\sqrt{2[\gamma_k^2 L_k^2 - (2-q)\gamma_k \ell_k + 1 - q]_+}}\right\} \tag{10}$$

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

where having the first term active (for instance if $\gamma_k L_k \leq 1$) for two consecutive updates already ensures an increase in the stepsize owing to the simple observation that $1/2 + \sqrt{1/2} > 1$.

While adaPG$^{q,\frac{q}{2}}$ has the ability to recover from a potentially bad choice of initial stepsizes γ_0, γ_{-1}, the behavior of the algorithm during the first iterations can be impacted negatively. To eliminate such scenarios, γ_0 can be refined by running offline proximal gradient updates. Specifically, starting from the initial point x^{-1}, γ_0 can be updated as the inverse of either one of (11b) or (11a) evaluated between x^{-1} and the obtained point. If the updated stepsize is substantially smaller than the original one, the same procedure may be repeated an additional time. Once a reasonable γ_0 is obtained, we suggest selecting γ_{-1} small enough such that $\sqrt{\frac{1}{q} + \frac{\gamma_0}{\gamma_{-1}}} \geq \frac{1}{\sqrt{2L_0}}$, ensuring that γ_1 would be proportional to the inverse of L_0. We remark that the choice of γ_{-1} does not affect the sequential convergence results of Theorem 3.7. It does nevertheless affect the constant in our sublinear rate results of Theorem 3.8, through possibly having a larger $\rho_{\max}$ therein, although this effect is a mere theoretical technicality with negligible practical implications.

3.1. Preliminary lemmas

This subsection collects some preliminary results adaptated from (Malitsky & Mishchenko, 2020; 2023; Latafat et al., 2023a;b) that hold true under convexity assumption without further restrictions. In particular, the next lemma can be viewed as a counterpart of the well known *firm* nonexpansiveness (FNE) of $\operatorname{prox}_{\gamma g}$, which is recovered when $\gamma_{k+1} = \gamma_k$, and offers a refinement of the nonexpansiveness-like inequality in (Malitsky & Mishchenko, 2023, Lem. 12) that follows after an application of Cauchy-Schwarz.

Lemma 3.1 (FNE-like inequality)**.** *Let Assumption 2.1.A2 hold and f be differentiable. For any $\gamma_k > 0$ and denoting $\rho_{k+1} := \gamma_{k+1}/\gamma_k$, iterates (5) satisfy the following:*

$$\begin{aligned}\tfrac{1}{\rho_{k+1}}\|x^{k+1} - x^k\|^2 &\leq \langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle \\ &\leq \rho_{k+1}\|H_k(x^{k-1}) - H_k(x^k)\|^2.\end{aligned}$$

By combining this inequality with the identity in Lemma 2.3 we obtain the following.

Corollary 3.2. *Let Assumptions 2.1.A1 and 2.1.A2 hold (possibly with $\nu = 0$). For any $\gamma_k > 0$, and with M_k and $\lambda_{k,\nu}$ as in (14) and (8), iterates (5) satisfy*

$$\|x^{k+1} - x^k\| \leq \tfrac{\gamma_{k+1}}{\gamma_k} M_k \|x^{k-1} - x^k\|. \tag{17}$$

We next present the main descent inequality taken from (Latafat et al., 2023b, Lem. 2.2). Its proof is nevertheless included in the appendix to demonstrate its independence of ν: it relies on mere use of convexity inequalities and *identities* involving L_k, ℓ_k as in Lemmas 2.3 and 3.1 to express norms and inner products in terms of $\|x^k - x^{k-1}\|^2$. It reveals that for any $q \geq 0$ and $x^\star \in \arg\min \varphi$, up to a proper stepsize selection, the function

$$\begin{aligned}\mathcal{U}_k^q(x^\star) := \tfrac{1}{2}\|x^k - x^\star\|^2 + \tfrac{1}{2}\|x^k - x^{k-1}\|^2 \\ + \gamma_k(1 + q\rho_k)P_{k-1}\end{aligned} \tag{18}$$

monotonically decreases, where we introduced the symbol

$$P_k := \varphi(x^k) - \min \varphi \tag{19}$$

for the sake of conciseness.

Lemma 3.3 (main inequality)**.** *Let Assumption 2.1 hold (possibly with $\nu = 0$), and consider a sequence generated by proximal gradient iterations (5) with $\gamma_k > 0$ and $\rho_{k+1} := \gamma_{k+1}/\gamma_k$. Then, for any $q \geq 0$, $x^\star \in \arg\min \varphi$, and $k \in \mathbb{N}$ the following holds:*

$$\begin{aligned}\mathcal{U}_{k+1}^q(x^\star) \leq \mathcal{U}_k^q(x^\star) - \gamma_k\big(1 + q\rho_k - q\rho_{k+1}^2\big)P_{k-1} \\ - \big\{\tfrac{1}{2} - \rho_{k+1}^2\big[\gamma_k^2 L_k^2 - \gamma_k \ell_k(2-q) \\ + 1 - q\big]\big\}\|x^k - x^{k-1}\|^2.\end{aligned} \tag{20}$$

Apparently, the stepsize update rule of adaPG$^{q,\frac{q}{2}}$ is designed so as to make the coefficients of P_{k-1} and $\|x^k - x^{k-1}\|^2$ on the right-hand side of (20) negative, so that the corresponding proximal gradient iterates monotonically decrease the value of $\mathcal{U}_k^q(x^\star)$.

Lemma 3.4 (basic properties of adaPG$^{q,\frac{q}{2}}$)**.** *Under Assumption 2.1 (possibly with $\nu = 0$), the following hold for the iterates generated by adaPG$^{q,\frac{q}{2}}$:*

1. *$(\mathcal{U}_k^q(x^\star))_{k\in\mathbb{N}}$ as defined in (18) decreases and converges to a finite value.*
2. *The sequence $(x^k)_{k\in\mathbb{N}}$ is bounded and admits at most one optimal limit point.*
3. *$P_K^{\min} \leq \mathcal{U}_0^q(x^\star)/\left(\sum_{k=1}^{K+1}\gamma_k\right)$ for every $K \geq 1$, where $P_k^{\min} := \min_{0\leq i\leq k} P_i$.*

Remark 3.5. The validity of Lemma 3.4.3 also when $\nu = 0$ hints that having $\sum_{k\in\mathbb{N}}\gamma_k = \infty$ suffices to infer convergence results for the proximal *sub*gradient method without differentiability of f. Whether, or under which conditions, this is really the case is currently an open problem.

3.2. Convergence and rates

Distinguishing between the iterates in which the stepsize is updated according to the first or the second element in the minimum of (10) will play a fundamental role in our analysis. For this reason, it is convenient to introduce the following notation:

$$K_1 := \left\{k \in \mathbb{N} \mid \gamma_{k+1} = \gamma_k\sqrt{\tfrac{1}{q} + \tfrac{\gamma_k}{\gamma_{k-1}}}\right\} \tag{21a}$$

and

$$K_2 := \mathbb{N} \setminus K_1. \tag{21b}$$

Unlike its locally Lipschitz counterpart ($\nu = 1$), in the Hölder setting, a global lower bound for the stepsize sequence $(\gamma_k)_{k\in\mathbb{N}}$ cannot be expected. Nevertheless, a lower bound for the *scaled* stepsizes $\lambda_{k,\nu}$ whenever $k \in K_2$ is sufficient to ensure convergence.

Lemma 3.6. *Let Assumption 2.1 hold (possibly with $\nu = 0$), and consider the iterates generated by adaPG$^{q,\frac{q}{2}}$. Then, with K_2 as in (21b), for every $k \in K_2$*

$$\lambda_{k,\nu} \geq \frac{1}{\sqrt{2}L_{k,\nu}\rho_{\max}} \quad \text{and} \quad \rho_{k+1} \geq \frac{1}{\sqrt{2}\lambda_{k,\nu}L_{k,\nu}}, \tag{22}$$

where $\rho_{\max} := \max\left\{\frac{1}{2}\left(1+\sqrt{1+4/q}\right), \frac{\gamma_0}{\gamma_{-1}}\right\}$.

The anticipated lower bound on $(\lambda_{k,\nu})_{k\in K_2}$ will follow from (13), once boundedness of the sequence $(x^k)_{k\in\mathbb{N}}$ generated by adaPG$^{q,\frac{q}{2}}$ is established.

Theorem 3.7 (convergence)**.** *Under Assumption 2.1, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by adaPG$^{q,\frac{q}{2}}$ converges to some $x^\star \in \arg\min\varphi$.*

While a global lower bound for the stepsizes $(\gamma_k)_{k\in\mathbb{N}}$ is not available, it is at the moment unclear whether one for the (entire) scaled sequence $(\lambda_{k,\nu})_{k\in\mathbb{N}}$ exists. Nevertheless, with P_k as in (19), a lower bound for an alternative scaled sequence $(\gamma_{k+1}P_k^{-(1-\nu)/\nu})_{k\in\mathbb{N}}$ does exist, thanks to which the following convergence rate can be achieved.

Theorem 3.8 (sublinear rate)**.** *Suppose that Assumption 2.1 holds. Then, the following sublinear rate holds for the iterates generated by adaPG$^{q,\frac{q}{2}}$:*

$$\min_{0\leq i\leq K} P_i \leq \max\left\{\frac{\mathcal{U}_0^q(x^\star)}{\gamma_0(K+1)}, \frac{C(q,\nu)\mathcal{U}_0^q(x^\star)^{\frac{1+\nu}{2}}L_{\Omega,\nu}}{(K+1)^\nu}\right\}$$

where $C(q,\nu) = \sqrt{2}\left(\sqrt{q}\right)^\nu\left(\sqrt{2}\rho_{\max}+1\right)^{1-\nu}$ and $L_{\Omega,\nu}$ is a ν-Hölder modulus for ∇f on a compact convex set Ω that contains all the iterates x^k.

Some remarks are in order regarding the convergence results of the method. First, the obtained rate matches the one of the standard convex Hölder smooth setting of (Bredies, 2008) and the one in the nonconvex case (Bolte et al., 2023, Prop. 9), while it is worse than the one of the Universal Primal Gradient Method in (Nesterov, 2015). Since our analysis relies upon a more involved Lyapunov function along with an adaptive stepsize rule, whether or not it can be tightened in order to obtain a better rate remains an interesting open question.

In the locally Lipschitz setting $\nu = 1$, the above rate matches the $O(1/(K+1))$ of (Latafat et al., 2023a, Thm. 1.1) (with $r = q/2$). Despite such a worst-case sublinear rate, the fast behavior of the algorithm in practice can be explained by utilization of large stepsizes and the bound in Lemma 3.4.3.

We also remark that under (local) strong convexity, up to modifying the stepsize update similarly to (Malitsky & Mishchenko, 2020, §2.3), a contraction can be established in terms of $\mathcal{U}_k^q(x^\star)$ in Lemma 3.3. We refer the reader to (Malitsky & Mishchenko, 2020) for this approach and postpone a more detailed analysis to a future work.

4. Experiments

We demonstrate the performance of adaPG$^{q,\frac{q}{2}}$ on a range of simulations for three different choices of q. We compare against the baseline and state-of-the-art methods: Nesterov's universal primal gradient method (NUPG) (Nesterov, 2015), its fast variant (F-NUPG) (Nesterov, 2015), as well as the recently proposed auto-conditioned fast gradient method (AC-FGM) (Li & Lan, 2023).

While adaPG$^{q,\frac{q}{2}}$ only involves gradient evaluations, the other methods additionally require evaluations of the objective. In all applications considered in this section, the

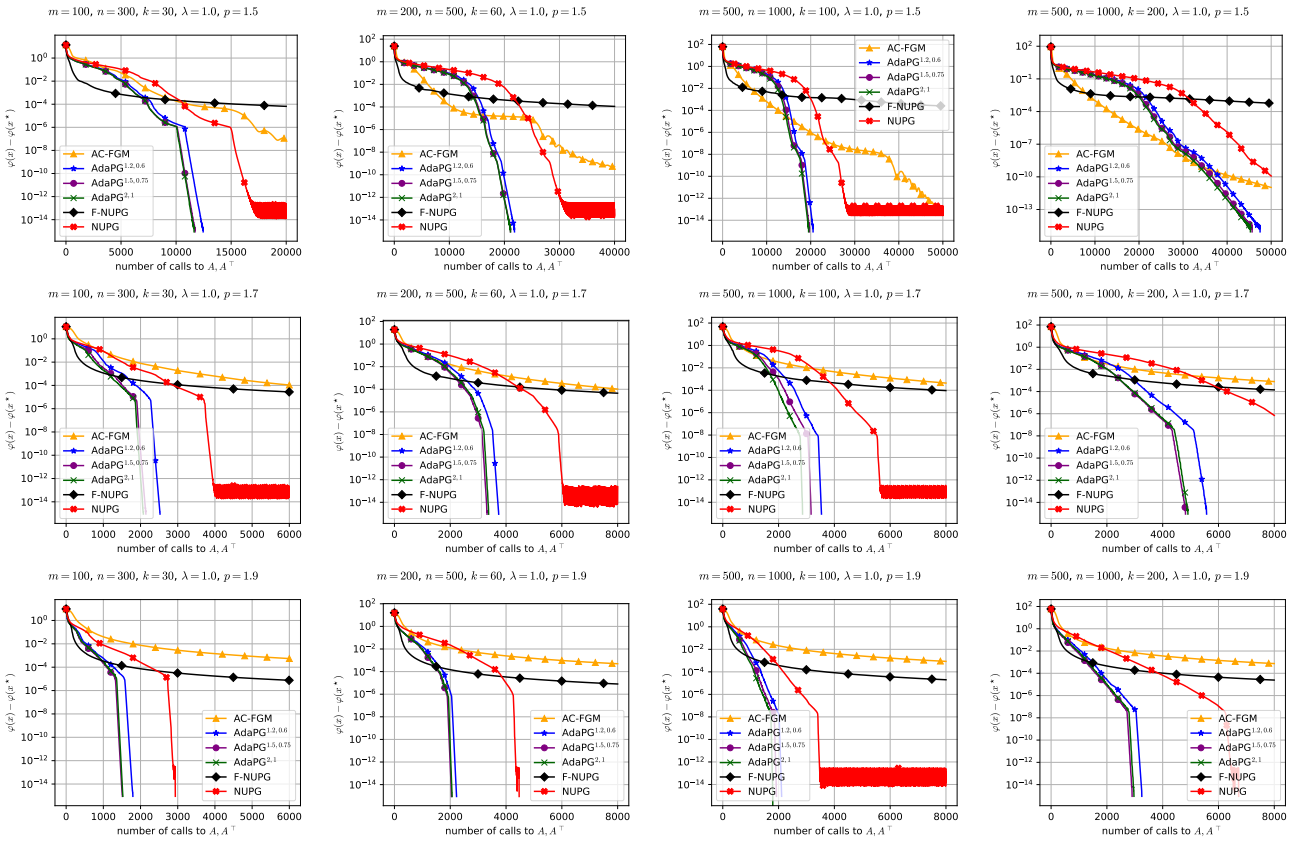

Figure 1. p-norm Lasso with varying powers p. It can be seen that adaPG$^{q,\frac{q}{2}}$ performs better than NUPG in all cases in terms of calls to $A, A^\top$. In this experiment adaPG$^{q,\frac{q}{2}}$ also performs consistently better than the accelerated algorithms AC-FGM and F-NUPG.

smooth part $f(x)$ takes the form $\psi(Ax)$ where A is the design matrix containing the data. Consequently, matrix vector products $y = Ax$ and $A^\top \nabla\psi(y)$, each of complexity $\mathcal{O}(mn)$, constitute the most costly operations. For the sake of a fair comparison, we store vectors that can be reused in subsequent evaluations, and plot the progress of the algorithm in terms of calls to A and $A^\top$. As a result, denoting with $\#_A$ the number of calls to A and $A^\top$, and with $\#_{\text{LS}} \geq 1$ the number of linesearch trials (including the successful ones), each iteration involves:

- NUPG: 1 call to ∇f and $\#_{\text{LS}}$ to f; by exploiting linearity, $\#_A = 1 + \#_{\text{LS}}$
- F-NUPG: $\#_{\text{LS}}$ calls to ∇f and $2 \times \#_{\text{LS}}$ to f; by exploiting linearity, $\#_A = 3 + \#_{\text{LS}}$
- adaPG$^{q,\frac{q}{2}}$: 1 call to ∇f; using linearity, $\#_A = 2$
- AC-FGM: 1 call to ∇f and 1 to f; by linearity, $\#_A = 2$.

Implementation details In all experiments the solvers use $x = 0$ as the initial point, and are executed with the same initial stepsize computed as follows. First, we ran an offline proximal gradient update starting from the initial point and computing the stepsize as the inverse of (11b) evaluated between the initial and the obtained point. This procedure was repeated one additional time in case the obtained stepsize was substantially smaller than the original one. In the case of AC-FGM, in addition, we used the procedure of (Li & Lan, 2023) which can also be found in Appendix D. The implementation for reproducing the experiments is publicly available.[2]

4.1. p-norm Lasso

In this experiment we consider a variant of Lasso in which the squared 2-norm is replaced with a p-norm raised to some power $p \in (1, 2)$:

$$\min_{x \in \mathbb{R}^n} \tfrac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1, \tag{23}$$

for $A \in [-1, 1]^{m \times n}$, $b \in \mathbb{R}^m$ with $n > m$ and $\lambda > 0$. The first term in (23) is differentiable with globally Hölder continuous gradient with order $\nu = p - 1$. The proximal mapping of the second term is the shrinkage operation which is computable in closed form. To assess the performance

[2] https://github.com/EmanuelLaude/universal-adaptive-proximal-gradient

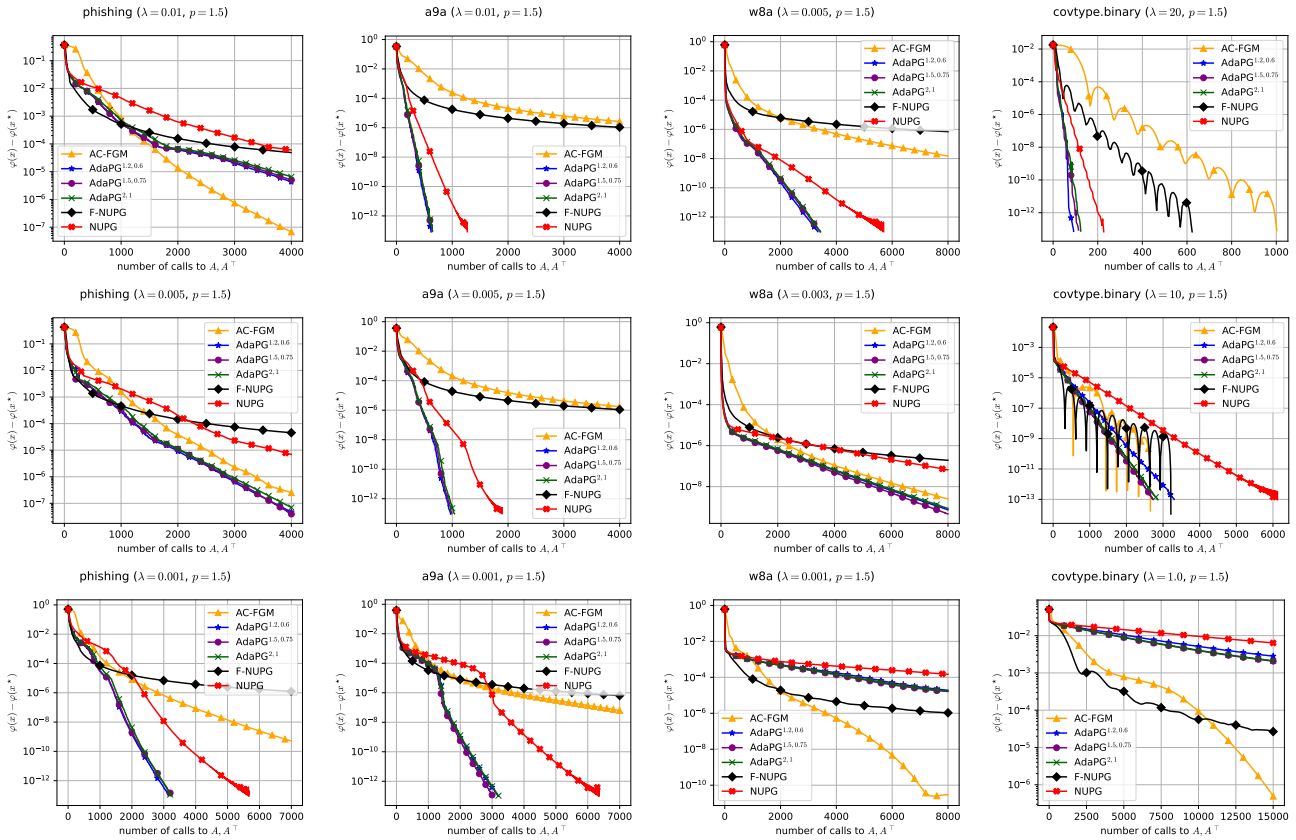

Figure 2. Classification with Hölder-smooth SVMs. It can be seen that adaPG$^{q,\frac{q}{2}}$ consistently outperforms NUPG in terms of calls to $A, A^\top$.

of the different algorithms we generate random instances of the problem using a p-norm version of the procedure provided in (Nesterov, 2013). In Figure 1 we depict convergence plots for the different methods applied to random instances with varying dimensions of $A \in \mathbb{R}^{m\times n}$, powers $p \in (1,2)$ and number of nonzero entries k of the solution $x^\star$. It can be seen that adaPG$^{q,\frac{q}{2}}$ performs consistently better than Nesterov's universal primal gradient (Nesterov, 2015) method NUPG with $\varepsilon = 1e^{-12}$ in terms of evaluations of forward and backward passes (calls to $A, A^\top$). In this experiment adaPG$^{q,\frac{q}{2}}$ also performs consistently better than the accelerated algorithms AC-FGM and F-NUPG.

4.2. Hölder-smooth SVMs with ℓ_1-regularization

In this subsection we assess the performance of the different algorithms on the task of classification using a p-norm version of the SVM model. For that purpose we consider a subset of the LibSVM binary classification benchmark that consists of a collection of datasets with varying number m of examples a_i, feature dimensions n and sparsity. Let $(a_j, b_j)_{j=1}^m$ with $b_j \in \{-1, 1\}$ and $a_j \in \mathbb{R}^n$ denote the collection of training examples. Then we consider the minimization problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{j=1}^m \frac{1}{p}\max\{0, 1 - b_j\langle a_j, x\rangle\}^p + \lambda\|x\|_1. \quad (24)$$

for $p \in (1,2)$. The loss function is globally Hölder smooth with order $\nu = p - 1$ while the proximal mapping of the second term can be solved in closed form. The results are depicted in Figure 2.

4.3. Logistic regression with p-norm regularization

In this subsection we consider classification with the logistic loss and a smooth p-norm regularizer, for some $p \in (1,2)$. The problem can be cast in convex composite minimization form as follows

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{j=1}^m \ln(1 - \exp(b_j\langle a_j, x\rangle)) + \lambda\tfrac{1}{p}\|x\|_p^p. \quad (25)$$

Unlike the previous classification model this yields a smooth optimization problem. Hence we perform gradient steps with respect to φ and choose the nonsmooth term $g \equiv 0$. The results are depicted in Figure 3.

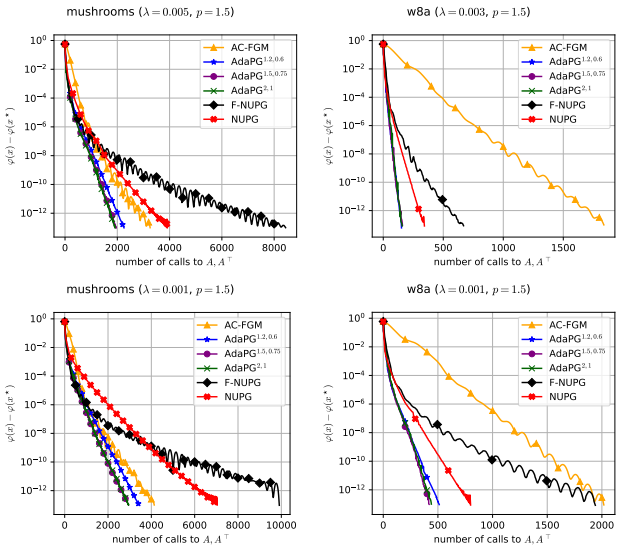

Figure 3. p-norm regularized logistic regression. In the present cases adaPG$^{q,\frac{q}{2}}$ performs better than the baseline algorithms.

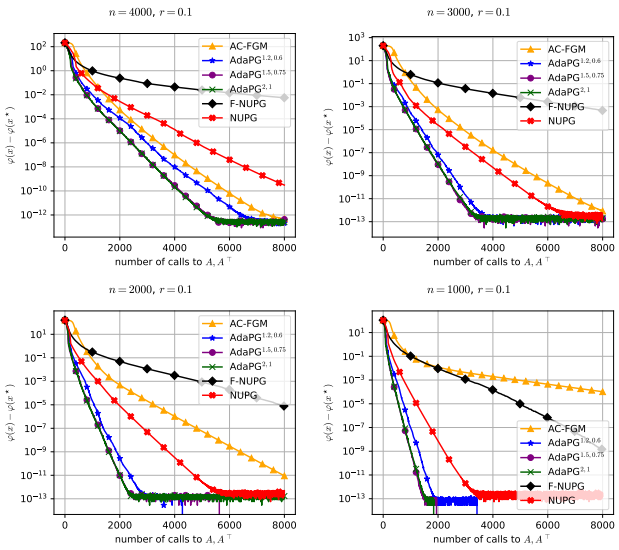

Figure 4. Mixture p-norm regression with ball constraint. It can be seen that adaPG$^{q,\frac{q}{2}}$ performs best on this comparison.

4.4. Mixture p-norm regression

In this final experiment we consider the following mixture model:

$$\min_{x \in \mathbb{B}_2(r)} \sum_{j=1}^{J} \frac{1}{p_j} \|A^j x - b^j\|_{p_j}^{p_j}, \tag{26}$$

where $\mathbb{B}_2(r)$ is the 2-norm ball with radius r and $p_j \in (1, 2]$. Since the p_j are not identical the smooth part in (26) is merely locally Hölder smooth. The nonsmooth part g is the indicator function of the set $\mathbb{B}_2(r)$. In Figure 4 we compare the performance for $J = 6$ with $p = (1.8, 1.7, 1.6, 1.5, 1.5, 1.5)$ and different values of n where the entries of $A^j \in [-1, 1]^{m_j \times n}$ and $b^j \in [-1, 1]^{m_j}$ are uniformly distributed between $[-1, 1]$ with $m = (400, 300, 400, 100, 100, 300)$.

5. Conclusions

In this paper we showed that adaptive proximal gradient methods are universal, in the sense that they converge under mere local Hölder gradient continuity of *any* order. This is achieved through a unified analysis of adaPG$^{q,\frac{q}{2}}$ that encapsulates existing methods for different values of the parameter $q \in [1, 2]$. Sequential convergence along with an $O(1/k^\nu)$ rate for the cost is established. Remarkably, the analysis and implementation of the algorithm does not require a-priori knowledge of the order $\nu \in (0, 1]$ of Hölder continuity. In other words, adaPG$^{q,\frac{q}{2}}$ and its analysis automatically adapts to the best possible choice of ν.

The validity of some of the auxiliary results for $\nu = 0$ is an encouraging indication that the algorithm could potentially cope with plain real valuedness of f, waiving differentiability assumptions. Whether this is really the case remains an interesting open question.

Moreover, our experiments demonstrate that adaPG$^{q,\frac{q}{2}}$ consistently outperforms NUPG on a diverse collection of challening convex optimization problems with both locally and globally Hölder smooth costs. In many cases we observe that adaPG$^{q,\frac{q}{2}}$ performs better than the accelerated algorithms F-NUPG and AC-FGM. We conjecture that adaPG$^{q,\frac{q}{2}}$ exploits a hidden Hölder growth that accelerated algorithms cannot take advantage from (as is known for the classical Euclidean case under metric subregularity). In future work we aim to extend our analysis to nonconvex and stochastic settings.

Acknowledgements

Work supported by: the Research Foundation Flanders 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.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Beck, A. *First-order methods in optimization*. SIAM, Philadelphia, PA, 2017.

Bochnak, J., Coste, M., and Roy, M.-F. *Real algebraic geometry*. Springer-Verlag Berlin Heidelberg, 1998.

Bolte, J., Glaudin, L., Pauwels, E., and Serrurier, M. The backtrack Hölder gradient method with application to min-max and min-min problems. *Open Journal of Mathematical Optimization*, 4:1–17, 2023.

Bredies, K. A forward–backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space. *Inverse Problems*, 25(1):015005, 2008.

Cartis, C., Gould, N. I., and Toint, P. L. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

Chartrand, R. and Yin, W. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE international conference on acoustics, speech and signal processing*, pp. 3869–3872. IEEE, 2008.

Devolder, O., Glineur, F., and Nesterov, Y. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

Doikov, N., Mishchenko, K., and Nesterov, Y. Super-universal regularized Newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024.

Forsythe, A. B. Robust estimation of straight line regression coefficients by minimizing pth power deviations. *Technometrics*, 14(1):159–166, 1972.

Ghadimi, S., Lan, G., and Zhang, H. Generalized uniformly optimal methods for nonlinear programming. *Journal of Scientific Computing*, 79:1854–1881, 2019.

Grimmer, B. On optimal universal first-order methods for minimizing heterogeneous sums. *Optimization Letters*, pp. 1–19, 2023.

Guan, W.-B. and Song, W. The forward–backward splitting method and its convergence rate for the minimization of the sum of two functions in Banach spaces. *Optimization Letters*, 15(5):1735–1758, 2021.

Hafiene, Y., Fadili, J., and Elmoataz, A. Nonlocal p-Laplacian variational problems on graphs. *arXiv: 1810.12817*, 2018.

Kamzolov, D., Dvurechensky, P., and Gasnikov, A. V. Universal intermediate gradient method for convex problems with inexact oracle. *Optimization Methods and Software*, 36(6):1289–1316, 2021.

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

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

Laude, E. and Patrinos, P. Anisotropic proximal gradient. *arXiv:2210.15531*, 2022.

Laude, E. and Patrinos, P. Anisotropic proximal point algorithm. *arXiv:2312.09834*, 2023.

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

Luque, J. *Nonlinear proximal point algorithms*. PhD thesis, Massachusetts Institute of Technology, Department of Civil Engineering, 1984.

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

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

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

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152 (1-2):381–404, 2015.

Nesterov, Y. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186:157–183, 2021a.

Nesterov, Y. Inexact accelerated high-order proximal-point methods. *Mathematical Programming*, pp. 1–26, 2021b.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4):773–810, 2021.

Oikonomidis, K. A., Bodard, A., Laude, E., and Patrinos, P. Power proximal point and augmented Lagrangian method. *arXiv:2312.12205*, 2023.

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

Stonyakin, F., Tyurin, A., Gasnikov, A., Dvurechensky, P., Agafonov, A., Dvinskikh, D., Alkousa, M., Pasechnyuk, D., Artamonov, S., and Piskunova, V. Inexact model: A framework for optimization and variational inequalities. *Optimization Methods and Software*, 36(6):1155–1201, 2021.

Yang, T. and Lin, Q. RSG: Beating subgradient method without smoothness and strong convexity. *The Journal of Machine Learning Research*, 19(1):236–268, 2018.

Yashtini, M. On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients. *Optimization letters*, 10:1361–1370, 2016.

Ying, Y. and Zhou, D.-X. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017. ISSN 1063-5203. doi: 10.1016/j.acha.2015.08.007.

Yurtsever, A., Tran Dinh, Q., and Cevher, V. A universal primal-dual convex optimization framework. *Advances in Neural Information Processing Systems*, 28, 2015.

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

A. Proofs of Section 2.1

***Proof of Fact 2.2** (Hölder-smoothness inequalities).* The proof of assertion 2.2.1 follows by the Cauchy-Schwarz inequality. For assertion 2.2.2 when $\nu > 0$ we refer the reader to (Yashtini, 2016, Lem. 1); although the reference assumes global Hölder continuity, the arguments therein only use Hölder continuity of ∇f on the segment $[x, y]$. For the case $\nu = 0$, for any $x, y \in E$ and $\nabla h(x) \in \partial h(x)$, $\nabla h(y) \in \partial h(y)$, we have

$$\begin{aligned} h(y) &\leq h(x) + \langle \nabla h(y), y - x\rangle \\ &= h(x) + \langle \nabla h(x), y - x\rangle + \langle \nabla h(y) - \nabla h(x), y - x\rangle \\ &\leq h(x) + \langle \nabla h(x), y - x\rangle + \|\nabla h(y) - \nabla h(x)\|\|y - x\| \\ &\leq h(x) + \langle \nabla h(x), y - x\rangle + L_{E,0}\|y - x\|, \end{aligned}$$

where the first inequality follows from convexity of h.

We now turn to the last two claims, and thus restrict to the case $\nu > 0$. We will only show assertion 2.2.4, as 2.2.3 in turn follows by exchanging the roles of x and y and summing the corresponding inequalities. The proof follows along the lines of (Beck, 2017, Thm. 5.8(iii)) and is included for completeness to highlight the need of the enlarged set $\overline{E}$. We henceforth fix $x, y \in E \subseteq \overline{E}$. Since ∇f is ν-Hölder continuous in $\overline{E}$ with modulus $L_{\overline{E},\nu}$, it follows from assertion 2.2.2 that

$$f(z) \leq f(y) + \langle \nabla f(y), z - y\rangle + \tfrac{L_{\overline{E},\nu}}{1+\nu}\|z - y\|^{1+\nu} \quad \forall z \in \overline{E}. \tag{27}$$

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

$$\begin{aligned} l_x(z) &= f(z) - f(x) - \langle \nabla f(x), z - x\rangle \\ &\overset{(27)}{\leq} f(y) + \langle \nabla f(y), z - y\rangle + \tfrac{L_{\overline{E},\nu}}{1+\nu}\|z - y\|^{1+\nu} - f(x) - \langle \nabla f(x), z - x\rangle \\ &= f(y) - f(x) - \langle \nabla f(x), y - x\rangle + \langle \nabla f(y) - \nabla f(x), z - y\rangle + \tfrac{L_{\overline{E},\nu}}{1+\nu}\|z - y\|^{1+\nu} \\ &= l_x(y) + \langle \nabla l_x(y), z - y\rangle + \tfrac{L_{\overline{E},\nu}}{1+\nu}\|z - y\|^{1+\nu}. \end{aligned}$$

Noticing that $\nabla l_x(x) = 0$, it follows from convexity that x is a global minimizer of l_x, hence that $\min l_x = l_x(x) = 0$. Let us denote $v \coloneqq \frac{1}{\|\nabla l_x(y)\|}\nabla l_x(y)$ and define $z = y - \|\nabla l_x(y)\|^{1/\nu} L_{\overline{E},\nu}^{-1/\nu} v$. Note that

$$\|z - y\| = \left(\tfrac{\|\nabla l_x(y)\|}{L_{\overline{E},\nu}}\right)^{\frac{1}{\nu}} = \left(\tfrac{\|\nabla f(y) - \nabla f(x)\|}{L_{\overline{E},\nu}}\right)^{\frac{1}{\nu}} \leq \|y - x\|,$$

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

$$\begin{aligned} 0 = \min l_x = l_x(x) &\leq l_x\big(y - \|\nabla l_x(y)\|^{1/\nu} L_{\overline{E},\nu}^{-1/\nu} v\big). \\ &\overset{(27)}{\leq} l_x(y) - \|\nabla l_x(y)\|^{1/\nu} L_{\overline{E},\nu}^{-1/\nu}\langle \nabla l_x(y), v\rangle + \tfrac{L_{\overline{E},\nu}}{1+\nu}\tfrac{1}{L_{\overline{E},\nu}^{1+1/\nu}}\|\nabla l_x(y)\|^{1+\frac{1}{\nu}} \\ &= f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \tfrac{\nu}{\nu+1}\tfrac{1}{L_{\overline{E},\nu}^{1/\nu}}\|\nabla f(x) - \nabla f(y)\|^{1+1/\nu}, \end{aligned}$$

as claimed. □

Proof of Lemma 2.3. Expanding the squares yields

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

From the identity $\gamma_k = \lambda_{k,\nu}\|x^k - x^{k-1}\|^{1-\nu} > 0$, the claimed expression follows. □

B. Proofs of Section 3.1

***Proof of Lemma 3.1** (FNE-like inequality).* Recall the subgradient characterization

$$\widetilde{\nabla} g(x^{k+1}) := \tfrac{x^k - x^{k+1}}{\gamma_{k+1}} - \nabla f(x^k) \in \partial g(x^{k+1}). \tag{28}$$

We have

$$\begin{aligned}\langle \tfrac{H_k(x^{k-1}) - H_k(x^k)}{\gamma_k}, x^k - x^{k+1}\rangle &= \langle \tfrac{H_k(x^{k-1}) - x^k}{\gamma_k} + \nabla f(x^k), x^k - x^{k+1}\rangle \\ &= \tfrac{1}{\gamma_{k+1}}\|x^{k+1} - x^k\|^2 + \langle x^k - x^{k+1}, \underbrace{\tfrac{H_k(x^{k-1}) - x^k}{\gamma_k}}_{\in \partial g(x^k)} - \underbrace{\tfrac{H_{k+1}(x^k) - x^{k+1}}{\gamma_{k+1}}}_{\in \partial g(x^{k+1})}\rangle \\ &\geq \tfrac{1}{\gamma_{k+1}}\|x^{k+1} - x^k\|^2,\end{aligned}$$

owing to monotonicity of ∂g. Invoking the Cauchy-Schwarz inequality completes the proof. □

***Proof of Lemma 3.3** (main inequality).* From the convexity inequality for g at x^{k+1} with subgradient $\widetilde{\nabla} g(x^{k+1})$ as in (28), we have that for any $x \in \operatorname{dom} g$

$$\begin{aligned}0 &\leq g(x) - g(x^{k+1}) + \langle \nabla f(x^k), x - x^{k+1}\rangle - \tfrac{1}{\gamma_{k+1}}\langle x^k - x^{k+1}, x - x^{k+1}\rangle \\ &= g(x) - g(x^{k+1}) + \langle \nabla f(x^k), x - x^{k+1}\rangle + \tfrac{1}{2\gamma_{k+1}}\big(\|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|x^k - x^{k+1}\|^2\big).\end{aligned} \tag{29}$$

On the other hand, using $\widetilde{\nabla} g(x^k) \in \partial g(x^k)$ yields that

$$0 \leq g(x^{k+1}) - g(x^k) + \langle \nabla f(x^{k-1}), x^{k+1} - x^k\rangle - \tfrac{1}{\gamma_k}\langle x^{k-1} - x^k, x^{k+1} - x^k\rangle. \tag{30}$$

We now rewrite the inner product in (29) is a way that allows us to exploit the above inequality:

$$\begin{aligned}\langle \nabla f(x^k), x - x^{k+1}\rangle &= \langle \nabla f(x^k), x - x^k\rangle + \langle \nabla f(x^k), x^k - x^{k+1}\rangle \\ &\leq f(x) - f(x^k) + \langle \nabla f(x^k), x^k - x^{k+1}\rangle.\end{aligned}$$

The last inner product can be controlled using (30):

$$\begin{aligned}\langle \nabla f(x^k), x^k - x^{k+1}\rangle &= \langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle + \langle \nabla f(x^{k-1}), x^k - x^{k+1}\rangle \\ &\leq \langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle + g(x^{k+1}) - g(x^k) - \tfrac{1}{\gamma_k}\langle x^{k-1} - x^k, x^{k+1} - x^k\rangle \\ &= g(x^{k+1}) - g(x^k) + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle \\ \text{(Lemma 3.1)} \quad &\leq g(x^{k+1}) - g(x^k) + \tfrac{\rho_{k+1}}{\gamma_k}\|H_k(x^{k-1}) - H_k(x^k)\|^2\end{aligned}$$

Combine this with the prior inequality and (29) to obtain

$$0 \leq \gamma_{k+1}(\varphi(x) - \varphi(x^k)) + \rho_{k+1}^2\|H_k(x^{k-1}) - H_k(x^k)\|^2 + \tfrac{1}{2}\big(\|x^k - x\|^2 - \|x^{k+1} - x\|^2 - \|x^k - x^{k+1}\|^2\big) \tag{31}$$

$$= \gamma_{k+1}(\varphi(x) - \varphi(x^k)) + \rho_{k+1}^2 M_k\|x^k - x^{k-1}\|^2 + \tfrac{1}{2}\|x^k - x\|^2 - \tfrac{1}{2}\|x^{k+1} - x\|^2 - \tfrac{1}{2}\|x^k - x^{k+1}\|^2, \tag{32}$$

where Lemma 2.3 was used in the last equality. Finally, recalling the definition of $\widetilde{\nabla} g(x^k)$ as in (28) we have

$$\nabla f(x^k) + \widetilde{\nabla} g(x^k) = \tfrac{1}{\gamma_k}\big(H_k(x^{k-1}) - H_k(x^k)\big) = \tfrac{x^{k-1} - x^k}{\gamma_k} + \nabla f(x^k) - \nabla f(x^{k-1}) \in \partial\varphi(x^k).$$

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

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

Combining the last two inequalities and letting $P_k(x) := \varphi(x^k) - \varphi(x)$ yields

$$\begin{aligned}&\tfrac{1}{2}\|x^{k+1} - x\|^2 + \gamma_{k+1}(1 + \vartheta_{k+1})P_k(x) + \tfrac{1}{2}\|x^k - x^{k+1}\|^2 \\ &\leq \tfrac{1}{2}\|x^k - x\|^2 + \gamma_{k+1}\vartheta_{k+1}P_{k-1}(x) - \rho_{k+1}\big(\vartheta_{k+1}(1 - \gamma_k\ell_k) - \rho_{k+1}M_k^2\big)\|x^{k-1} - x^k\|^2\end{aligned}$$

Letting $\vartheta_k = q\rho_k$ for all k, and setting $x = x^\star \in \arg\min\varphi$ establishes the claimed inequality. □

***Proof of Lemma 3.4** (basic properties of adaPG$^{q,\frac{q}{2}}$).*

- 3.4.1) The update rule for γ_k ensures that the coefficients of $\|x^k - x^{k-1}\|^2$ and P_{k-1} in (20) are negative, and that therefore $(\mathcal{U}_k^q(x^\star))_{k\in\mathbb{N}}$ is decreasing. Since $\mathcal{U}_k^q(x^\star) \geq 0$, the sequence converges to a finite (positive) value.

- 3.4.2) Boundedness follows by observing that $\frac{1}{2}\|x^k - x^\star\|^2 \leq \mathcal{U}_k^q(x^\star) \leq \mathcal{U}_0^q(x^\star)$ for all $k \geq 0$, where the last inequality owes to the previous assertion. In what follows, we let $L_{\Omega,\nu} < \infty$ be a ν-Hölder modulus for ∇f on a convex and compact set Ω that contains all the iterates x^k. In particular, $\ell_{k,\nu} \leq L_{k,\nu} \leq L_{\Omega,\nu}$ holds for every k. Suppose that x_∞ and x'_∞ are two optimal limit points, and observe that

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

  Since both $(\mathcal{U}_k^q(x_\infty))_{k\in\mathbb{N}}$ and $(\mathcal{U}_k^q(x'_\infty))_{k\in\mathbb{N}}$ are convergent, by taking the limit along the subsequences converging to x_∞ and x'_∞ we obtain $\langle x_\infty, x'_\infty - x_\infty\rangle = \langle x'_\infty, x'_\infty - x_\infty\rangle$, which after rearranging yields $\|x_\infty - x'_\infty\|^2 = 0$.

- 3.4.3) Since $\gamma_k(1 + q\rho_k - q\rho_{k+1}^2) \geq 0$ because of the update rule of γ_{k+1}, and with $P_k^{\min} \coloneqq \min_{0\leq i\leq k} P_i$ denoting the best-so-far cost at iteration k, a telescoping argument on (20) yields that

$$\begin{aligned} P_K^{\min}\sum_{k=1}^K \gamma_k(1 + q\rho_k - q\rho_{k+1}^2) &\leq \sum_{k=1}^K \gamma_k(1 + q\rho_k - q\rho_{k+1}^2)P_{k-1} \\ &\leq \mathcal{U}_1^q(x^\star) - \mathcal{U}_{K+1}^q(x^\star) \leq \mathcal{U}_1^q(x^\star) - \gamma_{K+1}(1 + q\rho_{K+1})P_K^{\min}, \end{aligned} \tag{33}$$

  hence that

$$\begin{aligned} P_K^{\min} &\leq \frac{\mathcal{U}_1^q(x^\star)}{\sum_{k=1}^{K+1}\gamma_k(1 + q\rho_k - q\rho_{k+1}^2) + \gamma_{k+1}(1 + q\rho_{K+1})} \\ &= \frac{\mathcal{U}_1^q(x^\star)}{\sum_{k=1}^{K}(\gamma_k + q\gamma_k\rho_k - q\gamma_{k+1}\rho_{k+1}) + \gamma_{k+1} + q\gamma_{k+1}\rho_{K+1}} = \frac{\mathcal{U}_1^q(x^\star)}{q\gamma_1\rho_1 + \sum_{k=1}^{K+1}\gamma_k}. \end{aligned}$$

  Further using the fact that $\mathcal{U}_1^q(x^\star) \leq \mathcal{U}_0^q(x^\star)$ by Lemma 3.3 completes the proof. □

C. Proofs of Section 3.2

Proof of Lemma 3.6. Owing to $\rho_{k+1} \leq \sqrt{1/q + \rho_k}$ as ensured in (1), it can be verified with a trivial induction argument that

$$\rho_k \leq \rho_{\max} \coloneqq \max\left\{\tfrac{1}{2}(1 + \sqrt{1 + 4/q}), \rho_0\right\} \quad \text{for all } k \geq 0. \tag{34}$$

If $k \in K_2$, then γ_{k+1} coincides with the second update in (10), and thus

$$\rho_{\max} \geq \rho_{k+1} = \frac{1}{\sqrt{2\left[\lambda_{k,\nu}^2 L_{k,\nu}^2 - (2-q)\lambda_{k,\nu}\ell_{k,\nu} + 1 - q\right]_+}} \geq \frac{1}{\sqrt{2}\lambda_{k,\nu}L_{k,\nu}}, \tag{35}$$

completing the proof. □

In our convergence analysis we will need the following lemma that extends (Latafat et al., 2023a, Lem. B.2) by allowing a vanishing stepsize. As a result it is only γ_{k+1} times the cost that can be ensured to converge to zero, which will nevertheless prove sufficient for our convergence analysis in the proof of Theorem 3.7.

Lemma C.1. *Suppose that a sequence $(x^k)_{k\in\mathbb{N}}$ converges to an optimal point $x^\star \in \arg\min\varphi$, and for every k let $\bar{x}^k \coloneqq \operatorname{prox}_{\gamma_{k+1}g}(x^k - \gamma_{k+1}\nabla f(x^k))$ with $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$ bounded. Then, $(\bar{x}^k)_{k\in\mathbb{N}}$ too converges to $x^\star$ and $(\gamma_{k+1}(\varphi(\bar{x}^k) - \min\varphi))_{k\in\mathbb{N}} \to 0$.*

Proof. By nonexpansiveness of the proximal mapping

$$\|\bar{x}^k - x^\star\| \leq \|x^k - x^\star - \gamma_{k+1}(\nabla f(x^k) - \nabla f(x^\star))\| \leq \|x^k - x^\star\| + \gamma_{k+1}\|\nabla f(x^k) - \nabla f(x^\star)\| \to 0,$$

where we used the fact that $x^\star = \operatorname{prox}_{\gamma_{k+1}g}(x^\star - \gamma_{k+1}\nabla f(x^\star))$ for any $\gamma_{k+1} > 0$ in the first inequality, and boundedness of γ_{k+1} in the last implication. Moreover, for every $k \in \mathbb{N}$ one has

$$\gamma_{k+1}(\varphi(\bar{x}^k) - \min\varphi) = \gamma_{k+1}(f(\bar{x}^k) + g(\bar{x}^k) - \min\varphi) \leq \gamma_{k+1}(f(\bar{x}^k) - f(x^\star)) - \langle x^k - \gamma_{k+1}\nabla f(x^k) - \bar{x}^k, x^\star - \bar{x}^k\rangle,$$

where in the inequality we used the subgradient characterization of the proximal mapping. The inner product vanishes since both x^k and $\bar{x}^k$ converge to $x^\star$, and the claim follows by continuity of f and lower semicontinuity of φ. □

***Proof of Theorem 3.7** (convergence).* We first show two intermediate claims.

Claim 3.7(a): If $\nu > 0$, then $\inf_{k\in\mathbb{N}} P_k = 0$, and in particular $(x^k)_{k\in\mathbb{N}}$ admits a (unique) optimal limit point.

If $\sup_{k\in\mathbb{N}}\gamma_k = \infty$, then we know from Lemma 3.4.3 that $\liminf_{k\to\infty} P_k = 0$. Suppose instead that $(\gamma_k)_{k\in\mathbb{N}}$ is bounded. Then, the set K_2 as in (21b) must be infinite. Let $L_{\Omega,\nu}$ be a ν-Hölder modulus for ∇f on a compact convex set Ω that contains all the iterates x^k, ensured to exist by Lemma 3.4.2. Since $L_{k,\nu} \leq L_{\Omega,\nu}$, it follows from Lemma 3.6 that

$$\lambda_{k,\nu} \geq \lambda_{\min,\nu} \coloneqq \frac{1}{\sqrt{2}L_{\Omega,\nu}\rho_{\max}} \quad \forall k \in K_2, \tag{36}$$

hence from (33) that

$$\sum_{k\in K_2} \|x^k - x^{k-1}\|^{1-\nu}(1 + q\rho_k - q\rho_{k+1}^2)P_{k-1} < \infty.$$

Noticing that $1 + q\rho_k - q\rho_{k+1}^2 = 0$ for $k \notin K_2$, necessarily $1 + q\rho_k - q\rho_{k+1}^2 \not\to 0$ as $K_2 \ni k \to \infty$ (or, equivalently, as $k \to \infty$), for otherwise $\liminf_{k\to\infty}\rho_k > 1$ and thus $\gamma_k \nearrow \infty$. Therefore, there exists an infinite set $\tilde{K}_2 \subseteq K_2$ such that $1 + q\rho_k - q\rho_{k+1}^2 \geq \varepsilon > 0$ for all $k \in \tilde{K}_2$, implying that

$$\sum_{k\in\tilde{K}_2} \|x^k - x^{k-1}\|^{1-\nu}P_{k-1} < \infty.$$

Thus, $\lim_{\tilde{K}_2\ni k\to\infty}\|x^k - x^{k-1}\| = 0$ (or $\liminf_{k\in\tilde{K}_2} P_{k-1} = 0$, in which case there is nothing to show). For any $x^\star \in \arg\min\varphi$ we thus have

$$\begin{aligned} 0 \leq P_k = \varphi(x^k) - \min\varphi &\leq \langle x^k - x^\star, \tfrac{x^{k-1}-x^k}{\gamma_k} - \big(\nabla f(x^{k-1}) - \nabla f(x^k)\big)\rangle \\ &\leq \|x^k - x^\star\|\Big(\tfrac{1}{\gamma_k}\|x^{k-1} - x^k\| + \|\nabla f(x^{k-1}) - \nabla f(x^k)\|\Big) \\ &= \|x^k - x^\star\|\Big(\tfrac{1}{\lambda_{k,\nu}\|x^{k-1}-x^k\|^{1-\nu}}\|x^{k-1} - x^k\| + L_{k,\nu}\|x^{k-1} - x^k\|^\nu\Big) \\ &\leq \|x^k - x^\star\|\|x^{k-1} - x^k\|^\nu\Big(\tfrac{1}{\lambda_{\min,\nu}} + L_{\Omega,\nu}\Big) \quad \forall k \in \tilde{K}_2. \end{aligned} \tag{37}$$

Since $\nu > 0$, by taking the limit as $\tilde{K}_2 \ni k \to \infty$ we obtain that $\lim_{\tilde{K}_2\ni k\to\infty} P_k = 0$.

Claim 3.7(b): If $\nu > 0$ and $(\gamma_k)_{k\in\mathbb{N}}$ is bounded, then $(x^k)_{k\in\mathbb{N}}$ converges to a solution.

Suppose first that $(\gamma_k)_{k\in\mathbb{N}}$ is bounded. Consider a subsequence $(x^k)_{k\in K}$ such that $\lim_{K\ni k\to\infty} P_k = 0$, which exists and converges to a solution $x^\star$ by Claim 3.7(a). Since $(\gamma_k)_{k\in\mathbb{N}}$ is bounded, in light of Lemma C.1 also $x^{k+1} \to x^\star$ and, in turn, $x^{k+2} \to x^\star$ as well. Then,

$$\mathcal{U}_{k+1}^q(x^\star) = \tfrac{1}{2}\|x^{k+1} - x^\star\|^2 + \tfrac{1}{2}\|x^{k+1} - x^k\|^2 + \gamma_k(1 + q\rho_{k+1})P_k \to 0 \quad \text{as } K \ni k \to \infty,$$

and thus $\frac{1}{2}\|x^k - x^\star\| \leq \mathcal{U}_k^q(x^\star) \to 0$ as $k \to \infty$, since the entire sequence $(\mathcal{U}_k^q(x^\star))_{k\in\mathbb{N}}$ is convergent.

To conclude the proof ot the theorem, it remains to show that also in case $(\gamma_k)_{k\in\mathbb{N}}$ is unbounded the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a solution. To this end, let us suppose now that γ_k is not bounded. This case requires requires a few more technical steps, which can nevertheless almost verbatim be adapted from the proof of (Latafat et al., 2023a, Thm. 2.4(ii)).

See also (Latafat et al., 2023b, Thm. 2.3(iii)) for an alternative argument; we emphasize that the difference with both aforementioned works is that the stepsize sequence is not guaranteed to be bounded away from zero.

We start by observing that Claim 3.7(a) and Lemma 3.4.2 ensure that an optimal limit point $x^\star \in \arg\min \varphi$ exists. It then suffices to show that $\mathcal{U}_k^q(x^\star)$ converges to zero. To arrive to a contradiction, suppose that this is not the case, that is, that $U := \lim_{k\to\infty} \mathcal{U}_k^q(x^\star) > 0$. We shall henceforth proceed by intermediate claims that follow from this condition, eventually arriving to a contradictory conclusion.

Contradiction claim 1: For any $K \subseteq \mathbb{N}$, $\lim_{K\ni k\to\infty} x^k = x^\star$ holds iff $\lim_{K\ni k\to\infty} \gamma_{k+1} = \infty$.*
The implication "$\Leftarrow$" follows from

$$\gamma_k P_{k-1} \leq \mathcal{U}_k^q(x^\star) < \infty$$

since $(x^k)_{k\in\mathbb{N}}$ is bounded and $x^\star$ is its unique optimal limit point.
Suppose now that $(x^k)_{k\in K} \to x^\star$. To arrive to a contradiction, up to possibly extracting another subsequence suppose that $(\gamma_{k+1})_{k\in K} \to \bar{\gamma} \in [0,\infty)$. Then, it follows from Lemma C.1 that $(x^{k+1})_{k\in K} \to x^\star$ and $(\gamma_{k+1}P_{k+1})_{k\in K} \to 0$. As shown in (34)

$$\rho_k \leq \rho_{\max} \quad \text{for all } k \geq 0 \tag{38}$$

which in turn implies $(\gamma_{k+2}P_{k+1})_{k\in K} \to 0$ and that $(\gamma_{k+2})_{k\in K}$ is also bounded, we may iterate and infer that also $(x^{k+2})_{k\in K}$ converges to $x^\star$. Recalling the definition of $\mathcal{U}_k^q$ in (18),

$$\mathcal{U}_{k+2}^q(x^\star) := \tfrac{1}{2}\|x^{k+2} - x^\star\|^2 + \tfrac{1}{2}\|x^{k+2} - x^{k+1}\|^2 + \gamma_{k+2}(1 + q\rho_{k+2})P_{k+1} \to 0,$$

contradicting $U = \lim_{K\ni k\to\infty} \mathcal{U}_{k+2}^q(x^\star) > 0$.

Contradiction claim 2: Suppose that $(x^k)_{k\in K} \to x^\star$; then also $(x^{k-1})_{k\in K} \to x_\star$.*
It follows from the previous claim that $\lim_{K\ni k\to\infty} \gamma_{k+1} = \infty$. Because of (38), one must also have $\lim_{K\ni k\to\infty} \gamma_k = \infty$. Invoking again the previous claim, by the arbitrarity of the index set K the assertion follows.

Contradiction claim 3: Suppose that $(x^k)_{k\in K} \to x^\star$; then $(\gamma_{k-1}L_{k-1})_{k\in K} \to \infty$ and $(\rho_k)_{k\in K} \to 0$.*
Using the previous claim twice, $x^{k-1}, x^{k-2} \to x^\star$ as $K \ni k \to \infty$. In particular

$$\lim_{K\ni k\to\infty} \|x^{k-1} - x^{k-2}\|^2 = 0. \tag{39}$$

From the expression (18) of $\mathcal{U}_k^q$ we then have

$$\lim_{K\ni k\to\infty} \gamma_k(1 + q\rho_k)P_{k-1} = U, \tag{40}$$

where we remind that by contradiction assumption $U := \lim_{k\to\infty} \mathcal{U}_k^q(x^\star) > 0$. Denoting $C := \rho_{\max}(1 + q\rho_{\max})$, we have

$$\begin{aligned}\gamma_{k-1}P_{k-1} &\leq \|x^{k-1} - x^\star\|\big(\|x^{k-1} - x^{k-2}\| + \gamma_{k-1}\|\nabla f(x^{k-1}) - \nabla f(x^{k-2})\|\big)\\ &= \|x^{k-1} - x^\star\|\big(\|x^{k-1} - x^{k-2}\| + \gamma_{k-1}L_{k-1}\|x^{k-1} - x^{k-2}\|^\nu\big)\end{aligned}$$

for every k. Then, by (40)

$$\begin{aligned}0 < U = \lim_{K\ni k\to\infty} \gamma_k(1 + q\rho_k)P_{k-1} &= \liminf_{K\ni k\to\infty} \rho_k(1 + q\rho_k)\gamma_{k-1}P_{k-1}\\ &\leq \rho_{\max}(1 + q\rho_{\max}) \liminf_{K\ni k\to\infty} \|x^{k-1} - x^\star\|\big(\|x^{k-1} - x^{k-2}\| + \gamma_{k-1}L_{k-1}\|x^{k-1} - x^{k-2}\|^\nu\big)\end{aligned}$$

which yields the first claim owing to (39) and $\nu > 0$.
This along with the update rule (10) implies

$$\rho_k \leq \frac{1}{2\big[\gamma_{k-1}^2 L_{k-1}^2 - (2-q)\gamma_{k-1}\ell_{k-1} + 1 - q\big]} \to 0,$$

as claimed.

Having shown the above claims, the proof is concluded as in (Latafat et al., 2023a, Thm. 2.4(ii)) by constructing a specific unbounded stepsize sequence and using claims 1* and 3* to obtain the sought contradiction. □

***Proof of Theorem 3.8** (sublinear rate).* The existence of Ω as in the statement follows from boundedness of $(x^k)_{k\in\mathbb{N}}$, see Lemma 3.4.2. We proceed by intermediate claims.

Claim 3.8(a): If $\lambda_{k,\nu} \le \frac{1}{L_{\Omega,\nu}}$, then $P_k \le P_{k-1}$.

Let $\widetilde{\nabla}\varphi(x^k) := \nabla f(x^k) + \widetilde{\nabla} g(x^k)$, where $\widetilde{\nabla} g$ is as in (28). Then, $\widetilde{\nabla}\varphi(x^k) \in \partial\varphi(x^k)$ and thus

$$\begin{aligned}\varphi(x^{k-1}) &\ge \varphi(x^k) + \langle \widetilde{\nabla}\varphi(x^k), x^{k-1} - x^k\rangle \\ &= \varphi(x^k) + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^{k-1} - x^k\rangle \\ &= \varphi(x^k) + \tfrac{1}{\gamma_k}\|x^k - x^{k-1}\|^2 - L_{k,\nu}\|x^{k-1} - x^k\|^{1+\nu} \\ &= \varphi(x^k) + \big(\tfrac{1}{\lambda_{k,\nu}} - L_{k,\nu}\big)\|x^{k-1} - x^k\|^{1+\nu},\end{aligned}$$

establishing the claim.

We next aim at establishing a lower bound on the stepsize sequence in terms of $P_k^{\frac{1-\nu}{\nu}}$. To simplify the exposition, we now fix $x^\star \in \arg\min\varphi$ and denote

$$\tilde{C}^q_\nu(\nu) := \sqrt{2\mathcal{U}^q_1(x^\star)}\big(\tfrac{1}{\nu} + L_{\Omega,\nu}\big). \tag{41}$$

Claim 3.8(b): For every $k \in \mathbb{N}$ it holds that $P_k \le \tilde{C}^q_\nu(\lambda_{k,\nu})\|x^k - x^{k-1}\|^\nu$.

We begin by observing that

$$\widetilde{\nabla}\varphi(x^k) := \tfrac{1}{\gamma_k}\big(H_k(x^{k-1}) - H_k(x^k)\big) \in \partial\varphi(x^k)$$

owing to (28). Combined with (16) and (8) it follows that

$$\|\widetilde{\nabla}\varphi(x^k)\| \le \big(\tfrac{1}{\lambda_{k,\nu}} + L_{k,\nu}\big)\|x^k - x^{k-1}\|^\nu \le \big(\tfrac{1}{\lambda_{k,\nu}} + L_{\Omega,\nu}\big)\|x^k - x^{k-1}\|^\nu.$$

Moreover, by convexity,

$$P_k = \varphi(x^k) - \min\varphi \le \langle \widetilde{\nabla}\varphi(x^k), x^k - x^\star\rangle \le \|\widetilde{\nabla}\varphi(x^k)\|\|x^k - x^\star\| \le \overbrace{\sqrt{2\mathcal{U}^q_1(x^\star)}\big(\tfrac{1}{\lambda_{k,\nu}} + L_{\Omega,\nu}\big)}^{\tilde{C}^q_\nu(\lambda_{k,\nu})}\|x^k - x^{k-1}\|^\nu$$

as claimed, where the last inequality uses the fact that $\frac{1}{2}\|x^k - x^\star\|^2 \le \mathcal{U}^q_k(x^\star) \le \mathcal{U}^q_1(x^\star)$.

We next analyze two possible cases for any iteration index $k \ge 0$.

Claim 3.8(c): For any $k \in K_2$, $\gamma_{k+1} \ge \frac{1}{\sqrt{2}L_{\Omega,\nu}}\Big(\frac{P_k}{\tilde{C}^q_\nu(\lambda_{\min,\nu})}\Big)^{\frac{1-\nu}{\nu}}$.

Since $L_{k,\nu} \le L_{\Omega,\nu}$, as shown in Lemma 3.6 $\lambda_{k,\nu} \ge \lambda_{\min,\nu} = \frac{1}{\sqrt{2}L_{\Omega,\nu}\rho_{\max}}$ holds for every $k \in K_2$. Moreover, by definition of K_2,

$$\begin{aligned}\gamma_{k+1} &= \frac{\gamma_k}{\sqrt{2\Big[\lambda^2_{k,\nu}L^2_{k,\nu} - (2-q)\lambda_{k,\nu}\ell_{k,\nu} + 1 - q\Big]_+}} \\ &= \frac{\lambda_{k,\nu}\|x^k - x^{k-1}\|^{1-\nu}}{\sqrt{2\Big[\lambda^2_{k,\nu}L^2_{k,\nu} - (2-q)\lambda_{k,\nu}\ell_{k,\nu} + 1 - q\Big]_+}} \ge \frac{\|x^k - x^{k-1}\|^{1-\nu}}{\sqrt{2}L_{k,\nu}} \ge \frac{\|x^k - x^{k-1}\|^{1-\nu}}{\sqrt{2}L_{\Omega,\nu}} \quad \forall k \in K_2.\end{aligned} \tag{42}$$

By using the lower bound $\|x^k - x^{k-1}\|^{1-\nu} \ge \Big(\frac{P_k}{\tilde{C}^q_\nu(\lambda_{k,\nu})}\Big)^{\frac{1-\nu}{\nu}} \ge \Big(\frac{P_k}{\tilde{C}^q_\nu(\lambda_{\min,\nu})}\Big)^{\frac{1-\nu}{\nu}}$ in Claim 3.8(b) raised to the power $\frac{1-\nu}{\nu}$ the claim follows.

Claim 3.8(d): For any $k \in K_1$, $\gamma_{k+1} \ge \begin{cases}\frac{1}{\sqrt{2q}L_{\Omega,\nu}}\Big(\frac{P_k}{\tilde{C}^q_\nu(\lambda_{\min,\nu})}\Big)^{\frac{1-\nu}{\nu}} & \text{if } (K_2 \ne \emptyset \text{ and}) \ k \ge \min K_2, \\ \big(1 + \frac{1}{q}\big)^{\frac{k}{2}}\gamma_0 & \text{otherwise.}\end{cases}$

Let

$$K_{2,<k} := K_2 \cap \{0, 1, \dots, k-1\}$$

denote the (possibly empty) set of all iteration indices up to $k-1$ such that the first term in (10) is strictly larger than the second one.
If $K_{2,<k} = \emptyset$, then $\rho_{t+1} = \sqrt{1/q + \rho_t}$ holds for all $t \leq k$, which inductively gives $\rho_t^2 \geq 1 + \frac{1}{q}$ for all $t = 1, \dots, K$ (since $\rho_0 \geq 1$). We then have

$$\gamma_{k+1}^2 = \gamma_0^2 \prod_{t=1}^{k} \rho_{t+1}^2 \geq (1 + \tfrac{1}{q})^k \gamma_0^2. \tag{43}$$

Suppose instead that $K_{2,<k} \neq \emptyset$, and let $n_{2,k}$ denote its largest element:

$$n_{2,k} \coloneqq \max K_{2,<k} = \max \left\{ i < k \mid \gamma_{i+1} < \gamma_i \sqrt{\tfrac{1}{q} + \rho_i} \right\}.$$

Observe that the update rule $\rho_{i+1} = \sqrt{\frac{1}{q} + \rho_i}$ implies that $\rho_{i+2} \geq 1$ holds whenever $i, i+1 \in K_1$. In fact, $\rho_{i+1} = \sqrt{\frac{1}{q} + \rho_i} \geq \frac{1}{\sqrt{q}}$ holds for every $i \in K_1$, in turn implying that

$$i, i+1 \in K_1 \quad \Rightarrow \quad \rho_{i+2} \geq \sqrt{\tfrac{1}{q} + \sqrt{\tfrac{1}{q}}} \geq \sqrt{\tfrac{1}{2} + \sqrt{\tfrac{1}{2}}} > 1.$$

In particular,

$$i, i+1, \dots, j \in K_1 \quad \Rightarrow \quad \textstyle\prod_{t=i+1}^{j+1} \rho_t \geq \frac{1}{\sqrt{q}} \tag{44}$$

(this being also trivially true for an empty product, since $q \geq 1$).
We consider two possible subcases:

◇ First, suppose that the index $j \coloneqq \max \left\{ n_{2,k} \leq i \leq k \mid \lambda_{i,\nu} > \frac{1}{L_{\Omega,\nu}} \right\}$ exists. Schematically,

$$\overbrace{n_{2,k}}^{\in K_2}, \overbrace{\dots, j, \underbrace{\dots, k}_{\lambda_{i,\nu} \leq \frac{1}{L_{\Omega,\nu}}}}^{\in K_1} \quad \text{and} \quad \lambda_{j,\nu} > \tfrac{1}{L_{\Omega,\nu}}. \tag{45}$$

By definition of $n_{2,k}$, all indices between j and k are in K_1, and thus

$$\begin{aligned}\gamma_{k+1} = \gamma_j \prod_{i=j+1}^{k+1} \rho_i &\overset{(44)}{\geq} \tfrac{1}{\sqrt{q}} \gamma_j = \tfrac{1}{\sqrt{q}} \lambda_{j,\nu} \|x^j - x^{j-1}\|^{1-\nu} \\ &\overset{(45)}{>} \tfrac{1}{\sqrt{q}} \tfrac{1}{L_{\Omega,\nu}} \|x^j - x^{j-1}\|^{1-\nu} \overset{\text{Claim 3.8(b)}}{\geq} \tfrac{1}{\sqrt{q}} \tfrac{1}{L_{\Omega,\nu}} \left(\tfrac{P_j}{\tilde{C}_\nu^q(\lambda_{j,\nu})} \right)^{\frac{1-\nu}{\nu}} \overset{(45)}{>} \tfrac{1}{\sqrt{q}} \tfrac{1}{L_{\Omega,\nu}} \left(\tfrac{P_j}{\tilde{C}_\nu^q(1/L_{\Omega,\nu})} \right)^{\frac{1-\nu}{\nu}}.\end{aligned}$$

Since $\lambda_{i,\nu} \leq \frac{1}{L_{\Omega,\nu}}$ holds for all $i = j+1, \dots, k$, it follows from Claim 3.8(a) that $P_k \leq P_j$, and thus

$$\gamma_{k+1} \geq \tfrac{1}{\sqrt{q}} \tfrac{1}{L_{\Omega,\nu}} \left(\tfrac{P_k}{\tilde{C}_\nu^q(1/L_{\Omega,\nu})} \right)^{\frac{1-\nu}{\nu}}. \tag{46}$$

◇ Alternatively, it holds that $\lambda_{j,\nu} \leq \frac{1}{L_{\Omega,\nu}}$ for all $j = n_{2,k}, \dots, k$, and in particular by virtue of Claim 3.8(b) we have that $P_k \leq P_{n_{2,k}}$. Arguing as before,

$$\gamma_{k+1} = \gamma_{n_{2,k}+1} \prod_{i=n_{2,k}+1}^{k+1} \rho_i \overset{(44)}{\geq} \tfrac{1}{\sqrt{q}} \gamma_{n_{2,k}+1} \overset{\text{Claim 3.8(c)}}{\geq} \tfrac{1}{\sqrt{2q} L_{\Omega,\nu}} \left(\tfrac{P_{n_{2,k}}}{\tilde{C}_\nu^q(\lambda_{\min,\nu})} \right)^{\frac{1-\nu}{\nu}} \geq \tfrac{1}{\sqrt{2q} L_{\Omega,\nu}} \left(\tfrac{P_k}{\tilde{C}_\nu^q(\lambda_{\min,\nu})} \right)^{\frac{1-\nu}{\nu}}. \tag{47}$$

Combining (46) and (47)

$$\gamma_{k+1} \geq \min \left\{ \frac{1}{\tilde{C}_\nu^q(1/L_{\Omega,\nu})^{\frac{1-\nu}{\nu}}}, \frac{1}{\sqrt{2}\tilde{C}_\nu^q(\lambda_{\min,\nu})^{\frac{1-\nu}{\nu}}} \right\} \frac{P_k^{\frac{1-\nu}{\nu}}}{\sqrt{q} L_{\Omega,\nu}} = \frac{1}{\sqrt{2q} L_{\Omega,\nu} \tilde{C}_\nu^q(\lambda_{\min,\nu})^{\frac{1-\nu}{\nu}}} P_k^{\frac{1-\nu}{\nu}},$$

where the identity uses the fact that the minimum is attained at the first element, having $\tilde{C}_\nu^q(\nu)$ decreasing in $\nu > 0$ and $\frac{1}{L_{\Omega,\nu}} \geq \lambda_{\min,\nu} = \frac{1}{\sqrt{2}\rho_{\max} L_{\Omega,\nu}}$ (since $\rho_{\max} \geq 1$).

Finally, combining Claims 3.8(c) and 3.8(d) and noting that

$$\frac{1}{\sqrt{2q}L_{\Omega,\nu}}\left(\frac{P_k}{\tilde{C}_\nu^q(\lambda_{\min,\nu})}\right)^{\frac{1-\nu}{\nu}} = \frac{1}{\sqrt{2q}L_{\Omega,\nu}^{1/\nu}}\left(\frac{1}{\sqrt{2\mathcal{U}_1^q(x^\star)}(\sqrt{2}\rho_{\max}+1)}\right)^{\frac{1-\nu}{\nu}} P_k^{\frac{1-\nu}{\nu}},$$

we conclude that

$$\gamma_{k+1} \geq \begin{cases} \dfrac{1}{\mathcal{U}_1^q(x^\star)^{\frac{1-\nu}{2\nu}} C(q,\nu)^{\frac{1}{\nu}}} P_k^{\frac{1-\nu}{\nu}} & \text{if } (K_2 \neq \emptyset \text{ and}) \ k \geq \min K_2, \\ \left(1+\frac{1}{q}\right)^{\frac{k}{2}}\gamma_0 & \text{otherwise} \end{cases}$$

holds for any $k \in \mathbb{N}$, where

$$C(q,\nu) = \sqrt{2}L_{\Omega,\nu}\sqrt{q}^{\nu}\left(1+\sqrt{2}\rho_{\max}\right)^{1-\nu}$$

is as in the statement. Denoting $k_0 = \min K_2 - 1$ if $K_2 \neq \emptyset$ and 0 otherwise, the sum of stepsizes can be lower bounded by

$$\begin{aligned}
\textstyle\sum_{k=1}^{K+1}\gamma_k = \sum_{k=1}^{k_0}\gamma_k + \sum_{k=k_0}^{K+1}\gamma_k &\geq \gamma_0\sum_{k=1}^{k_0}\left(1+\tfrac{1}{q}\right)^{\frac{k}{2}} + \frac{1}{\mathcal{U}_1^q(x^\star)^{\frac{1-\nu}{2\nu}}C(q,\nu)^{\frac{1}{\nu}}}\sum_{k=k_0}^{K+1}P_{k-1}^{\frac{1-\nu}{\nu}} \\
&\geq \gamma_0\sum_{k=1}^{k_0}\left(1+\tfrac{1}{q}\right)^{\frac{k}{2}} + \frac{K+1-k_0}{\mathcal{U}_1^q(x^\star)^{\frac{1-\nu}{2\nu}}C(q,\nu)^{\frac{1}{\nu}}}\left(\min_{k\leq K}P_k\right)^{\frac{1-\nu}{\nu}} \\
&\geq \gamma_0 k_0 + \frac{K+1-k_0}{\mathcal{U}_1^q(x^\star)^{\frac{1-\nu}{2\nu}}C(q,\nu)^{\frac{1}{\nu}}}\left(\min_{k\leq K}P_k\right)^{\frac{1-\nu}{\nu}} \\
&\geq \min\left\{\gamma_0, \frac{1}{\mathcal{U}_1^q(x^\star)^{\frac{1-\nu}{2\nu}}C(q,\nu)^{\frac{1}{\nu}}}\left(\min_{k\leq K}P_k\right)^{\frac{1-\nu}{\nu}}\right\}(K+1).
\end{aligned}$$

Therefore, in light of Lemma 3.4.3, for every $K \geq 1$ we have

$$\mathcal{U}_1^q(x^\star) \geq \min_{k\leq K}P_k\sum_{k=1}^{K+1}\gamma_k \geq \min\left\{\gamma_0\min_{k\leq K}P_k, \frac{1}{\mathcal{U}_1^q(x^\star)^{\frac{1-\nu}{2\nu}}C(q,\nu)^{\frac{1}{\nu}}}\left(\min_{k\leq K}P_k\right)^{\frac{1}{\nu}}\right\}(K+1).$$

Equivalently, for every $K \geq 1$ it holds that

$$\text{either} \quad \min_{k\leq K}P_k \leq \frac{\mathcal{U}_1^q(x^\star)}{\gamma_0(K+1)} \quad \text{or} \quad \min_{k\leq K}P_k \leq \frac{\mathcal{U}_1^q(x^\star)^{\frac{1+\nu}{2}}C(q,\nu)}{(K+1)^\nu}.$$

Further using the fact that $\mathcal{U}_1^q(x^\star) \leq \mathcal{U}_0^q(x^\star)$ by Lemma 3.3 results in the claimed bound. □

D. Implementation details of AC-FGM

In this section we describe the specific implementation of the auto-conditioned fast gradient method (AC-FGM) (Li & Lan, 2023), which in our notation reads

$$\begin{aligned}
z^{k+1} &= \operatorname{prox}_{\gamma_{k+1}g}(y^k - \gamma_{k+1}\nabla f(x^k)) \\
y^{k+1} &= (1-\beta_{k+1})y^k + \beta_{k+1}z^{k+1} \\
x^{k+1} &= (z^{k+1} + \tau_{k+1}x^k)/(1+\tau_{k+1}).
\end{aligned}$$

Regarding the positive sequences $(\gamma_k)_{k\in\mathbb{N}}, (\beta_k)_{k\in\mathbb{N}}, (\tau_k)_{k\in\mathbb{N}}$, we use the update rule described in (Li & Lan, 2023, Cor. 3) and as such in Corollary 2 of the same paper. We choose $\beta_1 = 0$ and $\beta_k = \beta = \frac{1-\sqrt{3}}{2}$ for all $k \geq 2$, ensure that $\gamma_1 \in [\frac{\beta}{4(1-\beta)c_1}, \frac{1}{3c_1}]$ and set $\gamma_2 = \frac{\beta}{2c_1}$ and $\gamma_{k+1} = \min\{\frac{\tau_{k-1}+1}{\tau_k}\gamma_k, \frac{\beta\tau_k}{4c_k}\}$ for all $k \geq 2$. Finally, $\tau_1 = 0$, $\tau_2 = 2$ and $\tau_{k+1} = \tau_k + \frac{\alpha}{2} + \frac{2(1-\alpha)\gamma_k c_k}{\beta\tau_k}$, for $k \geq 2$ and some $\alpha \in [0,1]$. We chose $\alpha = 0$, since this configuration consistently

outperfomed the others in our experiments. The sequence $(c_k)_{k\in\mathbb{N}}$ is the so-called local Lipschitz estimate and is defined as in (Li & Lan, 2023, Eq. (3.9)):

$$c_k = \begin{cases} \frac{\sqrt{\|x^1-x^0\|^2\|\nabla f(x^1)-\nabla f(x^0)\|^2+(\epsilon/4)^2}-\epsilon/4}{\|x^1-x^0\|^2} & \text{if } k=1, \\ \frac{\|\nabla f(x^k)-\nabla f(x^{k-1})\|^2}{2[f(x^{k-1})-f(x^k)-\langle\nabla f(x^k),x^{k-1}-x^k\rangle]+\epsilon/\tau_k} & \text{otherwise.} \end{cases}$$

where ϵ is the predefined desired accuracy of the algorithm.