Mathematical Programming
https://doi.org/10.1007/s10107-023-02040-5

FULL LENGTH PAPER

Series A

Complementary composite minimization, small gradients in general norms, and applications

Jelena Diakonikolas[1] · **Cristóbal Guzmán**[2]

Received: 18 June 2021 / Accepted: 21 November 2023

Abstract

Composite minimization is a powerful framework in large-scale convex optimization, based on decoupling of the objective function into terms with structurally different properties and allowing for more flexible algorithmic design. We introduce a new algorithmic framework for *complementary composite minimization*, where the objective function decouples into a (weakly) smooth and a uniformly convex term. This particular form of decoupling is pervasive in statistics and machine learning, due to its link to regularization. The main contributions of our work are summarized as follows. First, we introduce the problem of complementary composite minimization in general normed spaces; second, we provide a unified accelerated algorithmic framework to address broad classes of complementary composite minimization problems; and third, we prove that the algorithms resulting from our framework are near-optimal in most of the standard optimization settings. Additionally, we show that our algorithmic framework can be used to address the problem of making the gradients small in general normed spaces. As a concrete example, we obtain a nearly-optimal method for the standard ℓ_1 setup (small gradients in the ℓ_∞ norm), essentially matching the bound of Nesterov (Optima Math Optim Soc Newsl 88:10–11, 2012) that was previously known only for the Euclidean setup. Finally, we show that our composite methods are broadly applicable to a number of regression and other classes of optimization prob-

JD was supported by the NSF grant CCF-2007757 and by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation. CG was partially supported by INRIA through the INRIA Associate Teams project, CORFO through the Clover 2030 Engineering Strategy—14ENI-26862, and FONDECYT Project 1210362.

✉ Jelena Diakonikolas
jelena@cs.wisc.edu

Cristóbal Guzmán
crguzmanp@mat.uc.cl

[1] University of Wisconsin-Madison, Madison, USA

[2] Institute for Mathematical and Computational Engineering, Faculty of Mathematics and School of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

Published online: 05 January 2024

lems, where regularization plays a key role. Our methods lead to complexity bounds that are either new or match the best existing ones.

Keywords Composite minimization · Gradient norm minimization · Linear convergence · Regression

Mathematics Subject Classification 90C06 · 90C25 · 65K05

1 Introduction

No function can be both smooth and strongly convex with respect to an ℓ_p norm and have a dimension-independent condition number, unless $p = 2$.

This is a basic fact from convex analysis[1] and the primary reason why in the existing literature smooth and strongly convex optimization is typically considered only for Euclidean (or, slightly more generally, Hilbert) spaces. In fact, it is not only that moving away from $p = 2$ the condition number becomes dimension-dependent, but that the dependence on the dimension is polynomial for all examples of functions we know of, unless p is trivially close to two. Thus, it is tempting to assert that dimension-independent linear convergence (i.e., with logarithmic dependence on the inverse accuracy $1/\epsilon$) is reserved for Euclidean (or nearly-Euclidean) spaces, which has long been common wisdom within the optimization community.

We show that this separation between Euclidean and non-Euclidean setups is not so clear-cut, and it is in fact possible to attain linear convergence even in ℓ_p (or, more generally, in normed vector) spaces, as long as the objective function can be decomposed into two functions with complementary properties. In particular, we show that if the objective function can be written in the following *complementary composite* form

$$\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x}), \tag{1}$$

where f is convex and L-smooth w.r.t. a (not necessarily Euclidean) norm $\|\cdot\|$ and ψ is m-strongly convex w.r.t. the same norm and "simple," meaning that the optimization problems of the form[2]

$$\min_{\mathbf{x}} \langle \mathbf{z}, \mathbf{x} \rangle + \psi(\mathbf{x}) \tag{2}$$

[1] More generally, it is known that the existence of a continuous uniformly convex function with growth bounded by the squared norm implies that the space has an equivalent 2-uniformly convex norm [12]; furthermore, using duality [63], we conclude that the existence of a smooth and strongly convex function implies that the space has equivalent 2-uniformly convex and 2-uniformly smooth norms, a rare property for a normed space (the most notable examples of spaces that are simultaneously 2-uniformly convex and 2-uniformly smooth are Hilbert spaces; see e.g., [7] for related definitions and more details).

[2] This oracle should be contrasted with the proximal oracle, which would require solving problems of the form $\min_{\mathbf{x}}\{\frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + \psi(\mathbf{x})\}$ and is primarily used with Euclidean norms. Equation (2) reduces to a linear optimization/Frank-Wolfe oracle when ψ is the indicator of a convex polytope. For our analysis, it also suffices to have an oracle of the form $\min_{\mathbf{x}}\{\frac{1}{2}\|\mathbf{x} - \mathbf{z}\|^2 + \psi(\mathbf{x}) + \phi(\mathbf{x})\}$, where ϕ is strongly convex w.r.t. $\|\cdot\|$.

can be solved efficiently for any linear functional $\mathbf{z}$, then $\bar{f}(\mathbf{x})$ can be minimized to accuracy $\epsilon > 0$ in $O\left(\sqrt{\frac{L}{m}}\log(\frac{L\phi(\bar{\mathbf{x}}^\star)}{\epsilon})\right)$ iterations, where $\bar{\mathbf{x}}^\star = \operatorname{argmin}_{\mathbf{x}} \bar{f}(\mathbf{x})$. As in other standard first-order iterative methods, each iteration requires one call to the gradient oracle of f and one call to a solver for the problem from Eq. (2). To the best of our knowledge, such a result was previously known only for Euclidean spaces [51].

This is the basic variant of our result. We also consider more general setups in which f is only weakly smooth (with Hölder-continuous gradients) and ψ is uniformly convex (see Sect. 1.2 for specific definitions and useful properties). We refer to the resulting objective functions $\bar{f}$ as *complementary composite objective functions* (as functions f and ψ that constitute $\bar{f}$ have complementary properties) and to the resulting optimization problems as *complementary composite optimization problems*. The algorithmic framework, based on the Approximate Duality Gap Technique (ADGT) [29], that we consider for complementary composite optimization in Sect. 2 is near-optimal (optimal up to logarithmic or poly-logarithmic factors) in terms of iteration complexity in most of the standard optimization settings, which we certify by providing near-matching oracle complexity lower bounds in Sect. 4. On a conceptual level, the extension of ADGT to complementary composite settings that handles both uniform convexity and weak smoothness without much additional technical work is another contribution of our work.[3] We now summarize some further implications of our results.

Small gradients in ℓ_p and $\mathscr{S}_p$ norms. The original motivation for complementary composite optimization in our work comes from making the gradients of smooth functions small in non-Euclidean norms. This is a fundamental optimization question, whose study was initiated in [53] and that is still far from being well-understood. Prior to this work, (near)-optimal algorithms were known only for the Euclidean (ℓ_2) and ℓ_∞ setups.[4]

For the Euclidean setup, there are two main results: due to [43] and due to [53]. The algorithm of [43] is iteration-complexity-optimal; however, the methodology by which this algorithm was obtained is crucially Euclidean, as it relies on numerical solutions to semidefinite programs, whose formulation is made possible by assuming that the norm of the space is inner-product-induced. An alternative approach, due to [53], is to apply the fast gradient method to a regularized function $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$ for a sufficiently small $\lambda > 0$, where f is the smooth function whose gradient we hope to minimize. Under the appropriate choice of $\lambda > 0$, the resulting algorithm is near-optimal (optimal up to a logarithmic factor).

As discussed earlier, applying the fast gradient method directly to a regularized function as in [53] is out of question for $p \neq 2$, as the resulting regularized objective function cannot simultaneously be smooth and strongly convex w.r.t. $\|\cdot\|_p$ without its condition number growing with the problem dimension. This is where the framework of complementary composite optimization proposed in our work comes into play. Our result also generalizes to normed matrix spaces endowed with $\mathscr{S}_p$ (Schatten-p)

[3] We note that ADGT could already handle basic composite cases (without uniform convexity of ψ) [29] and weakly smooth cases [28]; however, previous analyses did not allow exploiting uniform convexity of ψ.

[4] In the ℓ_∞ setup, a non-Euclidean variant of gradient descent is optimal in terms of iteration complexity.

norms.[5] As a concrete example, our approach leads to near-optimal complexity results in the ℓ_1 and $\mathscr{S}_1$ (a.k.a. nuclear norm) setups, where the gradient is minimized in the ℓ_∞, respectively, $\mathscr{S}_\infty$ (a.k.a. spectral), norm.

It is important to note here why strongly convex regularizers are not sufficient in general and what motivated us to consider the more general uniformly convex functions ψ. While for $p \in (1, 2]$ choosing $\psi(\mathbf{x}) = \frac{1}{2}\|\cdot\|_p^2$ (which is $(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$; see [41, 48]) is sufficient, when $p > 2$ the strong convexity parameter of $\frac{1}{2}\|\cdot\|_p^2$ w.r.t. $\|\cdot\|_p$ is bounded above by $1/d^{1-\frac{2}{p}}$. This is not only true for $\frac{1}{2}\|\cdot\|_p^2$, but for any convex function bounded above by a constant on a unit ℓ_p-ball; see e.g., [25, Example 5.1]. Thus, in this case, we work with $\psi(\mathbf{x}) = \frac{1}{p}\|\mathbf{x}\|_p^p$, which is only uniformly convex.

Lower complexity bounds. We complement the development of algorithms for complementary composite minimization and minimizing the norm of the gradient with lower bounds for the oracle complexity of these problems. Our lower bounds leverage recent lower bounds for weakly smooth convex optimization from [27, 38]. These existing results suffice for proving lower bounds for minimizing the norm of the gradient, and certify the near-optimality of our approach for the smooth (i.e., with Lipschitz continuous gradient) setting, when $1 \le p \le 2$. On the other hand, proving lower bounds for complementary convex optimization requires the design of an appropriate oracle model; namely, one that takes into account that our algorithm accesses the gradient oracle of f and solves subroutines of type (2) w.r.t. ψ. With this model in place, we combine constructions from uniformly convex nonsmooth lower bounds [42, 60] with local smoothing [27, 38] to provide novel lower bounds for complementary composite minimization. The resulting bounds show that our algorithmic framework is nearly optimal (i.e., optimal up to poly-logarithmic factors w.r.t. dimension, target accuracy, regularity constants of the objective, and initial distance to optimum) for all interesting regimes of parameters.

Applications. The importance of complementary composite optimization and making the gradients small in ℓ_p and $\mathscr{S}_p$ norms is perhaps best exhibited by considering some of the classical problems that are frequently used in statistics and machine learning. It turns out that considering these problems within the complementary composite framework not only leads to faster algorithms in general, but also reveals some interesting properties of the solutions. For example, an application of our framework to risk minimization problems leads to statistically optimal rates of the order $\frac{1}{\sqrt{n}}$ in ℓ_p spaces for $p \in (1, 2]$, while simultaneously guaranteeing that the ℓ_p norm of the output predictor is within a constant factor of the minimum ℓ_p norm over all minimizers of the (unregularized) risk function. When p is close to 1, the latter property can be interpreted as enforcing sparsity, similar to LASSO.

Section 5 provides several illustrative examples of problems that can be addressed using our framework, including lasso, elastic net, empirical risk minimization, solving positive semidefinite linear systems with maximum constraint violation guarantee, ℓ_p regression (with standard and correlated errors), and related spectral variants. It is important to note that a single algorithmic framework suffices for addressing all of

[5] $\mathscr{S}_p$ norm of a matrix $\mathbf{A}$ is defined as the ℓ_p norm of $\mathbf{A}$'s singular values.

these problems. Most of the results we obtain in this way are either conjectured or known to be unimprovable. Furthermore, to illustrate the broad applicability of our methods, we also explore the consequences of minimizing the norm of the gradient for the discrete optimal transport problem. In this case, our space of interest is ℓ_∞, and we make use of simple quadratic regularization. The resulting arithmetic complexity of our method matches many of the recently developed methods for this problem.

1.1 Further related work

Nonsmooth convex optimization problems with the composite structure of the objective function $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$, where f is smooth and convex, but ψ is nonsmooth, convex, and "simple," are well-studied in the optimization literature [11, 37, 39, 51, 57, and references therein]. The main benefit of exploiting the composite structure lies in the ability to recover accelerated rates for nonsmooth problems. One of the most celebrated results in this domain are the FISTA algorithm from [11], and a method based on composite gradient mapping proposed in [51], which demonstrated that accelerated convergence (with rate $1/k^2$) is possible for this class of problems.

By comparison, the literature on complementary composite minimization is scarce. For example, in [51] it was proved that in a Euclidean space complementary composite optimization with smooth f and strongly convex ψ admits a linear convergence rate. The algorithm proposed there is different from ours, as it relies on the use of composite gradient mapping, for which the proximal operator of ψ (solution to problems of the form $\min_{\mathbf{x}}\{\psi(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2\}$ for all $\mathbf{z}$; compare to Eq. (2)) is assumed to be efficiently computable. A more general setting of complementary composite optimization with non-Euclidean norms can be handled by [4]; however it still requires computing a non-Euclidean version of the proximal operator of ψ, of the form $\min_{\mathbf{x}}\{\langle \mathbf{z}, \mathbf{x}\rangle + \psi(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\|^2\}$ for any $\mathbf{z}$, $\mathbf{x}_0$. In addition to being primarily applicable to Euclidean spaces, such assumptions about the existence of a generalized proximal operator further restrict the class of functions that can be efficiently optimized compared to our approach (see Sect. 2.2 for a further discussion). Another composite algorithm where linear convergence has been proved is the celebrated method from [19], where proximal steps are taken w.r.t. both terms in the composite model (f and ψ). In the case where f is smooth and ψ is strongly convex, a linear convergence rate can be established. Notice their result works with analog assumptions to ours, but it is only applicable to the Euclidean setting.

Beyond the realm of Euclidean norms, linear convergence results have been established for functions that are *relatively smooth and relatively strongly convex* [8, 9, 46]. The class of complementary composite functions does not fall into this category. Further, while we show accelerated rates (with square-root dependence on the appropriate notion of the condition number) for complementary composite optimization, such results are not attainable for relatively smooth relatively strongly convex optimization (see Appendix A for a proof). In a work independent and parallel to ours, [22, Appendix E] obtained a linear convergence result that handles complementary composite setting in which f is smooth and ψ is strongly convex, with respect to an

arbitrary norm, under a similar assumption about "simplicity" of ψ to ours, as stated in (2).

The problem of minimizing the norm of the gradient has become a central question in optimization and its applications in machine learning, mainly motivated by nonconvex settings, where the norm of the gradient is useful as a stopping criterion. However, the norm of the gradient is also useful in linearly constrained convex optimization problems, where the norm of the gradient of a Fenchel dual is useful in controlling the feasibility violation in the primal [53]. Our approach for minimizing the norm of the gradient is inspired by the regularization approach proposed in [53]. As discussed earlier, this regularization approach is not directly applicable to non-Euclidean settings, and is where our complementary composite framework becomes crucial.

Finally, our work is inspired by and uses fundamental results about the geometry of high-dimensional normed spaces; in particular, the fact that for ℓ_p and $\mathscr{S}_p$ spaces the optimal constants of uniform convexity are known [7]. These results imply that powers of the respective norm are uniformly convex, which suffices for our regularization. Moreover, those functions have explicitly computable convex conjugates (problems as in Eq. (2) can be solved in closed form), which is crucial for our algorithms to work.

1.2 Notation and preliminaries

Throughout the paper, we use boldface letters to denote vectors and italic letters to denote scalars.

We consider real finite-dimensional normed vector spaces $\mathbf{E}$, endowed with a norm $\|\cdot\|$, and denoted by $(\mathbf{E}, \|\cdot\|)$. The space dual to $(\mathbf{E}, \|\cdot\|)$ is denoted by $(\mathbf{E}^*, \|\cdot\|_*)$, where $\|\cdot\|_*$ is the norm dual to $\|\cdot\|$, defined in the usual way by $\|\mathbf{z}\|_* = \sup_{\mathbf{x}\in\mathbf{E}:\|\mathbf{x}\|\leq 1} \langle \mathbf{z}, \mathbf{x}\rangle$, where $\langle \mathbf{z}, \mathbf{x}\rangle$ denotes the evaluation of a linear functional $\mathbf{z}$ on a point $\mathbf{x} \in \mathbf{E}$. As a concrete example, we may consider the ℓ_p space $(\mathbb{R}^d, \|\cdot\|_p)$, where $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$, $1 \leq p \leq \infty$. The space dual to $(\mathbb{R}^d, \|\cdot\|_p)$ is isometrically isomorphic to the space $(\mathbb{R}^d, \|\cdot\|_{p_*})$, where $\frac{1}{p} + \frac{1}{p_*} = 1$. Throughout, given $1 \leq p \leq \infty$, we refer to $p_* = \frac{p}{p-1}$ as the conjugate exponent to p (notice that $1 \leq p_* \leq \infty$, and $\frac{1}{p} + \frac{1}{p_*} = 1$). The (closed) $\|\cdot\|$-norm ball centered at $\mathbf{x}$ with radius $R > 0$ is denoted by $\mathcal{B}_{\|\cdot\|}(\mathbf{x}, R)$. We start by recalling some standard definitions from convex analysis.

Definition 1 A function $f : \mathbf{E} \to \mathbb{R}$ is said to be (L, κ)-weakly smooth w.r.t. a norm $\|\cdot\|$, where $L > 0$ and $\kappa \in (1, 2]$, if its gradients are $(L, \kappa - 1)$ Hölder continuous, i.e., if

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbf{E}): \quad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|^{\kappa-1}.$$

We denote the class of (L, κ)-weakly smooth functions w.r.t. $\|\cdot\|$ by $\mathcal{F}_{\|\cdot\|}(L, \kappa)$.

Note that when $\kappa = 1$, the function may not be differentiable. Since we will only be working with functions that are proper, convex, and lower semicontinuous, we will still have that f is subdifferentiable on the interior of its domain [56, Theorem 23.4].

The definition of (L, κ)-weakly smooth functions then boils down to the bounded variation of the subgradients.

Definition 2 A function $\psi : \mathbf{E} \to \mathbb{R}$ is said to be q-uniformly convex w.r.t. a norm $\|\cdot\|$ and with constant λ (and we refer to such functions as (λ, q)-uniformly convex), where $\lambda \geq 0$ and $q \geq 2$, if $\forall \alpha \in (0, 1)$:

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbf{E}) : \quad \psi((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)\psi(\mathbf{x}) + \alpha\psi(\mathbf{y}) - \frac{\lambda}{q}\alpha(1-\alpha)\|\mathbf{y}-\mathbf{x}\|^q.$$

We denote the class of (λ, q)-uniformly convex functions w.r.t. $\|\cdot\|$ by $\mathcal{U}_{\|\cdot\|}(\lambda, q)$.

When ψ is only subdifferentiable (but not differentiable), we make a mild assumption that the subgradient oracle of ψ is consistent, i.e., that it returns the same element of $\partial\psi(\mathbf{x})$ whenever queried at the same point $\mathbf{x}$.

Observe that when $\lambda = 0$, uniform convexity reduces to standard convexity, while for $\lambda > 0$ and $q = 2$ we recover the definition of strong convexity. We only consider functions that are lower semicontinuous, convex, and proper. These properties suffice for a function to be subdifferentiable on the interior of its domain. It is then not hard to show that if ψ is (λ, q)-uniformly convex w.r.t. a norm $\|\cdot\|$ and $\mathbf{g}_\mathbf{x} \in \partial\psi(\mathbf{x})$ is its subgradient at a point $\mathbf{x}$, we have

$$(\forall \mathbf{y} \in \mathbf{E}) : \quad \psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \mathbf{g}_\mathbf{x}, \mathbf{y}-\mathbf{x}\rangle + \frac{\lambda}{q}\|\mathbf{y}-\mathbf{x}\|^q. \tag{3}$$

Definition 3 Let $\psi : \mathbf{E} \to \mathbb{R} \cup \{+\infty\}$. The convex conjugate of ψ, denoted by ψ^*, is defined by

$$(\forall \mathbf{z} \in \mathbf{E}^*) : \quad \psi^*(\mathbf{z}) = \sup_{\mathbf{x}\in\mathbf{E}}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}.$$

Recall that the convex conjugate of any function is convex. Some simple examples of conjugate pairs of functions that will be useful for our analysis are: (i) univariate functions $\frac{1}{p}|\cdot|^p$ and $\frac{1}{p_*}|\cdot|^{p_*}$, where $1 < p < \infty$ (see, e.g., [13, Exercise 4.4.2]) and (ii) functions $\frac{1}{2}\|\cdot\|^2$ and $\frac{1}{2}\|\cdot\|_*^2$, where norms $\|\cdot\|$ and $\|\cdot\|_*$ are dual to each other (see, e.g., [15, Example 3.27]). The latter example can be easily adapted to prove that the functions $\frac{1}{p}\|\cdot\|^p$ and $\frac{1}{p_*}\|\cdot\|_*^{p_*}$ are conjugates of each other, for $1 < p < \infty$.

The following auxiliary facts will be useful for our analysis.

Fact 1 *Let $\psi : \mathbf{E} \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower semicontinuous, and let ψ^* be its convex conjugate. Then ψ^* is proper, convex, and lower semicontinuous (and thus subdifferentiable on the interior of its domain) and $\forall \mathbf{z} \in \operatorname{int}\operatorname{dom}(\psi^*)$: $\mathbf{g} \in \partial\psi^*(\mathbf{z})$ if and only if $\mathbf{g} \in \operatorname{argmax}_{\mathbf{x}\in\mathbb{R}^d}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$.*

The following proposition will be repeatedly used in our analysis, and we prove it here for completeness.

Proposition 1 *Let* $(\mathbf{E}, \|\cdot\|)$ *be a normed space with* $\|\cdot\|^2 : \mathbf{E} \to \mathbb{R}$ *differentiable, and let* $1 < q < \infty$. *Then*

$$\left\|\nabla\left(\frac{1}{q}\|\mathbf{x}\|^q\right)\right\|_* = \|\mathbf{x}\|^{q-1} = \|\mathbf{x}\|^{q/q_*},$$

where $q_* = \frac{q}{q-1}$ *is the exponent conjugate to* q.

Proof We notice that $\|\cdot\|^2$ is differentiable if and only if $\|\cdot\|^q$ is differentiable [64, Thm. 3.7.2]. Since the statement clearly holds for $\mathbf{x} = \mathbf{0}$, in the following we assume that $\mathbf{x} \neq \mathbf{0}$. Next, write $\frac{1}{q}\|\cdot\|^q$ as a composition of functions $\frac{1}{q}|\cdot|^{q/2}$ and $\|\cdot\|^2$. Applying the chain rule of differentiation, we now have:

$$\nabla\left(\frac{1}{q}\|\mathbf{x}\|^q\right) = \frac{1}{2}\left(\|\mathbf{x}\|^2\right)^{\frac{q}{2}-1}\nabla\left(\|\mathbf{x}\|^2\right) = \|\mathbf{x}\|^{q-2}\nabla\left(\frac{1}{2}\|\mathbf{x}\|^2\right).$$

It remains to argue that $\left\|\nabla\left(\frac{1}{2}\|\mathbf{x}\|^2\right)\right\|_* = \|\mathbf{x}\|$. This immediately follows by Fact 1, as $\frac{1}{2}\|\cdot\|^2$ and $\frac{1}{2}\|\cdot\|_*^2$ are convex conjugates of each other. □

We also state here a lemma that allows approximating weakly smooth functions by weakly smooth functions of a different order. A variant of this lemma (for $p = 2$) first appeared in [26], while the more general version stated here is from [25].

Lemma 1 *Let* $f : \mathbf{E} \to \mathbb{R}$ *be a function that is* (L, κ)*-weakly smooth w.r.t. some norm* $\|\cdot\|$. *Then for any* $\delta > 0$ *and*

$$M \geq \left[\frac{2(p-\kappa)}{p\kappa\delta}\right]^{\frac{p-\kappa}{\kappa}} L^{\frac{p}{\kappa}} \tag{4}$$

we have

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbf{E}): \quad f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{M}{p}\|\mathbf{y} - \mathbf{x}\|^p + \frac{\delta}{2}.$$

Finally, the following lemma will be useful when bounding the gradient norm in Sect. 3 (see also [64, Section 3.5]).

Lemma 2 *Let* $f : \mathbf{E} \to \mathbb{R}$ *be a function that is convex and* (L, κ)*-weakly smooth w.r.t. some norm* $\|\cdot\|$. *Then:*

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbf{E}): \frac{\kappa - 1}{L^{\frac{1}{\kappa-1}}\kappa}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_*^{\frac{\kappa}{\kappa-1}} \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle.$$

Proof Let $h(\mathbf{x})$ be any (L, κ)-weakly smooth function and let $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} h(\mathbf{x})$. As h is (L, κ)-weakly smooth, we have for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$h(\mathbf{y}) \leq h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{\kappa}\|\mathbf{y} - \mathbf{x}\|^\kappa.$$

Fixing $\mathbf{x} \in \mathbb{R}^d$ and minimizing both sides of the last inequality w.r.t. $\mathbf{y} \in \mathbb{R}^d$, it follows that

$$h(\mathbf{x}^\star) \leq h(\mathbf{x}) - \frac{L^{1-\kappa_*}}{\kappa_*} \|\nabla h(\mathbf{x})\|_*^{\kappa_*}, \tag{5}$$

where we have used that the functions $\frac{1}{\kappa}\|\cdot\|^\kappa$ and $\frac{1}{\kappa_*}\|\cdot\|_*^{\kappa_*}$ are convex conjugates of each other.

To complete the proof, it remains to apply Eq. (5) to function $h_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{y}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$ for any fixed $\mathbf{x} \in \mathbb{R}^d$, and observe that $h_{\mathbf{x}}(\mathbf{y})$ is convex, (L, κ)-weakly smooth, and minimized at $\mathbf{y} = \mathbf{x}$. □

2 Complementary composite minimization

In this section, we consider minimizing complementary composite functions, which are of the form

$$\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x}), \tag{6}$$

where f is (L, κ)-weakly smooth w.r.t. some norm $\|\cdot\|$, $\kappa \in (1, 2]$, and ψ is (λ, q)-uniformly convex w.r.t. the same norm, for some $q \geq 2$, $\lambda \geq 0$. We assume that the feasible set $\mathcal{X} \subseteq \mathbf{E}$ is closed, convex, and nonempty.

2.1 Algorithmic framework and convergence analysis

The algorithmic framework we consider is a generalization of AGD+ from [21],[6] stated as follows:

Generalized AGD+

$$\begin{aligned}
\mathbf{x}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1} \\
\mathbf{v}_k &= \operatorname*{argmin}_{\mathbf{u}\in\mathcal{X}} \left\{ \sum_{i=0}^{k} a_i \langle \nabla f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i\rangle + A_k \psi(\mathbf{u}) + m_0 \phi(\mathbf{u}) \right\} \\
\mathbf{y}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k, \\
\mathbf{y}_0 &= \mathbf{v}_0,\ \mathbf{x}_0 \in \mathcal{X},
\end{aligned} \tag{7}$$

where m_0 and the sequence of positive numbers $\{a_k\}_{k\geq 0}$ are parameters of the algorithm specified in the convergence analysis below, $A_k = \sum_{i=0}^k a_i$, and we take $\phi(\mathbf{u})$ to be

[6] The same method is also known as the method of similar triangles, introduced independently in [37]. This method is also related to Tseng's accelerated proximal gradient method [61], and can be seen as its "lazy" or dual averaging counterpart.

a function that satisfies $\phi(\mathbf{u}) \geq \frac{1}{q}\|\mathbf{u} - \mathbf{x}_0\|^q$. For example, if $\lambda > 0$, we can take $\phi(\mathbf{u}) = \frac{1}{\lambda} D_\psi(\mathbf{u}, \mathbf{x}_0)$. Observe also that for this choice of ϕ, the minimization problem defining $\mathbf{v}_k$ is of the form of Eq. (2). When $\lambda = 0$, we take ϕ to be $(1, q)$-uniformly convex.

The convergence analysis relies on the approximate duality gap technique (ADGT) of [29]. The main idea is to construct an upper estimate $G_k \geq \bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star)$ of the true optimality gap, where $\bar{\mathbf{x}}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \bar{f}(\mathbf{u})$, and then argue that $A_k G_k \leq A_{k-1} G_{k-1} + E_k$, which in turn implies:

$$\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) \leq \frac{A_0 G_0}{A_k} + \frac{\sum_{i=1}^k E_i}{A_k}.$$

Thus, as long as $A_0 G_0$ is bounded and the cumulative error $\sum_{i=1}^k E_i$ is either bounded or increasing slowly compared to A_k, the optimality gap of the sequence $\mathbf{y}_k$ converges to the optimum at rate $(1 + \sum_{i=1}^k E_i)/A_k$. The goal is, of course, to make A_k as fast-growing as possible, but that turns out to be limited by the requirement that $A_k G_k$ be non-increasing or slowly increasing compared to A_k.

The gap G_k is constructed as the difference $U_k - L_k$, where $U_k \geq \bar{f}(\mathbf{y}_k)$ is an upper bound on $\bar{f}(\mathbf{y}_k)$ and $L_k \leq \bar{f}(\bar{\mathbf{x}}^\star)$ is a lower bound on $\bar{f}(\bar{\mathbf{x}}^\star)$. In this particular case, we make the following choices:

$$U_k = f(\mathbf{y}_k) + \frac{1}{A_k} \sum_{i=0}^k a_i \psi(\mathbf{v}_i).$$

As $\mathbf{y}_k = \frac{1}{A_k} \sum_{i=0}^k a_i \mathbf{v}_i$, we have, by Jensen's inequality: $U_k \geq f(\mathbf{y}_k) + \psi(\mathbf{y}_k) = \bar{f}(\mathbf{y}_k)$, i.e., U_k is a valid upper bound on $\bar{f}(\mathbf{y}_k)$.

For the lower bound, we use the following inequalities:

$$\begin{aligned}
\bar{f}(\bar{\mathbf{x}}^\star) \geq\ & \frac{1}{A_k} \sum_{i=0}^k a_i f(\mathbf{x}_i) + \frac{1}{A_k} \sum_{i=0}^k a_i \left\langle \nabla f(\mathbf{x}_i), \bar{\mathbf{x}}^\star - \mathbf{x}_i \right\rangle + \psi(\bar{\mathbf{x}}^\star) \\
& + \frac{m_0}{A_k} \phi(\bar{\mathbf{x}}^\star) - \frac{m_0}{A_k} \phi(\bar{\mathbf{x}}^\star) \\
\geq\ & \frac{1}{A_k} \sum_{i=0}^k a_i f(\mathbf{x}_i) - \frac{m_0}{A_k} \phi(\bar{\mathbf{x}}^\star) \\
& + \frac{1}{A_k} \min_{\mathbf{u}\in\mathcal{X}} \left\{ \sum_{i=0}^k a_i \left\langle \nabla f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i \right\rangle + A_k \psi(\mathbf{u}) + m_0 \phi(\mathbf{u}) \right\} \\
=:\ & L_k,
\end{aligned}$$

where the first inequality uses

$$f(\bar{\mathbf{x}}^\star) \geq \frac{1}{A_k} \sum_{i=0}^k a_i f(\mathbf{x}_i) + \frac{1}{A_k} \sum_{i=0}^k a_i \left\langle \nabla f(\mathbf{x}_i), \bar{\mathbf{x}}^\star - \mathbf{x}_i \right\rangle,$$

by convexity of f.

We start by bounding the initial (scaled) gap $A_0 G_0$.

Lemma 3 (Initial Gap) *For any* $\delta_0 > 0$ *and* $M_0 = \left[\frac{2(q-\kappa)}{q\kappa\delta_0}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}$, *if* $A_0 M_0 = m_0$, *then*

$$A_0 G_0 \leq m_0 \phi(\bar{\mathbf{x}}^\star) + \frac{A_0 \delta_0}{2}.$$

Proof By definition, and using that $a_0 = A_0$,

$$\begin{aligned} A_0 G_0 = \; & A_0\Big(f(\mathbf{y}_0) + \psi(\mathbf{v}_0) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{v}_0 - \mathbf{x}_0\rangle - \psi(\mathbf{v}_0) - \frac{m_0}{A_0}\phi(\mathbf{v}_0)\Big) \\ & + m_0 \phi(\bar{\mathbf{x}}^\star) \\ = \; & A_0(f(\mathbf{y}_0) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{y}_0 - \mathbf{x}_0\rangle) - m_0\phi(\mathbf{y}_0) + m_0\phi(\bar{\mathbf{x}}^\star) \end{aligned}$$

where the second line is by $\mathbf{y}_0 = \mathbf{v}_0$.

By assumption, $\phi(\mathbf{u}) \geq \frac{1}{q}\|\mathbf{u} - \mathbf{x}_0\|^q$, for all $\mathbf{u}$, and, in particular, $\phi(\mathbf{y}_0) \geq \frac{1}{q}\|\mathbf{y}_0 - \mathbf{x}_0\|_q^q$. On the other hand, by (L, κ)-weak smoothness of f and using Lemma 1, we have that (below $M_0 = \left[\frac{2(q-\kappa)}{q\kappa\delta_0}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}$):

$$f(\mathbf{y}_0) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{y}_0 - \mathbf{x}_0\rangle \leq \frac{M_0}{q}\|\mathbf{y}_0 - \mathbf{x}_0\|_q^q + \frac{\delta_0}{2}.$$

Therefore:

$$A_0 G_0 \leq \big(A_0 M_0 - m_0\big)\frac{\|\mathbf{y}_0 - \mathbf{x}_0\|^q}{q} + m_0\phi(\bar{\mathbf{x}}^\star) + \frac{A_0\delta_0}{2} = m_0\phi(\bar{\mathbf{x}}^\star) + \frac{A_0\delta_0}{2}, \tag{8}$$

as $m_0 = A_0 M_0$. □

The next step is to bound $A_k G_k - A_{k-1} G_{k-1}$, as in the following lemma.

Lemma 4 (Gap Evolution) *Given arbitrary* $\delta_k > 0$ *and* $M_k = \left[\frac{2(q-\kappa)}{q\kappa\delta_k}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}$, *if* $\frac{a_k{}^q}{A_k{}^{q-1}} \leq \frac{\max\{\lambda A_{k-1}, m_0\}}{M_k}$ *then*

$$A_k G_k - A_{k-1} G_{k-1} \leq \frac{A_k \delta_k}{2}.$$

Proof To bound $A_k G_k - A_{k-1} G_{k-1}$, we first bound $A_k U_k - A_{k-1} U_{k-1}$ and $A_k L_k - A_{k-1} L_{k-1}$. By definition of U_k,

$$\begin{aligned} A_k U_k - A_{k-1} U_{k-1} = \; & A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}) + a_k \psi(\mathbf{v}_k) \\ = \; & A_k(f(\mathbf{y}_k) - f(\mathbf{x}_k)) + A_{k-1}(f(\mathbf{x}_k) - f(\mathbf{y}_{k-1})) \\ & + a_k f(\mathbf{x}_k) + a_k \psi(\mathbf{v}_k). \end{aligned} \tag{9}$$

For the lower bound, define the function under the minimum in the definition of the lower bound as $h_k(\mathbf{u}) := \sum_{i=0}^{k} a_i \langle \nabla f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i \rangle + A_k \psi(\mathbf{u}) + m_0 \phi(\mathbf{u})$, so that we have:

$$A_k L_k - A_{k-1} L_{k-1} = a_k f(\mathbf{x}_k) + h_k(\mathbf{v}_k) - h_{k-1}(\mathbf{v}_{k-1}). \tag{10}$$

Observe first that

$$h_k(\mathbf{v}_k) - h_{k-1}(\mathbf{v}_k) = a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle + a_k \psi(\mathbf{v}_k). \tag{11}$$

On the other hand, using the definition of Bregman divergence and the fact that Bregman divergence is blind to constant and linear terms, we can bound $h_{k-1}(\mathbf{v}_k) - h_{k-1}(\mathbf{v}_{k-1})$ as

$$\begin{aligned} h_{k-1}(\mathbf{v}_k) - h_{k-1}(\mathbf{v}_{k-1}) &= \langle \nabla h_{k-1}(\mathbf{v}_{k-1}), \mathbf{v}_k - \mathbf{v}_{k-1} \rangle + D_{h_{k-1}}(\mathbf{v}_k, \mathbf{v}_{k-1}) \\ &\geq A_{k-1} D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) + m_0 D_\phi(\mathbf{v}_k, \mathbf{v}_{k-1}), \end{aligned}$$

where the second line is by $\mathbf{v}_{k-1}$ being the minimizer of h_{k-1}. Combining with Eqs. (10) and (11), we have:

$$\begin{aligned} A_k L_k - A_{k-1} L_{k-1} \geq\ & a_k f(\mathbf{x}_k) + a_k \psi(\mathbf{v}_k) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle \\ & + A_{k-1} D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) + m_0 D_\phi(\mathbf{v}_k, \mathbf{v}_{k-1}). \end{aligned} \tag{12}$$

Combining Eqs. (9) and (12), we can now bound $A_k G_k - A_{k-1} G_{k-1}$ as

$$\begin{aligned} A_k G_k - A_{k-1} G_{k-1} \leq\ & A_k (f(\mathbf{y}_k) - f(\mathbf{x}_k)) + A_{k-1}(f(\mathbf{x}_k) - f(\mathbf{y}_{k-1})) \\ & - a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle \\ & - A_{k-1} D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) - m_0 D_\phi(\mathbf{v}_k, \mathbf{v}_{k-1}) \\ \leq\ & A_k (f(\mathbf{y}_k) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \rangle) \\ & - A_{k-1} D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) - m_0 D_\phi(\mathbf{v}_k, \mathbf{v}_{k-1}), \end{aligned}$$

where we have used $f(\mathbf{x}_k) - f(\mathbf{y}_{k-1}) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{y}_{k-1} \rangle$ (by convexity of f) and the definition of $\mathbf{y}_k$ from Eq. (7). Similarly as for the initial gap, we now use the weak smoothness of f and Lemma 1 to write:

$$\begin{aligned} f(\mathbf{y}_k) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \rangle &\leq \frac{M_k}{q} \|\mathbf{y}_k - \mathbf{x}_k\|^q + \frac{\delta_k}{2} \\ &= \frac{M_k}{q} \frac{a_k{}^q}{A_k{}^q} \|\mathbf{v}_k - \mathbf{v}_{k-1}\|^q + \frac{\delta_k}{2}, \end{aligned}$$

where $M_k = \left[\frac{2(q-\kappa)}{q\kappa\delta_k}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}$ and the equality is by $\mathbf{y}_k - \mathbf{x}_k = \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{v}_{k-1})$, which follows by the definition of algorithm steps from Eq. (7).

On the other hand, as ψ is (λ, q)-uniformly convex, we have that

$$D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) \geq \frac{\lambda}{q}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^q.$$

Further, if $\lambda = 0$, we have that $D_\phi(\mathbf{v}_k, \mathbf{v}_{k-1}) \geq \frac{1}{q}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^q$. Thus:

$$\begin{aligned} A_k G_k - A_{k-1} G_{k-1} &\leq \left(M_k \frac{{a_k}^q}{{A_k}^{q-1}} - \max\{\lambda A_{k-1}, m_0\}\right) \frac{\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^q}{q} + \frac{A_k \delta_k}{2} \\ &\leq \frac{A_k \delta_k}{2}, \end{aligned}$$

as $\frac{{a_k}^q}{{A_k}^{q-1}} \leq \frac{\max\{\lambda A_{k-1}, m_0\}}{M_k}$. □

We are now ready to state and prove the main result from this section.

Theorem 2 *Let $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$, where f is convex and (L, κ)-weakly smooth w.r.t. a norm $\|\cdot\|$, $\kappa \in (1, 2]$, and ψ is q-uniformly convex with constant $\lambda \geq 0$ w.r.t. the same norm for some $q \geq 2$. Let $\bar{\mathbf{x}}^\star$ be the minimizer of $\bar{f}$. Let $\mathbf{x}_k, \mathbf{v}_k, \mathbf{y}_k$ evolve according to Eq. (7) for an arbitrary initial point $\mathbf{x}_0 \in \mathcal{X}$, where $A_0 M_0 = m_0$, ${a_k}^q \leq \frac{\max\{\lambda {A_{k-1}}^q, m_0 {A_k}^{q-1}\}}{M_k}$ for $k \geq 1$, and $M_k = \left[\frac{2(q-\kappa)}{q\kappa\delta_k}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}$, for $\delta_k > 0$ and $k \geq 0$. Then, $\forall k \geq 1$:*

$$\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) \leq \frac{2A_0 M_0 \phi(\bar{\mathbf{x}}^\star) + \sum_{i=0}^k A_i \delta_i}{2A_k}.$$

In particular, for any $\epsilon > 0$, setting $\delta_k = \frac{a_k}{A_k}\epsilon$, for $k \geq 0$, $a_0 = A_0 = 1$, and ${a_k}^q = \frac{\max\{\lambda {A_{k-1}}^q, m_0 {A_k}^{q-1}\}}{M_k}$ for $k \geq 1$, we have that $\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) \leq \epsilon$ after at most

$$k = O\left(\min\left\{\left(\frac{1}{\epsilon}\right)^{\frac{q-\kappa}{q\kappa-q+\kappa}}\left(\max\left\{\frac{L^{\frac{q}{\kappa}}}{\lambda}, 1\right\}\right)^{\frac{\kappa}{q\kappa-q+\kappa}} \log\left(\frac{L\phi(\bar{\mathbf{x}}^\star)}{\epsilon}\right), \left(\frac{L}{\epsilon}\right)^{\frac{q}{q\kappa-q+\kappa}}\left(\phi(\bar{\mathbf{x}}^\star)\right)^{\frac{\kappa}{q\kappa-q+\kappa}}\right\}\right)$$

iterations.

Proof The first part of the theorem follows immediately by combining Lemma 3 and Lemma 4.

For the second part, we have

$$\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) \leq \frac{A_0 M_0 \phi(\bar{\mathbf{x}}^\star)}{A_k} + \frac{\epsilon}{2},$$

so all we need to show is that, under the step size choice from the theorem statement, we have $\frac{A_0 M_0 \phi(\bar{\mathbf{x}}^\star)}{A_k} \leq \frac{\epsilon}{2}$.

As $A_0 = a_0 = 1$, we have that $\delta_0 = \epsilon$ and

$$M_0 = \left[\frac{2(q-\kappa)}{q\kappa\epsilon}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}. \tag{13}$$

It remains to bound the growth of A_k. In this case, by theorem assumption, we have ${a_k}^q = \frac{\max\{\lambda {A_{k-1}}^q, m_0 {A_k}^{q-1}\}}{M_k}$. Thus, (i) $\frac{{a_k}^q}{{A_{k-1}}^q} \geq \frac{\lambda}{M_k}$ and (ii) $\frac{{a_k}^q}{{A_k}^{q-1}} \geq \frac{m_0}{M_k}$, and the growth of A_k can be bounded below as the maximum of growths determined by these two cases.

Consider $\frac{{a_k}^q}{{A_{k-1}}^q} \geq \frac{\lambda}{M_k}$ first. As $\delta_k = \frac{a_k}{A_k}\epsilon$ and $M_k = \left[\frac{2(q-\kappa)}{q\kappa\delta_k}\right]^{\frac{q-\kappa}{\kappa}} L^{\frac{q}{\kappa}}$, the condition $\frac{{a_k}^q}{{A_k}^{q-1}A_{k-1}} \geq \frac{\lambda}{M_k}$ can be equivalently written as:

$$\frac{{a_k}^{q-\frac{q}{\kappa}+1}}{{A_{k-1}}^{q-\frac{q}{\kappa}+1}} \geq \left[\frac{2(q-\kappa)}{q\kappa\epsilon}\right]^{-\frac{q-\kappa}{\kappa}} \frac{\lambda}{L^{\frac{q}{\kappa}}}.$$

Hence,

$$\frac{a_k}{A_{k-1}} \geq \left[\frac{2(q-\kappa)}{q\kappa\epsilon}\right]^{-\frac{q-\kappa}{q\kappa-q+\kappa}} \left(\frac{\lambda}{L^{\frac{q}{\kappa}}}\right)^{\frac{\kappa}{q\kappa-q+\kappa}}.$$

As $a_k = A_k - A_{k-1}$, it follows that $\frac{A_k}{A_{k-1}} \geq 1 + \left[\frac{2(q-\kappa)}{q\kappa\epsilon}\right]^{-\frac{q-\kappa}{q\kappa-q+\kappa}} \left(\frac{\lambda}{L^{\frac{q}{\kappa}}}\right)^{\frac{\kappa}{q\kappa-q+\kappa}}$, further leading to

$$A_k \geq \left(1 + \left[\frac{2(q-\kappa)}{q\kappa\epsilon}\right]^{-\frac{q-\kappa}{q\kappa-q+\kappa}} \left(\frac{\lambda}{L^{\frac{q}{\kappa}}}\right)^{\frac{\kappa}{q\kappa-q+\kappa}}\right)^k.$$

On the other hand, the condition $\frac{{a_k}^q}{{A_k}^{q-1}} \geq \frac{m_0}{M_k}$ can be equivalently written as:

$$\frac{{a_k}^{\frac{q\kappa-q}{\kappa}+1}}{{A_k}^{\frac{q\kappa-q}{\kappa}}} \geq \frac{m_0}{L^{\frac{q}{\kappa}}}\left[\frac{q\kappa\epsilon}{2(q-\kappa)}\right]^{\frac{q-\kappa}{\kappa}} = 1,$$

where we have used the definition of m_0, which implies

$$A_k = \Omega\left(k^{\frac{q\kappa-q+\kappa}{\kappa}}\right), \tag{14}$$

and further leads to the claimed bound on the number of iterations. □

Let us point out some special cases of the bound from Theorem 2. When f is smooth ($\kappa = 2$) and ψ is q-uniformly convex, assuming $L^{q/2} \geq \lambda$, the bound simplifies to

$$k = O\left(\min\left\{\left(\frac{1}{\epsilon}\right)^{\frac{q-2}{q+2}}\left(\frac{L^{\frac{q}{2}}}{\lambda}\right)^{\frac{2}{q+2}} \log\left(\frac{L\phi(\bar{\mathbf{x}}^\star)}{\epsilon}\right),\ \left(\frac{L}{\epsilon}\right)^{\frac{q}{q+2}}\left(\phi(\bar{\mathbf{x}}^\star)\right)^{\frac{2}{q+2}}\right\}\right). \tag{15}$$

In particular, if ψ is strongly convex ($q = 2$), we recover the same bound as in the Euclidean case:

$$k = O\left(\min\left\{\sqrt{\frac{L}{\lambda}}\log\left(\frac{L\phi(\bar{\mathbf{x}}^\star)}{\epsilon}\right),\ \sqrt{\frac{L\phi(\bar{\mathbf{x}}^\star)}{\epsilon}}\right\}\right). \tag{16}$$

Note that this result uses smoothness of f and strong convexity of ψ with respect to the same but arbitrary norm $\|\cdot\|$. Because we do not require the same function to be simultaneously smooth and strongly convex w.r.t. $\|\cdot\|$, the resulting "condition number" $\frac{L}{\lambda}$ can be dimension-independent even for non-Euclidean norms (in particular, this will be possible for any ℓ_p norm with $p \in (1, 2]$). Observe further that it is generally possible for the "condition number" $\frac{L}{\lambda}$ to be smaller than one. In that case, as the convergence bound in Theorem 2 depends on $\max\{\frac{L}{\lambda}, 1\}$, the bound becomes independent of the condition number (although it still depends on L).

Because $\bar{f}$ is q-uniformly convex, Theorem 2 also implies a bound on $\|\mathbf{y}_k - \bar{\mathbf{x}}^\star\|$ whenever $\lambda > 0$, as follows.

Corollary 1 *Under the same assumptions as in Theorem 2, and assuming, in addition, that $\lambda > 0$, we have that $\|\mathbf{y}_k - \bar{\mathbf{x}}^\star\| \le \bar{\epsilon}$ after at most*

$$k = O\left(\left(\frac{q}{\lambda\bar{\epsilon}^q}\right)^{\frac{q-\kappa}{q\kappa-q+\kappa}}\left(\frac{L^{\frac{q}{\kappa}}}{\lambda}\right)^{\frac{\kappa}{q\kappa-q+\kappa}}\log\left(\frac{qL\phi(\bar{\mathbf{x}}^\star)}{\bar{\epsilon}^q\lambda}\right)\right)$$

iterations.

Proof By q-uniform convexity of $\bar{f}$ and $\mathbf{0} \in \partial f(\bar{\mathbf{x}}^\star)$ (as $\bar{\mathbf{x}}^\star$ minimizes $\bar{f}$), we have

$$\|\mathbf{y}_k - \bar{\mathbf{x}}^\star\|^q \le \frac{q}{\lambda}(\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star)).$$

Thus, it suffices to apply the bound from Theorem 2 with the accuracy parameter $\epsilon = \frac{\lambda\bar{\epsilon}^q}{q}$. □

2.2 Computational considerations

At a first glance, the result from Theorem 2 may seem of limited applicability, as there are potentially four different parameters (L, κ, λ, q) that one would need to tune. However, we now argue that this is not a constraining factor. First, for most of the applications in which one would be interested in using this framework, function ψ is a regularizing function with known uniform convexity parameters λ and q (see Sect. 5 for several illustrative examples). Second, the knowledge of parameters L and κ is not necessary for our results; we presented the analysis assuming the knowledge of these parameters to not over-complicate the exposition.

In particular, the only place in the analysis where the (L, κ) smoothness of f is used is in the inequality

$$f(\mathbf{y}_k) \le f(\mathbf{x}_k) + \langle\nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle + \frac{M_k}{q}\|\mathbf{y}_k - \mathbf{x}_k\|^q + \frac{\delta_k}{2}. \tag{17}$$

But instead of explicitly computing the value of M_k based on L, κ, one could maintain an estimate of M_k, double it whenever the inequality from Eq. (17) is not satisfied, and recompute all iteration-k variables. This is a standard *line-search* trick employed in optimization (see, e.g., [52]). Observe that, due to (L, κ)-weak smoothness of f and Lemma 1, there exists a sufficiently large M_k for any value of δ_k. In particular, under the choice $\delta_k = \frac{a_k}{A_k}\epsilon$ from Theorem 3, the total number of times that M_k can get doubled is logarithmic in all of the problem parameters, which means that it can be absorbed in the overall convergence bound from Theorem 2.

Finally, the described algorithm (Generalized AGD+ from Eq. (7)) can be efficiently implemented only if the minimization problems defining $\mathbf{v}_k$ can be solved efficiently (preferably in closed form, or with $\tilde{O}(d)$ arithmetic operations). This is indeed the case for most problems of interest. In particular, when ψ is uniformly convex, we will typically take $\phi(\mathbf{u})$ to be the Bregman divergence $D_\psi(\mathbf{u}, \mathbf{x}_0)$. Then, the computation of $\mathbf{v}_k$ boils down to solving problems of the form (2), i.e., $\min_{\mathbf{u}\in\mathcal{X}}\{\langle \mathbf{z}, \mathbf{x}\rangle + \psi(\mathbf{x})\}$, for a given $\mathbf{z}$. Such problems are efficiently solvable whenever the convex conjugate of $\psi + I_\mathcal{X}$, where $I_\mathcal{X}$ is the indicator function of the closed convex set $\mathcal{X}$, is efficiently computable, in which case the minimizer is $\nabla(\psi + I_\mathcal{X})^*(\mathbf{z})$. In particular, for $\mathcal{X} = \mathbf{E}$ and $\psi(\mathbf{x}) = \frac{1}{q}\|\cdot\|^q$, $q > 1$, (a common choice for our applications of interest; see Sect. 5), the minimizer is computable in closed form as $\nabla\left(\frac{1}{q_*}\|\mathbf{z}\|_*^{q_*}\right)$, where $q_* = \frac{q}{q-1}$ is the exponent dual to q. This should be compared to the computation of proximal maps needed in [4, 51], where the minimizer would be the gradient of the infimal convolution of ψ and the squared norm, for which there are much fewer efficiently computable examples. Note that such an assumption would be sufficient for our algorithm to work in the Euclidean case (by taking $\phi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u} - \mathbf{x}_0\|_2^2$); however, it is not necessary.

3 Minimizing the gradient norm in ℓ_p and Sch_p spaces

We now show how to use the result from Theorem 2 to obtain near-optimal convergence bounds for minimizing the norm of the gradient. In particular, assuming that f is (L, κ)-weakly smooth w.r.t. $\|\cdot\|_p$, to obtain the desired results, we apply Theorem 2 to function $\bar{f}(\cdot) = f(\cdot) + \lambda\psi_p(\cdot)$, where

$$\psi_p(\mathbf{x}) = \begin{cases} \frac{1}{2(p-1)}\|\mathbf{x} - \mathbf{x}_0\|_p^2, & \text{if } p \in (1, 2], \\ \frac{1}{p}\|\mathbf{x} - \mathbf{x}_0\|_p^p, & \text{if } p \in (2, +\infty). \end{cases} \tag{18}$$

Function ψ_p is then $(1, \max\{2, p\})$-uniformly convex. The proof of strong convexity of ψ_p when $1 < p \leq 2$ can be found, e.g., in [10, Example 5.28]. For $2 < p < +\infty$, ψ_p is a separable function, hence its p-uniform convexity can be proved from the duality between uniform convexity and uniform smoothness [63] and direct computation. These ℓ_p results also have spectral analogues, given by the Schatten spaces $\mathscr{S}_p = (\mathbb{R}^{d\times d}, \|\cdot\|_{\mathscr{S},p})$. Here, the functions below can be proved to be $(1, \max\{2, p\})$-uniformly convex, which is a consequence of sharp estimates of uniform convexity

for Schatten spaces [7, 41]

$$\Psi_{\mathscr{S},p}(\mathbf{x}) = \begin{cases} \frac{1}{2(p-1)}\|\mathbf{x}-\mathbf{x}_0\|^2_{\mathscr{S},p}, & \text{if } p \in (1,2], \\ \frac{1}{p}\|\mathbf{x}-\mathbf{x}_0\|^p_{\mathscr{S},p}, & \text{if } p \in (2,+\infty). \end{cases} \tag{19}$$

Finally, both for ℓ_1 and $\mathscr{S}_1$ spaces, our algorithms can work on the equivalent norm with parameter $p = \ln d/(\ln d - 1)$. The cost of this change of norm is at most logarithmic in d for the diameter and strong convexity constants. Similarly, our results also extend to the case $p = \infty$, by similar considerations (here, using exponent $p = \ln d$).

To obtain the results for the norm of the gradient in ℓ_p spaces, we can apply Theorem 2 with $\phi(\mathbf{x}) = \psi_p(\mathbf{x})$, where ψ_p is specified in Eq. (18). The result is summarized in the following theorem. The same result can be obtained for $\mathscr{S}_p$ spaces, by following the same argument as in Theorem 3 below, which we omit for brevity.

Theorem 3 *Let f be a convex, (L,κ)- weakly smooth function w.r.t. a norm $\|\cdot\|_p$, where $p \in (1,\infty)$. Then, for any $\epsilon > 0$, Generalized AGD+ from Eq. (7), initialized at some point $\mathbf{x}_0 \in \mathbb{R}^d$ and applied to $\bar{f} = f + \lambda\psi_p$, where ψ_p is specified in Eq. (18),*

$$\lambda = \begin{cases} \frac{\epsilon(p-1)}{2\|\mathbf{x}^\star-\mathbf{x}_0\|_p}, & \text{if } p \in (1,2], \\ \frac{\epsilon}{2\|\mathbf{x}^\star-\mathbf{x}_0\|_p^{p-1}}, & \text{if } p \in (2,\infty), \end{cases}$$

and with the choice $\phi = \psi_p$, constructs a point $\mathbf{y}_k$ with $\|\nabla f(\mathbf{y}_k)\|_{p_*} \le \epsilon$ in at most

$$k = \begin{cases} O\left(\left(\frac{2L}{\epsilon}\right)^{\frac{\kappa}{(\kappa-1)(3\kappa-2)}}\left(\frac{\kappa^{2\kappa}}{(\kappa-1)^{2\kappa}}\frac{\|\mathbf{x}^\star-\mathbf{x}_0\|_p^2}{(p-1)^\kappa}\right)^{\frac{1}{3\kappa-2}}\log\left(\frac{L\|\mathbf{x}^\star-\mathbf{x}_0\|_p}{(p-1)\epsilon}\right)\right), & \text{if } p \in (1,2], \\ O\left(\left(\frac{2L\|\mathbf{x}^\star-\mathbf{x}_0\|_p}{\epsilon}\right)^{\frac{\kappa(p-1)}{p\kappa-p+\kappa}}\left(\frac{\kappa}{\kappa-1}\right)^{\frac{p}{p\kappa-p+\kappa}}\log\left(\frac{L\|\mathbf{x}^\star-\mathbf{x}_0\|_p^p}{\epsilon}\right)\right), & \text{if } p \in (2,\infty), \end{cases}$$

iterations. In particular, when $\kappa = 2$ (i.e., when f is L-smooth):

$$k = \begin{cases} \widetilde{O}\left(\sqrt{\frac{L\|\mathbf{x}^\star-\mathbf{x}_0\|_p}{\epsilon}}\right), & \text{if } p \in (1,2], \\ \widetilde{O}\left(\left(\frac{L\|\mathbf{x}^\star-\mathbf{x}_0\|_p}{\epsilon}\right)^{\frac{2(p-1)}{p+2}}\right), & \text{if } p \in (2,\infty), \end{cases}$$

where $\widetilde{O}$ hides logarithmic factors in L, $\|\mathbf{x}-\mathbf{x}_0\|_p$, $\frac{1}{p-1}$ and $1/\epsilon$.

Proof Let us first relate $\|\bar{\mathbf{x}}^\star - \mathbf{x}_0\|_p$ to $\|\mathbf{x}^\star - \mathbf{x}_0\|_p$, where $\bar{\mathbf{x}}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d}\bar{f}(\mathbf{x})$, $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. By the definition of $\bar{f}$:

$$\begin{aligned} 0 &\le \bar{f}(\mathbf{x}^\star) - \bar{f}(\bar{\mathbf{x}}^\star) \\ &= f(\mathbf{x}^\star) - f(\bar{\mathbf{x}}^\star) + \lambda\psi_p(\mathbf{x}^\star) - \lambda\psi_p(\bar{\mathbf{x}}^\star) \\ &\le \lambda\psi_p(\mathbf{x}^\star) - \lambda\psi_p(\bar{\mathbf{x}}^\star). \end{aligned}$$

It follows that

$$\psi_p(\bar{\mathbf{x}}^\star) \leq \psi_p(\mathbf{x}^\star).$$

Thus, using the definition of ψ_p,

$$\|\bar{\mathbf{x}}^\star - \mathbf{x}_0\|_p \leq \|\mathbf{x}^\star - \mathbf{x}_0\|_p. \tag{20}$$

By triangle inequality and $\bar{\mathbf{x}}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} \bar{f}(\mathbf{x})$ (which implies $\nabla \bar{f}(\mathbf{x}^\star) = \mathbf{0}$),

$$\begin{aligned}
\|\nabla f(\mathbf{y}_k)\|_{p_*} &\leq \|\nabla f(\mathbf{y}_k) - \nabla f(\bar{\mathbf{x}}^\star)\|_{p_*} + \|\nabla f(\bar{\mathbf{x}}^\star)\|_{p_*} \\
&= \|\nabla f(\mathbf{y}_k) - \nabla f(\bar{\mathbf{x}}^\star)\|_{p_*} + \|\nabla \bar{f}(\bar{\mathbf{x}}^\star) - \lambda \nabla \psi_p(\bar{\mathbf{x}}^\star)\|_{p_*} \\
&= \|\nabla f(\mathbf{y}_k) - \nabla f(\bar{\mathbf{x}}^\star)\|_{p_*} + \lambda \|\nabla \psi_p(\bar{\mathbf{x}}^\star)\|_{p_*}.
\end{aligned} \tag{21}$$

As f is convex and (L, κ) weakly smooth, using Lemma 2, we also have:

$$\begin{aligned}
\frac{\kappa - 1}{L^{\frac{1}{\kappa-1}}\kappa} \|\nabla f(\mathbf{y}_k) - \nabla f(\bar{\mathbf{x}}^\star)\|_{p_*}^{\frac{\kappa}{\kappa-1}} &\leq f(\mathbf{y}_k) - f(\bar{\mathbf{x}}^\star) - \left\langle \nabla f(\bar{\mathbf{x}}^\star), \mathbf{y}_k - \bar{\mathbf{x}}^\star \right\rangle \\
&= \bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) - \lambda \psi_p(\mathbf{y}_k) + \lambda \psi_p(\bar{\mathbf{x}}^\star) \\
&\quad - \left\langle \nabla \bar{f}(\bar{\mathbf{x}}^\star) - \lambda \nabla \psi_p(\bar{\mathbf{x}}^\star), \mathbf{y}_k - \bar{\mathbf{x}}^\star \right\rangle \\
&= \bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) - \lambda \Big(\psi_p(\mathbf{y}_k) - \psi_p(\bar{\mathbf{x}}^\star) \\
&\quad - \left\langle \nabla \psi_p(\bar{\mathbf{x}}^\star), \mathbf{y}_k - \bar{\mathbf{x}}^\star \right\rangle \Big) \\
&\leq \bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star),
\end{aligned} \tag{22}$$

where the second line uses $\bar{f} = f + \psi_p$, the third line follows by $\nabla \bar{f}(\bar{\mathbf{x}}^\star) = 0$ (as $\bar{\mathbf{x}}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} \bar{f}(\mathbf{x})$), and the last inequality is by convexity of ψ_p.

From Eqs. (21) and (22), to obtain $\|\nabla f(\mathbf{y}_k)\|_{p_*} \leq \epsilon$, it suffices that $\lambda \|\nabla \psi_p(\bar{\mathbf{x}}^\star)\|_{p_*} \leq \frac{\epsilon}{2}$ and $\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) \leq \left(\frac{\epsilon}{2}\right)^{\frac{\kappa}{\kappa-1}} \frac{\kappa-1}{L^{\frac{1}{\kappa-1}}\kappa}$.

The first condition determines the value of λ. Using Proposition 1, $\lambda \|\nabla \psi_p(\bar{\mathbf{x}}^\star)\|_{p_*} \leq \frac{\epsilon}{2}$ is equivalent to

$$\begin{cases} \frac{\lambda}{p-1} \|\bar{\mathbf{x}}^\star - \mathbf{x}_0\|_p \leq \frac{\epsilon}{2}, & \text{if } p \in (1, 2] \\ \lambda \|\bar{\mathbf{x}}^\star - \mathbf{x}_0\|_p^{p-1} \leq \frac{\epsilon}{2}, & \text{if } p \in (2, \infty). \end{cases}$$

Using Eq. (20), it suffices that:

$$\lambda = \begin{cases} \frac{\epsilon(p-1)}{2\|\mathbf{x}^\star - \mathbf{x}_0\|_p}, & \text{if } p \in (1, 2], \\ \frac{\epsilon}{2\|\mathbf{x}^\star - \mathbf{x}_0\|_p^{p-1}}, & \text{if } p \in (2, \infty). \end{cases} \tag{23}$$

Using the choice of λ from Eq. (23), it remains to apply Theorem 2 to bound the number of iterations until $\bar{f}(\mathbf{y}_k) - \bar{f}(\bar{\mathbf{x}}^\star) \leq \left(\frac{\epsilon}{2}\right)^{\frac{\kappa}{\kappa-1}} \frac{\kappa-1}{L^{\frac{1}{\kappa-1}}\kappa}$. Applying Theorem 2, we

have:

$$k = O\left(\left(\frac{2^{\frac{\kappa}{\kappa-1}} L^{\frac{1}{\kappa-1}} \kappa}{\epsilon^{\frac{\kappa}{\kappa-1}} (\kappa-1)}\right)^{\frac{q-\kappa}{q\kappa-q+\kappa}} \left(\frac{L^{\frac{q}{\kappa}}}{\lambda}\right)^{\frac{\kappa}{q\kappa-q+\kappa}} \log\left(\frac{2^{\frac{\kappa}{\kappa-1}} L^2 \kappa \psi_p(\bar{\mathbf{x}}^\star)}{\epsilon^{\frac{\kappa}{\kappa-1}} (\kappa-1)}\right)\right).$$

It remains to plug in the choice of λ from Eq. (23), $q = \max\{p, 2\}$, and simplify. □

Although the proof of Theorem 3 appears fairly simple, we now discuss why it is not obvious in light of similar existing results for the Euclidean norm [53]. In Euclidean settings, one uses an accelerated method to minimize the smooth and strongly convex objective $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$ for $\lambda = \Theta\left(\frac{\epsilon}{\|\mathbf{x}^* - \mathbf{x}_0\|_2}\right)$. Because the regularized function $\bar{f}$ in this setting is $(L+\lambda)$-smooth, we can conclude that the same algorithm ensures $\|\nabla \bar{f}(\mathbf{x})\|_2 \le \epsilon/2$ within $O(\sqrt{\frac{L}{\lambda}} \log(\frac{L\|\mathbf{x}^* - \mathbf{x}_0\|_2}{\epsilon}))$ iterations. To guarantee that $\|\nabla f(\mathbf{x})\|_2 \le \epsilon$, it then suffices to use the triangle inequality:

$$\|\nabla f(\mathbf{x})\|_2 \le \|\nabla \bar{f}(\mathbf{x})\|_2 + \lambda\|\mathbf{x} - \mathbf{x}_0\|_2. \tag{24}$$

This approach does not directly extend to ℓ_p norms (even when $\|\cdot\|_p^2$ is strongly convex, that is, when $p \in (1, 2)$), for the following reasons. First, the composite function $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_0\|_p^2$ is not smooth, so it is unclear how to directly argue that $\|\nabla \bar{f}(\mathbf{x})\|_{p^*}$ is small when $\bar{f}(\mathbf{x}) - \bar{f}(\bar{\mathbf{x}}^*)$ is small. This is the reason why our proof never uses an equivalent of (24) but instead applies the less obvious triangle inequality from (21) and then leverages the definitions of $\bar{f}$ and $\bar{\mathbf{x}}^*$. The second term that comes from applying either of the two triangle inequalities, $\|\nabla\left(\frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_0\|_p^2\right)\|_{p^*}$, can be bounded using our Proposition 1 and it is crucial here that the regularizer we use is of the form $\frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_0\|_p^2$; otherwise, for an alternative regularizer chosen as the Bregman divergence of a strongly convex function, we would need to use smoothness, which for p not trivially close to 2 would scale polynomially with the dimension, leading to polynomial in d scaling of the parameter λ (and consequently the convergence bound). Second, even efficiently minimizing the complementary composite objectives of this form, as discussed in the introduction, was not clear how to do prior to this work. Finally, it is crucial that in (22) we relate the norm of the gradient distance to the optimality gap w.r.t. $\bar{f}$ (as opposed to f); otherwise, to have $f(\mathbf{y}_k) - f(\mathbf{x}^*)$ that scales with ϵ^2, we would need $\lambda \propto \frac{\epsilon^2}{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}$, leading to a worse complexity bound resulting from the application of Theorem 2.

Remark 1 Observe that, as the gradient norm minimization relies on the application of Theorem 2, the knowledge of parameters L and κ is not needed, as discussed in Sect. 2.2. The only parameter that needs to be determined is λ, which cannot be known in advance, as it would require knowing the initial distance to optimum $\|\mathbf{x}^\star - \mathbf{x}_0\|$. However, tuning λ can be done at the cost of an additional $\log(\frac{\lambda}{\lambda_0})$ multiplicative factor in the convergence bound. In particular, one could start with a large estimate of λ (say, $\lambda = \lambda_0 = 1$), run the algorithm, and halt and restart with $\lambda \leftarrow \lambda/2$ each time $\|\nabla \bar{f}(\mathbf{y}_k)\|_* \le 2\epsilon$ but $\|\nabla f(\mathbf{y}_k)\|_* > \epsilon$. This condition is sufficient because, when λ is of the correct order, $\lambda\|\nabla\psi(\mathbf{y}_k)\|_* = O(\lambda\|\nabla\psi(\bar{\mathbf{x}}^\star)\|_*) = O(\epsilon)$, $\|\nabla f(\mathbf{y}_k)\|_* \le \epsilon$, and $\|\nabla \bar{f}(\mathbf{y}_k)\|_* \le \|\nabla f(\mathbf{y}_k)\|_* + \lambda\|\nabla\psi(\mathbf{y}_k)\|_* = O(\epsilon)$.

Remark 2 The approach conducted in this section is more broadly applicable; in fact, it is not even necessary to use the same norm to a power in order to regularize the objective. In Sect. 5 we explore an alternative regularization for minimizing the norm of the gradient, in the context of discrete optimal transport. This example also shows the importance of considering non-Euclidean norms for minimizing the norm of the gradient; particularly, the importance of quantifying the gradient error in the ℓ_1-norm, as this permits a dimension-independent error in a rounding procedure proposed in [6].

4 Lower bounds

In this section, we address the question of the optimality of our algorithmic framework, in a formal oracle model of computation. We first study the question of minimizing the norm of the gradient, which follows from a simple reduction to the complexity of minimizing the objective function and for which nearly tight lower bounds are known. In this case, the lower bounds show that our resulting algorithms are nearly optimal when $q = \kappa = 2$.

Regarding the complexity of minimizing the norm of the gradient, in cases where either we have weaker smoothness ($\kappa < 2$) or larger uniform convexity exponent ($q > 2$), we observe the presence of polynomial gaps in the complexity w.r.t. $1/\epsilon$. One natural question regarding the aforementioned gaps is whether this is due to a possible suboptimality of the algorithm used for complementary composite minimization (see Eq. (7)), or due to the approach of minimizing the norm of the gradient via the composite model itself. Here, we discard the first possibility, showing sharp lower bounds for complementary composite optimization in a new composite oracle model. Our lower bounds show that the complementary composite minimization algorithms are optimal up to factors which depend at most logarithmically on the initial distance to the optimal solution, the target accuracy, and dimension.

Before proceeding to the specific results, we provide a short summary of the classical oracle complexity in convex optimization and some techniques that will be necessary for our results. For more detailed information on the subject, we refer the reader to the thorough monograph [48]. In the oracle model of convex optimization, we consider a class of objectives $\mathcal{F}$, comprised of functions $f : \mathbf{E} \to \mathbb{R}$; an oracle $\mathcal{O} : \mathcal{F} \times \mathbf{E} \to \mathbf{F}$ (where $\mathbf{F}$ is a vector space); and a target accuracy, $\epsilon > 0$. An algorithm $\mathcal{A}$ can be described by a sequence of functions $(\mathcal{A}_k)_{k\in\mathbb{N}}$, where $\mathcal{A}_{k+1} : (\mathbf{E} \times \mathbf{F})^{k+1} \to \mathbf{E}$, so that the algorithm sequentially interacts with the oracle querying points

$$\mathbf{x}^{k+1} = \mathcal{A}_{k+1}(\mathbf{x}^0, \mathcal{O}(f, \mathbf{x}^0), \ldots, \mathbf{x}^k, \mathcal{O}(f, \mathbf{x}^k)).$$

The running time of algorithm $\mathcal{A}$ is given by the minimum number of queries to achieve some measure of accuracy (with accuracy parameter $\epsilon > 0$), and will be denoted by $T(\mathcal{A}, f, \epsilon)$. The most classical example in optimization is achieving additive optimality gap bounded by ϵ:

$$T(\mathcal{A}, f, \epsilon) = \inf\{k \geq 0 : f(\mathbf{x}^k) \leq f^\star + \epsilon\},$$

but other relevant goal for our work is achieving a (dual) norm of the gradient bounded above by ϵ, i.e.,

$$T(\mathcal{A},\, f, \epsilon) = \inf\{k \geq 0 : \|\nabla f(\mathbf{x}^k)\|_* \leq \epsilon\}.$$

Given a measure of efficiency T, the *worst-case oracle complexity* for a problem class $\mathcal{F}$ endowed with oracle $\mathcal{O}$, is given by

$$\text{Compl}(\mathcal{F}, \mathcal{O}, \epsilon) = \inf_{\mathcal{A}} \sup_{f \in \mathcal{F}} T(\mathcal{A},\, f, \epsilon).$$

4.1 Lower complexity bounds for minimizing the norm of the gradient

We now provide lower complexity bounds for minimizing the norm of the gradient. For the sake of simplicity, we can think of these lower bounds for the gradient oracle $\mathcal{O}(f, x) = \nabla f(\mathbf{x})$, but we point out they work more generally for arbitrary *local oracles* (more on this in the next subsection).

In short, we reduce the problem of making the gradient small to that of approximately minimizing the objective.

Proposition 2 *Let $f : \mathbf{E} \to \mathbb{R}$ be a convex and differentiable function, with a global minimizer $\mathbf{x}^*$. If $\|\nabla f(\mathbf{x})\|_* \leq \epsilon$ and $\|\mathbf{x} - \mathbf{x}^*\| \leq R$, then $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \epsilon R$.*

Proof By convexity of f,

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \leq \|\nabla f(\mathbf{x})\|_* \|\mathbf{x} - \mathbf{x}^*\| \leq \epsilon R,$$

where the second inequality is by duality of norms $\|\cdot\|$ and $\|\cdot\|_*$. □

For the classical problem of minimizing the objective function value, lower complexity bounds for ℓ_p-setups have been previously studied in both constrained [38] and unconstrained [27] settings. We summarize those results in the following theorem,[7]

Theorem 4 (From [27, 38]) *Let $1 \leq p \leq \infty$, and consider the problem class of unconstrained minimization with objectives in the class $\mathcal{F}_{\mathbb{R}^d, \|\cdot\|_p}(\kappa, L)$, whose minima are attained in $\mathcal{B}_{\|\cdot\|_p}(0, R)$. Then, the complexity of achieving additive optimality gap ϵ, for any local oracle, is bounded below by:*

- $\Omega\left(\left(\frac{LR^\kappa}{\epsilon[\ln d]^{\kappa-1}}\right)^{\frac{2}{3\kappa-2}}\right)$, *if* $1 \leq p < 2$;
- $\Omega\left(\left(\frac{LR^\kappa}{\epsilon \min\{p, \ln d\}^{\kappa-1}}\right)^{\frac{p}{\kappa p+\kappa-p}}\right)$, *if* $2 \leq p < \infty$; *and,*
- $\Omega\left(\left(\frac{LR^\kappa}{\epsilon[\ln d]^{\kappa-1}}\right)^{\frac{1}{\kappa-1}}\right)$, *if* $p = \infty$.

[7] More precisely, to obtain this result one can use the p-norm smoothing construction from [38, Section 2.3] in combination with the norm term used in [27, Eq. (3)]. This leads to a smooth objective over an unconstrained domain that provides a hard function class. We also remark for the interested reader that the lower bounds from Theorem 4 also apply to randomized algorithms, at least in the case $2 \leq p < \infty$ and with a more restrictive dimension regime. For details we refer to [27].

The dimension d for the lower bound to hold must be at least as large as the lower bound itself.

By combining the reduction from Proposition 2 with the lower bounds for function minimization from Theorem 4, we can now immediately obtain lower bounds for minimizing the (dual of the) ℓ_p norm of the gradient, as follows.

Corollary 2 *Let $1 \leq p \leq \infty$, and consider the problem class with objectives in $\mathcal{F}_{\mathbb{R}^d, \|\cdot\|_p}(\kappa, L)$, whose minima are attained in $\mathcal{B}_{\|\cdot\|_p}(0, R)$. Then, the complexity of achieving the dual norm of the gradient bounded by ϵ, for any local oracle, is bounded below by:*

- $\Omega\left(\left(\frac{LR^{\kappa-1}}{\epsilon[\ln d]^{\kappa-1}}\right)^{\frac{2}{3\kappa-2}}\right)$, *if* $1 \leq p < 2$;
- $\Omega\left(\left(\frac{LR^{\kappa-1}}{\epsilon \min\{p, \ln d\}^{\kappa-1}}\right)^{\frac{p}{\kappa p+\kappa-p}}\right)$, *if* $2 \leq p < \infty$; *and,*
- $\Omega\left(\left(\frac{LR^{\kappa-1}}{\epsilon[\ln d]^{\kappa-1}}\right)^{\frac{1}{\kappa-1}}\right)$, *if* $p = \infty$.

The dimension d for the lower bound to hold must be at least as large as the lower bound itself.

Comparing to the upper bounds from Theorem 3, it follows that for $p \in (1, 2]$ and $\kappa = 2$, our bound is optimal up to a $\log(d)\log(\frac{LR}{(p-1)\epsilon})$ factor; i.e., it is near-optimal. Recall that the upper bound for $p = 1$ can be obtained by applying the result from Theorem 3 with $p = \log(d)/[\log d - 1]$. When $p > 2$ and $\kappa = 2$, our upper bound is larger than the lower bound by a factor $\left(\frac{LR}{\epsilon}\right)^{\frac{p-2}{p+2}} \log\left(\frac{LR}{\epsilon}\right) (\min\{p, \log(d)\})^{\frac{p}{p+2}}$. The reason for the suboptimality in the $p > 2$ regime comes from the polynomial in $1/\epsilon$ factors in the upper bound for complementary composite minimization from Sect. 2, and it is a limitation of the regularization approach used in this work to obtain bounds for the norm of the gradient. In particular, we believe that it is not possible to obtain tighter bounds via an alternative analysis by using the same regularization approach. Thus, it is an interesting open problem to obtain tight bounds for $p > 2$, and it may require developing completely new techniques. Similar complexity gaps are encountered when $\kappa < 2$; however, it is reasonable to suspect that here the lower bounds are not sharp. In particular, when $\kappa = 1$, the objective function is not guaranteed to be differentiable (see the discussion below Definition 1 in Sect. 1.2) and points with small subgradients may be a zero measure set,[8] making them difficult (if not impossible) to attain. Notice that this is not reflected in the lower bound for $p < \infty$, while for $p = \infty$ the lower bound rules out oracle complexity guarantees of the form $\min\{d \log(1/\epsilon), \text{poly}(1/\epsilon)$ as $\kappa \to 1$ (suppressing the dependence on parameters other than d and ϵ), which are otherwise attainable for κ bounded away from one. Therefore, two interesting questions arise: first, how to strengthen these lower bounds for weakly smooth function classes; and second, can we study milder

[8] As a specific example, consider $f(\mathbf{x}) = \|\mathbf{x}\|_1$. Here, the norm of subgradient is constant and bounded away from zero for any $\mathbf{x} \neq \mathbf{0}$ and even at $\mathbf{x}^\star = \mathbf{0}$ the subgradient oracle could return any point from $[-1, 1]^d$.

accuracy measures in the weakly smooth case? For example, a recent line of work has focused on the task of approximating near stationary points, by use of proximal mappings [31, 32]. We leave these interesting open questions for future research.

4.2 Lower complexity bounds for complementary composite minimization

We investigate the (sub)optimality of the composite minimization algorithm in an oracle complexity model. To accurately reflect how our algorithms work (namely, using gradient information on the smooth term and regularized proximal subproblems w.r.t. the uniformly convex term), we introduce a new problem class and oracle for the complementary composite problem. We observe that existing constructions in the literature of lower bounds for nonsmooth uniformly convex optimization (e.g., [42, 60]) apply to our composite setting for $\kappa = 1$. The main idea of the lower bounds in this section is to combine these constructions with local smoothing, to obtain composite functions that match our assumptions.

Assumptions 5 Consider the problem class $\mathcal{P}(\mathcal{F}_{\|\cdot\|}(L, \kappa), \mathcal{U}_{\|\cdot\|}(\lambda, q), R)$, given by composite objective functions

$$(P_{f,\psi}) \quad \min_{\mathbf{x}\in\mathbf{E}}[\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})],$$

with the following assumptions:

(A.1) $f \in \mathcal{F}_{\|\cdot\|}(L, \kappa)$;
(A.2) $\psi \in \mathcal{U}_{\|\cdot\|}(\lambda, q)$; and,
(A.3) the optimal solution of $(P_{f,\psi})$ is attained within $\mathcal{B}_{\|\cdot\|}(0, R)$.

The problem class is additionally endowed with oracles $\mathcal{O}_{\mathcal{F}}$ and $\mathcal{O}_{\mathcal{U}}$, for function classes $\mathcal{F}_{\|\cdot\|}(L, \kappa)$ and $\mathcal{U}_{\|\cdot\|}(\lambda, q)$, respectively; which satisfy

(O.1) $\mathcal{O}_{\mathcal{F}}$ is a local oracle: if $f, g \in \mathcal{F}_{\|\cdot\|}(L, \kappa)$ are such that there exists $r > 0$ such that they coincide in a neighborhood $\mathcal{B}_{\|\cdot\|}(\mathbf{x}, r)$, then $\mathcal{O}_{\mathcal{F}}(\mathbf{x}, f) = \mathcal{O}_{\mathcal{F}}(\mathbf{x}, g)$; and,
(O.2) $\mathcal{U}_{\|\cdot\|}(\lambda, q)$ is any oracle (not necessarily local).

In brief, we are interested in the oracle complexity of achieving ϵ-optimality gap for the family of problems $(P_{f,\psi})$, where $f \in \mathcal{F}_{\|\cdot\|}(L, \kappa)$ is endowed with a local oracle, $\psi \in \mathcal{U}_{\|\cdot\|}(\lambda, q)$ is endowed with any oracle, and the optimal solution of problem $(P_{f,\psi})$ lies in $\mathcal{B}_{\|\cdot\|}(0, R)$. A simple observation is that in the case $\lambda = 0$, our model coincides with the classical oracle mode, which was discussed in the previous section. The goal now is to prove a more general lower complexity bound for the composite model.

Before proving the theorem, we first provide some building blocks in this construction, borrowed from past work [27, 38]. In particular, our lower bound works generally for q-uniformly convex and *locally smoothable* spaces.

Assumptions 6 Given the normed space $(\mathbf{E}, \|\cdot\|)$, we consider the following properties:

(B.1) $\psi(\mathbf{x}) = \frac{1}{q}\|\mathbf{x}\|^q$ is q-uniformly convex with constant $\bar{\lambda}$ w.r.t. $\|\cdot\|$.

(B.2) The space $(\mathbf{E}, \|\cdot\|)$ is $(\kappa, \eta, \eta, \bar{\mu})$-locally smoothable. That is, there exists a mapping $\mathcal{S} : \mathcal{F}_{(\mathbf{E},\|\cdot\|)}(0,1) \to \mathcal{F}_{(\mathbf{E},\|\cdot\|)}(\kappa, \overline{\mu})$ (denoted as the smoothing operator in [27, Definition 2]), such that $\|\mathcal{S}f - f\|_\infty \leq \eta$, and this operator preserves the equality of functions when they coincide in a ball of radius 2η; i.e., if $f|_{\mathcal{B}_{\|\cdot\|}(0,2\eta)} = g|_{\mathcal{B}_{\|\cdot\|}(0,2\eta)}$ then $\mathcal{S}f|_{\mathcal{B}_{\|\cdot\|}(0,\eta)} = \mathcal{S}g|_{\mathcal{B}_{\|\cdot\|}(0,\eta)}$.

(B.3) There exists $\Delta > 0$ and vectors $\mathbf{z}^1, \ldots, \mathbf{z}^M \in \mathbf{E}$ with $\|\mathbf{z}^i\|_* \leq 1$, such that for all $s_1, \ldots, s_M \in \{-1, +1\}^M$

$$\inf_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}_M} \left\| \sum_{i \in [M]} \alpha_i s_i \mathbf{z}^i \right\|_* \geq \Delta, \tag{25}$$

where $\boldsymbol{\Delta}_M = \{\boldsymbol{\alpha} \in \mathbb{R}^M_+ : \sum_i \alpha_i = 1\}$ is the discrete probability simplex in M-dimensions.

The three assumptions in Assumptions 6 are common in the literature, and can be intuitively understood as follows. Assumption (B.1) is the existence of a simple function that we can use as the uniformly convex term in the composite model. Assumption (B.2) appeared in [38], and provides a simple way to reduce the complexity of smooth convex optimization to its nonsmooth counterpart. We emphasize that there is a canonical way to construct smoothing operators, which is stated in Observation 7 below. Finally, Assumption (B.3) comes from the hardness constructions in nonsmooth convex optimization from [48], which are given by piecewise linear objectives that are learned one by one by an adversarial argument. The fact that the resulting piecewise linear function has a sufficiently negative optimal value (for any adversarial choice of signs) can be directly obtained by minimax duality from Eq. (25).

Note that d-dimensional ℓ_p spaces (denoted by ℓ_p^d) satisfy the assumptions above when $2 \leq p < \infty$.

Observation 7 (From [38]) *Let $2 \leq p < \infty$ and $\eta > 0$, and consider the space $\ell_p^d = (\mathbb{R}^d, \|\cdot\|_p)$. We now verify the Assumptions 6 for $q = p$, $\bar{\lambda} = 1$, $\bar{\mu} = 2^{2-\kappa}(\min\{p, \ln d\}/\eta)^{\kappa-1}$ and $\Delta = 1/M^{1/p}$. Indeed,*

- (B.1) *The p-uniform convexity of ψ was discussed after Eq. (18).*
- (B.2) *The smoothing operator can be obtained by infimal convolution, with kernel function $\phi(\mathbf{x}) = 2\|\mathbf{x}\|_r^2$ (with $r = \min\{p, 3\ln d\}$. We recall that the infimal convolution of two functions f and ϕ is given by*

$$(f \square \phi)(\mathbf{x}) = \inf_{h \in \mathcal{B}_p(0,1)} [f(\mathbf{x} + \mathbf{h}) + \phi(\mathbf{h})].$$

  The infimal convolution above can be adapted to obtain arbitrary uniform approximation to f and the preservation of equality of functions (see [38, Section 2.2] for details).
- (B.3) *Letting $\mathbf{z}^i = \mathbf{e}_i$, $i \in [M]$, be the first M canonical vectors, we have*

$$\left\| \sum_{i \in [M]} \alpha_i s_i \mathbf{z}^i \right\|_{p_*} = \|\boldsymbol{\alpha}\|_{p_*} \geq M^{1/p_* - 1}\|\boldsymbol{\alpha}\|_1 = M^{-1/p}.$$

This bound is achieved when $\alpha_i = 1/M$, for all i.

Before proving the result for ℓ_p-spaces, we provide a general lower complexity bound for the composite setting, which we later apply to derive the lower bounds for ℓ_p setups.

Lemma 5 *Let $(\mathbf{E}, \|\cdot\|)$ be a normed space that satisfies Assumptions 6 and let*

$$\mathcal{P}(\mathcal{F}_{\|\cdot\|}(L,\kappa), \mathcal{U}_{\|\cdot\|}(\lambda, q), R)$$

be a class of complementary composite problems that satisfies Assumptions 5. Suppose the following relations between parameters are satisfied:

(a) $2qL\bar{\lambda}/[\lambda\bar{\mu}] \le R^{q-1}$.
(b) $(M+3)\eta \le 4R$.
(c) $\frac{L}{4\bar{\mu}}(M+7)\eta \le \frac{1}{2q_*}\left(\frac{L\Delta}{\bar{\mu}}\right)^{q_*}\left(\frac{\bar{\lambda}}{\lambda}\right)^{\frac{1}{q-1}}$.

Then, the worst-case optimality gap for the problem class where algorithms are constrained to make M queries to oracles $\mathcal{O}_{\mathcal{F}}, \mathcal{U}_{\|\cdot\|}(\lambda, q)$, is bounded below by

$$\frac{1}{2q_*}\left(\frac{L\Delta}{\bar{\mu}}\right)^{q_*}\left(\frac{\bar{\lambda}}{\lambda}\right)^{\frac{1}{q-1}}.$$

Proof Given $M \in \mathbb{N}$, scalars $\delta_1, \dots, \delta_M > 0$, and $s_1, \dots, s_M \in \{-1, +1\}$, we consider the functions

$$f_s(\mathbf{x}) = \frac{L}{\bar{\mu}}\mathcal{S}\Big(\max_{i\in[M]}[\langle s_i\mathbf{z}^i, \cdot\rangle - \delta_i]\Big)(\mathbf{x}),$$

and $\bar{f}_s(\mathbf{x}) = f_s(\mathbf{x}) + (\lambda/\bar{\lambda})\psi(\mathbf{x})$, where ψ is given by Assumption (B.1).

We now show the composite objective $\bar{f}_s$ satisfies Assumption 5. Properties (A.1) and (A.2) are clearly satisfied. Regarding (A.3), we prove next that the optimum of these functions lies in $\mathcal{B}_{\|\cdot\|}(0, R)$. For this, notice that by Assumption (B.2):

$$\begin{aligned}\bar{f}_s(\mathbf{x}) &\ge \frac{L}{\bar{\mu}}\max_{i\in[M]}[\langle s_i\mathbf{z}^i, \mathbf{x}\rangle - \delta_i] - \frac{L\eta}{\bar{\mu}} + \frac{\lambda}{q\bar{\lambda}}\|\mathbf{x}\|^q \\ &\ge \|\mathbf{x}\|\left[\frac{\lambda}{\bar{\lambda}q}\|\mathbf{x}\|^{q-1} - \frac{L}{\bar{\mu}}\right] - \frac{L}{\bar{\mu}}(\eta + \max_i \delta_i).\end{aligned}$$

We will later show that $\eta + \max_i \delta_i \le (M+3)\eta/4 \le R$ (the last inequality by (b)), hence for $\|\mathbf{x}\| \ge R$

$$\bar{f}_s(\mathbf{x}) \ge \left(\frac{\lambda}{\bar{\lambda}q}\|\mathbf{x}\|^{q-1} - \frac{2L}{\bar{\mu}}\right)\|\mathbf{x}\| \ge 0,$$

where the last inequality follows from (a). To conclude the verification of Assumption (A.3), we now prove that $\min_{\mathbf{x}\in\mathbb{R}} \bar{f}_s(\mathbf{x}) < 0$. By Assumption (B.2):

$$\begin{aligned}
\inf_{\mathbf{x}\in\mathbf{E}} \bar{f}(\mathbf{x}) &\leq \inf_{\mathbf{x}\in\mathbf{E}} \Big(\frac{L}{\bar{\mu}} \max_{i\in[M]} [\langle s_i \mathbf{z}^i, x\rangle - \delta_i] + \frac{L}{\bar{\mu}}\eta + \frac{\lambda}{q\bar{\lambda}} \|\mathbf{x}\|^q \Big) \\
&= \max_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}_M} \inf_{x\in\mathbf{E}} \Big(\Big\langle \frac{L}{\bar{\mu}} \sum_{i\in[M]} \alpha_i s_i \mathbf{z}^i, x \Big\rangle + \frac{\lambda}{q\bar{\lambda}} \|\mathbf{x}\|^q - \frac{L}{\bar{\mu}} \sum_{i\in[M]} \alpha_i \delta_i + \frac{L}{\bar{\mu}}\eta \Big) \\
&= \max_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}_M} -\frac{1}{q_*} \Big(\frac{L}{\bar{\mu}} \Big)^{q_*} \Big(\frac{\bar{\lambda}}{\lambda} \Big)^{\frac{1}{q-1}} \Big\| \sum_{i\in[M]} \alpha_i s_i \mathbf{z}^i \Big\|_*^{q_*} - \frac{L}{\bar{\mu}} \sum_{i\in[M]} \alpha_i \delta_i + \frac{L}{\bar{\mu}}\eta \\
&= -\frac{1}{q_*} \Big(\frac{L}{\bar{\mu}} \Big)^{q_*} \Big(\frac{\bar{\lambda}}{\lambda} \Big)^{\frac{1}{q-1}} \Delta^{q_*} + \frac{L}{\bar{\mu}}\eta.
\end{aligned}$$

Notice that the second step above follows from the Sion Minimax Theorem [59]. We conclude that the optimal value of $(P_{f,\psi})$ is negative by (c).

Following the arguments provided in [38, Proposition 2], one can prove that for any algorithm interacting with oracle $\mathcal{O}_{\mathcal{F}}$, after M steps there exists a choice of $s_1, \ldots, s_M \in \{-1, +1\}^M$ such that

$$\min_{t\in[M]} f_s(\mathbf{x}^t) \geq \frac{L}{\bar{\mu}} [-\eta - \max_{i\in[M]} \delta_i];$$

further, for this adversarial argument it suffices that $\min_{i\in[M]} \delta_i = 0$, and $\max_{i\in[M]} \delta_i \geq (M-1)\eta/4$.

We conclude that the optimality gap after M steps is bounded below by

$$\begin{aligned}
\min_{t\in[M]} \bar{f}_s(\mathbf{x}^t) - \min_{\mathbf{x}\in\mathbf{E}} \bar{f}_s(\mathbf{x}) &\geq -\frac{L}{4\bar{\mu}}(M+7)\eta + \frac{1}{q_*} \Big(\frac{L}{\bar{\mu}} \Big)^{q_*} \Big(\frac{\bar{\lambda}}{\lambda} \Big)^{\frac{1}{q-1}} \Delta^{q_*} \\
&\geq \frac{1}{2q_*} \Big(\frac{L\Delta}{\bar{\mu}} \Big)^{q_*} \Big(\frac{\bar{\lambda}}{\lambda} \Big)^{\frac{1}{q-1}},
\end{aligned}$$

where we used the third bound from the statement. □

We now proceed to the lower bounds for ℓ_p-setups, with $2 \leq p \leq \infty$.

Theorem 8 *Consider the space $\ell_p^d = (\mathbb{R}^d, \|\cdot\|_p)$, where $2 \leq p < \infty$. Then, the oracle complexity of problem class $\mathcal{P} := \mathcal{P}(\mathcal{F}_{\|\cdot\|}(L,\kappa), \mathcal{U}_{\|\cdot\|}(\lambda, p), R)$, comprised of composite problems in the form $(P_{f,\psi})$ under Assumptions 5, is bounded below by*

$$\begin{cases}
\Big\lfloor \sqrt{\frac{L}{2\lambda}} - 7 \Big\rfloor, & \text{if } p = \kappa = 2,\ \epsilon < 2\sqrt{2\lambda L} R^2 \min\{\frac{2\lambda}{L}, 1\}, \\
\frac{C(p,\kappa)}{\min\{p, \ln d\}^{2(\kappa-1)}} \Big(\frac{L^p}{\lambda^\kappa \epsilon^{p-\kappa}} \Big)^{\frac{1}{\kappa p + \kappa - p}}, & \text{if } 1 \leq \kappa < p,\ p \in [2,\infty],\ \text{and } \lambda \geq \tilde{\lambda}.
\end{cases}$$

where $C(p,\kappa) := \left(\left(\frac{p-1}{p}\right)^{\kappa(p-1)} 2^{\frac{(p-\kappa)(1-2p)+(\kappa-1)p(2p-3)}{(p-1)}}\right)^{\frac{1}{\kappa p+\kappa-p}}$ *is bounded below by an absolute constant, and*

$$\begin{aligned}\tilde{\lambda} := \bar{C}\max\Bigg\{ & \min\{p,\ln d\}^3\left(\frac{\epsilon^\kappa}{LR}\right)^{\frac{1}{\kappa-1}}, \\ & \min\{p,\ln d\}^5\left(\frac{\epsilon^p}{L^{(p+1)}R^{\frac{(p-1)(\kappa p+\kappa-p)}{(\kappa-1)}}}\right)^{\frac{\kappa-1}{\kappa p+1-p}}\Bigg\},\end{aligned} \tag{26}$$

where $\bar{C} > 0$ *is a universal constant.*

In particular, our lower bounds show that the algorithm presented in the previous section—particularly the rates stated in Theorem 2—are nearly optimal. In the case $p = \kappa = 2$, the gap between upper and lower bounds is only given by a factor which grows at most logarithmically in $L\phi(\bar{\mathbf{x}}^*)/\epsilon$, and in the case $\kappa < p$, the gap is $O\big(\log(L\phi(\bar{\mathbf{x}}^*)/\epsilon)/\min\{p,\ln d\}^{\Theta(1)}\big)$. In both cases, the gaps are quite moderate, so the proposed algorithm is proved to be nearly optimal. Finally, we would also like to emphasize that the constant $C(p,\kappa) = \Theta(1)$, as a function of $1 < \kappa \le 2$ and $2 \le p \le \infty$. Therefore, the lower bounds also apply to the case $p = \infty$.

Proof (of Theorem 8) By Observation 7, in the case of ℓ_p^d, with $2 \le p < \infty$, Assumptions 6 are satisfied if $q = p$, $\Delta = 1/M^{1/p}$, $\overline{\lambda} = 1$, and $\bar{\mu} = 2^{2-\kappa}(\min\{p,\ln d\}/\eta)^{\kappa-1}$ (for given $\eta > 0$). This way, hypotheses (a), (b), (c) in Lemma 5 become

(a) $\eta \le \frac{\min\{p,\ln d\}}{2}\left(\frac{\lambda R^{p-1}}{pL}\right)^{\frac{1}{\kappa-1}}$.
(b) $(M+3)\eta \le 4R$.
(c) $\eta^{p-\kappa} \le \frac{2^{p+\kappa-3}L}{p_*^{(p-1)}\min\{p,\ln d\}^{(\kappa-1)}\lambda M(M+7)^{(p-1)}}$.

Case 1: $p = \kappa = 2$. In order to satisfy (c), it suffices to choose $M = \left\lfloor\sqrt{\frac{L}{2\lambda}} - 7\right\rfloor$.

Given such a choice, to satisfy (a), (b) of the lemma, we can choose

$$\eta = \min\left\{\frac{\lambda R}{2L}, \frac{4R}{M+3}\right\} \ge R\sqrt{\frac{2\lambda}{L}}\min\left\{\frac{1}{4}\sqrt{\frac{2\lambda}{L}}, 4\right\}.$$

Now, under the conditions imposed above, the lemma provides an optimality gap lower bound of

$$\frac{1}{4\lambda}\left(\frac{L\eta}{2\sqrt{M}}\right)^2 \ge 2\sqrt{2\lambda L}R^2\min\left\{\frac{2\lambda}{L}, 1\right\}.$$

In conclusion, if $\epsilon < 2\sqrt{2\lambda L}R^2\min\{2\lambda/L, 1\}$, then

$$\mathrm{Compl}(\mathcal{P}, (\mathcal{O}_\mathcal{F}, \mathcal{O}_\psi), \epsilon) \ge \left\lfloor\sqrt{\frac{L}{2\lambda}}\right\rfloor - 1.$$

Case 2: $p > \kappa$ (**where** $1 < \kappa \leq 2$, $2 \leq p < \infty$). Here, to ensure (a) and (b), it suffices that

$$\eta \leq \min\left\{\frac{4R}{M+3}, \frac{\min\{p, \ln d\}}{2}\left(\frac{\lambda R^{p-1}}{p}\right)^{\frac{1}{\kappa-1}}\right\}. \tag{27}$$

We later certify these conditions hold. On the other hand, for (c) it suffices to let

$$\eta = \left[\left(\frac{p-1}{p}\right)^{p-1} \frac{2^{p+\kappa-3}L}{\lambda \min\{p, \ln d\}^{\kappa-1} M(M+7)^{p-1}}\right]^{\frac{1}{p-\kappa}}.$$

Then by Lemma 5 the optimality gap is bounded below as

$$\frac{1}{2p_*}\left(\frac{L^p \eta^{p(\kappa-1)}}{2^{p(2-\kappa)}\lambda M \min\{p, \ln d\}^{p(\kappa-1)}}\right)^{\frac{1}{p-1}}$$
$$= \left[C(p,\kappa)^{\kappa p+\kappa-p} \cdot \frac{L^p}{\min\{p, \ln d\}^{\frac{p(\kappa-1)(\kappa p-2\kappa+1)}{p-1}} \lambda^\kappa (M+7)^{\kappa p+\kappa-p}}\right]^{\frac{1}{p-\kappa}},$$

where $C(p,\kappa) := \left(\left(\frac{p-1}{p}\right)^{\kappa(p-1)} 2^{\frac{(p-\kappa)(1-2p)+(\kappa-1)p(2p-3)}{(p-1)}}\right)^{\frac{1}{\kappa p+\kappa-p}}$. In particular, if ϵ is smaller than the gap above, resolving for M gives

$$\mathrm{Compl}(\mathcal{P}, (\mathcal{O}_\mathcal{F}, \mathcal{O}_\psi), \epsilon) \geq M = \frac{C(p,\kappa)}{\min\{p, \ln d\}^{2(\kappa-1)}}\left(\frac{L^p}{\lambda^\kappa \epsilon^{p-\kappa}}\right)^{\frac{1}{\kappa p+\kappa-p}}, \tag{28}$$

where we further simplified the bound, noting that $\frac{p(\kappa-1)(\kappa p-2\kappa+1)}{(p-1)(p\kappa+\kappa-p)} \leq 2(\kappa-1)$.

Now, given the chosen value of M, we will verify that Eq. (27) holds. For this, we note that Eq. (27) is implied by the following pair of inequalities

$$\lambda \geq C'(p,\kappa) \min\{p, \ln d\}^{(\kappa-1)(2\kappa-1)}\left(\frac{\epsilon^\kappa}{LR}\right)^{\frac{1}{\kappa-1}} \tag{29}$$

$$\lambda \geq C''(p,\kappa) \min\{p, \ln d\}^5 \left(\frac{\epsilon^p}{L^{(p+1)} R^{\frac{(p-1)(\kappa p+\kappa-p)}{(\kappa-1)}}}\right)^{\frac{\kappa-1}{\kappa p+1-p}} \tag{30}$$

with $C'(p,\kappa), C''(p,\kappa) \geq \bar{C} > 0$, are bounded below by a universal positive constant. Therefore, there exists a universal constant $\bar{C} > 0$ such that if λ satisfies Eqs. (29) and (30) where $C'(p,\kappa), C''(p,\kappa)$ are replaced by $\bar{C}$, then the lower complexity bound from Eq. (28) holds. □

Remark 3 Observe that the lower bounds from Theorem 8 apply only when λ is sufficiently large, which is consistent with the behavior of our algorithm, which for small values of λ obtains iteration complexity matching the classical smooth setting (as if we ignore the uniform convexity of the objective).

5 Applications

We now provide some illustrative applications of the results from Sects. 2 and 3 to different regression problems. In typical applications, the data matrix $\mathbf{A}$ is assumed to have fewer rows than columns, so that the system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b}$ is the vector of labels, is underdetermined, and one seeks a sparse solution $\mathbf{x}^\star$ that provides a good linear fit between the data and the labels.

5.1 Elastic net

One of the simplest applications of our framework is to the elastic net regularization, introduced by [65]. Elastic net regularized problems are of the form:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 + \lambda_1\|\mathbf{x}\|_1,$$

i.e., the elastic net regularization combines the lasso and ridge regularizers. Function f is assumed to be $(L, 2)$-weakly smooth (i.e., L-smooth) w.r.t. the Euclidean norm $\|\cdot\|_2$. It is typically chosen as either the linear least squares loss or the logistic loss.

We can apply results from Sect. 2 to this problem for $q = \kappa = 2$, choosing $\psi(\mathbf{x}) = \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 + \lambda_1\|\mathbf{x}\|_1$ and $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$. Observe that our algorithm only needs to solve subproblems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ \langle \mathbf{z}, \mathbf{x}\rangle + \frac{\lambda''}{2}\|\mathbf{x}\|_2^2 + \lambda'\|\mathbf{x}\|_1 \right\},$$

for fixed vectors $\mathbf{z} \in \mathbb{R}^d$ and fixed parameters λ', λ'', which is computationally inexpensive, as the problem under the min is separable.

Applying Theorem 2, the elastic net regularized problems can be solved to any accuracy $\epsilon > 0$ using

$$k = O\left(\min\left\{ \sqrt{\frac{L}{\lambda_2}} \log\left(\frac{L\|\mathbf{x}^\star - \mathbf{x}_0\|_2}{\epsilon} \right), \sqrt{\frac{L\|\mathbf{x}^\star - \mathbf{x}_0\|_2^2}{\epsilon}} \right\} \right)$$

iterations, where $\mathbf{x}^\star \in \mathbb{R}^d$ is the problem minimizer. We note in passing that an upper bound of the same order can be obtained by the composite accelerated method in [51].

5.2 Risk minimization with strongly convex ℓ_p-norm regularization

In this section, we argue that the results obtained in this paper are useful for solving certain regularized empirical risk problems and that the particular choice of regularizers proposed here leads to non-asymptotic consistency and generalization bounds. We emphasize that most of the theory used in what follows (particularly regarding regularization, stability and generalization) is classical (e.g., [14]), and we only provide the necessary tools for completeness.

Concretely, our main object of interest are population risk minimization problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x}), \quad \mathcal{L}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(\mathbf{x};\mathbf{z})], \tag{31}$$

where $\mathcal{D}$ is an unknown distribution, individual loss functions ℓ are convex, and we are given a set of n i.i.d. samples $\mathcal{S} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ from $\mathcal{D}$. For this problem to be (computationally and statistically) tractable, further assumptions are required, which we state in Assumption 9.

To address (31), we apply AGD+ to the following regularized empirical version of the problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathcal{L}_{\mathcal{S}}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_p^2, \tag{32}$$

where $\mathcal{L}_{\mathcal{S}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x};\mathbf{z}_i)$ is the empirical risk, λ is a regularization parameter (specified in the analysis below), and $p \in (1, 2]$. We let $\hat{\mathbf{x}}$ denote the output of AGD+, invoked with accuracy parameter $\epsilon_n > 0$, specified later in this subsection. Further, we let $\hat{\mathbf{x}}^*$ denote the unique[9] solution to (32). We further use $\mathcal{X}_{\mathcal{S}}^*$ to denote the set of minimizers of the empirical loss $\mathcal{L}_{\mathcal{S}}$ and denote the minimum value of the empirical loss $\mathcal{L}_{\mathcal{S}}$ by $\mathcal{L}_{\mathcal{S}}^*$.

First, we provide a justification for the regularization used in (32), from an optimization perspective, in the following proposition. In particular, we argue that the utilized regularization ensures that the solution output by AGD+ has ℓ_p-norm almost as small as the smallest norm among empirical minimizers, $\min_{\mathbf{x}\in\mathcal{X}_{\mathcal{S}}^*}\|\mathbf{x}\|_p$, while requiring only a modest amount of computation. When p is close to 1, we can interpret this property as enforcing sparsity of the output solution, similar to LASSO.

Proposition 3 *Given (32), assume that* $\mathcal{X}_{\mathcal{S}}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d}\mathcal{L}_{\mathcal{S}}(\mathbf{x})$ *is non-empty. Let* $\hat{\mathbf{x}}^*$ *denote the solution to (32). Then:*

$$\|\hat{\mathbf{x}}^*\|_p \le \min_{\mathbf{x}^*\in\mathcal{X}_{\mathcal{S}}^*}\|\mathbf{x}^*\|_p, \quad \textit{and}$$
$$\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*) - \mathcal{L}_{\mathcal{S}}^* \le \frac{\lambda}{2}\min_{\mathbf{x}^*\in\mathcal{X}_{\mathcal{S}}^*}\|\mathbf{x}^*\|_p^2.$$

As a consequence, if $\hat{\mathbf{x}}$ *is the solution output by AGD+ for optimality gap* ϵ_n*, then*

$$\|\hat{\mathbf{x}}\|_p \le \min_{\mathbf{x}^*\in\mathcal{X}_{\mathcal{S}}^*}\|\mathbf{x}^*\|_p + \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}, \quad \textit{and}$$
$$\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}) - \mathcal{L}_{\mathcal{S}}^* \le \frac{\lambda}{2}\min_{\mathbf{x}^*\in\mathcal{X}_{\mathcal{S}}^*}\|\mathbf{x}^*\|_p^2 + \epsilon_n.$$

[9] The solution to (32) is unique as the problem is strongly convex, due to the regularizer.

Proof The first inequality follows by (20), already proved within the proof of Theorem 3, as $\|\hat{\mathbf{x}}\|_p \le \|\mathbf{x}^*\|_p$ holds for any $\mathbf{x}^* \in \mathcal{X}_{\mathcal{S}}^*$. For the second inequality, using the assumption that $\hat{\mathbf{x}}^*$ is the solution to (32), we have, for all $\mathbf{x}^* \in \mathcal{X}_{\mathcal{S}}^*$,

$$\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*) + \frac{\lambda}{2}\|\hat{\mathbf{x}}^*\|_p^2 - \left(\mathcal{L}_{\mathcal{S}}(\mathbf{x}^*) + \frac{\lambda}{2}\|\mathbf{x}^*\|_p^2\right) \le 0. \tag{33}$$

Hence, the second inequality follows by rearranging (33), using that $\frac{\lambda}{2}\|\hat{\mathbf{x}}^*\|_p^2 \ge 0$, and taking the minimum of both sides over $\mathbf{x}^* \in \mathcal{X}_{\mathcal{S}}^*$.

For the third inequality, as $\mathcal{L}_{\mathcal{S}}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_p^2$ is $\lambda(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$ (as $\mathcal{L}_{\mathcal{S}}(\mathbf{x})$ is convex and $\frac{1}{2}\|\mathbf{x}\|_p^2$ is $(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$), $\hat{\mathbf{x}}^*$ minimizes $\mathcal{L}_{\mathcal{S}}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_p^2$, and $\hat{\mathbf{x}}$ is an ϵ_n-approximate solution to (32), we have

$$\frac{\lambda(p-1)}{2}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}^*\|_p^2 \le f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}}^*) \le \epsilon_n.$$

Hence, we have $\|\hat{\mathbf{x}} - \hat{\mathbf{x}}^*\|_p \le \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}$. The claimed inequality now follows using triangle inequality and the first part of the proposition, since

$$\|\hat{\mathbf{x}}\|_p \le \|\hat{\mathbf{x}}^*\|_p + \|\hat{\mathbf{x}} - \hat{\mathbf{x}}^*\|_p \le \min_{\mathbf{x}^* \in \mathcal{X}_{\mathcal{S}}^*} \|\mathbf{x}^*\|_p + \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}.$$

For the final part, let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}_{\mathcal{S}}^*} \|\mathbf{x}\|_p$. Then

$$\begin{aligned}
\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}) - \mathcal{L}_{\mathcal{S}}^* = \; & \mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}) + \frac{\lambda}{2}\|\hat{\mathbf{x}}\|_p^2 - \mathcal{L}_{\mathcal{S}}^* - \frac{\lambda}{2}\|\mathbf{x}^*\|_p^2 - \frac{\lambda}{2}\|\hat{\mathbf{x}}\|_p^2 + \frac{\lambda}{2}\|\mathbf{x}^*\|_p^2 \\
\le \; & \mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}) + \frac{\lambda}{2}\|\hat{\mathbf{x}}\|_p^2 - \mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*) - \frac{\lambda}{2}\|\hat{\mathbf{x}}^*\|_p^2 + \frac{\lambda}{2}\|\mathbf{x}^*\|_p^2 \\
\le \; & \epsilon_n + \frac{\lambda}{2}\|\mathbf{x}^*\|_p^2,
\end{aligned}$$

where the first inequality follows from $\hat{\mathbf{x}}^*$ being the minimizer of (32) and $\frac{\lambda}{2}\|\hat{\mathbf{x}}\|_p^2 \ge 0$, and the last inequality is by the definition of ϵ_n. □

Note that Proposition 3 allows us to treat the empirical problem (32) as if it were a constrained optimization problem, with constraint set $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_p \le \min_{\mathbf{x}^* \in \mathcal{X}_{\mathcal{S}}^*} \|\mathbf{x}^*\|_p + \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}\}$, as the predictor $\hat{\mathbf{x}}$ is guaranteed to lie in this set.

We now specify the assumptions used for obtaining statistical and computational guarantees.

Assumptions 9 Given the population risk minimization problem (31), the following all hold

(C.1) For any set $\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ of empirical samples, the set of empirical minimizers $\mathcal{X}_{\mathcal{S}}^*$ is non-empty and $\min_{\mathbf{x} \in \mathcal{X}_{\mathcal{S}}^*} \|\mathbf{x}\|_p \le B$, where $B < \infty$;

(C.2) The loss function ℓ is differentiable and satisfies that for any two samples $\mathbf{z}, \mathbf{z}'$ drawn from $\mathcal{D}$ and $\mathbf{x}$ such that $\|\mathbf{x}\|_p \leq B$, we have $\|\nabla \ell(\mathbf{x}; \mathbf{z}) - \nabla \ell(\mathbf{x}; \mathbf{z}')\|_* \leq M$, where $M < \infty$;
(C.3) The loss function ℓ is L-smooth for some $L < \infty$.
(C.4) For any $\mathbf{x}$ such that $\|\mathbf{x}\|_p \leq 2B$, $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[|\ell(\mathbf{x}; \mathbf{z}) - \mathcal{L}(\mathbf{x})|] \leq G$.

Assumption (C.1) ensures learnability of the population risk minimization problem, and some variant of it is necessary. It is usually enforced via a stronger condition that the minimization is performed over a bounded convex set. Assumption (C.2) is looser than the assumption about Lipschitz-continuity of ℓ that is typically enforced in the literature on stability and generalization. Further, due to Assumptions (C.1) and (C.3) and the criterion $\|\mathbf{x}\|_p \leq B$ in its statement, Assumption (C.2) holds as in this case $\|\nabla \ell(\mathbf{x}, \mathbf{z})\|_* \leq M$ is bounded (by $2LB$). Assumption (C.3) is made for computational tractability via the complexity bound of AGD+, and can be replaced with an assumption that ℓ is (κ, L)-weakly smooth for any $\kappa \in [1, 2]$, with all of the analysis still being applicable and just by invoking the appropriate complexity bound for this class (observe that constant M in Assumption (C.2) can still be bounded by $2LB^{\kappa-1}$ in this case). However, for concreteness and simplicity of exposition, we carry out the analysis under the assumption that ℓ is L-smooth. Finally, Assumption (C.4) is made to be able to apply [58, Theorem 8], which relates uniform replace one stability (as in Lemma 6) to consistency and generalization. Note that this assumption can be omitted, since it is automatically satisfied for any non-degenerate problem, as ℓ is a (Lipschitz) continuous function and as such bounded on the compact domain $\|\mathbf{x}\|_p \leq 2B$. In particular, by Assumption (C.1), ℓ has at least one minimizer $\mathbf{x}^*$ that belongs to the ℓ_p-ball $\|\mathbf{x}\|_p \leq B$. As ℓ is L-smooth (by Assumption (C.3)), $\sup_{\mathbf{x}:\|\mathbf{x}\|_p\leq 2B}\{\ell(\mathbf{x}) - \ell(\mathbf{x}^*)\} \leq \sup_{\mathbf{x}:\|\mathbf{x}\|_p\leq 2B} \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|_p^2 \leq \frac{9LB^2}{2}$. Thus, $G \leq \frac{9LB^2}{2}$.

The key property that allows us to prove consistency and generalization bounds is uniform replace one (RO) stability. For completeness, we first define uniform RO stability, consistency, and generalization, and then move on to proving the claimed bounds. The definitions provided below can be found in, e.g., [58].

Definition 4 Let A be a rule that given a sample $\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ and problem (31) outputs a predictor $A(\mathcal{S})$. A is said to be uniform-RO stable with rate $\epsilon_{\text{st}}(n)$, if for all possible sets $\mathcal{S}^{(i)} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_i', \mathbf{z}_{i+1}, \ldots \mathbf{z}_n\}$ that replace the i^{th} sample $\mathbf{z}_i$ by some $\mathbf{z}_i'$ and for any $\mathbf{z}'$ from the support of $\mathcal{D}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} |\ell(A(\mathcal{S}); \mathbf{z}') - \ell(A(\mathcal{S}^{(i)}); \mathbf{z}')| \leq \epsilon_{\text{st}}(n).$$

Definition 5 A learning rule A is said to be consistent with rate $\epsilon_{\text{cons}}(n)$ under distribution $\mathcal{D}$ if for all $n \geq 1$,

$$\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^n}[\mathcal{L}(A(\mathcal{S})) - \inf_{\mathbf{x}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x})] \leq \epsilon_{\text{cons}}(n).$$

A learning problem is learnable if there exists a learning rule A that is consistent with rate $\epsilon_{\text{st}}(n)$ under any distribution $\mathcal{D}$ and where $\epsilon_{\text{st}}(n) \overset{n\to\infty}{\longrightarrow} 0$. If it exists, such a rule is then called a universally consistent learning rule.

Definition 6 A rule A is said to generalize with rate $\epsilon_{\text{gen}}(n)$ under distribution $\mathcal{D}$ if for all $n \geq 1$,

$$\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^n}[|\mathcal{L}(A(\mathcal{S})) - \mathcal{L}_{\mathcal{S}}(A(\mathcal{S}))|] \leq \epsilon_{\text{gen}}(n).$$

A rule is said to universally generalize with rate $\epsilon_{\text{gen}}(n)$ if it generalizes with rate $\epsilon_{\text{gen}}(n)$ irrespective of the distribution over the given support.

To prove the consistency and generalization rates, we prove the following lemma that certifies uniform RO stability of the outputs of AGD+ (under a suitable choice of ϵ_n and λ). We then obtain the consistency and generalization rates as an application of [58, Theorem 8].

Lemma 6 *Given the population risk minimization problem (31), let $\hat{\mathbf{x}}_{\mathcal{S}}$ be an ϵ_n-approximate solution for the regularized empirical problem formulation in (32). Then $\hat{\mathbf{x}}_{\mathcal{S}}$ is a uniform RO stable learning rule with rate*

$$\epsilon_{\text{st}}(n) = L\left(2B + \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}\right)\left(\frac{M}{n\lambda(p-1)} + 2\sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}\right), \tag{34}$$

under any distribution that satisfies Assumption (C.1) and for any loss function ℓ that satisfies Assumptions (C.2) and (C.3).

Proof Let $\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \overset{i.i.d.}{\sim} \mathcal{D}$. Let $\mathcal{S}^{(i)}$ be any set obtained from $\mathcal{S}$ by replacing the i^{th} element $(\mathbf{z}_i)$ by an independent sample $\mathbf{z}_i' \sim \mathcal{D}$. Let $\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}$ be the minimum ℓ_p-norm solution to (32) with sample $\mathcal{S}^{(i)}$ and $\hat{\mathbf{x}}_{\mathcal{S}^{(i)}}$ be the approximate solution to the same problem output by AGD+. Similarly, let $\hat{\mathbf{x}}^*_{\mathcal{S}}$ be the minimum ℓ_p-norm solution to (32) with sample $\mathcal{S}$ and $\hat{\mathbf{x}}_{\mathcal{S}}$ be the approximate solution to the same problem output by AGD+. To prove the lemma, we first bound $\|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p$ and then use Proposition 3 with smoothness of ℓ to conclude that $\hat{\mathbf{x}}_{\mathcal{S}}$ is uniform RO stable.

As $\hat{\mathbf{x}}^*_{\mathcal{S}}$ minimizes $\mathcal{L}_{\mathcal{S}}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_p^2$, we have $\nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) + \nabla\left(\frac{1}{2}\|\hat{\mathbf{x}}^*_{\mathcal{S}}\|_p^2\right) = \mathbf{0}$, and, similarly, $\nabla\mathcal{L}_{\mathcal{S}^{(i)}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}) + \nabla\left(\frac{1}{2}\|\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p^2\right) = \mathbf{0}$. Further, as $\mathcal{L}_{\mathcal{S}}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_p^2$ is $\lambda(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$,

$$\begin{aligned}
&\left\langle \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) - \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle \\
&\quad\geq\ \lambda(p-1)\|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p^2 - \left\langle \nabla\left(\frac{1}{2}\|\hat{\mathbf{x}}^*_{\mathcal{S}}\|_p^2\right) - \nabla\left(\frac{1}{2}\|\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p^2\right), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle \\
&\quad=\ \lambda(p-1)\|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p^2 + \left\langle \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) - \nabla\mathcal{L}_{\mathcal{S}^{(i)}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle
\end{aligned} \tag{35}$$

Further, observe that by definition of $\mathcal{L}_{\mathcal{S}}$, we have $\nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) - \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}) = \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) - \nabla\mathcal{L}_{\mathcal{S}^{(i)}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}) + \frac{1}{n}(\nabla\ell(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}, \mathbf{z}'_i) - \nabla\ell(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}, \mathbf{z}_i))$. Hence:

$$\begin{aligned}
&\left\langle \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) - \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle \\
&\quad \le \left\langle \nabla\mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}^*_{\mathcal{S}}) - \nabla\mathcal{L}_{\mathcal{S}^{(i)}}(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle \\
&\qquad + \frac{1}{n}\left\langle \nabla\ell(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}, \mathbf{z}'_i) - \nabla\ell(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}, \mathbf{z}_i), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle
\end{aligned} \tag{36}$$

Hence, combining (35) and (36), we have

$$\begin{aligned}
\lambda(p-1)\|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p^2 &\le \frac{1}{n}\left\langle \nabla\ell(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}, \mathbf{z}'_i) - \nabla\ell(\hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}, \mathbf{z}_i), \hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}} \right\rangle \\
&\le \frac{M}{n}\|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p.
\end{aligned}$$

Thus, we conclude that

$$\|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p \le \frac{M}{n\lambda(p-1)},$$

and, consequently, using strong convexity and the guarantee of AGD+ as in the proof of Proposition 3, we have

$$\begin{aligned}
\|\hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}_{\mathcal{S}^{(i)}}\|_p &\le \|\hat{\mathbf{x}}^*_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p + \|\hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}^*_{\mathcal{S}}\|_p + \|\hat{\mathbf{x}}_{\mathcal{S}^{(i)}} - \hat{\mathbf{x}}^*_{\mathcal{S}^{(i)}}\|_p \\
&\le \frac{M}{n\lambda(p-1)} + 2\sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}.
\end{aligned} \tag{37}$$

Finally, let $\mathbf{x}^*_1$ be the minimizer of $\ell(\cdot; \mathbf{z}')$ with the minimum ℓ_p norm. By Assumption (C.1), $\|\mathbf{x}^*_1\|_p \le B$. Hence, using smoothness of ℓ, we have

$$\begin{aligned}
|\ell(\hat{\mathbf{x}}_{\mathcal{S}}; \mathbf{z}') - \ell(\hat{\mathbf{x}}_{\mathcal{S}^{(i)}}; \mathbf{z}')| &\le \left\langle \nabla\ell(\hat{\mathbf{x}}_{\mathcal{S}}; \mathbf{z}'), \hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}_{\mathcal{S}^{(i)}} \right\rangle \\
&\le \|\nabla\ell(\hat{\mathbf{x}}_{\mathcal{S}}; \mathbf{z}') - \nabla\ell(\mathbf{x}^*_1; \mathbf{z}')\|_{p^*}\|\hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}_{\mathcal{S}^{(i)}}\|_p \\
&\le L\|\hat{\mathbf{x}}_{\mathcal{S}} - \mathbf{x}^*_1\|_p\|\hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}_{\mathcal{S}^{(i)}}\|_p \\
&\le L(\|\hat{\mathbf{x}}_{\mathcal{S}}\|_p + \|\mathbf{x}^*_1\|_p)\|\hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}_{\mathcal{S}^{(i)}}\|_p \\
&\le L\Big(2B + \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}\Big)\|\hat{\mathbf{x}}_{\mathcal{S}} - \hat{\mathbf{x}}_{\mathcal{S}^{(i)}}\|_p \\
&\le L\Big(2B + \sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}\Big)\Big(\frac{M}{n\lambda(p-1)} + 2\sqrt{\frac{2\epsilon_n}{\lambda(p-1)}}\Big),
\end{aligned}$$

where the second to last inequality uses the third part of Proposition 3 and the last inequality uses (37). The quantity $-(\ell(\hat{\mathbf{x}}_{\mathcal{S}}; \mathbf{z}') - \ell(\hat{\mathbf{x}}_{\mathcal{S}^{(i)}}; \mathbf{z}'))$ can be bounded using a similar argument, which is omitted for brevity. □

We are now ready to state and prove computational and statistical guarantees for a learning rule $\hat{\mathbf{x}}_{\mathcal{S}}$ defined as an ϵ_n-approximate solution to (32), which can be computed using AGD+.

Theorem 10 *Given the population risk minimization problem (31) that satisfies Assumptions (C.1)–(C.4), let $\hat{\mathbf{x}}_{\mathcal{S}}$ be an ϵ_n-approximate solution for the regularized empirical problem formulation in (32), where ϵ_n and λ are defined by:*

$$\epsilon_n = \min\left\{\frac{B^2\lambda(p-1)}{2}, \frac{M^2}{8n^2\lambda(p-1)}\right\}, \quad \lambda = 2\sqrt{3}\sqrt{\frac{LM}{Bnp(p-1)}}. \tag{38}$$

Then $\hat{\mathbf{x}}_{\mathcal{S}}$ can be computed with

$$\begin{aligned} k &= O\left(\sqrt{\frac{L}{\lambda}}\log\left(\frac{LB}{\epsilon_n}\right)\right) \\ &= O\left(\left(\frac{nBL}{M}\right)^{1/4}\log\left(\frac{nBL}{p-1}\right)\right) \end{aligned} \tag{39}$$

iterations of AGD+ and it is consistent and generalizes under $\mathcal{D}$ with rates

$$\epsilon_{\text{cons}}(n) = 2\sqrt{3}\sqrt{\frac{p}{p-1}\cdot\frac{LM}{n}}B^{3/2}$$

$$\epsilon_{\text{gen}}(n) = 2\sqrt{3}\sqrt{\frac{p}{p-1}\cdot\frac{LM}{n}}B^{3/2} + \frac{2G}{\sqrt{n}}.$$

Proof The bound on the number of iterations of AGD+ required to compute $\hat{\mathbf{x}}_{\mathcal{S}}$ follows as a direct application of Theorem 2.

For the consistency and generalization rates, we apply [58, Theorem 8], by which

$$\begin{aligned} \epsilon_{\text{cons}}(n) &\le \epsilon_{\text{st}}(n) + \epsilon_{\text{erm}}(n), \\ \epsilon_{\text{gen}}(n) &\le \epsilon_{\text{st}}(n) + \epsilon_{\text{erm}}(n) + \frac{2G}{\sqrt{n}}, \end{aligned}$$

where

$$\epsilon_{\text{erm}}(n) := \mathcal{L}_{\mathcal{S}}(\hat{\mathbf{x}}) - \mathcal{L}^*_{\mathcal{S}} \le \frac{\lambda}{2}B^2 + \epsilon_n$$

(by Proposition 3) and $\epsilon_{\text{st}}(n)$ was bounded in Lemma 6. The claimed bounds now follow after plugging in the choice of λ, ϵ_n from the theorem statement, and simplifying. □

A few remarks are in order regarding the practical use of the proposed learning rule based on (32). In Theorem 10, we choose ϵ_n and λ according to the problem parameters

specified in Assumptions (C.1)-(C.4). However, parameters B, M, L are usually not available at the input. Further, if the minimum loss $\mathcal{L}_{\mathcal{S}}^*$ is not known, we cannot use $f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq \epsilon_n$ as a stopping criterion in AGD+. (Recall that, as discussed in Sect. 2.2, knowledge of L is not required for running AGD+, as L can be adaptively estimated.) In such situations, it suffices to let $\lambda = \Theta\Big(\sqrt{\frac{1}{np(p-1)}}\Big)$, $\epsilon_n = \Theta\big(\frac{p-1}{n^2}\big)$, and run AGD+ until the number of iterations reaches the estimate $C\sqrt{\frac{L}{\lambda}}\log(nL/(p-1))$ for some constant C and the adaptively estimated value of L. It is simple to verify that under this choice, $\epsilon_{\text{cons}}(n)$ and $\epsilon_{\text{gen}}(n)$ grow as $\frac{1}{\sqrt{n(p-1)}}$ as functions of n and p (albeit with a worse, though still polynomial, dependence on the remaining parameters).

Finally, when $p = 1 + \frac{c}{\log(d)}$ for $c > 0$, we have $\|\cdot\|_p \leq \|\cdot\|_1 \leq (1+c)\|\cdot\|_p$. Hence, the proposed regularized empirical risk minimization approach can be used as an alternative to LASSO under this choice of p, based on the discussion preceding Proposition 3. Note that, as shown in [58, Section 4.3], [36, Section 3.3], for general loss functions and under the assumptions from this section (Assumptions 9), the same consistency and generalization rates cannot be obtained using LASSO (regardless of whether it is formulated using ℓ_1 regularization or via bounding the optimization problem using ℓ_1 ball constraints).

5.3 Solving symmetric PSD linear systems with maximum constraint violation guarantee

When solving linear systems $\mathbf{A}\mathbf{x} = \mathbf{b}$, we are often interested in the maximum constraint violation as opposed to the ℓ_2 norm of the error vector $\mathbf{A}\mathbf{x} - \mathbf{b}$, $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$, obtained by minimizing the quadratic function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$. When $\mathbf{A}$ is symmetric and positive semidefinite (PSD), a common approach to solving linear systems is by minimizing the quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$. The gradient of this quadratic function is precisely the error vector $\mathbf{A}\mathbf{x} - \mathbf{b}$ for the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, thus in this case we are interested in minimizing the gradient of f.

If one uses a Euclidean first-order approach to minimize the gradient of f, then the resulting gradient oracle complexity to obtain $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \epsilon$ is

$$\Theta\Big(\min\Big\{d, \sqrt{\frac{\|\mathbf{A}\|_2\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\epsilon}}\Big\}\Big),$$

where $\mathbf{x}_0$ is an initial point and $\mathbf{x}^*$ is a solution to the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ [49, 50]. Note that $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty$ can be as large as $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ in the worst case. On the other hand, applying our result from Theorem 3 with $p = 1 + \frac{c}{\log(d)}$, $c > 0$, we get that $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty \leq \epsilon$ with gradient oracle complexity $\widetilde{O}\big(\sqrt{\frac{\|\mathbf{A}\|_{p\to p^*}\|\mathbf{x}_0 - \mathbf{x}^*\|_p}{c\epsilon}}\big) = \widetilde{O}\big(\sqrt{\frac{\max_{1\leq i,j\leq d} A_{ij}\|\mathbf{x}_0 - \mathbf{x}^*\|_1}{c\epsilon}}\big)$. If the system $\mathbf{A}\mathbf{x} - \mathbf{b}$ has sparse solutions, then selecting $\mathbf{x}_0 = \mathbf{0}$, the obtained bound can be smaller by a factor $\sqrt{d}$ for constant c, as $\|\mathbf{A}\|_2$ can be as large as $d\max_{1\leq i,j\leq d} A_{ij}$.[10] Further, as a consequence of the results from

[10] This bound is tight for the matrix of all ones.

Sect. 3, our algorithm enforces small ℓ_p-norm of the output solutions (and thus a small ℓ_1 norm). In particular, due to Eq. (20), when solving the regularized problem from Sect. 3 to obtain a solution with small gradient norm, it is guaranteed that $\|\bar{\mathbf{x}}\|_p^* \leq \|\mathbf{x}^*\|_p$, where $\bar{\mathbf{x}}^*$ is the solution to the regularized problem and $\mathbf{x}^*$ is any solution to the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$. As a consequence, $\|\bar{\mathbf{x}}\|_p^* \leq \min_{\mathbf{x}:\mathbf{A}\mathbf{x}=\mathbf{b}} \|\mathbf{x}\|_p$. Since the regularized problem is strongly convex, we further have that the output solution $\mathbf{y}_k$ satisfies $\|\mathbf{y}_k - \bar{\mathbf{x}}^*\|_p^2 \leq \frac{1}{2\lambda}(f(\mathbf{y}_k) - f(\bar{\mathbf{x}}^*))$. If we slightly change the target error of Generalized AGD+ in Theorem 3 to guarantee that $f(\mathbf{y}_k) - f(\bar{\mathbf{x}}^*) \leq \frac{(p-1)\epsilon^3}{2}$ (which only affects the terms under the log factor and does not change the resulting asymptotic complexity), we get that $\|\mathbf{y}_k - \bar{\mathbf{x}}^*\|_p \leq \epsilon$. As a consequence, using the triangle inequality, we have

$$\|\mathbf{y}_k\|_p \leq \|\mathbf{y}_k - \bar{\mathbf{x}}^*\|_p + \|\bar{\mathbf{x}}^*\|_p \leq (1+\epsilon)\|\bar{\mathbf{x}}^*\|_p \leq (1+\epsilon) \min_{\mathbf{x}:\mathbf{A}\mathbf{x}=\mathbf{b}} \|\mathbf{x}\|_p.$$

Finally, using properties of ℓ_p norms and our choice of p, we have that

$$\|\mathbf{y}_k\|_1 \leq \Big(1 + O(c+\epsilon)\Big) \min_{\mathbf{x}:\mathbf{A}\mathbf{x}=\mathbf{b}} \|\mathbf{x}\|_1.$$

5.4 ℓ_p regression

Standard ℓ_p-regression problems have as their goal finding a vector $\mathbf{x}^\star$ that minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$, where $p \geq 1$. When $p = 1$ or $p = \infty$, this problem can be solved using linear programming. More generally, when $p \notin \{1, \infty\}$, the problem is nonlinear, and multiple approaches have been developed for solving it, including, e.g., a homotopy-based solver [16], solvers based on iterative refinement [1, 3], and solvers based on the classical method of iteratively reweighted least squares [2, 34]. Such solvers typically rely on fast linear system solves and attain logarithmic dependence on the inverse accuracy $1/\epsilon$, at the cost of iteration count scaling polynomially with one of the dimensions of $\mathbf{A}$ (typically the lower dimension, which is equal to the number of rows m), each iteration requiring a constant number of linear system solves.

Here, we consider algorithmic setups in which the iteration count is dimension-independent and no linear system solves are required, but the dependence on $1/\epsilon$ is polynomial. First, for standard ℓ_p-regression problems, we can use a non-composite variant of the algorithm (with $\psi(\cdot) = 0$), while relying on the fact that the function $\frac{1}{q}\|\cdot\|_p^q$ with $q = \min\{2, p\}$ is $(1, p)$-weakly smooth for $p \in (1, 2)$ and $(p-1, 2)$-weakly smooth for $p \geq 2$. Using this fact, it follows that the function

$$f_p(\mathbf{x}) = \frac{1}{q}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^q$$

is (L_p, q)-weakly smooth w.r.t. $\|\cdot\|_p$, with $L_p = \max\{p-1, 1\}\|\mathbf{A}\|_{p\to p_*}^{q-1}$. On the other hand, function $\phi(\mathbf{x}) = \frac{1}{\bar{q}\min\{p-1,1\}}\|\mathbf{x} - \mathbf{x}_0\|_p^{\bar{q}}$, where $\bar{q} = \max\{2, p\}$ is $(1, \bar{q})$-uniformly convex w.r.t. $\|\cdot\|_p$. Thus, applying Theorem 2, we find that we can construct a point $\mathbf{y}_k \in \mathbb{R}^d$ such that $f_p(\mathbf{y}_k) - f_p(\mathbf{x}^\star) \leq \epsilon$, where $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f_p(\mathbf{x})$, with

at most

$$k = \begin{cases} O\left(\left(\frac{\|\mathbf{A}\|_{p\to p_*}^{p-1}}{\epsilon}\right)^{\frac{2}{3p-2}}\left(\frac{\|\mathbf{x}^\star-\mathbf{x}_0\|_p^2}{p-1}\right)^{\frac{p}{3p-2}}\right), & \text{if } p\in(1,2) \\ O\left(\left(\frac{(p-1)\|\mathbf{A}\|_{p\to p_*}}{\epsilon}\right)^{\frac{p}{p+2}}\left(\frac{\|\mathbf{x}^\star-\mathbf{x}_0\|_p^p}{p}\right)^{\frac{2}{p+2}}\right), & \text{if } p\geq 2 \end{cases}$$

iterations of Generalized AGD+. The same result can be obtained by applying the iteration complexity-optimal algorithms for smooth minimization over ℓ_p-spaces [25, 47].

More interesting for our framework is the ℓ_p regression on correlated errors, described in the following.

ℓ_p-regression on correlated errors. As argued in [17], there are multiple reasons why minimizing the correlated errors $\mathbf{A}^T(\mathbf{Ax}-\mathbf{b})$ in place of the standard errors $\mathbf{Ax}-\mathbf{b}$ is more meaningful for many applications. First, unlike standard errors, correlated errors are invariant to orthonormal transformations of the data. Indeed, if $\mathbf{U}$ is a matrix with orthonormal columns, then $(\mathbf{UA})^T(\mathbf{UAx}-\mathbf{Ub}) = \mathbf{A}^T(\mathbf{Ax}-\mathbf{b})$, but the same cannot be established for the standard error $\mathbf{Ax}-\mathbf{b}$. Other reasons involve ensuring that the model includes explanatory variables that are highly correlated with the data, which is only possible to argue when working with correlated errors (see [17] for more information).

Within our framework, minimization of correlated errors in ℓ_p-norms can be reduced to making the gradient small in the ℓ_p-norm; i.e., to applying results from Sect. 3. In particular, consider the function:

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2.$$

The gradient of this function is precisely the vector of correlated errors, i.e., $\nabla f(\mathbf{x}) = \mathbf{A}^T(\mathbf{Ax}-\mathbf{b})$. Further, function f is L_{p_*}-smooth w.r.t. $\|\cdot\|_{p_*}$, where $L_{p_*} = \|\mathbf{A}^T\mathbf{A}\|_{p_*\to p}$.

Applying the results from Theorem 3, it follows that, for any $\epsilon > 0$, we can construct a vector $\mathbf{y}_k \in \mathbb{R}^d$ with $\|\mathbf{A}^T(\mathbf{Ay}_k-\mathbf{b})\|_p \leq \epsilon$, where $\frac{1}{p}+\frac{1}{p_*}=1$, with at most

$$k = \begin{cases} \widetilde{O}\left(\left(\frac{\|\mathbf{A}^T\mathbf{A}\|_{p_*\to p}\|\mathbf{x}^\star-\mathbf{x}_0\|_{p_*}}{\epsilon}\right)^{\frac{2}{3p-2}}\right), & \text{if } p\in(1,2) \\ \widetilde{O}\left(\sqrt{\frac{\|\mathbf{A}^T\mathbf{A}\|_{p_*\to p}\|\mathbf{x}^\star-\mathbf{x}_0\|_{p_*}}{\epsilon}}\right), & \text{if } p>2 \end{cases}$$

iterations of generalized AGD+, where $\widetilde{O}$ hides a factor that is logarithmic in $1/\epsilon$ and where each iteration takes time linear in the number of non-zeros of $\mathbf{A}$. We are not aware of results of this type in the literature.

5.5 Spectral variants of regression problems

The algorithms we propose in this work are not limited to ℓ_p settings, but apply more generally to uniformly convex spaces. A notable example of such spaces are the *Schatten spaces*, $\mathscr{S}_p := (\mathbb{R}^{d\times d}, \|\cdot\|_{\mathscr{S},p})$, where $\|\mathbf{X}\|_{\mathscr{S},p} = (\sum_{j\in[d]} \sigma_j(\mathbf{X})^p)^{1/p}$, where $\sigma_1(\mathbf{X}), \ldots, \sigma_d(\mathbf{X})$ are the singular values of $\mathbf{X}$. In particular, the aforementioned ℓ_p-regression problems have their natural spectral counterparts, e.g., given a linear operator $\mathcal{A} : \mathbb{R}^{d\times d} \to \mathbb{R}^k$, and $\mathbf{b} \in \mathbb{R}^k$,

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times d}} \frac{1}{s}\|\mathcal{A}\mathbf{X} - \mathbf{b}\|_q^s + \frac{\lambda}{r}\|\mathbf{X}\|_{\mathscr{S},p}^r.$$

The most popular example of such a formulation comes from the nuclear norm relaxation for *low-rank matrix completion* [20, 54, 55]. We observe that the exact formulation of the problem may vary, but by virtue of Lagrangian relaxation we can interchangeably consider these different formulations as equivalent (modulo appropriate choice of regularization/constraint parameter choice).
To apply our algorithms to Schatten norm settings, we observe the functions below are $(1, r)$-uniformly convex, with $r = \max\{2, p\}$:

$$\Psi_{\mathscr{S},p}(\mathbf{X}) = \begin{cases} \frac{1}{2(p-1)}\|\mathbf{X}\|_{\mathscr{S},p}^2, & \text{if } p \in (1, 2], \\ \frac{1}{p}\|\mathbf{X}\|_{\mathscr{S},p}^p, & \text{if } p \in (2, +\infty). \end{cases}$$

On the other hand, notice that more generally than regression problems, for composite objectives

$$f(\mathbf{X}) + \lambda\Psi_{\mathscr{S},p}(\mathbf{X} - \mathbf{X}_0),$$

if the function f is unitarily invariant and convex, there is a well-known formula for its subdifferential, based on the subdifferential of its vector counterpart (there is a one-to-one correspondence between unitarily invariant functions $\mathbb{R}^{d\times d}$ and absolutely symmetric functions on $\mathbb{R}^d$) [44]. Even if f is not unitarily invariant, in the case of regression problems the gradients can be computed explicitly. On the other hand, the regularizer $\Psi_{\mathscr{S},p}$ admits efficiently computable solutions to problems from Eq. (2), given its unitary invariance (see, e.g., [10, Section 7.3.2]).
Iteration complexity bounds obtained with these regularizers are analogous to those obtained in the ℓ_p setting. On the other hand, the lower complexity bounds proved in Sect. 4 also apply to Schatten spaces by diagonal embedding from ℓ_p^d, hence all the optimality/suboptimality results established for ℓ_p carry over into $\mathscr{S}_p$.

5.6 Entropy-regularized optimal transport

Consider the entropic regularization [35] of the discrete optimal transport problem and its dual [24, 45]. Given a transport cost $\mathbf{C} \in \mathbb{R}_+^{m\times n}$, marginals $\boldsymbol{\mu} \in \Delta_m$, $\boldsymbol{\nu} \in \Delta_n$,[11] and regularization parameter $r > 0$, let

$$(P^r) \quad \min_{\mathbf{X}\in\mathbb{R}_+^{m\times n}} \{\langle \mathbf{C}, \mathbf{X}\rangle + r\langle \mathbf{X}, \ln(\mathbf{X}) - \mathbf{U}\rangle : \mathbf{X}\mathbf{1} = \boldsymbol{\mu},\ \mathbf{X}^\top\mathbf{1} = \boldsymbol{\nu},\ \langle \mathbf{U}, \mathbf{X}\rangle = 1\}$$

$$(D^r) \quad \min_{\mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^n} \varphi((\mathbf{u}, \mathbf{v})) := \Big\{ r \ln\Big(\sum_{i,j} \exp\big(\frac{1}{r}[u_i + v_j - c_{ij}]\big)\Big) - \langle \boldsymbol{\mu}, \mathbf{u}\rangle - \langle \boldsymbol{\nu}, \mathbf{v}\rangle \Big\}.$$

where $\mathbf{1}$ denotes the all-ones vector (of the corresponding dimension) and $\mathbf{U} \in \mathbb{R}^{m\times n}$ is the all-ones matrix, $\langle\cdot,\cdot\rangle$ applied to vectors is the standard inner product, which for matrices is the Frobenius inner product, and $\ln(\cdot)$ applied to a matrix denotes the component-wise application of the natural logarithm. Note that $\mathbf{u}, \mathbf{v}$ are the dual variables associated with the marginal constraints $\mathbf{X}\mathbf{1} = \boldsymbol{\mu}$ and $\mathbf{X}^\top\mathbf{1} = \boldsymbol{\nu}$, respectively. Further, we emphasize that the dual objective φ is L-smooth with respect to the $\|\cdot\|_\infty$ norm with $L = 1/r$ [10].

Observe that by denoting $\mathbf{B}(\mathbf{u}, \mathbf{v}) = \big(\exp([u_i + v_j - c_{ij}]/r)\big)_{i,j\in[n]}$ and $\mathbf{X}(\mathbf{u}, \mathbf{v}) = \mathbf{B}(\mathbf{u}, \mathbf{v})/\langle \mathbf{U}, \mathbf{B}(\mathbf{u}, \mathbf{v})\rangle$, then

$$\nabla\varphi((\mathbf{u}, \mathbf{v})) = \big(\mathbf{X}(\mathbf{u}, \mathbf{v})\mathbf{1} - \boldsymbol{\mu}\,,\ \mathbf{X}(\mathbf{u}, \mathbf{v})^\top\mathbf{1} - \boldsymbol{\nu}\big)^\top.$$

In particular, a global minimum of (D^r) induces a feasible solution for (P^r), and a dual solution with small (ℓ_1-norm of the) gradient implies a primal solution with small (ℓ_1-norm) infeasibility. The utility of finding such nearly feasible transports is related to the possibility of constructing (exactly) feasible transports with small additional transport cost.

Lemma 7 (From [6]) *There exists an algorithm that runs in time $O(mn)$ which takes as input $\mathbf{X} \in \Delta_{m\times n}$ (infeasible w.r.t. the marginal constraints), and produces a $\hat{\mathbf{X}} \in \Delta_{m\times n}$, which is feasible for the marginal constraints, and such that $\|\mathbf{X} - \hat{\mathbf{X}}\|_1 \le 2[\|\mathbf{X}\mathbf{1} - \boldsymbol{\mu}\|_1 + \|\mathbf{X}^\top\mathbf{1} - \boldsymbol{\nu}\|_1]$.*

We conclude that, in order to solve an (unregularized) optimal transport problem, it suffices to find a (regularized) dual solution with small norm of the gradient. More specifically, noticing that if $\mathbf{X}^r$ is an optimal solution for (P^r), then $\langle \mathbf{C}, \mathbf{X}^r\rangle - \langle \mathbf{C}, \mathbf{X}^0\rangle \le r\ln(mn)$ [23, 62], and therefore in order to obtain an ϵ optimal solution for (P^0) (the unregularized problem), it suffices to choose $r = \epsilon/[2\ln(mn)]$ and target accuracy for the norm of the gradient $\|\nabla\varphi((\mathbf{u}, \mathbf{v}))\|_1 \le \varepsilon/[4\|\mathbf{C}\|_\infty] =: \delta$, where $\|\mathbf{C}\|_\infty = \max_{i,j} |c_{ij}|$.

We now use the methods developed in previous sections to compute such a solution. For this purpose, we endow the space $\mathbb{R}^{m+n}$ with the ℓ_∞ norm, and we use the regularizer $\psi(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2$, which is 1-strongly convex w.r.t. ℓ_∞-norm. Notice that this

[11] Without loss of generality, we may assume that both probability distributions have full support, thus $\mu_i, \nu_j > 0$ for all i, j.

setting does not exactly coincide with that of Sect. 3, in particular since ψ is not the ℓ_∞ norm to some power. Nevertheless, the same rationale used in the aforementioned section shows that running AGD+ on the regularized objective $\bar{\varphi} := \varphi + \lambda\psi$ with λ set such that $\lambda\|\nabla\psi((\mathbf{u}^*, \mathbf{v}^*))\|_1 \leq \delta/2$ will provide the desired vector with ℓ_1 norm of the gradient δ with complexity $O\big(\sqrt{\frac{L}{\lambda}}\log\big(\frac{L\psi((\mathbf{u}^*,\mathbf{v}^*))}{\delta}\big)\big)$.

Our last task is then to provide an a-priori bound on the ℓ_∞-norm of the optimal dual solution. In this respect, multiple results can be found in the literature [18, 23, 33]. The following is particularly useful for our purposes.

Proposition 4 (Adapted from [45]) *There exists an optimal solution* $(\mathbf{u}^*, \mathbf{v}^*)$ *for* (D^r) *such that* $\|(\mathbf{u}^*, \mathbf{v}^*)\|_\infty \leq r[\ln\max\{\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}\} + \ln(mn)] + \|\mathbf{C}\|_\infty$, *where* $\bar{\boldsymbol{\mu}} = \max_i 1/\mu_i$, *and* $\bar{\boldsymbol{\nu}} = \max_j 1/\nu_j$.

Finally, the resulting oracle complexity to obtain accuracy δ in the norm of the gradient is given by

$$
\begin{aligned}
&O\left(\sqrt{\frac{L}{\lambda}}\log\Big(\frac{L\psi((\mathbf{u}^*,\mathbf{v}^*)}{\delta}\Big)\right)\\
&= O\left(\sqrt{\tfrac{2\|(\mathbf{u}^*,\mathbf{v}^*)\|_1}{r\delta}}\log\Big(\tfrac{\|\mathbf{C}\|_\infty\|(\mathbf{u}^*,\mathbf{v}^*)\|_2^2}{r\varepsilon}\Big)\right)\\
&= O\left(\sqrt{(m+n)\log(mn)}\ln\Big(\tfrac{(m+n)\|\mathbf{C}\|_\infty}{\varepsilon}\Big)\Big[\tfrac{\|\mathbf{C}\|_\infty}{\varepsilon} + \sqrt{\tfrac{\|\mathbf{C}\|_\infty\varepsilon\log\max\{\bar{\mu},\bar{\nu}\}}{\log(mn)}}\Big]\right).
\end{aligned}
$$

Noticing that each step of this method requires arithmetic complexity $O(mn)$, we finally obtain a total complexity of $\tilde{O}((m+n)^{5/2}\|C\|_\infty/\varepsilon)$, which matches the state of the art of the existing practically scalable methods (see discussions in [18]). The only existing method that obtains improved complexity is based on area convexity [40], which unfortunately is not competitive unless the dimension is extremely large. To summarize, this example shows how our methods can directly reproduce existing complexity bounds that have been obtained by arguably much more sophisticated and ad-hoc methods.

6 Conclusion and future work

We presented a general algorithmic framework for *complementary composite optimization*, where the objective function is the sum of two functions with complementary properties—(weak) smoothness and uniform/strong convexity. The framework has a number of interesting applications, including in making the gradient of a smooth function small in general norms and in different regression problems that frequently arise in machine learning. We also provided lower bounds that certify near-optimality of our algorithmic framework for the majority of standard ℓ_p and $\mathscr{S}_p$ setups.

Some challenging questions for future work remain. First, the regularization-based approach that we employed for gradient norm minimization leads to near-optimal oracle complexity bounds only when the objective function is smooth and the norm of the space is strongly convex (i.e., when the p_*-norm of the gradient is sought for

$p_* \geq 2$). The primary reason for this is that these are the only settings in which the complementary composite minimization leads to linear convergence. As the bounds we obtain for complementary composite minimization are near-tight, this represents a fundamental limitation of direct regularization-based approach. It is an open question whether the non-tight bounds for gradient norm minimization can be improved using some type of recursive regularization, as in [5]. Of course, there are clear challenges in trying to generalize such an approach to non-Euclidean norms, caused by the fundamental limitation that non-Euclidean norms cannot be simultaneously smooth and strongly convex with a dimension-independent condition number, as discussed at the beginning of the paper. Another interesting question is whether there exist direct (not regularization-based) algorithms for minimizing general gradient norms and that converge with (near-)optimal oracle complexity.

Acknowledgements The authors would like to thank Carlos Sing-Long and Adrien Taylor for valuable discussions and feedback on a first version of this paper. We would also like to thank Juan Pablo Contreras, Roberto Cominetti and Mario Bravo for insightful discussions on optimal transport and its entropic regularization.

A Impossibility of acceleration in the relatively smooth and relatively strongly convex setting

Below we exploit a simple reduction based on regularization to prove a lower bound on relatively smooth and relatively strongly convex optimization. The problem we will reduce to is relatively smooth convex optimization, for which tight lower complexity bounds are known [30].

Proposition 5 *The complexity of L-relatively smooth and μ-relative strongly convex minimization is bounded below by $\Omega\left(\frac{L}{\mu}\right)$.*

Proof Suppose that the oracle complexity of solving relatively smooth and relatively strongly convex functions is $o\left(\frac{L}{\mu}\right)$. Then, given a function f which is L relatively smooth w.r.t. h, consider the optimization problems

$$(P_0) \quad \min_x f(x),$$
$$(P_\lambda) \quad \min_x f(x) + \lambda h(x).$$

Note that (P_λ) is λ-relative strongly convex and $(L+\lambda)$-relatively smooth, both w.r.t. h. Proceeding as in the proof of Theorem 3, we have that $h(x^\lambda) \leq h(x^0)$, where x^0 and x^λ are optimal solutions for (P_0) and (P_λ), respectively. Hence, it suffices to solve (P_λ) to accuracy $\varepsilon/2$ and $\lambda = \varepsilon/[2\,h(x^0)]$, to obtain a solution with accuracy ε for problem (P_0).
Now, using an optimal algorithm for problem $(P_{\varepsilon/[2h(x^0)]})$, we have by our assumption that its oracle complexity is at most

$$o\left(\frac{L+\lambda}{\lambda}\right) = o\left(\frac{Lh(x^0)}{\varepsilon}\right),$$

a contradiction with the lower bound from [30]. □

References

1. Adil, D., Kyng, R., Peng, R., Sachdeva, S: Iterative refinement for ℓ_p-norm regression. In: Proc. ACM-SIAM SODA'19 (2019)
2. Adil, D., Peng, R., Sachdeva, S.: Fast, provably convergent IRLS algorithm for p-norm linear regression. In: Proc. NeurIPS'19 (2019)
3. Adil, D., Sachdeva, S.: Faster p-norm minimizing flows, via smoothed q-norm problems. In: Proc. ACM-SIAM SODA'20 (2020)
4. Allen-Zhu, Z.: Katyusha: the first direct acceleration of stochastic gradient methods. J. Mach. Learn. Res. **18**(1), 8194–8244 (2017)
5. Allen-Zhu, Z.: How to make the gradients small stochastically: even faster convex and nonconvex SGD. In: Proc. NeurIPS'18 (2018)
6. Altschuler, J., Niles-Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
7. Ball, K., Carlen, E.A., Lieb, E.H.: Sharp uniform convexity and smoothness inequalities for trace norms. Invent. Math. **115**(1), 463–482 (1994)
8. Bauschke, H.H., Bolte, J., Chen, J., Teboulle, M., Wang, X.: On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. J. Optim. Theory Appl. **182**(3), 1068–1087 (2019)
9. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017)
10. Beck, A.: First-Order Methods in Optimization. MOS-SIAM Series on Optimization, SIAM, New Delhi (2017)
11. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
12. Borwein, J., Guirao, A.J., Hájek, P., Vanderwerff, J.: Uniformly convex functions on Banach spaces. Proc. AMS **137**(3), 1081–1091 (2009)
13. Borwein, J.M., Zhu, Q.J.: Techniques of Variational Analysis. Springer, New York (2004)
14. Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Learn. Res. **2**(Mar), 499–526 (2002)
15. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
16. Bubeck, S., Cohen, M.B., Lee, Y.T., Li, Y.: An homotopy method for l_p regression provably beyond self-concordance and in input-sparsity time. In: Proc. ACM STOC'18 (2018)
17. Candés, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)
18. Chambolle, A., Contreras, J.P.: Accelerated Bregman primal-dual methods applied to optimal transport and Wasserstein Barycenter problems. SIAM J. Math. Data Sci. **4**, 1369–1395 (2022)
19. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
20. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
21. Cohen, M., Diakonikolas, J., Orecchia, L.: On acceleration with noise-corrupted gradients. In: Proc. ICML'18, pp. 1019–1028 (2018)
22. Cohen, M.B., Sidford, A., Tian, K.: Relative Lipschitzness in extragradient methods and a direct recipe for acceleration. In: 12th Innovations in Theoretical Computer Science Conference (ITCS 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2021)
23. Cominetti, R., Martín, J.S.: Asymptotic analysis of the exponential penalty trajectory in linear programming. Math. Program. **67**, 169–187 (1994)
24. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Burges, C.J.C., Bottou, L. Ghahramani, Z., Weinberger, K.Q. (eds) NIPS, pp. 2292–2300 (2013)
25. d'Aspremont, A., Guzmán, C., Jaggi, M.: Optimal affine-invariant smooth minimization algorithms. SIAM J. Optim. **28**(3), 2384–2405 (2018)
26. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
27. Diakonikolas, J., Guzmán, C.: Lower bounds for parallel and randomized convex optimization. J. Mach. Learn. Res. **21**(5), 1–31 (2020)

28. Diakonikolas, J., Orecchia, L.: Accelerated extra-gradient descent: a novel accelerated first-order method. In: 9th Innovations in Theoretical Computer Science Conference (ITCS 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2018)
29. Diakonikolas, J., Orecchia, L.: The approximate duality gap technique: a unified theory of first-order methods. SIAM J. Optim. **29**(1), 660–689 (2019)
30. Dragomir, R.-A., Taylor, A., d'Aspremont, A., Bolte, J.: Optimal complexity and certification of Bregman first-order methods. arXiv:1911.08510 (2019)
31. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. Math. Program. **185**(1–2), 357–383 (2021)
32. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Oper. Res. **43**(3), 919–948 (2018)
33. Dvurechensky, P., Gasnikov, A., Kroshnin, A.: Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In: International Conference on Machine Learning, pp. 1367–1376. PMLR (2018)
34. Ene, A., Vladu, A.: Improved convergence for ℓ_1 and ℓ_∞ regression via iteratively reweighted least squares. In: Proc. ICML'19 (2019)
35. Fang, S.-C.: An unconstrained convex programming view of linear programming. ZOR Methods Model. Oper. Res. **36**(2), 149–161 (1992)
36. Feldman, V.: Generalization of erm in stochastic convex optimization: The dimension strikes back. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
37. Gasnikov, A.V., Nesterov, Y.E.: Universal method for stochastic composite optimization problems. Comput. Math. Math. Phys. **58**(1), 48–64 (2018)
38. Guzmán, C., Nemirovski, A.: On lower complexity bounds for large-scale smooth convex optimization. J. Complex. **31**(1), 1–14 (2015)
39. He, N., Juditsky, A.B., Nemirovski, A.: Mirror Prox algorithm for multi-term composite minimization and semi-separable problems. Comput. Optim. Appl. **61**(2), 275–319 (2015)
40. Jambulapati, A., Sidford, A., Tian, K.: A direct tildeO(1/epsilon) iteration parallel algorithm for optimal transport. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp. 11355–11366 (2019)
41. Juditsky, A. Nemirovski, A.S.: Large deviations of vector-valued martingales in 2-smooth normed spaces. arXiv:0809.0813 (2008)
42. Juditsky, A., Nesterov, Y.: Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. Stoch. Syst. **4**(1), 44–80 (2014)
43. Kim, D., Fessler, J.A.: Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. J. Optim. Theory Appl. **188**, 192–219 (2020)
44. Lewis, A.S.: The convex analysis of unitarily invariant matrix functions. J. Convex Anal. **2**(1/2), 173–183 (1995)
45. Lin, T., Ho, N., Jordan, M.I.: On the efficiency of entropic regularized algorithms for optimal transport. J. Mach. Learn. Res. **23**(137), 1–42 (2022)
46. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. **28**(1), 333–354 (2018)
47. Nemirovskii, A.S., Nesterov, Y.E.: Optimal methods of smooth convex minimization. USSR Comput. Math. Math. Phys. **25**(2), 21–30 (1985)
48. Nemirovskii, A.S., Yudin: Problem Complexity and Method Efficiency in Optimization. Wiley, Hoboken (1983)
49. Nemirovsky, A.S.: On optimality of krylov's information when solving linear operator equations. J. Complex. **7**(2), 121–130 (1991)
50. Nemirovsky, A.S.: Information-based complexity of linear operator equations. J. Complex. **8**(2), 153–175 (1992)
51. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
52. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. **152**(1–2), 381–404 (2015)
53. Nesterov, Y.: How to make the gradients small. Optima Math. Optim. Soc. Newsl. **88**, 10–11 (2012)

54. Nesterov, Y., Nemirovski, A.: On first-order algorithms for ℓ_1/nuclear norm minimization. Acta Numer. **22**, 509 (2013)
55. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
56. Tyrrell Rockafellar, R.: Convex Analysis. Princeton Mathematical Series. Princeton University Press, Princeton (1970)
57. Scheinberg, K., Goldfarb, D., Bai, X.: Fast first-order methods for composite convex optimization with backtracking. Found. Comput. Math. **14**(3), 389–417 (2014)
58. Shalev-Shwartz, S., Shamir, O., Srebro, N., Sridharan, K.: Learnability, stability and uniform convergence. J. Mach. Learn. Res. **11**, 2635–2670 (2010)
59. Sion, M.: On general minimax theorems. Pac. J. Math. **8**(1), 171–176 (1958)
60. Srebro, N., Sridharan, K.: On convex optimization, fat shattering and learning. unpublished note (2012)
61. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Manuscript, 1 (2008)
62. Weed, J.: An explicit analysis of the entropic penalty in linear programming. In: Bubeck, S., Perchet, V., Rigollet, P. (eds.) Conference on Learning Theory, COLT 2018, Stockholm, Sweden, 6–9 July 2018, Volume 75 of Proceedings of Machine Learning Research, pp. 1841–1855 (2018)
63. Zalinescu, C.: On uniformly convex functions. J. Math. Anal. Appl. **95**, 344–374 (1983)
64. Zalinescu, C.: Convex Analysis in General Vector Spaces. World Scientific, Singapore (2002)
65. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**(2), 301–320 (2005)