CrossMark

FULL LENGTH PAPER

Oracle complexity of second-order methods for smooth convex optimization

Yossi Arjevani[1] · **Ohad Shamir**[1] · **Ron Shiff**[1]

Abstract Second-order methods, which utilize gradients as well as Hessians to optimize a given function, are of major importance in mathematical optimization. In this work, we prove tight bounds on the oracle complexity of such methods for smooth convex functions, or equivalently, the worst-case number of iterations required to optimize such functions to a given accuracy. In particular, these bounds indicate when such methods can or cannot improve on gradient-based methods, whose oracle complexity is much better understood. We also provide generalizations of our results to higher-order methods.

Keywords Smooth convex optimization · Oracle complexity

Mathematics Subject Classification 90C25 · 65K05 · 49M37

1 Introduction

We consider an unconstrained optimization problem of the form

$$\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}), \tag{1}$$

✉ Ron Shiff
ron.shiff1@gmail.com

Yossi Arjevani
yossi.arjevani@weizmann.ac.il

Ohad Shamir
ohad.shamir@weizmann.ac.il

[1] Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel

where f is a generic smooth and convex function. A natural and fundamental question is how efficiently can we optimize such functions.

We study this question through the well-known framework of oracle complexity [13], which focuses on iterative methods relying on local information. Specifically, it is assumed that the algorithm's access to the function f is limited to an oracle, which given a point $\mathbf{w}$, returns the values and derivatives of the function f at $\mathbf{w}$. This naturally models standard optimization approaches to unstructured problems such as (1), and allows one to study their efficiency, by bounding the number of oracle calls required to reach a given optimization error. Different classes of methods can be distinguished by the type of oracle they use. For example, gradient-based methods (such as gradient descent or accelerated gradient descent) rely on a first-order oracle, which returns gradients, whereas methods such as the Newton method rely on a second-order oracle, which returns gradients as well as Hessians.

The theory of *first-order* oracle complexity is quite well developed [12,13,15]. For example, if the dimension is unrestricted, f in (1) has μ_1-Lipschitz gradients, and the algorithm makes its first oracle query at a point $\mathbf{w}_1$, then the worst-case number of queries T required to attain a point $\mathbf{w}_T$ satisfying $f(\mathbf{w}_T) - \min_\mathbf{w} f(\mathbf{w}) \leq \epsilon$ is

$$\Theta\left(\sqrt{\frac{\mu_1 D^2}{\epsilon}}\right), \tag{2}$$

where D is an upper bound on the distance between $\mathbf{w}_1$ and the nearest minimizer of f. Moreover, if the function f is also λ-strongly convex for some $\lambda > 0$[1], then the oracle complexity bound is

$$\Theta\left(\sqrt{\frac{\mu_1}{\lambda}} \cdot \log\left(\frac{\mu_1 D^2}{\epsilon}\right)\right). \tag{3}$$

Both bounds are achievable using accelerated gradient descent [14].

However, these bounds do not capture the attainable performance of *second-order* methods, which rely on gradient as well as Hessian information. This is a central class of optimization methods, including the well-known Newton method and its many variants. Clearly, since these methods rely on Hessians as well as gradients, their oracle complexity can only be better than first-order methods. On the flip side, the per-iteration computational complexity is generally higher, in order to process the additional Hessian information (especially in high-dimensional problems where the Hessian matrix may be very large). Thus, it is natural to ask how much does this added per-iteration complexity pay off in terms of oracle complexity.

To answer this question, one needs good oracle complexity lower bounds for second-order methods, which establish the limits of attainable performance using any such algorithm. Perhaps surprisingly, such results do not seem to currently exist in the literature, and clarifying the oracle complexity of such methods was posed as an important open question (see for example [16]). The goal of this paper is to address this gap.

---

[1] Assuming f is twice-differentiable, this corresponds to $\nabla^2 f(\mathbf{w}) \succeq \lambda I$ for all $\mathbf{w}$.

Specifically, we prove that when the dimension is sufficiently large, for the class of convex functions with μ_1-Lipschitz gradients and μ_2-Lipschitz Hessians, the worst-case oracle complexity of any deterministic algorithm is

$$\Omega\left(\min\left\{\sqrt{\frac{\mu_1 D^2}{\epsilon}}, \left(\frac{\mu_2 D^3}{\epsilon}\right)^{2/7}\right\}\right). \tag{4}$$

This bound is tight up to constants, as it is matched by a combination of existing methods in the literature (see discussion below). Moreover, if we restrict ourselves to functions which are λ-strongly convex, we prove an oracle complexity lower bound of

$$\Omega\left(\left(\min\left\{\sqrt{\frac{\mu_1}{\lambda}}, \left(\frac{\mu_2}{\lambda}D\right)^{2/7}\right\} + \log\log_{18}\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right)\right). \tag{5}$$

Moreover, we establish that this bound is tight up to logarithmic factors (independent of ϵ), utilizing a novel adaptation of the A-NPE algorithm proposed in [10] (see "Appendix A"). These new lower bounds have several implications:

- Perhaps unexpectedly, (5) establishes that one cannot avoid in general a polynomial dependence on geometry-dependent "condition numbers" of the form μ_1/λ or $\mu_2 D/\lambda$, even with second-order methods. This is despite the ability of such methods to favorably alter the geometry of the problem (for example, the Newton method is well-known to be affine invariant).
- To improve on the oracle complexity of first-order methods for strongly-convex problems ((3)) by more than logarithmic factors, one cannot avoid a polynomial dependence on the initial distance D to the optimum. This is despite the fact that the dependence on D with first-order methods is only logarithmic. In fact, when D is sufficiently large (of order $\frac{\mu_1^{7/4}}{\mu_2\lambda^{3/4}}$ or larger), second-order methods cannot improve on the oracle complexity of first-order methods by more than logarithmic factors.
- In the convex case, second-order methods are again no better than first-order methods in certain parameter regimes (i.e., when $\mu_2 \geq \mu_1^{7/4}\sqrt{D}/\epsilon^{3/4}$), despite the availability of more information.

Finally, we show how our proof techniques can be generalized, to establish lower bounds for methods employing higher-order derivatives. In particular, for methods using all derivatives up to order k, we show that for convex functions with μ_k-Lipschitz k-th order derivatives, the oracle complexity is

$$\Omega\left(\left(\frac{\mu_k D^{k+1}}{(k+1)!k\epsilon}\right)^{2/(3k+1)}\right).$$

This directly generalizes (2) for $k = 1$, and (4) when $k = 2$ and μ_1 is unrestricted.

We note that in this paper, we focus on deterministic algorithms, in view of the fact that all state-of-the-art algorithms for our problem are deterministic, and following most existing literature on oracle complexity lower bounds. However, we believe our

results can be readily extended to randomized algorithms as well (see Remark 1 for more details).

1.1 Related work

Below, we review some pertinent results in the context of second-order methods. Related results in the context of k-th order methods are discussed in Sect. 2.2.

Perhaps the most well-known and fundamental second-order method is the Newton method, which relies on iterations of the form $\mathbf{w}_{t+1} = \mathbf{w}_t - (\nabla^2 f(\mathbf{w}))^{-1} \nabla f(\mathbf{w})$ (see e.g., [6]). It is well-known that this method exhibits *local* quadratic convergence, in the sense that if f is strictly convex, and the method is initialized close enough to the optimum $\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w})$, then $\mathcal{O}(\log\log(1/\epsilon))$ iterations suffice to reach a solution $\mathbf{w}$ such that $f(\mathbf{w}) - f(\mathbf{w}^*) \leq \epsilon$. However, in order to get global convergence (starting from an arbitrary point not necessarily close to the optimum), one needs to make some algorithmic modifications. For the Newton method with a line search, the number of iterations can be upper bounded by $\mathcal{O}\left(\frac{\mu_1^2 \mu_2^2}{\lambda^5}(f(\mathbf{w}_1) - f(\mathbf{w}^*)) + \log\log_2\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right)$, where μ_1, μ_2 are the Lipschitz parameters of the gradients and Hessians respectively, and assuming the function is λ-strongly convex [9]. Note that the first term captures the initial phase required to get sufficiently close to $\mathbf{w}^*$, and is easily the dominant term in the bound (unless ϵ is exceedingly small). If f is self-concordant [2], this can be improved to $\mathcal{O}\left((f(\mathbf{w}_1) - f(\mathbf{w}^*)) + \log\log_2\left(\frac{1}{\epsilon}\right)\right)$, independent of the strong convexity and Lipschitz parameters ([17]). Unfortunately, not all practically relevant objective functions are self-concordant (or at least cannot be made self-concordant without strongly affecting other problem parameters).

Returning to our setting of generic convex and smooth functions, and focusing on strongly convex functions for now, the best existing upper bounds (we are aware of) were obtained for cubic-regularized variants of the Newton method [8,16,18]. The existing analysis (see section 6 of [16]) implies an oracle complexity bound of at most $\mathcal{O}\left(\left(\frac{\mu_2}{\lambda}D\right)^{1/3} + \log\log_2\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right)$, where $D = \|\mathbf{w}_1 - \mathbf{w}^*\|$ is the distance from the initialization point $\mathbf{w}_1$ to the optimum $\mathbf{w}^*$. However, as we show in "Appendix A", a better oracle complexity bound can be obtained, by adapting the A-NPE method proposed in [10] and analyzed for convex functions, to the strongly convex case. The resulting complexity upper bound is

$$\mathcal{O}\left(\left(\frac{\mu_2}{\lambda}D\right)^{2/7}\log\left(\frac{\mu_1\mu_2^2 D^2}{\lambda^3}\right) + \log\log_2\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right). \tag{6}$$

An alternative to the above is to use a hybrid scheme, starting with accelerated gradient descent (which is an optimal *first-order* method for strongly convex functions with Lipschitz gradients) and when close enough to the optimal solution, switch to a Newton based method The required number of iterations can be shown to be at most

---

[2] That is, for any vectors $\mathbf{v}, \mathbf{w}$, the function $g(t) = f(\mathbf{w} + t\mathbf{v})$ satisfies $|g'''(t)| \leq 2g''(t)^{3/2}$.

$$\mathcal{O}\left(\sqrt{\frac{\mu_1}{\lambda}} \cdot \log\left(\frac{\mu_1 \mu_2^2 D^2}{\lambda^3}\right) + \log\log_2\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right), \tag{7}$$

where $D = \|\mathbf{w}_1 - \mathbf{w}^*\|$ (see [15,16]). Clearly, by taking the best of (6) and (7) (depending on the parameters), one can theoretically attain an oracle complexity which is the minimum of (6) and (7). This minimum matches (up to a logarithmic factors) the lower bound in (5), which we establish in this paper.

It is interesting to note that the bounds in (6) and (7) are not directly comparable: The first bound has a polynomial dependence on μ_2/λ and $\|\mathbf{w}_1 - \mathbf{w}^*\|$, and a logarithmic dependence on μ_1, whereas the second bound has a polynomial dependence on μ_1/λ, logarithmic dependence on $\|\mathbf{w}_1 - \mathbf{w}^*\|$, and a logarithmic dependence on μ_2. In a rather wide parameter regime (e.g. when D is reasonably large, as often occurs in practice), the bound of the hybrid scheme can be better than that of pure second-order methods. In light of this, [16] raised the question of whether second-order schemes are indeed useful at the initial stage of the optimization process, for these types of problems. Our results indicate that indeed, in certain parameter regimes, this is not the case.

Analogous results can be obtained for convex (not necessarily strongly convex) smooth functions. The best upper bounds we are aware are for the (second-order) A-NPE method of [10], and for the (first order) accelerated gradient descent method, which are

$$\mathcal{O}\left(\left(\frac{\mu_2 D^3}{\epsilon}\right)^{2/7}\right) \quad \text{and} \quad \mathcal{O}\left(\sqrt{\frac{\mu_1 D^2}{\epsilon}}\right)$$

respectively. Clearly, by taking the best of these two methods (depending on the problem parameters), one can attain an oracle complexity equal to the minimum of the two bounds. This is matched (up to constants) by the lower bound in (4), which we establish in this paper.

Finally, we discuss the few existing lower bounds known for second-order methods. If μ_2 is not bounded (i.e.,the Hessians are not Lipschitz), it is easy to show that Hessian information is not useful. Specifically, the lower bound of (2) for first-order methods will then also apply to second-order methods, and in fact, to any method based on local information (see [13, section 7.2.6] and [4]). Of course, this lower bound does not apply to second-order methods when μ_2 is bounded. In our setting, it is also possible to prove an $\Omega(\log\log(1/\epsilon))$ lower bound, even in one dimension [13, section 8.1.1], but this does not capture the dependence on the strong convexity and Lipschitz parameters. Some algorithm-specific lower bounds in the context of non-convex optimization are provided in [7]. Finally, we were recently informed of a new work ([1], yet unpublished at the time of writing), which uses a clean and elegant smoothing approach, to derive second- and higher-order oracle lower bounds directly from known first-order oracle lower bounds, as well as extensions to randomized algorithms. However, the resulting bounds are not as tight as ours.

2 Main results

In this section, we formally present our main results, starting with second-order oracle complexity bounds (Sect. 2.1), and then discussing extensions to higher-order oracles (Sect. 2.2).

2.1 Second-order oracle

We consider a *second-order* oracle, which given a point $\mathbf{w}$ returns the function's value $f(\mathbf{w})$, its gradient $\nabla f(\mathbf{w})$ and its Hessian $\nabla^2 f(\mathbf{w})$, and algorithms, which produce a sequence of points $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T$, with each $\mathbf{w}_t$ being some deterministic function of the oracle's responses at $\mathbf{w}_1, \ldots, \mathbf{w}_{t-1}$. Our main results (for strongly convex and convex functions f respectively) are provided below.

Theorem 1 *For any positive $\mu_1, \mu_2, \lambda, D, \epsilon$ such that $\frac{\mu_1}{\lambda} \geq c_1, \frac{\mu_2}{\lambda} D \geq c_2$ and $\epsilon < \frac{c_3 \lambda^3}{\mu_2^2}$ (for some positive universal constants c_1, c_2, c_3), and any algorithm as above with initialization point $\mathbf{w}_1$, there exists a function $f : \mathbb{R}^d \to \mathbb{R}$ (for some finite d) such that*

- *f is λ-strongly convex, twice-differentiable, has μ_1-Lipschitz gradients and μ_2-Lipschitz Hessians, and has a global minimum $\mathbf{w}^*$ satisfying $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq D$.*
- *For some universal constant $c > 0$, the index T required to ensure $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ is at least*

$$c \cdot \left(\min \left\{ \sqrt{\frac{\mu_1}{\lambda}} \, , \, \left(\frac{\mu_2}{\lambda} D \right)^{2/7} \right\} + \log \log_{18} \left(\frac{\lambda^3 / \mu_2^2}{\epsilon} \right) \right).$$

Theorem 2 *For any positive $\mu_1, \mu_2, D, \epsilon$ and any algorithm as above with initialization point $\mathbf{w}_1$, there exists a function $f : \mathbb{R}^d \to \mathbb{R}$ (for some finite d) such that*

- *f is convex, twice-differentiable, has μ_1-Lipschitz gradients and μ_2-Lipschitz Hessians, and has a global minimum $\mathbf{w}^*$ satisfying $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq D$.*
- *For some universal constant $c > 0$, the index T required to ensure $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ is at least*

$$c \cdot \min \left\{ \sqrt{\frac{\mu_1 D^2}{\epsilon}} \, , \, \left(\frac{\mu_2 D^3}{\epsilon} \right)^{2/7} \right\}.$$

We emphasize that the theorems focus on the high-dimensional setting, where the dimension d is not necessarily fixed and may depend on other problem parameters (more specifically, we require the dimension to be at least $2T$). Also, we note that the parameter constraints in Theorem 1 are purely for technical reasons (they imply that the different terms in the bound are at least some positive constant), and can probably be relaxed somewhat.

Let us compare these theorems to the upper bounds discussed in the related work section, which are

$$\mathcal{O}\left(\min\left\{\sqrt{\frac{\mu_1}{\lambda}}, \left(\frac{\mu_2 D}{\lambda}\right)^{2/7}\right\} \cdot \log\left(\frac{\mu_1 \mu_2^2 D^2}{\lambda^3}\right) + \log\log_2\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right).$$

in the strongly convex case, and

$$\mathcal{O}\left(\min\left\{\sqrt{\frac{\mu_1 D^2}{\epsilon}}, \left(\frac{\mu_2 D^3}{\epsilon}\right)^{2/7}\right\}\right).$$

in the convex case. Our bound in the convex case is tight up to constants, and in the strongly convex case, up to a $\log(\mu_1 \mu_2^2 D^2/\lambda^3)$ factor. We conjecture that some such logarithmic factor (possibly a smaller one) is indeed necessary, in order to get a tight interpolation to the $\Omega(\sqrt{\mu_1/\lambda} \cdot \log(\mu_1 D^2/\epsilon))$ lower bound of first-order methods as $\mu_2 \to \infty$ (see [13, section 7.2.6] and [4]), and that it can be recovered with a more careful analysis of our construction. However, this involves some non-trivial technical challenges, which we leave to future work.

Remark 1 (Randomized Algorithms) In this paper, we focus on deterministic algorithms, where each point $\mathbf{w}_t$ produced is a deterministic function of the oracle's responses at $\mathbf{w}_1, \ldots, \mathbf{w}_{t-1}$. This follows most existing works on oracle complexity lower bounds, and is without much loss of generality, since all existing state-of-the-art algorithms for this problem are deterministic. In any case, we believe our bounds can be readily extended to randomized algorithms, using the same techniques that extend first-order oracle complexity lower bounds to randomized algorithms (see [13] and the explicit construction in [20]). In a nutshell, for deterministic algorithms we can tailor the "hard" function to the algorithm, whereas for randomized algorithms, we construct some fixed distribution over "hard" functions, so that any algorithm (random or not) will fail to achieve a certain optimization error with high probability. Since the formal proof is considerably more involved than in the deterministic case, we leave its formal derivation to future work.

2.2 Higher order oracles

In addition to first-order and second-order oracles, it is of interest to understand what can be achieved with methods employing higher order derivatives. It turns out that the techniques we use to establish our second-order lower bounds can be easily generalized to such higher-order methods.

More explicitly, we consider methods which can be modelled as interacting with a *k-th order* oracle, which given a point $\mathbf{w}$ returns the function's value and all of its derivatives up to order k, namely, $f(\mathbf{w}), \nabla f(\mathbf{w}), \nabla^2 f(\mathbf{w}), \ldots, \nabla^k f(\mathbf{w})$. Given access to such an oracle, the method produces a sequence of points $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T$ as before (where each $\mathbf{w}_t$ is a deterministic function of the previous oracle responses).

For simplicity, we will focus here on the case of convex functions (not necessarily strongly convex), where the k-th order derivative is Lipschitz continuous.

Theorem 3 *For any positive integer k, positive μ_k, D, ϵ, and algorithm based on a k-th order oracle as above, there exists a function $f : \mathbb{R}^d \to \mathbb{R}$ (for some finite d) such that*

- *f is convex, k times differentiable, k-order smooth (i.e., $\|\nabla^k f(\mathbf{u}) - \nabla^k f(\mathbf{v})\| \leq \mu_k \|\mathbf{u} - \mathbf{v}\|$) and has a global minimum $\mathbf{w}^*$ satisfying $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq D$.*
- *For some universal constant $c > 0$, the index T required to ensure $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ is at least*

$$c \left(\frac{\mu_k D^{k+1}}{(k+1)!k\epsilon} \right)^{2/(3k+1)}.$$

Note that this result directly generalizes existing results for first-order oracles ($k = 1$), as well as our results for second-order oracles ($k = 2$, when μ_1 is unrestricted).

Finally, we compare our lower bound to the best upper bound we are aware of, established by [5] using a high-order method with oracle complexity of

$$\mathcal{O}\left(\sqrt{k} \left(\frac{f(\mathbf{w}_1) - f(\mathbf{w}^*)}{\epsilon} + \frac{\mu_k D^{k+1}}{(k+1)!\epsilon} \right)^{1/(k+1)} \right).$$

Note that the upper bound contains an additional $(f(\mathbf{w}_1) - f(\mathbf{w}^*))/\epsilon$ term, and moreover, the exponent (as a function of k) is larger than ours ($1/(k+1)$ vs. $2/(3k+1)$). Based on our results, we know that this upper bound is loose in the $k = 2$ case, so we conjecture that it is indeed loose for all k, and can be improved.

3 Proof ideas

The proofs of our theorems are based on a careful modification of a known lower bound construction for first-order methods (see [15]). That construction uses quadratic functions, which in the convex case and ignoring various parameters, have a basic structure of the form

$$f_T(\mathbf{w}) = f_T(w_1, w_2, \ldots) = w_1^2 + \sum_{j=1}^{T-1} (w_j - w_{j+1})^2 + w_T^2 - w_1$$

(more precisely, one considers $f_T(V\mathbf{w})$ for a certain orthogonal matrix V, and use additional parameters depending on the smoothness). A crucial ingredient of the proof is that the function $x \mapsto x^2$ has a value and derivative of zero at the origin, which allows us to construct a function which "hides" information from an algorithm relying solely on values and gradients. This can be shown to lead to an optimization error lower bound of the form $\min_{\mathbf{w}} f_T(\mathbf{w}) - \min_{\mathbf{w}} f_{2T}(\mathbf{w})$ after T oracle queries, which for first-order methods leads to an $\Omega(\mu_1 D^2/T^2)$ lower bound on the error, translating to an

$\Omega(\sqrt{\mu_1 D^2/\epsilon})$ lower bound on T. However, this construction leads to trivial bounds for second-order methods, since given the Hessian and a gradient of a quadratic function at just a single point, one can already compute the exact minimizer.

Our approach to handle second-order (and more generally, k-order) methods is quite simple: Instead of $x \mapsto x^2$, we rely on mappings of the form $x \mapsto |x|^{k+1}$, and use functions with the basic structure

$$f_T(\mathbf{w}) = |w_1|^{k+1} + \sum_{j=1}^{T-1} |w_j - w_{j+1}|^{k+1} + |w_T|^{k+1} - w_1.$$

The intuition is that $x \mapsto |x|^{k+1}$ has a value *and first k derivatives* of zero at the origin, and therefore variants of the function above can be used to "hide" information from the algorithm, even if it can receive Hessians or higher-order derivatives of the function. Another motivation for choosing such functions is that they are generally not self-concordant, and therefore the upper bounds relevant to self-concordant functions do not apply. We rely on this construction and arguments similar to those of first-order oracle lower bounds, to get our results.

In the derivation of our results for second-order methods, there are two technical challenges that need to be overcome: The first is that f_T, as defined above (for $k = 2$), can be shown to have globally Lipschitz Hessians, but not globally Lipschitz gradients as required by our theorems. To tackle this, we replace the mapping $x \mapsto |x|^3$ by a more complicated mapping, which is cubic close to the origin and quadratic further away. This necessarily complicates the proof. The second challenge is that due to the cubic terms, computing the minimizer of f_T and its minimal value is more challenging than in first-order lower bounds, especially in the strongly convex case (where we are unable to even find a closed-form expression for the minimizer, and resort to bounds instead). Again, this makes the analysis more complicated.

We conclude this section by sketching how our bounds can be derived in case of second-order methods, and in the simplest possible setting, where we wish to obtain an $\Omega((D^3/\epsilon)^{2/7})$ lower bound for the class of convex functions with Lipschitz Hessians (and no assumptions on the Lipschitz parameter of the gradients), assuming the algorithm makes its first query at the origin:

Proposition 1 *For any positive μ_2, D, ϵ and any algorithm with an initialization point $\mathbf{w}_1$ in the origin, there exists a function $f : \mathbb{R}^d \to \mathbb{R}$ (for some finite d) such that*

- *f is convex, twice-differentiable, has μ_2-Lipschitz Hessians, and has a global minimum $\mathbf{w}^*$ satisfying $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq D$.*
- *The index T required to ensure $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ is $\Omega((D^3/\epsilon)^{2/7})$.*

Proof Sketch Consider the function f_T in this class of the form

$$f_T(\mathbf{w}) = |w_1|^3 + \sum_{j=1}^{T-1} |w_j - w_{j+1}|^3 + |w_T|^3 - 3\gamma \cdot w_1,$$

where γ is a parameter to be chosen later. Computing the derivatives and setting to zero, and arguing that the minimizer must have non-negative coordinates, we get that the optimum satisfies

$$w_1^2 + (w_1 - w_2)^2 = \gamma, \quad w_T^2 = (w_{T-1} - w_T)^2$$

and

$$\forall j = 2, 3, \ldots, T - 1, \quad (w_{j-1} - w_j)^2 = (w_j - w_{j+1})^2.$$

It can be verified that this is satisfied by $w_j = (T + 1 - j)\sqrt{\frac{\gamma}{T^2+1}}$ for all $j = 1, 2, \ldots, T$, and that this is the unique minimizer of f_T as a function on $\mathbb{R}^T$. Moreover, assuming $\gamma \leq D^2/T$, the norm of this minimizer (and hence the initial distance to it from the algorithm's first query point, by assumption) is at most D as required. Plugging this $\mathbf{w}$ into f_T, we get that

$$\min_{\mathbf{w}} f_T(\mathbf{w}) = -\frac{2\gamma^{3/2} T}{\sqrt{T^2 + 1}}.$$

Now, using arguments very similar to those in the first-order oracle complexity lower bounds in [15], it is possible to construct a function for which the optimization error of the algorithm is lower bounded by $\min_{\mathbf{w}} f_T(\mathbf{w}) - \min_{\mathbf{w}} f_{2T}(\mathbf{w})$. By the calculations above, this in turn equals

$$2\gamma^{3/2} \left(\frac{2T}{\sqrt{4T^2 + 1}} - \frac{T}{\sqrt{T^2 + 1}} \right) = 2\gamma^{3/2} \left(\frac{1}{\sqrt{1 + \frac{1}{4T^2}}} - \frac{1}{\sqrt{1 + \frac{1}{T^2}}} \right).$$

Using the fact that $\frac{1}{\sqrt{1+x}} \approx 1 - \frac{1}{2}x$ for small x, this equals $\Omega(\gamma^{3/2}/T^2)$. Choosing γ on the order of D^2/T (as required earlier to satisfy the norm constraint on the minimizer), we get a lower bound of $\Omega(D^3/T^{7/2})$ on the optimization error ϵ, or equivalently, a lower bound of $\Omega((D^3/\epsilon)^{2/7})$ on T. $\qquad\square$

4 Proof of Theorem 1

We will assume without loss of generality that the algorithm initializes at $\mathbf{w}_1 = \mathbf{0}$ (if that is not the case, one can simply replace the "hard" function $f(\mathbf{w})$ below by $f(\mathbf{w} - \mathbf{w}_1)$, and the same proof holds verbatim). Thus, the theorem requires that our function has a minimizer $\mathbf{w}^*$ satisfying $\|\mathbf{w}^*\| \leq D$.

Let Δ, γ be parameters to be chosen later[3]. Define $g : \mathbb{R} \mapsto \mathbb{R}$ as

$$g(x) = \begin{cases} \frac{1}{3}|x|^3 & |x| \leq \Delta \\ \Delta x^2 - \Delta^2 |x| + \frac{1}{3}\Delta^3 & |x| > \Delta, \end{cases}$$

---

[3] Ultimately, we will choose $\gamma = \min\left\{ \left(\frac{3(\mu_1 - \lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(12\lambda)^6}{2^4 \mu_2^6}} \right\}$ and $\Delta = \sqrt{\gamma}$, see Subsection 4.3.

which is easily verified to be convex and twice continuously differentiable, and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$ be orthogonal unit vectors in $\mathbb{R}^d$ which will be specified later. Letting the number of iterations T be fixed, we consider the function

$$f(\mathbf{w}) = \frac{\mu_2}{12} \left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{v}_i, \mathbf{w} \rangle - \langle \mathbf{v}_{i+1}, \mathbf{w} \rangle) - \gamma \langle \mathbf{v}_1, \mathbf{w} \rangle \right) + \frac{\lambda}{2} \|\mathbf{w}\|^2,$$

where $\tilde{T} \geq \max \left\{ 4\gamma \left(\frac{\mu_2}{6\lambda} \right)^2 + 1, 2T, \frac{\gamma \mu_2}{6\lambda} + 1 \right\}$ is some sufficiently large number, and the dimension d is at least $2\tilde{T}$.

The proof is constructed of several parts: First, we analyze properties of the global minimum of f (Sect. 4.1). Then, we prove the oracle complexity lower bound in Sect. 4.2 (depending on Δ, γ), and finally, in Sect. 4.3, we choose the parameters so that f indeed has the various geometric properties specified in the theorem.

4.1 Minimizer of f

The goal of this subsection is to prove the following proposition, which characterizes key properties of the global minimum of f:

Proposition 2 *Suppose that* $\gamma \geq 10^4 \left(\frac{\lambda}{\mu_2} \right)^2$ *and* $\Delta \geq \sqrt{\gamma}$. *Then for any unit orthogonal* $\mathbf{v}_1, \mathbf{v}_2, \ldots$, *the function* f *has a unique minimizer* $\mathbf{w}^*$ *which satisfies the following:*

1. For any $t \in \{1, 2, \ldots, \tilde{T}\}$, it holds that
   $\langle \mathbf{v}_t, \mathbf{w}^* \rangle \geq \max \left\{ 0, \frac{\gamma^{3/4}}{7\sqrt{12\lambda/\mu_2}} + \sqrt{\gamma} \left(\frac{1}{2} - t \right) \right\}$.
2. There exists some $t_0 \leq \tilde{T}/2$ such that for all indices $k \in \{0, 1, \ldots, \tilde{T} - t_0\}$, it holds that $\langle \mathbf{v}_{t_0+k}, \mathbf{w}^* \rangle \geq \frac{108\lambda}{\mu_2} \cdot (18)^{-2^k}$.
3. $\|\mathbf{w}^*\|^2 \leq \frac{2\gamma^{7/4}}{(12\lambda/\mu_2)^{3/2}}$.

Since f is strongly convex, its global minimizer is unique and well-defined. To prove the proposition, we will consider the simpler strongly-convex function

$$\tilde{f}(\mathbf{w}) = \frac{1}{3} \sum_{i=1}^{\tilde{T}-1} |w_i - w_{i+1}|^3 + \frac{\tilde{\lambda}}{2} \|\mathbf{w}\|^2 - \gamma \cdot w_1, \tag{8}$$

where $\tilde{\lambda} := \frac{12\lambda}{\mu_2}$, and prove that its minimizer $\tilde{\mathbf{w}}^*$ satisfies the following:

1. For any $t \in \{1, 2, \ldots, \tilde{T}\}$, it holds that $\tilde{w}_t^* \geq \max \left\{ 0, \frac{\gamma^{3/4}}{7\sqrt{\tilde{\lambda}}} + \sqrt{\gamma} \left(\frac{1}{2} - t \right) \right\}$ (Lemma 2).
2. There exists some $t_0 \leq \tilde{T}/2$ such that for all $k \in \{0, 1, \ldots, \tilde{T} - t_0\}$, it holds that $\tilde{w}_{t_0+k}^* \geq 9\tilde{\lambda} \cdot (18)^{-2^k}$ (Lemma 3).
3. $\sum_{i=1}^{\tilde{T}} \tilde{w}_i^{*2} \leq \frac{2\gamma^{7/4}}{\tilde{\lambda}^{3/2}}$ (Lemma 4)

We then argue that the minimizer $\mathbf{w}^*$ of f satisfies $\langle \mathbf{v}_i, \mathbf{w}^* \rangle = \tilde{w}_i^*$ for all $i = 1, 2, \ldots, \tilde{T}$ (Lemma 5), and that $\|\mathbf{w}^*\|^2 = \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^* \rangle^2$ (Lemma 6), from which Proposition 2 follows.

We begin with the following technical key result:

Lemma 1 *It holds that* $\tilde{w}_1^* \geq \tilde{w}_2^* \geq \cdots \geq \tilde{w}_{\tilde{T}}^* \geq 0$, *and*

$$\tilde{w}_{t+1}^* = w_t^* - \sqrt{\gamma - \tilde{\lambda} \sum_{j=1}^{t} \tilde{w}_j^*} \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\}.$$

Moreover, $\sum_{j=1}^{\tilde{T}} \tilde{w}_j^* = \frac{\gamma}{\tilde{\lambda}}$.

Proof We begin by showing that $\tilde{w}_j^* \geq 0$ for all j, first for $j = 1$ and then for $j > 1$. To do so, note that $\tilde{f}(\mathbf{0}) = 0$ yet $\nabla \tilde{f}(\mathbf{0}) = -\gamma \cdot \mathbf{e}_1 \neq \mathbf{0}$, and therefore $\mathbf{0}$ is a sub-optimal point. Thus, we must have $\tilde{f}(\tilde{\mathbf{w}}^*) < 0$. The only negative term in the definition of $\tilde{f}(\cdot)$ is $-\gamma \cdot w_1$, so we must have $\tilde{w}_1^* > 0$. We now argue that $\tilde{w}_j^* \geq 0$ for all $j > 1$: Otherwise, let $\mathbf{w}$ be the vector which equals $w_j = |\tilde{w}_j^*|$ for all j, and note that $w_1 = \tilde{w}_1^*$ since we just showed $\tilde{w}_1^* > 0$. Based on this, it is easily verified that

$$\tilde{f}(\mathbf{w}) - \tilde{f}(\tilde{\mathbf{w}}^*) = \frac{1}{3} \sum_{i=1}^{\tilde{T}-1} \left(\left| |\tilde{w}_i^*| - |\tilde{w}_{i+1}^*| \right|^3 - |\tilde{w}_i^* - \tilde{w}_{i+1}^*|^3 \right) \leq 0,$$

which means that $\mathbf{w}$ is the (unique) minimum of $\tilde{f}$, hence $\mathbf{w} = \tilde{\mathbf{w}}^*$. By definition of $\mathbf{w}$, this implies $\tilde{w}_j^* = |\tilde{w}_j^*|$ for all j, hence $\tilde{w}_j^* \geq 0$ for all j.

We now turn to prove that $\tilde{w}_j^*$ is monotonically decreasing in j. Suppose on the contrary that this is not the case, and let j_0 be the smallest index for which $\tilde{w}_{j_0}^* < \tilde{w}_{j_0+1}^*$, and let $\delta := \tilde{w}_{j_0+1}^* - \tilde{w}_{j_0}^* > 0$. Define the vector $\mathbf{w}$ to be

$$w_i = \begin{cases} \tilde{w}_i^* & i \leq j_0 \\ \max\{0, \tilde{w}_i^* - \delta\} & d \geq i > j_0 \end{cases}.$$

Note that this vector must be different than $\mathbf{w}$, as $w_{j_0+1} = \max\{0, \tilde{w}_{j_0+1}^* - \delta\} = \max\{0, \tilde{w}_{j_0}^*\} = \tilde{w}_{j_0}^* = w_{j_0}$, hence $w_{j_0+1} = w_{j_0}$ yet $\tilde{w}_{j_0+1}^* > \tilde{w}_{j_0}^*$ by assumption. On the other hand, it is easily verified that $|w_i - w_{i+1}|^3 \leq |\tilde{w}_i^* - \tilde{w}_{i+1}^*|^3$ and $w_i^2 \leq (\tilde{w}_i^*)^2$ for all[4] i, and therefore $\tilde{f}(\mathbf{w}) \leq \tilde{f}(\tilde{\mathbf{w}}^*)$. But since $\tilde{\mathbf{w}}^*$ is the unique global minimizer and $\mathbf{w} \neq \tilde{\mathbf{w}}^*$, we get a contradiction, so we must have $\tilde{w}_j^*$ monotonically decreasing for all j.

---

[4] This is trivially true for $i < j_0$. For $i = j_0$, we have $|w_{j_0} - w_{j_0+1}|^3 = 0 < |\tilde{w}_{j_0}^* - \tilde{w}_{j_0+1}^*|^3$ and $w_{j_0}^2 = (\tilde{w}_{j_0}^*)^2$. For $i > j_0$, we have $|w_i - w_{i+1}|^3 = |\max\{0, \tilde{w}_i^* - \Delta\} - \max\{0, \tilde{w}_{i+1}^* - \Delta\}|^3 \leq |(\tilde{w}_i^* - \Delta) - (\tilde{w}_{i+1}^* - \Delta)|^3 = |\tilde{w}_i^* - \tilde{w}_{i+1}^*|^3$, and moreover, $w_i^2 = \max\{0, \tilde{w}_i^* - \Delta\}^2$, which is 0 (hence $\leq (\tilde{w}_i^*)^2$) if $\tilde{w}_i^* \leq \Delta$ and less than $(\tilde{w}_i^*)^2$ if $\tilde{w}_i^* > \Delta$.

We now turn to prove the recursive relation $\tilde{w}_{t+1}^* = w_t^* - \sqrt{\gamma - \tilde{\lambda} \sum_{j=1}^t \tilde{w}_j^*}$. By differentiating $\tilde{f}$ and setting to zero (and using the fact that $\tilde{w}_j^*$ is monotonically decreasing in j), we get that

$$(\tilde{w}_1^* - \tilde{w}_2^*)^2 = \gamma - \tilde{\lambda} w_1^*, \quad (\tilde{w}_{\tilde{T}-1}^* - \tilde{w}_{\tilde{T}}^*)^2 = \tilde{\lambda} w_{\tilde{T}}^* \tag{9}$$

and

$$(\tilde{w}_t^* - \tilde{w}_{t+1}^*)^2 = (\tilde{w}_{t-1}^* - \tilde{w}_t^*)^2 - \tilde{\lambda} \tilde{w}_t^* \quad \forall t \in \{2, 3, \ldots, \tilde{T} - 1\}. \tag{10}$$

By unrolling this recursive form, we get

$$(\tilde{w}_t^* - \tilde{w}_{t+1}^*)^2 = \gamma - \tilde{\lambda} \sum_{j=1}^t \tilde{w}_j^* \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\},$$

from which the equation

$$\tilde{w}_{t+1}^* = w_t^* - \sqrt{\gamma - \tilde{\lambda} \sum_{j=1}^t \tilde{w}_j^*} \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\} \tag{11}$$

follows, again using the monotonicity of $\tilde{w}_t^*$ in t.

It remains to prove that $\sum_{j=1}^{\tilde{T}} \tilde{w}_j^* = \frac{\gamma}{\tilde{\lambda}}$. By summing both sides of (10) from $t = 2$ to $t = \tilde{T} - 1$ we have that:

$$(\tilde{w}_{\tilde{T}-1}^* - \tilde{w}_{\tilde{T}}^*)^2 = (\tilde{w}_1^* - \tilde{w}_2^*)^2 - \tilde{\lambda} \sum_{t=2}^{\tilde{T}-1} \tilde{w}_t^*$$

So by using (9) we get the desired equality. □

Lemma 2 *For any $t \in \{1, 2, \ldots, \tilde{T}\}$,*

$$\tilde{w}_t^* \geq \max \left\{ 0, \frac{\gamma^{3/4}}{7\sqrt{\tilde{\lambda}}} + \sqrt{\gamma} \left(\frac{1}{2} - t \right) \right\}.$$

Proof By the displayed equation in Lemma 1, we clearly have $\tilde{w}_{t+1}^* \geq \tilde{w}_t^* - \sqrt{\gamma}$ for all $t \leq \tilde{T} - 1$, and therefore

$$\tilde{w}_t^* \geq \tilde{w}_1^* - (t-1)\sqrt{\gamma} \quad \forall t \in \{1, 2, \ldots, \tilde{T}\}. \tag{12}$$

Using the facts that $\tilde{w}_t^*$ is also always non-negative, that $\tilde{T} \geq \frac{\gamma \mu_2}{6\lambda} + 1 = \frac{\gamma}{\tilde{\lambda}} + 1 \geq \tilde{w}_1^* + 1$, and by Lemma 1,

$$\frac{\gamma}{\tilde{\lambda}} = \sum_{t=1}^{\tilde{T}} \tilde{w}_t^* \geq \sum_{t=1}^{\tilde{T}} \max\{0, \tilde{w}_1^* - (t-1)\sqrt{\gamma}\} = \sum_{t=1}^{\lfloor \tilde{w}_1^*/\sqrt{\gamma}+1 \rfloor} \left(\tilde{w}_1^* - (t-1)\sqrt{\gamma} \right)$$

$$= \left\lfloor \frac{\tilde{w}_1^*}{\sqrt{\gamma}} + 1 \right\rfloor \tilde{w}_1^* - \sqrt{\gamma} \frac{\left(\left\lfloor \frac{\tilde{w}_1^*}{\sqrt{\gamma}} + 1 \right\rfloor - 1 \right) \left\lfloor \frac{\tilde{w}_1^*}{\sqrt{\gamma}} + 1 \right\rfloor}{2} \geq \frac{(\tilde{w}_1^*)^2}{\sqrt{\gamma}} - \sqrt{\gamma} \frac{\frac{\tilde{w}_1^*}{\sqrt{\gamma}} \left(\frac{w_1^*}{\sqrt{\gamma}} + 1 \right)}{2},$$

which implies that $(\tilde{w}_1^*)^2 - \sqrt{\gamma} \cdot w_1^* - \frac{2\gamma^{3/2}}{\tilde{\lambda}} \leq 0$, which implies in turn

$$\tilde{w}_1^* \leq \frac{\sqrt{\gamma} + \sqrt{\gamma + 8\gamma^{3/2}/\tilde{\lambda}}}{2} \leq \frac{\sqrt{\gamma} + \sqrt{\gamma} + \sqrt{8\gamma^{3/2}/\tilde{\lambda}}}{2} = \sqrt{\gamma} + \sqrt{\frac{2\gamma^{3/2}}{\tilde{\lambda}}}. \quad (13)$$

On the other hand, again by Lemma 1, we know that

$$\tilde{w}_{t+1}^* \leq \tilde{w}_t^* - \frac{\sqrt{\gamma}}{2}, \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\} : \sum_{j=1}^t \tilde{w}_j^* \leq \frac{3\gamma}{4\tilde{\lambda}},$$

and hence

$$\tilde{w}_{t+1}^* \leq \tilde{w}_1^* - \frac{t\sqrt{\gamma}}{2}, \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\} : \sum_{j=1}^t \tilde{w}_j^* \leq \frac{3\gamma}{4\tilde{\lambda}}. \quad (14)$$

Let $t_0 \in \{1, 2, \ldots, \tilde{T}\}$ be the smallest index such that $\sum_{j=1}^{t_0} \tilde{w}_j^* > \frac{3\gamma}{4\tilde{\lambda}}$ (such an index must exist since $\sum_{j=1}^{\tilde{T}} \tilde{w}_j^* = \frac{\gamma}{\tilde{\lambda}}$). Since $\frac{3\gamma}{4\tilde{\lambda}} < \sum_{j=1}^{t_0} \tilde{w}_j^* \leq t_0 \tilde{w}_1^* \leq t_0 \left(\sqrt{\gamma} + \sqrt{2\gamma^{3/2}/\tilde{\lambda}} \right)$ by (13), it follows that

$$t_0 \geq \frac{3\gamma}{4\tilde{\lambda} \left(\sqrt{\gamma} + \sqrt{2\gamma^{3/2}/\tilde{\lambda}} \right)} = \frac{3\sqrt{\gamma}}{4 \left(\tilde{\lambda} + \sqrt{2\gamma^{1/2}\tilde{\lambda}} \right)}.$$

According to (14) and the fact that $\tilde{w}_{t_0}^* \geq 0$, it follows that

$$0 \leq \tilde{w}_{t_0}^* \leq \tilde{w}_1^* - \frac{(t_0 - 1)\sqrt{\gamma}}{2},$$

and hence

$$\tilde{w}_1^* \geq \frac{(t_0 - 1)\sqrt{\gamma}}{2} \geq \frac{3\gamma}{8(\tilde{\lambda} + \sqrt{2\gamma^{1/2}\tilde{\lambda}})} - \frac{\sqrt{\gamma}}{2}.$$

Using this and (12), it follows that for all $t \leq \tilde{T}$,

$$\tilde{w}_t^* \geq \frac{3\gamma}{8(\tilde{\lambda} + \sqrt{2\gamma^{1/2}\tilde{\lambda}})} + \sqrt{\gamma} \left(\frac{1}{2} - t \right).$$

Since we assumed $\gamma \geq 10^4 (\lambda/\mu_2)^2 > (12\lambda/\mu_2)^2 = \tilde{\lambda}^2$, we have $\tilde{\lambda} < \sqrt{\gamma^{1/2}\tilde{\lambda}}$, so the above can be lower bounded by the simpler expression $\gamma^{3/4}/7\sqrt{\tilde{\lambda}} + \sqrt{\gamma}(1/2 - t)$. Since we also know that $\tilde{w}_t^*$ is non-negative, the result follows. □

Lemma 3 *There exists an index* $t_0 \leq \tilde{T}/2$ *such that*

$$\tilde{w}_{t_0+k}^* \geq 9\tilde{\lambda} \cdot (18)^{-2^k} \quad \forall k \in \{0, 1, \ldots, \tilde{T} - t_0\}$$

Proof By Lemma 1, it holds for any $t \in \{1, 2, \ldots, \tilde{T} - 1\}$ that

$$\tilde{w}_t^* - \tilde{w}_{t+1}^* = \sqrt{\gamma - \tilde{\lambda} \sum_{j=1}^t \tilde{w}_j^*} = \sqrt{\gamma - \tilde{\lambda}\left(\frac{\gamma}{\tilde{\lambda}} - \sum_{j=t+1}^{\tilde{T}} \tilde{w}_j^*\right)} = \sqrt{\tilde{\lambda} \sum_{j=t+1}^{\tilde{T}} \tilde{w}_j^*}.$$

(15)

In particular, since $\tilde{w}_j^* \geq 0$ for all $j \leq \tilde{T}$, it follows that $\tilde{w}_t^* \geq \sqrt{\tilde{\lambda}\sum_{j=t+1}^{\tilde{T}} \tilde{w}_j^*} \geq \sqrt{\tilde{\lambda}\tilde{w}_{t+1}^*}$, and therefore

$$\tilde{w}_{t+1}^* \leq \frac{1}{\tilde{\lambda}}(\tilde{w}_t^*)^2 \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\}.$$

(16)

Let $t \leq \tilde{T} - 1$ be any index such that[5] $\tilde{w}_{t+1}^* \leq \frac{\tilde{\lambda}}{2}$. By (16), this implies that

$$\sum_{j=t+1}^{\tilde{T}} \tilde{w}_j^* \leq \sum_{k=0}^{\tilde{T}-t-1} \tilde{\lambda}\left(\frac{\tilde{w}_{t+1}^*}{\tilde{\lambda}}\right)^{2^k} = \sum_{k=0}^{\tilde{T}-t-1} \tilde{w}_{t+1}^* \left(\frac{\tilde{w}_{t+1}^*}{\tilde{\lambda}}\right)^{2^k - 1}$$

$$\leq \sum_{k=0}^{\tilde{T}-t-1} \tilde{w}_{t+1}^* \left(\frac{1}{2}\right)^{2^k - 1} < 2 \cdot \tilde{w}_{t+1}^*.$$

Using the inequality above together with (15) and the monotonicity of $\tilde{w}_t^*$, we get that for all $t \leq \tilde{T} - 1$ such that $\tilde{w}_{t+1}^* \leq \frac{\tilde{\lambda}}{2}$,

$$\tilde{w}_t^* = \tilde{w}_{t+1}^* + \sqrt{\tilde{\lambda} \sum_{j=t+1}^{\tilde{T}} \tilde{w}_j^*} \leq \tilde{w}_{t+1}^* + \sqrt{2\tilde{\lambda}\tilde{w}_{t+1}^*} = \sqrt{\tilde{w}_{t+1}^*}\left(\sqrt{\tilde{w}_{t+1}^*} + \sqrt{2\tilde{\lambda}}\right)$$

$$\leq \sqrt{\tilde{w}_{t+1}^*}\left(\sqrt{\frac{\tilde{\lambda}}{2}} + \sqrt{2\tilde{\lambda}}\right) \leq 3\sqrt{\tilde{\lambda}\tilde{w}_{t+1}^*}.$$

---

[5] Such an index must exist: By assumption, $\tilde{T} \geq 2\gamma \left(\frac{\mu_2}{6\lambda}\right)^2 = \frac{2\gamma}{\tilde{\lambda}^2}$, so by Lemma 1, $\frac{\gamma}{\tilde{\lambda}} = \sum_{t=1}^{\tilde{T}} \tilde{w}_t^* \geq \tilde{T}\tilde{w}_{\tilde{T}}^* \geq \frac{2\gamma}{\tilde{\lambda}^2}\tilde{w}_{\tilde{T}}^*$, hence $\tilde{w}_{\tilde{T}} \leq \tilde{\lambda}/2$.

This chain of inequalities implies that

$$w_{t+1} \geq \frac{(\tilde{w}_t^*)^2}{9\tilde{\lambda}} \quad \forall t \in \{1, 2, \dots, \tilde{T} - 1\} : \tilde{w}_{t+1}^* \leq \frac{\tilde{\lambda}}{2}.$$

Let $t_0 \leq \tilde{T}/2$ denote the unique index that satisfies $\tilde{w}_{t_0}^* > \frac{\tilde{\lambda}}{2}$, as well as $\tilde{w}_{t_0+1}^* \leq \frac{\tilde{\lambda}}{2}$ for all t between t_0 and $\tilde{T} - 1$ [6]. Using the displayed inequality above, we get that for any $k \leq \tilde{T} - t_0$,

$$\tilde{w}_{t_0+k}^* \geq \frac{(\tilde{w}_{t_0+k-1}^*)^2}{9\tilde{\lambda}} \geq \frac{(\tilde{w}_{t_0+k-2}^*)^4}{(9\tilde{\lambda})^3} \geq \cdots \geq 9\tilde{\lambda} \left(\frac{\tilde{w}_{t_0}^*}{9\tilde{\lambda}}\right)^{2^k} > 9\tilde{\lambda} \left(\frac{\tilde{\lambda}/2}{9\tilde{\lambda}}\right)^{2^k},$$

so we get $\tilde{w}_{t_0+k}^* \geq 9\tilde{\lambda} \cdot (18)^{-2^k}$ as required. $\square$

Lemma 4 $\sum_{i=1}^{\tilde{T}} (\tilde{w}_i^*)^2 \leq 2\gamma^{7/4}/\tilde{\lambda}^{3/2}$

Proof We need to upper bound the squared Euclidean norm of $(\tilde{w}_1^*, \dots, \tilde{w}_{\tilde{T}}^*)$. Note that for any vector $\mathbf{w}$, we have $\|\mathbf{w}\|^2 = \sum_i w_i^2 \leq (\max_i |w_i|) \sum_i |w_i| = \|\mathbf{w}\|_\infty \|\mathbf{w}\|_1$. Thus, by Lemma 1, (13), and the assumption that $\gamma \geq 10^4 (\lambda/\mu_2)^2 > 277\tilde{\lambda}^2$, the squared norm is at most

$$\left(\sqrt{\gamma} + \sqrt{\frac{2\gamma^{3/2}}{\tilde{\lambda}}}\right) \frac{\gamma}{\tilde{\lambda}} = \left(1 + \sqrt{\frac{2\gamma^{1/2}}{\tilde{\lambda}}}\right) \frac{\gamma^{3/2}}{\tilde{\lambda}} \leq \left(\sqrt{\frac{\gamma^{1/2}}{\sqrt{277\tilde{\lambda}}}} + \sqrt{\frac{2\gamma^{1/2}}{\tilde{\lambda}}}\right) \frac{\gamma^{3/2}}{\tilde{\lambda}},$$

which is at most $2\sqrt{\gamma^{1/2}/\tilde{\lambda}} \cdot \gamma^{3/2}/\tilde{\lambda} = 2\gamma^{7/4}/\tilde{\lambda}^{3/2}$ $\square$

Lemma 5 $\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w})$ *satisfies* $\langle \mathbf{v}_i, \mathbf{w}^* \rangle = \tilde{w}_i^*$ *for all* $i = 1, \dots, \tilde{T}$, *where* $\tilde{\mathbf{w}}^* = \arg\min_{\mathbf{w}} \tilde{f}(\mathbf{w})$.

Proof First, we argue that $\tilde{\mathbf{w}}^*$, which minimizes

$$\tilde{f}(\mathbf{w}) = \frac{1}{3} \sum_{i=1}^{\tilde{T}-1} |w_i - w_{i+1}|^3 + \frac{\tilde{\lambda}}{2} \|\mathbf{w}\|^2 - \gamma \cdot w_1 ,$$

---

[6] Since $\tilde{w}_t^*$ monotonically decrease in t, such an index must exist: On the one hand, $\tilde{w}_1^*$ can be verified to be at least $\tilde{\lambda} > \tilde{\lambda}/2$ (by Lemma 2 and the assumption $\gamma \geq 10^4 (\lambda/\mu_2)^2$, hence $\gamma \geq 277\tilde{\lambda}^2$). On the other hand, if we let t_1 be the largest index $\leq \tilde{T}$ satisfying $\tilde{w}_{t_1}^* > \tilde{\lambda}/2$, we have by Lemma 1 that $\frac{\gamma}{\tilde{\lambda}} \geq \sum_{t=1}^{t_1} \tilde{w}_t^* \geq t_1 \tilde{w}_{t_1}^* > \frac{t_1 \tilde{\lambda}}{2}$, which implies that $t_1 \leq \frac{2\gamma}{\tilde{\lambda}^2}$, which is less than $\tilde{T}/2$ by the assumption on $\tilde{T}$ being large enough. Therefore, t_0 is at most $\tilde{T}/2$ as well.

also minimizes

$$\hat{f}(\mathbf{w}) = \sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) + \frac{\tilde{\lambda}}{2} \|\mathbf{w}\|^2 - \gamma \cdot w_1.$$

To see this, note that $\tilde{f}$ and $\hat{f}$ differ only in that $g(x)$ is replaced by $\frac{1}{3}|x|^3$. By definition of g, we have that $g(x)$ and $\frac{1}{3}|x|^3$ coincide for any $|x| \leq \Delta$, from which it is easily verified that f and $\tilde{f}$ have the same values and gradients at any $\mathbf{w}$ for which $|w_i - w_{i+1}| \leq \Delta$ for all $i \leq \tilde{T} - 1$. By Lemma 1 and the assumption $\Delta \geq \sqrt{\gamma}$, the global minimizer $\tilde{\mathbf{w}}^*$ of $\tilde{f}$ belongs to this set, and therefore $\nabla \tilde{f}(\tilde{\mathbf{w}}^*) = \nabla \hat{f}(\tilde{\mathbf{w}}^*) = \mathbf{0}$. But $\hat{f}$ is strongly convex, hence has a unique point (the global minimizer) at which the gradient of $\hat{f}$ is zero, hence $\tilde{\mathbf{w}}^*$ is indeed the global minimizer of $\hat{f}$.

Next, since the global minimizer is invariant to multiplying the function by a fixed positive factor, we get that $\tilde{\mathbf{w}}^*$ is also the global minimizer of

$$\frac{\mu_2}{12} \hat{f}(\mathbf{w}) = \frac{\mu_2}{12} \left(\sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) + \frac{\tilde{\lambda}}{2} \|\mathbf{w}\|^2 - \gamma \cdot w_1 \right)$$

$$= \frac{\mu_2}{12} \left(\sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) - \gamma \cdot w_1 \right) + \frac{\lambda}{2} \|\mathbf{w}\|^2,$$

where in the last step we used the fact that $\tilde{\lambda} = 12\lambda/\mu_2$. Recalling that

$$f(\mathbf{w}) = \frac{\mu}{12} \left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{v}_i, \mathbf{w} \rangle - \langle \mathbf{v}_{i+1}, \mathbf{w} \rangle) - \langle \mathbf{v}_1, \mathbf{w} \rangle \right) + \frac{\lambda}{2} \|\mathbf{w}\|^2,$$

and that $\mathbf{v}_1, \mathbf{v}_2, \ldots$ are orthogonal, we can write $f(\mathbf{w})$ as $\frac{\mu}{12} \cdot \hat{f}(V\mathbf{w})$, where V is any orthogonal matrix with the first $\tilde{T}$ columns being $\mathbf{v}_1, \ldots, \mathbf{v}_{\tilde{T}}$. Therefore, the minimizer $\mathbf{w}^*$ of f satisfies $V\mathbf{w}^* = (\langle \mathbf{v}_1, \mathbf{w}^* \rangle, \langle \mathbf{v}_2, \mathbf{w}_2^* \rangle, \ldots) = \tilde{\mathbf{w}}^*$. $\qquad\square$

Lemma 6 $\|\mathbf{w}^*\|^2 = \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^* \rangle^2$

Proof $f(\mathbf{w})$ is a function which can be written in the form $h(\langle \mathbf{v}_1, \mathbf{w} \rangle, \langle \mathbf{v}_2, \mathbf{w} \rangle, \ldots, \langle \mathbf{v}_{\tilde{T}}, \mathbf{w} \rangle) + \frac{\lambda}{2} \|\mathbf{w}\|^2$, so by the Representer theorem, its minimizer $\mathbf{w}^*$ must lie in the span of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$. Moreover, since these vectors are orthogonal and of unit norm, we have $\mathbf{w}^* = \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^* \rangle \mathbf{v}_i$, and thus $\|\mathbf{w}^*\|^2 = \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^* \rangle^2$. $\qquad\square$

4.2 Oracle complexity lower bound

In this subsection, we prove the following oracle complexity lower bound, depending on the free parameter γ:

Proposition 3 *Assume that $\epsilon < \min\left\{\frac{108^2 \cdot \lambda^3}{\mu_2^2}, \frac{\gamma\lambda}{8}\right\}$. Under the conditions of Proposition 2, it is possible to choose the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$ in the function f, such that the number of iterations T required to have $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ is at least*

$$\max\left\{\frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}}, \ \log_2\log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2\epsilon}\right) - 1\right\}$$

To prove the theorem, we will need the following key lemma, which establishes that oracle information at certain points $\mathbf{w}$ do not leak any information on some of the $\mathbf{v}_1, \mathbf{v}_2, \ldots$ vectors.

Lemma 7 *For any $\mathbf{w} \in \mathbb{R}^d$ orthogonal to $\mathbf{v}_t, \mathbf{v}_{t+1}, \ldots, \mathbf{v}_{\tilde{T}}$, it holds that $f(\mathbf{w}), \nabla f(\mathbf{w}), \nabla^2 f(\mathbf{w})$ do not depend on $\mathbf{v}_{t+1}, \mathbf{v}_{t+2}, \ldots, \mathbf{v}_{\tilde{T}}$.*

Proof Since the regularization term $\frac{\lambda}{2}\|\mathbf{w}\|^2$ doesn't depend on $\mathbf{v}_{t+1}, \mathbf{v}_{t+2}, \ldots, \mathbf{v}_{\tilde{T}}$ we can define $h(\mathbf{w}) := f(\mathbf{w}) - \frac{\lambda}{2}\|\mathbf{w}\|^2$ and prove the result on $h(\mathbf{w})$. Using the definition of h and differentiating, we have that

$$h(\mathbf{w}) = \frac{\mu_2}{12}\left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \gamma\langle \mathbf{v}_1, \mathbf{w}\rangle\right)$$

$$\nabla h(\mathbf{w}) = \frac{\mu_2}{12}\left(\sum_{i=1}^{\tilde{T}-1} g'(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)(\mathbf{v}_i - \mathbf{v}_{i+1}) - \gamma\mathbf{v}_1\right)$$

$$\nabla^2 h(\mathbf{w}) = \frac{\mu_2}{12}\left(\sum_{i=1}^{\tilde{T}-1} g''(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)(\mathbf{v}_i - \mathbf{v}_{i+1})(\mathbf{v}_i - \mathbf{v}_{i+1})^T\right)$$

By the assumption $\langle \mathbf{v}_t, \mathbf{w}\rangle = \langle \mathbf{v}_{t+1}, \mathbf{w}\rangle = \ldots = 0$, and the fact that $g(0) = g'(0) = g''(0) = 0$, we have that $g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) = g'(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) = g''(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) = 0$ for all $i \in \{t, t+1, \ldots, \tilde{T}-1\}$. Therefore, it is easily verified that the expressions above indeed do not depend on $\mathbf{v}_{t+1}, \mathbf{v}_{t+2}, \ldots, \mathbf{v}_{\tilde{T}}$. $\square$

Let us now fix any number of iterations $T \leq \tilde{T}$. Using the previous results, we now argue that given any deterministic algorithm, we can choose $\mathbf{v}_1, \ldots, \mathbf{v}_{\tilde{T}}$ so that one of them will necessarily be orthogonal to $\mathbf{w}_T$, the algorithm's output after T iterations:

Lemma 8 *For any deterministic algorithm, it is possible to choose $\mathbf{v}_1, \ldots, \mathbf{v}_{\tilde{T}}$, so that $\mathbf{v}_T, \mathbf{v}_{T+1}, \ldots, \mathbf{v}_{\tilde{T}}$ are orthogonal to both $\mathbf{w}_1, \ldots, \mathbf{w}_T$ and $\mathbf{v}_1, \ldots, \mathbf{v}_T$.*

Proof The proof is constructive, and uses an argument similar to other oracle complexity lower bounds in the literature. It proceeds as follows: $\square$

– First, we compute $\mathbf{w}_1$ (which is possible since the algorithm is deterministic and $\mathbf{w}_1$ is chosen before any oracle calls are made).

- We pick $\mathbf{v}_1$ to be some unit vector orthogonal to $\mathbf{w}_1$. Assuming $\mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$ will also be orthogonal to $\mathbf{w}_1$ (which will be ensured by the construction which follows), we have by Lemma 7 that the information $F(\mathbf{w}_1), \nabla F(\mathbf{w}_1), \nabla^2 F(\mathbf{w}_1)$ provided by the oracle to the algorithm does not depend on $\{\mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}\}$, and thus depends only on $\mathbf{v}_1$ which was already fixed. Since the algorithm is deterministic, this fixes the next query point $\mathbf{w}_2$.
- For $t = 2, 3, \ldots, T - 1$, we repeat the process above: We compute $\mathbf{w}_t$, and pick $\mathbf{v}_t$ to be some unit vectors orthogonal to $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_t$, as well as all previously constructed $\mathbf{v}$'s (this is always possible since the dimension is sufficiently large). By Lemma 7, as long as all vectors thus constructed are orthogonal to $\mathbf{w}_t$, the information $\{F(\mathbf{w}_t), \nabla F(\mathbf{w}_t), \nabla^2 F(\mathbf{w}_t)\}$ provided to the algorithm does not depend on $\mathbf{v}_{t+1}, \ldots, \mathbf{v}_{\tilde{T}}$, and only depends on $\mathbf{v}_1, \ldots, \mathbf{v}_t$ which were already determined. Therefore, the next query point $\mathbf{w}_{t+1}$ is fixed.
- At the end of the process, we pick $\mathbf{v}_T, \mathbf{v}_{T+1}, \ldots, \mathbf{v}_{\tilde{T}}$ to be some unit vectors orthogonal to all previously chosen $\mathbf{v}$'s as well as $\mathbf{w}_1, \ldots, \mathbf{w}_T$ (this is possible since the dimension is large enough). □

Using the facts that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$ are orthogonal (and thus act as an orthonormal basis to a subspace of $\mathbb{R}^d$), that $\mathbf{w}_T$ is orthogonal to $\mathbf{v}_T, \mathbf{v}_{T+1}, \ldots \mathbf{v}_{\tilde{T}}$, and that $t_0 + T \leq \frac{\tilde{T}}{2} + \frac{\tilde{T}}{2} = \tilde{T}$ (where t_0 is as defined in Proposition 2), we have $\|\mathbf{w}_T - \mathbf{w}^*\|^2 \geq \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}_T - \mathbf{w}^* \rangle^2 \geq \langle \mathbf{v}_{t_0+T}, \mathbf{w}_T - \mathbf{w}^* \rangle^2 = \langle \mathbf{v}_{t_0+T}, \mathbf{w}^* \rangle^2$. By Proposition 2, we can lower bound the above by $\frac{108^2 \lambda^2}{\mu_2^2} \cdot (18)^{-2(T+1)}$. Using the strong convexity of f, we therefore get

$$f(\mathbf{w}_T) - f(\mathbf{w}^*) \geq \frac{\lambda}{2} \cdot \|\mathbf{w}_T - \mathbf{w}^*\|^2 \geq \frac{108^2 \cdot \lambda^3}{\mu_2^2} (18)^{-2(T+1)}.$$

To make the right-hand side smaller than ϵ, T must satisfy $(18)^{-2(T+1)} \leq \frac{\mu_2^2 \epsilon}{108^2 \cdot \lambda^3}$, which is equivalent to $2^{(T+1)} \geq \log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2 \epsilon}\right)$. Assuming $\epsilon < \frac{108^2 \cdot \lambda^3}{\mu_2^2}$, then

$$T \geq \log_2 \log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2 \epsilon}\right) - 1.$$

We now turn to argue that we can also lower bound T by $\frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}}$. Otherwise, suppose by contradiction that we can have $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ for some $T < \frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}}$. From Proposition 2 we know that $\langle \mathbf{v}_T, \mathbf{w}^* \rangle \geq \frac{\gamma^{3/4}}{7\sqrt{12\lambda/\mu_2}} + \sqrt{\gamma}\left(\frac{1}{2} - T\right)$, so as before, we have that $\|\mathbf{w}_T - \mathbf{w}^*\|^2 \geq \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}_T - \mathbf{w}^* \rangle^2 \geq \langle \mathbf{v}_T, \mathbf{w}_T - \mathbf{w}^* \rangle^2 = \langle \mathbf{v}_T, \mathbf{w}^* \rangle^2$, and thus

$$f(\mathbf{w}_T) - f(\mathbf{w}^*) \geq \frac{\lambda}{2} \cdot \|\mathbf{w}_T - \mathbf{w}^*\|^2 \geq \frac{\lambda}{2} \cdot \langle \mathbf{v}_T, \mathbf{w}^* \rangle^2$$

$$\geq \frac{\lambda}{2}\left(\frac{\gamma^{3/4}}{7\sqrt{12\lambda/\mu_2}} + \sqrt{\gamma}\left(\frac{1}{2} - T\right)\right)^2.$$

To make the right-hand side smaller than ϵ, T must satisfy $\left(\frac{\gamma^{3/4}}{7\sqrt{12\lambda/\mu_2}} + \sqrt{\gamma}\left(\frac{1}{2} - T\right)\right)^2 \leq \frac{2\epsilon}{\lambda}$, or equivalently $T \geq \frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}} + \frac{1}{2} - \sqrt{\frac{2\epsilon}{\gamma\lambda}}$. But since we assume $\epsilon < \frac{\gamma\lambda}{8}$, this is at least $\frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}}$, contradicting our earlier assumption.

Overall, we showed that T is lower bounded by both $\frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}}$, as well as $\log_2 \log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2 \epsilon}\right) - 1$, hence proving Proposition 3.

4.3 Setting the γ, Δ Parameters

In the following lemma, we establish the strong convexity and smoothness parameters of f (depending on the parameter Δ which is still free at this point).

Lemma 9 *f is λ-strongly convex and twice-differentiable, with μ_2-Lipschitz Hessians and $\left(\frac{2\mu_2\Delta}{3} + \lambda\right)$-Lipschitz gradients.*

Proof Since f is a sum of convex, twice-differentiable functions and the λ-strongly convex function $\frac{\lambda}{2}\|\mathbf{w}\|^2$, it is clearly λ-strongly convex and twice-differentiable. Thus, it only remains to calculate the Lipschitz parameter of the gradients and Hessians.

To simplify the proof, we note that Lipschitz smoothness is a property invariant to the coordinate system used, so we can assume without loss of generality that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$ correspond to the standard basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{\tilde{T}}$, and consider the Lipschitz properties of the function

$$\hat{f}(\mathbf{w}) = \frac{\mu_2}{12}\left(\sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) - \gamma \cdot w_1\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

By definition of g, it is easily verified that $g''(x) = 2 \cdot \min\{\Delta, |x|\}$, which is a 2-Lipschitz function bounded in $[0, 2\Delta]$. This implies that $g'(x)$ is 2Δ-Lipschitz. Letting $\mathbf{r}_i := \mathbf{e}_i - \mathbf{e}_{i+1}$, we can write $\hat{f}$ as $\frac{\mu_2}{12}\left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{r}_i, \mathbf{w}\rangle) - \gamma \cdot w_1\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. Differentiating twice, we get $\nabla^2 \hat{f}(\mathbf{w}) = \frac{\mu_2}{12}\sum_{i=1}^{\tilde{T}-1} g''(\langle \mathbf{r}_i, \mathbf{w}\rangle) \cdot \mathbf{r}_i \mathbf{r}_i^\top + \lambda I$. Since this is a sum of positive-semidefinite matrices with non-negative coefficients (as we showed that $g''(x) \in [0, 2\Delta]$ for all x), it follows that its spectral norm is at most $\frac{\mu_2\Delta}{6} \cdot \left\|\sum_{i=1}^{\tilde{T}-1} \mathbf{r}_i \mathbf{r}_i^\top\right\| + \lambda$, and the first term equals

$$\frac{\mu_2\Delta}{6} \cdot \max_{\mathbf{x}} \frac{\sum_{i=1}^{\tilde{T}-1}\langle \mathbf{r}_i, \mathbf{x}\rangle^2}{\|\mathbf{x}\|^2} = \frac{\mu_2\Delta}{6} \cdot \max_{\mathbf{x}} \frac{\sum_{i=1}^{\tilde{T}-1}(x_i - x_{i+1})^2}{\|\mathbf{x}\|^2}$$

$$\leq \frac{\mu_2\Delta}{6} \cdot \max_{\mathbf{x}} \frac{\sum_{i=1}^{\tilde{T}-1}(2x_i^2 + 2x_{i+1}^2)}{\sum_{i=1}^d x_i^2} \leq \frac{2\mu_2\Delta}{3}.$$

Overall, we showed that $\|\nabla^2 \hat{f}(\mathbf{w})\| \leq \frac{2\mu_2 \Delta}{3} + \lambda$, so the gradients of f are $\left(\frac{2\mu_2 \Delta}{3} + \lambda\right)$-Lipschitz.

It remains to show that $\nabla^2 \hat{f}(\mathbf{w})$ is μ_2-Lipschitz. Using the formula for $\nabla^2 \hat{f}(\mathbf{w})$, and recalling that $g''(x)$ is 2-Lipschitz, and $\|\mathbf{r}_i\| = \sqrt{2}$ by definition, we have that for any $\mathbf{w}, \tilde{\mathbf{w}}$,

$$
\begin{aligned}
\|\nabla^2 \hat{f}(\mathbf{w}) - \nabla^2 \hat{f}(\tilde{\mathbf{w}})\| &= \frac{\mu_2}{12} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} (g''(\langle \mathbf{r}_i, \mathbf{w}\rangle) - g''(\langle \mathbf{r}_i, \tilde{\mathbf{w}}\rangle)) \cdot \mathbf{r}_i \mathbf{r}_i^\top \right\| \\
&\leq \frac{\mu_2}{12} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} |g''(\langle \mathbf{r}_i, \mathbf{w}\rangle) - g''(\langle \mathbf{r}_i, \tilde{\mathbf{w}}\rangle)| \cdot \mathbf{r}_i \mathbf{r}_i^\top \right\| \\
&\leq \frac{\mu_2}{12} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} 2|\langle \mathbf{r}_i, \mathbf{w} - \mathbf{w}'\rangle| \cdot \mathbf{r}_i \mathbf{r}_i^\top \right\| \leq \frac{\mu_2}{12} \cdot 2\sqrt{2} \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\| \cdot \left\| \sum_{i=1}^{\tilde{T}-1} \mathbf{r}_i \mathbf{r}_i^\top \right\|.
\end{aligned}
$$

Using the same calculations as earlier, we have $\left\| \sum_{i=1}^{\tilde{T}-1} \mathbf{r}_i \mathbf{r}_i^\top \right\| \leq 4$, and therefore we showed overall that $\|\nabla^2 \hat{f}(\mathbf{w}) - \nabla^2 \hat{f}(\tilde{\mathbf{w}})\| \leq \frac{\mu_2 \cdot 8\sqrt{2}}{12} \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\| < \mu_2 \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|$, hence $\nabla^2 \hat{f}(\mathbf{w})$ is μ_2-Lipschitz. $\qquad\square$

We now collect the ingredients necessary to fix γ, Δ and hence prove our theorem. Combining the previous lemma, Proposition 2 and Proposition 3, and recalling that we want f to have μ_1-Lipschitz gradients and μ_2-Lipschitz Hessians, with an optimizer $\mathbf{w}^*$ satisfying $\|\mathbf{w}^*\| \leq D$, we have an oracle complexity lower bound of the form

$$
T \geq \max\left\{ \frac{\gamma^{1/4}}{7\sqrt{12\lambda/\mu_2}}, \ \log_2 \log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2 \epsilon}\right) - 1 \right\}, \tag{17}
$$

assuming the following conditions: $\gamma \geq 10^4 \left(\frac{\lambda}{\mu_2}\right)^2, \Delta \geq \sqrt{\gamma}, \epsilon < \min\left\{\frac{108^2 \cdot \lambda^3}{\mu_2^2}, \frac{\gamma\lambda}{8}\right\}$, $\frac{2\mu_2 \Delta}{3} + \lambda \leq \mu_1, \sqrt{\frac{2\gamma^{7/4}}{(12\lambda/\mu_2)^{3/2}}} \leq D$. Picking $\Delta = \sqrt{\gamma}$, using the fact that $\mu_1 \geq \lambda$ (as any λ-strongly convex function must have gradients with Lipschitz parameter at least λ), and rewriting the last two conditions, this is equivalent to

$$
\gamma \geq 10^4 \left(\frac{\lambda}{\mu_2}\right)^2, \epsilon < \min\left\{\frac{108^2 \cdot \lambda^3}{\mu_2^2}, \frac{\gamma\lambda}{8}\right\}, \gamma \leq \left(\frac{3(\mu_1 - \lambda)}{2\mu_2}\right)^2, \gamma \leq \sqrt[7]{\frac{D^8 (12\lambda)^6}{2^4 \mu_2^6}}.
$$

Since the first condition needs to hold anyway, we can allow ourself to make the second condition stronger, by substituting $10^4 (\lambda/\mu_2)^2$ in lieu of γ in the second condition. Doing this, simplifying, and merging the last two conditions, the set of condition above is implied by requiring

$$\gamma \geq 10^4 \left(\frac{\lambda}{\mu_2}\right)^2, \; \epsilon < \frac{10^4 \lambda^3}{8\mu_2^2}, \; \gamma \leq \min\left\{\left(\frac{3(\mu_1 - \lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(12\lambda)^6}{2^4\mu_2^6}}\right\}.$$

Clearly, to make the lower bound in (17) as large as possible, we should pick the largest possible γ, namely $\gamma = \min\left\{\left(\frac{3(\mu_1-\lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(12\lambda)^6}{2^4\mu_2^6}}\right\}$, and to ensure that the other conditions hold, require that

$$\min\left\{\left(\frac{3(\mu_1-\lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(12\lambda)^6}{2^4\mu_2^6}}\right\} \geq 10^4\left(\frac{\lambda}{\mu_2}\right)^2, \; \epsilon < \frac{10^4\lambda^3}{8\mu_2^2}.$$

Simplifying a bit, these two conditions are implied by requiring

$$\frac{\mu_1}{\lambda} \geq 68, \; \frac{\mu_2}{\lambda}D \geq 694, \; \epsilon < \frac{10^4\lambda^3}{8\mu_2^2}, \tag{18}$$

Finally, let us plug our choice of $\gamma = \min\left\{\left(\frac{3(\mu_1-\lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(12\lambda)^6}{2^4\mu_2^6}}\right\}$ into the lower bound in (17). We thus get an oracle complexity lower bound of

$$\max\left\{\frac{\sqrt{\mu_2}}{7\sqrt{12\lambda}}\min\left\{\sqrt{\frac{3(\mu_1-\lambda)}{2\mu_2}}, \frac{D^{2/7}(12\lambda)^{3/14}}{2^{1/7}\mu_2^{3/14}}\right\}, \log_2\log_{18}\left(\frac{108^2\lambda^3}{\mu_2^2\epsilon}\right) - 1\right\}$$

$$= \max\left\{\min\left\{\frac{1}{14}\sqrt{\frac{\mu_1-\lambda}{2\lambda}}, \frac{(D\mu_2/12\lambda)^{2/7}}{7 \cdot 2^{1/7}}\right\}, \log_2\log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2\epsilon}\right) - 1\right\},$$

under the conditions of (18).

To simplify the bound a bit, we note that we can lower bound $\mu_1 - \lambda$ by $\frac{67}{68}\mu_1$ (possible by (18)), and lower bound $\log_2\log_{18}\left(\frac{108^2 \cdot \lambda^3}{\mu_2^2\epsilon}\right) - 1$ by $\frac{1}{2}\log\log_{18}\left(\frac{\lambda^3}{\mu_2^2\epsilon}\right)$, by assuming that $\epsilon \leq c\lambda^3/\mu_2^2$ for some small enough c (in other words, increasing the constant in the third condition in (18)). Finally, using the fact that $\max\{a, b\} \geq (a + b)/2$, the result in the theorem follows.

5 Proof of Theorem 2

The proof follows the lines of Proposition 1 and its proof sketch, but with a more complicated construction (as we need to capture the dependence on the Lipschitz parameters of both the gradients and the Hessians).

Similarly to the strongly convex case, we will assume without loss of generality that the algorithm initializes at $\mathbf{w}_1 = \mathbf{0}$, since otherwise one can simply replace the "hard" function $f(\mathbf{w})$ below by $f(\mathbf{w} - \mathbf{w}_1)$, and the same proof holds verbatim. Thus, the theorem requires that our function has a minimizer $\mathbf{w}^*$ satisfying $\|\mathbf{w}^*\| \leq D$.

Define $g : \mathbb{R} \mapsto \mathbb{R}$ as

$$
g(x) = \begin{cases} \frac{1}{3}|x|^3 & |x| \le \Delta \\ \Delta x^2 - \Delta^2 |x| + \frac{1}{3}\Delta^3 & |x| > \Delta, \end{cases}
$$

where $\Delta := \frac{3\mu_1}{2\mu_2}$. The function g can be easily verified to be twice continuously differentiable. Assume that $d \ge 2\bar{T}$, and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\bar{T}}$ be orthogonal unit vectors in $\mathbb{R}^d$ which will be specified later. Given $\bar{T}$, and letting $\gamma > 0$ be a parameter to be specified later, define the function $f_{\bar{T}}$ as

$$
f_{\bar{T}}(\mathbf{w}) = \frac{\mu_2}{12} \left(g(\langle \mathbf{v}_1, \mathbf{w} \rangle) + g(\langle \mathbf{v}_{\bar{T}}, \mathbf{w} \rangle) + \sum_{i=1}^{\bar{T}-1} g(\langle \mathbf{v}_i, \mathbf{w} \rangle - \langle \mathbf{v}_{i+1}, \mathbf{w} \rangle) - \gamma \langle \mathbf{v}_1, \mathbf{w} \rangle \right)
$$

$$
= \frac{\mu_2}{12} \left(g(\langle \mathbf{v}_1, \mathbf{w} \rangle) + g(\langle \mathbf{v}_{\bar{T}}, \mathbf{w} \rangle) + \sum_{i=1}^{\bar{T}-1} g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w} \rangle) - \gamma \langle \mathbf{v}_1, \mathbf{w} \rangle \right)
$$

This function is easily shown to be convex and twice-differentiable, with μ_1-Lipschitz gradients and μ_2-Lipschitz Hessians (the proof is identical to the proof of Lemma 9). Our goal will be to show a lower bound on the optimization error using this type of function.

5.1 Minimizer of $f_{\bar{T}}$

In this subsection, we analyze the properties of a minimizer of $f_{\bar{T}}$. To that end, we introduce the following function in $\mathbb{R}^{\bar{T}}$:

$$
\hat{f}_{\bar{T}}(\mathbf{w}) = g(w_1) + g(w_{\bar{T}}) + \sum_{i=1}^{\bar{T}-1} g(w_i - w_{i+1}) - \gamma w_1.
$$

It is easily verified that the minimal values of $\frac{\mu_2}{12} \hat{f}_{\bar{T}}$ and $f_{\bar{T}}$ are the same, and moreover, if $\hat{\mathbf{w}} \in \mathbb{R}^{\bar{T}}$ is a minimizer of $\hat{f}_{\bar{T}}$, then $\mathbf{w}^* = \sum_{j=1}^{\bar{T}} \hat{w}_j^* \cdot \mathbf{v}_j \in \mathbb{R}^d$ is a minimizer of $f_{\bar{T}}$, and with the same Euclidean norm as $\hat{\mathbf{w}}^*$.

We begin with the following technical lemma:

Lemma 10 $\hat{f}_{\bar{T}}$ *has a unique minimizer* $\hat{\mathbf{w}}^* \in \mathbb{R}^{\bar{T}}$, *which satisfies*

$$
\hat{w}_t^* = \delta \cdot (\bar{T} + 1) \cdot \left(1 - \frac{t}{\bar{T} + 1} \right),
$$

for all $t = 1, 2, \ldots, \bar{T}$, *where* δ *is non-negative and independent of t. Moreover,*

$$
g'(\hat{w}_1^*) + g'(\hat{w}_{\bar{T}}^*) = \gamma.
$$

Proof Taking the derivative and setting to zero, we get that the

$$g'(\hat{w}_1^*) + g'(\hat{w}_1^* - \hat{w}_2^*) = \gamma, \quad g'(\hat{w}_{\bar{T}-1}^* - \hat{w}_{\bar{T}}^*) = g'(\hat{w}_{\bar{T}}^*)$$

as well as

$$g'(\hat{w}_{j-1}^* - \hat{w}_j^*) = g'(\hat{w}_j^* - \hat{w}_{j+1}^*)$$

for all $j \in \{2, 3, \ldots, \bar{T} - 1\}$. By definition of g, it is easily verified that g' is a strictly monotonic (hence invertible) function, so the above implies $\hat{w}_{j-1}^* - \hat{w}_j^* = \hat{w}_j^* - \hat{w}_{j+1}^*$ for all $j \in \{2, 3, \ldots, \bar{T} - 1\}$, as well as $\hat{w}_{\bar{T}-1}^* - \hat{w}_{\bar{T}}^* = \hat{w}_{\bar{T}}^*$. From this, it follows by straightforward induction that $\hat{w}_{\bar{T}+1-t}^* = t \cdot \hat{w}_{\bar{T}}^*$, from which the first displayed equation in the lemma follows. This also implies $g'(\bar{T} \hat{w}_{\bar{T}}^*) + g'(\hat{w}_{\bar{T}}^*) = \gamma$, and since g' is strictly monotonic, we have that $\hat{w}_{\bar{T}}^*$ is uniquely defined, and since the other coordinates of $\hat{\mathbf{w}}^*$ are also uniquely defined given $\hat{w}_{\bar{T}}^*$, we get that $\hat{\mathbf{w}}^*$ is unique. Finally, δ (and hence $\hat{w}_t^*$ for all t) is necessarily non-negative, since otherwise $\hat{w}_1^*$ is negative, which would imply $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*) > 0$, even though $\hat{f}_{\bar{T}}(\mathbf{0}) = 0$, violating the fact that $\hat{\mathbf{w}}^*$ minimizes $\hat{f}_{\bar{T}}$.   □

The main technical result in this subsection is the following proposition, which characterizes $\|\hat{\mathbf{w}}^*\|$ and $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*)$ under various parameter regimes. By the discussion above and definition of $f_{\bar{T}}$, we have

$$\|\mathbf{w}^*\| = \|\hat{\mathbf{w}}^*\| \quad \text{and} \quad f_{\bar{T}}(\mathbf{w}^*) = \frac{\mu_2}{12} \cdot \hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*), \tag{19}$$

which will be used in the remainder of the proof of our theorem.

Proposition 4 *The values of the minimizer of $\hat{f}_{\bar{T}}$ and the corresponding minimum $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*)$ for different γ regimes (which depend on Δ and $\bar{T}$) are summarized in Table 1.*

Table 1 Properties of $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*)$ and $\|\hat{\mathbf{w}}^*\|^2$ for different γ regimes

	Regimes		
	$\gamma \leq \dfrac{\Delta^2\left(1+\bar{T}^2\right)}{\bar{T}^2}$	$\dfrac{\Delta^2\left(1+\bar{T}^2\right)}{\bar{T}^2} < \gamma \leq 2\Delta^2\bar{T}$	$\gamma > 2\Delta^2\bar{T}$
$\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*)$	$-\dfrac{2\mu_2\gamma^{3/2}\bar{T}}{3\sqrt{(1+\bar{T}^2)}}$	$\frac{1}{3}\bar{T}\delta^3 + \Delta\bar{T}^2\delta^2 - \bar{T}\left(\Delta^2+\gamma\right)\delta + \frac{\Delta^3}{3}$ where $\delta = -\Delta\bar{T} + \Delta\bar{T}\sqrt{1 + \frac{\gamma+\Delta^2}{\Delta^2\bar{T}^2}}$	$-\dfrac{\bar{T}\left(\gamma+2\Delta^2\right)^2}{4\Delta(\bar{T}+1)} + \dfrac{(\bar{T}+1)\Delta^3}{3}$
$\|\hat{\mathbf{w}}^*\|^2$	$\leq \dfrac{\gamma(1+\bar{T})^3}{3(1+\bar{T}^2)}$	$\leq \dfrac{\left(\gamma+\Delta^2\right)^2(\bar{T}+1)^3}{12\Delta^2\bar{T}^2}$	$\leq \dfrac{(\bar{T}+1)\left(\gamma+2\Delta^2\right)^2}{12\Delta^2}$

Proof To prove the proposition, we will consider three regimes, depending on $\bar{T}, \delta, \Delta$ (Table 1): Namely, $\bar{T}\delta \leq \Delta$, $\frac{\Delta}{\bar{T}} < \delta \leq \Delta$ and $\delta > \Delta$. We will show that each regime corresponds to one of the three regimes specified in the proposition, and prove the relevant bounds.

Case 1: $\bar{T}\delta \leq \Delta$. In that case, $\hat{w}_1^*$, $\hat{w}_{\bar{T}}^*$ as well as $\hat{w}_i^* - \hat{w}_{i+1}^*$ for all $i = 2, \ldots, \bar{T} - 1$ in the definition of $\hat{f}_{\bar{T}}$ all lie in the interval where g is a cubic function. Using Lemma 10, $g'(w_1^*) + g'(w_{\bar{T}}^*) = w_1^{*2} + w_{\bar{T}}^{*2} = \gamma$, hence $\delta^2 \bar{T}^2 + \delta^2 = \gamma$ and $\delta = \sqrt{\frac{\gamma}{1+\bar{T}^2}}$. Therefore, our condition $\bar{T}\delta \leq \Delta$ is exactly equivalent to $\gamma \leq \frac{\Delta^2(1+\bar{T}^2)}{\bar{T}^2}$, namely the first regime discussed in the proposition. We now establish the relevant bounds. By plugging the optimal solution $\hat{\mathbf{w}}^*$ in $\hat{f}_{\bar{T}}(\mathbf{w})$, we have that

$$\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*) = -\frac{2\gamma^{3/2}\bar{T}}{3\sqrt{(1+\bar{T}^2)}}$$

and

$$\|\hat{\mathbf{w}}^*\|_2^2 = \sum_{t=1}^{\bar{T}} \hat{w}_t^{*2} = \frac{\gamma(1+\bar{T})^2}{(1+\bar{T}^2)} \sum_{t=1}^{\bar{T}} \left(1 - \frac{t}{1+\bar{T}}\right)^2 \leq \frac{\gamma(1+\bar{T})^3}{3(1+\bar{T}^2)},$$

where in the calculation above we used fact $\sum_{t=1}^{\bar{T}} t^2 \leq \int_1^{\bar{T}+1} t^2 dt < \frac{(\bar{T}+1)^3}{3}$.

Case 2: $\frac{\Delta}{\bar{T}} < \delta \leq \Delta$. In this case, by Lemma 10, $\hat{w}_{\bar{T}}^* \leq \Delta$ but $\hat{w}_1^* > \Delta$. Therefore, in the definition $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*)$, $g(\hat{w}_1^*)$ lies in the quadratic region of g, whereas $g(\hat{w}_{\bar{T}}^*)$ and $g'(\hat{w}_i^* - \hat{w}_{i+1}^*)$ for all i lies in the cubic region of g. As a result, $g'(w_1^*) + g'(w_{\bar{T}}^*) = 2\Delta w_1^* - \Delta^2 + w_{\bar{T}}^{*2} = \gamma$. Plugging in $w_{\bar{T}}^* = \delta$ and $w_1^* = \bar{T} \cdot \delta$, we get $\delta^2 + 2\Delta\delta \cdot \bar{T} - (\gamma + \Delta^2) = 0$, and therefore (using the fact $\delta \geq 0$, see Lemma 10), $\delta = -\Delta\bar{T} + \Delta\bar{T}\sqrt{1 + \frac{\gamma+\Delta^2}{\Delta^2\bar{T}^2}}$. This, plus the assumption $\frac{\Delta}{\bar{T}} < \delta \leq \Delta$, is equivalent to $\frac{\Delta^2(1+\bar{T}^2)}{\bar{T}^2} < \gamma \leq 2\Delta^2\bar{T}$, hence showing that we are indeed in the second regime as specified in our proposition.

Turning to calculate the relevant bounds, we have $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*) = \bar{T}\delta^3 + \Delta^2\bar{T}^2\delta^2 - \bar{T}(\Delta^2 + \gamma)\delta + \frac{\Delta^3}{3}$. Moreover, $\|\hat{\mathbf{w}}^*\|^2 = \delta^2(\bar{T} + 1)^2 \sum_{t=1}^{\bar{T}} \left(1 - \frac{t}{1+\bar{T}}\right)^2$, which by definition of δ above and the inequality $\sqrt{1+x} \leq 1 + \frac{1}{2}x$ for all $x \geq 0$, is at most $\frac{(\gamma+\Delta^2)^2(\bar{T}+1)^3}{12\Delta^2\bar{T}^2}$.

Case 3: $\delta > \Delta$. In this case, by Lemma 10, we have $\hat{w}_1^* > \hat{w}_{\bar{T}}^* = \hat{w}_i^* - \hat{w}_{i+1}^* > \Delta$, which implies that in the definition of $\hat{f}_t(\hat{\mathbf{w}}^*)$, these terms all lie in the quadratic region of g. Therefore, $g'(w_1^*) + g'(w_{\bar{T}}^*) = 2\Delta w_1^* - \Delta^2 + 2\Delta w_{\bar{T}}^* - \Delta^2 = \gamma$, and thus $2\Delta(\bar{T} + 1)\delta = \gamma + 2\Delta^2$, or equivalently $\delta = \frac{\gamma+2\Delta^2}{2\Delta(\bar{T}+1)}$. Note that this, plus our assumption $\delta > \Delta$, is equivalent to $\gamma > 2\Delta^2\bar{T}$, which shows that we are indeed in the third regime as specified in our proposition. Turning to calculate $\|\hat{\mathbf{w}}^*\|$ and $\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*)$, we have

$$\hat{f}_{\bar{T}}(\hat{\mathbf{w}}^*) = -\frac{\bar{T}\left(\gamma + 2\Delta^2\right)^2}{4\Delta(\bar{T}+1)} + \frac{(\bar{T}+1)\Delta^3}{3},$$

and $\|\hat{\mathbf{w}}^*\|^2 = \frac{(\gamma+2\Delta^2)^2}{4\Delta^2}\sum_{t=1}^{\bar{T}}\left(1 - \frac{t}{1+\bar{T}}\right)^2 \leq \frac{(\bar{T}+1)(\gamma+2\Delta^2)^2}{12\Delta^2}.$ □

5.2 Oracle complexity lower bound

Given the expressions on the optimal value of $\hat{f}_{\bar{T}}$, derived in the previous subsection, we turn to explain how the oracle complexity lower bound is derived. The argument is very similar to the strongly convex case (proof of Theorem 1, Sect. 4.2). Specifically, consider the case when $\bar{T} = 2T$, given by

$$f_{2T}(\mathbf{w}) = \frac{\mu_2}{12}\left(g(\langle \mathbf{v}_1, \mathbf{w}\rangle) + g(\langle \mathbf{v}_{2T}, \mathbf{w}\rangle) + \sum_{i=1}^{2T-1} g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \gamma\langle \mathbf{v}_1, \mathbf{w}\rangle\right).$$

Given an algorithm, we choose $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{2T}$ to be orthogonal unit vectors, so that each $\mathbf{v}_t$ is orthogonal to the first t points $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_t$ computed by the algorithm (this is possible by an argument identical to Lemma 8).

With this choice, it is easily verified that $f_{2T}(\mathbf{w}_T)$ equals

$$\frac{\mu_2}{12}\left(g(\langle \mathbf{v}_1, \mathbf{w}_T\rangle) + g(\langle \mathbf{v}_T, \mathbf{w}_T\rangle) + \sum_{i=1}^{T-1} g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}_T\rangle) - \gamma\langle \mathbf{v}_1, \mathbf{w}_T\rangle\right),$$

which is clearly greater than $\min_{\mathbf{w}} f_T(\mathbf{w})$, where f_T is defined with the same $\mathbf{v}_1, \ldots, \mathbf{v}_T$. Therefore, we can lower bound the optimization error $f_{2T}(\mathbf{w}_T) - \min_{\mathbf{w}} f_{2T}(\mathbf{w})$ by $\min_{\mathbf{w}} f_T(\mathbf{w}) - \min_{\mathbf{w}} f_{2T}(\mathbf{w})$. Moreover, by (19), this equals $\frac{\mu_2}{12}\left(\min_{\mathbf{w}} \hat{f}_T(\mathbf{w}) - \min_{\mathbf{w}} \hat{f}_{2T}(\mathbf{w})\right)$. Using Proposition 4, we can now plug in these minimal values, depending on the various parameter regimes, and get an oracle complexity lower bound. Computing these bounds and parameter regimes (while picking the free parameter γ appropriately) is performed in the next subsection.

5.3 Setting the γ Parameter

To simplify notation, we let $\hat{f}_T^*$ and $\hat{f}_{2T}^*$ be shorthand for $\min_{\mathbf{w}} \hat{f}_T(\mathbf{w})$ and $\min_{\mathbf{w}} \hat{f}_{2T}(\mathbf{w})$ respectively, with minimizers $\hat{\mathbf{w}}_T^*$ and $\hat{\mathbf{w}}_{2T}^*$. We will consider three regimes, depending on the relationships between D, Δ, T.

5.3.1 Case 1: $\frac{D^2}{48\Delta^2 T^3} \leq \frac{1}{T^2}$

In this setting, we choose $\gamma = \frac{D^2}{48T}$. Using this and the assumption on the parameters, we get that $\gamma \leq \Delta^2 < \frac{\Delta^2(1+4T^2)}{4T^2} < \frac{\Delta^2(1+T^2)}{T^2}$, and therefore, we are in the first regime for both f_T and f_{2T} as specified in Proposition 4. Plugging in the bound on

$\|\hat{\mathbf{w}}^*\|^2$ in that regime, and using the fact that $\Delta^2 \leq \gamma$ by the assumption above, we have $\|\hat{\mathbf{w}}_{2T}^*\|_2^2 \leq \frac{\gamma(1+2T)^3}{3(1+4T^2)} \leq \frac{27D^2T^3}{576T^2} \leq D^2$ as required.

Using the results from Proposition 4 for the first regime we can compute the optimization error bound

$$\hat{f}_T^* - \hat{f}_{2T}^* = \frac{4\gamma^{3/2}T}{3\sqrt{(1+4T^2)}} - \frac{2\gamma^{3/2}T}{3\sqrt{(1+T^2)}}$$

$$= \frac{2\gamma^{3/2}}{3}\left(\frac{1}{\sqrt{\left(1+\frac{1}{4T^2}\right)}} - \frac{1}{\sqrt{\left(1+\frac{1}{T^2}\right)}}\right)$$

$$\geq \frac{2\gamma^{3/2}}{3}\left(1 - \frac{1}{8T^2} - \left(1 - \frac{1}{2T^2} + \frac{3}{8T^4}\right)\right)$$

$$= \frac{2\gamma^{3/2}}{3}\left(\frac{3}{8T^2} - \frac{3}{8T^4}\right) = \frac{\gamma^{3/2}\left(T^2 - 1\right)}{4T^4} \geq \frac{D^3}{1331T^{7/2}}$$

Where in the first inequality we used the fact that $1 - \frac{1}{2}x \leq \frac{1}{\sqrt{1+x}} \leq 1 - \frac{1}{2}x + \frac{3}{8}x^2$ for all $x \geq 0$ and for the last inequality we assumed that $T \geq 2$. In the case that $T = 1$, the final result still holds. Hence, the suboptimality is at least $\frac{\mu_2 D^3}{16,000T^{7/2}}$.

5.3.2 Case 2: $\frac{1}{T^2} < \frac{D^2}{48\Delta^2T^3} \leq 1$

In this setting, we choose $\gamma = \frac{D\Delta}{\sqrt{12T}}$. Using this and the assumption on the parameters, we get that $\frac{\Delta^2(1+T^2)}{T^2} < \gamma < 2\Delta^2T$, and therefore, we are in the second regime for both f_T and f_{2T} as specified in Proposition 4. Plugging in the bound on $\|\hat{\mathbf{w}}^*\|^2$ in that regime, and using the fact that $\Delta^2 < \gamma$ by the assumption above, we have $\|\hat{\mathbf{w}}_{2T}^*\|^2 \leq \frac{(\gamma+\Delta^2)^2(2T+1)^3}{48\Delta^2T^2} \leq \frac{\gamma^2(2T+1)^3}{12\Delta^2T^2} = \frac{D^2(2T+1)^3}{144T^3}$, which is at most D^2 as required.

Turning to compute the optimization error bound, and letting δ_T, δ_{2T} denote the quantity δ in Proposition 4 for $\hat{f}_T$ and $\hat{f}_{2T}$ respectively, we have

$$f_T^* - f_{2T}^* = (2\delta_{2T} - \delta_T)\left(T\left(\Delta^2 + \gamma\right) - \Delta T^2\left(2\delta_{2T} + \delta_T\right)\right) + \frac{1}{3}T\left(\delta_T^3 - 2\delta_{2T}^3\right). \tag{20}$$

To continue, we use the following auxiliary lemma:

Lemma 11 $(2\delta_{2T} - \delta_T)\left(T\left(\Delta^2 + \gamma\right) - \Delta T^2\left(2\delta_{2T} + \delta_T\right)\right) \geq 0$

Proof First we will prove that $T\left(\Delta^2 + \gamma\right) - \Delta T^2\left(2\delta_{2T} + \delta_T\right) \geq 0$: Since $\delta_T = -\Delta T + \Delta T\sqrt{1 + \frac{\gamma+\Delta^2}{\Delta^2T^2}}$ and using $\sqrt{1+x} \leq 1 + \frac{1}{2}x$ for $x \geq 0$ we have that $\delta_T \leq$

$\frac{(\gamma+\Delta^2)}{2\Delta T}$, so $T\left(\Delta^2+\gamma\right)-\Delta T^2\left(2\delta_{2T}+\delta_T\right) \geq T\left(\Delta^2+\gamma\right)-\Delta T^2\frac{(\gamma+\Delta^2)}{\Delta T}=0$. To complete the proof, it remains to show that $2\delta_{2T}-\delta_T \geq 0$. We have

$$2\delta_{2T}-\delta_T = -4\Delta T+4\Delta T\sqrt{1+\frac{\gamma+\Delta^2}{4\Delta^2 T^2}}+\Delta T-\Delta T\sqrt{1+\frac{\gamma+\Delta^2}{\Delta^2 T^2}}$$

Define $\alpha := \frac{\gamma+\Delta^2}{\Delta^2 T^2} \geq 0$. Hence, we need to prove $-4+4\sqrt{1+\frac{1}{4}\alpha}+1-\sqrt{1+\alpha} \geq 0$. By straightforward manipulations this is equivalent to $6+3\alpha \geq 6\sqrt{1+\alpha}$, which is true since $\sqrt{1+\alpha} \leq 1+\frac{1}{2}\alpha$.   □

With this lemma, we can lower bound the optimization error in (20) by

$$\frac{1}{3}T\left(\delta_T^3-2\delta_{2T}^3\right). \tag{21}$$

To continue, we note that by definition of δ_T, δ_{2T} and the fact that $1+\frac{1}{2}x-\frac{1}{8}x^2 \leq \sqrt{1+x} \leq 1+\frac{1}{2}x$, we have $\frac{(\gamma+\Delta^2)}{2\Delta T}-\frac{(\gamma+\Delta^2)^2}{8\Delta^3 T^3} \leq \delta_T \leq \frac{(\gamma+\Delta^2)}{2\Delta T}$. Therefore,

$$\begin{aligned}
\delta_T-\sqrt[3]{2}\delta_{2T} &\geq \frac{\left(\gamma+\Delta^2\right)}{2\Delta T}-\frac{\left(\gamma+\Delta^2\right)^2}{8\Delta^3 T^3}-\frac{\sqrt[3]{2}\left(\gamma+\Delta^2\right)}{4\Delta T} \\
&\geq \frac{\left(\gamma+\Delta^2\right)}{20\Delta T}+\frac{\left(\gamma+\Delta^2\right)}{8\Delta T}\left(1-\frac{\gamma+\Delta^2}{\Delta^2 T^2}\right) \geq \frac{\left(\gamma+\Delta^2\right)}{20\Delta T}.
\end{aligned}$$

Using this inequality, and the fact $(a-b)^3 \leq a^3-b^3$ for $a \geq b \geq 0$, we can lower bound (21) by $\frac{1}{3}T\left(\delta_T-\sqrt[3]{2}\delta_{2T}\right)^3 \geq \frac{(\gamma+\Delta^2)^3}{60\Delta^3 T^2} \geq \frac{D^3}{2500 T^{7/2}}$. Hence, the suboptimality is at least $\frac{\mu_2 D^3}{30000 T^{7/2}}$.

5.3.3 Case 3: $\frac{D^2}{48\Delta^2 T^3} > 1$

In this setting, we choose $\gamma = \frac{D\Delta}{\sqrt{3T}}$. Using this and the assumption on the parameters, we get that $\gamma > 4\Delta^2 T$, and therefore, we are in the third regime for both f_T and f_{2T} as specified in Proposition 4. Plugging in the bound on $\|\hat{\mathbf{w}}_{2T}^*\|^2$ in that regime, and using the fact that $2\Delta^2 < \gamma$ by the assumption above, we have $\|\hat{\mathbf{w}}_{2T}^*\|^2 \leq \frac{(\gamma+2\Delta^2)^2(2T+1)}{12\Delta^2} \leq \frac{4\gamma^2(2T+1)}{12\Delta^2} = \frac{D^2(2T+1)}{9T} \leq D^2$. Now, by the assumptions that $T\Delta^3 < \frac{\Delta D^2}{48T^2}$ and by using the fact that $1-x \leq \frac{1}{1+x} \leq 1-x+x^2$ for all $x \geq 0$, the optimization error bound is

$$\hat{f}_T^*-\hat{f}_{2T}^* \geq \frac{2T\left(\gamma+2\Delta^2\right)^2}{4\Delta(2T+1)}-\frac{T\left(\gamma+2\Delta^2\right)^2}{4\Delta(T+1)}+\frac{(T+1)\Delta^3}{3}-\frac{(2T+1)\Delta^3}{3}$$

$$= \frac{(\gamma + 2\Delta^2)^2}{4\Delta} \left(\frac{1}{1 + \frac{1}{2T}} - \frac{1}{1 + \frac{1}{T}} \right) - \frac{T\Delta^3}{3}$$

$$\geq \frac{D^2\Delta}{12T} \left(\frac{1}{2T} - \frac{1}{T^2} \right) - \frac{T\Delta^3}{3} \geq \frac{D^2\Delta}{72T^2} - \frac{D^2\Delta}{144T^2} = \frac{D^2\Delta}{144T^2}$$

In the last inequality we assumed that $T \geq 3$. For $T = 1, 2$ it can be easily verified that the inequality $\hat{f}_T^* - \hat{f}_{2T}^* \geq \frac{D^2\Delta}{144T^2}$ holds. Hence, using $\Delta = \frac{3\mu_1}{2\mu_2}$ the suboptimality is at least $\frac{\mu_1 D^2}{576T^2}$.

5.4 Wrapping up

Combining the three cases from the previous subsection, we see that we get the following lower bound

$$f_{2T}(\mathbf{w}_T) - \min_{\mathbf{w}} f_{2T}(\mathbf{w}) \geq \begin{cases} \frac{\mu_2 D^3}{30000 T^{7/2}} & \frac{D^2}{48\Delta^2 T^3} \leq 1 \\ \frac{\mu_1 D^2}{576 T^2} & \frac{D^2}{48\Delta^2 T^3} > 1 \end{cases}.$$

Thus, we get that $f_{2T}(\mathbf{w}_T) - \min_{\mathbf{w}} f_{2T}(\mathbf{w}) \geq \min \left\{ \frac{\mu_2 D^3}{30,000 T^{7/2}}, \frac{\mu_1 D^2}{576 T^2} \right\}$. Equating these bounds to ϵ, and solving for T, the theorem follows.

6 Proof of Theorem 3

The Proof of Theorem 3 will follow the same outline of the proof of 2. We are again going to assume without loss of generality that $\mathbf{w}_1 = 0$, and we will thus require that $\|\mathbf{w}^*\| \leq D$ (see discussion in the Proof of Theorem 2). We define $g : \mathbb{R} \mapsto \mathbb{R}$ as $g(x) = \frac{1}{k+1}|x|^{k+1}$ and

$$f_T(\mathbf{w}) = \frac{\mu_k}{k! 2^{\frac{k+3}{2}}} \left(g(\langle \mathbf{v}_1, \mathbf{w} \rangle) + g(\langle \mathbf{v}_T, \mathbf{w} \rangle) + \sum_{i=1}^{T-1} g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w} \rangle) - \gamma \langle \mathbf{v}_1, \mathbf{w} \rangle \right).$$

Lemma 12 $f_T(\mathbf{w})$ *is k-times differentiable, with μ_k-Lipschitz k-th order derivative tensor.*

Proof Similarly to Lemma 9, we can assume without loss of generality, that the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T$ correspond to the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T$, so we can examine the Lipschitz properties of

$$\hat{f}(\mathbf{w}) = \frac{\mu_k}{k! 2^{\frac{k+3}{2}}} \left(g(w_1) + g(w_T) + \sum_{i=1}^{T-1} g(w_i - w_{i+1}) - \gamma w_1 \right).$$

Letting $\mathbf{r}_0 = \mathbf{e}_1$, $\mathbf{r}_T = \mathbf{e}_T$ and $\mathbf{r}_i = \mathbf{e}_i - \mathbf{e}_{i+1}$ for all $1 \le i \le T - 1$. Differentiating k times, we have that

$$\nabla^{(k)} \hat{f}(\mathbf{w}) = \frac{\mu_k}{k! 2^{\frac{k+3}{2}}} \left(g^{(k)}(w_1) \mathbf{r}_0^{\otimes k} + g^{(k)}(w_T) \mathbf{r}_T^{\otimes k} + \sum_{i=1}^{T-1} g^{(k)}(\langle \mathbf{r}_i, \mathbf{w} \rangle) \mathbf{r}_i^{\otimes k} \right)$$

$$= \frac{\mu_k}{k! 2^{\frac{k+3}{2}}} \left(\sum_{i=0}^{T} g^{(k)}(\langle \mathbf{r}_i, \mathbf{w} \rangle) \mathbf{r}_i^{\otimes k} \right),$$

Where $\mathbf{v}^{\otimes p} = \underbrace{\mathbf{v} \otimes \mathbf{v} \otimes \cdots \mathbf{v}}_{p \text{ times}}$. Since $g^{(k)}(x) = k! x$, we get that $\| \nabla^{(k)} \hat{f}(\mathbf{w}) - \nabla^{(k)} \hat{f}(\tilde{\mathbf{w}}) \|$ equals

$$\frac{\mu_k}{k! 2^{\frac{k+3}{2}}} \left\| \sum_{i=0}^{T} \left(g^{(k)}(\langle \mathbf{r}_i, \mathbf{w} \rangle) - g^{(k)}(\langle \mathbf{r}_i, \tilde{\mathbf{w}} \rangle) \right) \mathbf{r}_i^{\otimes k} \right\| = \frac{\mu_k}{2^{\frac{k+3}{2}}} \left\| \sum_{i=0}^{T} \langle \mathbf{r}_i, \mathbf{w} - \tilde{\mathbf{w}} \rangle \mathbf{r}_i^{\otimes k} \right\|. \tag{22}$$

Note that for a k-th order symmetric tensor T, the operator norm equals $\|T\| = \max_{\|\mathbf{x}\|=1} \left| \sum_{i_1, i_2, \ldots, i_k} T_{i_1, i_2, \ldots, i_k} x_{i_1} x_{i_2} \ldots x_{i_k} \right|$ (see for example [11]). So,

$$\left\| \sum_{i=0}^{T} \langle \mathbf{r}_i, \mathbf{w} - \tilde{\mathbf{w}} \rangle \mathbf{r}_i^{\otimes k} \right\| = \max_{\|\mathbf{x}\|=1} \left| \sum_{i_1, i_2, \ldots, i_k} \sum_{i=0}^{T} \langle \mathbf{r}_i, \mathbf{w} - \tilde{\mathbf{w}} \rangle r_{i, i_1} r_{i, i_2} \ldots r_{i, i_k} x_{i_1} x_{i_2} \ldots x_{i_k} \right|$$

$$= \max_{\|\mathbf{x}\|=1} \left| \sum_{i=0}^{T} \langle \mathbf{r}_i, \mathbf{w} - \tilde{\mathbf{w}} \rangle \sum_{i_1} r_{i, i_1} x_{i_1} \ldots \sum_{i_k} r_{i, i_k} x_{i_k} \right| = \max_{\|\mathbf{x}\|=1} \left| \sum_{i=0}^{T} \langle \mathbf{r}_i, \mathbf{w} - \tilde{\mathbf{w}} \rangle \langle \mathbf{r}_i, \mathbf{x} \rangle^k \right|$$

$$\le 2^{\frac{k}{2}} \max_{\|\mathbf{x}\|=1} \sum_{i=0}^{T} |\langle \mathbf{r}_i, \mathbf{w} - \tilde{\mathbf{w}} \rangle| \left| \left\langle \frac{\mathbf{r}_i}{\sqrt{2}}, \mathbf{x} \right\rangle \right|^k \le 2^{\frac{k+1}{2}} \| \mathbf{w} - \tilde{\mathbf{w}} \| \max_{\|\mathbf{x}\|=1} \sum_{i=0}^{T} \left\langle \frac{\mathbf{r}_i}{\sqrt{2}}, \mathbf{x} \right\rangle^2$$

$$= 2^{\frac{k-1}{2}} \| \mathbf{w} - \tilde{\mathbf{w}} \| \max_{\|\mathbf{x}\|=1} x_1^2 + x_T^2 + \sum_{i=1}^{T-1} (x_i - x_{i+1})^2$$

$$\le 2^{\frac{k-1}{2}} \| \mathbf{w} - \tilde{\mathbf{w}} \| \max_{\|\mathbf{x}\|=1} x_1^2 + x_T^2 + 2 \sum_{i=1}^{T-1} x_i^2 + 2 \sum_{i=2}^{T} x_i^2 \le 2^{\frac{k+3}{2}} \| \mathbf{w} - \tilde{\mathbf{w}} \|$$

where in the first inequality we used $\|\mathbf{r}_i\| \le \sqrt{2}$ for all i. Plugging this into (22) we have $\left\| \nabla^{(k)} \hat{f}(\mathbf{w}) - \nabla^{(k)} \hat{f}(\tilde{\mathbf{w}}) \right\| \le \mu_k \| \mathbf{w} - \tilde{\mathbf{w}} \|$ as required. $\qquad \square$

6.1 Minimizer of f_T

In order to derive the complexity bound, we will first analyze $\hat{f}_T$, which is a simplified version of f_T, as defined in Sect. 5.1. It is easily verified that $\min_{\mathbf{w}} f_T(\mathbf{w}) = \frac{\mu_k}{k!2^{\frac{k+3}{2}}} \cdot \min_{\mathbf{w}} \mathbf{w} \hat{f}_T(\hat{\mathbf{w}})$, and moreover, if $\hat{\mathbf{w}} \in \mathbb{R}^T$ is a minimizer of $\hat{f}_T$, then $\mathbf{w}^* = \sum_{j=1}^{T} \hat{w}_j^* \cdot \mathbf{v}_j \in \mathbb{R}^d$ is a minimizer of f_T, and with the same Euclidean norm as $\hat{\mathbf{w}}^*$.

Using an identical proof to Lemma 10 we can have that $\hat{f}_T$ has a unique minimizer $\hat{\mathbf{w}}^* \in \mathbb{R}^T$, which satisfies $\hat{w}_t^* = \delta \cdot (T+1) \cdot \left(1 - \frac{t}{T+1}\right)$ for some $\delta > 0$ and all $t = 1, 2, \ldots, T$, and $g'(w_1^*) + g'(w_T^*) = (\delta T)^k + \delta^k = \delta^k(T^k + 1) =.$ Hence, $\delta = \left(\frac{\gamma}{T^k+1}\right)^{\frac{1}{k}}$. Plugging this and performing some algebraic manipulations, we get

$$
\begin{aligned}
\hat{f}_T(\hat{\mathbf{w}}^*) &= \frac{1}{k+1}(\delta T)^{k+1} + \frac{T}{k+1}\delta^{k+1} - \gamma\delta T \\
&= \frac{T}{k+1}\left(T^k + 1\right)\left(\frac{\gamma}{T^k+1}\right)^{\frac{k+1}{k}} - \gamma T \left(\frac{\gamma}{T^{p-1}+1}\right)^{\frac{1}{p-1}} = -\frac{kT\gamma^{\frac{k+1}{k}}}{(k+1)\left(T^k+1\right)^{\frac{1}{k}}}
\end{aligned}
\tag{23}
$$

$$
\|\hat{\mathbf{w}}^*\|_2^2 = \sum_{t=1}^{T} \hat{w}_t^{*2} = \delta^2(1+T)^2 \sum_{t=1}^{T}\left(1 - \frac{t}{1+T}\right)^2 \leq \frac{1}{3}\left(\frac{\gamma}{T^k+1}\right)^{\frac{2}{k}}(1+T)^3,
\tag{24}
$$

where we used $\sum_{t=1}^{T}(1 - \frac{t}{1+T})^2 \leq \frac{1}{3}(1+T)$ as in Proposition 4.

6.2 Oracle complexity lower bound

The derivation of the lower complexity bound will be exactly the same as in Sect. 5.2. In that subsection, we showed that we can lower bound the optimization error $f_{2T}(\mathbf{w}_T) - \min_{\mathbf{w}} f_{2T}(\mathbf{w})$ by $\min_{\mathbf{w}} f_T(\mathbf{w}) - \min_{\mathbf{w}} f_{2T}(\mathbf{w})$. Using the fact that $f_T(\mathbf{w}^*) = \frac{\mu_k}{k!2^{\frac{k+3}{2}}} \cdot \hat{f}_T(\hat{\mathbf{w}}^*)$, this equals $\frac{\mu_k}{k!2^{\frac{k+3}{2}}}(\min_{\mathbf{w}} \hat{f}_T(\mathbf{w}) - \min_{\mathbf{w}} \hat{f}_{2T}(\mathbf{w}))$. Letting f_T^* and $\hat{f}_T^*$ to be the minimal values of f_T and $\hat{f}_T$ respectively, and by using equation (23),

$$
\begin{aligned}
\hat{f}_T^* - \hat{f}_{2T}^* &= \frac{k\gamma^{\frac{k+1}{k}}}{(k+1)\left(1 + \frac{1}{(2T)^k}\right)^{\frac{1}{k}}} - \frac{k\gamma^{\frac{k+1}{k}}}{(k+1)\left(1 + \frac{1}{T^k}\right)^{\frac{1}{k}}} \\
&\geq \frac{k\gamma^{\frac{k+1}{k}}}{k+1}\left(1 - \frac{1}{k(2T)^k} - \left(1 - \frac{1}{kT^k} + \frac{k+1}{2k^2T^{2k}}\right)\right) \\
&= \frac{\gamma^{\frac{k+1}{k}}}{(k+1)kT^k}\left(1 - \frac{1}{2^k} - \frac{k+1}{2kT^k}\right) \geq \frac{\gamma^{\frac{k+1}{k}}}{6(k+1)kT^k}
\end{aligned}
$$

The last inequality holds for $k = 1, T \geq 3, k = 2, T \geq 2$ or $k \geq 3, T \geq 1$. It can be verified that for the other cases, the inequality $\hat{f}_T^* - \hat{f}_{2T}^* \geq \frac{\gamma^{\frac{k+1}{k}}}{6(k+1)kT^k}$ holds.

Since we want $f_T^* - f_{2T}^*$ to be as large as possible, we will set γ to be as large as possible, under the constraint that $\|\mathbf{w}_{2T}^*\| \leq D$. By (24) we can choose $\gamma = \frac{3^{\frac{k}{2}} D^k (1+(2T)^k)}{(1+2T)^{\frac{3k}{2}}}$, so

$$\hat{f}_T^* - \hat{f}_{2T}^* \geq \frac{3^{\frac{k+1}{2}} D^{k+1} \left(1 + (2T)^k\right)^{\frac{k+1}{k}}}{6(k+1)kT^k(1+2T)^{\frac{3(k+1)}{2}}} \geq \frac{2^{k+1} D^{k+1}}{6 \cdot 3^{k+1}(k+1)kT^{\frac{3k+1}{2}}}.$$

Thus, according to the discussion in Sect. 6.2, the final bound is $f_T^* - f_{2T}^* \geq \frac{\mu_k \sqrt{2}^{k+1} D^{k+1}}{12 \cdot 3^{k+1}(k+1)!kT^{\frac{3k+1}{2}}}$, and the number of iterations required for having $\min_{\mathbf{w}} f_T(\mathbf{w}) - \min_{\mathbf{w}} f_{2T}(\mathbf{w}) < \epsilon$, T_ϵ must satisfy

$$T_\epsilon \geq \left(\frac{\mu_k \sqrt{2}^{k+1} D^{k+1}}{12 \cdot 3^{k+1}(k+1)!k\epsilon}\right)^{\frac{2}{3k+1}} \geq c \left(\frac{\mu_k D^{k+1}}{(k+1)!k\epsilon}\right)^{\frac{2}{3k+1}}$$

for an appropriate numerical constant c.

Appendix A: An improved second-order oracle complexity bound for strongly convex functions

In this section, we show how the A-NPE algorithm of [10], which is a second-order method analyzed for smooth convex functions, can be used to yield near-optimal performance if the function is also strongly convex. Rather than directly adapting their analysis, which is non-trivial, we use a simple restarting scheme, which allows one to convert an algorithm for the convex setting, to an algorithm in the strongly convex setting[7].

Our algorithm is described as follows: In the first phase, we apply a generic restarting scheme (based on [3, Subsction 4.2]), where we repeatedly run A-NPE for a bounded number of steps, followed by restarting the algorithm, running it from the last iterate obtained. By strong convexity, we show that each such epoch reduces the suboptimality by a constant factor. Once we reach a point sufficiently close to the global optimum, we switch to the second phase, where we use the cubic-regularized Newton method to get a quadratic convergence rate.

---

[7] We note that the *reverse* direction, of adapting strongly convex optimization algorithms to the convex case, is more common in the literature, and can be achieved using regularization or more sophisticated approaches [2].

To formalize this, let us first analyze the convergence rate of the first phase. We assume that we use the algorithm described in [10, Subsection 7.4][8]. By [10, Theorem 6.4 and Theorem 3.10], we have that the t'th iterate satisfies

$$\|\mathbf{w}_t - \mathbf{w}_*\| \le D \quad \text{and} \quad f(\mathbf{w}_t) - f(\mathbf{w}^*) \le \frac{c\mu_2\|\mathbf{w}_1 - \mathbf{w}^*\|^3}{t^{7/2}},$$

where μ_2 is the Lipschitz constant of $\nabla^2 f$, $\mathbf{w}_1$ is the initialization point, $\mathbf{w}^*$ is the unique minimizer (due to strong convexity) of f, D bounds $\|\mathbf{w}_1 - \mathbf{w}^*\|$ from above, and $c > 0$ is some universal constant. Since f is also assumed to be λ-strongly convex, we have $\frac{\lambda}{2}\|\mathbf{w}_1 - \mathbf{w}^*\|^2 \le f(\mathbf{w}_1) - f(\mathbf{w}^*)$, hence $f(\mathbf{w}_t) - f(\mathbf{w}^*)$ is at most

$$\frac{c\mu_2\|\mathbf{w}_1 - \mathbf{w}^*\|^3}{t^{7/2}} \le \frac{2c\mu_2\|\mathbf{w}_1 - \mathbf{w}^*\|(f(\mathbf{w}_1) - f(\mathbf{w}^*))}{\lambda t^{7/2}} \le \frac{2c\mu_2 D(f(\mathbf{w}_1) - f(\mathbf{w}^*))}{\lambda t^{7/2}}.$$

Thus, running the algorithm for $\tau = \left(\frac{4c\mu_2 D}{\lambda}\right)^{2/7}$ iterations, we see that $f(\mathbf{w}_t) - f(\mathbf{w}^*) \le (f(\mathbf{w}_1) - f(\mathbf{w}^*))/2$. Now, since the distance from $\mathbf{w}_t$ to $\mathbf{w}^*$ is also smaller than D, we may initialize the algorithm at the last iterate returned by the previous run and run it for τ iterations to reduce $f(\mathbf{w}_t) - f(\mathbf{w}^*)$ in, yet again, a factor of 2. Applying the algorithm for T iterations (and restarting the algorithmic parameters after every τ iterations) yields $f(\mathbf{w}_T) - f(\mathbf{w}^*) \le \frac{f(\mathbf{w}_1) - f(\mathbf{w}^*)}{2^{T/\tau}}$. Equivalently, to obtain an ϵ-optimal solution, we need at most $\left(\frac{4c\mu_2 D}{\lambda}\right)^{2/7} \log_2\left(\frac{f(\mathbf{w}_1) - f(\mathbf{w}^*)}{\epsilon}\right)$ oracle calls (note that this restarting scheme can be applied also on uniform convex functions of any order, as defined in, e.g., [19]).

Next, after performing a number of iterations sufficiently large to obtain high accuracy solutions, we proceed to the second phase of the algorithm where cubic-regularized Newton steps are applied (see [16]). According to that analysis, after reducing the optimization error to below $\lambda^3/4\mu_2^2$, the number of cubic-regularized Newton steps required to achieve an ϵ-suboptimal solution is $\mathcal{O}\left(\log\log_2\left(\frac{\lambda^3}{\mu_2^2\epsilon}\right)\right)$. Thus, using the μ_1-Lipschitzness of the gradient to bound $f(\mathbf{w}_1) - f(\mathbf{w}^*)$ from above by $\mu_1 D^2/2$, we get that the overall number of iterations is at most $\mathcal{O}\left(\left(\frac{\mu_2 D}{\lambda}\right)^{2/7} \log_2\left(\frac{\mu_1\mu_2^2 D^2}{\lambda^3}\right) + \log\log_2\left(\frac{\lambda^3}{\mu_2^2\epsilon}\right)\right)$.

References

1. Agarwal, N., Hazan, E.: Lower bounds for higher-order optimization. Working draft (2017)
2. Allen-Zhu, Z., Hazan, E.: Optimal black-box reductions between optimization objectives. In: Advances in Neural Information Processing Systems, pp. 1614–1622 (2016)
3. Arjevani, Y., Shamir, O.: On the iteration complexity of oblivious first-order optimization algorithms. In: International Conference on Machine Learning, pp. 908–916 (2016)

---

[8] Specifically, since in our framework we do not limit computational resources, we assume that the minimization problem in Eq. (6.1) of [10] can be solved exactly.

4. Arjevani, Y., Shamir, O.: Oracle complexity of second-order methods for finite-sum problems. arXiv preprint arXiv:1611.04982 (2016)
5. Baes, M.: Estimate Sequence Methods: Extensions and Approximations. Institute for Operations Research, ETH, Zürich (2009)
6. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
7. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems. SIAM J. Optim. **20**(6), 2833–2852 (2010)
8. Cartis, C., Gould, N.I., Toint, P.L.: Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. Optim Methods Softw. **27**(2), 197–219 (2012)
9. Kantorovich, L.V.: Functional analysis and applied mathematics. Uspekhi Matematicheskikh Nauk **3**(6), 89–185 (1948)
10. Monteiro, R.D., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM J. Optim. **23**(2), 1092–1125 (2013)
11. Mu, C., Hsu, D., Goldfarb, D.: Successive rank-one approximations for nearly orthogonally decomposable symmetric tensors. SIAM J. Matrix Anal. Appl. **36**(4), 1638–1659 (2015)
12. Nemirovski, A.: Efficient methods in convex programming—lecture notes (2005)
13. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
14. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **27**(2), 372–376 (1983)
15. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Berlin (2004)
16. Nesterov, Y.: Accelerating the cubic regularization of newton method on convex problems. Math. Program. **112**(1), 159–181 (2008)
17. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia (1994)
18. Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
19. Vladimirov, A., Nesterov, Y.E., Chekanov, Y.N.: On uniformly convex functionals. Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet **3**, 12–23 (1978)
20. Woodworth, B., Srebro, N.: Lower bound for randomized first order convex optimization. arXiv preprint arXiv:1709.03594 (2017)