Oracle Complexity of Second-Order Methods for Smooth Convex Optimization

Ohad Shamir Ron Shiff
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
{ohad.shamir,ron.shiff}@weizmann.ac.il

May 23, 2017

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 study the oracle complexity of such methods, or equivalently, the number of iterations required to optimize a function to a given accuracy. Focusing on smooth and convex functions, we derive (to the best of our knowledge) the first algorithm-independent lower bounds for such methods. These bounds shed light on the limits of attainable performance with second-order methods, and in which cases they can or cannot require less iterations than gradient-based methods, whose oracle complexity is much better understood.

1 Introduction

Consider an unconstrained optimization problem of the form

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

where $\mathbf{w}$ takes values in Euclidean space, and 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 [Nemirovsky and Yudin, 1983], 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 Eq. (1), and allows one to study their efficiency, by bounding the number of oracle calls required to reach a given optimization accuracy. 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 also returns gradients as well as Hessians.

The theory of *first-order* oracle complexity is quite well developed [Nemirovsky and Yudin, 1983, Nesterov, 2004, Nemirovski, 2005]. For example, if the dimension is unrestricted, f in Eq. (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[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 [Nesterov, 1983].

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 methods in optimization, which includes the Newton method and its many variants. Clearly, since they rely on Hessians as well as gradients, the oracle complexity of second-order methods 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 very natural to ask how much does this added per-iteration complexity pays off in terms of oracle complexity.

To answer this question, one needs good oracle complexity lower bounds for second-order methods, which establishes 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 Nesterov, 2008). The goal of this paper is to address this gap.

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

$$\Omega\left(\left(\min\left\{\sqrt{\frac{\mu_1}{\lambda}}\,,\ \left(\frac{\mu_2}{\lambda}D\right)^{2/7}\right\} + \log\log_{18}\left(\frac{\lambda^3/\mu_2^2}{\epsilon}\right)\right)\right). \tag{4}$$

for λ-strongly convex functions, and

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

for convex functions. As we discuss in more detail later on, these bounds have several implications in light of existing algorithms and upper bounds in the literature:

- In the context of strongly convex functions, Eq. (4) 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 fact that Hessian information can be used to alter the geometry of the problem (for example, the Newton method is well-known to be affine invariant).

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

- To improve on the oracle complexity of first-order methods for strongly-convex problems (Eq. (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 moderately 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, despite the availability of more information.

- Neglecting a logarithmic factor, the only difference between our lower bounds and the existing upper bounds in the literature (as discussed later on), is that the $\left(\frac{\mu_2}{\lambda} D\right)^{2/7}$ term is replaced by $\left(\frac{\mu_2}{\lambda} D\right)^{1/3}$ in the strongly convex case, and $\left(\frac{\mu_2 D^3}{\epsilon}\right)^{2/7}$ by $\left(\frac{\mu_2 D^3}{\epsilon}\right)^{1/3}$ in the convex case. We do not know whether this gap in the exponent (between $2/7 = 0.28..$ and $1/3 = 0.33..$) is due to a looseness in our bounds, or whether the existing algorithms can be improved.

1.1 Related Work

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. Boyd and Vandenberghe [2004]). It is well-known that this method exhibits *local* quadratic convergence, in the sense that if it is initialized close enough to the optimum, and f is strictly convex, then $\mathcal{O}(\log\log(1/\epsilon))$ iterations suffice to reach an ϵ-optimal solution. 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, such as introducing a step size parameter or line search, employing trust region methods, or adding various types of regularization (see for example Conn et al. [2000] and references therein). Despite the huge literature on the subject, the worst-case global convergence behavior of these methods is not well understood Nesterov and Polyak [2006]. For the Newton method with a line search, the number of iterations to get a point $\mathbf{w}$ such that $f(\mathbf{w}) - f(\mathbf{w}^*) \leq \epsilon$, where $\mathbf{w}^* \in \arg\min_{\mathbf{w}} f(\mathbf{w})$, can be upper bounded from above by

$$\mathcal{O}\left(\frac{\mu_1^2 \mu_2^2}{\lambda^5}(f(\mathbf{w}_1) - f(\mathbf{w}^*)) + \log_2\log_2\left(\frac{2\lambda^3/\mu_2^2}{\epsilon}\right)\right),$$

where μ_1, μ_2 are the Lipschitz parameters of the gradients and Hessians respectively, and assuming the function is λ-strongly convex (Kantorovich [1948], see also Boyd and Vandenberghe [2004]). Note that the first term captures the initial phase required to get sufficiently close to $\mathbf{w}^*$, whereas the second term captures the quadratically convergent phase. Although the final convergence is rapid, the first phase is the dominant one 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 (Nesterov and Nemirovskii [1994]). Unfortunately, not all practically relevant objective functions are self-concordant. For example, loss functions

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

common in machine learning applications, such as the logistic loss $x \mapsto \log(1+\exp(-x))$, are not self-concordant[3], and our own results utilize the simple but not self-concordant function $x \mapsto |x|^3$.

Returning to our setting of generic convex and smooth functions, and focusing on strongly convex functions for now, the best dimension-free upper bounds we are aware of, using second order methods, were obtained for cubic-regularized variants of the Newton method, where at each iteration one essentially minimizes a quadratic approximation of the function at the current point, regularized by a cubic function. This approach was proposed in Nesterov and Polyak [2006], with an oracle complexity upper bound of

$$\mathcal{O}\left(\sqrt{\frac{\mu_2}{\lambda}\bar{D}} + \log\log_4\left(\frac{2\lambda^3/9\mu_2^2}{\epsilon}\right)\right),$$

where $\bar{D} = \max_{\mathbf{w}}\{\|\mathbf{w}-\mathbf{w}^*\| : f(\mathbf{w}) \le f(\mathbf{w}_0)\}$. This was further improved in Nesterov [2008], where an accelerated cubic-regularized variant of the Newton method was proposed, with an analysis (in section 6 of that paper) which implies a bound of

$$\mathcal{O}\left(\left(\frac{\mu_2}{\lambda}D\right)^{1/3} + \log\log_2\left(\frac{2\lambda^3/\mu_2^2}{\epsilon}\right)\right), \tag{6}$$

where $D = \|\mathbf{w}_1 - \mathbf{w}^*\|$ is the distance from the initialization point $\mathbf{w}_1$ to the optimum $\mathbf{w}^*$. See also Cartis et al. [2012] for another treatment of such cubic-regularized methods.

An alternative to the above is to use a two-phase scheme, starting with accelerated gradient descent (which is an optimal *first-order* method for strongly convex functions with Lipschitz gradients) and then switching to a Newton method when close enough to the optimal solution. The required number of iterations is then

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

where $D = \|\mathbf{w}_1 - \mathbf{w}^*\|$ (see Nesterov [2004, 2008]). Note that the bounds in Eq. (6) and Eq. (7) are not directly comparable: The first bound has a cube-root dependence on μ_2/λ, no dependence on μ_1, and a polynomial dependence on $\|\mathbf{w}_1 - \mathbf{w}^*\|$, whereas the second has a square-root dependence on μ_1/λ, logarithmic dependence on μ_2, and a logarithmic dependence on $\|\mathbf{w}_1 - \mathbf{w}^*\|$. In a rather wide parameter regime (e.g. when D is reasonably large, as often occurs in practice), the bound of the simple two-phase scheme can be better than that of the cubic-regularized Newton method. In light of this, Nesterov [2008] raised the question of whether second-order schemes are indeed useful at the initial stage of the optimization process, for these types of problems.

Analogous results can be obtained for convex (not necessarily strongly convex) smooth functions. Using an appropriate analysis of the accelerated cubic-regularized Newton method [Nesterov, 2008], one can attain a bound of

$$\mathcal{O}\left(\left(\frac{\mu_2 D^3}{\epsilon}\right)^{1/3}\right). \tag{8}$$

Moreover, using an optimal first-order method (such as accelerated gradient descent), one can attain a bound of

$$\mathcal{O}\left(\sqrt{\frac{\mu_1 D^2}{\epsilon}}\right). \tag{9}$$

[3]These can often be made self-concordant by re-scaling, smoothing and adding regularization (e.g. Bach [2010]), but even when possible, these modifications strongly affect the $f(\mathbf{w}_1) - f(\mathbf{w}^*)$ term in the bound, and prevents it from being independent of the strong convexity and Lipschitz parameters.

Clearly, by taking the best of the two approaches (depending on the problem parameters), one can attain the minimum of those two bounds.

In terms of lower bounds for second-order methods, it appears that not much is currently known. If μ_2 is not bounded (i.e. the Hessians are not Lipschitz), it is easy to show that Hessian information is not useful, and the lower bound of Eq. (2) for first-order methods also apply to second-order methods, and indeed, to any method based on local information (see Nemirovsky and Yudin [1983, section 7.2.6] and Arjevani and Shamir [2016]). 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 [Nemirovsky and Yudin, 1983, 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 Cartis et al. [2010].

2 Main Results

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 $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_T$, 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*

$$\frac{\mu_1}{\lambda} \geq c_1 \ , \ \frac{\mu_2}{\lambda} D \geq c_2 \ , \ \epsilon < \frac{c_3 \lambda^3}{\mu_2^2}$$

(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 $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq D$.*
- *The number of calls T to a second-order oracle, required to ensure $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$, is at least*

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

for some universal constant $c > 0$.

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 $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq D$.*
- *The number of calls T to a second-order oracle, required to ensure $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$, is at least*

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

for some universal constant $c > 0$.

We emphasize that the theorems provide a *dimension-free* oracle complexity lower bound, in the sense that the dimension d can be chosen sufficiently large. This is opposed to the dimension-dependent setting, where d is fixed and the attainable oracle complexity can have a different behavior (Nemirovsky and Yudin [1983], Nemirovski [2005]).

Let us compare these theorems to the upper bounds discussed in the related work section. Starting with the strongly convex case, and taking the minimum of Eq. (6) (achieved with the accelerated cubic-regularized Newton method) and Eq. (7) (achieved with accelerated gradient descent followed by the Newton method), we have the upper bound

$$\mathcal{O}\left(\min\left\{\sqrt{\frac{\mu_1}{\lambda}}\cdot\log\left(\frac{\mu_1\mu_2^2 D}{\lambda^3}\right), \left(\frac{\mu_2}{\lambda}D\right)^{1/3}\right\} + \log\log_2\left(\frac{2\lambda^3/\mu_2^2}{\epsilon}\right)\right).$$

Comparing this to Thm. 1, we see that there are two potential sources for looseness (other than constants). First, the exponent of the $\mu_2 D/\lambda$ term is $2/7$ ($= 0.285..$) rather than $1/3$ ($= 0.333..$). At this point, we do not know if this is due to a looseness in our lower bound, or if it is actually achievable by a better algorithm than those currently known. A second potential looseness is the missing logarithmic factor $\log(\mu_1\mu_2^2 D/\lambda^3)$, which is probably an artifact of the analysis (indeed, when $\mu_2 \to \infty$, one should recover the $\Omega(\sqrt{\mu_1/\lambda}\cdot \log(\mu_1 D^2/\epsilon))$ lower-bound of first order methods, see Nemirovsky and Yudin [1983, section 7.2.6] and Arjevani and Shamir [2016]). We believe this logarithmic factor can be recovered by a more careful analysis of our construction. However, this involves several technical intricacies which we leave to future work.

Based on these lower and upper bounds, one can make the following observations:

- The lower bound captures the two phases common in second-order methods such as the Newton method: An initial slow convergence from the initialization point to the local neighborhood of the optimum (captured by the $\min\left\{\sqrt{\frac{\mu_1}{\lambda}}\ ,\ \left(\frac{\mu_2}{\lambda}D\right)^{2/7}\right\}$ term), followed by a fast local quadratic convergence to the optimum (captured by the second term, which is doubly-logarithmic in the accuracy ϵ).

- Unless ϵ is exceedingly small, the oracle complexity is dominated by the geometry-dependent terms μ_1/λ and $\mu_2 D/\lambda$. This is despite the fact that second-order methods can use Hessian information to alter the geometry of the problem (for example, the Newton method is well-known to be affine invariant).

- If $\mu_2 D/\lambda$ is moderately large (specifically, if D is order of $\frac{\mu_1^{7/4}}{\mu_2\lambda^{3/4}}$ or larger), then the lower bound becomes at least $\sqrt{\mu_1/\lambda}$, which is no betters what can be obtained with first-order methods up to logarithmic factors (see Eq. (3)). Since D often scales inversely with the strong convexity of the problem (e.g. since the strong convexity is due to a regularization term), this is a rather broad and reasonable regime.

- On the other hand, if $\mu_2 D/\lambda$ is smaller than $\sqrt{\mu_1/\lambda}$, then the oracle complexity can be significantly better than that of first-order methods, but this still comes at the inevitable price of a polynomial dependence on the distance D from the optimum. In contrast, first order methods have only a logarithmic dependence on D (see Eq. (3)).

Similar types of conclusions can be drawn in the convex case. Taking the minimum of Eq. (8) and Eq. (9), we get an oracle complexity upper bound of

$$\mathcal{O}\left(\min\left\{\sqrt{\frac{\mu_1 D^2}{\epsilon}}\ ,\ \left(\frac{\mu_2 D^3}{\epsilon}\right)^{1/3}\right\}\right).$$

Comparing this to Thm. 2, we see there is a potential gap in terms of the exponent ($1/3$ vs. $2/7$, which may point at the possibility of a better algorithm or a stronger lower bound). Moreover, if $\mu_2 D^3/\epsilon$ is large enough (specifically, if $\mu_2 \geq \mu_1^{7/4}\sqrt{D}/\epsilon^{3/4}$), then according to the lower bound, the complexity of second-order methods is not significantly better than what can be obtained with first-order methods.

3 Proof Ideas

The proofs of our theorems are based on a careful modification of the standard lower bound construction for first order methods (see Nemirovsky and Yudin [1983], Nesterov [2004], Nemirovski [2005]). Such bounds are based on 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 uses additional parameters in the definition of f_T). 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 methods is quite simple: Instead of $x \mapsto x^2$, we rely on cubic functions of the form $x \mapsto |x|^3$, and functions with the basic structure

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

The intuition is that $x \mapsto |x|^3$ has both values, first derivative *and second derivative* 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 of the function. Another motivation for choosing a cubic function for our construction is that it is 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.

Unfortunately, there are two technical challenges that need to be overcome: The first is that f_T as defined above can be shown to have globally Lipschitz Hessians, but not globally Lipschitz gradients as required by our theorems. To tackle this, we replace $x \mapsto |x|^3$ by a more complicated function, 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 the simplest possible case, where we wish to obtain an $\Omega((D^3/\epsilon)^{2/7})$ lower bound for the class of convex functions with $\mathcal{O}(1)$-Lipschitz

Hessians (and no assumptions on the Lipschitz parameter of the gradients), assuming the algorithm makes its first query at the origin. In that case, 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

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

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 $\min_{\mathbf{w}} f_T(\mathbf{w})$ equals

$$\begin{aligned} f_T(\mathbf{w}) &= T^3 \left(\frac{\gamma}{T^2+1}\right)^{3/2} + (T-1)\left(\frac{\gamma}{T^2+1}\right)^{3/2} + \left(\frac{\gamma}{T^2+1}\right)^{3/2} - 3\gamma T\sqrt{\frac{\gamma}{T^2+1}} \\ &= T(T^2+1)\left(\frac{\gamma}{T^2+1}\right)^{3/2} - 3\frac{\gamma^{3/2}T}{\sqrt{T^2+1}} \ = \ -\frac{2\gamma^{3/2}T}{\sqrt{T^2+1}} \ . \end{aligned}$$

Now, using arguments very similar to those in first-order oracle complexity lower bounds, it is possible to construct a function for which the optimization error of the algorithm is lower bounded by

$$\begin{aligned} \min_{\mathbf{w}} f_T(\mathbf{w}) - \min_{\mathbf{w}} f_{2T}(\mathbf{w}) &= 2\gamma^{3/2}\left(\frac{2T}{\sqrt{4T^2+1}} - \frac{T}{\sqrt{T^2+1}}\right) \\ &= 2\gamma^{3/2}\left(\frac{1}{\sqrt{1+\frac{1}{4T^2}}} - \frac{1}{\sqrt{1+\frac{1}{T^2}}}\right). \end{aligned}$$

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.

4 Proof of Thm. 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. Define $g : \mathbb{R} \mapsto \mathbb{R}$ as

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

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

$$f(\mathbf{w}) = \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{v}_i, \mathbf{w}\rangle - \langle \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \gamma\langle \mathbf{v}_1, \mathbf{w}\rangle\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2 ,$$

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 (Subsection 4.1). Then, we prove the oracle complexity lower bound in Subsection 4.2 (depending on Δ, γ), and finally, in Subsection 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 1. *Suppose that* $\gamma \geq 10^4\left(\frac{\lambda}{\mu_2}\right)^2$ *and* $\Delta \geq \sqrt{\gamma}$. *Then* f *has a unique minimizer* $\mathbf{w}^*$ *which satisfies the following:*

1. *For any* $t \in \{1, 2, \ldots, \tilde{T}\}$, *it holds that* $\langle \mathbf{v}_t, \mathbf{w}^*\rangle \geq \max\left\{0\,,\ \frac{\gamma^{3/4}}{7\sqrt{6\lambda/\mu_2}} + \sqrt{\gamma}\left(\frac{1}{2} - t\right)\right\}$.
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* $\langle \mathbf{v}_{t_0+k}, \mathbf{w}^*\rangle \geq \frac{54\lambda}{\mu_2}\cdot (18)^{-2^k}$.
3. $\|\mathbf{w}^*\|^2 \leq \frac{2\gamma^{7/4}}{(6\lambda/\mu_2)^{3/2}}$.

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

$$\tilde{f}(\mathbf{w}) = \frac{1}{3}\sum_{i=1}^{\tilde{T}-1} |w_i - w_{i+1}|^3 + \frac{\tilde{\lambda}}{2}\|\mathbf{w}\|^2 - \gamma \cdot w_1, \tag{10}$$

where

$$\tilde{\lambda} := \frac{6\lambda}{\mu_2} ,$$

and prove that its minimizer $\tilde{\mathbf{w}}^*$ satisfies the following:

1. For any $t \in \{1, 2, \ldots, \tilde{T}\}$, it holds that $\tilde{w}_t^* \geq \max\left\{0\,,\ \frac{\gamma^{3/4}}{7\sqrt{\lambda}} + \sqrt{\gamma}\left(\frac{1}{2} - t\right)\right\}$ (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 1 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*

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

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 $w_j \geq 0$ 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.

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

$$(\tilde{w}_1^* - \tilde{w}_2^*)^2 = \gamma - \tilde{\lambda} w_1^* \quad , \quad (\tilde{w}_{\tilde{T}-1}^* - \tilde{w}_{\tilde{T}}^*)^2 = \tilde{\lambda} w_{\tilde{T}}^* \tag{11}$$

and

$$(\tilde{w}_t^* - \tilde{w}_{t+1}^*)^2 = (\tilde{w}_{t-1}^* - \tilde{w}_t^*)^2 - \tilde{\lambda} \tilde{w}_t^* \quad \forall t \in \{2, 3, \ldots, \tilde{T} - 1\} \,. \tag{12}$$

[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 $\tilde{w}_i^* > \Delta$.

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

$$\tilde{w}_{t+1}^* = w_t^* - \sqrt{\gamma - \tilde{\lambda} \sum_{j=1}^{t} \tilde{w}_j^*} \quad \forall t \in \{1, 2, \ldots, \tilde{T} - 1\} \tag{13}$$

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 Eq. (12) 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 Eq. (11) 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

$$\tilde{w}_t^* \geq \tilde{w}_1^* - (t-1)\sqrt{\gamma} \quad \forall t \in \{1, 2, \ldots, \tilde{T}\} \,. \tag{14}$$

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,

$$\begin{aligned}
\frac{\gamma}{\tilde{\lambda}} \;&=\; \sum_{t=1}^{\tilde{T}} \tilde{w}_t^* \;\geq\; \sum_{t=1}^{\tilde{T}} \max\{0, \tilde{w}_1^* - (t-1)\sqrt{\gamma}\} \;=\; \sum_{t=1}^{\lfloor \tilde{w}_1^*/\sqrt{\gamma}+1 \rfloor} (\tilde{w}_1^* - (t-1)\sqrt{\gamma}) \\
&=\; \left\lfloor \frac{\tilde{w}_1^*}{\sqrt{\gamma}} + 1 \right\rfloor \cdot \tilde{w}_1^* - \sqrt{\gamma} \frac{\left(\left\lfloor \frac{\tilde{w}_1^*}{\sqrt{\gamma}} + 1 \right\rfloor - 1 \right) \left\lfloor \frac{\tilde{w}_1^*}{\sqrt{\gamma}} + 1 \right\rfloor}{2} \;\geq\; \frac{(\tilde{w}_1^*)^2}{\sqrt{\gamma}} - \sqrt{\gamma} \frac{\frac{\tilde{w}_1^*}{\sqrt{\gamma}} \left(\frac{w_1^*}{\sqrt{\gamma}} + 1 \right)}{2},
\end{aligned}$$

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

$$\tilde{w}_1^* \leq \frac{\sqrt{\gamma} + \sqrt{\gamma + 8\gamma^{3/2}/\tilde{\lambda}}}{2} \leq \frac{\sqrt{\gamma} + \sqrt{\gamma} + \sqrt{8\gamma^{3/2}/\tilde{\lambda}}}{2} = \sqrt{\gamma} + \sqrt{\frac{2\gamma^{3/2}}{\tilde{\lambda}}}. \tag{15}$$

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

$$\tilde{w}^*_{t+1} \le \tilde{w}^*_1 - \frac{t\sqrt{\gamma}}{2}\,, \quad \forall t \in \{1,2,\dots,\tilde{T}-1\} : \sum_{j=1}^{t} \tilde{w}^*_j \le \frac{3\gamma}{4\tilde{\lambda}}. \tag{16}$$

Let $t_0 \in \{1,2,\dots,\tilde{T}\}$ 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 $\frac{3\gamma}{4\tilde{\lambda}} < \sum_{j=1}^{t_0} \tilde{w}^*_j \le t_0 \tilde{w}^*_1 \le t_0 \left(\sqrt{\gamma} + \sqrt{2\gamma^{3/2}/\tilde{\lambda}}\right)$ by Eq. (15), it follows that

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

According to Eq. (16) and the fact that $\tilde{w}^*_{t_0} \ge 0$, it follows that

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

and hence

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

Using this and Eq. (14), it follows that for all $t \le \tilde{T}$,

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

Since we assumed $\gamma \ge 10^4(\lambda/\mu_2)^2 > (6\lambda/\mu_2)^2 \ge \tilde{\lambda}^2$, 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 $t_0 \le \tilde{T}/2$ such that*

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

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

$$\tilde{w}^*_t \;=\; \tilde{w}^*_{t+1} + \sqrt{\gamma - \tilde{\lambda}\sum_{j=1}^{t} \tilde{w}^*_j} \;=\; \tilde{w}^*_{t+1} + \sqrt{\gamma - \tilde{\lambda}\left(\frac{\gamma}{\tilde{\lambda}} - \sum_{j=t+1}^{\tilde{T}} \tilde{w}^*_j\right)} \;=\; \tilde{w}^*_{t+1} + \sqrt{\tilde{\lambda}\sum_{j=t+1}^{\tilde{T}} \tilde{w}^*_j}\,. \tag{17}$$

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

$$\tilde{w}^*_{t+1} \;\le\; \frac{1}{\tilde{\lambda}}(\tilde{w}^*_t)^2 \quad \forall t \in \{1,2,\dots,\tilde{T}-1\}\,. \tag{18}$$

Let $t \le \tilde{T}-1$ be any index such that[5] $\tilde{w}^*_{t+1} \le \frac{\tilde{\lambda}}{2}$. By Eq. (18), this implies that

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

Using the inequality above together with Eq. (17) and the monotonicity of $\tilde{w}^*_t$, we get that for all $t \le \tilde{T}-1$ such that $\tilde{w}^*_{t+1} \le \frac{\tilde{\lambda}}{2}$,

$$\begin{aligned}\tilde{w}^*_t \;&=\; \tilde{w}^*_{t+1} + \sqrt{\tilde{\lambda}\sum_{j=t+1}^{\tilde{T}} \tilde{w}^*_j} \;\le\; \tilde{w}^*_{t+1} + \sqrt{2\tilde{\lambda}\tilde{w}^*_{t+1}} \;=\; \sqrt{\tilde{w}^*_{t+1}}\left(\sqrt{\tilde{w}^*_{t+1}} + \sqrt{2\tilde{\lambda}}\right)\\ &\le\; \sqrt{\tilde{w}^*_{t+1}}\left(\sqrt{\frac{\tilde{\lambda}}{2}} + \sqrt{2\tilde{\lambda}}\right) \;\le\; 3\sqrt{\tilde{\lambda}\tilde{w}^*_{t+1}}.\end{aligned}$$

This chain of inequalities implies that

$$w_{t+1} \ge \frac{(\tilde{w}^*_t)^2}{9\tilde{\lambda}} \qquad \forall t \in \{1,2,\ldots,\tilde{T}-1\} : \tilde{w}^*_{t+1} \le \frac{\tilde{\lambda}}{2}.$$

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

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

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

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

Proof. We need to upper bound the squared Euclidean norm of $(\tilde{w}^*_1,\ldots,\tilde{w}^*_{\tilde{T}})$. Note that for any vector $\mathbf{w}$, we have $\|\mathbf{w}\|^2 = \sum_i w_i^2 \le (\max_i |w_i|)\sum_i |w_i| = \|\mathbf{w}\|_\infty\|\mathbf{w}\|_1$. Thus, by Lemma 1, Eq. (15), and the assumption that $\gamma \ge 10^4(\lambda/\mu_2)^2 > 277\tilde{\lambda}^2$, the squared norm is at most

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

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

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

[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 $\gamma \ge 10^4(\lambda/\mu_2)^2$, hence $\gamma \ge 277\tilde{\lambda}^2$). On the other hand, if we let t_1 be the largest index $\le \tilde{T}$ satisfying $\tilde{w}^*_{t_1} > \tilde{\lambda}/2$, we have by Lemma 1 that $\frac{\gamma}{\tilde{\lambda}} \ge \sum_{t=1}^{t_1}\tilde{w}^*_t \ge t_1\tilde{w}^*_{t_1} > \frac{t_1\tilde{\lambda}}{2}$, which implies that $t_1 \le \frac{2\gamma}{\tilde{\lambda}^2}$, 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.

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

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

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

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

$$\begin{aligned}\frac{\mu_2}{6}\hat{f}(\mathbf{w}) &= \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) + \frac{\tilde{\lambda}}{2}\|\mathbf{w}\|^2 - \gamma \cdot w_1\right) \\ &= \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) - \gamma \cdot w_1\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2,\end{aligned}$$

where in the last step we used the fact that $\tilde{\lambda} = 6\lambda/\mu_2$. Recalling that

$$f(\mathbf{w}) = \frac{\mu}{6}\left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{v}_i, \mathbf{w}\rangle - \langle \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \langle \mathbf{v}_1, \mathbf{w}\rangle\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

and that $\mathbf{v}_1, \mathbf{v}_2, \ldots$ are orthogonal, we can write $f(\mathbf{w})$ as $\frac{\mu}{6} \cdot \hat{f}(V\mathbf{w})$, 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 $V\mathbf{w}^* = (\langle \mathbf{v}_1, \mathbf{w}^*\rangle, \langle \mathbf{v}_2, \mathbf{w}_2^*\rangle, \ldots) = \tilde{\mathbf{w}}^*$. □

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

$$\|\mathbf{w}^*\|^2 = \left\langle \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^*\rangle \mathbf{v}_i, \sum_{j=1}^{\tilde{T}} \langle \mathbf{v}_j, \mathbf{w}^*\rangle \mathbf{v}_j \right\rangle = \sum_{i,j=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^*\rangle \langle \mathbf{v}_j, \mathbf{w}^*\rangle \langle \mathbf{v}_i, \mathbf{v}_j\rangle = \sum_{i=1}^{\tilde{T}} \langle \mathbf{v}_i, \mathbf{w}^*\rangle^2.$$

□

4.2 Oracle Complexity Lower Bound

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

Proposition 2. *Assume that $\epsilon < \min\left\{\frac{54^2\cdot\lambda^3}{\mu_2^2}, \frac{\gamma\lambda}{8}\right\}$. Under the conditions of Proposition 1, 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*

$$\max\left\{\frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}}\,,\ \log_2\log_{18}\left(\frac{54^2\cdot\lambda^3}{\mu_2^2\epsilon}\right) - 1\right\}$$

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 $h(\mathbf{w}) \triangleq f(\mathbf{w}) - \frac{\lambda}{2}\|\mathbf{w}\|^2$ and prove the result on $h(\mathbf{w})$. Using the definition of h and differentiating, we have that

$$h(\mathbf{w}) = \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \langle \mathbf{v}_1, \mathbf{w}\rangle\right)$$

$$\nabla h(\mathbf{w}) = \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g'(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)(\mathbf{v}_i - \mathbf{v}_{i+1}) - \mathbf{v}_1\right)$$

$$\nabla^2 h(\mathbf{w}) = \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g''(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)(\mathbf{v}_i - \mathbf{v}_{i+1})(\mathbf{v}_i - \mathbf{v}_{i+1})^T\right)$$

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 $\mathbf{v}_{t+1}, \mathbf{v}_{t+2}, \ldots, \mathbf{v}_{\tilde{T}}$. □

Let us now fix any number of iterations $T \leq \tilde{T}$. Using the previous results, we can provide a way to pick $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{\tilde{T}}$ for any deterministic algorithm, such that we can provide a lower bound for the number of second-order oracle calls.

- 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 1), 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 1, we can lower bound the above by

$$\frac{54^2 \lambda^2}{\mu_2^2} \cdot (18)^{-2^{(T+1)}}.$$

Using the strong convexity of f, we therefore get

$$f(\mathbf{w}_T) - f(\mathbf{w}^*) \geq \frac{\lambda}{2} \cdot \|\mathbf{w}_T - \mathbf{w}^*\|^2 \geq \frac{54^2 \cdot \lambda^3}{\mu_2^2} (18)^{-2^{(T+1)}}.$$

To make the right-hand side smaller than ϵ, T must satisfy

$$(18)^{-2^{(T+1)}} \leq \frac{\mu_2^2 \epsilon}{54^2 \cdot \lambda^3}$$

Which is equivalent to

$$2^{(T+1)} \geq \log_{18} \left(\frac{54^2 \cdot \lambda^3}{\mu_2^2 \epsilon} \right) .$$

Assuming $\epsilon < \frac{54^2 \cdot \lambda^3}{\mu_2^2}$, then

$$T \;\geq\; \log_2 \log_{18} \left(\frac{54^2 \cdot \lambda^3}{\mu_2^2 \epsilon} \right) - 1 .$$

We now turn to argue that we can also lower bound T by $\frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}}$. Otherwise, suppose by contradiction that we can have $f(\mathbf{w}_T) - f(\mathbf{w}^*) \leq \epsilon$ for some $T < \frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}}$. From Proposition 1 we know that

$$\langle \mathbf{v}_T, \mathbf{w}^* \rangle \;\geq\; \frac{\gamma^{3/4}}{7\sqrt{6\lambda/\mu_2}} + \sqrt{\gamma} \left(\frac{1}{2} - T \right) ,$$

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

$$f(\mathbf{w}_T) - f(\mathbf{w}^*) \;\geq\; \frac{\lambda}{2} \cdot \|\mathbf{w}_T - \mathbf{w}^*\|^2 \;\geq\; \frac{\lambda}{2} \cdot \langle \mathbf{v}_T, \mathbf{w}^* \rangle^2 \;\geq\; \frac{\lambda}{2} \left(\frac{\gamma^{3/4}}{7\sqrt{6\lambda/\mu_2}} + \sqrt{\gamma} \left(\frac{1}{2} - T \right) \right)^2 .$$

To make the right-hand side smaller than ϵ, T must satisfy

$$\left(\frac{\gamma^{3/4}}{7\sqrt{6\lambda/\mu_2}} + \sqrt{\gamma} \left(\frac{1}{2} - T \right) \right)^2 \leq \frac{2\epsilon}{\lambda},$$

or equivalently

$$T \;\geq\; \frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}} + \frac{1}{2} - \sqrt{\frac{2\epsilon}{\gamma\lambda}}.$$

But since we assume $\epsilon < \frac{\gamma\lambda}{8}$, this is at least $\frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}}$, contradicting our earlier assumption.

Overall, we showed that T is lower bounded by both $\frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}}$, as well as $\log_2 \log_{18} \left(\frac{54^2 \cdot \lambda^3}{\mu_2^2 \epsilon} \right) - 1$, hence proving Proposition 2.

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 8. *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

$$\hat{f}(\mathbf{w}) = \frac{\mu_2}{6} \left(\sum_{i=1}^{\tilde{T}-1} g(w_i - w_{i+1}) - \gamma \cdot w_1 \right) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

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

$$\hat{f}(\mathbf{w}) = \frac{\mu_2}{6}\left(\sum_{i=1}^{\tilde{T}-1} g(\langle \mathbf{r}_i, \mathbf{w}\rangle) - \gamma \cdot w_1\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

Differentiating twice, we get

$$\nabla^2 \hat{f}(\mathbf{w}) = \frac{\mu_2}{6}\sum_{i=1}^{\tilde{T}-1} g''(\langle \mathbf{r}_i, \mathbf{w}\rangle) \cdot \mathbf{r}_i \mathbf{r}_i^\top + \lambda I.$$

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

$$\frac{\mu_2 \Delta}{3} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} \mathbf{r}_i \mathbf{r}_i^\top \right\| + \lambda,$$

and the first term equals

$$\begin{aligned}
\frac{\mu_2\Delta}{3} \cdot \sqrt{\max_{\mathbf{x}} \frac{\sum_{i=1}^{\tilde{T}-1}\langle \mathbf{r}_i, \mathbf{x}\rangle^2}{\|\mathbf{x}\|^2}} &= \frac{\mu_2\Delta}{3} \cdot \sqrt{\max_{\mathbf{x}} \frac{\sum_{i=1}^{\tilde{T}-1}(x_i - x_{i+1})^2}{\|\mathbf{x}\|^2}} \\
&\leq \frac{\mu_2\Delta}{3} \cdot \sqrt{\max_{\mathbf{x}} \frac{\sum_{i=1}^{\tilde{T}-1}(2x_i^2 + 2x_{i+1}^2)}{\sum_{i=1}^{d} x_i^2}} \leq \frac{\mu_2\Delta}{3} \cdot \sqrt{4} = \frac{2\mu_2\Delta}{3}.
\end{aligned}$$

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 $\mathbf{w}, \tilde{\mathbf{w}}$,

$$\begin{aligned}
\|\nabla^2 \hat{f}(\mathbf{w}) - \nabla^2 \hat{f}(\tilde{\mathbf{w}})\| &= \frac{\mu_2}{6} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} (g''(\langle \mathbf{r}_i, \mathbf{w}\rangle) - g''(\langle \mathbf{r}_i, \tilde{\mathbf{w}}\rangle)) \cdot \mathbf{r}_i \mathbf{r}_i^\top \right\| \\
&\leq \frac{\mu_2}{6} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} |g''(\langle \mathbf{r}_i, \mathbf{w}\rangle) - g''(\langle \mathbf{r}_i, \tilde{\mathbf{w}}\rangle)| \cdot \mathbf{r}_i \mathbf{r}_i^\top \right\| \\
&\leq \frac{\mu_2}{6} \cdot \left\| \sum_{i=1}^{\tilde{T}-1} 2|\langle \mathbf{r}_i, \mathbf{w} - \mathbf{w}'\rangle| \cdot \mathbf{r}_i \mathbf{r}_i^\top \right\| \\
&\leq \frac{\mu_2}{6} \cdot 2\sqrt{2} \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\| \cdot \left\| \sum_{i=1}^{\tilde{T}-1} \mathbf{r}_i \mathbf{r}_i^\top \right\|.
\end{aligned}$$

Using the same calculations as earlier, we have $\left\|\sum_{i=1}^{\tilde{T}-1} \mathbf{r}_i \mathbf{r}_i^\top\right\| \leq 2$, and therefore we showed overall that

$$\|\nabla^2 \hat{f}(\mathbf{w}) - \nabla^2 \hat{f}(\tilde{\mathbf{w}})\| \leq \frac{\mu_2 \cdot 4\sqrt{2}}{6} \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\| < \mu_2 \cdot \|\mathbf{w} - \tilde{\mathbf{w}}\|,$$

hence $\nabla^2 \hat{f}(\mathbf{w})$ is μ_2-Lipschitz. □

We now collect the ingredients necessary to fix γ, Δ and hence prove our theorem. Combining the previous lemma, Proposition 1 and Proposition 2, 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

$$T \;\geq\; \max\left\{\frac{\gamma^{1/4}}{7\sqrt{6\lambda/\mu_2}}\;,\; \log_2\log_{18}\left(\frac{54^2\cdot\lambda^3}{\mu_2^2\epsilon}\right)-1\right\}, \tag{19}$$

assuming the following conditions:

$$\gamma \geq 10^4\left(\frac{\lambda}{\mu_2}\right)^2\;,\;\; \Delta \geq \sqrt{\gamma}\;,\;\; \epsilon < \min\left\{\frac{54^2\cdot\lambda^3}{\mu_2^2}, \frac{\gamma\lambda}{8}\right\}\;,\;\; \frac{2\mu_2\Delta}{3}+\lambda \leq \mu_1\;,\;\; \sqrt{\frac{2\gamma^{7/4}}{(6\lambda/\mu_2)^{3/2}}} \leq D\;.$$

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

$$\gamma \geq 10^4\left(\frac{\lambda}{\mu_2}\right)^2\;,\;\; \epsilon < \min\left\{\frac{54^2\cdot\lambda^3}{\mu_2^2}, \frac{\gamma\lambda}{8}\right\}\;,\;\; \gamma \leq \left(\frac{3(\mu_1-\lambda)}{2\mu_2}\right)^2\;,\;\; \gamma \leq \sqrt[7]{\frac{D^8(6\lambda)^6}{2^4\mu_2^6}}\;.$$

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

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

Clearly, to make the lower bound in Eq. (19) as large as possible, we should pick the largest possible γ, namely $\gamma = \min\left\{\left(\frac{3(\mu_1-\lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(6\lambda)^6}{2^4\mu_2^6}}\right\}$, and to ensure that the other conditions hold, require that

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

Simplifying a bit, these two conditions are implied by requiring

$$\frac{\mu_1}{\lambda} \geq 68\;,\;\; \frac{\mu_2}{\lambda}D \geq 1167\;,\;\; \epsilon < \frac{10^4\lambda^3}{8\mu_2^2}, \tag{20}$$

Finally, let us plug our choice of $\gamma = \min\left\{\left(\frac{3(\mu_1-\lambda)}{2\mu_2}\right)^2, \sqrt[7]{\frac{D^8(6\lambda)^6}{2^4\mu_2^6}}\right\}$ into the lower bound in Eq. (19). We thus get an oracle complexity lower bound of

$$\begin{aligned}
&\max\left\{\min\left\{\frac{\sqrt{\mu_2}}{7\sqrt{6\lambda}}\cdot\sqrt{\frac{3(\mu_1-\lambda)}{2\mu_2}}\;,\; \frac{\sqrt{\mu_2}}{7\sqrt{6\lambda}}\cdot\frac{D^{2/7}(6\lambda)^{3/14}}{2^{1/7}\mu_2^{3/14}}\right\}\;,\; \log_2\log_{18}\left(\frac{54^2\cdot\lambda^3}{\mu_2^2\epsilon}\right)-1\right\} \\
&= \max\left\{\min\left\{\frac{1}{14}\sqrt{\frac{\mu_1-\lambda}{\lambda}}\;,\; \frac{(D\mu_2/6\lambda)^{2/7}}{7\cdot 2^{1/7}}\right\}\;,\; \log_2\log_{18}\left(\frac{54^2\cdot\lambda^3}{\mu_2^2\epsilon}\right)-1\right\},
\end{aligned}$$

under the conditions of Eq. (20).

To simplify the bound a bit, we note that we can lower bound $\mu_1-\lambda$ by $\frac{67}{68}\mu_1$ (possible by Eq. (20)), and lower bound $\log_2\log_{18}\left(\frac{54^2\cdot\lambda^3}{\mu_2^2\epsilon}\right)-1$ 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 Eq. (20)). Finally, using the fact that $\max\{a,b\} \geq (a+b)/2$, the result in the theorem follows.

5 Proof of Thm. 2

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

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

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

$$f_T(\mathbf{w}) = \frac{\mu_2}{6}\left(g(\langle \mathbf{v}_1, \mathbf{w}\rangle) + g(\langle \mathbf{v}_T, \mathbf{w}\rangle) + \sum_{i=1}^{T-1} g(\langle \mathbf{v}_i, \mathbf{w}\rangle - \langle \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \gamma\langle \mathbf{v}_1, \mathbf{w}\rangle\right).$$

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 8). Our goal will be to show a lower bound on the optimization error using this type of function.

5.1 Minimizer of f_T

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

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

It is easily verified that the minimal values of $\hat{f}_T$ and f_T are the same, 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 $\hat{\mathbf{w}}^*$.

We begin with the following technical lemma:

Lemma 9. *$\hat{f}_T$ 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 all $t = 1, 2, \ldots, T$, where δ is non-negative and independent of t. Moreover,

$$g'(\hat{w}_1^*) + g'(\hat{w}_T^*) = \gamma\,.$$

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

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

as well as

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

for all $j \in \{2, 3, \ldots, T-1\}$. 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 $j \in \{2, 3, \ldots, T-1\}$, as well as $\hat{w}^*_{T-1} - \hat{w}^*_T = \hat{w}^*_T$. From this, it follows by straightforward induction that $\hat{w}^*_{T+1-t} = t \cdot \hat{w}^*_T$, from which the first displayed equation in the lemma follows. This also implies $g'(T\hat{w}^*_T) + g'(\hat{w}^*_T) = \gamma$, and since g' is strictly monotonic, we have that $\hat{w}^*_T$ is uniquely defined, and since the other coordinates of $\hat{\mathbf{w}}^*$ are also uniquely defined given $\hat{w}^*_T$, 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 $\hat{f}_T(\hat{\mathbf{w}}^*) > 0$, even though $\hat{f}_T(\mathbf{0}) = 0$, violating the fact that $\hat{\mathbf{w}}^*$ minimizes $\hat{f}_T$. □

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

$$\|\mathbf{w}^*\| = \|\hat{\mathbf{w}}^*\| \quad \text{and} \quad f_T(\mathbf{w}^*) = \frac{\mu_2}{6} \cdot \hat{f}_T(\hat{\mathbf{w}}^*)\,, \tag{21}$$

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

Proposition 3. *The function $\hat{f}_T$ and its minimizer $\hat{\mathbf{w}}^*$ has the following properties, depending on the values of γ, Δ, T:*

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

Proof. To prove the proposition, we will consider three regimes, depending on T, δ, Δ: Namely, $T\delta \le \Delta$, $\frac{\Delta}{T} < \delta \le \Delta$ 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: $T\delta \le \Delta$. In that case, $\hat{w}^*_1, \hat{w}^*_T$ as well as $\hat{w}^*_i - \hat{w}^*_{i+1}$ for all $i = 2, \ldots, T-1$ in the definition of $\hat{f}_T$ all lie in the interval where g is a cubic function. Using Lemma 9,

$$g'(w^*_1) + g'(w^*_T) = w^{*2}_1 + w^{*2}_T = \gamma$$

hence

$$\delta^2 T^2 + \delta^2 = \gamma$$

and

$$\delta = \sqrt{\frac{\gamma}{1+T^2}}.$$

Therefore, our condition $T\delta \le \Delta$ is exactly equivalent to $\gamma \le \frac{\Delta^2\left(1+T^2\right)}{T^2}$, namely the first regime discussed

in the proposition. We now establish the relevant bounds:

$$\begin{aligned}\hat{f}_T(\hat{\mathbf{w}}^*) =& \frac{1}{3}\left(\frac{\sqrt{\gamma}(1+T)}{\sqrt{(1+T^2)}}\right)^3\left[\left(1-\frac{1}{1+T}\right)^3+\left(1-\frac{T}{1+T}\right)^3\right] \\ &+\frac{1}{3}(T-1)\left(\sqrt{\frac{\gamma}{(1+T^2)}}\right)^3-\frac{\gamma^{3/2}T}{\sqrt{(1+T^2)}} \\ =& \frac{1}{3}\left(\sqrt{\frac{\gamma}{(1+T^2)}}\right)^3\left(T^3+1+T-1\right)-\frac{\gamma^{3/2}T}{\sqrt{(1+T^2)}} \\ =& \frac{\gamma^{3/2}T}{\sqrt{(1+T^2)}}\left(\frac{1}{3}-1\right)\end{aligned}$$

and

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

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

Case 2: $\frac{\Delta}{T} < \delta \leq \Delta$. In this case, by Lemma 9, $\hat{w}_T^* \leq \Delta$ but $\hat{w}_1^* > \Delta$. Therefore, in the definition $\hat{f}_T(\hat{\mathbf{w}}^*)$, $g(\hat{w}_1^*)$ lies in the quadratic region of g, whereas $g(\hat{w}_T^*)$ and $g'(\hat{w}_i^* - \hat{w}_{i+1}^*)$ for all i lies in the cubic region of g. As a result,

$$g'(w_1^*)+g'(w_T^*) = 2\Delta w_1^* - \Delta^2 + w_T^{*2} = \gamma.$$

Plugging in $w_T^* = \delta$ and $w_1^* = T\cdot\delta$, we get

$$\delta^2 + 2\Delta\delta\cdot T - \left(\gamma+\Delta^2\right) = 0\,,$$

and therefore (using the fact $\delta \geq 0$, see Lemma 9),

$$\delta = -\Delta T + \Delta T\sqrt{1+\frac{\gamma+\Delta^2}{\Delta^2T^2}}\,.$$

This, plus the assumption $\frac{\Delta}{T} < \delta \leq \Delta$, is equivalent to $\frac{\Delta^2\left(1+T^2\right)}{T^2} < \gamma \leq 2\Delta^2T$, hence showing that we are indeed in the second regime as specified in our proposition. Turning to calculate the relevant bounds, we have

$$\hat{f}_T(\hat{\mathbf{w}}^*) = T\delta^3 + \Delta^2T^2\delta^2 - T\left(\Delta^2+\gamma\right)\delta + \frac{\Delta^3}{3}\,.$$

Moreover,

$$\|\hat{\mathbf{w}}^*\|^2 \;=\; \delta^2(T+1)^2\sum_{t=1}^{T}\left(1-\frac{t}{1+T}\right)^2,$$

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 $\frac{\left(\gamma+\Delta^2\right)^2(T+1)^3}{12\Delta^2T^2}$.

Case 3: $\delta > \Delta$. In this case, by Lemma 9, we have $\hat{w}_1^* > \hat{w}_T^* = \hat{w}_i^* - \hat{w}_{i+1}^* > \Delta$, which implies that in the definition of $\hat{f}_t(\hat{\mathbf{w}}^*)$, these terms all lie in the quadratic region of g. Therefore,

$$g'(w_1^*) + g'(w_T^*) = 2\Delta w_1^* - \Delta^2 + 2\Delta w_T^* - \Delta^2 = \gamma \,,$$

and thus

$$2\Delta(T+1)\delta = \gamma + 2\Delta^2 \,,$$

or equivalently

$$\delta = \frac{\gamma + 2\Delta^2}{2\Delta(T+1)} \,.$$

Note that this, plus our assumption $\delta > \Delta$, is equivalent to $\gamma > 2\Delta^2 T$, which shows that we are indeed in the third regime as specified in our proposition. Turning to calculate $\|\hat{\mathbf{w}}^*\|$ and $\hat{f}_T(\hat{\mathbf{w}}^*)$, we have

$$\begin{aligned}
\hat{f}_T(\hat{\mathbf{w}}^*) &= \Delta\hat{w}_1^{*2} - \Delta^2\hat{w}_1^* + \frac{1}{3}\Delta^3 + \Delta\hat{w}_T^{*2} - \Delta^2\hat{w}_T^* + \frac{1}{3}\Delta^3 + \\
&\qquad \sum_{i=1}^{T-1}\left(\Delta(\hat{w}_i^* - \hat{w}_{i+1}^*)^2 - \Delta^2(\hat{w}_i^* - \hat{w}_{i+1}^*) + \frac{1}{3}\Delta^3\right) - \gamma\hat{w}_1 \\
&= T(T+1)\Delta\delta^2 - T(\gamma + 2\Delta^2)\delta + \frac{(T+1)\Delta^3}{3} \\
&= \frac{T(\gamma+2\Delta^2)^2}{4\Delta(T+1)} - \frac{T\left(\gamma+2\Delta^2\right)^2}{2\Delta(T+1)} + \frac{(T+1)\Delta^3}{3} \\
&= -\frac{T\left(\gamma+2\Delta^2\right)^2}{4\Delta(T+1)} + \frac{(T+1)\Delta^3}{3} \,,
\end{aligned}$$

and

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

□

5.2 Oracle Complexity Lower Bound

Given the expressions on the optimal value of $\hat{f}_T$, 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 Thm. 1, subsection 4.2): Specifically, consider the function f_{2T}, given by

$$f_{2T}(\mathbf{w}) = \frac{\mu_2}{6}\left(g(\langle\mathbf{v}_1,\mathbf{w}\rangle) + g(\langle\mathbf{v}_{2T},\mathbf{w}\rangle) + \sum_{i=1}^{2T-1} g(\langle\mathbf{v}_i,\mathbf{w}\rangle - \langle\mathbf{v}_{i+1},\mathbf{w}_T\rangle) - \gamma\langle\mathbf{v}_1,\mathbf{w}\rangle\right).$$

Given an algorithm, we choose $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T$ 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, since the gradients and Hessians of f_{2T} at each $\mathbf{w}_t$ reveals no information on future $\mathbf{v}_t$'s – see Lemma 7). Also, we let $\mathbf{v}_{T+1}, \ldots, \mathbf{v}_{2T}$ equal $\mathbf{v}_T$.

With this choice, it is easily verified that

$$f_{2T}(\mathbf{w}_T) = \frac{\mu_2}{6}\left(g(\langle \mathbf{v}_1, \mathbf{w}\rangle) + g(\langle \mathbf{v}_T, \mathbf{w}\rangle) + \sum_{i=1}^{T-1} g(\langle \mathbf{v}_i, \mathbf{w}_T\rangle - \langle \mathbf{v}_{i+1}, \mathbf{w}_T\rangle) - \gamma\langle \mathbf{v}_1, \mathbf{w}_T\rangle\right),$$

which is clearly no better 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 Eq. (21), this equals

$$\frac{\mu_2}{6}\left(\min_{\mathbf{w}} \hat{f}_T(\mathbf{w}) - \min_{\mathbf{w}} \hat{f}_{2T}(\mathbf{w})\right).$$

Using proposition 3, 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

$$\gamma = \frac{D^2}{12T}$$

Using this and the assumption on the parameters, we get that $\gamma \leq \frac{\Delta^2(1+4T^2)}{4T^2}T$, and therefore, we are in the first regime for both f_T and f_{2T} as specified in proposition 3. 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

$$\|\hat{\mathbf{w}}_{2T}^*\|_2^2 \leq \frac{\gamma(1+2T)^3}{3\left(1+4T^2\right)} \leq \frac{27D^2T^3}{144T^2} \leq D^2$$

as required.

Using the results from proposition 3 for the first regime we can compute the optimization error bound

$$\begin{aligned}
\hat{f}_T^* - \hat{f}_{2T}^* &= \frac{2\gamma^{3/2}T}{3\sqrt{(1+4T^2)}} - \frac{\gamma^{3/2}T}{3\sqrt{(1+T^2)}} \\
&= \frac{\gamma^{3/2}}{3}\left(\frac{1}{\sqrt{\left(1+\frac{1}{4T^2}\right)}} - \frac{1}{\sqrt{\left(1+\frac{1}{T^2}\right)}}\right) \\
&\geq \frac{\gamma^{3/2}}{3}\left(1 - \frac{1}{8T^2} - \left(1 - \frac{1}{2T^2} + \frac{3}{8T^4}\right)\right) \\
&= \frac{\gamma^{3/2}}{3}\left(\frac{3}{8T^2} - \frac{3}{8T^4}\right) = \frac{\gamma^{3/2}\left(T^2-1\right)}{8T^4} \\
&\geq \frac{D^3}{672T^{7/2}}
\end{aligned}$$

Where in the last inequality we used the facts 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 $x \geq 0$. Hence, the suboptimality is at least $\frac{\mu_2 D^3}{4032T^{7/2}}$.

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

In this setting, we choose

$$\gamma = \frac{D\Delta}{\sqrt{12T}}. \tag{22}$$

Using this and the assumption on the parameters, we get that $\frac{\Delta^2(1+T^2)}{T^2} < \gamma < 2\Delta^2 T$, and therefore, we are in the second regime for both f_T and f_{2T} as specified in proposition 3. 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

$$\|\hat{\mathbf{w}}^*_{2T}\|^2 \;\leq\; \frac{\left(\gamma + \Delta^2\right)^2 (2T+1)^3}{48\Delta^2 T^2} \;\leq\; \frac{\gamma^2 (2T+1)^3}{12\Delta^2 T^2} \;=\; \frac{D^2(2T+1)^3}{144T^3} \;\leq\; D^2$$

as required.

Turning to compute the optimization error bound, and letting δ_T, δ_{2T} denote the quantity δ in proposition 3 for $\hat{f}_T$ and $\hat{f}_{2T}$ respectively, we have

$$f^*_T - f^*_{2T} \;=\; (2\delta_{2T} - \delta_T)\left(T\left(\Delta^2 + \gamma\right) - \Delta T^2 (2\delta_{2T} + \delta_T)\right) + \frac{1}{3}T\left(\delta_T^3 - 2\delta_{2T}^3\right). \tag{23}$$

To continue, we use the following auxiliary lemma:

Lemma 10. $(2\delta_{2T} - \delta_T)\left(T\left(\Delta^2 + \gamma\right) - \Delta T^2 (2\delta_{2T} + \delta_T)\right) \geq 0$

Proof. First we will prove that $T\left(\Delta^2 + \gamma\right) - \Delta T^2 (2\delta_{2T} + \delta_T) \geq 0$.

Since $\delta_T = -\Delta T + \Delta T\sqrt{1 + \frac{\gamma+\Delta^2}{\Delta^2 T^2}}$ and using $\sqrt{1+x} \leq 1 + \frac{1}{2}x$ for $x \geq 0$ we have that:

$$\delta_T \leq \frac{\left(\gamma + \Delta^2\right)}{2\Delta T}$$

So

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

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:

$$\begin{aligned}
&-4 + 4\sqrt{1 + \frac{1}{4}\alpha} + 1 - \sqrt{1+\alpha} \geq 0 \\
\iff\; &4\sqrt{1 + \frac{1}{4}\alpha} \geq 3 + \sqrt{1+\alpha} \\
\iff\; &16\left(1 + \frac{1}{4}\alpha\right) \geq 9 + 6\sqrt{1+\alpha} + 1 + \alpha \\
\iff\; &6 + 3\alpha \geq 6\sqrt{1+\alpha}
\end{aligned}$$

Which is true since $\sqrt{1+\alpha} \leq 1 + \frac{1}{2}\alpha$. $\square$

With this lemma, we can lower bound the optimization error in Eq. (23) by

$$\frac{1}{3}T\left(\delta_T^3 - 2\delta_{2T}^3\right). \tag{24}$$

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

$$\frac{\left(\gamma+\Delta^2\right)}{2\Delta T} - \frac{\left(\gamma+\Delta^2\right)^2}{8\Delta^3 T^3} \leq \delta_T \leq \frac{\left(\gamma+\Delta^2\right)}{2\Delta T}.$$

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 Eq. (24) by

$$\frac{1}{3}T\left(\delta_T - \sqrt[3]{2}\delta_{2T}\right)^3 \geq \frac{\left(\gamma+\Delta^2\right)^3}{60\Delta^3 T^2} \geq \frac{D^3}{2500T^{7/2}}.$$

Hence, the suboptimality is at least $\frac{\mu_2 D^3}{15000T^{7/2}}$.

5.3.3 Case 3:

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 3. 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

$$\|\hat{\mathbf{w}}_{2T}^*\|^2 \leq \frac{\left(\gamma+2\Delta^2\right)^2(2T+1)}{12\Delta^2} \leq \frac{4\gamma^2(2T+1)}{12\Delta^2} = \frac{D^2(2T+1)}{9T} \leq D^2$$

Now, since by assumption $T\Delta^3 < \frac{\Delta D^2}{48T^2}$ the optimization error bound is

$$\hat{f}_T^* - \hat{f}_{2T}^* \geq \frac{\left(\gamma+2\Delta^2\right)^2}{8\Delta T} - \frac{T\Delta^3}{3} \geq \frac{D^2\Delta}{24T^2} - \frac{D^2\Delta}{144T^2} \geq \frac{\Delta D^2}{30T^2}$$

Hence, the suboptimality is at least $\frac{\mu_1 D^2}{180T^2}$.

5.4 Wrapping Up

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

$$f_{2T}(\mathbf{w}_T) - \min_{\mathbf{w}} f_{2T}(\mathbf{w}) \geq \begin{cases} \frac{\mu_2 D^3}{15000T^{7/2}} & \frac{D^2}{48\Delta^2 T^3} \leq 1 \\ \frac{\mu_1 D^2}{180T^2} & \frac{D^2}{48\Delta^2 T^3} > 1 \end{cases} .$$

Thus, we get that

$$f_{2T}(\mathbf{w}_T) - \min_{\mathbf{w}} f_{2T}(\mathbf{w}) \geq \min \left\{ \frac{\mu_2 D^3}{15000T^{7/2}}, \frac{\mu_1 D^2}{180T^2} \right\} .$$

Equating these bounds to ϵ, and solving for T, the theorem follows.

Acknowledgements

We thank Yurii Nesterov for several helpful comments on a preliminary version of this paper.

References

Yossi Arjevani and Ohad Shamir. Oracle complexity of second-order methods for finite-sum problems. *arXiv preprint arXiv:1611.04982*, 2016.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems. *Siam journal on optimization*, 20(6):2833–2852, 2010.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optimization Methods and Software*, 27(2): 197–219, 2012.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Leonid Vital'evich Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

Arkadi Nemirovski. Efficient methods in convex programming – lecture notes, 2005.

Arkadi Nemirovsky and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

Yurii Nesterov. Accelerating the cubic regularization of newtons method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.