A. Proofs

A.1. Auxiliary Lemmas

The following lemma was essentially proven in (Lan, 2015; Nesterov, 2013), but we provide a proof for completeness:

Lemma 1. *Fix $\alpha, \beta \geq 0$, and consider the following function on $\mathbb{R}^d$:*

$$F(\mathbf{w}) = \frac{\alpha}{8}\left(w_1^2 + \sum_{i=1}^{d-1}(w_i - w_{i+1})^2 + (a_{\tilde{\kappa}} - 1)w_d^2 - w_1\right) + \frac{\beta}{2}\|\mathbf{w}\|^2,$$

and $a_{\tilde{\kappa}} = \frac{\sqrt{\tilde{\kappa}}+3}{\sqrt{\tilde{\kappa}}+1}$ where $\tilde{\kappa} = \frac{\alpha+\beta}{\beta}$ is the condition number of F. Then F is β strongly convex, $(\alpha+\beta)$-smooth, and has a unique minimum at $(q, q^2, q^3, \ldots, q^d)$ where $q = \frac{\sqrt{\tilde{\kappa}}-1}{\sqrt{\tilde{\kappa}}+1}$.

Proof. The function is equivalent to

$$F(\mathbf{w}) = \frac{\alpha}{8}\left(\mathbf{w}^\top A\mathbf{w} - w_1\right) + \frac{\beta}{2}\|\mathbf{w}\|^2,$$

where

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & a_{\tilde{\kappa}} \end{pmatrix}.$$

Since A is symmetric, all its eigenvalues are real. Therefore, by Gershgorin circle theorem and the fact that $a_{\tilde{\kappa}} \in [1, 2]$ (since $\tilde{\kappa} \geq 1$), we have that all the eigenvalues of A lie in $[0, 4]$. Thus, the eigenvalues of $\nabla^2 F = (\alpha/4)A + \beta I$ lie in $[\beta, \alpha+\beta]$, implying that F is β-strongly convex and $(\alpha+\beta)$-smooth.

It remains to compute the optimum of F. By differentiating F and setting to zero, we get that the optimum $\mathbf{w}$ must satisfy the following set of equations:

$$\begin{aligned} & w_2 - 2\cdot\frac{\tilde{\kappa}+1}{\tilde{\kappa}-1}\cdot w_1 + 1 = 0 \\ & w_{i+1} - 2\cdot\frac{\tilde{\kappa}+1}{\tilde{\kappa}-1}\cdot w_i + w_{i-1} = 0 \quad \forall\, i = 2, \ldots, d-1 \\ & \left(a_{\tilde{\kappa}} + \frac{4}{\tilde{\kappa}-1}\right)w_d - w_{d-1} = 0. \end{aligned}$$

It is easily verified that this is satisfied by the vector $(q, q^2, q^3, \ldots, q^d)$, where $q = \frac{\sqrt{\tilde{\kappa}}-1}{\sqrt{\tilde{\kappa}}+1}$. Since F is strongly convex, this stationary point must be the unique global optimum of F. □

Lemma 2. *For some $q \in (0, 1)$ and positive d, define*

$$g(z) = \begin{cases} q^{2(z+1)} & z < d \\ 0 & z \geq d \end{cases}.$$

Let l be a non-negative random variable, and suppose $d \geq 2\mathbb{E}[l]$. Then $\mathbb{E}[g(l)] \geq \frac{1}{2}q^{2\mathbb{E}[l]+2}$.

Proof. Since $q \in (0, 1)$, the function $z \mapsto q^z$ is convex for non-negative z and monotonically decreasing. Therefore, by definition of g and Jensen's inequality, we have

$$\mathbb{E}[g(l)] = \Pr(l < d)\cdot\mathbb{E}[q^{2(l+1)}|l < d] + \Pr(l \geq d)\cdot 0 \geq \Pr(l < d)\cdot q^{\mathbb{E}[2(l+1)]}.$$

Using Markov's inequality to derive $\Pr(l < d) = 1 - \Pr(l \geq d) \geq 1 - \frac{\mathbb{E}[l]}{d} \geq \frac{1}{2}$, concludes the proof. □

A.2. Proof of Thm. 1

The proof is inspired by a technique introduced in (Woodworth and Srebro, 2016) for analyzing randomized first-order methods, in which a quadratic function is “locally flattened” in order to make first-order (gradient) information non-informative. We use a similar technique to make *second-order* (Hessian) information non-informative, hence preventing second-order methods from having an advantage over first-order methods.

Given a (deterministic) algorithm and a bound T on the number of oracle calls, we construct the function F in the following manner. We first choose some dimension $d \geq 2T$. We then define

$$\kappa = \frac{\mu}{8\lambda} \quad , \quad q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1},$$

and choose $r > 0$ sufficiently small so that

$$\frac{T\mu r^2}{8\lambda} \leq 1 \quad \text{and} \quad \sqrt{\frac{T\mu r^2}{16\lambda}} \leq \frac{1}{2}q^T.$$

We also let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be orthonormal vectors in $\mathbb{R}^d$ (to be specified later). We finally define our function as

$$F(\mathbf{w}) = H(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

where

$$H(\mathbf{w}) = \frac{\lambda(\kappa-1)}{8}\left(\langle \mathbf{v}_1, \mathbf{w}\rangle^2 + \sum_{i=1}^{T-1}\phi_r(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) + (a_\kappa - 1)\phi_r(\langle \mathbf{v}_T, \mathbf{w}\rangle) - \langle \mathbf{v}_1, \mathbf{w}\rangle\right),$$

$a_\kappa = \frac{\sqrt{\kappa}+3}{\sqrt{\kappa}+1}$, and

$$\phi_r(z) = \begin{cases} 0 & |z| \leq r \\ 2(|z|-r)^2 & r < |z| \leq 2r \\ z^2 - 2r^2 & |z| > 2r \end{cases}.$$

It is easy to show that ϕ_r is 4-smooth and satisfies $0 \leq z^2 - \phi_r(z) \leq 2r^2$ for all z.

First, we establish that F is indeed strongly convex and smooth as required:

Lemma 3. *F as defined above is λ-strongly convex and μ-smooth.*

Proof. Since ϕ_r is convex, and the composition of a convex and linear function is convex, we have that $\mathbf{w} \mapsto \phi_r(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)$ are convex for all i, as well as $\mathbf{w} \mapsto \langle \mathbf{v}_1, \mathbf{w}\rangle^2$ and $\mathbf{w} \mapsto \phi_r(\langle \mathbf{v}_T, \mathbf{w}\rangle)$. Therefore, $H(\mathbf{w})$ is convex. As a result, F is λ-strongly convex due to the $\frac{\lambda}{2}\|\mathbf{w}\|^2$ term. As to smoothness, note first that $H(\mathbf{w})$ can be equivalently written as $\tilde{H}(V\mathbf{w})$, where V is some orthogonal $d \times d$ matrix with the first T rows equal to $\mathbf{v}_1, \ldots, \mathbf{v}_T$, and

$$\tilde{H}(\mathbf{x}) = \frac{\lambda(\kappa-1)}{8}\left(x_1^2 + \sum_{i=1}^{T-1}\phi_r(x_i - x_{i+1}) + (a_\kappa - 1)\phi_r(x_T) - x_1\right).$$

Therefore, $\nabla^2 F(\mathbf{w}) = \nabla^2 H(\mathbf{w}) + \lambda I = V^\top \nabla^2 \tilde{H}(V\mathbf{w})V + \lambda I$. It is easily verified that $\nabla^2 \tilde{H}$ at any point (and in particular $V\mathbf{w}$) is tridiagonal, with each element having absolute value at most $2\lambda(\kappa-1)$. Therefore, using the orthogonality of V and the fact that $(a+b)^2 \leq 2(a^2+b^2)$,

$$\begin{aligned}
\sup_{\mathbf{x}:\|\mathbf{x}\|=1} \mathbf{x}^\top \nabla^2 F(\mathbf{w})\mathbf{x} &= \sup_{\mathbf{x}:\|\mathbf{x}\|=1} \mathbf{x}^\top (V^\top \nabla^2 \tilde{H}(V\mathbf{w})V + \lambda I)\mathbf{x} \\
&= \sup_{\mathbf{x}:\|\mathbf{x}\|=1} \mathbf{x}^\top \nabla^2 \tilde{H}(V\mathbf{w})\mathbf{x} + \lambda \\
&\leq \sup_{\mathbf{x}:\|\mathbf{x}\|=1} 2\lambda(\kappa-1)\left(\sum_{i=1}^{d} x_i^2 + 2\sum_{i=1}^{d-1}|x_i x_{i+1}|\right) + \lambda \\
&\leq \sup_{\mathbf{x}:\|\mathbf{x}\|=1} 2\lambda(\kappa-1)\sum_{i=1}^{d-1}(|x_i|+|x_{i+1}|)^2 + \lambda \\
&\leq \sup_{\mathbf{x}:\|\mathbf{x}\|=1} 4\lambda(\kappa-1)\sum_{i=1}^{d-1}(x_i^2+x_{i+1}^2) + \lambda \\
&\leq 8\lambda(\kappa-1)+\lambda \;\leq\; 8\lambda\kappa.
\end{aligned}$$

Plugging in the definition of κ, this equals μ. Therefore, the spectral norm of the Hessian of F at any point is at most μ, and therefore F is μ-smooth. □

By construction, the function F also has the following key property:

Lemma 4. *For any $\mathbf{w} \in \mathbb{R}^d$ orthogonal to $\mathbf{v}_t, \mathbf{v}_{t+1}, \ldots, \mathbf{v}_T$ (for some $t \in \{1,2,\ldots,T-1\}$), it holds that $F(\mathbf{w}), \nabla F(\mathbf{w}), \nabla^2 F(\mathbf{w})$ do not depend on $\mathbf{v}_{t+1}, \mathbf{v}_{t+2}, \ldots, \mathbf{v}_T$.*

Proof. Recall that F is derived from H by adding a $\frac{\lambda}{2}\|\mathbf{w}\|^2$ term, which clearly does not depend on $\mathbf{v}_1, \ldots, \mathbf{v}_T$. Therefore, it is enough to prove the result for $H(\mathbf{w}), \nabla H(\mathbf{w}), \nabla^2 H(\mathbf{w})$. By taking the definition of H and differentiating, we have that $H(\mathbf{w})$ is proportional to

$$\langle \mathbf{v}_1, \mathbf{w}\rangle^2 + \sum_{i=1}^{T-1}\phi_r(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) + (a_\kappa - 1)\phi_r(\langle \mathbf{v}_T, \mathbf{w}\rangle) - \langle \mathbf{v}_1, \mathbf{w}\rangle,$$

$\nabla H(\mathbf{w})$ is proportional to

$$2\langle \mathbf{v}_1, \mathbf{w}\rangle \mathbf{v}_1 + \sum_{i=1}^{T-1}\phi_r'(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)(\mathbf{v}_i - \mathbf{v}_{i+1}) + (a_\kappa - 1)\phi_r'(\langle \mathbf{v}_T, \mathbf{w}\rangle)\mathbf{v}_T - \mathbf{v}_1,$$

and $\nabla^2 H(\mathbf{w})$ is proportional to

$$2\mathbf{v}_1\mathbf{v}_1^\top + \sum_{i=1}^{T-1}\phi_r''(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)(\mathbf{v}_i - \mathbf{v}_{i+1})(\mathbf{v}_i - \mathbf{v}_{i+1})^\top + (a_\kappa - 1)\phi_r''(\langle \mathbf{v}_T, \mathbf{w}\rangle)\mathbf{v}_T\mathbf{v}_T^\top.$$

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

With this lemma at hand, we now turn to describe how $\mathbf{v}_1, \ldots, \mathbf{v}_T$ are constructed:

- 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 $\mathbf{v}_2, \ldots, \mathbf{v}_T$ will also be orthogonal to $\mathbf{w}_1$ (which will be ensured by the construction which follows), we have by Lemma 4 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 $\{\mathbf{v}_2, \ldots, \mathbf{v}_T\}$, 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 4, 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 $\mathbf{v}_{t+1}, \ldots, \mathbf{v}_T$, 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 $\mathbf{v}_T$ to be some unit vector orthogonal to all previously chosen $\mathbf{v}$'s as well as $\mathbf{w}_1, \ldots, \mathbf{w}_T$.

Based on this construction, the following lemma is self-evident:

Lemma 5. *It holds that* $\langle \mathbf{w}_T, \mathbf{v}_T \rangle = 0$.

Based on this lemma, we now turn to argue that $\mathbf{w}_T$ must be a sub-optimal point. We first establish the following result:

Lemma 6. *Letting* $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F(\mathbf{w})$*, it holds that*

$$\left\| \mathbf{w}^\star - \sum_{i=1}^{T} q^i \mathbf{v}_i \right\| \;\le\; \sqrt{\frac{T\mu r^2}{16\lambda}}$$

where $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

Proof. Let F_r denote F, where we make the dependence on the parameter r explicit. We first argue that

$$\sup_{\mathbf{w}\in\mathbb{R}^d} |F_r(\mathbf{w}) - F_0(\mathbf{w})| \;\le\; \frac{T\mu r^2}{32}. \tag{7}$$

This is because

$$|F_r(\mathbf{w}) - F_0(\mathbf{w})| \;\le\; \frac{\lambda(\kappa-1)}{8}\Bigg(\sum_{i=1}^{T-1} |\phi_r(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle) - \phi_0(\langle \mathbf{v}_i - \mathbf{v}_{i+1}, \mathbf{w}\rangle)| \\ + |\phi_r(\langle \mathbf{v}_T, \mathbf{w}\rangle) - \phi_0(\langle \mathbf{v}_T, \mathbf{w}\rangle)|\Bigg),$$

and since $\sup_{z\in\mathbb{R}} |\phi_r(z) - \phi_0(z)| = \sup_{z\in\mathbb{R}} |\phi_r(z) - z^2| \le 2r^2$, the above is at most $\frac{\lambda(\kappa-1)}{4} Tr^2 \le \frac{\lambda\kappa}{4} Tr^2$. Recalling that $\kappa = \mu/8\lambda$, Eq. (7) follows.

Let $\mathbf{w}_r = \arg\min F_r(\mathbf{w})$. By λ-strong convexity of F_0 and F_r,

$$F_0(\mathbf{w}_r) - F_0(\mathbf{w}_0) \;\ge\; \frac{\lambda}{2}\|\mathbf{w}_r - \mathbf{w}_0\|^2 \quad , \quad F_r(\mathbf{w}_0) - F_r(\mathbf{w}_r) \;\ge\; \frac{\lambda}{2}\|\mathbf{w}_0 - \mathbf{w}_r\|^2.$$

Summing the two inequalities and using Eq. (7),

$$\lambda\|\mathbf{w}_r - \mathbf{w}_0\|^2 \;\le\; F_0(\mathbf{w}_r) - F_r(\mathbf{w}_r) + F_r(\mathbf{w}_0) - F_0(\mathbf{w}_0) \;\le\; \frac{T\mu r^2}{16}\ ,$$

and therefore

$$\|\mathbf{w}_r - \mathbf{w}_0\|^2 \;\le\; \frac{T\mu r^2}{16\lambda}. \tag{8}$$

By definition, $\mathbf{w}_r = \mathbf{w}^\star$ from the statement of our lemma, so it only remains to prove that $\mathbf{w}_0 = \arg\min F_0(\mathbf{w})$ equals $\sum_{i=1}^{T} q^i \mathbf{v}_i$. To see this, note that $F_0(\mathbf{w})$ can be equivalently written as $\tilde{F}(V\mathbf{w})$, where V is some orthogonal $d \times d$ matrix with its first T rows equal to $\mathbf{v}_1, \ldots, \mathbf{v}_T$, and

$$\tilde{F}(\mathbf{x}) = \frac{\lambda(\kappa-1)}{8}\left(x_1^2 + \sum_{i=1}^{T-1}(x_i - x_{i+1})^2 + (a_\kappa - 1)x_T^2 - w_1\right) + \frac{\lambda}{2}\|\mathbf{x}\|^2.$$

By an immediate corollary of Lemma 1, $\tilde{F}(\cdot)$ is minimized at $(q, q^2, \ldots, q^T, 0, \ldots, 0)$, where $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, and therefore $F(\mathbf{w}) = \tilde{F}(V\mathbf{w})$ is minimized at $V^\top(q, q^2, \ldots, q^T, 0, \ldots, 0)$, which equals $\sum_{i=1}^{T} q^i \mathbf{v}_i$ as required. □

Note that this lemma also allows us to bound the norm of $\mathbf{w}^\star = \arg\min F(\mathbf{w})$, since it implies that

$$\|\mathbf{w}^\star\| \;\leq\; \left\|\sum_{i=1}^{T} q^i \mathbf{v}_i\right\| + \sqrt{\frac{T\mu r^2}{16\lambda}},$$

and since $(a+b)^2 \leq 2a^2 + 2b^2$ and $q < 1$, we have

$$\begin{aligned}
\|\mathbf{w}^\star\|^2 \;&\leq\; 2\left\|\sum_{i=1}^{T} q^i \mathbf{v}_i\right\|^2 + \frac{T\mu r^2}{8\lambda} \;=\; 2\sum_{i=1}^{T} q^{2i} + \frac{T\mu r^2}{8\lambda}\\
&\leq\; 2\sum_{i=1}^{\infty} q^{2i} + \frac{T\mu r^2}{8\lambda} \;=\; \frac{2q^2}{1-q^2} + \frac{T\mu r^2}{8\lambda}\\
&\leq\; \frac{2}{1-q} + \frac{T\mu r^2}{8\lambda} = \sqrt{\kappa} + 1 + \frac{T\mu r^2}{8\lambda},
\end{aligned}$$

which is at most $\sqrt{\kappa} + 2 \leq 3\sqrt{\kappa}$, since we assume that c is sufficiently small so that $\frac{T\mu r^2}{8\lambda} \leq 1$, and that $\kappa = \mu/8\lambda \geq 1$.

The proof of the theorem follows by combining Lemma 5 and Lemma 6. Specifically, Lemma 5 (which states that $\langle \mathbf{w}_T, \mathbf{v}_T\rangle = 0$) and the fact that $\mathbf{v}_1, \ldots, \mathbf{v}_T$ are orthonormal tells us that

$$\begin{aligned}
\left\|\mathbf{w}_T - \sum_{i=1}^{T} q^i \mathbf{v}_i\right\|^2 \;&=\; \left\|\left(\mathbf{w}_T - \sum_{i=1}^{T-1} q^i \mathbf{v}_i\right) - q^T \mathbf{v}_T\right\|^2 \;=\; \left\|\mathbf{w}_T - \sum_{i=1}^{T-1} q^i \mathbf{v}_i\right\|^2 + \|q^T \mathbf{v}_T\|^2\\
&\geq\; \|q^T \mathbf{v}_T\|^2 \;=\; q^{2T},
\end{aligned}$$

and hence

$$\left\|\mathbf{w}_T - \sum_{i=1}^{T} q^i \mathbf{v}_i\right\| \;\geq q^T.$$

On the other hand, Lemma 6 states that

$$\left\|\mathbf{w}^\star - \sum_{i=1}^{T} q^i \mathbf{v}_i\right\| \;\leq\; \sqrt{\frac{T\mu r^2}{16\lambda}}.$$

Combining the last two displayed equations by the triangle inequality, we get that

$$\|\mathbf{w}_T - \mathbf{w}^\star\| \;\geq\; q^T - \sqrt{\frac{T\mu r^2}{16\lambda}}.$$

By the assumption that c is sufficiently small so that $\sqrt{\frac{T\mu r^2}{16\lambda}} \leq \frac{1}{2}q^T$, the left hand side is at least $\frac{1}{2}q^T$. Squaring both sides, we get

$$\|\mathbf{w}_T - \mathbf{w}^\star\|^2 \;\geq\; \frac{1}{4}q^{2T},$$

so by strong convexity of F,

$$F(\mathbf{w}_T) - F(\mathbf{w}^\star) \;\geq\; \frac{\lambda}{2}\|\mathbf{w}_T - \mathbf{w}^\star\|^2 \;\geq\; \frac{\lambda}{8}q^{2T}.$$

Plugging in the value of q, we get

$$F(\mathbf{w}_T) - F(\mathbf{w}^\star) \;\geq\; \frac{\lambda}{8}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T}.$$

On the other hand, we showed earlier that $\|\mathbf{w}^\star\|^2 \leq 3\sqrt{\kappa}$, so by smoothness, $F(\mathbf{0}) - F(\mathbf{w}^\star) \leq \frac{\mu}{2}\|\mathbf{w}^\star\|^2 \leq \frac{3\mu}{2}\sqrt{\kappa}$. Therefore,

$$\frac{F(\mathbf{w}_T) - F(\mathbf{w}^\star)}{F(\mathbf{0}) - F(\mathbf{w}^\star)} \;\geq\; \frac{\lambda}{12\mu\sqrt{\kappa}}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T}$$

To make the right-hand side less than ϵ, T must be such that

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T} \leq \frac{12\mu\sqrt{\kappa}\epsilon}{\lambda},$$

which is equivalent to

$$2T\cdot\log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right) \geq \log\left(\frac{\lambda}{12\mu\sqrt{\kappa}\epsilon}\right).$$

Since $\log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right) = \log\left(1+\frac{2}{\sqrt{\kappa}-1}\right) \leq \frac{2}{\sqrt{\kappa}-1}$, it follows that T must be such that

$$\frac{4T}{\sqrt{\kappa}-1} \;\geq\; \log\left(\frac{\lambda}{12\mu\sqrt{\kappa}\epsilon}\right).$$

Plugging in $\kappa = \mu/8\lambda$ and simplifying a bit, we get that

$$T \;\geq\; \frac{1}{4}\left(\sqrt{\frac{\mu}{8\lambda}}-1\right)\cdot\log\left(\frac{\sqrt{8}(\lambda/\mu)^{3/2}}{12\epsilon}\right),$$

from which the result follows.

A.3. Proof of Thm. 2

We will define a randomized choice of quadratic functions $f_1,\ldots,f_n$, and prove a lower bound on the expected optimization error of any algorithm (where the expectation is over both the algorithm and the randomized functions). This implies that for any algorithm, the same lower bound (in expectation over the algorithm only) holds for some deterministic choice of $f_1,\ldots,f_n$.

There will actually be two separate constructions, one leading to a lower bound of $\Omega(n)$, and one leading to a lower bound of $\Omega\left(\sqrt{\frac{n\mu}{\lambda}}\cdot\log\left(\frac{(\lambda/\mu)^{3/2}\sqrt{n}}{\epsilon}\right)\right)$. Choosing the construction which leads to the larger lower bound, the theorem follows.

A.3.1. An $\Omega(n)$ Lower Bound

Starting with the $\Omega(n)$ lower bound, let δ_i, where $i\in\{1,\ldots,n\}$, be chosen uniformly at random from $\{-1,+1\}$, and define

$$f_i(\mathbf{w}) = -\delta_i w_1 + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

Clearly, these are λ-smooth (and hence μ-smooth) functions, as well as λ-strongly convex. Also, the optimum of $F(\mathbf{w}) = \frac{\mu}{n}\sum_{i=1}^n f_i(\mathbf{w})$ equals $\mathbf{w}^\star = \left(\frac{1}{n\lambda}\sum_{i=1}^n \delta_i\right)\mathbf{e}_1$, where $\mathbf{e}_1$ is the first unit vector. As a result, $\|\mathbf{w}^\star\|^2 = \frac{1}{\lambda^2}\left(\frac{1}{n}\sum_{i=1}^n\delta_i\right)^2$, so by λ-smoothness of F

$$F(\mathbf{0})-F(\mathbf{w}^\star) \;\leq\; \frac{\lambda}{2}\|\mathbf{w}^\star\|^2 \;=\; \frac{1}{2\lambda}\left(\frac{1}{n}\sum_{i=1}^n\delta_i\right)^2.$$

Since δ_i are i.i.d., we have by Hoeffding's bound that with probability at least $3/4$, $\left|\frac{1}{n}\sum_{i=1}^n\delta_i\right|$ is at most $\sqrt{2\log(8/3)/n}\leq\sqrt{2/n}$. Plugging into the equation above, we get that with probability at least $3/4$,

$$F(\mathbf{0})-F(\mathbf{w}^\star) \;\leq\; \frac{1}{\lambda n}. \tag{9}$$

Turning to lower bound $F(\mathbf{w}_T)-F(\mathbf{w}^\star)$, we have by strong convexity that

$$\begin{aligned} F(\mathbf{w}_T)-F(\mathbf{w}^\star) \;&\geq\; \frac{\lambda}{2}\|\mathbf{w}_T-\mathbf{w}^\star\|^2 \;\geq\; \frac{\lambda}{2}(w_{T,1}-w_1^\star)^2 \\ &=\; \frac{1}{2\lambda}\left(\lambda w_{T,1}-\frac{1}{n}\sum_{i=1}^n\delta_i\right)^2. \end{aligned}$$

Now, if at most $\lfloor n/2 \rfloor$ indices $\{1,\ldots,n\}$ were queried by the algorithm, then the $\mathbf{w}_T$ returned by the algorithm must be independent of at least $\lceil n/2 \rceil$ random variables $\delta_{j_1},\ldots,\delta_{j_{\lceil n/2 \rceil}}$ (for some distinct indices $j_1, j_2, \ldots$ depending on the algorithm's behavior, but independent of the values of $\delta_{j_1},\ldots,\delta_{j_{\lceil n/2 \rceil}}$). Therefore, conditioned on $j_1,\ldots,j_{\lceil n/2 \rceil}$ and the values of $\delta_{j_1},\ldots,\delta_{j_{\lceil n/2 \rceil}}$, the expression above can be written as

$$\frac{1}{2\lambda}\left(\eta - \frac{1}{n}\sum_{i \notin \{j_1,\ldots,j_{\lceil n/2 \rceil}\}} \delta_i\right)^2 ,$$

where η is a fixed quantity independent of the values of δ_i for $i \notin \{j_1,\ldots,j_{\lceil n/2 \rceil}\}$. By a standard anti-concentration argument, with probability at least $3/4$, this expression will be at least $\frac{1}{2\lambda}\left(\frac{c'}{\sqrt{n}}\right)^2 = \frac{c'^2}{2\lambda n}$ for some universal positive $c' > 0$. Since this is true for any $j_1,\ldots,j_{\lceil n/2 \rceil}$ and $\delta_{j_1},\ldots,\delta_{j_{\lceil n/2 \rceil}}$, we get that with probability at least $3/4$ over $\delta_1,\ldots,\delta_n$,

$$F(\mathbf{w}_T) - F(\mathbf{w}^\star) \;\geq\; \frac{c'^2}{2\lambda n}.$$

Combining this with Eq. (9) using a union bound, we have that with probability at least $1/2$,

$$\frac{F(\mathbf{w}_T) - F(\mathbf{w}^\star)}{F(\mathbf{0}) - F(\mathbf{w}^\star)} \;\geq\; \frac{c'^2 \lambda n}{2\lambda n} \;=\; \frac{c'^2}{2}.$$

As a result, since the ratio above is always a non-negative quantity,

$$\mathbb{E}\left[\frac{F(\mathbf{w}_T) - F(\mathbf{w}^\star)}{F(\mathbf{0}) - F(\mathbf{w}^\star)}\right] \;\geq\; \frac{c'^2}{4}.$$

Using the assumption stated in the theorem (taking $c = c'^2/4$), we have that the right hand side cannot be smaller than ϵ, unless more than $\lfloor n/2 \rfloor = \Omega(n)$ oracle calls are made.

A.3.2. An $\Omega\left(\sqrt{\frac{n\mu}{\lambda}} \cdot \log\left(\frac{(\lambda/\mu)^{3/2}\sqrt{n}}{\epsilon}\right)\right)$ Lower Bound

We now turn to prove the $\Omega\left(\sqrt{\frac{n\mu}{\lambda}} \cdot \log\left(\frac{\lambda}{\epsilon}\right)\right)$ lower bound, using a different function construction: Let $j_1,\ldots,j_{d-1}$ be chosen uniformly and independently at random from $\{1,\ldots,n\}$, and define

$$f_i(\mathbf{w}) = \frac{\mu-\lambda}{8}\left(\sum_{l=1}^{d-1} \mathbf{1}_{j_l = i}(w_l - w_{l+1})^2 + \frac{1}{n}\left(w_1^2 + (a_\kappa - 1)w_d^2 - w_1\right)\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{10}$$

where $\mathbf{1}_A$ is the indicator of the event i. Note that these are all λ-strongly convex functions, as all terms in their definition are convex in $\mathbf{w}$, and there is an additional $\frac{\lambda}{2}\|\mathbf{w}\|^2$ term. Moreover, they are also μ-smooth: To see this, note that $\nabla^2 f_i(\mathbf{w}) \preceq \frac{(\mu-\lambda)}{4}A + \lambda I \preceq \mu I$, where $A \preceq 4I$ is as defined in the proof of Lemma 1.

The average function $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w})$ equals

$$F(\mathbf{w}) = \frac{\mu-\lambda}{8n}\left(w_1^2 + \sum_{i=1}^{d-1}(w_i - w_{i+1})^2 + (a_\kappa - 1)w_d^2 - w_1\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{11}$$

Therefore, by Lemma 1, the smoothness parameter of F is $(\mu - \lambda)/n + \lambda \leq \mu$, the global minimum $\mathbf{w}^\star$ of F equals $(q, q^2, \ldots, q^d)$, where $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ and

$$\kappa = \frac{\frac{\mu-\lambda}{n} + \lambda}{\lambda} = \frac{\frac{\mu}{\lambda} - 1}{n} + 1.$$

Note that since $q < 1$ and $\kappa \geq 1$, the squared norm of $\mathbf{w}^\star$ is at most

$$\sum_{i=1}^{d} q^{2i} \;\leq \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1-q^2} \leq \frac{1}{1-q} = \frac{\sqrt{\kappa}+1}{2} \;\leq\; \sqrt{\kappa}, \tag{12}$$

hence by smoothness,

$$F(\mathbf{0}) - F(\mathbf{w}^\star) \;\leq\; \frac{\mu}{2}\|\mathbf{w}^\star\|^2 \;\leq\; \frac{\mu}{2}\sqrt{\kappa}. \tag{13}$$

With these preliminaries out of the way, we now turn to compute a lower bound on the expected optimization error. The proof is based on arguing that $\mathbf{w}_T$ can only have a first few coordinates being non-zero. To see how this gives a lower bound, let $l_T \in \{1,\ldots,d\}$ be the largest index of a non-zero coordinate of $\mathbf{w}_T$ (or 0 if $\mathbf{w}_T = \mathbf{0}$). By definition of $\mathbf{w}^\star$, we have

$$\|\mathbf{w}_T - \mathbf{w}^\star\|^2 \geq \sum_{i=l_T+1}^{d} q^{2i} \geq g(l_T),$$

where

$$g(z) = \begin{cases} q^{2(z+1)} & z < d \\ 0 & z \geq d \end{cases}. \tag{14}$$

By strong convexity of F, this implies that

$$F(\mathbf{w}_T) - F(\mathbf{w}^\star) \;\geq\; \frac{\lambda}{2}\|\mathbf{w}_T - \mathbf{w}^\star\|^2 \;\geq\; \frac{\lambda}{2} g(l_T).$$

Finally, taking expectation over the randomness of $j_1,\ldots,j_{d-1}$ above (and over the internal randomness of the algorithm, if any), applying Lemma 2, and choosing the dimension $d = \lceil 2\mathbb{E}[l_T]\rceil$ (which we will later show to equal the value specified in the theorem), we have

$$\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}^\star)\right] \;\geq\; \frac{\lambda}{4} q^{4\mathbb{E}[l_T]+4} \;=\; \frac{\lambda}{4}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2\mathbb{E}[l_T]+2}.$$

Combined with Eq. (13), this gives

$$\mathbb{E}\left[\frac{F(\mathbf{w}_T) - F(\mathbf{w}^\star)}{F(\mathbf{0}) - F(\mathbf{w}^\star)}\right] \;\geq\; \frac{\lambda}{2\mu\sqrt{\kappa}}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2\mathbb{E}[l_T]+2}. \tag{15}$$

Thus, it remains to upper bound $\mathbb{E}[l_T]$.

To get a bound, we rely on the following key lemma (where $\mathbf{e}_i$ is the i-th unit vector, and recall that $\mathcal{W}_t$ defines the set of allowed query points $\mathbf{w}_t$, and $j_1,\ldots,j_d$ are the random indices used in constructing $f_1,\ldots,f_n$):

Lemma 7. *For all t, it holds that $\mathcal{W}_t \subseteq \text{span}\{\mathbf{e}_d, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_{\ell_t}\}$ for all t, where ℓ_t is defined recursively as follows: $\ell_1 = 1$, and ℓ_{t+1} equals the largest number in $\{1,\ldots,d-1\}$ such that $\{j_{\ell_t}, j_{\ell_t+1},\ldots,j_{\ell_{t+1}-1}\} \subseteq \{i_t, i_{t-1},\ldots,i_{\max\{1,t-\lfloor n/2\rfloor+1\}}\}$ (and $\ell_{t+1} = \ell_t$ if no such number exists).*

As will be seen later, ℓ_T (which is a random variable as a function of the random indices $j_1,\ldots,j_d$) upper-bounds the number of non-zero coordinates of $\mathbf{w}_T$, and therefore we can upper bound $\mathbb{E}[l_T]$ by $\mathbb{E}[\ell_T]$.

Proof. The proof is by induction over t. Since $\mathcal{W}_1 = \{\mathbf{0}\} \subseteq \text{span}(\mathbf{e}_d)$, the result trivially holds for $t = 1$. Now, suppose that $\mathcal{W}_t \subseteq \text{span}\{\mathbf{e}_d, \mathbf{e}_1, \ldots, \mathbf{e}_{\ell_t}\}$ for some t and ℓ_t. Note that in particular, this means that $\mathbf{w}_t$ is non-zero only in its first ℓ_t coordinates. By definition of f_i for any i,

$$\nabla f_i(\mathbf{w}) = \frac{\lambda n(\kappa-1)}{8}\left(2\sum_{l=1}^{d-1}\mathbf{1}_{j_l=i}(w_l - w_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}) + \frac{1}{n}\left(2w_1\mathbf{e}_1 + 2(a_\kappa - 1)w_d\mathbf{e}_d - \mathbf{e}_1\right)\right) + \lambda\mathbf{w}$$

$$\nabla^2 f_i(\mathbf{w}) = \frac{\lambda n(\kappa-1)}{8}\left(\sum_{l=1}^{d-1}\mathbf{1}_{j_l=i}(2E_{l,l} - E_{l+1,l} - E_{l,l+1}) + \frac{1}{n}\left(2E_{1,1} + 2(a_\kappa - 1)E_{d,d}\right)\right) + \lambda I,$$

where $E_{r,s}$ is the $d \times d$ which is all zeros, except for an entry of 1 in location (r,s). It is easily seen that these expressions imply the following:

- If $j_{\ell_t} \neq i_t$, then $\nabla f_{i_t}(\mathbf{w}_t) \in \text{span}\{\mathbf{e}_d, \mathbf{e}_1, \ldots, \mathbf{e}_{\ell_t}\}$, otherwise $\nabla f_{i_t}(\mathbf{w}_t) \in \text{span}\{\mathbf{e}_d, \mathbf{e}_1, \ldots, \mathbf{e}_{\ell_t+1}\}$.
- For any $\mathbf{w}$ and $l \in \{1, \ldots, d-1\}$, if $j_l \neq i$, then $\nabla^2 f_i(\mathbf{w})$ is block-diagonal, with a block in the first $l \times l$ entries. In other words, any entry (r, s) in the matrix, where $r \leq l$ and $s > l$ (or $r > l$ and $s \leq l$) is zero.
- As a result, if $j_l \notin \{i_t, i_{t-1}, \ldots, i_{\max\{1, t-\lfloor n/2 \rfloor+1\}}\}$, then $\sum_{\tau=\max\{1,t-\lfloor n/2\rfloor+1\}}^{t} \alpha_\tau \nabla^2 f_{i_\tau}(\mathbf{w}_\tau)$, for arbitrary scalars τ, is block-diagonal with a block in the first $l \times l$ entries. The same clearly holds for any matrix with the same block-diagonal structure.

Together, these observations imply that the operations specified in Assumption 1 can lead to vectors outside $\text{span}\{\mathbf{e}_d, \mathbf{e}_1, \ldots, \mathbf{e}_{\ell_t}\}$, only if $j_{\ell_t} \in \{i_t, i_{t-1}, \ldots, i_{\max\{1,t-\lfloor n/2\rfloor+1\}}\}$. Moreover, these vectors must belong to $\text{span}\{\mathbf{e}_d, \mathbf{e}_1, \ldots, \mathbf{e}_{\ell_{t+1}}\}$, where ℓ_{t+1} is as specified in the lemma: By definition, $j_{\ell_{t+1}}$ is not in $\{i_t, i_{t-1}, \ldots, i_{\max\{1,t-\lfloor n/2\rfloor+1\}}\}$, and therefore all relevant Hessians have a block in the first $\ell_{t+1} \times \ell_{t+1}$ entries, hence it is impossible to create a vector with non-zero coordinates (using the operations of Assumption 1) beyond the first ℓ_{t+1}. □

Since $\mathbf{w}_T \subseteq \mathcal{W}_T$, the lemma above implies that $\mathbb{E}[l_T]$ from Eq. (15) (where l_T is the largest index of a non-zero coordinate of $\mathbf{w}_T$) can be upper-bounded by $\mathbb{E}[\ell_T]$, where the expectation is over the random draw of the indices $j_1, \ldots, j_{d-1}$. This can be bounded using the following lemma:

Lemma 8. *It holds that* $\mathbb{E}[\ell_T] \leq 1 + \frac{2(T-1)}{n}$.

Proof. By definition of ℓ_t and linearity of expectation, we have

$$\mathbb{E}[\ell_T] = \mathbb{E}\left[\sum_{t=1}^{T-1} (\ell_{t+1} - \ell_t)\right] + \ell_1 = \sum_{t=1}^{T-1} \mathbb{E}[\ell_{t+1} - \ell_t] + 1. \tag{16}$$

Let us consider any particular term in the sum above. Since $\ell_{t+1} - \ell_t$ is a non-negative integer, we have

$$E[\ell_{t+1} - \ell_t] \;=\; \Pr(\ell_{t+1} > \ell_t) \cdot \mathbb{E}\left[\ell_{t+1} - \ell_t \mid \ell_{t+1} > \ell_t\right].$$

By definition of ℓ_t, the event $\ell_{t+1} > \ell_t$ can occur only if $j_{\ell_t} \notin \{i_{t-1}, i_{t-2}, \ldots, i_{\max\{1,t-\lfloor n/2\rfloor\}}\}$, yet $j_{\ell_t} \in \{i_t, i_{t-1}, \ldots, i_{\max\{1,t-\lfloor n/2\rfloor+1\}}\}$. This is equivalent to $j_{\ell_t} = i_t$ (that is, in iteration t we happened to choose the index j_{ℓ_t} of the unique individual function, which contains the block linking coordinate ℓ_t and $\ell_t + 1$, hence allowing us to "advance" and have more non-zero coordinates). But since the algorithm is oblivious, i_t is fixed whereas j_{ℓ_t} is chosen uniformly at random, hence the probability of this event is $1/n$. Therefore, $\Pr(\ell_{t+1} > \ell_t) \leq 1/n$. Turning to the conditional expectation of $\ell_{t+1} - \ell_t$ above, it equals the expected number of indices $j_{\ell_t}, j_{\ell_t+1}, \ldots$ belonging to $\{i_t, i_{t-1}, \ldots, i_{\max\{1,t-\lfloor n/2\rfloor+1\}}\}$, conditioned on j_{ℓ_t} belonging to that set. But since the i indices are fixed and the j indices are chosen uniformly at random, this equals one plus the expected number of times where a randomly drawn $j \in \{1, \ldots, n\}$ belongs to $\{i_t, i_{t-1}, \ldots, i_{t-\lfloor n/2\rfloor+1}\}$. Since this set contains at most $\lfloor n/2 \rfloor$ distinct elements in $\{1, \ldots, n\}$, this is equivalent to (one plus) the expectation of a geometric random variable, where the success probability is at most $1/2$. By a standard derivation, this is at most $1 + \frac{1/2}{1-1/2} = 2$. Plugging into the displayed equation above, we get that

$$\mathbb{E}[\ell_{t+1} - \ell_t] \;\leq\; \frac{1}{n} \cdot 2 \;=\; \frac{2}{n},$$

and therefore the bound in Eq. (16) is at most $\frac{2(T-1)}{n} + 1$ as required. □

Plugging this bound into Eq. (15), we get

$$\mathbb{E}\left[\frac{F(\mathbf{w}_T) - F(\mathbf{w}^\star)}{F(\mathbf{0}) - F(\mathbf{w}^\star)}\right] \;\geq\; \frac{\lambda}{2\mu\sqrt{\kappa}} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{\frac{4(T-1)}{n}+4}.$$

To make the right-hand side less than ϵ, T must be such that

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{\frac{4(T-1)}{n}+4} \leq \frac{2\mu\sqrt{\kappa}\epsilon}{\lambda},$$

which is equivalent to

$$\left(\frac{4(T-1)}{n}+4\right)\log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)\geq\log\left(\frac{\lambda}{2\mu\sqrt{\kappa}\epsilon}\right).$$

Since $\log\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)=\log\left(1+\frac{2}{\sqrt{\kappa}-1}\right)\leq\frac{2}{\sqrt{\kappa-1}}$ (see, e.g., Lemma 12 in (Arjevani and Shamir, 2016a)), it follows that T must be such that

$$\left(\frac{4(T-1)}{n}+4\right)\frac{2}{\sqrt{\kappa-1}}\ \geq\ \log\left(\frac{\lambda}{2\mu\sqrt{\kappa}\epsilon}\right).$$

Plugging in $\kappa=\frac{\frac{\mu}{\lambda}-1}{n}+1$, we get that

$$T\ \geq\ 1+\frac{n}{4}\left(\frac{\sqrt{\frac{\frac{\mu}{\lambda}-1}{n}}}{2}\cdot\log\left(\frac{\lambda}{2\mu\epsilon\sqrt{\frac{\frac{\mu}{\lambda}-1}{n}+1}}\right)-4\right).$$

Using asymptotic notation the right-hand side equals

$$\Omega\left(\sqrt{n(\mu/\lambda-1)}\log\left(\frac{(\lambda/\mu)^{3/2}\sqrt{n}}{\epsilon}\right)\right).$$

as required. The bound on the dimension d follows from the fact that we chose it to be $\mathcal{O}(\mathbb{E}[l_T])=\mathcal{O}(1+T/n)$, and to make the lower bound valid it is enough to pick some $T=\mathcal{O}\left(\sqrt{\frac{n\mu}{\lambda}}\cdot\log\left(\frac{(\lambda/\mu)^{3/2}\sqrt{n}}{\epsilon}\right)\right)$.

A.4. Proof of Thm. 3

Recall that the proof of Thm. 2 essentially shows that for any (possibly stochastic) index-oblivious optimization algorithm there exists some 'bad' assignment of the $d-1$ blocks $j_1,\ldots,j_{d-1}$ whose corresponding $f_i:\mathbb{R}^d\to\mathbb{R}$ (see Eq. (10)) form a functions which is hard-to-optimize. When considering non-oblivious (i.e., adaptive) algorithms this construction fails as soon as the algorithm obtains the Hessians of all the individual functions (potentially, after n second-order oracle queries). Indeed, knowing the Hessians of f_i, one can devise an index-schedule which gains at least one coordinate at every iteration, as opposed to $1/n$ on average for the oblivious case. Thus, in order to tackle the non-oblivious case, we form a function over some D-dimensional space which 'contains' all the n^{d-1} sub-problems at one and the same time (clearly, to carry out our plans we must pick D which grows exponentially fast with d, the dimension of the sub-problems). This way, any index-schedule, oblivious or adaptive, must 'fit' all the n^{d-1} sub-problems well, and as such, bound to a certain convergence rate which we analyze below.

Denote $[n]=\{1,\ldots,n\}$, set $D=n^{d-1}d$ and define for any $\mathbf{j}\in[n]^{d-1}$ the following,

$$f_i^{\mathbf{j}}:\mathbb{R}^d\to\mathbb{R},\qquad \mathbf{w}\mapsto\frac{\mu-\lambda}{8}\left(\sum_{l=1}^{d-1}\mathbf{1}_{j_l=i}(w_l-w_{l+1})^2+\frac{1}{n}\left(w_1^2+(a_\kappa-1)w_d^2-w_1\right)\right)+\frac{\lambda}{2}\|\mathbf{w}\|^2,$$

$$Q^{\mathbf{j}}:\mathbb{R}^D\to\mathbb{R}^d,\qquad \mathbf{u}\mapsto\sum_{l=1}^{d}\mathbf{u}^\top\mathbf{e}_{\#\mathbf{j}d+l}$$

where $\#\mathbf{j}$ enumerates the n^{d-1} tuples $[n]^{d-1}$ from 0 to $n^{d-1}-1$. Note that $f_i^{\mathbf{j}}$ are defined exactly as in Eq. (10), only here we make the dependence on $\mathbf{j}$ explicit. The individual functions are defined as follows:

$$f_i(\mathbf{u})=\sum_{\mathbf{j}\in[n]^{d-1}}f_i^{\mathbf{j}}(Q^{\mathbf{j}}\mathbf{u}).$$

Note that,

$$\nabla^2 f_i(\mathbf{u})=\sum_{\mathbf{j}\in[n]^{d-1}}(Q^{\mathbf{j}})^\top\nabla^2 f_i^{\mathbf{j}}(Q^{\mathbf{j}}\mathbf{u})Q^{\mathbf{j}}.$$

Since $\nabla^2 f_i$ are block-diagonal, we have $\Lambda(\nabla^2 f_i) = \bigcup_{\mathbf{j}} \Lambda(\nabla^2 f_i^{\mathbf{j}})$, where $\Lambda(\cdot)$ denotes the spectrum of a given matrix. Thus, since $f_i^{\mathbf{j}}$ are μ-smooth and λ-strongly convex (see proof of Thm. 2), we see that f_i is also μ-smooth and λ-strongly convex.

As for the average function $\Phi(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{u})$, it is easily verified that for any fixed $\mathbf{j} \in [n]^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^n f_i^{\mathbf{j}}(Q^{\mathbf{j}}\mathbf{u}) = F(Q^{\mathbf{j}}\mathbf{u}),$$

where F is as defined in Eq. (11). Thus,

$$\Phi(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{j}\in[n]^{d-1}} f_i^{\mathbf{j}}(Q^{\mathbf{j}}\mathbf{u}) = \sum_{\mathbf{j}\in[n]^{d-1}} F(Q^{\mathbf{j}}\mathbf{u}).$$

To compute the minimizer of Φ, we compute the first-order derivative:

$$\begin{aligned}\nabla\Phi(\mathbf{u}) &= \nabla\left(\sum_{\mathbf{j}\in[n]^{d-1}} F(Q^{\mathbf{j}}\mathbf{u})\right)\\ &= \sum_{\mathbf{j}\in[n]^{d-1}} \nabla\Big(F(Q^{\mathbf{j}}\mathbf{u})\Big)\\ &= \sum_{\mathbf{j}\in[n]^{d-1}} (Q^{\mathbf{j}})^\top \nabla F(Q^{\mathbf{j}}\mathbf{u}).\end{aligned}$$

Thus, by setting $\mathbf{u}^* = \sum_{\mathbf{j}} (Q^{\mathbf{j}})^\top \mathbf{w}^*$, where $\mathbf{w}^*$ is the minimizer of F as in Lemma 1, we get

$$\nabla\Phi(\mathbf{u}^*) = \sum_{\mathbf{j}\in[n]^{d-1}} (Q^{\mathbf{j}})^\top \nabla F\Big(Q^{\mathbf{j}}\sum_{\mathbf{j}}(Q^{\mathbf{j}})^\top \mathbf{u}^*\Big) = \sum_{\mathbf{j}\in[n]^{d-1}} (Q^{\mathbf{j}})^\top \nabla F(\mathbf{w}^*) = 0.$$

Note that, by Eq. (13), $\|\mathbf{u}^*\|^2 = n^{d-1}\|\mathbf{w}^*\|^2 \le n^{d-1}\sqrt{\kappa}$. Hence, by smoothness,

$$\Phi(\mathbf{0}) - \Phi(\mathbf{u}^\star) \;\le\; \frac{\mu}{2}\|\mathbf{u}^\star\|^2 \;\le\; \frac{\mu}{2}n^{d-1}\sqrt{\kappa}. \tag{17}$$

To derive the analytical properties of Φ, we compute the second derivative:

$$\begin{aligned}\nabla^2\Phi(\mathbf{u}) &= \sum_{\mathbf{j}\in[n]^{d-1}} \nabla((Q^{\mathbf{j}})^\top \nabla F(Q^{\mathbf{j}}\mathbf{u}))\\ &= \sum_{\mathbf{j}\in[n]^{d-1}} (Q^{\mathbf{j}})^\top \nabla(\nabla F(Q^{\mathbf{j}}\mathbf{u}))\\ &= \sum_{\mathbf{j}\in[n]^{d-1}} (Q^{\mathbf{j}})^\top \nabla^2 F(Q^{\mathbf{j}}\mathbf{u})Q^{\mathbf{j}}.\end{aligned}$$

Since $\nabla^2\Phi$ is a block-diagonal matrix, we have $\Lambda(\nabla^2\Phi) = \bigcup_{\mathbf{j}} \Lambda(\nabla^2 F) = \Lambda(\nabla^2 F)$. Thus, by Lemma 1, it follows that Φ is $((\mu-\lambda)/n + \lambda)$-smooth and λ-strongly convex.

With these preliminaries out of the way, we now turn to compute a lower bound on the expected optimization error. The proof follows by arguing that $\mathbf{u}_T$ can only have a first few coordinates being non-zero for each of the n^{d-1} sub-problems. To see how this gives a lower bound, let $l_T^{\mathbf{j}} \in \{1,\ldots,d\}$ be the largest index of a non-zero coordinate of $Q^{\mathbf{j}}\mathbf{u}_T$ (or 0 if

$Q^{\mathbf{j}}\mathbf{u}_T = \mathbf{0}$). By the definition of $\mathbf{u}^\star$ and by Eq. (12), we have

$$\begin{aligned}\|\mathbf{u}_T - \mathbf{u}^*\|^2 &= \|\sum_{\mathbf{j}}(Q^{\mathbf{j}})^\top Q^{\mathbf{j}}\mathbf{u}_T - \sum_{\mathbf{j}}(Q^{\mathbf{j}})^\top \mathbf{w}^*\|^2 \\ &= \|\sum_{\mathbf{j}}(Q^{\mathbf{j}})^\top (Q^{\mathbf{j}}\mathbf{u}_T - \mathbf{w}^*)\|^2 \\ &= \sum_{\mathbf{j}} \|Q^{\mathbf{j}}\mathbf{u}_T - \mathbf{w}^*\|^2 \\ &\geq \sum_{\mathbf{j}} g(l_T^{\mathbf{j}}),\end{aligned}$$

where g is defined in Eq. (14). By the strong convexity of F, this implies that

$$\Phi(\mathbf{u}_T) - \Phi(\mathbf{u}^\star) \;\geq\; \frac{\lambda}{2}\|\mathbf{u}_T - \mathbf{u}^\star\|^2 \;\geq\; \frac{\lambda}{2}\sum_{\mathbf{j}} g(l_T^{\mathbf{j}}).$$

We now proceed along the same lines as in the proof of Thm. 2. First, to upper bound $l_T^{\mathbf{j}}$ (note that, g is monotonically decreasing), we use the following generalized version of Lemma 7 (whose proof is a straightforward adaptation of the proof of Lemma 7):

Lemma 9. *Under Assumption 1, for all t, it holds that*

$$\mathcal{U}_t \subseteq \mathit{span}\left\{\bigcup_{\mathbf{j}\in[n]^{d-1}} \{\mathbf{e}_{\#\mathbf{j}d+d}, \mathbf{e}_{\#\mathbf{j}d+1}, \mathbf{e}_{\#\mathbf{j}d+2}, \mathbf{e}_{\#\mathbf{j}d+3}, \ldots, \mathbf{e}_{\#\mathbf{j}d+\ell_t^{\mathbf{j}}}\}\right\}$$

for all t, where $\ell_t^{\mathbf{j}}$ is defined recursively as follows: $\ell_1^{\mathbf{j}} = 1$, and $\ell_{t+1}^{\mathbf{j}}$ equals the largest number in $\{1, \ldots, d-1\}$ such that $\{j_{\ell_t^{\mathbf{j}}}, j_{\ell_t^{\mathbf{j}}+1}, \ldots, j_{\ell_{t+1}^{\mathbf{j}}-1}\} \subseteq \{i_t, i_{t-1}, \ldots, i_{\max\{1, t-\lfloor n/2\rfloor+1\}}\}$ (and $\ell_{t+1}^{\mathbf{j}} = \ell_t^{\mathbf{j}}$ if no such number exists).

As in the proof of Thm. 2, $\ell_T^{\mathbf{j}}$ bound $l_T^{\mathbf{j}}$ from above (for any given choice of $i_1, \ldots, i_T$), and since d is chosen so that

$$\frac{1}{n^{d-1}}\sum_{\mathbf{j}} \ell_T^{\mathbf{j}} \leq \frac{d}{2}, \tag{18}$$

we may take expectation over the internal randomness of the algorithm (if any), and combine it with (17) and Lemma 11 and Lemma 10 below to get

$$\begin{aligned}\mathbb{E}\left[\frac{\Phi(\mathbf{u}_T) - \Phi(\mathbf{u}^\star)}{\Phi(\mathbf{0}) - \Phi(\mathbf{u}^\star)}\right] &\geq \mathbb{E}\left[\frac{\lambda}{\mu\sqrt{\kappa}n^{d-1}}\sum_{\mathbf{j}} g(l_T^{\mathbf{j}})\right] \geq \mathbb{E}\left[\frac{\lambda}{\mu\sqrt{\kappa}n^{d-1}}\sum_{\mathbf{j}} g(\ell_T^{\mathbf{j}})\right] \\ &\geq \mathbb{E}\left[\frac{\lambda}{2\mu\sqrt{\kappa}} g\left(\frac{1}{n^{d-1}}\sum_{\mathbf{j}} \ell_T^{\mathbf{j}}\right)\right] \geq \frac{\lambda}{2\mu\sqrt{\kappa}}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{\frac{4(T-1)}{n}+4}.\end{aligned}$$

Following the same derivation as in the proof of Thm. 2, we get that T must be of order of

$$\Omega\left(\sqrt{n(\mu/\lambda - 1)}\log\left(\frac{(\lambda/\mu)^{3/2}\sqrt{n}}{\epsilon}\right)\right),$$

as required. The bound on d follows from the fact that we chose it to satisfy Inequality (18) through the following condition,

$$2\left(1 + \frac{2(T-1)}{n}\right) \leq d,$$

and to make the lower bound valid it is enough to pick some $T = \mathcal{O}\left(\sqrt{\frac{n\mu}{\lambda}} \cdot \log\left(\frac{(\lambda/\mu)^{3/2}\sqrt{n}}{\epsilon}\right)\right)$. Thus, we have that d is $\tilde{\mathcal{O}}(1 + \sqrt{\mu/\lambda n})$, implying $D = n^{d-1}d = n^{\tilde{\mathcal{O}}\left(1+\sqrt{\mu/\lambda n}\right)}$.

Lemma 10. *For any fixed sequence* $\mathbf{i} := i_1, \ldots, i_T \in [n]$ *of individual functions chosen during a particular execution of an optimization algorithm which satisfies Assumption 2, it holds that,*

$$\frac{1}{n^{d-1}} \sum_{\mathbf{j}} \ell_T^{\mathbf{j}} \leq 1 + \frac{2(T-1)}{n}.$$

Proof. By Lemma 9, $\ell_{t+1}^{\mathbf{j}}$ depends only on j_p for $\ell_t^{\mathbf{j}} \leq p \leq \ell_{t+1}^{\mathbf{j}}$. Thus, we may define

$$A_s = \left|\left\{(j_1, \ldots, j_s) \mid \ell_t^{(j_1,\ldots,j_s,*)} = s,\ \ell_{t+1}^{(j_1,\ldots,j_s,*)} > s\right\}\right|, \quad s \in [d],$$
$$B_s = \left|\left\{(j_1, \ldots, j_s) \mid \ell_t^{(j_1,\ldots,j_s,*)} = s,\ \ell_{t+1}^{(j_1,\ldots,j_s,*)} = s\right\}\right|, \quad s \in [d].$$

Intuitively, A_s and B_s count how many tuples $(j_1, \ldots, j_s)$, under a given choice of $i_1, \ldots, i_T$, allow at most s non-zero coordinates after t iterations, with one major difference: in A_s we want to allow the algorithm to make a progress after $t+1$ iterations (equivalently, $j_s = i_t$), whereas in B_s we want the algorithm to have the same number of s non-zero coordinates after $t+1$ (equivalently, $j_s \neq i_t$). One can easily verify the following:

$$\sum_{s=1}^{d} (A_s + B_s) n^{d-s-1} = n^{d-1},$$
$$B_s = (n-1)A_s.$$

The first equality may be obtained by splitting the space of all $[n]^{d-1}$ tuples into a group of disjoint sets characterized by the maximal number of non-zero coordinates the algorithm may gain by the t iteration. The second equality is a simple consequence of the way j_s is being constrained by A_s and B_s. This yields,

$$\sum_{s=1}^{d} A^s n^{-s} = n^{-1}. \tag{19}$$

Denoting $\mathcal{I} := \{i_t, i_{t-1}, \ldots, i_{\max\{t-\lfloor n/2\rfloor+1,1\}}\}$, we get that for any $1 \leq s \leq d-1$ and $1 \leq k \leq d-s$,

$$\begin{aligned}
&\left|\left\{\mathbf{j} \mid \ell_t^{\mathbf{j}} = s,\ \ell_{t+1}^{\mathbf{j}} = s+k\right\}\right| \\
&\quad = \left|\left\{(j_1, \ldots, j_{s-1}) \mid \ell_t^{(j_1,\ldots,j_{s-1},i_t,*)} = s\right\}\right| \cdot \left|\left\{(j_{s+1}, \ldots, j_{s+k}) \mid j_{s+1}, \ldots, j_{s+k-1} \in \mathcal{I},\ j_{s+k} \notin \mathcal{I}\right\}\right| \cdot n^{d-s-k-1} \\
&\quad = A_s |\mathcal{I}|^{k-1}(n - |\mathcal{I}|) n^{d-s-k-1}
\end{aligned}$$

This allows us to bound from above the average $\ell_{t+1}^{\mathbf{j}} - \ell_t^{\mathbf{j}}$ over $\mathbf{j}$ as follows,

$$\begin{aligned}
\frac{1}{n^{d-1}} \sum_{\mathbf{j}} (\ell_{t+1}^{\mathbf{j}} - \ell_t^{\mathbf{j}}) &= \frac{1}{n^{d-1}} \sum_{s=1}^{d-1} \sum_{k=1}^{d-s} \left|\left\{\mathbf{j} \mid \ell_t^{\mathbf{j}} = s,\ \ell_{t+1}^{\mathbf{j}} = s+k\right\}\right| k \\
&= \frac{1}{n^{d-1}} \sum_{s=1}^{d-1} \sum_{k=1}^{d-s} A_s |\mathcal{I}|^{k-1}(n - |\mathcal{I}|) n^{d-s-k-1} k \\
&= \sum_{s=1}^{d-1} A_s n^{-s} \sum_{k=1}^{d-s} |\mathcal{I}|^{k-1}(n - |\mathcal{I}|) n^{-k} k \\
&= \sum_{s=1}^{d-1} A_s n^{-s} \left(1 - \frac{|\mathcal{I}|}{n}\right) \sum_{k=1}^{d-s} \left(\frac{|\mathcal{I}|}{n}\right)^{k-1} k \\
&= \sum_{s=1}^{d-1} A_s n^{-s} \left(1 - \frac{|\mathcal{I}|}{n}\right) \sum_{k=1}^{\infty} \left(\frac{|\mathcal{I}|}{n}\right)^{k-1} k.
\end{aligned}$$

By standard manipulations of power series we have,

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \implies \sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}.$$

Combining this with Eq. (19) and the fact that $|\mathcal{I}| \leq n/2$ yields,

$$\frac{1}{n^{d-1}} \sum_{\mathbf{j}} (\ell_{t+1}^{\mathbf{j}} - \ell_t^{\mathbf{j}}) \leq \sum_{s=1}^{d-1} A_s n^{-s} \left(1 - \frac{|\mathcal{I}|}{n}\right)^{-1} \leq 2 \sum_{s=1}^{d-1} A_s n^{-s} \leq \frac{2}{n},$$

which, in turn, gives

$$\begin{aligned} \frac{1}{n^{d-1}} \sum_{\mathbf{j}} \ell_T^{\mathbf{j}} &= \frac{1}{n^{d-1}} \sum_{\mathbf{j}} \left(\sum_{t=1}^{T-1} (\ell_{t+1}^{\mathbf{j}} - \ell_t^{\mathbf{j}}) + \ell_1^{\mathbf{j}} \right) \\ &= \sum_{t=1}^{T-1} \frac{1}{n^{d-1}} \sum_{\mathbf{j}} (\ell_{t+1}^{\mathbf{j}} - \ell_t^{\mathbf{j}}) + \frac{1}{n^{d-1}} \sum_{\mathbf{j}} \ell_1^{\mathbf{j}} \\ &\leq \frac{2(T-1)}{n} + 1. \end{aligned}$$

□

Lemma 11. *For some $q \in (0,1)$ and positive d, define*

$$g(z) = \begin{cases} q^{2(z+1)} & z < d \\ 0 & z \geq d \end{cases}.$$

Let $a_1, \ldots, a_p$ be a sequence of non-negative reals, such that

$$\frac{1}{p} \sum_{i=1}^{p} a_i \leq \frac{d}{2},$$

then

$$\frac{1}{p} \sum_{i}^{p} g(a_i) \geq \frac{1}{2} g\left(\frac{1}{p} \sum_{i=1}^{p} a_i \right).$$

Proof. Since $q \in (0,1)$, the function $z \mapsto q^z$ is convex for non-negative z. Therefore, by the definition of g and Jensen's inequality we have

$$\begin{aligned} \frac{1}{p} \sum_{i}^{p} q(a_i) &= \frac{|\{i : a_i < d\}|}{p} \frac{1}{|\{i : a_i < d\}|} \sum_{\{i : a_i < d\}} g(a_i) \\ &\geq \frac{|\{i : a_i < d\}|}{p} g\left(\frac{1}{|\{i : a_i < d\}|} \sum_{\{i : a_i < d\}} a_i \right). \end{aligned}$$

Note that,

$$\frac{d}{2} \geq \frac{1}{p} \sum_{i=1}^{p} a_i = \frac{1}{p} \sum_{\{i : a_i < d\}} a_i + \frac{1}{p} \sum_{\{i : a_i \geq d\}} a_i \geq \frac{d}{p} |\{i | a_i \geq d\}| \implies \frac{|\{i | a_i < d\}|}{p} \geq \frac{1}{2}.$$

Therefore, together with the fact that g decreases monotonically and that

$$\frac{1}{|\{i : a_i < d\}|} \sum_{\{i : a_i < d\}} a_i \leq \frac{1}{p} \sum_{i=1}^{p} a_i,$$

we get

$$\frac{1}{p} \sum_{i}^{p} q(a_i) \geq \frac{1}{2} g \left(\frac{1}{p} \sum_{i=1}^{p} a_i \right).$$

□