## A. Optimization of a strongly convex smooth functions

The most accessible derivation of this classic lower bound (Nesterov, 2004) relies on the simplifying assumption that the successive points $x_k$ lie in the span of the gradients previously returned by the oracle. This section provides a derivation of the lower bound that does not rely on this assumption and is critical for Theorem 1 where no such assumptions are made.

This section considers algorithms that produces an approximate solution of the optimization problem

$$x_f^* \;=\; \arg\min_{x\in\ell_2} \; f(x) \;=\; \frac{\mu}{2}\|x\|^2 + g(x) \qquad \text{where} \;\; f(x) \in \mathcal{S}^{\mu,L}(\ell_2)\;. \tag{9}$$

using, as sole knowledge of function $f$, an oracle that returns the value $f(x)$ and the gradient $f'(x)$ on points successively determined by the algorithm. Note that this writing of $f$ is without loss of generality, since any $\mu$-strongly convex function can be written in the form (9) where $g$ is convex.

**Remark 2** *We could equivalently consider an oracle that reveals $g(x_k)$ and $g'(x_k)$ instead of $f(x_k)$ and $f'(x_k)$ because these quantities can be computed from each other (since $\mu$ is known.)*

At a high-level, our proof will have the following structure. We will first establish that any algorithm for solving the minimization problem (9) for all $f \in \mathcal{S}^{\mu,L}(\ell_2)$ will be forced to play the point $x_K$ in the span of the previous iterates and gradients. This essentially shows that the restriction made by Nesterov is not too serious. The second part of the proof constructs a resisting oracle for such algorithms whose final query point falls within the span of the previous responses. Combining these ingredients, we obtain the desired lower bound.

### A.1. Restriction of final solution to span

Consider an algorithm that calls the oracle on $K > 1$ successive points $x_0, \ldots, x_{K-1}$. The first part of the proof describes how to pick the best possible $x_K$ on the basis of the oracle answers and the algorithm's queries.

**Definition 6** *For any $\gamma \geq 0$, let $\mathcal{S}_\gamma^{\mu,L}(\ell_2)$ be the set of all functions $f \in \mathcal{S}^{\mu,L}(\ell_2)$ that reach their minimum in a point $x_f^*$ such that $\|x_f^*\| = \gamma$.*

**Definition 7** *Let $\mathcal{G}_\gamma^f \subset \mathcal{S}_\gamma^{\mu,L}(\ell_2)$ be the set of the functions of $\mathcal{S}_\gamma^{\mu,L}(\ell_2)$ whose values and gradients coincide with those of $f$ on points $x_0 \ldots x_{K-1}$. Let $H_\gamma^f \in \ell_2$ be the set of their minima.*

When the function $f$ is clear from the context, we will drop the superscript for brevity. Since all functions in $\mathcal{G}_\gamma$ are compatible with the values returned by the oracle, running our algorithm on any of them would perform the same calls to the oracle and obtain the same answers. Therefore, in order to offer the best guarantee on $\|x_K - x_f^*\|^2$ without further knowledge of the function $f(x)$, our algorithm must choose $x_K$ to be the center of the smallest ball containing $H_\gamma$.

**Definition 8** *Let $P = \mathrm{Span}\{x_0 \ldots x_{K-1}, f'(x_0) \ldots f'(x_{K-1})\}$. Let $\Pi_P(x)$ be the orthogonal projection of point $x$ on $P$ and let $M_p(x) = 2\Pi_P(x) - x$ be its mirror image with respect to $P$.*

Stated differently, we know that any point $x$ can be decomposed into $\Pi_P(x)$ and $\Pi_{P^\perp}(x)$ such that $x = \Pi_P(x) + \Pi_{P^\perp}(x)$. Then the above definition yields $M_P(x) = \Pi_P(x) - \Pi_{P^\perp}(x)$, which is the natural reflection of $x$ with respect to the subspace $P$.

**Proposition 3** *The set $H_\gamma$ is symmetric with respect to $P$.*

**Proof** Consider an arbitrary point $x_h^* \in H_\gamma$ which minimizes a function $h \in \mathcal{G}_\gamma$. Since function $x \mapsto h(M_P(x))$ also belongs to $\mathcal{G}_\gamma$, its minimum $M_p(x_h^*)$ also belongs to $H_\gamma$. ■

**Corollary 2** *The center of the smallest ball enclosing $H_\gamma$ belongs to $P$.*

We are now in a position to present the main ingredient of our proof that allows us to state a more general result than Nesterov. In particular, we demonstrate that the assumption made by Nesterov about the iterates lying in the span of previous gradients can be made almost without loss of generality. The key distinction is that we can only make it on the step $K$, where the algorithm is constrained to produce a good answer, while Nesterov assumes it on all iterates, somewhat restricting the class of admissible algorithms.

**Lemma 3** *For any $\gamma > 0$ and any algorithm $\mathsf{A}$ that performs $K \geq 1$ calls of the oracle and produces an approximate solution $x_K^{\mathsf{A}}(f)$ of problem (9), there is an algorithm $\mathsf{B}$ that performs $K$ calls of the oracle and produces an approximate solution $x_K^{\mathsf{B}}(f) \in \mathrm{Span}\{x_0 \dots x_{K-1}, f'(x_0) \dots f'(x_{K-1})\}$ for all $f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)$ such that*

$$\sup_{f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)} \|x_K^{\mathsf{B}} - x_f^*\|^2 \;\leq\; \sup_{f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)} \|x_K^{\mathsf{A}} - x_f^*\|^2 \;. \tag{10}$$

**Proof** Consider an algorithm $\mathsf{B}$ that first runs algorithm $\mathsf{A}$ and then returns the center of the smallest ball enclosing $H_\gamma^f$ as $x_K^{\mathsf{B}}(f)$. Corollary 2 ensures that $x_K^{\mathsf{B}}(f)$ belongs to the posited span. This choice of $x_K^{\mathsf{B}}(f)$ also ensures that $\sup_{\bar{x} \in H_\gamma^f} \|x_K^{\mathsf{B}}(f) - \bar{x}\| \leq \sup_{\bar{x} \in H_\gamma^f} \|x_K^{\mathsf{A}}(f) - \bar{x}\|$. Equivalently, $\sup_{g \in G_\gamma^f} \|x_K^{\mathsf{B}}(g) - x_g^*\| \leq \sup_{g \in G_\gamma^f} \|x_K^{\mathsf{A}}(g) - x_g^*\|$, where we use the fact that $x_K^{\mathsf{B}}(g) = x_K^{\mathsf{B}}(f)$ and $x_K^{\mathsf{A}}(g) = x_K^{\mathsf{A}}(f)$ because function $g \in G_\gamma^f$ coincides with $f$ on $x_0 \dots x_{K-1}$. Therefore,

$$\sup_{f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)} \; \sup_{g \in G_\gamma^f} \|x_K^{\mathsf{B}}(g) - x_g^*\| \;\leq\; \sup_{f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)} \; \sup_{g \in G_\gamma^f} \|x_K^{\mathsf{A}}(g) - x_g^*\| \;.$$

This inequality implies (10) because $\mathcal{S}_\gamma^{\mu,L}(\ell_2) = \cup_{f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)} G_\gamma^f$. ■

Lemma 3 means that we can restrict the analysis to algorithms that pick their final estimate $x_K$ in the subspace $P$ that results from the execution of the algorithm. In order to establish a lower bound for such an algorithm, it is sufficient to construct a function $f_K$ whose minimum is located sufficiently far away from this subspace. We construct this function by running the algorithm against a *resisting oracle*, which is quite standard in these lower bound proofs. Each call to the resisting oracle picks a new objective function $f_k$ among all the $\mathcal{S}^{\mu,L}(\ell_2)$ functions that agree with the values and gradients returned by all previous calls to the oracle. This constraint ensures that the optimization algorithm would have reached the same state if it had been run against function $f_k$ instead of the resisting oracle.

### A.2. Construction of a resisting oracle

We start by defining the basic structure of the function which will be used by our oracle to construct hard problem instances. This structure is identical to that used by Nesterov.

**Definition 9 (Nesterov)** *Fix $\rho > 0$ and let $N_{\mu,L}$ denote the function*

$$N_{\mu,L}(x) = \frac{L-\mu}{8}\left( (x[1])^2 + \sum_{i=1}^{\infty} (x[i+1] - x[i])^2 - 2\,\rho\, x[1] \right) + \frac{\mu}{2}\|x\|^2 \;.$$

**Proposition 4** *$N_{\mu,L} \in \mathcal{S}^{\mu,L}(\ell_2)$ and reaches its minimum in $x_N^* = (\rho\, q^i)_{i=1}^{\infty}$ with $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.*

**Proof** The assertions $\mu I \preceq N''_{\mu,L} \preceq LI$ and $N'_{\mu,L}(x_N^*) = 0$ follow from direct calculation, as shown in Nesterov (2004, p. 67). ■

**Remark 3** *We can arbitrarily choose the value of $\|x_N^*\|$ by appropriately selecting $\rho$.*

We also need some other properties of the function, which are also present in Nesterov's analysis.

**Proposition 5** *Let $[e_1, e_2, \dots]$ be the canonical basis of $\ell_2$ and let $R_k = \mathrm{Span}(e_1 \dots e_k)$.*

$$x \in R_k \;\Rightarrow\; N'_{\mu,L}(x) \in R_{k+1} \;.$$

**Proof** Through a direct calculation, it is easy to verify that

$$\frac{\partial}{\partial x[i]} N_{\mu,L}(x) = \begin{cases} \frac{L-\mu}{4}\left(x[1] + \left(x[1] - x[2] - 2\rho\right) + \mu x[1]\right. & \text{for } i = 1, \\ \frac{L-\mu}{4}\left(2x[i] - x[i+1] - x[i-1]\right) & \text{for } i > 1. \end{cases}$$

The statement directly follows from this. ■

We now recall our earlier definition of the matrix notation for orthonormal families in Definition 5. The resisting oracle we construct will apply the function $N_{\mu,L}$ to appropriately rotated versions of the point $x$, that is, it constructs functions of the form $N_{\mu,L}(S^\top x)$, where the orthonormal operators $S$ will be constructed appropriately to ensure that the optimal solution is sufficiently far away from the span of algorithm's queries and the oracle's responses. Before we define the oracle, we need to define the relevant orthogonalization operations.

**Definition 10 (Gram-Schmidt)** *Given a finite orthonormal family $S$ and a vector $v$, the Gram-Schmidt operator* $\text{Gram}(S,v)$ *augments the orthonormal family, ensuring that $v$ lies in the new span.*

$$\text{Gram}(S,v) = \begin{cases} S & \text{if } v \in \text{Span}(S) \\ \left[S, \frac{v - SS^\top v}{\|v - SS^\top v\|}\right] & \text{otherwise} \end{cases}$$

Our resisting oracle incrementally constructs orthonormal families $S_k$ and defines the functions $f_k(x)$ as the application of function $N_{\mu,L}$ to the coordinates of $x$ expressed an orthonormal basis of $\ell_2$ constructed by completing $S_k$.

**Definition 11 (Resisting oracle)** *Let $S_{-1}$ be an empty family of vectors. Each call $k = 0 \dots K-1$ of the resisting oracle performs the following computations and returns $y_k = f_k(x_k)$ and $g_k = f_k'(x_k)$.*

$$\begin{aligned} S_k &= \text{Gram}(\text{Gram}(S_{k-1}, x_k), v_k) \quad \text{for some } v_k \notin \text{Span}(S_{k-1}, x_k). && (11) \\ \bar{S}_k &= [S_k, \dots] && (12) \\ y_k &= f_k(x_k) \ = \ N_{\mu,L}(\bar{S}_k^\top x_k) && (13) \\ g_k &= f_k'(x_k) \ = \ \bar{S}_k N_{\mu,L}'(\bar{S}_k^\top x_k) && (14) \end{aligned}$$

Step (11) augments $S_{k-1}$ to ensure that $\text{Span}(S_k)$ contains both $x_k$ and an arbitrary additional vector. This construction ensures that $\dim(S_k) \leq 2k+2$. Step (12) nominally constructs an orthonormal basis $\bar{S}_k$ of $\ell_2$ by completing $S_k$. This is mostly for notational convenience because the additional basis vectors have no influence on the results produced by oracle. Step (13) computes the value of $y_k = f_k(x_k)$ by applying the function $N_{\mu,L}$ to the coordinates $\bar{S}_k^\top x_k$ of vector $x_k$ in basis $\bar{S}_k$. Since $x_k$ belongs to the span of the first $\dim(S_k)-1$ basis vectors, $\bar{S}_k^\top x_k \in R_{\dim(S_k)-1}$. Finally, step (14) computes the gradient $g_k = f_k'(x_k)$. Note that $g_k \in S_k$ because proposition 5 ensures that $N_{\mu,L}'(\bar{S}_k^\top x_k) \in R_{\dim(S_k)}$.

**Proposition 6** *The resisting oracle satisfies the following properties:*

*(a)* $\quad S_k = \text{Span}\{x_0 \dots x_{K-1}, f'(x_0) \dots f'(x_{K-1})\} \quad \dim(S_k) \leq 2k+2$ .

*(b)* $\quad \forall\, i < k \quad y_i = f_k(x_i) \quad g_i = f_k'(x_i)$ .

**Proof** Property (a) holds by construction (see discussion above). Property (b) holds because both $x_i$ and $g_i$ belong to $\text{Span}(S_i)$. Therefore $y_i = f_k(x_i)$ because $S_i^\top x_i = S_k^\top x_i$ and $g_i = f_k'(x_i)$ because $N_{\mu,L}'(\bar{S}_k^\top x_i) = N_{\mu,L}'(\bar{S}_i^\top x_i) \in R_{\dim(S_i)}$. ■

### A.3. Proof of Theorem 2

We now have all the ingredients to establish the main result of this appendix on the complexity of optimizing smooth and strongly convex functions. Given our work so far, we know that the solution $x_K$ lives in a $2K+2$ dimensional subspace of $\ell_2$. We also know that our resisting oracle constructs orthonormal operators $S_k$, so that the optimal solution of the function $f$ being constructed can be as far away as possible from this subspace. The next proposition, which almost establishes the theorem, essentially quantifies just how far the optimum lies from this span.

**Proposition 7** *The minimum $x^*$ of function $f_{K-1}$ satisfies*

$$\mathrm{dist}[\, x^*, \mathrm{Span}(S_{K-1})\,] \;\geq\; \|x^*\|\; q^{2K} \quad \text{with } \; q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \;\text{ and }\; \kappa = \frac{\mu}{L}\;.$$

**Proof** Any vector $x \in \mathrm{Span}(S_{K-1})$ is such that $(\bar{S}_{K-1})^\top x \;\in\; R_{\dim(S_{K-1})} \;\subset\; R_{2K}$.

Meanwhile, equation (13) and Proposition 4 imply that $(\bar{S}_{K-1})^\top x^* = (\rho\, q^i)_{i=1}^{\infty}$. Therefore

$$\|x^* - x\|^2 = \|(\bar{S}_{K-1})^\top x^* - (\bar{S}_{K-1})^\top x\|^2 \;\geq\; \sum_{i=2K+1}^{\infty} (\rho\, q^i - 0)^2 = q^{4K} \sum_{i=1}^{\infty} (\rho q^i)^2 = q^{4K} \|x^*\|^2\;. \qquad \blacksquare$$

Proposition 7 and Lemma 3 then directly yield the theorem. Indeed, the theorem is trivial when $K=0$. Consider otherwise an algorithm $\mathsf{B}$ known to pick its answer $x_K^{\mathsf{B}}$ in $\mathrm{Span}(x_0 \dots x_{K-1}, f'(x_0) \dots f'(x_K)$. For an appropriate choice of constant $\rho$, Proposition 7 constructs a function that satisfies the theorem. Finally, for any algorithm $\mathsf{A}$, lemma 3 implies that there is a function $f \in \mathcal{S}_\gamma^{\mu,L}(\ell_2)$ such that $\|x_f^* - x_K^{\mathsf{A}}\| \geq \|x_f^* - x_K^{\mathsf{B}}\|$ .

Lemma 2 then yields the corollary.

**Corollary 3** *In order to guarantee that $\|x^* - x_K\| \leq \varepsilon \|x^*\|$ for $\varepsilon < 1$, any first order black box algorithm for the optimization of $f \in \mathcal{S}^{\mu,L}(\ell_2)$ must perform at least $K = \Omega(\sqrt{\kappa - 1}\, \log(1/\varepsilon))$ calls to the oracle.*

Since this lower bound is established in the case where $\mathcal{X} = \ell_2$, it should be interpreted as the best *dimension independent* guarantee that can be offered by a first order black box algorithm for the optimization of $L$-smooth $\mu$-strongly convex functions.