On the Quality of First-Order Approximation of Functions with Hölder Continuous Gradient

Guillaume O. Berger[1] (iD) · **P.-A. Absil[1]** · **Raphaël M. Jungers[1]** · **Yurii Nesterov[1]**

Abstract

We show that Hölder continuity of the gradient is not only a sufficient condition, but also a necessary condition for the existence of a global upper bound on the error of the first-order Taylor approximation. We also relate this global upper bound to the Hölder constant of the gradient. This relation is expressed as an interval, depending on the Hölder constant, in which the error of the first-order Taylor approximation is guaranteed to be. We show that, for the Lipschitz continuous case, the interval cannot be reduced. An application to the norms of quadratic forms is proposed, which allows us to derive a novel characterization of Euclidean norms.

1 Introduction

The purpose of this paper is to investigate the relation between two properties of a real-valued function, which play crucial roles in optimization. The first property of interest is the Hölder continuity of the gradient, which means that the variation of the gradient of the function between two points is upper bounded by a power (with

✉ Guillaume O. Berger
guillaume.berger@uclouvain.be

P.-A. Absil
pa.absil@uclouvain.be

Raphaël M. Jungers
raphael.jungers@uclouvain.be

Yurii Nesterov
yurii.nesterov@uclouvain.be

[1] UCLouvain, Louvain-la-Neuve, Belgium

exponent smaller than or equal to one) of the distance between the two points; up to some multiplicative constant, called the Hölder constant of the gradient. The second property is that there exists a global upper bound on the error of the first-order Taylor approximation of the function. This global upper bound takes the form of a power (with exponent between one and two) of the distance between the point of interest and the reference point for the Taylor approximation; the power of the distance can be scaled by a multiplicative constant, called the approximation parameter of the function.

The class of functions with Hölder continuous gradient is ubiquitous in optimization. Indeed, the vast majority of first-order optimization methods (e.g., the gradient descent) requires Hölder continuity of the gradient to compute the optimal step size and to assert the convergence of the method to a stationary point [1–4]. It is well known that, if a function has Hölder continuous gradient, then there is a global upper bound on the error of the first-order Taylor approximation of the function at any point. Actually, it appears that the majority of the above-mentioned developments only make use of the Hölder continuity of the gradient as a convenient sufficient condition to ensure the global upper bound on the error of the first-order Taylor approximation.

For example, in the global convergence analysis found in [2, Section 3], the function is assumed to have Hölder continuous gradient, and the very first step is a lemma, stating that this property implies the existence of a global upper bound on the error of the first-order Taylor approximation of the function. In fact, the main complexity result [2, Corollary 2] can be obtained by assuming the global upper bound on the error of the first-order Taylor approximation, while disregarding the Hölder continuity of the gradient.

Another evidence of the prominent importance of the second property (global upper bound on the error of the first-order Taylor approximation) over the first one (Hölder continuity of the gradient) is that it is a generalization of the second property—and not of the first one—that is used as an assumption in [5] to generalize global complexity bounds for the minimization of functions defined on Riemannian manifolds.

As already mentioned, the first property is sufficient for the second one. The following questions then naturally arise. Is it also necessary? And, if it is, how can we relate the Hölder constant of the gradient to the approximation parameter arising in the second property? To the best of the authors' knowledge, an answer to this second question is available in the literature only for convex functions with Hölder exponent of the gradient equal to one (the particular case of Hölder continuity with exponent equal to one is generally referred to as Lipschitz continuity); see, e.g., [6, Theorem 2.1.5]. However, this second question is important for its implications in optimization: The upper bound on the error of the first-order Taylor approximation is used to compute the step size and to estimate the global rate of convergence of first-order optimization methods; the more accurately we know this bound, the better we can choose the step size and estimate the rate of convergence (see, e.g., Example 3.1).

In this paper, we provide an answer to the above questions. We show in Theorem 4.1 that a function satisfies the second property, if *and only if* it satisfies the first one. We also provide an interval, depending on the Hölder constant of the gradient, in which the approximation parameter of the function is guaranteed to be. We show that, in the case of functions with Lipschitz continuous gradient and quadratic upper bound on the first-order Taylor approximation, this interval is tight (see Example 5.1). We also

provide a more detailed analysis, when the domain of the function is endowed with a Euclidean norm, i.e., a norm induced by a scalar product. Finally, we apply these results to quadratic functions; see Sect. 6. This allows us to obtain a novel characterization of Euclidean norms.

The paper is organized as follows. In Sect. 2, we introduce notation and definitions. The questions we address are motivated in Sect. 3 with an example that demonstrates their importance in optimization. In Sect. 4, we prove the equivalence between the two properties described above, and explain the relation between the Hölder constant of the gradient and the approximation parameter. In Sect. 5, we particularize the results of Sect. 4 to the case of functions with Lipschitz continuous gradient, and we show that the bounds, derived in this specific case, are tight. The application to quadratic functions is presented in Sect. 6.

2 Notation and Preliminaries

In the sequel, E is a real finite-dimensional normed vector space with norm $\|\cdot\|$. The norm $\|\cdot\|$ is said to be *Euclidean*, if it is induced by a scalar product. Equivalently, $\|\cdot\|$ is Euclidean, if and only if every pair of vectors $u, v \in E$ satisfies the parallelogram law $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$ (Jordan–von Neumann theorem [7]). In this paper, we will consider both cases of Euclidean and non-Euclidean norms.

On $\mathbb{R}$, we denote the absolute value by $|\cdot|$. The *dual* of E is the space of linear maps from E to $\mathbb{R}$ and is denoted by E^*. We denote by $\langle \varphi, h \rangle$ the image of $h \in E$ by $\varphi \in E^*$. We endow E^* with the *dual norm*, $\|\cdot\|_*$, defined by

$$\|\varphi\|_* = \max \{\langle \varphi, h \rangle : h \in E, \ \|h\| = 1\}.$$

Let f be a function from E to $\mathbb{R}$. In the sequel, we will always assume that f is differentiable. For $x \in E$, we denote by $f'(x)$ the derivative—which we also term *gradient* as in [1]—of f at x. Note that $f'(x) \in E^*$. (Thus, it acts on vectors $h \in E$, and $\langle f'(x), h \rangle$ is sometimes called the directional derivative of f at x, in the direction h.)

The first property of interest is defined by:

Definition 2.1 (*ν-Hölder continuous gradient*) Let $0 \le \nu \le 1$. We say that f has *ν-Hölder continuous gradient*, if there exists an $M \ge 0$, such that, for every $x, y \in E$,

$$\|f'(x) - f'(y)\|_* \le M\|x - y\|^\nu. \tag{1}$$

In this case, it is easily seen that there exists a smallest M satisfying inequality (1) for every $x, y \in E$. This M is called the *ν-Hölder constant* of the gradient of f and is denoted by $M_f(\nu)$:

$$M_f(\nu) := \sup_{x \ne y} \frac{\|f'(x) - f'(y)\|_*}{\|x - y\|^\nu} < \infty.$$

Note that not every function has ν-Hölder continuous gradient for some $\nu \in [0, 1]$.[1] The next definition particularizes the former for $\nu = 1$.

Definition 2.2 (*Lipschitz continuous gradient*) We say that f has *Lipschitz continuous gradient*, if f has 1-Hölder continuous gradient. In this case, $M_f(1)$ is called the *Lipschitz constant* of the gradient of f.

As for the second property of interest in this paper, it reads as:

Definition 2.3 (ν-*approximable*) We will say that f is ν-*approximable*, if there exists an $L \geq 0$, such that, for every $x, y \in E$,

$$|f(y) - f(x) - \langle f'(x), y - x \rangle| \leq \frac{L}{1 + \nu} \|y - x\|^{1+\nu}. \tag{2}$$

The smallest such L will be called the ν-*approximation parameter* of f and will be denoted by $L_f(\nu)$:

$$L_f(\nu) := (1 + \nu) \sup_{x \neq y} \frac{|f(y) - f(x) - \langle f'(x), y - x \rangle|}{\|y - x\|^{1+\nu}} < \infty.$$

Remark 2.1 Clearly, Definitions 2.1–2.3 are invariant under translation and addition of linear functions. In other words, if $g(x) = f(x + a) + \langle \varphi, x \rangle + c$ for some $a \in E$, $\varphi \in E^*$ (i.e., φ is a linear map from E to $\mathbb{R}$), and $c \in \mathbb{R}$, then g is ν-approximable (resp. has ν-Hölder continuous gradient), if and only if f is ν-approximable (resp. has ν-Hölder continuous gradient); moreover, $M_g(\nu) = M_f(\nu)$ and $L_g(\nu) = L_f(\nu)$.

The following proposition is a classical result in optimization. It states that, if f has ν-Hölder continuous gradient for some $0 \leq \nu \leq 1$, then f is ν-approximable. Moreover, the ν-Hölder constant is an upper bound on the ν-approximation parameter:

Proposition 2.1 *Let f have ν-Hölder continuous gradient. Then, f is ν-approximable, and $L_f(\nu) \leq M_f(\nu)$.*

Proof See [2, Lemma 1], for instance.   □

3 Motivation

We present an example of optimization method, which uses inequality (2) to assert the convergence of the method to a stationary point. To compute the step size, the method requires an L satisfying (2). For a given such L, the global complexity of the method is proportional to $L^{1/\nu}$. It is thus *beneficial to choose L as small as possible*, ideally $L = L_f(\nu)$. The results in the next section provide an interval (or a specific value), depending on $M_f(\nu)$, in which $L_f(\nu)$ is guaranteed to be.

Example 3.1 (from [2]) Let f be a real-valued function, defined on $E = \mathbb{R}^n$ (with any norm $\|\cdot\|$), and satisfying (2) for some $L \geq 0$ and some[2] $0 < \nu \leq 1$ (i.e., f

---

[1] For example, $x \in \mathbb{R} \mapsto x^3$, or $x \in \mathbb{R} \mapsto x^2 \sin(1/x^2)$ (with continuous extension at 0).

[2] Note that the method requires $\nu > 0$. In fact, finding a descent direction for a non-smooth non-convex function is NP-hard [1], and thus, it is reasonable to ask that $\nu > 0$.

is ν-approximable with parameter $L_f(\nu) \le L$). Suppose that there is a lower bound $f_* \in \mathbb{R}$ on $f \colon f(x) \ge f_*$ for all $x \in E$. We use the gradient method to find a stationary point of f. The step size depends on L and ν, and on the norm of the gradient at each iteration point.

The method goes as follows. (For more details, we refer the reader to [2].) We start from a point $x_0 \in \mathbb{R}^n$. For $k = 0, 1, 2, \ldots$, let x_k be the iterate at step k. The norm of $f'(x_k)$ is denoted by $n_k = \|f'(x_k)\|_*$. Let $d_k \in \mathbb{R}^n$ be such that $\|d_k\| = 1$ and $\langle f'(x_k), d_k \rangle = n_k$ (i.e., $-d_k$ is a steepest descent direction).

Fix $0 < \xi < 1$. Define the step size at iteration k as $h_k = \xi \left(\frac{1+\nu}{L}\right)^{1/\nu} n_k^{1/\nu}$. Then, the next iterate is defined by $x_{k+1} = x_k - h_k d_k$. From (2), we have

$$f(x_{k+1}) \le f(x_k) - \langle f'(x_k), h_k d_k \rangle + \frac{L}{1+\nu} \|h_k d_k\|^{1+\nu}.$$

This gives

$$f(x_{k+1}) \le f(x_k) - \xi \left(\frac{1+\nu}{L}\right)^{1/\nu} n_k^{1+1/\nu} + \xi^{1+\nu} \left(\frac{1+\nu}{L}\right)^{1/\nu} n_k^{1+1/\nu}$$

$$= f(x_k) - \xi(1 - \xi^\nu) \left(\frac{1+\nu}{L}\right)^{1/\nu} n_k^{1+1/\nu}.$$

Summing over k from 0 to K, we get

$$\xi(1 - \xi^\nu) \left(\frac{1+\nu}{L}\right)^{1/\nu} \sum_{k=0}^{K} n_k^{1+1/\nu} \le f(x_0) - f(x_{K+1}) \le f(x_0) - f_*.$$

We conclude that

$$\min_{0 \le k \le K} \|f'(x_k)\|_*^{1+1/\nu} \le \frac{1}{K+1} \left(\frac{L}{1+\nu}\right)^{1/\nu} \frac{f(x_0) - f_*}{\xi(1 - \xi^\nu)}.$$

In particular, if we choose $\xi = \left(\frac{1}{1+\nu}\right)^{1/\nu}$, we obtain

$$\min_{0 \le k \le K} \|f'(x_k)\|_*^{1+1/\nu} \le \frac{1}{K+1} \frac{1+\nu}{\nu} L^{1/\nu} (f(x_0) - f_*).$$

The objective is to converge to a quasi-stationary point: For some fixed $\epsilon > 0$, we want to find an $\bar{x}$ such that $\|f'(\bar{x})\|_* \le \epsilon$. Then, the above-presented method stops after a number of iterations not greater than

$$\left\lceil \frac{1}{\epsilon^{1+1/\nu}} \frac{1+\nu}{\nu} L^{1/\nu} (f(x_0) - f_*) \right\rceil, \tag{3}$$

where $\lceil \cdot \rceil$ denotes the ceiling operator.

In summary, with the above example, we have shown that:

– The existence of an L satisfying (2) (i.e., being ν-approximable) is sufficient to assert the convergence of the method to a quasi-stationary point. Contrary to the way it is presented in many textbooks, it is not necessary to resort to the ν-Hölder continuous gradient assumption (1).
– The bound (3) on the number of iterations shows that the knowledge of the smallest L satisfying (2) (i.e., $L_f(\nu)$) allows us to have a better bound on the total number of iterations required by the method.

As we will see in the next section, a lower bound on $L_f(\nu)$ can be obtained from the ν-Hölder constant of the gradient. In the above developments, the derivation of (3) requires only a global upper bound on the difference between the function and its first-order approximation, and not on the absolute value thereof, as in (2). A more general class of functions can be analyzed, if we allow for different parameters in the lower bound and in the upper bound on the difference between the function and its first-order approximation. For this class of functions, we obtain similar conclusions as for the case of ν-approximable functions (see next section as well).

4 Main Results

The main contributions of this paper are summarized in Theorems 4.1–4.2. Theorem 4.1 shows that ν-approximability implies ν-Hölder continuity of the gradient, and also provides an upper bound on the Hölder constant of the gradient. From this upper bound, we obtain an interval, depending on $M_f(\nu)$, in which the value of $L_f(\nu)$ is guaranteed to be. We also show that a smaller interval can be obtained, if we further assume that the norm $\|\cdot\|$, on the domain of f, is Euclidean.

Theorem 4.1 *Let $0 \leq \nu \leq 1$. Let f be a differentiable function from E (with norm $\|\cdot\|$) to $\mathbb{R}$. Suppose that f is ν-approximable, and denote by $L_f(\nu)$ its ν-approximation parameter. Then, f has ν-Hölder continuous gradient with ν-Hölder constant $M_f(\nu)$ satisfying*[3]

$$L_f(\nu) \leq M_f(\nu) \leq 2^{1-\nu} \left(\frac{1+\nu}{\nu} \right)^{\nu} L_f(\nu). \tag{4}$$

Moreover, if we further assume that $\|\cdot\|$ is Euclidean, then

$$L_f(\nu) \leq M_f(\nu) \leq \frac{2^{1-\nu}}{\sqrt{1+\nu}} \left(\frac{1+\nu}{\nu} \right)^{\nu/2} L_f(\nu). \tag{5}$$

See Fig. 1 for a comparison of (4) and (5).

Proof Let us fix arbitrary $\bar{x}, \bar{y} \in E$. From Remark 2.1, we may assume, without loss of generality, that $\bar{x} = 0$, $f(\bar{x}) = 0$, and $f'(\bar{x}) = 0$. Let $\varphi = f'(\bar{y})$. Then, for

---

[3] With the convention that $0^0 = 1$, hence, $\left(\frac{1+\nu}{\nu} \right)^{\nu} = \left(\frac{1+\nu}{\nu} \right)^{\nu/2} = 1$, when $\nu = 0$.
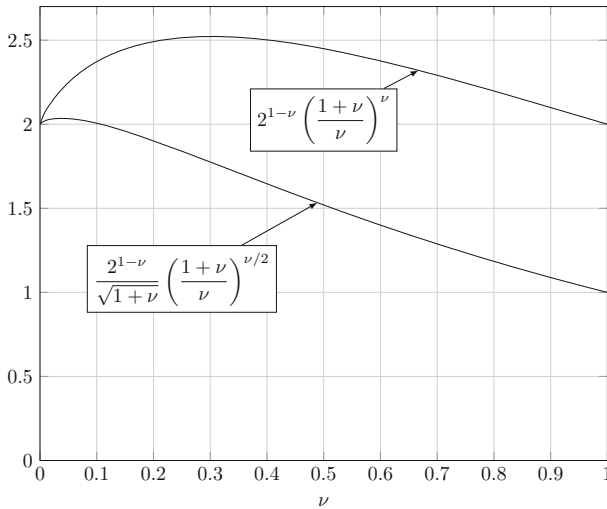
Fig. 1 Comparison of the coefficients appearing in the upper bounds on the ν-Hölder constant in (4) and (5) in Theorem 4.1

any two directions $z_1, z_2 \in E$, we have (from (2) with $(x, y) = (0, z_1)$ and with $(x, y) = (\bar{y}, z_1)$)

$$\frac{-L_f(\nu)}{1+\nu}\|z_1\|^{1+\nu} \le f(z_1) \le f(\bar{y}) + \langle \varphi, z_1 - \bar{y}\rangle + \frac{L_f(\nu)}{1+\nu}\|z_1 - \bar{y}\|^{1+\nu},$$

and (from (2) with $(x, y) = (0, z_2)$ and with $(x, y) = (\bar{y}, z_2)$)

$$\frac{L_f(\nu)}{1+\nu}\|z_2\|^{1+\nu} \ge f(z_2) \ge f(\bar{y}) + \langle \varphi, z_2 - \bar{y}\rangle - \frac{L_f(\nu)}{1+\nu}\|z_2 - \bar{y}\|^{1+\nu}.$$

Subtracting the rightmost and leftmost sides of the first set of inequalities from the second one, we get

$$\langle \varphi, z_2 - z_1 \rangle \le \frac{L_f(\nu)}{1+\nu}\left(\|z_1\|^{1+\nu} + \|z_2\|^{1+\nu} + \|z_1 - \bar{y}\|^{1+\nu} + \|z_2 - \bar{y}\|^{1+\nu}\right).$$

Let $\hat{v} \in E$ be such that $\|\hat{v}\| = 1$ and $\langle \varphi, \hat{v}\rangle = \|\varphi\|_*$. (From the compactness of $\{x \in E : \|x\| = 1\}$, it is always possible to find such a $\hat{v}$.) Then, let $\alpha \ge 0$, and define $z_1 = \frac{\bar{y} - \alpha\hat{v}}{2}$ and $z_2 = \frac{\bar{y} + \alpha\hat{v}}{2}$. This gives

$$\alpha \langle \varphi, \hat{v}\rangle \le 2\frac{L_f(\nu)}{1+\nu}\left(\left\|\frac{\bar{y} - \alpha\hat{v}}{2}\right\|^{1+\nu} + \left\|\frac{\bar{y} + \alpha\hat{v}}{2}\right\|^{1+\nu}\right). \tag{6}$$

Non-Euclidean case Let $r = \|\bar{y}\|$. The right-hand side of (6) can be bounded as follows:

$$\left\|\frac{\bar{y} - \alpha\hat{v}}{2}\right\|^{1+\nu} + \left\|\frac{\bar{y} + \alpha\hat{v}}{2}\right\|^{1+\nu} \leq 2\left(\frac{r+\alpha}{2}\right)^{1+\nu}.$$

If $\nu > 0$, then, from (6) and the above inequality, we obtain (letting $\alpha = \frac{r}{\nu}$)

$$\langle\varphi, \hat{v}\rangle \leq 4\frac{L_f(\nu)}{1+\nu}\frac{\nu}{r}\left(\frac{r(1+\nu)}{2\nu}\right)^{1+\nu} = 2L_f(\nu)\left(\frac{1+\nu}{2\nu}\right)^{\nu}r^{\nu}.$$

Hence, we have the conclusion in the case of $\nu > 0$. On the other hand, if $\nu = 0$, we get $M_f(0) \leq 2L_f(0)$, by letting $\alpha \to +\infty$ in (6).

Euclidean case Now, assume that $\|\cdot\|$ is Euclidean. Then, using the inequality $a^p + b^p \leq 2^{1-p}(a+b)^p$ for every $a, b \geq 0$ and $p \in [0, 1]$, (resulting from the concavity of $x \mapsto x^p$), we get the following upper bound on the right-hand side of (6):

$$\left\|\frac{\bar{y} - \alpha\hat{v}}{2}\right\|^{1+\nu} + \left\|\frac{\bar{y} + \alpha\hat{v}}{2}\right\|^{1+\nu} \leq 2^{(1-\nu)/2}\left(\left\|\frac{\bar{y} - \alpha\hat{v}}{2}\right\|^2 + \left\|\frac{\bar{y} + \alpha\hat{v}}{2}\right\|^2\right)^{(1+\nu)/2}.$$

Let $r = \|\bar{y}\|$. Using the parallelogram identity, we obtain

$$\left\|\frac{\bar{y} - \alpha\hat{v}}{2}\right\|^{1+\nu} + \left\|\frac{\bar{y} + \alpha\hat{v}}{2}\right\|^{1+\nu} \leq 2^{(1-\nu)/2}\left(\frac{r^2 + \alpha^2}{2}\right)^{(1+\nu)/2}.$$

For $\nu > 0$, we get (letting $\alpha = \frac{r}{\sqrt{\nu}}$)

$$\langle\varphi, \hat{v}\rangle \leq L_f(\nu)\frac{2^{1-\nu}}{\sqrt{1+\nu}}\left(\frac{1+\nu}{\nu}\right)^{\nu/2}r^{\nu}.$$

This concludes the proof of the theorem.   □

We now move to the second main result. Observe that inequality (2) is equivalent to

$$\frac{-L}{1+\nu}\|y - x\|^{1+\nu} \leq f(y) - f(x) - \langle f'(x), y - x\rangle \leq \frac{L}{1+\nu}\|y - x\|^{1+\nu}. \quad (7)$$

In some situations, we may also want to have different values of L for the lower bound and the upper bound in (7). This arises, for example, if we consider convex functions. In this case, the lower bound is zero, since the first-order approximation of a convex function always lies below the graph of the function. This leads to the following theorem. (Note that, for this theorem, we do not make the distinction between the Euclidean and non-Euclidean case.)

Theorem 4.2 *Let f be a differentiable function from E (with norm $\|\cdot\|$) to $\mathbb{R}$. Let $0 \le \nu \le 1$, and suppose there exist $L^- \ge 0$ and $L^+ \ge 0$, such that, for every $x, y \in E$,*

$$\frac{-L^-}{1+\nu}\|x-y\|^{1+\nu} \le f(y) - f(x) - \langle f'(x), y-x \rangle \le \frac{L^+}{1+\nu}\|y-x\|^{1+\nu}. \quad (8)$$

Then, f has ν-Hölder continuous gradient, and the ν-Hölder constant $M_f(\nu)$ satisfies[4]

$$M_f(\nu) \le 2^{-\nu}\left(\frac{1+\nu}{\nu}\right)^{\nu}\left(L^- + L^+\right).$$

Proof The proof is similar to the proof of Theorem 4.1. Let us fix arbitrary $\bar{x}, \bar{y} \in E$. Without loss of generality, we may assume that $\bar{x} = 0$, $f(\bar{x}) = 0$, and $f'(\bar{x}) = 0$. Denote $\varphi = f'(\bar{y})$. Then, for any two directions $z_1, z_2 \in E$, we have (from (8) with $(x, y) = (0, z_1)$ and with $(x, y) = (\bar{y}, z_1)$)

$$\frac{-L^-}{1+\nu}\|z_1\|^{1+\nu} \le f(z_1) \le f(\bar{y}) + \langle \varphi, z_1 - \bar{y} \rangle + \frac{L^+}{1+\nu}\|z_1 - \bar{y}\|^{1+\nu},$$

and (from (8) with $(x, y) = (0, z_2)$ and with $(x, y) = (\bar{y}, z_2)$)

$$\frac{L^+}{1+\nu}\|z_2\|^{1+\nu} \ge f(z_2) \ge f(\bar{y}) + \langle \varphi, z_2 - \bar{y} \rangle - \frac{L^-}{1+\nu}\|z_2 - \bar{y}\|^{1+\nu}.$$

Subtracting the rightmost and leftmost sides of the first set of inequalities from the second one, we get

$$\langle \varphi, z_2 - z_1 \rangle \le \frac{L^-}{1+\nu}\left(\|z_1\|^{1+\nu} + \|z_2 - \bar{y}\|^{1+\nu}\right) + \frac{L^+}{1+\nu}\left(\|z_2\|^{1+\nu} + \|z_1 - \bar{y}\|^{1+\nu}\right).$$

Let $\hat{v} \in E$ be such that $\|\hat{v}\| = 1$ and $\langle \varphi, \hat{v} \rangle = \|\varphi\|_*$. Then, let $\alpha \ge 0$, and define $z_1 = \frac{\bar{y}-\alpha\hat{v}}{2}$ and $z_2 = \frac{\bar{y}+\alpha\hat{v}}{2}$. This gives

$$\alpha \langle \varphi, \hat{v} \rangle \le \frac{2}{1+\nu}\left(L^-\left\|\frac{\bar{y}-\alpha\hat{v}}{2}\right\|^{1+\nu} + L^+\left\|\frac{\bar{y}+\alpha\hat{v}}{2}\right\|^{1+\nu}\right). \quad (9)$$

Let $r = \|\bar{y}\|$. The right-hand side of (9) can be bounded as follows:

$$L^-\left\|\frac{\bar{y}-\alpha\hat{v}}{2}\right\|^{1+\nu} + L^+\left\|\frac{\bar{y}+\alpha\hat{v}}{2}\right\|^{1+\nu} \le (L^- + L^+)\left(\frac{r+\alpha}{2}\right)^{1+\nu}.$$

We conclude in the same way as for the proof of Theorem 4.1. $\square$

---

[4] With the convention that $0^0 = 1$, as in Theorem 4.1.

Corollary 4.1 *Let f be ν-approximable. If f is convex, then the upper bound in (4) can be improved by a factor $1/2$: $L_f(\nu) \leq M_f(\nu) \leq 2^{-\nu} \left(\frac{1+\nu}{\nu}\right)^\nu L_f(\nu)$.*

5 Application to Functions with Lipschitz Continuous Gradient

In this section, we particularize Theorem 4.1 and Corollary 4.1 to the classical case of $\nu = 1$ (Lipschitz continuity). Moreover, we obtain that the bounds are tight, meaning that there exist functions, such that $M_f(\nu)$ attains either the lower or upper bound in (4) when $\nu = 1$. Recall that f is 1-approximable, if there exists an $L \geq 0$, such that, for every $x, y \in E$,

$$|f(y) - f(x) - \langle f'(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2. \tag{10}$$

To simplify the notation, we will denote $L_f = L_f(1)$ and $M_f = M_f(1)$.

Corollary 5.1 *Let f be a differentiable function from E (with norm $\|\cdot\|$) to $\mathbb{R}$. Then, f is 1-approximable, if and only if the gradient of f is Lipschitz continuous. Moreover, we have the following bounds, depending on the assumptions on f and $\|\cdot\|$:*

1. *In general, we have $\frac{1}{2} M_f \leq L_f \leq M_f$.*
2. *If $\|\cdot\|$ is Euclidean, then $L_f = M_f$.*
3. *If f is convex, then $L_f = M_f$.*

The proof of the corollary directly follows from Theorem 4.1 for the first and second items. The third item can be easily proved with Corollary 4.1. (Let us also mention that the case of convex functions is also proved, with different arguments, in [6, Theorem 2.1.5], for example.)

We would like to know whether the bounds on L_f that we obtained in the first item of Corollary 5.1 are tight. The second and third items directly give that the upper bound $L_f = M_f$ is reached, when $\|\cdot\|$ is Euclidean, or when $f(\cdot)$ is convex. Hence, the main question is: "Is there a normed space $(E, \|\cdot\|)$ and a function f, defined on E, such that $L_f = \frac{1}{2} M_f$?" The answer is yes, as shown in the following example:

Example 5.1 Let $E = \mathbb{R}^2$ with the *infinity norm*, $\|x\| = \max\{|x^{(1)}|, |x^{(2)}|\}$ for all $x = (x^{(1)}, x^{(2)})^\top \in \mathbb{R}^2$. (We use superscripts to denote the components of vectors in $\mathbb{R}^2$.) Define $f(x) = (x^{(1)})^2 - (x^{(2)})^2$. The gradient of f at x satisfies $\langle f'(x), h \rangle = 2x^{(1)} h^{(1)} - 2x^{(2)} h^{(2)}$ for all $h = (h^{(1)}, h^{(2)})^\top \in E$.

We check that (10) holds with $L = 2$. Indeed, for any $x, y \in \mathbb{R}^2$,

$$\text{Left-hand side of (10)} = \left| \left[y^{(1)} - x^{(1)}\right]^2 - \left[y^{(2)} - x^{(2)}\right]^2 \right|$$

$$\leq \max\left\{ \left|y^{(1)} - x^{(1)}\right|^2, \left|y^{(2)} - x^{(2)}\right|^2 \right\}$$

$$= \|y - x\|^2.$$

On the other hand, if $x = (1, 1)^\top$, and $y = (0, 0)^\top$, then $\langle f'(x) - f'(y), h \rangle = 2h^{(1)} - 2h^{(2)}$ for every $h \in \mathbb{R}^2$. Taking $h = (1, -1)^\top$, we get

$$\|f'(x) - f'(y)\|_* \geq |\langle f'(x), h \rangle - \langle f'(y), h \rangle| = 4.$$

Since $\|h\| = 1$, it follows that $M_f \geq 4$. Hence, we have $L_f \leq 2 \leq \frac{1}{2} M_f$, which concludes the proof that $L_f = \frac{1}{2} M_f$ for some functions f.

Before ending this section, we would like to point out that there is still a gap between the first and the second items of Corollary 5.1. Indeed, in the second item, we state: if the norm on E is Euclidean, then $L_f = M_f$. On the other hand, if $\|\cdot\|$ is any norm, then we can ensure that $\frac{1}{2} M_f \leq L_f$. However, we do not say anything about the possibility of finding a space E with a non-Euclidean norm, but such that every function on E satisfies $L_f = M_f$. We will prove in Sect. 6 that *this situation is impossible*. In other words, if all functions on a space E (with norm $\|\cdot\|$) satisfy $L_f = M_f$, then $\|\cdot\|$ is Euclidean. In fact, it suffices to have $L_f = M_f$ for every quadratic function with rank 2 (to be defined below), to conclude that the norm $\|\cdot\|$ is Euclidean.

6 Application: Norm of Quadratic Functions

In this section, we apply the results of the previous section to quadratic functions, that is, functions defined by self-adjoint operators. We will use Corollary 5.1 to derive bounds on the norms of self-adjoint operators; see Proposition 6.1. We do not believe that the results presented in Proposition 6.1 are new. (For instance, the case of $\|\cdot\|$ Euclidean is a standard result in functional analysis on Hilbert spaces; see, e.g., [8].) However, we will use these results—and more precisely the fact that they are a consequence of Corollary 5.1—to show that it is impossible to have a space E with a non-Euclidean norm, but such that every function on E satisfies $L_f = M_f$ (cf. the last paragraph of the previous section).

Definition 6.1 (*Self-adjoint operator*) Let E be a real finite-dimensional vector space.

1. A *self-adjoint operator* on E is a linear map $B : E \to E^*$, $x \mapsto Bx$, satisfying $\langle Bx, y \rangle = \langle By, x \rangle$ for every $x, y \in E$.
2. If B is a self-adjoint operator on E, we let $Q_B : E \to \mathbb{R}$ be the *quadratic form* associated with B, defined by $Q_B(x) = \langle Bx, x \rangle$ for every $x \in E$.

The set of self-adjoint operators on E is a real (finite-dimensional) vector space. Given a norm $\|\cdot\|$ on E, we define the following norm on this space:

$$\|B\| = \max\{\langle Bx, y \rangle : \|x\| = \|y\| = 1\} = \max\{\|Bx\|_* : \|x\| = 1\}.$$

We also define the norm of a quadratic form Q_B:

$$\|Q_B\| = \max\{|Q_B(x)| : \|x\| = 1\} = \max\{|\langle Bx, x \rangle| : \|x\| = 1\}.$$

The quantities $\|B\|$ and $\|Q_B\|$ appear in several fields of mathematics:

Example 6.1 If $E = \mathbb{R}^n$, then B can be identified with a symmetric matrix $\tilde{B} \in \mathbb{R}^{n \times n}$. If $\|x\| = \sqrt{x^\top x}$ (canonical Euclidean norm), then $\|B\| = \|Q_B\| = \rho(\tilde{B})$, where $\rho(\tilde{B})$ is the largest absolute value of the eigenvalues of $\tilde{B}$.

If $\|\cdot\|$ is the ℓ_1-norm, i.e., $\|x\| = \sum_{i=1}^n |x^{(i)}|$, where $x = (x^{(1)}, \ldots, x^{(n)})^\top$, then $\|B\| = \max_{1 \leq i,j \leq n} |\tilde{B}^{(i,j)}|$, where $\tilde{B}^{(i,j)}$ is the (i, j)th entry of matrix $\tilde{B}$. On the other hand, the value of $\|Q_B\|$ is, in general, NP-hard to compute. This quantity appears, for instance, in the problem of determining the largest clique in a graph (Motzkin–Straus theorem [9]).

The quantity $\|Q_B\|$, with $\|x\| = \max_{1 \leq i \leq n} |x^{(i)}|$ (*infinity* norm), appears, for instance, in the MAXCUT problem [10, Section 4.3.3]. If $\tilde{B}$ is a positive semidefinite matrix (as it is the case in the MAXCUT problem), then we can show that $\|Q_B\| = \|B\|$ (see Proposition 6.1).

Clearly, $\|Q_B\| \leq \|B\|$. The following proposition, which follows from Corollary 5.1, provides a partial converse result.

Proposition 6.1 *Let E be a normed vector space with norm $\|\cdot\|$, and let B be a self-adjoint operator on E. Then, we have the following results, depending on the assumptions on E and on B:*

1. *In general, we have $\frac{1}{2}\|B\| \leq \|Q_B\| \leq \|B\|$.*
2. *If $\|\cdot\|$ is Euclidean, then $\|Q_B\| = \|B\|$.*
3. *If B is positive semidefinite,[5] then $\|Q_B\| = \|B\|$.*

Proof Let $f(x) = \frac{1}{2}Q_B(x) = \frac{1}{2}\langle Bx, x \rangle$. The gradient of f at x is $f'(x) = Bx \in E^*$. Let x, y be in E. The difference between $f(y)$ and its first-order approximation $f(x) + \langle f'(x), y - x \rangle$ is given by

$$\frac{1}{2}\langle By, y \rangle - \frac{1}{2}\langle Bx, x \rangle - \langle Bx, y - x \rangle = \frac{1}{2}\langle By, y \rangle + \frac{1}{2}\langle Bx, x \rangle - \langle Bx, y \rangle$$
$$= \frac{1}{2}\langle B(y - x), y - x \rangle$$
$$= \frac{1}{2}Q_B(y - x).$$

Hence, f is 1-approximable, and $L_f = \|Q_B\|$. It is also not hard to see that f has Lipschitz continuous gradient, and $M_f = \|B\|$. Thus, we obtain the first and second items of the proposition, from the first and second items of Corollary 5.1. Finally, f is convex, if and only if B is positive semidefinite. Hence, we get the desired results from the third item of Corollary 5.1. $\qquad\square$

The second item in Proposition 6.1 states that, when the norm on E is Euclidean, the two quantities $\|B\|$ and $\|Q_B\|$ coincide for every B. The following theorem shows that $\|Q_B\| = \|B\|$ for every self-adjoint operator B on a normed vector space E, only if the norm of E is Euclidean.[6] (The proof of the theorem is presented in "Appendix.")

---

[5] That is, $\langle Bx, x \rangle \geq 0$ for every $x \in E$.

[6] The result presented in Theorem 6.1 seems to be a novel (to the best of the authors' knowledge) characterization of Euclidean norms in the finite-dimensional case. (For a detailed survey of results on equivalent characterizations of Euclidean norms, we refer the reader to the celebrated book by Amir [11].)

Theorem 6.1 *Let $(E, \|\cdot\|)$ be a finite-dimensional normed vector space. Then, $\|\cdot\|$ is Euclidean, if and only if $\|Q_B\| = \|B\|$ for every self-adjoint operator B on E.*

Remark 6.1 From the proof of Theorem 6.1 (see "Appendix"), we can obtain a stronger version of this theorem: $\|\cdot\|$ is Euclidean, if and only if $\|Q_B\| = \|B\|$ for every self-adjoint operator B on E *with rank 2*, that is, for every B that can be expressed as $Bx = \langle \varphi, x \rangle \varphi \pm \langle \psi, x \rangle \psi$ for some $\varphi, \psi \in E^*$.

Finally, Theorem 6.1 allows us to answer the question in the last paragraph of the previous section:

Corollary 6.1 *Let $(E, \|\cdot\|)$ be a finite-dimensional normed vector space, and suppose that, for every real-valued function f on E, we have $L_f = M_f$ (using the notation of Corollary 5.1). Then, $\|\cdot\|$ is Euclidean.*

Proof If $L_f = M_f$ for every function f on E, this is also true for every quadratic function $f(x) = Q_B(x) = \langle Bx, x \rangle$, where B is any self-adjoint operator on E. Hence, using similar developments as in the proof of Proposition 6.1, we get that $\|Q_B\| = \|B\|$ for every self-adjoint operator B on E. Thus, Theorem 6.1 implies that $\|\cdot\|$ must be Euclidean. $\qquad\square$

7 Conclusions

We have shown that Hölder continuity of the gradient is a sufficient and *necessary* condition for a function to have a global upper bound on the error of its first-order Taylor approximation. We established the link between the parameter appearing in the global upper bound on the error of the first-order Taylor approximation and the Hölder constant of the gradient: This takes the form of an interval, depending on the Hölder constant, in which the approximation parameter is guaranteed to be.

For the Lipschitz case, an example shows that the interval cannot be shortened. On top of this, tighter bounds can be obtained, if we further assume that the function is convex, or if the underlying norm is Euclidean. In particular, for the Lipschitz case, if the norm is Euclidean, then the interval reduces to a single point, and Theorem 6.1 shows that the assumption that the norm is Euclidean is *not* conservative. We have not addressed in this paper the question of whether the interval for the general case (i.e., with Hölder exponent different than one) is tight or not. We leave it for further work.

Another follow-up topic is a generalization of Theorem 4.1 to higher-order derivatives and higher-order Taylor approximations. Those play central roles in higher-order optimization methods. By applying a similar argument as in the proof of Theorem 4.1, one can obtain an interval, depending on the Hölder constant of the derivative of a given order, in which the approximation parameter of the Taylor approximation (of order equal to the one of the derivatives) is guaranteed to be. Drawing upon this observation, the following aspects still need to be addressed: Are the obtained bounds tight? Can we obtain better bounds, if we further assume that the norm is Euclidean? We plan to study these questions in further work.

Appendix: Proof of Theorem 6.1

The proof relies on the fact that, if $\|\cdot\|$ is not Euclidean, then the *unit ball* defined by $\|\cdot\|$, i.e., $\{x \in E : \|x\| \leq 1\}$, is not equal to the ellipsoid with smallest volume containing this ball. Based on this ellipsoid, we will build a self-adjoint operator $B : E \to E^*$, such that $\|Q_B\| < \|B\|$. The notions of ellipsoid and (Lebesgue) volume are defined on $\mathbb{R}^n$ only. The following lemma implies, among other things, that there is no loss of generality in restricting to the case $E = \mathbb{R}^n$:

Lemma A.1 *Let E be a real vector space with norm $\|\cdot\|$, and let $A : E \to E'$ be a bijective linear map. Then, $\|\cdot\|$ is Euclidean, if and only if the norm $\|\cdot\|'$ on E', defined by $\|x\|' = \|A^{-1}x\|$, is Euclidean.*

Proof Straightforward from the definition of $\|\cdot\|$ being Euclidean, if and only if it is induced by a scalar product, i.e., if and only if there exists a self-adjoint operator $H : E \to E^*$, satisfying $\|x\|^2 = \langle Hx, x \rangle$ for all $x \in E$. □

Proof of Theorem 6.1 The "only if" part follows from Proposition 6.1. For the proof of the "if" part, let E be an n-dimensional vector space, and let $\|\cdot\|$ be a non-Euclidean norm on E. We will build a self-adjoint operator B on E, such that $\|Q_B\| < \|B\|$.

By Lemma A.1, we may assume that $E = \mathbb{R}^n$ and that $\|\cdot\|$ is a non-Euclidean norm on $\mathbb{R}^n$. We use superscripts to denote the components of vectors in $\mathbb{R}^n$: $x = (x^{(1)}, \ldots, x^{(n)})^\top$.

Let $K = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$. Because K is compact, convex, with non-empty interior, and symmetric with respect to the origin, the Löwner–John ellipsoid theorem [12,13] asserts that there exists a unique ellipsoid $\mathcal{E}$, with minimal volume, and such that $K \subseteq \mathcal{E}$. Moreover, $\mathcal{E}$ is centered at the origin, and K has n linearly independent vectors on the boundary of $\mathcal{E}$.

Let $L : \mathbb{R}^n \to \mathbb{R}^n$ be a linear isomorphism, such that $L\mathcal{E}$ is the Euclidean ball $\mathbb{B}^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$, where $\|x\|_2 = \sqrt{x^\top x}$ is the canonical Euclidean norm on $\mathbb{R}^n$. Let $\|x\|' = \|L^{-1}x\|$, and let $K' = \{x \in \mathbb{R}^n : \|x\|' \leq 1\}$. By Lemma A.1, $\|\cdot\|'$ is not Euclidean. Since $K' = LK$, it is clear that K' is compact, convex, with non-empty interior, and symmetric with respect to the origin. Moreover, K' is included in $\mathbb{B}^n$, and it has n linearly independent vectors on the boundary $\mathbb{S}^{n-1}$ of $\mathbb{B}^n$.

We will need the following lemma to conclude the proof of Theorem 6.1:

Lemma A.2 *There exist $u, v \in \mathbb{S}^{n-1} \cap K'$, not colinear, such that $\frac{u+v}{\|u+v\|_2} \notin K'$.*

We proceed with the proof of Theorem 6.1 (a proof of Lemma A.2 is provided at the end of this "Appendix"). Let u, v be as in Lemma A.2, and define $e_1 = \frac{u+v}{\|u+v\|_2}$ and $e_2 = \frac{u-v}{\|u-v\|_2}$. Note that these vectors are orthonormal (w.r.t. the inner product $x^\top y$).

Let $\kappa = \max\{|e_1^\top x| : x \in K'\}$. Since $|e_1^\top x| < 1$ for every $x \in \mathbb{B}^n \setminus \{\pm e_1\}$, and $\pm e_1 \notin K'$, we have that $\kappa < 1$. Moreover, $\kappa > 0$, since $\text{int}(K') \neq \varnothing$. Let $\tilde{B}$ be the self-adjoint operator on $\mathbb{R}^n$, defined by

$$\langle \tilde{B}x, y \rangle = \frac{1}{\kappa^2} \left(e_1^\top x\right)\left(e_1^\top y\right) - \left(e_2^\top x\right)\left(e_2^\top y\right)$$

for every $x, y \in \mathbb{R}^n$. Let $x \in K'$. Then,

$$-1 \leq -\left(e_2^\top x\right)^2 \leq \langle \tilde{B}x, x \rangle \leq \frac{1}{\kappa^2}\left(e_1^\top x\right)^2 \leq \frac{1}{\kappa^2}\kappa^2 = 1.$$

It follows that, for every $x \in \mathbb{R}^n$ with $x \neq 0$, $|\langle \tilde{B}x, x \rangle| = \|x\|'^2 \, |\langle \tilde{B}\frac{x}{\|x\|'}, \frac{x}{\|x\|'}\rangle| \leq \|x\|'^2$. Hence, $\|Q_{\tilde{B}}\| \leq 1$. Now, we will show that $|\langle \tilde{B}u, v \rangle| > \|u\|'\|v\|'$ (where u, v are as above). Therefore, let $\alpha = \|u + v\|_2$ and $\beta = \|u - v\|_2$. Observe that $u = \frac{\alpha e_1 + \beta e_2}{2}$ and $v = \frac{\alpha e_1 - \beta e_2}{2}$. Thus,

$$\langle \tilde{B}u, v \rangle = \frac{1}{\kappa^2}\frac{\alpha^2}{4} + \frac{\beta^2}{4} = \frac{1 - \kappa^2}{\kappa^2}\frac{\alpha^2}{4} + \frac{\alpha^2}{4} + \frac{\beta^2}{4}.$$

This shows that $\langle \tilde{B}u, v \rangle > 1$, since (by the parallelogram identity)

$$\frac{\alpha^2}{4} + \frac{\beta^2}{4} = \frac{1}{4}\left(\|u + v\|_2^2 + \|u - v\|_2^2\right) = 1,$$

$0 < \kappa < 1$, and $\alpha > 0$. Since $u, v \in K'$ (i.e., $\|u\|', \|v\|' \leq 1$), we have that $\|u\|'\|v\|' \leq 1 < |\langle \tilde{B}u, v \rangle|$. Thus, $\|\tilde{B}\| > 1$.

Finally, define the self-adjoint operator B on E by $\langle Bx, y \rangle = \langle \tilde{B}Lx, Ly \rangle$. It is clear, from the definition of $\|\cdot\|'$, that $|\langle Bx, x \rangle| \leq \|x\|^2$ for every $x \in E$ and $|\langle Bx, y \rangle| > \|x\|\|y\|$ for $x = L^{-1}u$ and $y = L^{-1}v$ (where u, v are as above). Hence, one gets $\|Q_B\| \leq 1 < \|B\|$. This concludes the proof of Theorem 6.1. $\qquad\square$

It remains to prove Lemma A.2. The following proposition, known as *Fritz John necessary conditions for optimality* will be useful in the proof of Lemma A.2:

Proposition A.1 (Fritz John necessary conditions [12]) *Let S be a compact metric space. Let $F(x)$ be a real-valued function on $\mathbb{R}^n$, and let $G(x, y)$ be a real-valued function defined for all $(x, y) \in \mathbb{R}^n \times S$. Assume that $F(x)$ and $G(x, y)$ are both differentiable with respect to x and that $F(x)$, $G(x, y)$, $\frac{\partial F}{\partial x}(x)$, and $\frac{\partial G}{\partial x}(x, y)$ are continuous on $\mathbb{R}^n \times S$. Let $R = \{x \in \mathbb{R}^n : G(x, y) \leq 0, \forall y \in S\}$, and suppose that R is non-empty.*

Let $x^* \in R$ be such that $F(x^*) = \max_{x \in R} F(x)$. Then, there is $m \in \{0, \ldots, n\}$, and points $y_1, \ldots, y_m \in S$, and nonnegative multipliers $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$, such that (i) $G(x^*, y_i) = 0$ for every $1 \leq i \leq m$, (ii) $\sum_{i=0}^m \lambda_i > 0$, and (iii)

$$\lambda_0 \frac{\partial F}{\partial x}(x^*) = \sum_{i=1}^m \lambda_i \frac{\partial G}{\partial x}(x^*, y_i).$$

We refer the reader to [12] for a proof of Proposition A.1.

Proof of Lemma A.2 Consider the following optimization problem:

$$
\begin{aligned}
\text{maximize} \quad & F(x) := \|x\|_2^2 \\
\text{subject to} \quad & G(x, y) := x^\top y - 1 \le 0 \quad \text{for every } y \in \mathbb{S}^{n-1} \cap K',
\end{aligned}
\tag{11}
$$

with variable $x \in \mathbb{R}^n$.

First, we show that (11) is bounded. Suppose the contrary, and, for every $k \ge 1$, let x_k be a feasible solution with $\|x_k\|_2 \ge k$. Let $\hat{x}_k = x_k / \|x_k\|_2$. Taking a subsequence if necessary, we may assume that $\hat{x}_k$ converges to some $\hat{x}_*$, with $\|\hat{x}_*\|_2 = 1$. Since $\hat{x}_k^\top y \le 1/\|x_k\|_2$ for every $y \in \mathbb{S}^{n-1} \cap K'$, we have that $\hat{x}_*^\top y \le 0$ for every $y \in \mathbb{S}^{n-1} \cap K'$. By symmetry of $\mathbb{S}^{n-1} \cap K'$, it follows that $\hat{x}_*^\top y = 0$ for every $x \in \mathbb{S}^{n-1} \cap K'$, a contradiction with the fact that $\mathbb{S}^{n-1} \cap K'$ contains n linearly independent vectors. Hence, the set of feasible solutions of (11) is bounded, and closed (as the intersection of closed sets), so that (11) has an optimal solution, say $\bar{x}$.

We will show that $\|\bar{x}\|_2 > 1$. Therefore, we use the fact that $K' \ne \mathbb{B}^n$.[7] Fix some $z \in \mathbb{S}^{n-1} \setminus K'$, and let $\eta = \max\{z^\top y : y \in K'\}$. Since $z^\top y < 1$ for every $y \in \mathbb{B}^n \setminus \{z\}$, and $z \notin K'$, we have that $\eta < 1$. Let $x = z/\eta$. From the definition of η, it is clear that x is a feasible solution of (11). Moreover, $\|x\|_2 = \eta^{-1} > 1$, so that $\|\bar{x}\|_2 \ge \|x\|_2 > 1$.

The gradient of F at $\bar{x}$ is equal to $2\bar{x}$. Then, Proposition A.1 asserts that there exist vectors $y_1, \dots, y_m \in \mathbb{S}^{n-1} \cap K'$, and nonnegative multipliers $\lambda_0, \lambda_1, \dots, \lambda_m \ge 0$, such that $\bar{x}^\top y_i = 1$ for every $1 \le i \le m$, $\sum_{i=0}^{m} \lambda_i > 0$, and $\lambda_0 \bar{x} = \sum_{i=1}^{m} \lambda_i y_i$. If $\lambda_0 = 0$, then $0 = \sum_{i=1}^{m} \lambda_i \bar{x}^\top y_i = \sum_{i=1}^{m} \lambda_i > 0$, a contradiction. Hence, $\lambda_0 > 0$. Suppose that $y_1, \dots, y_m$ are colinear. This implies that all y_i's must be parallel to $\bar{x}$ (because $\lambda_0 \bar{x} = \sum_{i=1}^{m} \lambda_i y_i$ and $\lambda_0 \bar{x} \ne 0$), and since they are in $\mathbb{S}^{n-1}$, we have that $y_i = \pm \bar{x}/\|\bar{x}\|_2$, so that $\bar{x}^\top y_i = \|\bar{x}\|_2 > 1$ or $-\bar{x}^\top y_i = \|\bar{x}\|_2 > 1$. This gives a contradiction, since $-y_i$ and y_i are both in $\mathbb{S}^{n-1} \cap K'$ (by the symmetry of $\mathbb{S}^{n-1} \cap K'$). Thus, there exist at least two non-colinear vectors $u, v \in \mathbb{S}^{n-1} \cap K'$ satisfying $\bar{x}^\top u = 1$ and $\bar{x}^\top v = 1$.

Let $e_1 = \frac{u+v}{\|u+v\|_2}$. Since u and v are not colinear, $\|u + v\|_2 < 2$, and thus, $\bar{x}^\top e_1 = 2/\|u + v\|_2 > 1$. This shows that $e_1 \notin \mathbb{S}^{n-1} \cap K'$. By definition, e_1 is in $\mathbb{S}^{n-1}$, so that $e_1 \notin K'$, concluding the proof of the lemma. □

References

1. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
2. Yashtini, M.: On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients. Optim. Lett. **10**(6), 1361–1370 (2016)
3. Cartis, C., Gould, N.I., Toint, P.L.: Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients. Optim. Methods Softw. **32**(6), 1273–1298 (2017)

---

[7] Indeed, a norm $\|\cdot\|$ is completely determined by its unit ball K, via the identity $\|x\| = \inf\{\alpha > 0 : x/\alpha \in K\}$. Since $\|\cdot\|'$ is not Euclidean, its unit ball cannot be equal to $\mathbb{B}^n$ (since $\mathbb{B}^n$ is the unit ball of the canonical Euclidean norm).

4. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. **152**(1–2), 381–404 (2015)
5. Boumal, N., Absil, P.A., Cartis, C.: Global rates of convergence for nonconvex optimization on manifolds. IMA J. Numer. Anal. **39**(1), 1–33 (2018)
6. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Berlin (2013)
7. Jordan, P., von Neumann, J.: On inner products in linear, metric spaces. Ann. Math. **36**(3), 719–723 (1935)
8. Friedman, A.: Foundations of Modern Analysis. Courier Corporation, North Chelmsford (1982)
9. Motzkin, T.S., Straus, E.G.: Maxima for graphs and a new proof of a theorem of Turán. Can. J. Math. **17**, 533–540 (1965)
10. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications, vol. 2. SIAM, Philadelphia (2001)
11. Amir, D.: Characterizations of Inner Product Spaces. Birkhauser Verlag, Basel (1986)
12. John, F.: Extremum problems with inequalities as subsidiary conditions. In: Giorgi, G., Kjeldsen, T. (eds.) Traces and Emergence of Nonlinear Programming. Birkhäuser, Basel (2014)
13. Ball, K.: Ellipsoids of maximal volume in convex bodies. Geom. Dedicata **41**(2), 241–250 (1992)