
MOX–Report No. 29/2009

# Finite Element Approximation of the Sobolev Constant

Paola F. Antonietti, Aldo Pratelli

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

# Finite Element Approximation of the Sobolev Constant *

Paola F. Antonietti[♯] Aldo Pratelli[♭]

September 22, 2009

[♯] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
paola.antonietti@polimi.it

[♭] Dipartimento di Matematica
Università di Pavia
via Ferrata 1, 27100 Pavia, Italy
aldo.pratelli@unipv.it

**Keywords**: Sobolev constant, finite elements, quantitative estimates

**AMS Subject Classification**: 46E35, 65N30.

**Abstract**

Denoting by $S$ the sharp constant in the Sobolev inequality in $\mathrm{W}_0^{1,2}(B)$, being $B$ the unit ball in $\mathbb{R}^3$, and denoting by $S_h$ its approximation in a suitable finite element space, we show that $S_h$ converges to $S$ as $h \searrow 0$ with a polynomial rate of convergence. We provide both an upper and a lower bound on the rate of convergence, and present some numerical results.

## 1 Introduction

The Sobolev inequality in $\mathbb{R}^n$ says that, given $1 \leq p < n$, one has

$$\|Df\|_{L^p(\mathbb{R}^n)} \geq S(p,n)\|f\|_{L^{p^\star}(\mathbb{R}^n)} \tag{1}$$

*The first author has been supported by *Azioni Integrate Italia–Spagna* through the project *Tecniche numeriche all'avanguardia e metodi di ottimizzazione di forma per problemi di fluido-dinamica*. The work of the second author is partially supported by the ERC Advanced Grant 2008 *Analytic Techniques for Geometric and Functional Inequalities*, and by the MEC through the 2008 project MTM2008-03541.

for every $f \in \mathrm{W}^{1,p}(\mathbb{R}^n)$, where as usual $p^\star = np/(n-p)$. For $1 < p < n$, the sharp constant $S(p,n)$ was found by Aubin and Talenti [2, 11], and it is given by

$$S(p,n) = \sqrt{\pi}\, n^{1/p} \left(\frac{n-p}{p-1}\right)^{(p-1)/p} \left(\frac{\Gamma(n/p)\Gamma(1+n-n/p)}{\Gamma(1+n/2)\Gamma(n)}\right)^{1/n}, \qquad (2)$$

where $\Gamma(\cdot)$ is the gamma function, *i.e.*, $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)dt$, $x \in \mathbb{R}$.

In particular, it is known [11, 7] that the optimal functions are exactly those of the form

$$g_{a,b,x_0}(x) = \frac{a}{\left(1 + b|x - x_0|^{p'}\right)^{(n-p)/p}}, \qquad (3)$$

with $a \in \mathbb{R} \setminus \{0\}$, $b \in \mathbb{R}^+$ and $x_0 \in \mathbb{R}^n$, so that (2) follows by direct calculation.

Passing from $\mathbb{R}^n$ to the case of a generic open set $\Omega$, the Sobolev inequality still holds for any $f \in \mathrm{W}^{1,p}_0(\Omega)$ with the same constant as in (1), *i.e.*,

$$\|Df\|_{L^p(\Omega)} \geq S(p,n)\|f\|_{L^{p^\star}(\Omega)} \qquad \forall f \in \mathrm{W}^{1,p}_0(\Omega)\,.$$

As we will show in Lemma 2.1, the constant $S(p,n)$ is again optimal for any set $\Omega$, however there are no minimizing functions unless in the case $\Omega = \mathbb{R}^n$.

**Remark 1.1** *The case $p = 1$ is very different, and much simpler, as it is easily shown that inequality (1) holds not only in $\mathrm{W}^{1,1}(\mathbb{R}^n)$ but also on $BV(\mathbb{R}^n)$, provided that the term $\|Df\|_{L^p(\mathbb{R}^n)}$ is replaced by the total variation of $f$. Moreover, the optimal functions are exactly those of the form*

$$\hat{g}_{a,\rho,x_0}(x) = a\chi_{B_\rho(x_0)}(x)\,,$$

*with $a \in \mathbb{R} \setminus \{0\}$, $\rho \in \mathbb{R}^+$ and $x_0 \in \mathbb{R}^n$. In words, the optimal functions are exactly the characteristic functions of balls (with any center, any positive radius and multiplied by any constant). A fundamental difference with the case $p > 1$ is that for $p = 1$ all the optimal functions are compactly supported, while for $p > 1$ all the optimal functions have the whole $\mathbb{R}^n$ as support.*

The aim of this paper is to consider a suitable approximation of problem (1), and to study the convergence of the corresponding discrete Sobolev constants towards the continuous one. For simplicity, we will work in the case $n = 3$ and $p = 2$, but all the proofs still hold true for any choice of $p$ and $n$ with just a straightforward modification of the calculations. Let $V_h$ be the set of the $\mathrm{W}^{1,2}_0$ (finite element) functions on the unit ball $B \subseteq \mathbb{R}^3$ which are continuous, piecewise linear and vanish on the boundary of $B$. Since $V_h \subseteq \mathrm{W}^{1,2}(\mathbb{R}^3)$ (clearly extending the functions to 0 out of $B$), the Sobolev inequality (1) holds for all discrete functions $f \in V_h$, and there exists a minimal constant $S_h$ such that

$$\|Df\|_{L^p(B)} \geq S_h\|f\|_{L^{p^\star}(B)} \qquad \forall\, f \in V_h\,.$$

Clearly, $S_h \geq S$, where for brevity we write $S = S(3,2)$. We will prove the following result.

**Theorem 1.1** *The constants $S_h$ converge to $S$ when $h \searrow 0$. More precisely,*

$$S + \frac{1}{C}\, h^{\gamma} \leq S_h \leq S + C h^{1/3}\,,$$

*for two constants $C$, $\gamma > 0$.*

**Remark 1.2** *We have chosen $B$ to be the unit ball in $\mathbb{R}^3$ just for the sake of convenience: many other choices are possible. For example, any open, bounded domain in $\mathbb{R}^3$ with Lipschitz boundary (for instance, the unit cube) is admissible.*

**Remark 1.3** *The estimate of Theorem 1.1 holds true for any*

$$\gamma > \frac{2 \cdot 26^2}{3},$$

*see the discussion at the end of the proof of Proposition 2.2 below.*

We mention that, in the framework of geometric-functional inequalities, the question whether or not optimal constants are the limit of their discrete approximations has been considered also in [4]. More precisely, they consider the Sobolev–Poincaré inequality in the context of discontinuous finite element spaces, and show that, under suitable assumptions, the discrete optimal constants converge to the continuous one (see [4, Proposition 7.1]).

### 1.1 A quantitative form of the Sobolev inequality

In this work we will need to use an improved version of the Sobolev inequality, recently shown in [5], which says that the functions of the form (3), that are known to be the only functions for which (1) is an equality, are also stable: this means that a function for which (1) is *almost* an equality must be *almost* of the form (3).

More precisely, for any function $f \in \mathrm{W}^{1,p}(\mathbb{R}^n)$ we define the *Sobolev deficit*

$$\delta(f) = \frac{\|Df\|_{L^p(\mathbb{R}^n)}}{\|f\|_{L^{p^\star}(\mathbb{R}^n)}} - S,$$

which says how far inequality (1) is from being an equality (in particular $f$ is optimal if and only if $\delta(f) = 0$). We also set the *Sobolev asimmetry*

$$\lambda(f) = \inf\left\{ \frac{\|f - g_{a,b,x_0}\|_{L^{p^\star}(\mathbb{R}^n)}}{\|f\|_{L^{p^\star}(\mathbb{R}^n)}} \,:\, \|f\|_{L^{p^\star}(\mathbb{R}^n)} = \|g_{a,b,x_0}\|_{L^{p^\star}(\mathbb{R}^n)} \right\},$$

where $g_{a,b,x_0}$ is defined in (3). In words, $\lambda(f)$ is the (renormalized) distance of $f$ from the set of the optimal functions, then by definition $f$ is optimal if and only if $\lambda(f) = 0$ (the set of the optimal functions is clearly closed).

Hence, the Sobolev inequality and the results of existence and uniqueness of the minimizers can be restated by saying that $\delta(f) = 0 \iff \lambda(f) = 0$, while

the quantitative (or stability) result from [5] says that if $\delta(f)$ is small then also $\lambda(f)$ must be small –notice that the opposite implication is clearly false. More precisely, the result is the following.

**Theorem 1.2 (Quantitative Sobolev Inequality)** *There are two constants $C$ and $\beta$ such that, for any $f \in \mathrm{W}^{1,p}(\mathbb{R}^n)$, one has*

$$\lambda(f) \leq C\delta(f)^{\beta}\,.$$

*In particular, one can take*

$$\beta = \frac{1}{\xi^2 p^\star}\,, \qquad \textit{where} \qquad \xi = 3 + 4p - \frac{3p+1}{n}\,. \tag{4}$$

We will use this result to obtain the lower estimate (see Section 2.3 below).

### 1.2 Finite element setting

In this section we set up some notation and recall some technical tools we will require in our analysis. We will define the discrete space $V_h \subseteq \mathrm{W}^{1,2}_0(B)$ of piecewise linear conforming finite elements by the polygonal approximation technique [8].

We will consider an approximation of $B$ given by a family of polyhedral domains $\{B_h\}$ inscribed in $B$: the parameter $0 < h < 1$, to be specified in a moment, will go to 0. For any $h$, we construct a partition $\mathcal{T}_h$ of $B_h$, more precisely, let $\overline{B}_h = \bigcup_{T\in\mathcal{T}_h} \overline{T}$, where each $T$ is the image of the reference tetrahedron $\widehat{T}$ in $\mathbb{R}^3$ through an affine linear mapping $F_T : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$, *i.e.*, $T = F_T(\widehat{T})$ for any $T \in \mathcal{T}_h$. We choose partitions $\mathcal{T}_h$ that are *regular* (see, for instance, [6]): this means that there exists a constant $\sigma > 0$ such that

$$\frac{h_T}{\rho_T} \leq \sigma \quad \forall T \in \mathcal{T}_h\,, \tag{5}$$

where $\rho_T$ is the radius of the biggest ball contained in $T$, and $h_T$ is the diameter of $T$. Our meshes are also *uniform*, *i.e.*, setting $h = \max_{T\in\mathcal{T}_h} h_T$, the ratio $h/h_T$ is uniformly bounded. Notice that, since $B$ is smooth, convex and the boundary vertices of $\overline{B}_h$ lie on $\partial B$, $B_h$ can be constructed in such a way that $|B \setminus B_h| \lesssim h^2$.

Next, for a fixed $\mathcal{T}_h$, we define the space $\widetilde{V}_h$ as

$$\widetilde{V}_h = \{f \in H^1_0(\overline{B}_h) \,:\, f \circ F_T \in \mathbb{P}^1(\widehat{T}) \quad \forall\, T \in \mathcal{T}_h\},$$

where $\mathbb{P}^1(\widehat{T})$ is the space of linear polynomials on $\widehat{T}$. Finally, we set $V_h$ to be the space of functions in $\widetilde{V}_h$ extended by zero in the skin $B \setminus B_h$. Notice that $V_h$ is a finite dimensional vectorial space, since any $f \in V_h$ is univocally determined by the values of $f$ at the internal vertices of the mesh (called interpolation nodes).

Observe also that $V_h \subseteq \mathrm{W}_0^{1,2}(B)$. Finally, we define the interpolation operator $\Pi_h^1 : C^0(\overline{B}) \longrightarrow V_h$ as

$$\Pi_h^1 f(x_i) = f(x_i),$$

where $x_i$ are the interpolation nodes. We recall the following classical result (see, for example, [10]): there exists a constant $C > 0$, independent of $h$, such that:

$$\|D(\Pi_h^1 f - f)\|_{L^2(\Omega)} \leq Ch\|D^2 f\|_{L^2(\Omega)}, \tag{6}$$

for any $f \in H^2(B)$.

**Remark 1.4** *There are many alternatives to the space $V_h$ we are considering as, for example, hexaedra, prisms or isoparametric elements (see, for example, [3, Sect. 4.7]): it is of primary importance that standard interpolation estimates as (6) hold. Our analysis could also be applied to other choices of approximation spaces.*

## 2 Proof of the main result

This section is devoted to show the main result, Theorem 1.1. We first recall some preliminary results, next we prove the upper estimate (see Proposition 2.1), and the lower estimate (see Proposition 2.2) which complete the proof of the theorem.

### 2.1 Preliminary results

In this section we list a couple of well-known results which we will need later. We point out that they are valid for any $n$ and any $1 < p < n$.

Let us start by taking any open set $\Omega \subseteq \mathbb{R}^n$: one may ask if there is a version of Sobolev inequality which holds also inside $\Omega$. This should mean that there exists a constant $S(p, n, \Omega)$ such that

$$\|Df\|_{L^p(\Omega)} \geq S(p, n, \Omega)\|f\|_{L^{p^\star}(\Omega)} \qquad \forall f \in \mathrm{W}_0^{1,p}(\Omega). \tag{7}$$

Notice that the right space is $\mathrm{W}_0^{1,p}(\Omega)$ instead of $\mathrm{W}^{1,p}(\Omega)$, because in the latter all the constant functions show that Sobolev inequality is not true, at least when $\Omega$ has finite measure.

**Lemma 2.1** *Inequality (7) holds true for all functions $f \in \mathrm{W}_0^{1,p}(\Omega)$. Moreover, the optimal constant in the inequality is $S(p, n, \Omega) = S(p, n)$. Finally, the inequality is strict for any non-zero function $f \in \mathrm{W}_0^{1,p}(\Omega)$, unless $\Omega = \mathbb{R}^n$.*

**Proof.** Since $\mathrm{W}_0^{1,p}(\Omega) \subseteq \mathrm{W}^{1,p}(\mathbb{R}^n)$ via the extension to 0 out of $\Omega$, for any function $f \in \mathrm{W}_0^{1,p}(\Omega)$ we already know that (1) holds true, so inequality (7) is valid with $S(p, n)$ which implies $S(p, n, \Omega) \geq S(p, n)$.

On the other hand, fix a radius $\rho > 0$ and consider the ball $B_\rho$ centered at 0 and with radius $\rho$. For any $\varepsilon > 0$, then, define the function

$$\zeta_\varepsilon(x) = \frac{1}{\left(1 + \frac{1}{\varepsilon}\,|x|^{p'}\right)^{(n-p)/p}} - \frac{1}{\left(1 + \frac{1}{\varepsilon}\,\rho^{p'}\right)^{(n-p)/p}} \;:$$

this is a smooth function on $B_\rho$ which vanishes on the boundary, so that

$$S(p,n,B_\rho) \le \frac{\|D\zeta_\varepsilon\|_{L^p(\Omega)}}{\|\zeta_\varepsilon\|_{L^{p^\star}(\Omega)}}\,.$$

Recalling now formula (3) for the optimal functions on $\mathbb{R}^n$, it is immediate to realize that

$$\frac{\|D\zeta_\varepsilon\|_{L^p(\Omega)}}{\|\zeta_\varepsilon\|_{L^{p^\star}(\Omega)}} - S(n,p) = \frac{\|D\zeta_\varepsilon\|_{L^p(\Omega)}}{\|\zeta_\varepsilon\|_{L^{p^\star}(\Omega)}} - \frac{\|Dg_{1,1/\varepsilon,0}\|_{L^p(\Omega)}}{\|g_{1,1/\varepsilon,0}\|_{L^{p^\star}(\Omega)}} \xrightarrow[\varepsilon\to 0]{} 0\,.$$

This implies that, for any $\rho > 0$, one has the equality $S(p,n,B_\rho) = S(p,n)$. Since the map $S(p,n,\cdot)$ is clearly decreasing with respect to the inclusion of sets and is not effected by a translation, and since any open set contains a ball, we deduce the equality $S(p,n,\Omega) = S(p,n)$ for all open sets $\Omega$.

Finally, suppose that there exists $\Omega \subseteq \mathbb{R}^n$ and $f \in \mathrm{W}^{1,p}_0(\Omega)$ such that

$$\frac{\|Df\|_{L^p(\Omega)}}{\|f\|_{L^{p^\star}(\Omega)}} = S(n,p)\,.$$

Then, still denoting by $f$ the extension to 0 out of $\Omega$, which belongs to $\mathrm{W}^{1,p}(\mathbb{R}^n)$, one has

$$\frac{\|Df\|_{L^p(\mathbb{R}^n)}}{\|f\|_{L^{p^\star}(\mathbb{R}^n)}} = \frac{\|Df\|_{L^p(\Omega)}}{\|f\|_{L^{p^\star}(\Omega)}} = S(n,p)$$

hence $f$ is optimal for the Sobolev inequality in $\mathbb{R}^n$. By the existence-uniqueness result of the optimizers, it must be $f = g_{a,b,x_0}$ for some suitable $a$, $b$, $x_0$. And since all the optimal functions have the whole $\mathbb{R}^n$ as support, we deduce that $\Omega = \mathbb{R}^n$, so for any other set the constant $S(p,n,\Omega)$ is an infimum but not a minimum and the thesis is achieved. □

We give now the definition of the radial symmetrization for functions.

**Definition 2.1** *For $0 \le f \in \mathrm{W}^{1,p}(\mathbb{R}^n)$, we define* radially symmetric rearrangement *of $f$ the radially symmetric decreasing function $f^\star : \mathbb{R}^n \to \mathbb{R}^+$ such that*

$$\left|\left\{x \in \mathbb{R}^n :\, f(x) > \rho\right\}\right| = \left|\left\{x \in \mathbb{R}^n :\, f^\star(x) > \rho\right\}\right| \qquad \forall \rho > 0\,.$$

The following property of the radial symmetrization is well known (refer to [9]).

**Theorem 2.1 (Polya–Szegö)** *For any $0 \le f \in \mathrm{W}^{1,p}(\mathbb{R}^n)$, one has $f^\star \in \mathrm{W}^{1,p}(\mathbb{R}^n)$ and*

$$\|f^\star\|_{L^p(\mathbb{R}^n)} = \|f\|_{L^p(\mathbb{R}^n)}\,, \qquad \|Df^\star\|_{L^p(\mathbb{R}^n)} \le \|Df\|_{L^p(\mathbb{R}^n)}\,.$$

Let us immediately notice the important consequence that Polya–Szegö Theorem has when studying the Sobolev inequality (1). Assume for a moment that we are looking for an optimizer of the inequality, and assume of course that we still don't know the exact formula (3): then, Polya–Szegö Theorem immediately suggests us to restrict our attention to radially symmetric decreasing function, which is extremely useful since it basically means to study one-dimensional decreasing functions instead of $n-$dimensional generic ones. Indeed, assume that $f$ is optimal for the Sobolev inequality: then,

$$\|Df^\star\|_{L^p(\mathbb{R}^n)} \leq \|Df\|_{L^p(\mathbb{R}^n)} = S(p,n)\|f\|_{L^{p^\star}(\mathbb{R}^n)} = S(p,n)\|f^\star\|_{L^{p^\star}(\mathbb{R}^n)}\,,$$

which means that also the radially symmetric decreasing function $f^\star$ is optimal.

We conclude with a useful notation that we will use extensively in the following.

**Definition 2.2** *Let $p = 2$, $n = 3$. Then, for any $a > 0$ we denote by $T_a$ the function $T_a = g_{a,b,x_0}$ in the sense of (3), where $x_0 \equiv 0$, and $b = b(a)$ is chosen so that*

$$\|T_a\|_{L^6(\mathbb{R}^3)} = 1\,.$$

Notice that the above definition is correct, since $b \mapsto \|g_{a,b,0}\|_{L^6(\mathbb{R}^3)}$ is a continuous and strictly decreasing function from $(0,+\infty)$ to itself, which tends to 0 (resp. $+\infty$) when $b$ goes to $+\infty$ (resp. 0).

## 2.2 Upper estimate

In this section we will show the upper estimate.

**Proposition 2.1** *There exists a constant $C$ such that $S_h \leq S + Ch^{1/3}$.*

**Proof.** We divide the proof in four steps.

**Step I.** *Setting of the main function.*
Let us fix a number $\alpha \in \mathbb{R}^+$, to be precised later, and set

$$a := \frac{1}{h^\alpha}\,.$$

According to Definition 2.2, $b = b(a)$ is defined in such a way that

$$\begin{aligned} 1 = \|g_{a,b,0}\|^6_{L^6(\mathbb{R}^3)} &= \int_{\rho=0}^{+\infty} \frac{a^6}{\left(1+b\rho^2\right)^3}\, 4\pi\rho^2\, d\rho \\ &= \frac{4\pi}{h^{6\alpha}} \int_{\rho=0}^{+\infty} \frac{1}{\left(1+b\rho^2\right)^3}\, \rho^2\, d\rho = \frac{\pi^2}{4h^{6\alpha}b^{3/2}}\,, \end{aligned}$$

so that we derive that

$$b = \frac{1}{h^{4\alpha}} \left(\frac{\pi}{2}\right)^{4/3}. \tag{8}$$

Let us now consider the function $\widetilde{T}_a \in \mathrm{W}^{1,2}_0(B)$ defined as

$$\widetilde{T}_a(x) = T_a(x) - T_a(1)\,,$$

where, with an abuse of notation, we have denoted by $T_a(1)$ the constant value of the (radially symmetric) function $T_a$ on the set $\{x \in \mathbb{R}^n : |x| = 1\}$. An immediate calculation tells us that

$$T_a(1) = \frac{a}{\sqrt{1+b}} = \frac{1}{h^\alpha\sqrt{1+(\pi/2)^{4/3}h^{-4\alpha}}} = \left(\frac{2}{\pi}\right)^{2/3} h^\alpha + O(h^{5\alpha}) \tag{9}$$

for $h \searrow 0$. We will show our bound on $S_h$ by making use of the function $\Pi^1_h\widetilde{T}_a$, which course belongs to $V_h$ by definition. Indeed, one has that

$$S_h \le \frac{\|D\Pi^1_h\big(\widetilde{T}_a\big)\|_{L^2(B)}}{\|\Pi^1_h\big(\widetilde{T}_a\big)\|_{L^6(B)}}\,. \tag{10}$$

In view of the interpolation estimate (6), it is clear that we need an upper bound for $\|D^2\widetilde{T}_a\|_{L^2(B)}$, an upper bound for $\|D\widetilde{T}_a\|_{L^2(B)}$, and a lower bound for $\|\widetilde{T}_a\|_{L^6(B)}$. The first one will be obtained in Step III, while for the second one it is enough to notice that

$$\|D\widetilde{T}_a\|_{L^2(B)} = \|DT_a\|_{L^2(B)} \le \|DT_a\|_{L^2(\mathbb{R}^3)} = S\,. \tag{11}$$

Finally, for the lower bound for $\|\widetilde{T}_a\|_{L^6(B)}$, we will use the fact that, since $\|T_a\|_{L^6(B)} \le \|T_a\|_{L^6(\mathbb{R}^3)} = 1$, one has

$$\|\widetilde{T}_a\|^6_{L^6(B)} \ge 1 - K_1 - K_2\,, \tag{12}$$

having defined

$$K_1 := \|T_a\|^6_{L^6(\mathbb{R}^3\setminus B)}\,, \qquad\qquad K_2 := 1 - \frac{\|\widetilde{T}_a\|^6_{L^6(B)}}{\|T_a\|^6_{L^6(B)}}\,.$$

In Step II we will estimate $\|\widetilde{T}_a\|_{L^6(B)}$ by giving bounds to $K_1$ and $K_2$.

**Step II.** *Estimate on* $\|\widetilde{T}_a\|_{L^6(B)}$.
In the set $\{|x| \ge 1\}$ one has

$$T_a(x) = \frac{a}{\sqrt{1+b|x|^2}} = h^\alpha\left(\frac{2}{\pi}\right)^{2/3}\frac{1}{|x|}\left(1+O(h^{4\alpha})\right).$$

Hence, one has

$$\begin{aligned} K_1 = \|T_a\|^6_{L^6(\mathbb{R}^3\setminus B)} &= \left(1+O(h^{4\alpha})\right)\int_{\mathbb{R}^3\setminus B} h^{6\alpha}\left(\frac{2}{\pi}\right)^4\frac{1}{|x|^6}\,dx \\ &= \frac{2^6}{3\pi^3}h^{6\alpha} + O(h^{10\alpha})\,. \end{aligned} \tag{13}$$

Thus, we obtained the estimate for $K_1$.

Concerning $K_2$, it is convenient to divide the unit ball $B$ in the internal ball $B_I = \{x \in B : T_a(x) \ge 1\}$ and the external part $B_E = B \setminus B_I$, and treating the two regions in a different way. Let us start taking $x \in B_E$: then, being $T_a(x) \le 1$, one has

$$\begin{aligned} \widetilde{T}_a(x)^6 &= \Big(T_a(x) - T_a(1)\Big)^6 = T_a(x)^6\left(1 - \frac{T_a(1)}{T_a(x)}\right)^6 \ge T_a(x)^6\left(1 - 6\,\frac{T_a(1)}{T_a(x)}\right) \\ &= T_a(x)^6 - 6T_a(1)T_a(x)^5 \ge T_a(x)^6 - 6T_a(1)\,, \end{aligned}$$

from which we deduce

$$\begin{aligned}\|T_a\|^6_{L^6(B_E)} - \|\widetilde{T}_a\|^6_{L^6(B_E)} &= \int_{B_E} T_a(x)^6 - \widetilde{T}_a(x)^6\,dx \\ &\leq 8\pi\left(\frac{2}{\pi}\right)^{2/3} h^\alpha + O(h^{5\alpha})\,,\end{aligned} \tag{14}$$

recalling (9). On the other hand, if $x \in B_I$, $T_a(x) \geq 1$ and we have

$$\begin{aligned}\widetilde{T}_a(x)^6 &= \Big(T_a(x) - T_a(1)\Big)^6 = T_a(x)^6\left(1 - \frac{T_a(1)}{T_a(x)}\right)^6 \geq T_a(x)^6\left(1 - 6\,\frac{T_a(1)}{T_a(x)}\right) \\ &\geq T_a(x)^6\Big(1 - 6T_a(1)\Big)\,,\end{aligned}$$

which gives

$$\begin{aligned}\|T_a\|^6_{L^6(B_I)} - \|\widetilde{T}_a\|^6_{L^6(B_I)} &= \int_{B_I} T_a(x)^6 - \widetilde{T}_a(x)^6\,dx \\ &\leq 6T_a(1)\int_{B_I} T_a(x)^6\,dx \\ &\leq 6\,\|T_a\|^6_{L^6(B)}\left(\frac{2}{\pi}\right)^{2/3} h^\alpha + O(h^{5\alpha})\,,\end{aligned} \tag{15}$$

again recalling (9). Moreover, notice that by (13) it is

$$\|T_a\|^6_{L^6(B)} = 1 - \|T_a\|^6_{L^6(\mathbb{R}^3\setminus B)} = 1 - \frac{h^{6\alpha}2^6}{3\pi^3} + O(h^{10\alpha}) = 1 + O(h^{6\alpha})\,. \tag{16}$$

Finally, putting together (14), (15) and (16), we obtain the estimate for $K_2$

$$\begin{aligned}K_2 &= 1 - \frac{\|\widetilde{T}_a\|^6_{L^6(B)}}{\|T_a\|^6_{L^6(B)}} = \frac{\|T_a\|^6_{L^6(B_E)} - \|\widetilde{T}_a\|^6_{L^6(B_E)} + \|T_a\|^6_{L^6(B_I)} - \|\widetilde{T}_a\|^6_{L^6(B_I)}}{\|T_a\|^6_{L^6(B)}} \\ &\leq 8\pi\left(\frac{2}{\pi}\right)^{2/3} h^\alpha + 6\,\|T_a\|^6_{L^6(B)}\left(\frac{2}{\pi}\right)^{2/3} h^\alpha + O(h^{5\alpha}) \\ &= \big(8\pi + 6\big)\left(\frac{2}{\pi}\right)^{2/3} h^\alpha + O(h^{5\alpha})\,.\end{aligned} \tag{17}$$

Recalling (12), from (13) and (17), we finally obtain

$$\|\widetilde{T}_a\|^6_{L^6(B)} \geq 1 - \big(8\pi + 6\big)\left(\frac{2}{\pi}\right)^{2/3} h^\alpha + O(h^{5\alpha})\,. \tag{18}$$

**Step III.** *Estimate on* $\|D^2\widetilde{T}_a\|_{L^2(B)}$.
To estimate the semi-norm $\|D\widetilde{T}_a\|_{L^2(B)}$ we start noticing that, since

$$T_a(x) = \frac{a}{\sqrt{1 + b|x|^2}} =: \varphi(|x|)\,,$$

one has

$$DT_a(x) = \varphi'(|x|)\,\frac{x}{|x|} = -\frac{ab|x|}{\big(1 + b|x|^2\big)^{3/2}}\,\frac{x}{|x|}\,,$$

and

$$D^2_{ij}T_a(x) = \varphi''(|x|)\frac{x_i x_j}{|x|^2} + \varphi'(|x|)\,\frac{|x|^2\delta_{ij} - x_i x_j}{|x|^3}\,.$$

Therefore,

$$\begin{aligned}\|D^2\widetilde{T}_a\|^2_{L^2(B)} = \|D^2 T_a\|_{L^2(B)} &= \int_B \left|D^2T_a(x)\right|^2 dx \approx \int_B \varphi''(|x|)^2 + \frac{\varphi'(|x|)^2}{|x|^2}\,dx \\ &\approx a^2b^2\int_B \frac{1}{\left(1+b|x|^2\right)^3}\,dx = a^2b^2\int_{t=0}^1 \frac{4\pi t^2}{\left(1+bt^2\right)^3}\,dt \approx a^2\sqrt{b} \\ &\approx \frac{1}{h^{4\alpha}}\,.\end{aligned} \tag{19}$$

**Step IV.** *Conclusion.*
We can now conclude: by (6) and (19) we get

$$\begin{aligned}\|D\Pi^1_h\big(\widetilde{T}_a\big) - D\widetilde{T}_a\|_{L^2(B)} = \big|\Pi^1_h(\widetilde{T}_a) - \widetilde{T}_a\big|_1 &\le Ch\|D^2\widetilde{T}_a\|_{L^2(B)} \\ &\le Ch^{1-2\alpha}\,.\end{aligned} \tag{20}$$

Since the space $V_h$ is contained in $\mathrm{W}^{1,2}_0(B)$, by Sobolev embeddings we also have that

$$\|\Pi^1_h\big(\widetilde{T}_a\big) - \widetilde{T}_a\|_{L^6(B)} \le C\big|\Pi^1_h\big(\widetilde{T}_a\big) - \widetilde{T}_a\big|_{2,B} \le Ch^{1-2\alpha}\,. \tag{21}$$

Finally, putting together the estimates (11), (18), (20) and (21), and using (10), we immediately get

$$S_h \le S + Ch^{1-2\alpha} + Ch^{\alpha}\,.$$

We then derive that the best choice for $\alpha$ is $\alpha = 1/3$, which leads to the thesis. □

## 2.3 Lower estimate

In this section we will show the lower estimate.

**Proposition 2.2** *There exist two positive constants $C$ and $\gamma$ such that*

$$S_h \ge S + \frac{1}{C}\,h^{\gamma}\,.$$

**Proof.** We recall that

$$S_h = \inf_{f\in V_h} \frac{\|Df\|_{L^2(B)}}{\|f\|_{L^6(B)}}\,,$$

so that, since $V_h$ is a finite dimension space and the ratio is an invariant if we multiply $f$ by a constant, the infimum is realized by some function $f_h \in V_h$ with $\|f_h\|_{L^6(B)} = 1$: that is,

$$S_h = \frac{\|Df_h\|_{L^2(B)}}{\|f_h\|_{L^6(B)}} = \|Df_h\|_{L^2(B)}\,.$$

For simplicity, let us still denote by $f_h$ its extension by 0 on $\mathbb{R}^n \setminus B$, which belongs to $\mathrm{W}^{1,2}(\mathbb{R}^n)$. By applying the quantitative Sobolev inequality (Theorem 1.2) to the function $f_h$, we deduce the existence of an optimal function $G = g_{a,b,x_0}$ such that

$$\|G\|_{L^6(\mathbb{R}^3)} = \|f_h\|_{L^6(\mathbb{R}^3)} = \|f_h\|_{L^6(B)} = 1\,,$$

and

$$\|f_h - G\|_{L^6(\mathbb{R}^3)} = \lambda(f_h) \le C\delta(f_h)^\beta = C(S_h - S)^\beta \,. \tag{22}$$

Then, to get a lower estimate for $S_h$, we will try to estimate from below the term $\|f_h - G\|_{L^6(\mathbb{R}^3)}$. It will be useful, in analogy with Proposition 2.1, to define $\alpha = \alpha(h)$ so that $a = 1/h^\alpha$ (recall that $a$, $b$ and $x_0$ are fixed since $G = g_{a,b,x_0}$). We fix now $\varepsilon > 0$ and we divide two cases, namely whether $\alpha$ is bigger or smaller than $1+\varepsilon$.

**Case I.** *If* $\alpha \le 1+\varepsilon$.
In this case, keep in mind the estimate (13) from Proposition 2.1, since in that construction we had by definition $x_0 = 0$, the estimate tells us that

$$\|G\|_{L^6(\mathbb{R}^3\setminus B_0)} \ge \frac{1}{C}\, h^\alpha \,,$$

where $B_0 = \{x \in \mathbb{R}^3 : |x - x_0| \le 1\}$. Moreover, being $G$ a radially symmetric decreasing function, one clearly has

$$\begin{aligned} \|f_h - G\|_{L^6(\mathbb{R}^3)} &\ge \|f_h - G\|_{L^6(\mathbb{R}^3\setminus B)} = \|G\|_{L^6(\mathbb{R}^3\setminus B)} \\ &\ge \|G\|_{L^6(\mathbb{R}^3\setminus B_0)} \ge \frac{1}{C}\, h^\alpha \ge \frac{1}{C}\, h^{1+\varepsilon} \,. \end{aligned} \tag{23}$$

Notice that to get this estimate we did not really use the assumption $\alpha \le 1+\varepsilon$, except of course in the last inequality: the estimate $\|f_h - G\|_{L^6} \ge h^\alpha/C$ is true for any value of $\alpha$, but it is interesting for our purpose only if $\alpha$ is big enough.

**Case II.** *If* $\alpha \ge 1+\varepsilon$.
In this second case we start noticing that, being $\|G\|_{L^6(\mathbb{R}^3)} = 1$, formula (8) still holds for $b$. Moreover, since we already know that $S_h - S \searrow 0$, then by (22)

$$\begin{aligned} 1 \gg C\big(S_h - S\big)^\beta &\ge \|f_h - G\|_{L^6(\mathbb{R}^3)} \ge \|f_h - G\|_{L^6(B)} \\ &\ge \|f_h\|_{L^6(B)} - \|G\|_{L^6(B)} = \|G\|_{L^6(\mathbb{R}^3\setminus B)} \,. \end{aligned}$$

This immediately implies that $x_0 \in B$. Let then $T^* \in \mathcal{T}_h$ be the tetrahedron containing $x_0$, and let $\tilde h$ be defined so that, splitting $T^*$ into the two parts $T_1$ and $T_2$ given by

$$T_1 := T^* \cap \{x \in \mathbb{R}^3 :\, |x - x_0| \le \tilde h\} \,, \qquad\qquad T_2 := T^* \setminus T_1 \,,$$

one has

$$\big|T_1\big| = \big|T_2\big| = \frac{\big|T^*\big|}{2} \,.$$

Since the mesh is *regular* and *uniform* (cf. Section 1.2), and $x_0 \in T^*$, it is

$$h \ge \tilde h \ge \frac{1}{C}\, h_{T^*} \ge \frac{1}{C}\, h \,.$$

Notice now that, by the formula (3) for $G$, for any $x$ such that $|x - x_0| \ge \tilde h$ one has, again using (8),

$$G(x) \le \frac{a}{\sqrt{1 + b\tilde h^2}} \approx h^{\alpha-1} \ll 1 \,, \tag{24}$$

since in the present case $\alpha \ge 1+\varepsilon$. We now use again the estimate (13), which tells us that

$$\|G\|_{L^6(\mathbb{R}^3\setminus B_0)} \le Ch^\alpha \,.$$

Moreover, writing $\widetilde{B} = \{x : \tilde{h} \leq |x - x_0| \leq 1\}$, one has also

$$\|G\|_{L^6(\tilde{B})} \leq \|G\|_{L^\infty(\tilde{B})} \left|\widetilde{B}\right|^{1/6} \leq Ch^{\alpha-1} \leq Ch^{\varepsilon}\,,$$

thanks to (24). Summarizing, the assumption $\alpha \geq 1+\varepsilon$ leads us to deduce that, defining $\mathcal{B}_h = \{x : |x - x_0| \leq \tilde{h}\}$,

$$\|G\|_{L^6(\mathcal{B}_h)} = \sqrt[6]{1 - \|G\|^6_{L^6(\mathbb{R}^3\setminus B_0)} - \|G\|^6_{L^6(\tilde{B})}} \approx 1\,.$$

Since the mesh is regular, this implies the existence of a positive constant $C^*$, depending only on the shape regularity constant of the mesh, such that

$$\|G\|_{L^6(T_1)} \geq C^*\,. \tag{25}$$

On the other hand, it is of course

$$\|G\|_{L^6(T_2)} \leq \|G\|_{L^6(\mathbb{R}^3\setminus\mathcal{B}_h)} \ll 1\,. \tag{26}$$

Notice now that, by an easy geometrical argument, there exists a geometric constant $\widetilde{C}$, depending only on the mesh, such that for any function $v \in V_h$ one has

$$\frac{1}{\widetilde{C}}\,\|v\|_{L^6(T_1)} \leq \|v\|_{L^6(T_2)} \leq \widetilde{C}\|v\|_{L^6(T_1)}\,. \tag{27}$$

It is now simple to guess that (25), (26) and (27) will lead to a lower bound for $\|f_h - G\|_{L^6(\mathbb{R}^3)}$. Indeed,

$$\|f_h - G\|^6_{L^6(\mathbb{R}^3)} \geq \|f_h - G\|^6_{L^6(T_1)} + \|f_h - G\|^6_{L^6(T_2)}\ :$$

then, if

$$\|f_h\|_{L^6(T_1)} \leq \frac{C^*}{2}\,,$$

we have

$$\|f_h - G\|^6_{L^6(\mathbb{R}^3)} \geq \|f_h - G\|^6_{L^6(T_1)} \geq \left(\frac{C^*}{2}\right)^6\,.$$

On the other hand, if

$$\|f_h\|_{L^6(T_1)} \geq \frac{C^*}{2}\,,$$

then

$$\|f_h - G\|^6_{L^6(\mathbb{R}^3)} \geq \|f_h - G\|^6_{L^6(T_2)} \geq \left|\|f_h\|_{L^6(T_2)} - \|G\|_{L^6(T_2)}\right|^6 \geq \left(\frac{C^*}{3\widetilde{C}}\right)^6$$

for $h \ll 1$. We can then conclude by saying that, in the case $\alpha \geq 1+\varepsilon$, there is a constant $\widehat{C}$ so that

$$\|f_h - G\|_{L^6(\mathbb{R}^3)} \geq \widehat{C} \tag{28}$$

for $h \ll 1$. Notice that the constant $\widehat{C}$, which is formally given by

$$\widehat{C} = \min\left\{\left(\frac{C^*}{2}\right)^6,\ \left(\frac{C^*}{3\widetilde{C}}\right)^6\right\},$$

does *not* depend on $\varepsilon$. What depends on $\varepsilon$ is how small $h$ needs to be in order the estimate (28) to hold true.

We can finally conclude the proof. By (23) and (28) we know that, in any case, if $h \ll 1$ then

$$\|f_h - G\|_{L^6(\mathbb{R}^3)} \geq \frac{1}{C}\, h^{1+\varepsilon} \ :$$

by (22), then, we have

$$S_h - S \geq \frac{1}{C}\, h^{(1+\varepsilon)/\beta}\,.$$

Thus, recalling formula (4) for $\beta$, the thesis is obtained for any

$$\gamma > \frac{1}{\beta} = \frac{2 \cdot 26^2}{3}\,.$$

Notice that, if $\gamma \searrow 1/\beta$, then the corresponding $C$ goes to $+\infty$. □

# 3 Dimensional reduction

Since a numerical estimate for a three-dimensional problem would be extremely slow and fairly accurate, in this section we show how to reduce our original problem to a one-dimensional one, which is meaningful since, how we already observed, the problem of finding extremals for Sobolev inequality is basically one dimensional. To do so, we will construct a suitable sequence of "spherical meshes" of the unit ball in $\mathbb{R}^3$.

## 3.1 Construction of spherical meshes

We now describe how to construct a sequence of "spherical meshes" $\mathcal{T}_{h_k}$, $k \in \mathbb{N}$, which will be made by "spherical tetrahedra". We consider the usual transformation $\Sigma : \mathbb{R}^3_{\rho,\theta,\varphi} \longrightarrow \mathbb{R}^3_{x,y,z}$ from spherical to Cartesian coordinates given by

$$\Sigma \begin{pmatrix} \rho \\ \theta \\ \varphi \end{pmatrix} := \begin{pmatrix} \rho\cos(\theta)\sin(\varphi) \\ \rho\sin(\theta)\sin(\varphi) \\ \rho\cos(\varphi) \end{pmatrix},$$

with $\rho \in \mathbb{R}^+$, $\theta \in [0, 2\pi)$ and $\varphi = [0, \pi)$. The *spherical tetrahedron* of vertices $A$,$B$,$C$,$D$ is the image under $\Sigma$ of the standard "straight" tetrahedron having vertices $A' = \Sigma^{-1}(A)$, $B' = \Sigma^{-1}(B)$, $C' = \Sigma^{-1}(C)$, $D' = \Sigma^{-1}(D)$ (see Figure 1). To obtain the spherical mesh $\mathcal{T}_{h_k}$, we will construct a standard mesh $\widehat{\mathcal{T}}_{h_k}$ made of "straight" tetrahedra; then, replacing each tetrahedron in $\widehat{\mathcal{T}}_{h_k}$ by the spherical tetrahedron with the same vertices, we get $\mathcal{T}_{h_k}$.

The meshes $\widehat{\mathcal{T}}_{h_k}$, which are shown in Figure 2 for $k = 1, \ldots, 6$, will be defined as follows.

1. We identify a finite number of concentric spheres in the ball $B$: in particular, at step $k$, we will consider all the spheres with radii $\ell/k$, $\ell = 1, \ldots, k$;

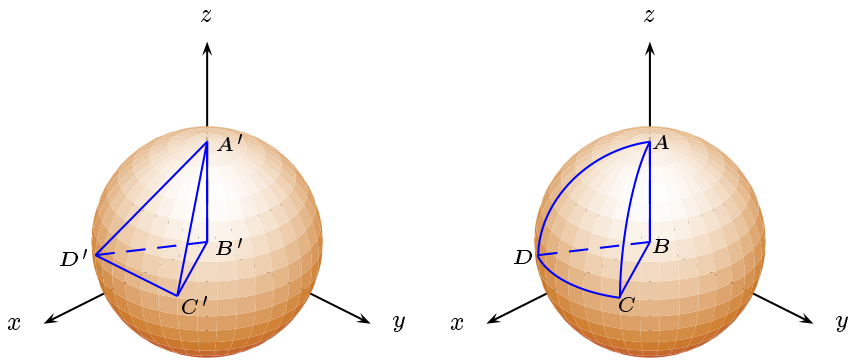



Figure 1: Sample of a *straight* and a *spherical* tetrahedron.

2. each sphere is approximated by a suitable triangular grid having all the vertices on the sphere;

3. the straight tetrahedra are obtained by suitably connecting the vertices of consecutive layers.

Let us now show how to generate a sequence $\{\mathcal{E}_\ell\}_{\ell\in\mathbb{N}}$ of shape regular and uniform triangulations approximating a given sphere $\partial B_\rho$ centered in the origin and with radius $\rho$. Our approach is similar to the one considered in [1] with some modifications due to the fact that at the end we are interested in constructing a three-dimensional mesh for the unit ball. Let $R = \{\rho\} \times [0, 2\pi) \times [0, \pi]$ be the parameter domain in the $(\rho, \theta, \varphi)$ coordinates. We consider triangulations of $R$ as the ones depicted in Figure 3 for $\ell = 1, 2, 3, 4$; for $\ell > 4$ the corresponding refinements are obtained analogously. Notice that the following properties hold:

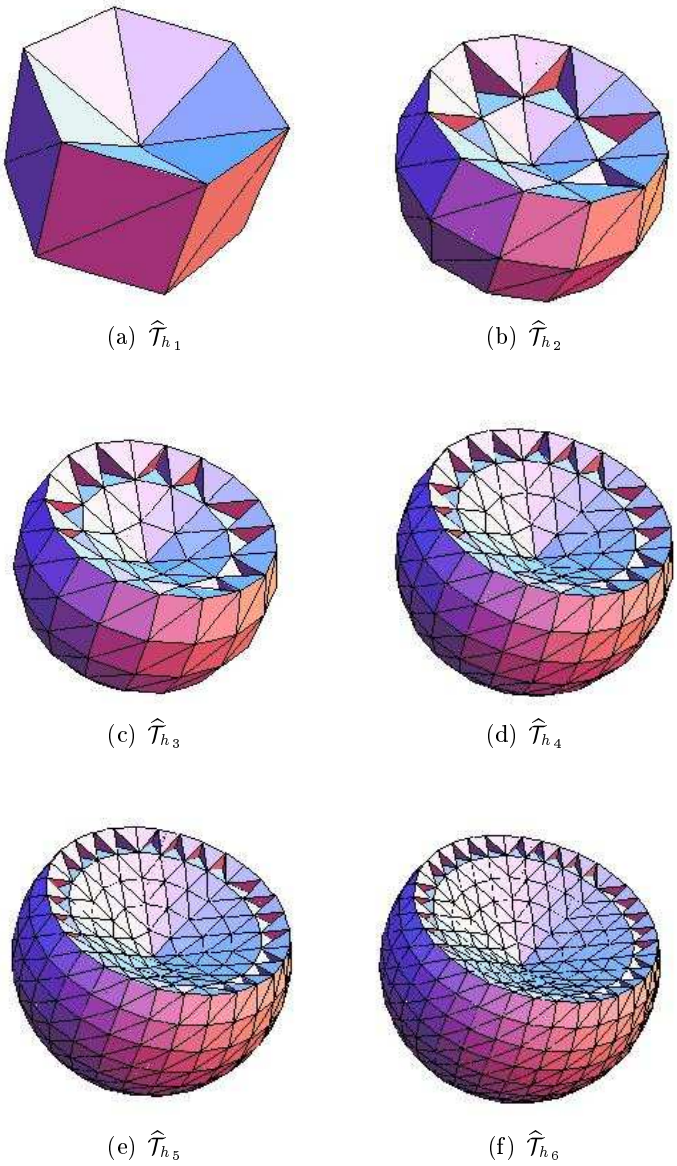
i) all the elements are triangular and have one edge parallel to the $\theta$-axis, except for pole elements (*i.e.*, the elements of the grid containing points with $\varphi$ equal to 0 or $\pi$) which are rectangular;

ii) at each pole there are six rectangles.

The searched grid which approximates $\partial B_\rho$ is then obtained as the triangular grid whose vertices are the image through $\Sigma$ of the vertices of $\mathcal{E}_\ell$ with the same connectivity matrix. Notice that this grid is made only by triangles because, when passing from spherical to cartesian coordinates, the rectangles near the poles become triangles.

Once we have constructed the sequence $\mathcal{E}_\ell$ of two-dimensional triangulations approximating $\partial B_\rho$, the corresponding three-dimensional mesh is obtain as follows:

- the initial grid $\widehat{\mathcal{T}}_{h_1}$ is obtained by constructing the mesh $\mathcal{E}_1$ with vertices lying on $\partial B_1$, and connecting all the boundary points with the origin (see Figure 2(a));



(a) $\widehat{\mathcal{T}}_{h_1}$ (b) $\widehat{\mathcal{T}}_{h_2}$

(c) $\widehat{\mathcal{T}}_{h_3}$ (d) $\widehat{\mathcal{T}}_{h_4}$

(e) $\widehat{\mathcal{T}}_{h_5}$ (f) $\widehat{\mathcal{T}}_{h_6}$

Figure 2: Sample of straight triangulations $\widehat{\mathcal{T}}_{h_k}$, $k = 1, \ldots, 6$.

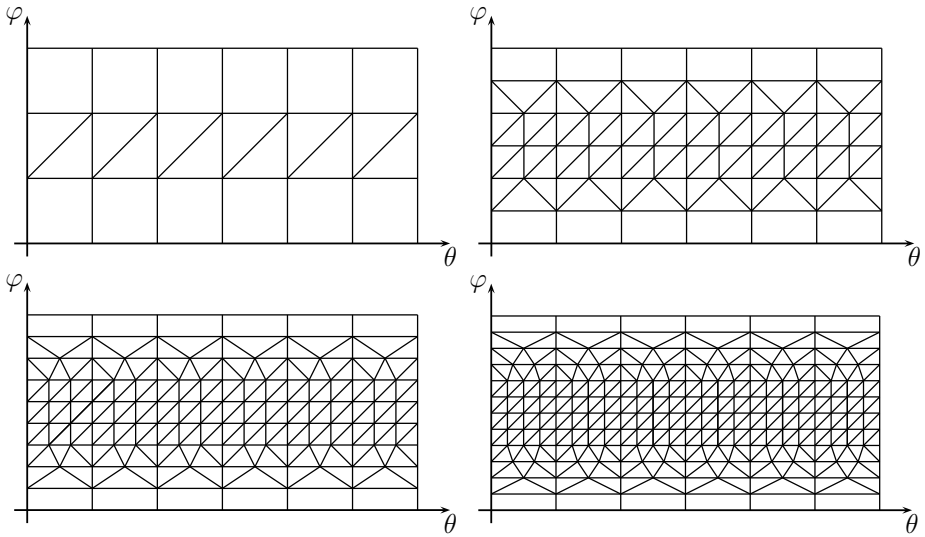



Figure 3: Triangulation of $\partial B_\rho$ in the parameter domain: levels $\ell = 1, 2$ (top), and levels $\ell = 3, 4$ (bottom).

- the second grid $\widehat{\mathcal{T}}_{h_2}$ is obtained by constructing the mesh $\mathcal{E}_1$ on $\partial B_{1/2}$ and the mesh $\mathcal{E}_2$ on $\partial B_1$. Next, we connect all the vertices on $\partial B_{1/2}$ with the origin, whereas the points on $\partial B_1$ and $\partial B_{1/2}$ are connected properly with each other to obtain a tetrahedral mesh as shown in Figure 2(b)).
- the grid $\widehat{\mathcal{T}}_{h_k}$ is obtained by constructing, for each $\ell = 1, \dots, k$, the mesh $\mathcal{E}_\ell$ on the sphere $\partial B_{\ell/k}$. Finally, the generated points that lie on two consecutive spheres are connected properly with each other to obtain a tetrahedral mesh: see Figures 2(c)–2(f)) for $k = 3, 4, 5, 6$.

The mesh $\mathcal{T}_{h_k}$ is then obtained from $\widehat{\mathcal{T}}_{h_k}$, as we said before, simply replacing all the straight tetrahedra of $\widehat{\mathcal{T}}_{h_k}$ with the spherical tetrahedra with the same vertices. Notice that this is a mesh on the whole ball $B$, not on an approximation $B_h$.

### 3.2 Reduction to a one-dimensional problem

Once we have our spherical meshes, we can consider a finite element approximation corresponding to them: more precisely, we can define the discret space $V_h \subseteq \mathrm{W}_0^{1,2}(B)$ as the set of those functions $f \in \mathrm{W}_0^{1,2}(B)$ which, for each spherical tetrahedron $T \in \mathcal{T}_h$, are affine on $T$ with respect to the spherical coordinates $\rho$, $\theta$ and $\varphi$. Notice that elements of $V_h$ are continuous and that an element of $V_h$ is completely known once one knows its values on the interpolation nodes: hence, $V_h$ is a finite-dimensional vectorial space. Notice also that, in this setting, the dimension of $V_h$ is *not* the number of the interpolation nodes which are inside the ball $B$, since not all the values of $f \in V_h$ at the interpolation nodes are

independent: this is due to the fact that the change of variables between spherical and cartesian coordinates is not one-to-one. More precisely, if a tetrahedron $T \in \mathcal{T}_h$ contains the origin and, inside $T$, $f$ is affine in $\rho$, $\theta$ and $\varphi$, it is clear that $f$ must be in fact affine only in $\rho$; hence in particular the values of $f$ at the interpolation nodes which are in the most internal sphere must be all equal. Analagously, a function $f \in V_h$ is affine only on $\rho$ and $\varphi$ in the tetrahedra which contain a polar point (*i.e.*, a point which is in the $z-$axis, or equivalently which has the $\varphi$ coordinate equal to 0 or $\pi$).

It is then possible to define the constant $S_h$ as the biggest constant for which the discrete Sobolev inequality

$$\|Df\|_{L^p(B)} \geq S_h \|f\|_{L^{p^\star}(B)} \qquad\qquad \forall f \in V_h$$

holds. The situation is completely analogous to the problem considered in the first sections, the only difference being the fact that meshes are now spherical instead of straight. In particular, the result of Theorem 1.1 can be proved in a completely similar way in this new setting. However, the problem is now easier to handle with since the spherical structure is better in order to approximate a problem which has radially symmetric solutions. Being more precise, let us call

$$\widehat{V}_h = \left\{ f \in V_h : f \text{ is radially symmetric} \right\}.$$

This set is not empty thanks to the fact that the mesh is made by spherical tetrahedra and the elements of $V_h$ are affine in the spherical coordinates: on the other hand, in the standard "straight" setting of the first sections there were no functions in $V_h$ which are radially symmetric (except the null function)! Notice that $\widehat{V}_h$ corresponds to all and only the functions of $V_h$ which have the same value at all the interpolation nodes having a given distance from the origin. Being $\widehat{V}_h$ a subspace of $V_h$, it is clear the $\widehat{S}_h \geq S_h$, where $\widehat{S}_h$ is of course the biggest constant for which one has

$$\|Df\|_{L^p(B)} \geq \widehat{S}_h \|f\|_{L^{p^\star}(B)} \qquad\qquad \forall f \in \widehat{V}_h \,.$$

Therefore, in order to check numerically the validity of our estimate $S_h \leq S + Ch^{1/3}$, it is enough to work with $\widehat{S}_h$ instead of $S_h$ (by the way, recalling Polya–Szegö Theorem 2.1 it is easy to guess that indeed $\widehat{S}_h - S \approx S_h - S$, so that in fact we will also check the lower estimate $S_h \geq S + C^{-1}h^\gamma$).

Finally, we can notice that, as anticipated before, the problem of evaluating numerically $\widehat{S}_h$ is much faster and more efficient than evaluating $S_h$: for a radially symmetric function $f(x) = u(|x|)$, indeed, one has clearly

$$\int_B |f(x)|^6 \, dx = \int_0^1 4\pi\rho^2 |u(\rho)|^6 \, d\rho \,,$$

and analogously

$$\int_B |Df(x)|^2 \, dx = \int_0^1 4\pi\rho^2 |u'(\rho)|^2 \, d\rho \;:$$

hence, the three-dimensional problem corresponding to $f$, which involves three-dimensional integrals, has reduced to a one-dimensional problem corresponding to $u$ and involving one-dimensional integrals.

In the next section, then, we are going to show our numerical results for this one-dimensional problem.

# 4 Numerical results

In this section we present some numerical results to validate our theoretical estimates. Since we have reduced ourselves to a one–dimensional problem, we are allowed to take our computational domain as $I = [0, 1]$. More precisely, associating to any radially symmetric function $f : B \to \mathbb{R}$ the corresponding $u : I \to \mathbb{R}$ so that $f(x) = u(|x|)$, we perform our numerical study working on $u$ instead of $f$. To this aim, let $\{\mathcal{T}_N\}_{N\geq 2}$ be a sequence of partitions of $I$ made by $N$ subintervals $I_i = [x_{i-1}, x_i]$, $i = 1, \dots, N$, with corresponding mesh size $h_N = \max_i |I_i|$.

Since we are going also to consider non-uniform and adaptively refined grids (cf. Section 4.2 and Section 4.3, below), the mesh size $h_N$ of the mesh is not the right parameter to study the behavior of the approximation error, being the number $N$ of intervals the correct one (as, of course, the time needed to get the numerical results only depends on $N$). For this reason, from now on we will call $S_N$ the approximation of the Sobolev constant $S$ on a grid $\mathcal{T}_N$ of $N$ intervals. Recall that for an equispaced grid with $N$ elements we have $h_N = N^{-1}$, saying that estimates in Theorem 1.1 can be restated as

$$S + \frac{1}{C}\left(\frac{1}{N}\right)^{\gamma} \leq S_h \leq S + C\left(\frac{1}{N}\right)^{1/3},$$

for two constants $C$, $\gamma > 0$.

For a fixed partition $\mathcal{T}_N$, we denote by $V_N$ the finite element space associated to $\mathcal{T}_N$, that reads now

$$V_N = \{u \in C^0(I) \,:\, u|_{I_i} \in \mathbb{P}^1(I_i) \quad \forall I_i \in \mathcal{T}_N, \quad u(1) = 0\}.$$

To represent functions in $V_N$ we use the standard set of Lagrange "hat" basis functions, *i.e.*, $V_N = \text{span}\{\chi_i,\ i = 0, \dots, N-1\}$, where $\chi_i(x_j) = \delta_{ij}$ for $j = 0, \dots, N-1$. Therefore, we write any $u_N \in V_N$ as $u_N = \sum_{i=0}^{N-1} u_i \chi_i$, and collect the expansion coefficients $u_i$, $i = 0, \dots, N-1$, in the vector $\mathbf{u} \in \mathbb{R}^N$.

We must consider, then, the following constrained optimization problem: find $\mathbf{u} \in \mathbb{R}^N$ realizing

$$\min_{\mathbf{u}\in\mathbb{R}^N} R_N(\mathbf{u}), \quad \text{subject to } u_0 = 1, \tag{29}$$

where $R_N(\mathbf{u})$ is the Rayleigh quotient given by

$$R_N(\mathbf{u}) = \frac{\left(\int_0^1 4\pi\rho^2 \left|u_N'(\rho)\right|^2 \, \mathrm{d}\rho\right)^{1/2}}{\left(\int_0^1 4\pi\rho^2 \left|u_N(\rho)\right|^6 \, \mathrm{d}\rho\right)^{1/6}}.$$

Notice that, being $R_N(\mathbf{u}) = R_N(\nu\mathbf{u})$ for any $\nu \in \mathbb{R}$ by definition, the assumption $u_0 = 1$ does not effect the minimization problem, but is just set in order to ensure convergence to our numerical procedure.

Thanks to the discussion in the previous section, we know that $S_N$ basically corresponds to the solution of problem (29). We have numerically solved (29) and compared our discrete approximation $S_N$ with the sharp constant (2).

In Section 4.1 we present some numerical results obtained with equispaced grids to validate our theoretical estimates. Then, due to the shape of the optimal functions (which decrease very rapidly from 1 to almost 0 near the origin, and then remain very close to 0 in most of the ball) we pass to consider non-equispaced grids clustered to 0. A first specific example of non-uniform grids, which shows that the convergence rate is much faster with respect to the case of equispaced grids, is made in Section 4.2. Finally, in Section 4.3 we present an adaptive algorithm which automatically generate the grids, and that provide an even faster rate of convergence.

### 4.1 Equispaced grids

We consider a sequence of equispaced uniform grids made of $N = 2^k$ elements, $k = 1, 2, \ldots, 9$, with corresponding mesh size $h_N = 2^{-k}$, and, at each step of refinement, we have solved the constrained optimization problem (29). In Table 1 we report the computed errors together with the computed convergence rates: we observe that $S_N - S \searrow 0$ as $N$ goes to $+\infty$, at a rate of 0.6 approximately. Observe that the convergence rate is in the range predicted by Theorem 1.1. In Figure 4 we report, for the refinement levels $k = 2, 3, \ldots, 9$, the computed optimal function $u_N$. For the sake of comparison, we also show the exact optimal function

$$u(\rho) = \frac{1}{\sqrt{1 + b\left|\rho\right|^2}}, \tag{30}$$

where, at each refinement level, the parameter $b$ in (30) has been chosen so that $\|f\|_{L^6(B)} = \|f_N\|_{L^6(B)}$, namely

$$\int_0^1 4\pi\rho^2 \left|u_N(\rho)\right|^6 \, \mathrm{d}\rho = \int_0^1 4\pi\rho^2 \left|u(\rho)\right|^6 \, \mathrm{d}\rho. \tag{31}$$

In Table 1 (fifth column) we also report $\|f_N\|_{L^6(B)}$. The selection of $b$ in (31) has been done numerically by the bisection method up the machine precision

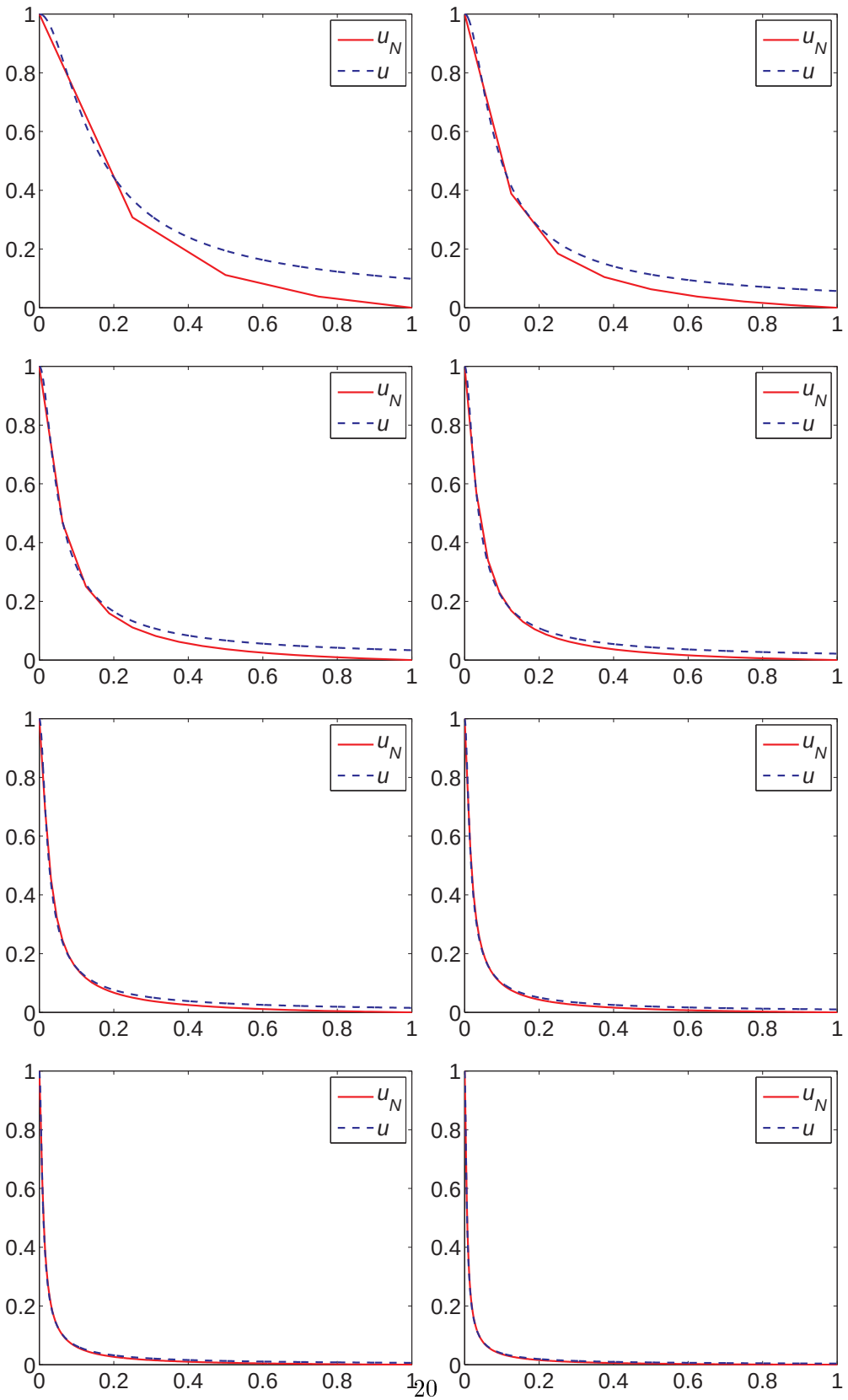





Figure 4: Approximated optimal function (solid line) and exact optimal function (dashed line) on equispaced grids for the refinement levels $k = 2, 3, \ldots, 9$.

Table 1: Equispaced grids. Error estimates and computed convergence rates; estimate of $\|f_N\|_{L^6(B)}$ and corresponding estimate of $b$ such that $\|f\|_{L^6(B)} = \|f_N\|_{L^6(B)}$.

| $k$ | $N$ | $S_N - S$ | rate | $\|f_N\|_{L^6(B)}$ | $b$ |
|---|---|---|---|---|---|
| 1 | 2 | 5.5294e-01 | - | 4.7878e-01 | 3.4565e+01 |
| 2 | 4 | 3.2139e-01 | 0.78279 | 3.6568e-01 | 1.0200e+02 |
| 3 | 8 | 1.9909e-01 | 0.69093 | 2.7725e-01 | 3.0897e+02 |
| 4 | 16 | 1.2868e-01 | 0.62960 | 2.1275e-01 | 8.9130e+02 |
| 5 | 32 | 8.4367e-02 | 0.60905 | 1.7156e-01 | 2.1078e+03 |
| 6 | 64 | 5.4019e-02 | 0.64322 | 1.4399e-01 | 4.2472e+03 |
| 7 | 128 | 3.4074e-02 | 0.66480 | 1.1684e-01 | 9.7981e+03 |
| 8 | 256 | 2.1515e-02 | 0.66333 | 9.2956e-02 | 2.4456e+04 |
| 9 | 512 | 1.3778e-02 | 0.64296 | 7.2168e-02 | 6.7314e+04 |

(cf. Table 1, last column). As it can be inferred from the results shown in Figure 4, as the mesh is refined we get better and better approximations.

It is easy to understand that, as $h$ goes to 0 or, equivalently, as $N$ goes to $+\infty$, the approximated solution $u_N$, decrease faster and faster near the origin (where by definition its value is always 1), and then it is very close to 0 in most of the interval $I$. This is clear from the geometry of the solution (cf. also Figure 4), and it can be inferred from the proof of Lemma 2.1. Therefore, to improve the approximation error and to save computational time, the grid points should be clustered near the origin where the solution undergoes a rapid variation. In other words, once the number $N$ of intervals of the grid is fixed, it appears quite reasonable that a grid more dense around the origin should give better approximation results.

Based on the above observation, we will now consider non-equispaced grids: we start in Section 4.2 with an arbitrary chosen grid, to show that even with this simple choice the convergence rate is improved, and then in Section 4.3 we will present an adaptive algorithm to get an automatically generation of the grids.

### 4.2 Non-equispaced grids: an example

The grid that we present in this section is very simple: we fix a positive parameter $\tau$ and we consider $N = 2^k$ intervals whose lengths are proportional to 1, $1+\tau$, $1+2\tau$, $\dots 1+(N-1)\tau$: this means that the points $x_i$ are given by the formula

$$x_i = \frac{i(2+\tau(i-1))}{N(2+\tau(N-1))}, \qquad i = 0, \dots, N.$$

Notice that, the equispaced grid correspond to $\tau = 0$, and when $\tau$ becomes bigger, then more points are clustered to the origin. Figure 5 shows a sample of the first four refinements ($k = 1, 2, 3, 4$) for $\tau = 1$.

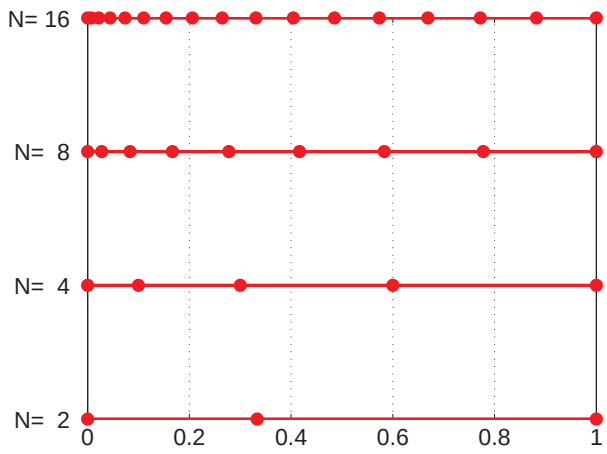



Figure 5: First four levels of non-equispaced grids for $\tau = 1$.

We have ran the same set of experiments as before: the results are shown in Table 2. We clearly observe an improvement in the approximation errors: with $N = 128$ we get a better approximation of $S$ than what we had in the equispaced case with $N = 512$. Also the computed convergence rate is quite better than in the equispaced case, namely 1 instead of 0.66. We have ran the same set of experiments with $\tau = 2, 3, 4$: the results are analogous to the ones reported in Table 2 and are omitted here for the sake of brevity.

Table 2: Non-equispaced grids. Error estimates and computed convergence rates; estimate of $\|f_N\|_{L^6(B)}$ and corresponding estimate of $b$ such that $\|f\|_{L^6(B)} = \|f_N\|_{L^6(B)}$.

| $k$ | $N$ | $S_N - S$ | rate | $\|f_N\|_{L^6(B)}$ | $b$ |
|---|---|---|---|---|---|
| 1 | 2 | 5.1876e-01 | - | 3.9518e-01 | 7.4740e+01 |
| 2 | 4 | 3.0466e-01 | 0.7679 | 2.4031e-01 | 5.4744e+02 |
| 3 | 8 | 1.6633e-01 | 0.8731 | 2.1834e-01 | 8.0339e+02 |
| 4 | 16 | 8.5380e-02 | 0.9621 | 1.5738e-01 | 2.9762e+03 |
| 5 | 32 | 4.3220e-02 | 0.9822 | 1.1682e-01 | 9.8043e+03 |
| 6 | 64 | 2.1770e-02 | 0.9894 | 8.4353e-02 | 3.6065e+04 |
| 7 | 128 | 1.0928e-02 | 0.9942 | 6.0105e-02 | 1.3991e+05 |
| 8 | 256 | 5.4981e-03 | 0.9911 | 4.1442e-02 | 6.1908e+05 |
| 9 | 512 | 2.8059e-03 | 0.9705 | 2.8469e-02 | 2.7799e+06 |

### 4.3 Non-equispaced grids: an adaptive refinement strategy

Finally we present an adaptive algorithm for the automatic refinement of the mesh. The refinement strategy follows this observation: in view of the classical

estimate (6) one can expect that, to estimate as good as possible a function $f$, more points of the grid are needed where the second derivative of $f$ is big. Recall also that, in our problem, we do not want to approximate a single unknown function: indeed, there is a whole $1-$parameter class of optimal functions, and of course whenever a different grid is selected then our approximated solution will be close to a different optimal one. Hence, the adaptively refined mesh is generated according to the following algorithm.

**Algorithm 4.1** `Given an initial grid with` $N_0$ `elements,`

1. `solve the constrained optimization problem` (29)`;`
2. `compute the parameter` $b$ `in` (30) `so that` (31) `is satisfied;`
3. `compute the quantities`

$$\eta_i = \int_{x_{i-1}}^{x_i} 4\pi\rho^2 \left|u''(\rho)\right|^2 \, d\rho, \qquad\qquad i = 1, \dots, N;$$

4. `employ the fixed fraction mesh refinement criterion, based on` $\eta_i$`, with refinement fraction set to` $25\%$`, to identify elements which will be refined;`
5. `refine elements marked for refinement.`

In Figure 6 we show the first six meshes generated by Algorithm 4.1 starting from an initial uniform grid made of $N_0 = 8$ elements, together with the corresponding zoom near the origin: as expected the adaptive algorithm cluster points near the origin. The computed errors $S_N - S$ together with the computed con-

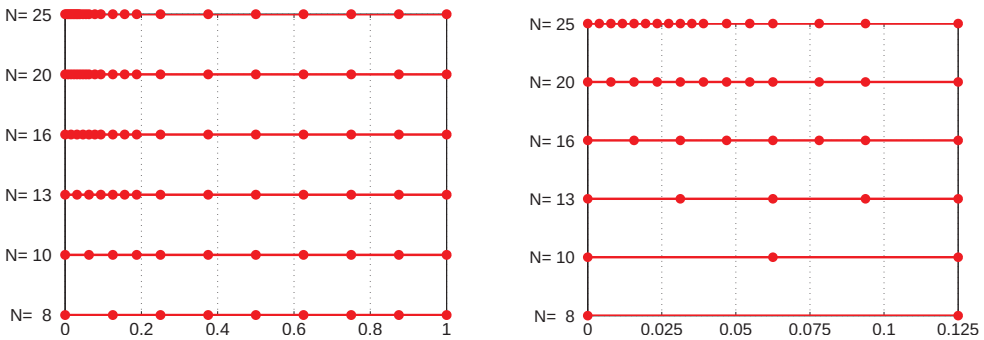



Figure 6: First six levels of adaptively refined grids (left), and corresponding zoom near the origin (right).

vergence rates are shown in Table 3. As before, we also report $\|f_N\|_{L^6(B)}$ and the corresponding estimate of $b$ such that $\|f\|_{L^6(B)} = \|f_N\|_{L^6(B)}$. The convergence rate is now around 1.5, so more than linear, indicating that the adaptive strategy provides better results than the ones obtained on general non-equispaced grids.

Table 3: Adaptively refined grids. Error estimates and computed convergence rates; estimate of $\|f_N\|_{L^6(B)}$ and corresponding estimate of $b$ such that $\|f\|_{L^6(B)} = \|f_N\|_{L^6(B)}$.

| level | $N$ | $S_N - S$ | rate | $\|f_N\|_{L^6(B)}$ | $b$ |
|---|---|---|---|---|---|
| 1 | 8 | 1.9909e-01 | - | 2.7724e-01 | 3.0899e+02 |
| 2 | 10 | 1.3432e-01 | 1.9609 | 2.1074e-01 | 9.2568e+02 |
| 3 | 13 | 9.0433e-02 | 1.6406 | 1.6706e-01 | 2.3441e+03 |
| 4 | 16 | 6.1186e-02 | 2.0122 | 1.3813e-01 | 5.0151e+03 |
| 5 | 20 | 4.0335e-02 | 1.9720 | 1.1178e-01 | 1.1695e+04 |
| 6 | 25 | 2.6498e-02 | 1.9672 | 8.8816e-02 | 2.9344e+04 |
| 7 | 31 | 1.7511e-02 | 1.9950 | 7.0312e-02 | 7.4711e+04 |
| 8 | 39 | 1.1505e-02 | 1.8826 | 5.5871e-02 | 1.8739e+05 |
| 9 | 49 | 7.5472e-03 | 1.8893 | 4.4451e-02 | 4.6771e+05 |
| 10 | 61 | 4.9593e-03 | 1.9520 | 3.5257e-02 | 1.1817e+06 |
| 11 | 76 | 3.3586e-03 | 1.7988 | 3.0061e-02 | 2.2361e+06 |
| 12 | 95 | 2.4773e-03 | 1.3799 | 2.5854e-02 | 4.0866e+06 |
| 13 | 119 | 1.8406e-03 | 1.3315 | 2.1575e-02 | 8.4274e+06 |
| 14 | 149 | 1.3621e-03 | 1.3492 | 1.8841e-02 | 1.4489e+07 |
| 15 | 186 | 1.0253e-03 | 1.2881 | 1.4799e-02 | 3.8069e+07 |
| 16 | 233 | 7.1047e-04 | 1.6361 | 1.1163e-02 | 1.1758e+08 |
| 17 | 291 | 5.0007e-04 | 1.5860 | 8.4020e-03 | 3.6641e+08 |

Finally, we compare the results obtained with the three set of meshes considered so far: namely, equispaced, non-equispaced, and adaptively refined grids: the computed errors versus the number of elements $N$ are shown in Figure 7 (loglog scale). Clearly, the results obtained on the sequence of adaptively refined grids outperform the ones obtained on both equispaced and non-equispaced meshes.

## 5 Conclusion

We have shown that the optimal constant in the discrete Sobolev inequality in $\mathrm{W}_0^{1,2}(B)$ approximates, with a polynomial rate of convergence, the optimal constant in the continuous version of the Sobolev inequality. The convergence is established providing both an upper and a lower bound on the rate of convergence. Numerical results including also an adaptive refinement strategy are also presented.

Possible future developments of our results may go in the following directions.

- The development of a better refinement strategy to construct the mesh: indeed, even though our method appears quite good, it could be made better since when $N$ increases also $b$ increases, and then the refinement of

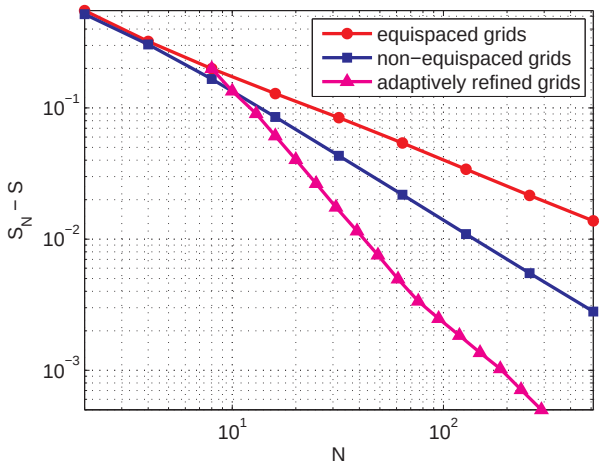



Figure 7: Computed error $S_N - S$ versus the number of elements: equispaced, non-equispaced, and adaptively refined grids.

the grid made at one level is surely good but it is not the best possible choice for the following levels.

- In this work we do not have to approximate a given function, but we are approximating a degenerating sequence of functions. Hence, it could be possible to adopt the same kind of strategy for situations where solutions do not exist or degenerate in some sense. For instance, in the case of problems with critical exponents, it may happen that a solution does not exist but the finite elements method gives an "approximated solution". It could be interesting to understand at which rate these approximated solutions explode or disappear when $h \searrow 0$.

- More general Sobolev embeddings and the related inequalities are extensively studied in the literature (among the recent works, we mention for instance [12]). Therefore, one could try to investigate the possible consequences that our kind of "polynomial rate of convergence" result has in these general kinds of problems.

## Acknowledgements

We are grateful to Prof. Enrique Zuazua for suggesting us the topic of the presented results, and for his valuable and constructive comments. Part of this work has been carried out at the Centro de Ciencias de Benasque Pedro Pascual during the summer school 2009 *"Partial differential equations, optimal design and numerics"*.

## References


[1] T. Apel and C. Pester. Clement–type interpolation on spherical domains–interpolation error estimates and application to a posteriori error estimation. *IMA J. Numer Anal*, 25:310–336, 2005.

[2] T. Aubin. Problèmes isopérimétriques et espaces de Sobolev. *J. Differential Geometry*, 11(4):573–598, 1976.

[3] S. C. Brenner and L. Scott. *The mathematical theory of finite element methods.* Springer, Amsterdam, 1994.

[4] A. Buffa and C. Ortner. Compact embeddings of broken Sobolev spaces and applications. IMA J. Numer. Anal, Advance Access published on July 4, 2008, 2008.

[5] A. Cianchi, N. Fusco, F. Maggi, and A. Pratelli. The sharp sobolev inequality in quantitative form. JEMS, (to appear).

[6] P. G. Ciarlet. *The finite element method for elliptic problems.* North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[7] D. Cordero-Erausquin, B. Nazaret, and C. Villani. A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities. *Adv. Math.*, 182(2):307–332, 2004.

[8] R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49:1–23, 1943.

[9] G. Pólya and G. Szegö. *Isoperimetric Inequalities in Mathematical Physics.* Princeton Univ. Press, Princeton, 1951.

[10] H. Strang. Approximation in the finite element method. *Numer. Math*, 19:81–98, 1972.

[11] G. Talenti. Best constant in Sobolev inequality. *Ann. Mat. Pura Appl. (4)*, 110:353–372, 1976.

[12] E. Zuazua. Log-Lipschitz regularity and uniqueness of the flow for a field in $(W^{n/p+1,p}_{\mathrm{loc}}(\mathbb{R}^n))^n$. *C. R. Math. Acad. Sci. Paris*, 335(1):17–22, 2002.

# MOX Technical Reports, last issues

**Dipartimento di Matematica "F. Brioschi",**
**Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)**

**29/2009** P.F. Antonietti, A. Pratelli:
*Finite Element Approximation of the Sobolev Constant*

**28/2009** L. Gaudio, A. Quarteroni:
*Spectral Element Discretization of Optimal Control Problems*

**27/2009** L. Mirabella, F. Nobile, A. Veneziani:
*A robust and efficient conservative technique for simulating three-dimensional sedimentary basins dynamics*

**26/2009** M. Longoni, A.C.I. Malossi, A. Villa:
*A robust and efficient conservative technique for simulating three-dimensional sedimentary basins dynamics*

**25/2009** P.E. Farrell, S. Micheletti, S. Perotto:
*An anisotropic Zienkiewicz-Zhu a posteriori error estimator for 3D applications*

**24/2009** F. Di Maio, P. Secchi, S. Vantini, E. Zio:
*Optimized Fuzzy C-Means Clustering and Functional Principal Components for Post-Processing Dynamic Scenarios in the Reliability Analysis of a Nuclear System*

**23/2009** L. Gerardo Giorda, F. Nobile, C. Vergara:
*Analysis and optimization of Robin-Robin partitioned procedures in Fluid-Structure Interaction Problems*

**22/2009** L. Formaggia, S. Minisini, P. Zunino:
*Modeling erosion controlled drug release and transport phenomena in the arterial tissue*

**21/2009** L. Bonaventura, S. Castruccio, L.M. Sangalli:
*A Bayesian approach to geostatistical interpolation with flexible variogram models*

**20/2009** N.Accoto, T. Rydén, P. Secchi:
*Bayesian hidden Markov models for performance-based regulation of continuity of electricity supply*