Greedy bisection generates optimally adapted triangulations

Jean-Marie Mirebeau, Albert Cohen

Greedy bisection generates optimally adapted triangulations

Jean-Marie Mirebeau and Albert Cohen

October 22, 2008

Abstract

We study the properties of a simple greedy algorithm introduced in [9] for the generation of data-adapted anisotropic triangulations. Given a function f, the algorithm produces nested triangulations $\mathcal{T}_N$ and corresponding piecewise polynomial approximations f_N of f. The refinement procedure picks the triangle which maximizes the local L^p approximation error, and bisect it in a direction which is chosen so to minimize this error at the next step. We study the approximation error in the L^p norm when the algorithm is applied to C^2 functions with piecewise linear approximations. We prove that as the algorithm progresses, the triangles tend to adopt an optimal aspect ratio which is dictated by the local hessian of f. For convex functions, we also prove that the adaptive triangulations satisfy the convergence bound $\|f - f_N\|_{L^p} \leq CN^{-1}\|\sqrt{|\det(d^2 f)|}\|_{L^\tau}$ with $\frac{1}{\tau} := \frac{1}{p} + 1$, which is known to be asymptotically optimal among all possible triangulations.

1 Introduction

In finite element approximation, a classical and important distinction is made between *uniform* and *adaptive* methods. In the first case the all the elements which constitute the mesh have comparable shape and size, while these attributes are allowed to vary strongly in the second case. An important feature of adaptive methods is the fact that the mesh is not fixed in advance but rather taylored to the properties of the function f to be approximated. Since the function approximating f is not picked from a fixed linear space, adaptive finite elements can be considered as an instance of *non-linear approximation*. Other instances include approximation by rational functions, or by N-term linear combinations of a basis or dictionary. We refer to [10] for a general survey on non-linear approximation.

In this paper, we focus our interest on *piecewise linear* finite element functions defined over triangulations of a bidimensional polygonal domain $\Omega \subset \mathbb{R}^2$. Given a triangulation $\mathcal{T}$ we denote by $V_\mathcal{T} := \{v \text{ s.t. } v_{|T} \in \Pi_1,\ T \in \mathcal{T}\}$ the associated finite element space. The norm in which we measure the approximation error is the L^p norm for $1 \leq p \leq \infty$ and we therefore do not require that the triangulations are conforming and that the functions of $V_\mathcal{T}$ are continuous between triangles. For a given function f we define

$$e_N(f)_{L^p} := \inf_{\#(\mathcal{T}) \leq N} \inf_{g \in V_\mathcal{T}} \|f - g\|_{L^p},$$

the best approximation error of f when using at most N elements. In adaptive finite element approximation, critical questions are:

1. Given a function f and a number $N > 0$, how can we characterize the *optimal mesh* for f with N elements corresponding to the above defined best approximation error.

2. What quantitative estimates are available for the best approximation error $e_N(f)_{L^p}$? Such estimates should involve the derivatives of f in a different way than for non-adaptive meshes.

3. Can we build by a simple algorithmic procedure a mesh $\mathcal{T}_N$ of cardinality N and a finite element function $f_N \in V_{\mathcal{T}_N}$ such that $\|f - f_N\|_{L^p}$ is comparable to $e_N(f)_{L^p}$?

While the optimal mesh is usually difficult to characterize exactly, it should satisfy two intuitively desirable features: (i) the triangulation should *equidistribute* the local approximation error between each triangle and (ii) the aspect ratio of a triangle T should be *isotropic* with respect to a distorted metric induced by the local value of the hessian $d^2 f$ on T (and therefore anisotropic in the sense of the euclidean metric). Under such prescriptions on the mesh, quantitative error estimates have recently been obtained in [7, 2] when f is a C^2 function. These estimates are of the form

$$e_N(f)_{L^p} \leq CN^{-1} \| \sqrt{|\det(d^2 f)|} \|_{L^\tau}, \quad \frac{1}{\tau} = \frac{1}{p} + 1, \tag{1.1}$$

where $\det(d^2 f)$ is the determinant of the 2×2 hessian matrix. They can be proved to be *asymptotically optimal* in the sense of a lower inequality of the form

$$\liminf_{N \to +\infty} N e_N(f)_{L^p} \geq c \| \sqrt{|\det(d^2 f)|} \|_{L^\tau}. \tag{1.2}$$

We refer in particular to [11] in which such estimates are generalized to higher order elements and proved to be optimal in the above sense.

From the computational viewpoint, a commonly used strategy for designing an optimal mesh consists therefore in evaluating the hessian $d^2 f$ and imposing that each triangle of the mesh is isotropic with respect to a metric which is properly related to its local value. We refer in particular to [5] where this program is executed using Delaunay mesh generation techniques. While these algorithms fastly produce anisotropic meshes which are naturally adapted to the approximated function, they suffer from two intrinsic limitations:

1. They use the data of $d^2 f$, and therefore do not apply to non-smooth or noisy functions.
2. They are non-hierarchical: for $N > M$, the triangulation $\mathcal{T}_N$ is not a refinement of $\mathcal{T}_M$.

In [9], an alternate strategy was proposed for the design of adaptive hierarchical meshes, based on a simple *greedy algorithm*: starting from an initial triangulation $\mathcal{T}_{N_0}$, the algorithm picks the triangle $T \in \mathcal{T}_k$ with the largest local L^p error. This triangle is then bisected from one of its vertex to the mid-point of the opposite edge. The choice of the vertex among the three options is the one that minimizes the new approximation error after bisection. The algorithm can be applied to any L^p function, smooth or not, in the context of piecewise polynomial approximation of any given order. In the case of piecewise linear approximation, numerical experiments in [9] indicate that this elementary strategy generates triangles with an optimal aspect ratio and approximations f_N such that $\|f - f_N\|_{L^p}$ satisfies the same estimate as $e_N(f)_{L^p}$ in (1.1).

The goal of this paper is to support these experimental observations by a rigorous analysis. Our paper is organized as follows:

In §2, we introduce notations which are used throughout the paper and collect some available approximation theory results for piecewise linear finite elements, making the distinction between (i) uniform, (ii) adaptive isotropic and (iii) adaptive anisotropic triangulations. In the last case, which is in the scope of this paper, we introduce a measure of non-degeneracy of a triangle T with respect to a quadratic form. We show that the optimal error estimate (1.1) is met when each triangles are non-degenerate with in the sense of the above measure with respect to the quadratic form given by the local hessian $d^2 f$. We end by briefly recalling the greedy algorithm which was introduced in [9]. In §3, we study the behavior of the refinement procedure when applied to a quadratic function q such that its associated quadratic form $\mathbf{q}$ is of positive or negative sign. A key observation is that the edge which is bisected is the longest with respect to the metric induced by $\mathbf{q}$. This allows us to prove that the triangles generated by the refinement procedure adopt an optimal aspect ratio in the sense of the non-degeneracy measure introduced in §2. In §4, we proceed to a similar analysis in the case where $\mathbf{q}$ is of mixed sign, also proving that the triangles adopt an optimal aspect ratio as they get refined. In §5, we study the behavior of the algorithm when applied to a general C^2 function f which is assumed to be strictly convex. We first prove a perturbation result, which show that when f is locally close to a quadratic function q the algorithm behaves in a similar manner as when applied to q. This allows us to show that the optimal convergence estimate (1.1) is met in the L^p norm. We do not know if a similar result holds when the algorithm is applied to an arbitrary C^2 function, although this seems plausible from the numerical experiments reported in [9].

2 Adaptive finite element approximation

2.1 Notations

We shall make use of a local approximation operator $\mathcal{A}_T$ from $L^p(T)$ onto Π_1. Here $1 \leq p \leq \infty$ is arbitrary but fixed. We define the local L^p approximation error

$$e_T(f)_p := \|f - \mathcal{A}_T f\|_{L^p(T)}.$$

An assumption which will be important in our analysis is that the operator $\mathcal{A}_T$ commutes with affine changes of variables:

$$\mathcal{A}_T(f) \circ \phi = \mathcal{A}_{\phi^{-1}(T)}(f \circ \phi),$$

for all affine transformation ϕ. We may consider for $\mathcal{A}_T$ the operator B_T of best $L^p(T)$ approximation which is defined by

$$\|f - B_T f\|_{L^p(T)} = \min_{\pi \in \Pi_m} \|f - \pi\|_{L^p(T)}.$$

However this operator is non-linear and not easy to compute when $p \neq 2$. We therefore restrict our attention to the following two options:

1. $\mathcal{A}_T = P_T$, the $L^2(T)$-orthogonal projection operator: $\int_T (f - \Pi_T f)\pi = 0$ for all $\pi \in \Pi_1$.
2. $\mathcal{A}_T = I_T$, the local interpolation operator: $I_T f(v_i) = f(v_i)$ with $\{v_0, v_1, v_2\}$ the vertices of T.

Unless explicitely stated, all our results are simultaneously valid when $\mathcal{A}_T$ is either P_T or I_T. Given a function f and a triangulation $\mathcal{T}_N$ with $N = \#(\mathcal{T}_N)$, we can associate a finite element approximation f_N defined on each $T \in \mathcal{T}_N$ by $f_N(x) = \mathcal{A}_T f(x)$. The global approximation error is given by

$$\|f - f_N\|_{L^p} = \Big(\sum_{T \in \mathcal{T}_N} e_T(f)_p^p \Big)^{\frac{1}{p}},$$

with the usual modification when $p = \infty$.

Here and throughout the paper, when q is a quadratic polynomial

$$q(x, y) = \sum_{\alpha+\beta \leq 2} a_{\alpha,\beta} x^\alpha y^\beta,$$

we denote by $\mathbf{q}$ the associated quadratic form : if $u = (x, y)$

$$\mathbf{q}(u) = \mathbf{q}(u, u) = \sum_{\alpha+\beta=2} a_{\alpha,\beta} x^\alpha y^\beta.$$

It is the restriction to the diagonal of a bilinear form $\mathbf{q}(u, v) = \langle Qu, v \rangle$ where the entries of the symmetric matrix Q are given by the coefficients $a_{\alpha,\beta}$. We define

$$\det(\mathbf{q}) = \det(Q).$$

If $\mathbf{q}$ is a positive or negative quadratic form, we define the $\mathbf{q}$-*metric*

$$|v|_{\mathbf{q}} := \sqrt{|\mathbf{q}(v)|}$$

which coincides with the euclidean norm when $\mathbf{q}(v) = x^2 + y^2$ for $v = (x, y)$. If $\mathbf{q}$ is a quadratic form of mixed sign, we define the associated positive form $|\mathbf{q}|$ which corresponds to the symmetric matrix $|Q|$ that has same eigenvectors as Q with eigenvalues $(|\lambda|, |\mu|)$ if (λ, μ) are the eigenvalues of Q. Note that $|\mathbf{q}|(u) \neq |\mathbf{q}(u)|$ and that one always has $|\mathbf{q}(u)| \leq |\mathbf{q}|(u)$.

2.2 From uniform to adaptive isotropic triangulations

A standard estimate in finite element approximation states that if $f \in W^{2,p}(\Omega)$ then

$$\inf_{g \in V_h} \|f - g\|_{L^p} \leq Ch^2 \|d^2 f\|_{L^p},$$

where V_h is the piecewise linear finite element space associated with a triangulation $\mathcal{T}_h$ of mesh size $h := \max_{T \in \mathcal{T}_h} \operatorname{diam}(T)$. If we restrict our attention to *uniform* triangulations, we have

$$N := \#(\mathcal{T}_h) \sim h^{-2}.$$

Therefore, denoting by $e_N^{\text{unif}}(f)_{L^p}$ the L^p approximation by a uniform triangulation of cardinality N, we can re-express the above estimate as

$$e_N^{\text{unif}}(f)_{L^p} \leq CN^{-1} \|d^2 f\|_{L^p}. \tag{2.3}$$

This estimate can be significantly improved when using adaptive partitions. We give here some heuristic arguments, which are based on the assumption that on each triangle T the relative variation of $d^2 f$ is small so that it can be considered as constant over T (which means that f is replaced by a quadratic function on each T), and we also indicate the available results which are proved more rigorously.

First consider *isotropic* triangulations, i.e. such that all triangles satisfy a uniform estimate

$$\rho_T = \frac{h_T}{r_T} \leq A, \tag{2.4}$$

where $h_T := \operatorname{diam}(T)$ and r_T is the radius of the largest disc contained in T. In such a case we start from the local approximation estimate on any T

$$e_T(f)_p \leq Ch_T^2 \|d^2 f\|_{L^p(T)},$$

and notice that

$$h_T^2 \|d^2 f\|_{L^p(T)} \sim |T| \, \|d^2 f\|_{L^p(T)} = \|d^2 f\|_{L^q(T)},$$

with $\frac{1}{q} := \frac{1}{p} + 1$ and $|T|$ the area of T, where we have used the isotropy assumption (2.4) in the equivalence and the fact that $d^2 f$ is constant over T in the equality. It follows that

$$e_T(f)_p \leq C \|d^2 f\|_{L^\tau(T)}, \quad \frac{1}{\tau} := \frac{1}{p} + 1$$

Assume now that we can construct adaptive isotropic triangulations $\mathcal{T}_N$ with $N := \#(\mathcal{T}_N)$ which *equidistributes* the local error in the sense that for some prescribed $\varepsilon > 0$

$$c\varepsilon \leq e_T(f)_p \leq \varepsilon, \tag{2.5}$$

with $c > 0$ a fixed constant independent of T and N. Then defining f_N as $\mathcal{A}_T(f)$ on each $T \in \mathcal{T}_N$, we have on the one hand

$$\|f - f_N\|_{L^p} \leq N^{1/p} \varepsilon,$$

and on the other hand, with $\frac{1}{\tau} := \frac{1}{p} + 1$,

$$N(c\varepsilon)^\tau \leq \sum_{T \in \mathcal{T}_N} \|f - f_N\|_{L^p(T)}^\tau \leq C^\tau \sum_{T \in \mathcal{T}_N} \|d^2 f\|_{L^\tau(T)}^\tau \leq C^\tau \|d^2 f\|_{L^\tau}^\tau.$$

Combining both, one obtains for $e_N^{\text{iso}}(f)_{L^p} := \|f - f_N\|_{L^p}$ the estimate

$$e_N^{\text{iso}}(f)_{L^p} \leq CN^{-1} \|d^2 f\|_{L^\tau}. \tag{2.6}$$

This estimate improves upon (2.3) since the rate N^{-1} is now obtained with the weaker smoothness condition $d^2 f \in L^\tau$ and since, even for smooth f, the quantity $\|d^2 f\|_{L^\tau}$ might be significantly smaller

than $\|d^2 f\|_{L^p}$. This type of result is classical in non-linear approximation and also occurs when we consider best N-term approximation in a wavelet basis.

The principle of error equidistribution suggests a simple *greedy algorithm* to build an adaptive isotropic triangulation for a given f, similar to our algorithm but where the bisection of the triangle T that that maximizes the local error $e_T(f)_p$ is systematically done from its *most recently created vertex* in order to preserve the estimate (2.4). Such an algorithm cannot exactly equilibrate the error in the sense of (2.5) and therefore does not lead to same the optimal estimate as in (2.6). However, it was proved in [4] that it satisfies

$$\|f - f_N\|_{L^p} \leq C|f|_{B^2_{\tau,\tau}} N^{-\frac{s}{2}},$$

for all τ such that $\frac{1}{\tau} < \frac{1}{p} + 1$. Here $B^2_{\tau,\tau}$ denotes the usual Besov space which is a natural substitute for $W^{2,\tau}$ when $\tau < 1$. Therefore this estimate is not far from (2.6).

2.3 Anisotropic triangulations: the optimal aspect ratio

We now turn to anisotropic adaptive triangulations, and start by discussing the optimal shape of a triangle T for a given function f at a given point. For this purpose, we again replace f by a quadratic function assuming that $d^2 f$ is constant over T. For such a $q \in \Pi_2$ and its associated quadratic form $\mathbf{q}$, we first derive an equivalent quantity for the local approximation error. Here and as well as in §3 and §4, we consider a triangle T and we denote by (a, b, c) its edge vectors oriented in clockwise or anticlockwise direction so that

$$a + b + c = 0.$$

Proposition 2.1 *The local L^p-approximation error satisfies*

$$e_T(q)_p = e_T(\mathbf{q})_p \sim |T|^{\frac{1}{p}} \max\{|\mathbf{q}(a)|, |\mathbf{q}(b)|, |\mathbf{q}(c)|\},$$

where the constant in the equivalence is independent of q, T and p.

Proof: The first equality is trivial since q and $\mathbf{q}$ differ by an affine function. Let T be an equilateral triangle of area $|T| = 1$, and edges a, b, c. Let E be the 3-dimensional vector space of all quadratic forms. Then the following quantities are norms on E, and thus equivalent:

$$e_T(\mathbf{q})_p \sim \max\{|\mathbf{q}(a)|, |\mathbf{q}(b)|, |\mathbf{q}(c)|\}.$$

Note that the constants in this equivalence are independent of p since all $L^p(T)$ norms are uniformly equivalent on E. If T is now an arbitrary triangle, we obtain the claimed equivalence with the same constants by a change of variable. ◇

In order to describe the optimal shape of a triangle T for the quadratic function q, we fix the area of $|T|$ and try to minimize the error $e_T(q)_p$ or equivalently $\max\{|\mathbf{q}(a)|, |\mathbf{q}(b)|, |\mathbf{q}(c)|\}$. The solution to this problem can be found by introducing for any $\mathbf{q}$ such that $\det(\mathbf{q}) \neq 0$ the following measure of *non-degeneracy* for T:

$$\rho_{\mathbf{q}}(T) := \frac{\max\{|\mathbf{q}(a)|, |\mathbf{q}(b)|, |\mathbf{q}(c)|\}}{|T|\sqrt{|\det(\mathbf{q})|}}. \tag{2.7}$$

It is easily checked that for any linear change of variable ϕ, we have

$$\rho_{\mathbf{q}\circ\phi}(T) = \rho_{\mathbf{q}}(\phi(T)). \tag{2.8}$$

This allows to reduce the study of $\rho_{\mathbf{q}}(T)$ to two elementary cases by change of variable:

1. The case where $\det(\mathbf{q}) > 0$ is reduced to $\mathbf{q}(x, y) = x^2 + y^2$. In this case we have $\rho_{\mathbf{q}}(T) = \frac{h_T^2}{|T|}$, which corresponds to a standard measure of shape regularity in the sense that its boundedness is equivalent to a property such as 2.4. This quantity is minimized when the triangle T is equilateral, with minimal value $\frac{4}{\sqrt{3}}$. For a general quadratic form $\mathbf{q}$ of positive sign, we obtain by change of variable that the minimal value $\frac{4}{\sqrt{3}}$ is obtained for triangles which are equilateral with respect to the metric $|\cdot|_{\mathbf{q}}$. More generally triangles with a good aspect ratio are those which are *isotropic with respect to this metric*. Of course, a similar conclusion hold for a quadratic form of negative sign.

2. The case where $\det(\mathbf{q}) < 00$ is reduced to $\mathbf{q}(x,y) = x^2 - y^2$. In this case, elementary yet tedious computations show that the quantity $\rho_{\mathbf{q}}(T)$ is minimized when T is a half of a square with sides parallel to the x and y axes, with minimal value 2. But we also notice that $\rho_{\mathbf{q}}(T)$ is left invariant by a linear transformation of T with eigenvalues $(\lambda, \frac{1}{\lambda})$ and eigenvectors $(1,1)$ and $(-1,1)$ for any $\lambda \neq 0$. Therefore, the triangles with a good aspect ratio are not necessarily isotropic. For a general quadratic form $\mathbf{q}$ of mixed sign, we find that triangles which are isotropic with respect to the metric $|\cdot|_{|\mathbf{q}|}$ have a good aspect ratio. But so do all triangles obtained from such isotropic triangles by a linear transformation with eigenvalues $(\lambda, \frac{1}{\lambda})$ and eigenvectors (u,v) in the *null cone* of $\mathbf{q}$, i.e. such that $\mathbf{q}(u) = \mathbf{q}(v) = 0$.

We leave aside the special case where $\det(\mathbf{q}) = 0$. In such a case, the triangles minimizing the error for a given area degenerate in the sense that they should be infinitely long and thin, aligned with the direction of the null eigenvalue of $\mathbf{q}$.

Summing up, we find that triangles with a good aspect ratio are characterized by the fact that $\rho_{\mathbf{q}}(T)$ is small. In addition, from Proposition 2.1 and the definition of $\rho_{\mathbf{q}}(T)$, we have

$$e_T(q)_p \sim |T|^{1+\frac{1}{p}} \sqrt{|\det(\mathbf{q})|}\rho_{\mathbf{q}}(T) = \|\sqrt{|\det(\mathbf{q})|}\|_{L^\tau(T)}\rho_{\mathbf{q}}(T), \;\; \frac{1}{\tau} := \frac{1}{p} + 1. \tag{2.9}$$

We now return to a function f such that $d^2 f$ is assumed to be constant on every $T \in \mathcal{T}_N$. Assuming that all triangles have a good aspect ratio in the sense that

$$\rho_{\mathbf{q}}(T) \leq C$$

for some fixed constant C and with $\mathbf{q}$ the value of $d^2 f$ over T, we find up to a change in C that

$$e_T(f)_p \leq C\|\sqrt{|\det(d^2 f)|}\|_{L^\tau(T)} \tag{2.10}$$

By a similar reasoning as with isotropic triangulations, we now obtain that if the triangulation $\mathcal{T}_N$ equidistributes the error in the sense of (2.5),

$$\|f - f_N\|_{L^p} \leq CN^{-1}\|\sqrt{|\det(d^2 f)|}\|_{L^\tau}, \tag{2.11}$$

and therefore (1.1) holds. This estimate improves upon (2.6) since the quantity $\|\sqrt{|\det(d^2 f)|}\|_{L^\tau}$ might be significantly smaller than $\|d^2 f\|_{L^\tau}$, in particular when f has some anisotropic features, such as sharp gradients along curved edges.

The above derivation of (1.1) is heuristic and non-rigourous. Clearly, this estimate cannot be valid as such since $\det(d^2 f)$ may vanish while the approximation error does not (consider for instance f depending only on a single variable). More rigorous versions were derived in [7] and [2]. In these results $|d^2 f|$ is typically replaced by a majorant $|d^2 f| + \varepsilon I$, avoiding that $A(f)$ vanishes. The estimate (1.1) can then be rigorously proved but holds for $N \geq N(\varepsilon, f)$ large enough. This limitation is unavoidable and reflects the fact that enough resolution is needed so that the hessian can be viewed as locally constant over each optimized triangle. Another formulation which is rigorously proved in [11] is of the form

$$\limsup_{N \to +\infty} N e_N(f)_{L^p} \leq C\|\sqrt{|\det(d^2 f)|}\|_{L^\tau},$$

where C is an absolute constant.

2.4 The greedy algorithm

Given a target function f, our algorithm iteratively builds triangulations $\mathcal{T}_N$ with $N = \#(\mathcal{T}_N)$ and finite element approximations f_N. The starting point is a coarse triangulation $\mathcal{T}_{N_0}$. Given $\mathcal{T}_N$, the algorithm selects the triangle T which maximizes the local error $e_T(f)_p$ among all triangles of $\mathcal{T}_N$, and bisects it from one of its vertex a_i towards the mid-point of the opposite edge. This give the new triangulation T_{N+1}.

The critical part of the algorithm lies in the choice of the edge $e \in \{a, b, c\}$ from which T is bisected. Denoting by $T_{e,1}$ and $T_{e,2}$ the two resulting triangles, we choose e as the minimizer of a *decision function*

$d_T(e,f)$, which role is to drive the generated triangles towards an optimal aspect ratio. While the most natural choice for $d_T(e,f)$ corresponds to the optimal split

$$d_T(e,f) = e_{T_{e,1}}(f)_p^p + e_{T_{e,2}}(f)_p^p,$$

we shall instead focus our attention on decision functions which are either based on the L^2 projection error,

$$d_T(e,f) = \|f - P_{T_{e,1}} f\|^2_{L^2(T_{e,1})} + \|f - P_{T_{e,2}} f\|^2_{L^2(T_{e,2})}, \tag{2.12}$$

or the sum of the L^∞ interpolation errors,

$$d_T(e,f) = \|f - I_{T_{e,1}} f\|_{L^\infty(T_{e,1})} + \|f - I_{T_{e,2}} f\|_{L^\infty(T_{e,2})}. \tag{2.13}$$

With these two choices the analysis of the algorithm is made simpler, due to the fact that we can derive explicit expressions when $f = q$ is a quadratic polynomial. We prove in §3 and §4 that both choices lead to triangles with an optimal aspect ratio in the sense of a small $\rho_{\mathbf{q}}(T)$. This leads us in §5 to a proof that the algorithm satisfies the optimal convergence estimate (2.11) in the case where f is C^2 and strictly convex.

3 Positive quadratic functions

In this section, we study the algorithm when applied to a quadratic polynomial q such that $\det(\mathbf{q}) > 0$. We shall assume without loss of generality that $\mathbf{q}$ is positive definite, since all our results extend in a trivial manner to the negative definite case.

We first establish that the refinement procedure - either based on the decision functions (2.12) or (2.13) - always selects the vertex opposite to the longest edge in the sense of the $\mathbf{q}$-metric $|\cdot|_{\mathbf{q}}$. This is used to prove that the refinement procedure produces triangles which tend to adopt an optimal aspect ratio.

3.1 The L^∞-based split

Let us denote by

$$\alpha_T(f) = \|f - I_T f\|_{L^\infty(T)},$$

the interpolation error in the sup norm. The decision function (2.13) can be re-expressed as

$$d_T(e,f) = \alpha_{T_{e,1}}(f) + \alpha_{T_{e,2}}(f). \tag{3.14}$$

Theorem 3.1 *If $|a|_{\mathbf{q}} > \max\{|b|_{\mathbf{q}}, |c|_{\mathbf{q}}\}$, then $d_T(a,q) < \min\{d_T(b,q), d_T(c,q)\}$. Therefore the refinement procedure based on* (3.14) *selects the longest edge in the sense the $\mathbf{q}$-metric.*

This theorem means that for the quadratic function $q(x,y) = x^2 + y^2$, our refinement procedure is equivalent to the *longest edge bisection* algorithm which has been extensively studied, see e.g. [13], and which is known to promote isotropic triangles. This gives us a first insight on why our algorithm might generates triangles which are locally adapted to the hessian without computing this quantity. In order to prove this result, we first study the interpolation error in more detail.

Proposition 3.2 *Let T be a triangle with edges a, b, c such that $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$, and let $w \in \mathbb{R}^2$ and $r > 0$ be the center and radius of the circumscribed circle for the $\mathbf{q}$-metric, i.e. such that $|v - w|_{\mathbf{q}} = r$ for all the vertices v of T. Then*

$$\frac{|a|^2_{\mathbf{q}}}{4} \leq \alpha_T(q) \leq r^2.$$

Right equality holds if T is acute, i.e. $\mathbf{q}(b,c) \leq 0$. Left equality holds if T is obtuse, i.e. $\mathbf{q}(b,c) \geq 0$.

Proof: At any point $u \in \mathbb{R}^2$, we have

$$(q - I_T q)(u) = |u - w|^2_{\mathbf{q}} - r^2.$$

This function is negative on T with maximal value 0 at the vertices. If T is acute, then its minimal value on T is $-r^2$ and is attained at $w \in T$. If T is not acute, then the minimum is attained at m_a, the midpoint of a, and if we choose a vertex v at one end of a, we obtain the value at the minimum by Pythagora's identity which gives

$$\begin{aligned}(q - I_T q)(m_a) &= |m_a - w|^2_{\mathbf{q}} - r^2 = |m_a - w|^2_{\mathbf{q}} - |v - w|^2_{\mathbf{q}} \\ &= -|v - m_a|^2_{\mathbf{q}} = -|a|^2_{\mathbf{q}}/4.\end{aligned}$$

$\diamond$

The dichotomy in the above result is illustrated in the case of the euclidean metric on figure 4. Note that it would be sufficient to establish the above proof in this particular case, since we can perform an affine coordinate change $\phi = Q^{-1/2}$ such that $\mathbf{q} \circ \phi$ is the standard euclidean form and that the L^∞ interpolation error is left invariant by this coordinate change.

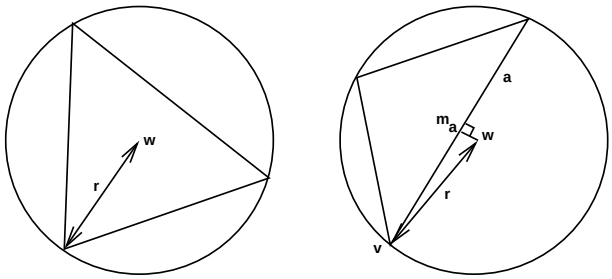

Figure 4: maximum point for the L^∞ interpolation error

We now prove the following result which clearly implies Theorem 4.5.

Proposition 3.3 *Let T be a triangle with edges $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$. We then have :*

$$\begin{aligned}d_T(b,q) - d_T(a,q) &\geq \frac{1}{4}(\mathbf{q}(a) - \mathbf{q}(b)) \\ d_T(c,q) - d_T(a,q) &\geq \frac{1}{4}(\mathbf{q}(a) - (\frac{|b|_{\mathbf{q}} + |c|_{\mathbf{q}}}{2})^2).\end{aligned}$$

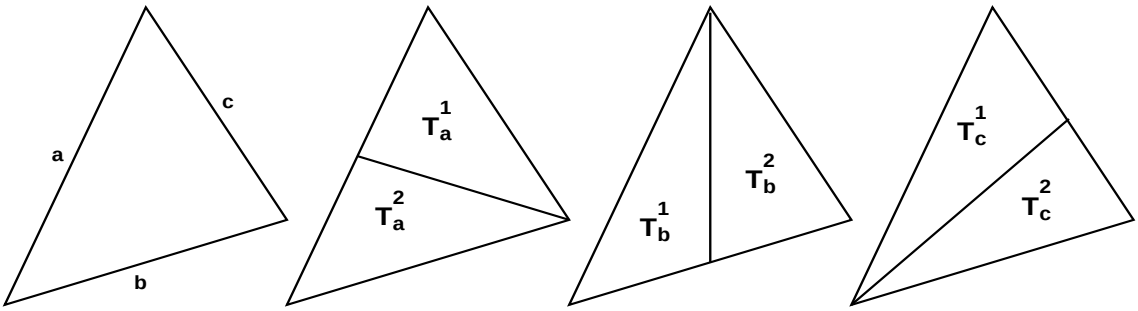

Figure 5: Notations in the proof of Proposition (3.3)

Proof: We introduce sub-triangles T_e^i, $i = 1, 2$ and $e = a, b, c$, as defined in Figure 5, which correspond to the three refinement scenarios. With such definitions, the following inequalities are easily derived from Proposition (3.2)

$$\begin{aligned}4\alpha_{T_a^2}(q) &= \mathbf{q}(b) \text{ (since } T_a^2 \text{ is obtuse)} \\ 4\alpha_{T_b^1}(q) &\geq \mathbf{q}(a) \\ 4\alpha_{T_c^1}(q) &\geq \mathbf{q}(a) \\ 4\alpha_{T_c^2}(q) &\geq \mathbf{q}(b)\end{aligned}$$

On the other hand, we shall prove

$$\alpha_{T_a^1}(q) \leq \alpha_{T_b^2}(q), \tag{3.15}$$

and

$$4\alpha_{T_a^1}(q) \leq \left(\frac{|b|_{\mathbf{q}} + |c|_{\mathbf{q}}}{2}\right)^2. \tag{3.16}$$

The proof of (3.15) follows from elementary geometric observations. Let L be a line which is parallel to c but does not contain it, and for $x \in L$ denote by T_x the triangle of vertices x and the end points of c. Denoting respectively by $u(x)$ and $v(x)$ the diameter of T_x and of its circumscribed circle for the $\mathbf{q}$-metric, we remark that these functions decrease monotonously as x tends to a point x_c which is the orthogonal projection (also for the $\mathbf{q}$-metric) of the mid-point of c onto L. Since the function $x \mapsto \alpha_{T_x}(q)$ is continuous in x and equal to $u(x)$ or $v(x)$ at all x, we conclude that this function also decreases monotonously as x tends to x_c. Applying this observation to the line that contains m_a and m_b the mid-points of a and b, and remarking that m_a is closer to x_c than m_b, we conclude that (3.15) holds.

From (3.15) and the first set of inequalities, we obtain the first statement of the theorem since

$$\begin{aligned} d_T(b,q) - d_T(a,q) &= \alpha_{T_b^1}(q) + \alpha_{T_b^2}(q) - \alpha_{T_a^2}(q) - \alpha_{T_a^1}(q) \\ &\geq \alpha_{T_b^1}(q) - \alpha_{T_a^2}(q) \geq \tfrac{1}{4}(\mathbf{q}(a) - \mathbf{q}(b)). \end{aligned}$$

The proof of (3.16) also follows from elementary geometric observations. In the case where T_a^1 is obtuse, one of its edges e is such that $4\alpha_{T_a^1} = \mathbf{q}(e)$, and (3.16) follows since $|e|_{\mathbf{q}} \leq \frac{1}{2}(|b|_{\mathbf{q}} + |c|_{\mathbf{q}})$ for all e, using triangle inequality. In the case where T_a^1 is acute, the center w of its circumscribed circle for the $\mathbf{q}$-metric is inside T_a^1, and its diameter is not larger than $\frac{1}{2}(|b|_{\mathbf{q}} + |c|_{\mathbf{q}})$ by convexity, as illustrated on Figure 6 when $\mathbf{q}$ is the euclidean metric.

From (3.16) and the first set of inequalities, we obtain the second statement of the theorem since

$$\begin{aligned} d_T(c,q) - d_T(a,q) &= \alpha_{T_c^2}(q) + \alpha_{T_c^1}(q) - \alpha_{T_a^2}(q) - \alpha_{T_a^1}(q) \\ &\geq \tfrac{1}{4}(\mathbf{q}(b) + \mathbf{q}(a) - \mathbf{q}(b) - (\tfrac{|b|_{\mathbf{q}}+|c|_{\mathbf{q}}}{2})^2) \\ &= \tfrac{1}{4}(\mathbf{q}(a) - (\tfrac{|b|_{\mathbf{q}}+|c|_{\mathbf{q}}}{2})^2). \end{aligned}$$

◇

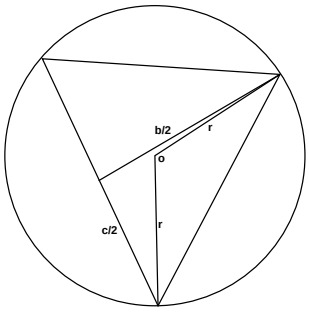

Figure 6: The case where T_a^1 is acute.

3.2 The L^2-based split

We now denote by

$$\beta_T(f) = \|f - \Pi_T f\|_{L^2(T)},$$

the orthogonal projection error in the L^2 norm. The decision function (2.12) now writes

$$d_T(e,f) = \beta_{T_{e,1}}(f)^2 + \beta_{T_{e,2}}(f)^2. \tag{3.17}$$

We shall prove that the refinement procedure based on (3.17) behaves in a similar way as (3.14).

Theorem 3.4 *If $(d,e) \in \{a,b,c\}$ are two edges such that $|d|_{\mathbf{q}} < |e|_{\mathbf{q}}$, then $d_T(e,q) < d_T(d,q)$. Therefore the refinement procedure based on* (3.17) *selects the longest edge in the sense of $|\cdot|_{\mathbf{q}}$.*

In order to prove this result, we first provide with an algebraic expression of $\beta_T(q)$ which is valid for any quadratic function q.

Proposition 3.5 *Let T be a triangle with edges a,b,c and area $|T|$, and let q be a quadratic function. Then*

$$\beta_T^2(q) = |T| \left(c_1(\mathbf{q}(a) + \mathbf{q}(b) + \mathbf{q}(c))^2 - c_2 \det(\mathbf{q}) |T|^2 \right). \tag{3.18}$$

with constants $c_1 = \frac{1}{1200}$ and $c_2 = c_1 \frac{64}{3} = \frac{4}{225}$.

Proof: We first prove (3.18) on the triangle R of vertices $\{(0,0),(0,1),(1,0)\}$. It is easy to compute the integrals on R of monomials $x^k y^l$, $k+l \leq 4$. Using these quantities, we can derive the orthogonal projection of a quadratic function thanks to a formal computing program, which gives us

$$\mathbf{q} = ux^2 + vy^2 + 2wxy \Rightarrow \Pi_R \mathbf{q} = -\frac{u+v+w}{10} + \frac{2x}{5}(2u+w) + \frac{2y}{5}(2v+w).$$

This yields the following expression for the L^2-squared error between q and its projection

$$\int_R (q - \Pi_R q)^2 = \frac{1}{300} \left(u^2 + (2uv)/3 + v^2 - 2uw - 2vw + (7w^2)/3 \right),$$

which is equivalent to (3.18).

For an arbitrary triangle T, using an affine bijective transformation ϕ from R to T, we have

$$\beta_T(q)^2 = J_\phi \beta_R(\tilde{q})^2,$$

where $\tilde{q} = q \circ \phi$ and J_ϕ is the constant jacobian of ϕ. Using the validity of (3.18) on R and the fact that $|T| = J_\phi |R|$, we thus obtain

$$\beta_T(q)^2 = |T| \left(c_1(\tilde{\mathbf{q}}(\tilde{a}) + \tilde{\mathbf{q}}(\tilde{b}) + \tilde{\mathbf{q}}(\tilde{c}))^2 - c_2 \det(\tilde{\mathbf{q}}) |R|^2 \right),$$

where $\tilde{\mathbf{q}}$ is the quadratic form associated to $\tilde{q}$ and $\tilde{e}$ denotes the edge segment of R mapped onto e by ϕ. Since $\tilde{\mathbf{q}}(\tilde{e}) = \mathbf{q}(e)$ and $\det(\tilde{\mathbf{q}}) = J_\phi^2 \det(\mathbf{q})$, we obtain (3.18) for T. ◇

We now prove the following result which clearly implies Theorem 3.4.

Corollary 3.6 *Let T be a triangle with edges a,b,c and area $|T|$, with $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}}$ and $|a|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$. Then,*

$$\begin{aligned} d_T(b,q) - d_T(a,q) &\geq \frac{5}{4} c_1 |T| (\mathbf{q}(a)^2 - \mathbf{q}(b)^2) \\ d_T(c,q) - d_T(a,q) &\geq \frac{5}{4} c_1 |T| (\mathbf{q}(a)^2 - \mathbf{q}(c)^2) \end{aligned}$$

Proof: The children triangles all have area $|T|/2$, and take their edges among a,b,c, $a/2,b/2,c/2$ and $\frac{a-b}{2}, \frac{b-c}{2}, \frac{c-a}{2}$ (recall that $a+b+c=0$), on which the quadratic form $\mathbf{q}$ can be expressed only in terms of $\mathbf{q}(a),\mathbf{q}(b),\mathbf{q}(c)$. In particular, for the last group $\frac{a-b}{2}, \frac{b-c}{2}, \frac{c-a}{2}$, we use the identity

$$\mathbf{q}(u+v) + \mathbf{q}(u-v) = 2\mathbf{q}(u) + 2\mathbf{q}(v),$$

which valid for all quadratic forms, and implies

$$\mathbf{q}(\frac{b-c}{2}) = \frac{\mathbf{q}(b) + \mathbf{q}(c)}{2} - \frac{\mathbf{q}(a)}{4}.$$

Using (3.18), this allows us to compute the local projection errors for the childrens of T. For example bisecting the edge a creates two children T' and T'' with edges $a/2, b, (c-b)/2$ and $a/2, c, (b-c)/2$, and therefore

$$\begin{aligned}\beta^2_{T'}(q) &= |T'|\left(c_1\left(\mathbf{q}(a/2)+\mathbf{q}(b)+\mathbf{q}(\frac{c-b}{2})\right)^2 - c_2\det(\mathbf{q})|T'|^2\right)\\ &= \frac{|T|}{2}\left(c_1\left(\frac{3\mathbf{q}(b)+\mathbf{q}(c)}{2}\right)^2 - c_2\det(\mathbf{q})\frac{|T|^2}{4}\right)\end{aligned}$$

and similarly

$$\beta^2_{T''}(q) = \frac{|T|}{2}\left(c_1\left(\frac{3\mathbf{q}(c)+\mathbf{q}(b)}{2}\right)^2 - c_2\det(\mathbf{q})\frac{|T|^2}{4}\right).$$

Adding up, we thus obtain

$$d_T(a,q) = |T|\left(c_1(3\mathbf{q}(b)+\mathbf{q}(c))^2 + c_1(\mathbf{q}(b)+3\mathbf{q}(c))^2 - 2c_2\det(\mathbf{q})|T|^2\right)/8. \tag{3.19}$$

Subtracting this from the analog expressions for $d_T(b,q)$ and $d_T(c,q)$, we obtain the announced inequalities. ◇

3.3 Convergence toward the optimal aspect ratio

We have proved that the refinement procedure - either based on the L^∞ or L^2 decision function - systematically picks the vertex opposite to the edge of largest length in the $\mathbf{q}$-metric. The purpose of this section is to study the iteration of several refinement steps and show that the generated triangle tend to adopt an optimal aspect ratio in the sense of the measure of non-degeneracy $\rho_{\mathbf{q}}(T)$ introduced in §2.

For this purpose, it will be convenient to introduce a close variant to $\rho_{\mathbf{q}}(T)$: if T is a triangle with edges a, b, c, such that $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$, we define

$$\sigma_{\mathbf{q}}(T) := \frac{\mathbf{q}(b)+\mathbf{q}(c)}{4|T|\sqrt{\det\mathbf{q}}} = \frac{|b|^2_{\mathbf{q}}+|c|^2_{\mathbf{q}}}{4|T|\sqrt{\det\mathbf{q}}}. \tag{3.20}$$

Using the inequalities $|b|^2_{\mathbf{q}}+|c|^2_{\mathbf{q}} \leq 2|a|^2_{\mathbf{q}}$ and $|a|^2_{\mathbf{q}} \leq 2(|b|^2_{\mathbf{q}}+|c|^2_{\mathbf{q}})$, we obtain the equivalence

$$\frac{\rho_{\mathbf{q}}(T)}{8} \leq \sigma_{\mathbf{q}}(T) \leq \frac{\rho_{\mathbf{q}}(T)}{2}. \tag{3.21}$$

Similar to $\rho_{\mathbf{q}}$, this quantity is invariant by a linear coordinate changes ϕ, in the sense that

$$\sigma_{\mathbf{q}\circ\phi}(T) = \sigma_{\mathbf{q}}(\phi(T)),$$

From (2.9) and (3.21) we can relate $\sigma_{\mathbf{q}}$ to the local approximation error.

Proposition 3.7 *The local L^p-approximation error satisfies*

$$e_T(q)_p = e_T(\mathbf{q})_p \sim |T|^{\frac{1}{\tau}}\sigma_{\mathbf{q}}(T), \quad \frac{1}{\tau} := \frac{1}{p}+1,$$

where the constants in the equivalence only depend on the choice of $\mathcal{A}_T$ between I_T, P_T or the best $L^p(T)$ approximation. The same holds with e_T replaced by α_T with $p=\infty$ or β_T with $p=2$.

Our next result shows that $\sigma_{\mathbf{q}}(T)$ is always reduced by the refinement procedure.

Proposition 3.8 *If T is a triangle with children T_1 and T_2 obtained by the refinement procedure for the quadratic function q, then*

$$\max\{\sigma_{\mathbf{q}}(T_1), \sigma_{\mathbf{q}}(T_2)\} \leq \sigma_{\mathbf{q}}(T).$$

Proof: Assuming that $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$, we know that the edge a is cut and that the children have area $|T|/2$ and edges $a/2, b, (c-b)/2$ and $a/2, (b-c)/2, c$ (recall that $a+b+c=0$). We then have

$$\begin{aligned} 2|T|\sqrt{\det \mathbf{q}}\ \sigma_{\mathbf{q}}(T_i) &\leq \mathbf{q}(a/2) + \mathbf{q}(\frac{b-c}{2}) && (3.22)\\ &= \mathbf{q}(\frac{b+c}{2}) + \mathbf{q}(\frac{b-c}{2}) && (3.23)\\ &= \frac{\mathbf{q}(b)+\mathbf{q}(c)}{2} && (3.24)\\ &= 2|T|\sqrt{\det \mathbf{q}}\ \sigma_{\mathbf{q}}(T). && (3.25) \end{aligned}$$

◇

Remark 3.9 *When* $\mathbf{q}$ *is the euclidean metric, the triangle that minimizes* $\sigma_{\mathbf{q}}$ *is the half square. This is consistent with the above result since it is the only triangle which is conformal to both of its children after one step of longest edge bisection.*

Remark 3.10 *It was already proved in [13] that longest edge bisection has the effect that the minimal angle in any triangle after an arbitrary number of refinements is at most twice the minimal angle of the initial triangle.*

Our next objective is to show that as we iterate the refinement process, the value of $\sigma_{\mathbf{q}}(T)$ becomes bounded independently of q for almost all generated triangles. For this purpose we introduce the following notation: if T is a triangle with edges such that $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$, we denote by $\psi_{\mathbf{q}}(T)$ the subtriangle of T obtained after bisection of a which contains the smallest edge c. We first establish inequalities between the measures $\sigma_{\mathbf{q}}$ and $\rho_{\mathbf{q}}$ applied to T and $\psi_{\mathbf{q}}(T)$.

Proposition 3.11 *Let* T *be a triangle, then*

$$\begin{aligned} \sigma_{\mathbf{q}}(\psi_{\mathbf{q}}(T)) &\leq \frac{5}{8}\rho_{\mathbf{q}}(T) && (3.26)\\ \rho_{\mathbf{q}}(\psi_{\mathbf{q}}(T)) &\leq \frac{\rho_{\mathbf{q}}(T)}{2}\left(1+\frac{16}{\rho^2_{\mathbf{q}}(T)}\right) && (3.27) \end{aligned}$$

Proof: We first prove (3.26). Obviously, $\psi_{\mathbf{q}}(T)$ contains one edge $s \in \{a,b,c\}$ from T, and one half edge $t \in \{\frac{a}{2}, \frac{b}{2}, \frac{c}{2}\}$ from T. Therefore

$$\sigma_{\mathbf{q}}(\psi_{\mathbf{q}}(T)) \leq \frac{|s|^2_{\mathbf{q}} + |t|^2_{\mathbf{q}}}{4|\psi_{\mathbf{q}}(T)|\sqrt{\det \mathbf{q}}} \leq \frac{|a|^2_{\mathbf{q}} + |\frac{a}{2}|^2_{\mathbf{q}}}{2|T|\sqrt{\det \mathbf{q}}} = \frac{5}{8}\rho_{\mathbf{q}}(T).$$

For the proof of (3.27), we restrict our attention to the case $\mathbf{q} = x^2+y^2$, without loss of generality thanks to the invariance formula (2.8).

Let T be a triangle with edges $|a| \geq |b| \geq |c|$. If h the width of T in the direction perpendicular to a, then

$$h = \frac{2|T|}{|a|} = \frac{2|a|}{\rho_{\mathbf{q}}(T)}.$$

The sub-triangle $\psi_{\mathbf{q}}(T)$ of T has edges $\frac{a}{2}, c, d$ where $d = \frac{b-c}{2}$, and the angles at the ends of $\frac{a}{2}$ are acute. Indeed

$$\langle \frac{a}{2}, c\rangle = \frac{1}{4}\left(|b|^2 - |a|^2 - |c|^2\right) \leq 0 \text{ and } \langle d, \frac{a}{2}\rangle = \frac{1}{4}\left(|c|^2 - |b|^2\right) \leq 0.$$

By Pythagora's theorem we thus find

$$\max\{\left|\frac{a}{2}\right|^2, |c|^2, |d|^2\} \leq \left|\frac{a}{2}\right|^2 + h^2 = \frac{|a|^2}{4}\left(1+\frac{16}{\rho^2_{\mathbf{q}}(T)}\right).$$

Dividing by the respective areas of T and $\psi_{\mathbf{q}}(T)$, we obtain the announced result. $\diamond$

Our next result shows that a significant reduction of $\sigma_{\mathbf{q}}$ occurs at least for one of the triangles obtained by three successive refinements, unless it has reached a small value of $\sigma_{\mathbf{q}}$.

Proposition 3.12 *Let T be a triangle such that $\sigma_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) \geq 5$. Then $\sigma_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) \leq 0.69\sigma_{\mathbf{q}}(T)$.*

Proof: The equation (3.26) tells us that $\rho_{\mathbf{q}}(\psi^2_{\mathbf{q}}(T)) \geq \frac{8}{5}\sigma_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) = 8$. Moreover, solving a second degree equation, the inequality (3.27) gives for any triangle S:

$$\text{If } \rho_{\mathbf{q}}(\psi_{\mathbf{q}}(S)) \geq 4 \text{ then } \rho_{\mathbf{q}}(S) \geq \rho_{\mathbf{q}}(\psi_{\mathbf{q}}(S)) + \sqrt{\rho_{\mathbf{q}}(\psi_{\mathbf{q}}(S))^2 - 16}.$$

Applying this to $S = \psi_{\mathbf{q}}(T)$ we find that $\rho_{\mathbf{q}}(\psi_{\mathbf{q}}(T)) \geq 14.9$. Applying it again to $S = T$ we obtain $\rho_{\mathbf{q}}(T) \geq 29.3$. Using again (3.27), it follows that

$$\frac{\rho(\psi^3_{\mathbf{q}}(T))}{\rho(T)} \leq \frac{1}{8}(1 + \frac{16}{\rho^2_{\mathbf{q}}(\psi^2_{\mathbf{q}}(T))})(1 + \frac{16}{\rho^2_{\mathbf{q}}(\psi_{\mathbf{q}}(T))})(1 + \frac{16}{\rho^2_{\mathbf{q}}(T)}) \leq 0.171$$

Finally, the inequalities (3.21) imply that

$$2\sigma_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) \leq \rho_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) \leq 0.171\rho_{\mathbf{q}}(T) \leq 0.171(8\sigma_{\mathbf{q}}(T))$$

which concludes the proof. $\diamond$

An immediate consequence of Propositions 3.8 and 3.12 is the following.

Corollary 3.13 *If $(T_i)_{i=1}^8$ are the eight children obtained from three successive refinement procedures from T for the function q, then*

- *for all i, $\sigma_{\mathbf{q}}(T_i) \leq \sigma_{\mathbf{q}}(T)$,*
- *there exists i such that $\sigma_{\mathbf{q}}(T_i) \leq 0.69\sigma_{\mathbf{q}}(T)$ or $\sigma_{\mathbf{q}}(T_i) \leq 5$.*

We are now ready to prove that most triangles tend to adopt an optimal aspect ratio as we iterate the refinement procedure.

Theorem 3.14 *Let T be a triangle, and $\mathbf{q}$ a positive definite quadratic function. Let $k = \frac{\ln \sigma_{\mathbf{q}}(T) - \ln 5}{-\ln(0.69)}$ Then after n applications of the refinement procedure starting from T, at most $Cn^k 7^{n/3}$ of the 2^n generated triangles satisfy $\sigma_{\mathbf{q}}(S) \geq 5$, where C is an absolute constant. Therefore the proportion of such triangles tends exponentially fast to 0 as $n \to +\infty$.*

Proof: If we prove the proposition for n multiple of 3, then it will hold for all n (with a larger constant) since $\sigma_{\mathbf{q}}$ decreases at each refinement step. We now assume that $n = 3m$, and consider the octree with root T obtained by only considering the triangles of generation $3k$ for $k = 0, \cdots, n$.

According to corollary 3.13, for each node of this tree, one of its eight children either checks $\sigma_{\mathbf{q}} \leq 5$ or has its non-degeneracy measure diminished by a factor $\theta := 0.69$. We remark that if $\sigma_{\mathbf{q}}$ is diminished at least k times on the path going from the root T to a leaf S, then $\sigma_{\mathbf{q}}(S) \leq 5$. As a consequence, the number $N(m)$ of triangles S which are such that $\sigma_{\mathbf{q}}(S) > 5$ within the generation level $n = 3m$ is bounded by the number of words in an eight letters alphabet $\{a_1, \cdots, a_8\}$ with length m and that use the letter a_8 at most k times, namely

$$N(m) \leq \sum_{l=0}^{k} \binom{m}{l} 7^{m-l} \leq Cm^k 7^m,$$

which is the announced result. $\diamond$

4 Quadratic functions of mixed sign

In this section, we study the algorithm when applied to a quadratic polynomial q such that $\det(\mathbf{q}) < 0$. We shall follow the same steps, and reach similar conclusions, as in the positive definite case, using a measure of non-degeneracy which is equivalent to $\rho_{\mathbf{q}}(T)$. If a triangle T has edges a, b, c such that $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$, we will still refer to a as the "longest" edge in the sense of $\mathbf{q}$, although $\mathbf{q}$ does not define a proper metric anymore.

The following inequalities that will be repeatedly used in this section can be derived when $\rho_{\mathbf{q}}(T)$ is large enough. We postpone their proof to the appendix.

Proposition 4.1 *Let T be a triangle such that $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$, and define $d = \frac{b-c}{2}$.*

$$\text{If } \rho_{\mathbf{q}}(T) \geq 4, \text{ then } \mathbf{q}(a)\mathbf{q}(b) \geq 0, \ |\mathbf{q}(a)| \geq 4|\mathbf{q}(c)| \text{ and } |\mathbf{q}(a)| \geq 4|\mathbf{q}(d)|. \tag{4.28}$$

$$\text{If } \rho_{\mathbf{q}}(T) \geq 8, \text{ then } |\mathbf{q}(a)| \leq \frac{3}{8}|\mathbf{q}(b) + \mathbf{q}(c)|. \tag{4.29}$$

4.1 The L^∞-based split

Theorem 4.2 *If $|\mathbf{q}(a)| > \max\{|\mathbf{q}(b)|, |\mathbf{q}(c)|\}$ and $\rho_{\mathbf{q}}(T) \geq 4$, then $d_T(a, q) < \min\{d_T(b, q), d_T(c, q)\}$. Therefore the refinement procedure based on* (3.14) *selects the longest edge in the sense of $\mathbf{q}$.*

This theorem is very similar to the one for positive quadratic functions. In order to prove it, we first study the interpolation error which has a simple form in this context.

Proposition 4.3 *Let T be a triangle with edges a, b, c. Then*

$$\alpha_T(q) = \frac{1}{4}\max\{|\mathbf{q}(a)|, |\mathbf{q}(b)|, |\mathbf{q}(c)|\}.$$

Proof: Let x_0 be the point of T at which the interpolation error is attained: $x_0 = \operatorname{argmax}_T |q - I_T q|$. If x_0 is in the interior of T, then it must be a local extremum of $q - I_T q$. However this function has only one critical point on $\mathbb{R}^2$, which is not an extremum since $\mathbf{q}$ has mixed signature. Therefore x_0 must lie an edge. On each edge of T, the function $q - I_T q$ is a one dimensional quadratic function vanishing at the endpoints. It follows that x_0 must lie in the middle of one edge and the result follows. ◇

Proposition 4.4 *Let T be a triangle with edges $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$ and verifying $\rho_{\mathbf{q}}(T) \geq 4$. Then*

$$\begin{aligned} d_T(b, q) - d_T(a, q) &\geq \frac{|\mathbf{q}(a)| - |\mathbf{q}(b)|}{8} \\ d_T(c, q) - d_T(a, q) &\geq \frac{|\mathbf{q}(a)|}{8} \end{aligned}$$

Proof: The bisection through the edge a creates two sub-triangles T_a^1, T_a^2 of edges respectively $\frac{a}{2}, b, d$ and $\frac{a}{2}, c, d$. Using the last two inequalities in (4.28) we obtain that $4\alpha_{T_a^1} = \max\{|\mathbf{q}(a/2)|, |\mathbf{q}(b)|\}$ and $4\alpha_{T_a^2} = |\mathbf{q}(a/2)|$. Therefore

$$4d_T(a, q) = \frac{|\mathbf{q}(a)|}{4} + \max\{\frac{|\mathbf{q}(a)|}{4}, |\mathbf{q}(b)|\}.$$

On the other hand, the choice of bisecting the edge b creates two subtriangles respectively containing the edges a and $\frac{b}{2}$, and the choice of bisecting the edge c creates two subtriangles respectively containing the edges a and b. This provides us with the lower bounds

$$\begin{aligned} 4d_T(b, q) &\geq |\mathbf{q}(a)| + \frac{|\mathbf{q}(b)|}{4}, \\ 4d_T(c, q) &\geq |\mathbf{q}(a)| + |\mathbf{q}(b)|. \end{aligned}$$

The proposition follows easily, distinguishing between the two cases $|\mathbf{q}(a)| \leq 4|\mathbf{q}(b)|$ and $|\mathbf{q}(a)| \geq 4|\mathbf{q}(b)|$. ◇

4.2 The L^2-based split

The same conclusions can be reached for the refinement procedure based on (3.17).

Theorem 4.5 *If* $|\mathbf{q}(a)| > \max\{|\mathbf{q}(b)|, |\mathbf{q}(c)|\}$ *and* $\rho_{\mathbf{q}}(T) \geq 4$, *then* $d_T(a,q) < \min\{d_T(b,q), d_T(c,q)\}$. *Therefore the refinement procedure based on* (3.17) *selects the longest edge in the sense of* $\mathbf{q}$.

Proof: The expression found in (3.19) remains valid when $\det(\mathbf{q}) < 0$. Substituting a by b or c and subtracting, we obtain

$$\begin{aligned} d_T(b,q) - d_T(a,q) &= \frac{5c_1}{2}|T|(\mathbf{q}(a) - \mathbf{q}(b))(s + \frac{\mathbf{q}(b)}{5}), \\ d_T(c,q) - d_T(a,q) &= \frac{5c_1}{2}|T|(\mathbf{q}(a) - \mathbf{q}(c))(s + \frac{\mathbf{q}(c)}{5}), \end{aligned}$$

where $s = \mathbf{q}(a) + \mathbf{q}(b) + \mathbf{q}(c)$. Using (4.28), we see that $s + \frac{\mathbf{q}(b)}{5}$, $s + \frac{\mathbf{q}(c)}{5}$, $\mathbf{q}(a) - \mathbf{q}(b)$ and $\mathbf{q}(a) - \mathbf{q}(c)$ all have the same sign as $\mathbf{q}(a)$ and are non-zero. It follows that $d_T(a,q) < \min\{d_T(b,q), d_T(c,q)\}$. ⋄

4.3 Convergence toward the optimal aspect ratio.

We have proved that the refinement procedure - either based on the L^∞ or L^2 decision function - systematically picks the longest edge in the sense of $\mathbf{q}$. Similarly to the positive definite case, we now study the iteration of several refinement steps and show that the generated triangles tend to adopt an optimal "aspect ratio" in the sense of the measure of non-degeneracy $\rho_{\mathbf{q}}(T)$ introduced in §2.

As in §3.3, we introduce a close variant to $\rho_{\mathbf{q}}(T)$. If T is a triangle with edges a, b, c, we define

$$\sigma_{\mathbf{q}}(T) := \frac{\min\left(|\mathbf{q}(a) + \mathbf{q}(b)|, |\mathbf{q}(b) + \mathbf{q}(c)|, |\mathbf{q}(c) + \mathbf{q}(a)|\right)}{4|T|\sqrt{|\det \mathbf{q}|}}. \tag{4.30}$$

Note that if $\mathbf{q}$ was a positive quadratic form, this definition is consistent with (3.20). We define our measure of non-degeneracy $\kappa_{\mathbf{q}}$ by

$$\kappa_{\mathbf{q}}(T) = \max(\sigma_{\mathbf{q}}(T), \frac{5}{2}). \tag{4.31}$$

We first show that the quantities $\kappa_{\mathbf{q}}$ and $\rho_{\mathbf{q}}$ are equivalent.

Proposition 4.6 *For any triangle* T, *one has*

$$2\sigma_{\mathbf{q}}(T) \leq \rho_{\mathbf{q}}(T), \tag{4.32}$$

and

$$\frac{4}{5}\kappa_{\mathbf{q}}(T) \leq \rho_{\mathbf{q}}(T) \leq \frac{32}{3}\kappa_{\mathbf{q}}(T). \tag{4.33}$$

Proof: The inequality (4.32) follows directly from the triangle inequality:

$$2|T|\sqrt{|\det \mathbf{q}|}\sigma_{\mathbf{q}}(T) \leq \frac{|\mathbf{q}(b) + \mathbf{q}(c)|}{2} \leq |\mathbf{q}(a)| \leq |T|\sqrt{|\det \mathbf{q}|}\rho_{\mathbf{q}}(T).$$

As mentionned earlier, $\rho_{\mathbf{q}}(T)$ is always larger than 2 and therefore (4.32) implies the left inequality in (4.33).

It remains to prove the right inequality in (4.33). If $\rho_{\mathbf{q}}(T) \leq 8$, it is immediate since $\sigma_{\mathbf{q}}(T) \geq \frac{5}{2}$ and $\frac{32}{3}\frac{5}{2} \geq 8$. If $\rho_{\mathbf{q}}(T) \geq 8$ we infer from (4.28) that $|\mathbf{q}(b) + \mathbf{q}(c)| \leq |\mathbf{q}(a) + \mathbf{q}(c)| \leq |\mathbf{q}(a) + \mathbf{q}(b)|$, and from (4.29) that $|\mathbf{q}(a)| \leq \frac{8}{3}|\mathbf{q}(b) + \mathbf{q}(c)|$. It follows that

$$\rho_{\mathbf{q}}(T) = \frac{|\mathbf{q}(a)|}{|T|\sqrt{|\det \mathbf{q}|}} \leq \frac{8}{3}\frac{|\mathbf{q}(b) + \mathbf{q}(c)|}{|T|\sqrt{|\det \mathbf{q}|}} = \frac{32}{3}\sigma_{\mathbf{q}}(T) \leq \frac{32}{3}\kappa_{\mathbf{q}}(T),$$

which concludes the proof. ◇

Similar to $\rho_{\mathbf{q}}$, the quantity $\kappa_{\mathbf{q}}$ is invariant by a linear coordinate changes ϕ, in the sense that

$$\kappa_{\mathbf{q}\circ\phi}(T) = \kappa_{\mathbf{q}}(\phi(T)).$$

Our next result shows that $\kappa_{\mathbf{q}}(T)$ is always reduced by the refinement procedure.

Proposition 4.7 *If T is a triangle with children T_1 and T_2 obtained by the refinement procedure for the quadratic function q, then*

$$\max\{\kappa_{\mathbf{q}}(T_1), \sigma_{\mathbf{q}}(T_2)\} \leq \kappa_{\mathbf{q}}(T).$$

Proof: Let us assume that a is the longest edge in the sense of $\mathbf{q}$. In the case where $\rho_{\mathbf{q}}(T) \geq 4$, we already noticed in the proof of Proposition 4.6 that

$$\sigma_{\mathbf{q}}(T) = \frac{|\mathbf{q}(b) + \mathbf{q}(c)|}{4|T|\sqrt{|\det \mathbf{q}|}}.$$

Moreover, according to the results established in §4.1 and 4.2, the edge a is selected by both decision functions. It follows that children T_i have edges $a/2, b, (c-b)/2$ and $a/2, (b-c)/2, c$ (recall that $a+b+c = 0$). We thus have

$$\begin{array}{rcl} 2|T|\sqrt{|\det \mathbf{q}|}\ \sigma_{\mathbf{q}}(T_i) & \leq & \left|\mathbf{q}(a/2) + \mathbf{q}(\frac{b-c}{2})\right| \\ & = & \left|\mathbf{q}(\frac{b+c}{2}) + \mathbf{q}(\frac{b-c}{2})\right| \\ & = & \frac{|\mathbf{q}(b)+\mathbf{q}(c)|}{2} \\ & = & 2|T|\sqrt{|\det \mathbf{q}|}\ \sigma_{\mathbf{q}}(T). \end{array}$$

We have proved that $\sigma_{\mathbf{q}}(T_i) \leq \sigma_{\mathbf{q}}(T)$, and it readily follows that $\kappa_{\mathbf{q}}(T_i) \leq \kappa_{\mathbf{q}}(T)$.

In the case where $\rho_{\mathbf{q}}(T) \leq 4$, we remark that T_i contains at least one edge from T, say $s \in \{a, b, c\}$ and one half-edge $t \in \{\frac{a}{2}, \frac{b}{2}, \frac{c}{2}\}$. This provides an upper bound for $\sigma_{\mathbf{q}}$:

$$\sigma_{\mathbf{q}}(T_i) \leq \frac{|\mathbf{q}(s) + \mathbf{q}(t)|}{2|T|\sqrt{|\det \mathbf{q}|}} \leq \frac{|\mathbf{q}(a)| + |\mathbf{q}(\frac{a}{2})|}{2|T|\sqrt{|\det \mathbf{q}|}} = \frac{5}{8}\rho_{\mathbf{q}}(T) \leq \frac{5}{2}. \tag{4.34}$$

Therefore $\kappa_{\mathbf{q}}(T_i) = \frac{5}{2} \leq \kappa_{\mathbf{q}}(T)$. ◇

Our next objective is to show that as we iterate the refinement process, the value of $\kappa_{\mathbf{q}}(T)$ becomes bounded independently of q for almost all generated triangles. If T is a triangle such that $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$ and if the edge a is cut (which is the case as soon as $\rho_{\mathbf{q}}(T) \geq 4$) we define $\psi_{\mathbf{q}}(T)$ as the subtriangle containing the edge c. We first prove a result which is analogous to Proposition 3.12.

Proposition 4.8 *If T is a triangle such that $\kappa_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) > \frac{5}{2}$, then $\kappa_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) \leq \frac{2}{3}\kappa_{\mathbf{q}}(T)$.*

Proof: Let S be a triangle such that $\kappa_{\mathbf{q}}(\psi_{\mathbf{q}}(S)) > \frac{5}{2}$. According to (4.34), one must have $\rho_{\mathbf{q}}(S) > 4$. Assuming that the edges of S satisfy $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$, since the three edges of $\psi_{\mathbf{q}}(S)$ are $\frac{a}{2}, c, \frac{b-c}{2}$, we infer from (4.28) that the longest edge of $\psi_{\mathbf{q}}(S)$ in the sense of $\mathbf{q}$ is $\frac{a}{2}$.

Since $\kappa_{\mathbf{q}}(\psi_{\mathbf{q}}(T)) \geq \kappa_{\mathbf{q}}(\psi_{\mathbf{q}}(T)) \geq \kappa_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) > \frac{5}{2}$, we can apply this observation to the triangles T, $\psi_{\mathbf{q}}(T)$ and $\psi^2_{\mathbf{q}}(T)$. Therefore, denoting by a the longest edge of T in the sense of $\mathbf{q}$, we find that $\frac{a}{8}$ is the longest edge of $\psi^3_{\mathbf{q}}(T)$. Since $|\psi^3_{\mathbf{q}}(T)| = |T|/8$, we obtain that $\rho_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) = \rho_{\mathbf{q}}(T)/8$. Using the results of Proposition 4.6, we thus have

$$\kappa_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) = \sigma_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) \leq \frac{1}{2}\rho_{\mathbf{q}}(\psi^3_{\mathbf{q}}(T)) = \frac{1}{16}\rho_{\mathbf{q}}(T) \leq \frac{2}{3}\kappa_{\mathbf{q}}(T),$$

which concludes the proof. ◇

An immediate consequence of Propositions 4.7 and 4.8 is the following.

Corollary 4.9 *If $(T_i)_{i=1}^8$ are the 8 children obtained from 3 successive refinement procedures from T for the function q,*

- $\forall i, \kappa_{\mathbf{q}}(T_i) \leq \kappa_{\mathbf{q}}(T)$
- $\exists i, \kappa_{\mathbf{q}}(T_i) \leq \frac{2}{3}\kappa_{\mathbf{q}}(T)$ *or* $\kappa_{\mathbf{q}}(T_i) = \frac{5}{2}$.

We finally obtain the following result which proof is exactly similar to the one of Theorem 3.14.

Theorem 4.10 *Let T be a triangle, and $\mathbf{q}$ a quadratic function of mixed type. Let $k = \frac{\ln(2\kappa_{\mathbf{q}}(T)/5)}{\ln 3 - \ln 2}$ Then after n applications of the refinement procedure starting from T, at most $Cn^k 7^{n/3}$ of the 2^n generated triangles are such that $\kappa_{\mathbf{q}}(S) > \frac{5}{2}$ where C is an absolute constant. Therefore the proportion of such triangles tends exponentially to 0 as $n \to +\infty$.*

5 The case of strictly convex functions

The goal of this section is to prove that the approximation error in the greedy algorithm satisfies the estimate (1.1) corresponding to an optimal triangulation. Our main result is so far limited to the case where f is strictly convex.

Theorem 5.1 *Let $f \in C^2(\Omega)$ be such that*

$$d^2 f(x) \geq \alpha I, \quad x \in \Omega$$

for some arbitrary but fixed $\alpha > 0$ independent of x. Let f_N be the approximant obtained by the greedy algorithm for the L^p metric, using the L^∞ decision function (2.13). Then, there exists $N_0 = N_0(f)$ such that for $N \geq N_0$,

$$\|f - f_N\|_{L^p} \leq CN^{-1} \|\sqrt{\det(d^2 f)}\|_{L^\tau}, \tag{5.35}$$

where $\frac{1}{\tau} = \frac{1}{p} + 1$ and where C is an absolute constant.

The extension of this result to strictly concave functions is immediate by a change of sign. Its extension to arbitrary C^2 functions is so far uncomplete, although plausible, as we explain in the concluding remarks of §6. The proof of Theorem 5.1 will use the fact that a strictly convex C^2 function is *locally* close to a quadratic function with positive definite hessian, which allows us to exploit the results obtained in §4 for these particular functions.

5.1 A perturbation result

Our first step towards the proof of Theorem 5.1 therefore consists in describing how the results obtained in §3 extend to arbitrary convex functions which are closeon a given triangle T to a quadratic function in the C^2 norm.

In the following, we fix $\alpha > 0$ and $0 < \varepsilon < \alpha$, and consider a pair of functions (q, f) defined on a triangle T, such that

- q is a quadratic polynomial such that $d^2 q \geq \alpha I$.
- f is a C^2 function and $\|d^2 f - d^2 q\|_{L^\infty(T)} \leq \varepsilon$ (with $\|\cdot\|$ the spectral norm).

We also define

$$\mu := \frac{\varepsilon}{\alpha} < 1.$$

Since the greedy algorithm is driven by the quantities $e_T(f)_p$, and $\alpha_T(f)$ or $\beta_T(f)$, we first need to show that these quantities do not differ too much from those associated with q.

Proposition 5.2 *There exists an absolute constant $C_e > 1$ such that*

$$(1 - C_e\mu)e_T(q)_p \leq e_T(f)_p \leq (1 + C_e\mu)e_T(q)_p. \tag{5.36}$$

The same holds for α_T and β_T in place of e_T with absolute constants C_α and C_β in place of C_e. Moreover,

$$(1 - \mu)\sqrt{\det \mathbf{q}} \leq \sqrt{\det d^2 f} \leq (1 + \mu)\sqrt{\det \mathbf{q}}. \tag{5.37}$$

Therefore, using Proposition 3.7*, and assuming that $\mu \leq c_e := \frac{1}{2C_e}$, we have with $\frac{1}{\tau} := 1 + \frac{1}{p}$,*

$$e_T(f)_p \sim e_T(q)_p \sim \sigma_{\mathbf{q}}(T)\|\sqrt{|\det \mathbf{q}|}\|_{L^\tau(T)} \sim \sigma_{\mathbf{q}}(T)\|\sqrt{\det d^2 f}\|_{L^\tau(T)},$$

with absolute constants in the equivalence.

Proof: Using Taylor formula, we know that there exists an affine function π such that

$$\|f - q - \pi\|_{L^\infty(T)} \leq \frac{1}{2}\varepsilon h_T^2.$$

We remark that for both $\mathcal{A}_T = I_T$ or P_T, the operator $I - \mathcal{A}_T$ is bounded in the $L^\infty(T)$ norm with spectral norm $C = \|I - \mathcal{A}_T\|$ independent of T (by affine change of coordinate). Note that $C = 1$ in the case $\mathcal{A}_T = I_T$. Since $\mathcal{A}_T\pi = \pi$, we thus have

$$\|(f - \mathcal{A}_T f) - (q - \mathcal{A}_T q)\|_{L^\infty(T)} \leq C\|f - q - \pi\|_{L^\infty(T)} \leq \frac{C}{2}\varepsilon h_T^2,$$

where C is an absolute constant. Therefore

$$\|(f - \mathcal{A}_T f) - (q - \mathcal{A}_T q)\|_{L^p(T)} \leq C\|f - q - \pi\|_{L^p(T)} \leq \frac{C}{2}|T|^{1/p}\varepsilon h_T^2,$$

It follows that

$$e_T(q)_p - \frac{C}{2}|T|^{1/p}\varepsilon h_T^2 \leq e_T(f)_p \leq e_T(q)_p + \frac{C}{2}|T|^{1/p}\varepsilon h_T^2.$$

On the other hand, we know from Proposition 2.1

$$e_T(q)_p \sim |T|^{1/p}\max\{\mathbf{q}(a), \mathbf{q}(b), \mathbf{q}(c)\} \geq |T|^{1/p}\alpha h_T^2,$$

we obtain (5.36).

For the proof of (5.37), we remark that if M and N are symmetric matrices such that $\|N\| \leq \varepsilon$ and $M \geq \alpha I$, with $0 < \varepsilon < \alpha$, then

$$(\lambda_1 - \varepsilon)(\lambda_2 - \varepsilon) \leq \det(M + N) \leq (\lambda_1 + \varepsilon)(\lambda_2 + \varepsilon),$$

where (λ_1, λ_2) are the eigenvalues of M which are both larger than α so that

$$(1 - \varepsilon/\alpha)^2 \leq \det(M + N)/\det M \leq (1 + \varepsilon/\alpha)^2.$$

Applying this to $M = d^2 q$ and $N = d^2 f - d^2 q$, this gives the desired result. ⋄

We have proved in the §4 that if q is a positive quadratic function the decision functions $d_T(e, q)$, either defined by (3.14) or (3.17), always prescribes to split towards the *longest edge* in the sense of the $\mathbf{q}$-metric. We now want to identify as much as possible the choice prescribed by $d_T(x, f)$. This motivates the following definition:

Definition 5.3 *Let T be a triangle with edges a, b, c. A δ-near longest edge bisection with respect to the $\mathbf{q}$-metric is a bisection of any edge $e \in \{a, b, c\}$ such that*

$$\mathbf{q}(e) \geq (1 - \delta)\max\{\mathbf{q}(a), \mathbf{q}(b), \mathbf{q}(c)\}$$

Using the closeness between f and q, we can prove that the bisection choice prescribed by $d_T(x, f)$ is of the above type, if μ is small enough.

Lemma 5.4 *The bisection of T prescribed by $d_T(x,f)$ is a $C_1\mu$-near longest edge bisection with respect to the* $\mathbf{q}$*-metric, where C_1 is an absolute constant.*

Proof: Let us assume $|a|_{\mathbf{q}} \geq |b|_{\mathbf{q}} \geq |c|_{\mathbf{q}}$. From Proposition 5.2, we have

$$|d_T(e,q) - d_T(e,f)| \leq C_0\mu d_T(e,q),$$

with d_T either defined by (3.14) or (3.17) and C_0 an absolute constant.

We shall first assume that $\mu \leq \frac{1}{2C_0}$. If $d_T(e,f) \leq d_T(a,f)$ for some edge e, we have

$$d_T(e,q) \leq \frac{1+C_0\mu}{1-C_0\mu} d_T(a,q) \leq (1+4C_0\mu) d_T(a,q).$$

We now want to show that $\mathbf{q}(e) \leq (1-C_1\mu)\max\{\mathbf{q}(a),\mathbf{q}(b),\mathbf{q}(c)\}$. We only need to consider the cases where $e=b$ or c.

For this purpose, we distinguish between the two types of decision functions. When d_T is defined by (3.14), we know from Proposition 3.3,

$$\begin{aligned} 4(d_T(b,q) - d_T(a,q)) &\geq \mathbf{q}(a) - \mathbf{q}(b) \\ 4(d_T(c,q) - d_T(a,q)) &\geq \mathbf{q}(a) - (\frac{|b|_{\mathbf{q}} + |c|_{\mathbf{q}}}{2})^2 \geq \frac{\mathbf{q}(a) - \mathbf{q}(c)}{2} \end{aligned}$$

On the other hand, according to the estimates obtained in the proof of Proposition 3.3, we have

$$d_T(a,q) = \alpha_{T_a^1}(q) + \alpha_{T_a^2}(q) \leq \frac{\mathbf{q}(b)}{2} \leq \frac{\mathbf{q}(a)}{2}.$$

Therefore, if $e=b$, since $d_T(b,q) \leq (1+4C_0\mu)d_T(a,q)$, it follows that

$$\mathbf{q}(a) - \mathbf{q}(b) \leq 4(d_T(b,q) - d_T(a,q)) \leq 16C_0\mu d_T(a,q) \leq 8C_0\mu\mathbf{q}(a).$$

Similarly, if $e=c$, since $d_T(c,q) \leq (1+4C_0\mu)d_T(a,q)$, it follows that

$$\mathbf{q}(a) - \mathbf{q}(c) \leq 8(d_T(b,q) - d_T(a,q)) \leq 32C_0\mu d_T(a,q) \leq 16C_0\mu\mathbf{q}(a).$$

We have thus obtained the desired result with $C_1 = 16C_0$. When d_T is defined by (3.17), we know from corollary 3.6 that for $e=b$ or c,

$$d_T(e,q) - d_T(a,q) \geq \frac{5}{4}c_1|T|(\mathbf{q}(a)^2 - \mathbf{q}(e)^2) \geq \frac{5}{4}c_1|T|\mathbf{q}(a)(\mathbf{q}(a) - \mathbf{q}(e)).$$

On the other hand, according to (3.19), we also have $d_T(a,q) \leq 4|T|c_1\mathbf{q}(a)^2$. Combining both, we obtain that

$$\mathbf{q}(a) - \mathbf{q}(e) \leq \frac{4}{5c_1|T|\mathbf{q}(a)}(d_T(e,q) - d_T(a,q)) \leq \frac{16C_0\mu}{5c_1|T|\mathbf{q}(a)} d_T(a,q) \leq \frac{64C_0\mu}{5}\mathbf{q}(a).$$

We have thus obtained the desired result with $C_1 = \frac{64}{5}C_0$.

Finally, we notice that in the case where $\mu > \frac{1}{2C_0}$, we have $C_1\mu > 1$ for both values of C_1 obtained above. In that case, the result is trivial since any bisection is a 1-near longest edge bisection. ⋄

We now introduce a perturbed version of the estimates describing the decay of the non-degeneracy measure which were obtained Corollary 3.13.

Proposition 5.5 *If $(T_i)_{i=1}^8$ are the eight children obtained from three successive refinement procedures from T for the function f, then*

- *for all i, $\sigma_{\mathbf{q}}(T_i) \leq \sigma_{\mathbf{q}}(T)(1+C_2\mu)$,*
- *there exists i such that $\sigma_{\mathbf{q}}(T_i) \leq 0.69\sigma_{\mathbf{q}}(T)(1+C_2\mu)$ or $\sigma_{\mathbf{q}}(T_i) \leq M$,*

where C_2 is an absolute constant and $M = 4(1 + C_2c_0)$.

Proof: Let $(T_i')_{i=1}^8$ be the eight children obtained from three successive refinement procedures from T for the function q. We know from Corollary 3.13 that $\sigma_{\mathbf{q}}(T_i') \leq \sigma_{\mathbf{q}}(T)$ for all i and that there exists i such that either $\sigma_{\mathbf{q}}(T_i') \leq \frac{3}{4}\sigma_{\mathbf{q}}(T)$ or $\sigma_{\mathbf{q}}(T_i') \leq 4$. The triangles T_i might differ from the T_i' but we shall prove that for a suitable ordering of the T_i,

$$|\sigma_{\mathbf{q}}(T_i) - \sigma_{\mathbf{q}}(T_i')| \leq C_2\mu\sigma_{\mathbf{q}}(T_i'), \tag{5.38}$$

which clearly implies our result, with $M = 4(1 + C_2c_0)$. The idea is to use the fact that the bisection of T prescribed by $d_T(x, f)$ is a near longest edge bisection with respect to the $\mathbf{q}$-metric as shown by Lemma 5.4 in order to prove that the triangles T_i and T_i' have similar shape. For this purpose, we introduce a distance between triangles: if T_1 and T_2 have edges a_1, b_1, c_1 and a_2, b_2, c_2 such that $\mathbf{q}(a_1) \geq \mathbf{q}(b_1) \geq \mathbf{q}(c_1)$ and $\mathbf{q}(a_2) \geq \mathbf{q}(b_2) \geq \mathbf{q}(c_2)$, we define

$$\delta_{\mathbf{q}}(T_1, T_2) = \max\{|\mathbf{q}(a_1) - \mathbf{q}(a_2)|, |\mathbf{q}(b_1) - \mathbf{q}(b_2)|, |\mathbf{q}(c_1) - \mathbf{q}(c_2)|\}.$$

Note that $\delta_{\mathbf{q}}$ is a distance up to rigid transformations. Using this distance, we now compare (R_1, U_1) and (R_2, U_2), the two pairs of childrens obtained from one refinement procedure from T_1 for the functions q and T_2 for the function f respectively. Up to a permutation, R_1 and U_1 have edge vectors $b_1, a_1/2, (c_1 - b_1)/2$ and $c_1, a_1/2, (b_1 - c_1)/2$. From Lemma 5.4, we see that two situations might occur for the pair (R_2, U_2):

- $\mathbf{q}(e) < (1 - C_1\mu)\mathbf{q}(a_2)$ with C_1 the constant in 5.4 for $e = b_2$ and c_2. In such a case the refinement procedure for f bisects T_2 towards a_2, so that up to a permutation, R_2 and U_2 have edge vectors $b_2, a_2/2, (c_2 - b_2)/2$ and $c_2, a_2/2, (b_2 - c_2)/2$. Using that $q((c - b)/2) = q(c)/2 + q(b)/2 - q(a)/4$ when $a + b + c = 0$, it clearly follows that

$$\max\{\delta_{\mathbf{q}}(R_1, R_2), \delta_{\mathbf{q}}(U_1, U_2)\} \leq \frac{5}{4}\delta_{\mathbf{q}}(T_1, T_2).$$

- $\mathbf{q}(e) \geq (1-C_1\mu)\mathbf{q}(a_2)$ with C_1 the constant in 5.4 for some $e = b_2$ or c_2. In such a case the refinement procedure for f may bisect T_2 say towards b_2, so that up to a permutation, R_2 and U_2 have edge vectors $a_2, b_2/2, (c_2 - a_2)/2$ and $c_2, b_2/2, (b_2 - c_2)/2$. But since $|\mathbf{q}(b_2) - \mathbf{q}(a_2)|_{\mathbf{q}} \leq C_1\mu\mathbf{q}(a_2)$, we obtain that

$$\max\{\delta_{\mathbf{q}}(R_1, R_2), \delta_{\mathbf{q}}(U_1, U_2)\} \leq \frac{5}{4}\delta_{\mathbf{q}}(T_1, T_2) + C_1\mu\mathbf{q}(a_2).$$

Applying the last estimate with $T_1 = T_2 = T$ and iterating it on the childrens and grand-childrens of T, we obtain that up to a suitable ordering the triangles $(T_i)_{i=1}^8$ and $(T_i')_{i=1}^8$ satisfy

$$\max_{i=1,\cdots,8} \delta_{\mathbf{q}}(T_i, T_i') \leq (1 + \frac{5}{4} + (\frac{5}{4})^2)C_1\mu\mathbf{q}(a) = \frac{61}{16}C_1\mu\mathbf{q}(a),$$

where a is the longest edge of T in the $\mathbf{q}$-metric. In order to conclude, we remark that according to the definition of $\sigma_{\mathbf{q}}$, and using the fact that T_i and T_i' have equal area, we have

$$|\sigma_{\mathbf{q}}(T_i) - \sigma_{\mathbf{q}}(T_i')| \leq \frac{2\delta(T_i, T_i')}{4|T_i|\sqrt{\det(\mathbf{q})}} \leq \frac{61}{8}C_1\mu\frac{q(a)}{4|T_i|\sqrt{\det(\mathbf{q})}} \leq \frac{61}{4}C_1\mu\sigma_{\mathbf{q}}(T).$$

We therefore obtain the desired result with $C_2 := \frac{61}{4}C_1$. ◇

5.2 Local optimality

Our next step towards the proof of Theorem 5.1 is to show that the triangulation produced by the greedy algorithm is locally optimal in the following sense: if the refinement procedure for the function f produces a triangle $T \in \mathcal{D}$ on which f is close enough to a quadratic function q, then the triangles which are generated from the refinement of T tend to adopt an optimal aspect ratio in the $\mathbf{q}$-metric, and a local version of the optimal estimate (1.1) holds on T.

We first prove that most triangles adopt an optimal aspect ratio as we iterate the refinement procedure. Our goal is thus to obtain a result similar to Theorem 3.14 which was restricted to quadratic functions. However, due to the perturbations by $C_2\mu$ that appear in Proposition 5.5, the formulation will be slightly different, yet sufficient for our purposes: we shall prove that the measure of non-degeneracy becomes bounded by an absolute constant in an average sense, as we iterate the refinement procedure.

For $r > 0$, we define the average r-th power of the measure of non-degeneracy of the 2^{3n} triangles obtained from T after $3n$ iterations of the refinement procedure:

$$\overline{\sigma_{\mathbf{q}}^r(n)} = \frac{1}{2^{3n}} \sum_{T' \in \mathcal{T}_n^u(T)} \sigma_{\mathbf{q}}^r(T'),$$

where $\mathcal{T}_n^u(T)$ is the triangulation of T which is built by iteratively applying the refinement procedure for the function f starting from T up to $3n$ generation levels. Note that $\#(\mathcal{T}_n^u(T)) = 2^{3n}$ and $|T'| = 2^{-3n}|T|$ for all $T' \in \mathcal{T}_n^u(T)$. We also define

$$\gamma(r,\mu) := \frac{1}{8}\Big(0.69(1+C_2\mu)\Big)^r + \frac{7}{8}(1+C_2\mu)^r,$$

where C_2 is the constant in Proposition 5.5. One easily checks that for any $r > 0$, there exists $\mu(r) > 0$ and $0 < \gamma(r) < 1$ such that $\gamma(r,\mu) \leq \gamma(r)$, if $0 < \mu < \mu(r)$.

Proposition 5.6 *Assume that $0 < \mu \leq \mu(r)$. We then have*

$$\overline{\sigma_{\mathbf{q}}^r(n)} \leq \sigma_{\mathbf{q}}^r(T)\gamma(r)^n + \frac{M^r}{8(1-\gamma(r))},$$

where M is the constant in Proposition 5.5. Therefore

$$\overline{\sigma_{\mathbf{q}}^r(n)} \leq C_3 := 1 + \frac{M^r}{8(1-\gamma(r))},$$

if $2^{3n} \geq 8\sigma_{\mathbf{q}}(T_0)^\lambda$ with $\lambda := \frac{3r\ln 2}{-\ln\gamma(r)}$.

Proof: Let us use the notations $u = \frac{3}{4}(1+C_2\mu)$ and $v = (1+C_2\mu)$. According to Proposition 5.5, we have

$$\overline{\sigma_{\mathbf{q}}^r(n)} \leq \mathbb{E}(\sigma_n^r),$$

where $\mathbb{E}$ is the expectation operator and σ_n is the Markov chain with value in $[1,+\infty[$ defined by

- $\sigma_{n+1} = \max\{\sigma_n u, M\}$ with probability $\alpha := \frac{1}{8}$,
- $\sigma_{n+1} = \sigma_n v$ with probability $\beta := \frac{7}{8}$,
- $\sigma_0 := \sigma_{\mathbf{q}}(T_0)$ with probability 1.

Denoting by μ_n the probability distribution of σ_n, we have

$$\begin{aligned}
\mathbb{E}(\sigma_{n+1}^r) &= \int_1^\infty \sigma^r d\mu_{n+1}(\sigma) \\
&= \int_1^\infty (\alpha(\max\{u\sigma, M\})^r + \beta(v\sigma)^r)\, d\mu_n(\sigma) \\
&= \alpha M^r \int_1^{M/u} d\mu_n(\sigma) + \alpha u^r \int_{M/u}^\infty \sigma^r d\mu_n(\sigma) + \beta v^r \int_1^{+\infty} \sigma^r d\mu_n(\sigma) \\
&\leq \alpha M^r + (\alpha u^r + \beta v^r)\mathbb{E}(\sigma_n^r) \\
&\leq \alpha M^r + \gamma(r)\mathbb{E}(\sigma_n^r)
\end{aligned}$$

By iteration, it follows that

$$\mathbb{E}(\sigma_n^r) \leq \mathbb{E}(\sigma_0^r)\gamma(r)^n + \frac{\alpha M^r}{1-\gamma(r)},$$

which gives the result. ◇

Our next goal is to show that the greedy algorithm initialized from T generates a triangulation which is a refinement of $\mathcal{T}_n^u(T)$ and therefore more accurate, yet with a similar amount of triangles. To this end, we apply the greedy algorithm with root T and stopping criterion given by the local error

$$\eta := \min_{T' \in \mathcal{T}_n^u(T)} e_{T'}(f)_p.$$

Therefore T' is splitted if and only if $e_{T'}(f)_p > \eta$. We denote by $\mathcal{T}_N(T)$ the resulting triangulation where N is its cardinality. From the definition of the stopping criterion, it is clear that $\mathcal{T}_N(T)$ is a refinement of $\mathcal{T}_n^u(T)$.

Proposition 5.7 *Assume that $\mu \leq frac12C_2$, where C_2 is the constant in Proposition 5.5, and define $r_0 := \frac{\ln 2}{\ln 4 - \ln 3} > 0$. We then have*

$$N \leq C_4 2^{3n} \overline{\sigma_{\mathbf{q}}^{r_0}(n)},$$

where C_4 is an absolute constant. Assuming in addition that $\mu \leq \mu(r_0)$ as in Proposition 5.6, we obtain that

$$N \leq C_5 2^{3n},$$

if $2^{3n} \geq 8\sigma_{\mathbf{q}}(T)^{\lambda}$ with $\lambda := \frac{3r_0 \ln 2}{-\ln \gamma(r_0)}$, and where $C_5 = C_3 C_4$.

Proof: Let T_1 be a triangle in $\mathcal{T}_n^u(T)$ and T_2 a triangle in $\mathcal{T}_N(T)$ such that $T_2 \subset T_1$. We shall give a bound on the number of splits k which were applied between T_1 and T_2, i.e. such that $|T_2| = 2^{-k}|T_1|$. We first remark that according to Proposition 5.2, we have

$$\eta \geq c \min_{T' \in \mathcal{T}_n^u(T)} |T'|^{1+\frac{1}{p}} \sigma_{\mathbf{q}}(T') \sqrt{\det \mathbf{q}} \geq c|T_1|^{1+\frac{1}{p}} \sqrt{\det \mathbf{q}},$$

where c is an absolute constant. On the other hand, using both Proposition 5.2 and 5.5, we obtain

$$\begin{aligned}
e_{T_2}(f)_{\mathbf{q}} &\leq C|T_2|^{1+\frac{1}{p}} \sigma_{\mathbf{q}}(T_2) \sqrt{\det \mathbf{q}} \\
&= |T_1|^{1+\frac{1}{p}} 2^{-k(1+\frac{1}{p})} \sigma_{\mathbf{q}}(T_2) \sqrt{\det \mathbf{q}} \\
&\leq C|T_1|^{1+\frac{1}{p}} \sigma_{\mathbf{q}}(T_1) \Big(2^{-(1+\frac{1}{p})}(1 + C_2\mu) \Big)^k \sqrt{\det \mathbf{q}}. \\
&\leq \tfrac{C}{c} \sigma_{\mathbf{q}}(T_1) \Big(\tfrac{1+C_2\mu}{2} \Big)^k \eta \\
&\leq \tfrac{C}{c} \sigma_{\mathbf{q}}(T_1) (\tfrac{3}{4})^k \eta,
\end{aligned}$$

where C is an absolute constant. Therefore we see that k is at most the smallest integer such that $\frac{C}{c}\sigma_{\mathbf{q}}(T_1)(\frac{3}{4})^k \leq 1$. It follows that the total number $n(T_1)$ of triangles $T_2 \in \mathcal{T}_N(T)$ which are contained in T_1 is bounded by

$$n(T_1) \leq 2^k \leq 2(\frac{C}{c}\sigma_{\mathbf{q}}(T_1))^{r_0},$$

and therefore

$$N' = \sum_{T_1 \in \mathcal{T}_n^u(T)} n(T_1) \leq 2(\frac{C}{c})^{r_0} \sum_{T_1 \in \mathcal{T}_n^u(T)} \sigma_{\mathbf{q}}(T_1)^{r_0} = C_4 2^{3n} \overline{\sigma_{\mathbf{q}}^{r_0}(n)},$$

with $C_4 = 2(\frac{C}{c})^{r_0}$. The fact that $N \leq C_5 2^{3n}$ when $2^{3n} \geq 8\sigma_{\mathbf{q}}(T)^{\lambda}$ with $\lambda := \frac{3r_0 \ln 2}{-\ln \gamma(r_0)}$ is an immediate consequence of Proposition 5.6. ◇

5.3 Optimal convergence estimates

Our last step towards the proof of Theorem 5.1 consists in deriving local error estimates for the greedy algorithm. For $\eta > 0$, we denote by f_η the approximant to f obtained by the greedy algorithm with

stopping criterion given by the local error η : a triangle T is splitted if and only if $e_T(f)_p > \eta$. The resulting triangulation is denoted by

$$\mathcal{T}_\eta = \mathcal{T}_N, \text{ with } N = N(\eta) = \#(\mathcal{T}_\eta).$$

For this N, we thus have $f_\eta = f_N$. For a given T generated by the refinement procedure such that $\eta \leq e_T(f)_p$, we also define

$$\mathcal{T}_\eta(T) = \{T' \subset T \, ; \, T' \in \mathcal{T}_\eta\}$$

the triangles in $\mathcal{T}_\eta$ which are contained in T and

$$N(T, \eta) = \#(\mathcal{T}_\eta(T)).$$

Our next result provides with estimates of the local error $\|f - f_\eta\|_{L^p(T)}$ and of $N(T, \eta)$ in terms of η, provided that μ is small enough.

Theorem 5.8 *Assume that $\mu \leq c_2 := \min\{\frac{1}{2C_2}, \mu(r_0)\}$, and that $\eta \leq \eta_0$, where*

$$\eta_0 = \eta_0(T) := \Big(\frac{|T|}{\sigma_{\mathbf{q}}(T)^\lambda}\Big)^{\frac{1}{\tau}} \sqrt{\det \mathbf{q}},$$

with $\lambda := \frac{3r_0 \ln 2}{-\ln \gamma(r_0)}$, and $\frac{1}{\tau} = \frac{1}{p} + 1$. Then

$$\|f - f_\eta\|_{L^p(T)} \leq \eta N(T, \eta)^{\frac{1}{p}}, \tag{5.39}$$

and

$$N(T, \eta) \leq C_6 \eta^{-\tau} \|\sqrt{\det d^2 f}\|^\tau_{L^\tau(T)}, \tag{5.40}$$

where C_6 is an absolute constant.

Proof: The first estimate is trivial since

$$\|f - f_\eta\|_{L^p(T)} = \Big(\sum_{T' \in \mathcal{T}_\eta(T)} e_{T'}(f)_p^p\Big)^{\frac{1}{p}} \leq \Big(\sum_{T' \in \mathcal{T}_\eta(T)} \eta^p\Big)^{\frac{1}{p}} = \eta N(T, \eta)^{\frac{1}{p}}.$$

In the case $p = \infty$, we trivially have

$$\|f - f_\eta\|_{L^\infty(T)} \leq \eta.$$

For the second estimate, we define $n_0 = n_0(T)$ the smallest positive integer such that $2^{3n_0(T)} \geq 8\sigma_{\mathbf{q}}(T)^\lambda$ with $\lambda := \frac{3r_0 \ln 2}{-\ln \gamma(r_0)}$. For any fixed $n \geq n_0$, we define

$$\eta_n := \min_{T' \in \mathcal{T}_n^u(T)} e_{T'}(f)_p.$$

We know from Proposition 5.7 that with the choice $\eta = \eta_n$

$$N(T, \eta_n) \leq C_5 2^{3n}. \tag{5.41}$$

On the other hand, we know from Proposition 5.6, that $\overline{\sigma_{\mathbf{q}}^{r_0}(n)} \leq C_3$, from which it follows that

$$\min_{T' \in \mathcal{T}_n^u(T)} \sigma_{\mathbf{q}}(T') \leq C_3^{\frac{1}{r_0}}.$$

According to Proposition 5.2, we also have

$$\eta_n \leq C \min_{T' \in \mathcal{T}_n^u(T)} |T'|^{1+\frac{1}{p}} \sigma_{\mathbf{q}}(T') \sqrt{\det \mathbf{q}} \leq C_3^{\frac{1}{r_0}} C \Big(\frac{|T|}{2^{3n}}\Big)^{\frac{1}{\tau}} \sqrt{\det \mathbf{q}},$$

where C is an absolute constant, which also reads

$$2^{3n} \leq C_3^{\frac{\tau}{r_0}} C^\tau \eta_n^{-\tau} |T| \sqrt{\det \mathbf{q}}^{\,\tau}.$$

Combining this with (5.41), we have obtained the estimate

$$N(T,\eta_n) \leq C_5 C_3^{\frac{\tau}{r_0}} C^\tau \eta_n^{-\tau} |T| \sqrt{\det \mathbf{q}}^\tau,$$

which by Proposition 5.2 is equivalent to (5.40) with $\eta = \eta_n$. In order to obtain (5.40) for all arbitrary values of η, we write that $\eta_{n+1} < \eta \leq \eta_n$ for some $n \geq n_0$, then

$$\begin{aligned} N(T,\eta) &\leq N(T,\eta_{n+1}) \\ &\leq C_5 2^{3(n+1)} \\ &\leq 8 C_5 C_3^{\frac{\tau}{r_0}} C^\tau \eta_n^{-\tau} |T| \sqrt{\det \mathbf{q}}^\tau \\ &\leq 8 C_5 C_3^{\frac{\tau}{r_0}} C^\tau \eta^{-\tau} |T| \sqrt{\det \mathbf{q}}^\tau, \end{aligned}$$

which by Proposition 5.2 is equivalent to (5.40). In the case where $\eta \geq \eta_{n_0}$, we simply write

$$\begin{aligned} N(T,\eta) &\leq N(T,\eta_{n_0}) \\ &\leq C_5 2^{3n_0} \\ &\leq 64 C_5 \sigma_{\mathbf{q}}(T)^\lambda \\ &= 64 C_5 \eta_0^{-\tau} |T| \sqrt{\det \mathbf{q}}^\tau \\ &\leq 64 C_5 \eta^{-\tau} |T| \sqrt{\det \mathbf{q}}^\tau, \end{aligned}$$

and we conclude in the same way. ◇

We remark that combining the estimates (5.39) and (5.40) in the above Theorem yields the optimal local convergence estimate

$$\|f - f_\eta\|_{L^p(T)} \leq C_6^{\frac{1}{\tau}} \|\sqrt{\det d^2 f}\|_{L^\tau(T)} N(T,\eta)^{-1}.$$

In order to obtain the global estimate of Theorem 5.1, we need to be ensured that after sufficiently many steps of the greedy algorithm, the target f can be well approximated by quadratic function $q = q(T)$ on each triangle T, so that our local results will apply on such triangles. This is ensured due to the following key result.

Proposition 5.9 *Let f be a C^2 function such that $\alpha I \leq d^2 f(x)$ for all $x \in \Omega$ and $\alpha > 0$ independent of x. Let $\mathcal{T}_N$ be the triangulation generated by the greedy algorithm applied to f using the L^∞ decision function given by (2.13). Then*

$$\lim_{N \to +\infty} \max_{T \in \mathcal{T}_N} \operatorname{diam}(T) = 0,$$

i.e. the diameter of all triangles tend to 0.

Proof: See appendix.

Remark 5.10 *We conjecture that this result is also true for the L^2 decision function given by (2.12), although we were not able to prove it. This is the only reason why the optimal convergence estimate in Theorem (5.1) is stated for the L^∞ based decision function.*

Proof of Theorem 5.1 Since $f \in C^2$, an immediate consequence of Proposition 5.9 is that for all $\mu > 0$, there exists

$$N_1 := N_1(f,\mu),$$

such that for all $T \in \mathcal{T}_{N_1}$, there exists a quadratic function q_T such that

$$\|d^2 f - d^2 q_T\|_{L^\infty(T)} \leq \varepsilon = \mu\alpha.$$

Therefore our local results apply on all $T \in \mathcal{T}_{N_1}$. Specifically, we choose

$$N_1 := N_1(f,c_2),$$

with c_2 the constant in Theorem 5.8. We then take

$$\eta \leq \eta_0 := \min_{T \in \mathcal{T}_{N_1}} \{e_T(f)_p, \Big(\frac{|T|}{\sigma_{\mathbf{q}_T}(T)^\lambda}\Big)^{\frac{1}{\tau}} \sqrt{\det \mathbf{q}_T}\},$$

We use the notations

$$f_\eta = f_N, \;\; \mathcal{T}_\eta = \mathcal{T}_N \;\; N = N(\eta) = \#(\mathcal{T}_\eta) = \#(\mathcal{T}_N),$$

for the approximants and triangulation obtained by the greedy algorithm with stopping criterion given by the local error η. Note that $\mathcal{T}_\eta$ is a refinement of $\mathcal{T}_{N_1}$, since $\eta \leq \min_{T \in \mathcal{T}_{N_1}} e_T(f)_p$, and therefore $N \geq N_1$. We obviously have

$$\|f - f_N\|_{L^p} \leq \eta N^{\frac{1}{p}}.$$

Using Theorem 5.8, we also have

$$N = \sum_{T \in \mathcal{T}_{N_1}} N(T, \eta) \leq C_6 \eta^{-\tau} \| \sqrt{\det d^2 f} \|^\tau_{L^\tau(\Omega)}$$

and therefore

$$\|f - f_N\| \leq C_6^{\frac{1}{\tau}} \| \sqrt{\det d^2 f} \|_{L^\tau(\Omega)} N^{-1},$$

which is the claimed estimate. Since we have assumed $\eta \leq \eta_0$, this estimate holds for

$$N > N_0,$$

where N_0 is largest value of N such that $e_T(f)_p \geq \eta_0$ for at least one $T \in \mathcal{T}_N$. $\diamond$

Remark 5.11 *In [9] a modification of the algorithm is proposed so that its convergence in the L^p norm is ensured for any function $f \in L^p(\Omega)$ (or $f \in C(\Omega)$ when $p = \infty$). However this modification is not needed in the proof of Theorem 5.1, due to the assumption that f is convex.*

6 Concluding remarks

In this work, we have shown that a simple greedy algorithm based on iterative bisection has the ability to generate adaptive triangulations for which the optimal convergence estimate (5.35) holds when the number of triangle is large enough. The essential reasons for this are that the algorithm *equidistributes* the local error and generates triangles which have an *optimal aspect ratio*.

So far, our analysis is limited to strictly convex functions, yet numerical results seem to indicate that (5.35) holds for more general C^2 functions. As pointed our in §2.2, we cannot expect that this estimate holds for all C^2 functions, since $\det(d^2 f) = 0$ when f is a function of the type $f(x, y) = g(ax+by)$ with g a univariate function, while the interpolation error is generally non-zero. In fact we conjecture the following:

$$\textit{For all } f \in C^2, \textit{ one has } \limsup_{N \to +\infty} N \|f - f_N\|_{L^p} \leq C \| \sqrt{|\det(d^2 f)|} \|_{L^\tau} \textit{ with } \tfrac{1}{\tau} := \tfrac{1}{p} + 1.$$

Our results of §4 show that the algorithm produces triangles with an optimal aspect ratio when applied to a quadratic polynomial q such that $\det(\mathbf{q}) < 0$. The main difficulties remaining to be solved in order to prove the above conjecture are in the subsequent perturbation analysis, as well as in proving an analog result to Proposition 5.9.

7 Appendix : proofs

7.1 Proof of Proposition 4.1

Let q be a quadratic function of mixed type, and T a triangle with edges a, b, c such that $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$. Up to a linear change of variables, we may assume that the quadratic part of q is $\mathbf{q}(x, y) = x^2 - y^2$. Up to a translation, linear rescaling and permutation between the x and y coordinates,

we may assume that the edge a is centered at 0 and such that $\mathbf{q}(a) = 4$. We write $\frac{a}{2} = (u, v)$, and assume that $u > 0$ without loss of generality. Then $1 = \mathbf{q}(\frac{a}{2}) = u^2 - v^2$, and there must be $\theta \in \mathbb{R}$ such that $(u, v) = (\cosh(\theta), \sinh(\theta))$.

We next observe that the linear transformation ϕ of matrix $\begin{pmatrix} \cosh(\theta) & -\sinh(\theta) \\ -\sinh(\theta) & \cosh(\theta) \end{pmatrix}$ leaves $\mathbf{q}$ invariant, in the sense that $\mathbf{q} \circ \phi = \mathbf{q}$, and verifies $\phi(\frac{a}{2}) = (1, 0)$. Up to such a transformation, we may therefore assume that $a = (2, 0)$. Therefore the two vertices of T corresponding to a are $(-1, 0)$ and $(1, 0)$. We denote the third vertex by (s, t). There is no loss of generality, finally, in assuming that $s \geq 0$ and $t > 0$. Note that $|T| = t$ and $\rho_{\mathbf{q}}(T) = \frac{4}{t}$.

We now specialise to the case where $\rho_{\mathbf{q}}(T) \geq 4$, which is equivalent to $t \leq 1$. Recall that the edges of T are such that $|\mathbf{q}(a)| \geq |\mathbf{q}(b)| \geq |\mathbf{q}(c)|$. The following lines show that $\mathbf{q}(1 + s, t) \geq |\mathbf{q}(s - 1, t)|$, which implies that $b = -(1 + s, t)$ and $c = (s - 1, t)$:

$$\begin{aligned} \mathbf{q}(1 + s, t) - \mathbf{q}(s - 1, t) &= 4s \geq 0 \\ \mathbf{q}(1 + s, t) + \mathbf{q}(s - 1, t) &= 2(1 + s^2 - t^2) \geq 0. \end{aligned}$$

In addition we see that $\mathbf{q}(b) \geq 0$ and we thus have proved that $\mathbf{q}(a)\mathbf{q}(b) \geq 0$. For the second and third inequality in (4.28) we remark that $\mathbf{q}(c) = (s - 1)^2 - t^2$ and $\mathbf{q}(d) = s^2 - t^2$. Clearly

$$-1 \leq -t^2 \leq \min\{\mathbf{q}(c), \mathbf{q}(d)\}.$$

If $0 \leq s \leq 1$, we also clearly have

$$\max\{\mathbf{q}(c), \mathbf{q}(d)\} \leq 1.$$

If $s \geq 1$ we have

$$\mathbf{q}(c) \leq \mathbf{q}(d) = s^2 - t^2 = (s + 1)^2 - t^2 - (2s + 1) = \mathbf{q}(b) - (2s + 1) \leq \mathbf{q}(a) - 3 = 1,$$

We thus always have

$$\max\{|\mathbf{q}(c)|, |\mathbf{q}(d)|\} \leq \frac{|\mathbf{q}(a)|}{4} = 1, \tag{7.42}$$

which concludes the proof of the inequalities (4.28).

Last, we specialize to the case where $\rho_{\mathbf{q}}(T) = \frac{4}{t} \geq 8$, equivalently $t \leq \frac{1}{2}$, to prove (4.29):

$$\mathbf{q}(b) + \mathbf{q}(c) = (s + 1)^2 - t^2 + (s - 1)^2 - t^2 = 2 + 2s^2 - 2t^2 \geq \frac{3}{2} = \frac{3}{8}\mathbf{q}(a).$$

7.2 Proof of Proposition 5.9

For any triangle T we denote by T_x the interval defined as the projection of T on the x axis, and by $|T_x|$ its length. We denote by I_T and I_{T_x} the two dimensional and one dimensional local interpolations operators on T and T_x respectively. It is clear that if g is a C^2 convex function of one variable and if $G(x, y) := g(x)$, then

$$\|G - I_T G\|_{L^\infty(T)} = \|g - I_{T_x} g\|_{L^\infty(T_x)}. \tag{7.43}$$

The following lemma compares the interpolation error on an interval of $\mathbb{R}$ and on a sub-interval.

Lemma 7.1 *Let $g \in C^2(\mathbb{R}, \mathbb{R})$ be such that $0 < m \leq g'' \leq M$. Let x_1, x_2, x_3 be real numbers satisfying $x_1 < \frac{x_1 + x_3}{2} \leq x_2 < x_3$, and denote $u := x_2 - x_1 \leq v := x_3 - x_1$. Then denoting by I_u and I_v the interpolation operators on the intervals $[x_1, x_2]$ and $[x_1, x_3]$ respectively, we have*

$$\|g - I_v g\|_{L^\infty([x_1, x_3])} \geq mv^2/8$$

and

$$\|g - I_u g\|_{L^\infty([x_1, x_2])} \leq (1 - \alpha\frac{v - u}{v})\|g - I_v g\|_{L^\infty([x_1, x_3])}$$

with $\alpha := \frac{1}{4}\sqrt{m/M}$.

Proof: Let us define $h_v = I_v g - g$. Since $h_v'' = -g''$ and $h_v(x_1) = h_v(x_3) = 0$, h_v can be represented as the integral

$$h_v(x) = \int_{x_1}^{x_3} K_v(x,y) g''(y) dy. \tag{7.44}$$

where the Green kernel K_v is given by

$$K_v(x,y) = \frac{1}{v} \begin{cases} (x - x_1)(x_3 - y) & \text{if } x \le y \\ (x_3 - x)(y - x_1) & \text{if } x \ge y \end{cases}$$

Of course, we have a similar representation of $h_u = I_u g - g$ with a kernel K_u.
The first part of the proposition immediately follows from

$$h_v(\frac{x_1 + x_3}{2}) \ge m \int_{x_1}^{x_3} K_v(\frac{x_1 + x_3}{2}, y) dy \ge mv^2/8.$$

In order to prove the second part, we shall compare the Green Kernels K_u and K_v. For this purpose, we define

$$\mu(x) = \frac{(x_2 - x)/u}{(x_3 - x)/v}.$$

For all $x, y \in [x_1, x_2]$, we thus have

$$\frac{K_u(x,y)}{K_v(x,y)} = \min\{\mu(x), \mu(y)\} \le \mu(x).$$

Therefore, defining $x_u := \text{argmax}_{t \in [x_1, x_2]} (I_u g - g)(t)$ and using (7.44), we obtain

$$\begin{array}{ll} \|g - I_u g\|_{L^\infty([x_1,x_2])} & = h_u(x_u) \\ & = \int_{x_1}^{x_2} K_u(x_u, y) g''(y) dy \\ & \le \mu(x_u) \int_{x_1}^{x_2} K_v(x_u, y) g''(y) dy \\ & \le \mu(x_u) \int_{x_1}^{x_3} K_v(x_u, y) g''(y) dy \\ & \le \mu(x_u) \|g - I_v g\|_{L^\infty([x_1,x_3])}. \end{array}$$

In order to conclude, we need to estimate $\mu(x_u)$. One easily checks by differentiation that μ is decreasing and concave on the interval $[x_1, x_2]$, and therefore for all $x \in [x_1, x_2]$

$$\mu(x) \le 1 - (x - x_1) \frac{v - u}{v^2}.$$

Differentiating the integral representation of h_u, we obtain

$$u h_u'(x_u) = -\int_{x_1}^{x_u} (y - x_1) g''(y) dy + \int_{x_u}^{x_2} (x_2 - y) g''(y) dy.$$

Since $h'(x_u) = 0$ and $0 < m \le g'' \le M$, we obtain

$$(x_2 - x_u)^2 m \le (x_u - x_1)^2 M.$$

Since $x_2 \ge \frac{x_1 + x_3}{2}$, this gives $x_u - x_1 \ge \sqrt{m/M} \frac{v}{4}$. Therefore,

$$\mu(x_u) \le 1 - \frac{1}{4} \sqrt{m/M} \frac{v - u}{v}.$$

$\diamond$

The following corollary uses the above lemma to compare the values of the L^∞-based decision functions for a convex function that depends only of one variable. For any vector $v \in \mathbb{R}^2$ we denote by v_x the absolute value of its x coordinate.

Corollary 7.2 *Let T be a triangle with edges a, b, c, such that $a_x \geq b_x \geq c_x$. Let $G(x, y) = g(x)$, where $g \in C^2$ and $0 < m \leq g'' \leq M$. Then, with d_T defined by (3.14),*

$$\begin{aligned} d_T(b, G) - d_T(a, G) &\geq C a_x (a_x - b_x), \\ d_T(c, G) - d_T(a, G) &\geq C a_x^2/2. \end{aligned}$$

with $C = \frac{m^{3/2}}{32\sqrt{M}}$.

Proof: We recall the notation $\alpha_T(f) := \|f - I_T f\|_{L^\infty}$. It follows from (7.43) that

$$\alpha_T(G) = \|g - I_{T_x} g\|_{L^\infty(T_x)}.$$

We label the extremities of a by $i \in \{1, 2\}$, and denote by $T^{i,s}$, $i \in \{1, 2\}$, $s \in \{a, b, c\}$ the child of T resulting from the bisection through the edge s in such a way that $T^{i,s}$ contains the extremity of a of label i. Then (up to exchanging the labels of the extremities of a),

$$\begin{array}{llll} |T_x^{1,a}| = b_x & & |T_x^{2,a}| = a_x/2, \\ |T_x^{1,b}| = a_x & & |T_x^{2,b}| = c_x + b_x/2, \\ |T_x^{1,c}| = b_x + c_x/2 & & |T_x^{2,c}| = a_x. \end{array}$$

In particular, we have $T_x^{2,a} \subset T_x^{2,b}$ and $T_x^{1,a} \subset T_x^{1,b}$ and therefore

$$\alpha_{T^{2,a}}(G) = \|g - I_{T_x^{2,a}} g\|_{L^\infty(T_x^{2,a})} \leq \|g - I_{T_x^{2,a}} g\|_{L^\infty(T_x^{2,b})} = \alpha_{T^{2,b}}(G),$$

and similarly $\alpha_{T^{1,a}}(G) \leq \alpha_{T^{1,c}}(G)$. Moreover, we can apply the previous lemma with $[x_1, x_2] = T_x^{1,a} \subset T_x^{1,b} = [x_1, x_3]$ or with $[x_1, x_2] = T_x^{2,a} \subset T_x^{2,c} = [x_1, x_3]$ which respectively leads to

$$\alpha_{T^{1,b}}(G) - \alpha_{T^{1,a}}(G) \geq \frac{m^{3/2}}{32\sqrt{M}} a_x (a_x - b_x),$$

and

$$\alpha_{T^{2,c}}(G) - \alpha_{T^{2,a}}(G) \geq \frac{m^{3/2}}{32\sqrt{M}} a_x^2/2.$$

This allows us to conclude since $d_T(s, G) = \alpha_{T_s^1} + \alpha_{T_s^2}$, for $s \in \{a, b, c\}$. ◇

Using the above result, we now prove that the decision function d_T tends to prescribe longest edge bisection with respect to the euclidean metric when the triangle T becomes too thin.

Corollary 7.3 *Let f be a convex function such that $m\,\mathrm{Id} \leq d^2 f \leq M\,\mathrm{Id}$, and let T be a triangle with measure of non-degeneracy $\sigma(T)$ for the euclidean metric and edges a, b, c, such that $|a| \geq |b| \geq |c|$. Then, with $K = 128(\frac{M}{m})^{3/2}$, if $\sigma(T) > 2K$, the bisection prescribed by the decision function $d_T(\cdot, f)$ is a δ-near longest edge bisection with respect to the euclidean metric (in the sense of definition 5.3), with $\delta := \frac{K}{\sigma(T)}$.*

Proof: We denote by $p(X)$ the affine orthogonal projection of a point $X \in \mathbb{R}^2$ onto the line which includes the edge a, and we denote by O the midpoint of a. We then define

$$\tilde{f}(X) = f(p(X)) + df_O(X - p(X)).$$

Then, using the notation $X(z) = p(X) + z(X - p(X))$, we have

$$\begin{aligned} f(X) - \tilde{f}(X) &= f(X) - f(p(X)) - df_O(X - p(X)), \\ &= \int_0^1 (df_{X(z)} - df_O)(X - p(X)) dz. \end{aligned}$$

But for $X \in T$, we have $X(z) \in T$ for all $z \in [0, 1]$, so that $|X(z) - O| \leq |a|$ and therefore

$$\|df_{X(z)} - df_O\| \leq M|a|.$$

We also have

$$|X - p(X)| \leq \frac{2|T|}{|a|} \leq \frac{|a|}{\sigma(T)}.$$

Therefore, $\|f - \tilde{f}\|_{L^\infty(T)} \leq M\frac{|a|^2}{\sigma(T)}$. We can apply the previous lemma to the function $\tilde{f}$, since it is the sum of a function of one variable and of an affine function which has no effect on the interpolation error. Assuming without loss of generality (up to a rotation) that a is parallel to the x axis, this gives us

$$\begin{aligned} d_T(b, \tilde{f}) - d_T(a, \tilde{f}) &\geq C|a|(|a| - b_x) \geq C|a|(|a| - |b|), \\ d_T(c, \tilde{f}) - d_T(a, \tilde{f}) &\geq C|a|^2/2. \end{aligned}$$

with $C = \frac{m^{3/2}}{32\sqrt{M}}$. Since the Lebesgue constant of the interpolation operator on any triangle is 2, we have

$$|d_T(x, f) - d_T(x, \tilde{f})| \leq 4||f - \tilde{f}||_{L^\infty(T)}$$

which implies the following inequalities

$$\begin{aligned} d_T(b, f) - d_T(a, f) &\geq C|a|(|a| - |b|) - 4M|a|^2/\sigma(T), \\ d_T(c, f) - d_T(a, f) &\geq |a|^2(C/2 - 4M/\sigma(T)). \end{aligned}$$

The second inequality shows that the edge c cannot be cut if $C/2 - 4M/\sigma(T) > 0$ which is equivalent to $\sigma(T) > 2K$. The first inequality shows that b may be cut provided that $C(|a| - |b|) - 4M|a|/\sigma(T) \leq 0$, i.e. $|b| \geq (1 - \frac{4M}{C\sigma(T)})|a|$ which shows the property of δ-near longest edge bisection with $\delta := \frac{K}{\sigma(T)}$. $\diamond$

The proof of Proposition 5.9 directly follows from this last result. Let us denote by $\mathrm{diam}(T)$ the diameter of any triangle T. If the size of the triangles generated by the greedy algorithm did not tend to zero, then there would be a sequence $(T_i)_{i\geq 0}$ of triangles such that T_{i+1} is one of the children of T_i, and $\mathrm{diam}(T_i) \to d > 0$ as $i \to \infty$. Since $|T_i| \to 0$, this also implies that $\sigma(T_i) \to +\infty$ as $i \to \infty$. We can therefore choose i large enough such that $\mathrm{diam}(T_i)^2 < \frac{4}{3}d^2$ and $1 + C_2\frac{K}{\sigma(T_j)} \leq \frac{3}{2}$ for all $j \geq i$, where C_2 is the constant in Proposition 5.5. According to this proposition, we have

$$\sigma(T_{i+3}) \leq \frac{3}{2}\sigma(T_i).$$

On the other hand, we obviously have for any triangle T,

$$\frac{\mathrm{diam}(T)^2}{8|T|} \leq \sigma(T) \leq \frac{\mathrm{diam}(T)^2}{2|T|},$$

from which it follows that

$$\mathrm{diam}(T_{i+3})^2 \leq 4\frac{|T_{i+3}|\sigma(T_{i+3})}{|T_i|\sigma(T_i)}\,\mathrm{diam}(T_i)^2 \leq \frac{3}{4}\,\mathrm{diam}(T_i)^2$$

Therefore, $\mathrm{diam}(T_{i+3})^2 < d^2$ which is a contradiction. This concludes the proof of Proposition 5.9.

References

[1] T. Apel, *Anisotropic finite elements: Local estimates and applications*, Series "Advances in Numerical Mathematics", Teubner, Stuttgart, 1999.

[2] V. Babenko, Y. Babenko, A. Ligun and A. Shumeiko, *On Asymptotical Behavior of the Optimal Linear Spline Interpolation Error of C^2 Functions*, East J. Approx. 12(1), 71–101, 2006.

[3] P. Binev, W. Dahmen and R. DeVore, *Adaptive Finite Element Methods with Convergence Rates*, Numerische Mathematik 97, 219–268, 2004.

[4] P. Binev, W. Dahmen, R. DeVore and P. Petrushev, *Approximation Classes for Adaptive Methods*, Serdica Math. J. 28, 391–416, 2002.

[5] H. Borouchaki, P.J. Frey, P.L. George, P. Laug and E. Saltel, *Mesh generation and mesh adaptivity: theory, techniques*, in Encyclopedia of computational mechanics, E. Stein, R. de Borst and T.J.R. Hughes ed., John Wiley & Sons Ltd., 2004.

[6] E. Candes and D. L. Donoho, *Curvelets and curvilinear integrals*, J. Approx. Theory. 113, 59–90, 2000.

[7] L. Chen, P. Sun and J. Xu, *Optimal anisotropic meshes for minimizing interpolation error in L^p-norm*, Math. of Comp. 76, 179–204, 2007.

[8] A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, *Tree-structured approximation and optimal encoding*, App. Comp. Harm. Anal. 11, 192–226, 2001.

[9] A. Cohen, N. Dyn, F. Hecht and J.-M. Mirebeau, *Adaptive multiresolution analysis based on anisotropic triangulations*, preprint, Laboratoire J.-L.Lions, submitted 2008.

[10] R. DeVore, *Nonlinear approximation*, Acta Numerica 51-150, 1998

[11] J.-M. Mirebeau, *Optimal bidimensional finite element meshes*, preprint, Laboratoire J.-L. Lions.

[12] P. Morin, R. Nochetto and K. Siebert, *Convergence of adaptive finite element methods*, SIAM Review 44, 631–658, 2002.

[13] M.C. Rivara, *New longest-edge algorithms for the renement and/or improvement of unstructured triangulations*, Int. J. Num. Methods 40, 3313–3324, 1997.