What Is a Good Linear Finite Element?
Interpolation, Conditioning, Anisotropy, and Quality Measures
(Preprint)

Jonathan Richard Shewchuk
jrs@cs.berkeley.edu

December 31, 2002

Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, CA 94720

Abstract

When a mesh of simplicial elements (triangles or tetrahedra) is used to form a piecewise linear approximation of a function, the accuracy of the approximation depends on the sizes and shapes of the elements. In finite element methods, the conditioning of the stiffness matrices also depends on the sizes and shapes of the elements. This article explains the mathematical connections between mesh geometry, interpolation errors, discretization errors, and stiffness matrix conditioning. These relationships are expressed by error bounds and element quality measures that determine the fitness of a triangle or tetrahedron for interpolation or for achieving low condition numbers. Unfortunately, the quality measures for these purposes do not fully agree with each other; for instance, small angles are bad for matrix conditioning but not for interpolation or discretization. The upper and lower bounds on interpolation error and element stiffness matrix conditioning given here are tighter than those usually seen in the literature, so the quality measures are likely to be unusually precise indicators of element fitness. Bounds are included for anisotropic cases, wherein long, thin elements perform better than equilateral ones. Surprisingly, there are circumstances wherein interpolation, conditioning, and discretization error are each best served by elements of different aspect ratios or orientations.

Supported in part by the National Science Foundation under Awards ACI-9875170, CMS-9980063, CCR-0204377, and EIA-9802069, and in part by a gift from the Okawa Foundation. The views and conclusions in this document are those of the author. They are not endorsed by, and do not necessarily reflect the position or policies of, the Okawa Foundation or the U. S. Government.

Keywords: finite element mesh, interpolation error, condition number, discretization error, anisotropy, quality measure

Contents

Comments Needed!

This preprint is surely blemished by a few errors and misconceptions. If you read part of it, and the final version hasn't yet appeared on my Web page, please send your comments to `jrs@cs.berkeley.edu`! I am especially interested in hearing about citations I've missed, insights I'm missing, and whether I'm emphasizing the right things. I expect this article will be under revision until about a year after the date below.

Thanks in advance for your help.

Jonathan Shewchuk, 31 December 2002.

1 Introduction

Interpolation, contouring, and finite element methods rely on the availability of meshes whose elements have the right shapes and sizes. The accuracy or speed of some applications can be compromised by just a few bad elements. Algorithms for mesh generation and mesh improvement are expected to produce elements whose "quality" is as good as possible. However, forty-odd years after the invention of the finite element method, our understanding of the relationship between mesh geometry, numerical accuracy, and stiffness matrix conditioning remains incomplete, even in the simplest cases. Engineering experience and asymptotic mathematical results have taught us that equilateral elements are usually good, and skinny or skewed elements are usually bad. However, there has been insufficient mathematical guidance for, say, choosing the better of two elements of intermediate quality, or choosing the aspect ratios of anisotropic elements.

This article examines triangular and tetrahedral meshes used for piecewise linear interpolation (including finite element methods with piecewise linear basis functions). The quality of a mesh depends on the application that uses it. Interpolation accuracy is important for most tasks. For finite element methods, discretization errors and the condition number of the global stiffness matrix are important too. The coming pages study how the shapes and sizes of linear elements affect these things. Error bounds and quality measures are provided here that estimate the influence of element geometry on accuracy and conditioning, and can guide numerical analysts and mesh generation algorithms in creating and evaluating meshes.

Interpolation on a triangular or tetrahedral mesh constructs a function that attempts to approximate some "true" function, whose exact identity might or might not be known. For example, a surveyor may know the altitude of the land at each point in a large sample, and use interpolation over a triangulation to approximate the altitude at points where readings were not taken. If a triangulation's sole purpose is as a basis for interpolation, the primary criterion of its fitness is how much the interpolated function differs from the true function. There are two types of *interpolation error* that matter for most applications: the difference between the interpolated function and the true function, and the difference between the gradient of the interpolated function and the gradient of the true function. Errors in the gradient can be surprisingly important, whether the application is rendering, mapmaking, or physical simulation, because they can compromise accuracy or create unwanted visual artifacts. In finite element methods they contribute to discretization errors.

If the true function is smooth, the error in the interpolated function can be reduced simply by making the triangles or tetrahedra smaller. However, the error in the gradient is strongly affected by the shape of the elements as well as their size, and this error is often the primary arbiter of element quality. The enemy is large angles: the error in the gradient can grow arbitrarily large as angles approach 180°. Bounds on the errors associated with piecewise linear interpolation are discussed and derived in Section 2.

If your application is the finite element method, then the condition number of the stiffness matrix associated with the method should be kept as small as possible. Poorly conditioned matrices affect linear equation solvers by slowing them down or introducing large roundoff errors into their results. Element shape has a strong influence on matrix conditioning, but unlike with interpolation errors, small angles are deleterious and large ones (alone) are not. The relationship between element shape and matrix conditioning depends on the partial differential equation being solved and the basis functions and test functions used to discretize it. Bounds on condition number must be derived on a case-by-case basis. The stiffness matrices associated with Poisson's equation on linear elements are studied in Section 3.

The *discretization error* is the difference between the approximation computed by the finite element method and the true solution. Like stiffness matrix condition numbers, discretization error depends in part on the partial differential equation and the method of discretization. However, discretization error is closely related to the interpolation errors, and is mitigated by elements whose shapes and sizes are selected to control the interpolation errors. The relationship is discussed in Section 4.

In some circumstances, the ideal element is *anisotropic*—elongated and oriented in an appropriate direction. The preferred aspect ratio and orientation of an element is determined by the nature of the interpolated function (for minimizing the interpolation and discretization errors) and the partial differential equation (for minimizing the stiffness matrix condition number and discretization error). Anisotropic interpolation is discussed in Sections 2.2 and 2.3, and the effects of anisotropic partial differential equations on stiffness matrix conditioning and discretization error are discussed in Sections 3.2 and 4.2. The preferred orientations for interpolation accuracy, matrix conditioning, and discretization accuracy often agree with each other, but not always, as Section 4.3 reveals.

Anisotropic elements can exhibit a surprising phenomenon I call *superaccuracy*: elements even longer and thinner than expected sometimes offer the best accuracy for the interpolated gradients or the discretization error. See Section 2.3.

Quality measures for evaluating and comparing elements are discussed in Section 6. These include measures of an element's fitness for interpolation and stiffness matrix formation (Section 6.1), and many other measures found in the literature (Section 6.3). The quality measures introduced in this article can be used in either an *a priori* or *a posteriori* fashion, and are designed to interact well with numerical optimization methods for mesh smoothing.

Error bounds and numerical quality measures are used in many ways: to expose the fundamental goals of mesh generation algorithms; to guide point placement in advancing front methods; to select elements in need of refinement or improvement by mesh smoothing or topological transformations; as a means to select the best configuration of elements when a topological transformation is considered; and as an objective function for smoothers based on numerical optimization. Of these uses, the last imposes the harshest requirements on a quality measure. Section 6.2 discusses how to use quality measures to help accomplish these goals.

Table 1 is a reference chart for the notation used in this article. Some of the quantities are illustrated in Figure 1. Triangle vertices and edges are numbered from 1 to 3, with vertex v_i opposite edge i. Tetrahedron vertices and faces are numbered from 1 to 4, with vertex v_i opposite face i.

Let t be a triangular or tetrahedral element. The value r_{circ} is the radius of the circumcircle or circumsphere of t, r_{in} is the radius of the incircle or insphere of t, and r_{mc} is the radius of the min-containment circle or sphere of t. The *circumcircle*, or circumscribed circle, of a triangle is the unique circle that passes through all three of its vertices, and the *circumsphere* of a tetrahedron passes through all four of its vertices. The *incircle*, or inscribed circle, of a triangle is the smallest circle that touches all three of its sides, and the *insphere* of a tetrahedron is the smallest sphere that touches all four of its triangular faces. The *min-containment circle* of a triangle is the smallest circle that encloses the triangle; its center is either the circumcenter of the triangle or a midpoint of one of its edges. The *min-containment sphere* of a tetrahedron is the smallest sphere that encloses it; its center is either the circumcenter of the tetrahedron, the circumcenter of one of its triangular faces, or a midpoint of one of its edges.

Numerically stable formulae for A, V, A_i, r_{circ}, r_{in}, and r_{mc} appear in Appendix A. For numerical computation, readers are urged to prefer these over most formulae they may have found elsewhere. Many formulae in the literature are either unnecessarily inaccurate, or unnecessarily slow.

Some of the quantities are *signed*, which means that they are negative for inverted elements. To say an element is *inverted* is to presuppose that it has a fixed topological orientation, defined by an ordering of its vertices. For instance, a triangle is inverted if its vertices are supposed to occur in counterclockwise order, but upon inspection occur in clockwise order. The topology of a mesh determines the orientation of each element relative to the orientations of all the others.

Table 1: Notation used in this article. Signed quantities are negative for inverted elements. Edge lengths are always nonnegative.

A	The signed area of a triangle.
V	The signed volume of a tetrahedron, or the signed measure of a d-dimensional simplicial element.
A_1, A_2, A_3, A_4	The unsigned areas of the faces of a tetrahedron, or the measures of the $(d-1)$-dimensional faces of a d-dimensional simplicial element.
A_{rms}	The root-mean-square face area of a tetrahedron, $\sqrt{\frac{1}{4}\sum_{i=1}^{4} A_i^2}$.
$A_{\max}, A_{\min}$	The unsigned areas of the largest and smallest faces of a tetrahedron.
$\ell_1, \ell_2, \ldots, \ell_e$	The edge lengths of an element, where e is 3 for a triangle or 6 for a tetrahedron.
ℓ_{ij}	The length of the edge connecting vertices v_i and v_j.
ℓ_{rms}	The root-mean-square edge length of an element, $\sqrt{\frac{1}{e}\sum_{i=1}^{e} \ell_i^2}$.
$\ell_{\min}, \ell_{\mathrm{med}}, \ell_{\max}$	The minimum, median, and maximum edge lengths of an element. (ℓ_{med} is defined for triangles only.)
$a_{\min}, a_{\mathrm{med}}, a_{\max}$	The minimum-, median-, and maximum-magnitude signed altitudes of an element. For a triangle, $a_{\min} = 2A/\ell_{\max}$, $a_{\mathrm{med}} = 2A/\ell_{\mathrm{med}}$, and $a_{\max} = 2A/\ell_{\min}$. For a tetrahedron, $a_{\min} = 3V/A_{\max}$ and $a_{\max} = 3V/A_{\min}$.
$h_{\min}$	The minimum signed *aspect* of an element: the smallest distance such that the element fits between two parallel lines or planes separated by a distance of $h_{\min}$. For a triangle, $h_{\min} = a_{\min}$.
r_{circ}	The signed *circumradius* of an element (the radius of its circumscribing circle or sphere). See Appendix A.2.
r_{in}	The signed *inradius* of an element (the radius of its inscribed circle or sphere). For a triangle, $r_{\mathrm{in}} = 2A/(\ell_1 + \ell_2 + \ell_3)$. For a tetrahedron, $r_{\mathrm{in}} = 3V/\sum_{i=1}^{4} A_i$.
r_{mc}	The unsigned radius of the *min-containment circle or sphere* of an element (the smallest circle or sphere that encloses the element). See Appendix A.3.
θ_i	The angle at vertex v_i of a triangle.
θ_{ij}	In a tetrahedron, the dihedral angle at the edge connecting vertices v_i and v_j.
$\theta_{\min}, \theta_{\max}$	For a triangle, the signed minimum and maximum angles. For a tetrahedron, the signed minimum and maximum dihedral angles.
$\min_{i,j} \sin\theta_{ij}$	The minimum sine of all angles of a triangle or dihedral angles of a tetrahedron. For a tetrahedron, $\min_{i,j} \sin\theta_{ij} = \min\{\sin\theta_{\min}, \sin\theta_{\max}\}$. For a triangle, $\min_i \sin\theta_i = \sin\theta_{\min}$ (because $\theta_{\max} \leq 180° - 2\theta_{\min}$).
ϕ_1, ϕ_2, ϕ_3, ϕ_4	The solid angles of a tetrahedron.
Z	An unsigned quantity associated with r_{circ}. See Appendix A.2.

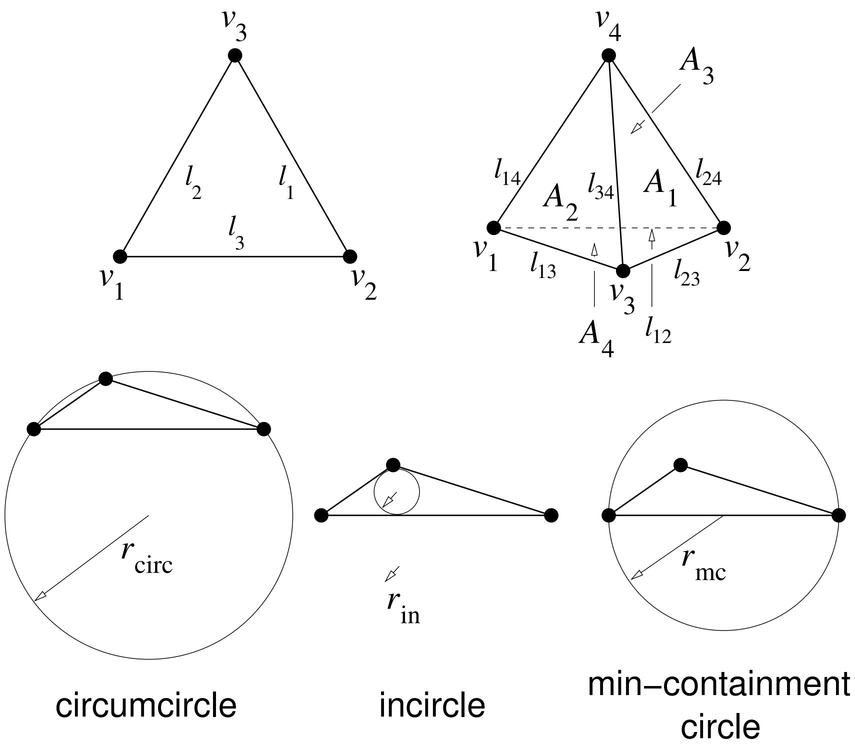


Figure 1: Quantities associated with triangles and tetrahedra.

2 Element Size, Element Shape, and Interpolation Error

The celebrated paper of Babuška and Aziz [4] demonstrates that the accuracy of finite element solutions on triangular meshes degrades seriously as angles are allowed to approach 180°, but the same is not true as angles are allowed to approach 0°, so long as the largest angles are not too large. In other words, small angles are not deleterious to the interpolation accuracy or the discretization error (although they may be deleterious to the stiffness matrix). Previously, researchers had believed that small angles must be prohibited. (Synge [48] proved a result similar to Babuška and Aziz's two decades earlier, but it was not well known until much later.)

The Babuška–Aziz paper is more often cited than understood, as it is cast in the language of functional analysis. Its results are asymptotic and offer little guidance in, say, how to compare two differently shaped elements of intermediate quality. This section presents error bounds (and Section 6.1 presents related quality measures) that can be used to accurately judge the size and shape of a linear element. These bounds are stronger than the classical bounds of approximation theory—asymptotically stronger in some cases. The bounds for triangles are tight to within a small constant factor. Section 2.1 describes the results and what they mean. The derivations, which appear in Section 2.4, require no knowledge of functional analysis.

2.1 Error Bounds for Interpolation

Let T be a triangular or tetrahedral mesh, and let $f(p)$ be a continuous scalar function defined over the mesh. Let $g(p)$ be a piecewise linear approximation to $f(p)$, where $g(v) = f(v)$ at each vertex v of T, and $g(p)$ is linear over any single element of T. Table 2 gives bounds on two types of interpolation error associated

Table 2: Bounds on interpolation error for a single element t. The function g is a linear approximation of f over t. All bounds assume that the magnitude of the directional second derivative of f does not exceed c_t anywhere in the element t. The "weaker but simpler upper bounds" are not asymptotically weaker; they are weaker than the stronger upper bounds by a factor of no more than three. Each lower bound implies that there exists some function f for which the error is at least as large as the lower bound.

	$\|f - g\|_\infty$	$\|\nabla f - \nabla g\|_\infty$
Upper bound, triangles	$c_t \dfrac{r_{\mathrm{mc}}^2}{2}$	$c_t \dfrac{\ell_{\max}\ell_{\mathrm{med}}\left(\ell_{\min} + 4r_{\mathrm{in}}\right)}{4A}$
Weaker but simpler upper bound, triangles	$c_t \dfrac{\ell_{\max}^2}{6}$	$c_t \dfrac{3\ell_{\max}\ell_{\mathrm{med}}\ell_{\min}}{4A}$
Lower bound, triangles	$c_t \dfrac{r_{\mathrm{mc}}^2}{2}$	$c_t \max\left\{r_{\mathrm{circ}}, a_{\max}, \sqrt{\ell_{\max}^2 - a_{\mathrm{med}}^2}\right\}$
Note: for triangles, $c_t \dfrac{\ell_{\max}\ell_{\mathrm{med}}\ell_{\min}}{4A} = c_t \dfrac{\ell_{\max}}{2\sin\theta_{\max}} = c_t r_{\mathrm{circ}}$		
Upper bound, tetrahedra	$c_t \dfrac{r_{\mathrm{mc}}^2}{2}$	$c_t \dfrac{\frac{1}{6V}\sum_{1\leq i<j\leq 4} A_i A_j \ell_{ij}^2 + \max_i \sum_{j\neq i} A_j \ell_{ij}}{\sum_{m=1}^4 A_m}$
Weaker but simpler upper bound, tetrahedra	$c_t \dfrac{3\ell_{\max}^2}{16}$	$c_t \dfrac{\sum_{1\leq i<j\leq 4} A_i A_j \ell_{ij}^2}{2V\sum_{m=1}^4 A_m}$
Lower bound, tetrahedra	$c_t \dfrac{r_{\mathrm{mc}}^2}{2}$	$c_t r_{\mathrm{circ}}$; see Section 2.4 for others
Note: for tetrahedra, $c_t \dfrac{\sum_{1\leq i<j\leq 4} A_i A_j \ell_{ij}^2}{6V\sum_{m=1}^4 A_m} = c_t \dfrac{\sum_{1\leq i<j\leq 4} \ell_{ij}^2 \ell_{kl}/\sin\theta_{kl}}{4\sum_{m=1}^4 A_m}$, where i, j, k, and l are distinct in each term of the summation		
Upper bound, higher-dimensional simplices	$c_t \dfrac{r_{\mathrm{mc}}^2}{2}$	$c_t \dfrac{\frac{1}{2dV}\sum_{1\leq i<j\leq d+1} A_i A_j \ell_{ij}^2 + \max_i \sum_{j\neq i} A_j \ell_{ij}}{\sum_{m=1}^{d+1} A_m}$

with g. The norm $\|f - g\|_\infty$ is defined to be the maximum pointwise interpolation error over the element t, $\max_{p\in t} |f(p) - g(p)|$. The norm $\|\nabla f - \nabla g\|_\infty$ is the maximum magnitude of the pointwise error in the interpolated gradient, $\max_{p\in t} |\nabla f(p) - \nabla g(p)|$. The bounds in Table 2 are derived in Section 2.4, but first they require some explanation.

If $f(p)$ is arbitrary, $g(p)$ can be an arbitrarily bad approximation of $f(p)$. The error can be bounded only if $f(p)$ is constrained in some way. A reasonable constraint, which yields the error bounds in Table 2, is to assume that $f(p)$ is smooth and the absolute curvature of $f(p)$ is bounded in each element t by some constant c_t (which may differ for each t). The *curvature* $f''_{\mathbf{d}}(p)$ of the function f at the point p along an arbitrary direction vector $\mathbf{d}$ is its directional second derivative along $\mathbf{d}$. Specifically, let the point p have coordinates (x, y, z), and consider the Hessian matrix

$$H(p) = \begin{bmatrix} \frac{\partial^2}{\partial x^2} f(p) & \frac{\partial^2}{\partial x \partial y} f(p) & \frac{\partial^2}{\partial x \partial z} f(p) \\ \frac{\partial^2}{\partial x \partial y} f(p) & \frac{\partial^2}{\partial y^2} f(p) & \frac{\partial^2}{\partial y \partial z} f(p) \\ \frac{\partial^2}{\partial x \partial z} f(p) & \frac{\partial^2}{\partial y \partial z} f(p) & \frac{\partial^2}{\partial z^2} f(p) \end{bmatrix}.$$

(For the two-dimensional case, delete the last row and column of $H(p)$.) To support matrix notation, each point or vector p is treated as a $d \times 1$ vector whose transpose p^T is a $1 \times d$ vector. The notation $\mathbf{d}^T H(p)\mathbf{d}$ denotes the scalar result of the matrix multiplication

$$\mathbf{d}^T H(p)\mathbf{d} = \begin{bmatrix} d_x & d_y & d_z \end{bmatrix} \begin{bmatrix} \frac{\partial^2}{\partial x^2} f(p) & \frac{\partial^2}{\partial x \partial y} f(p) & \frac{\partial^2}{\partial x \partial z} f(p) \\ \frac{\partial^2}{\partial x \partial y} f(p) & \frac{\partial^2}{\partial y^2} f(p) & \frac{\partial^2}{\partial y \partial z} f(p) \\ \frac{\partial^2}{\partial x \partial z} f(p) & \frac{\partial^2}{\partial y \partial z} f(p) & \frac{\partial^2}{\partial z^2} f(p) \end{bmatrix} \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix}.$$

For any unit direction vector $\mathbf{d}$, the directional curvature is $f''_{\mathbf{d}}(p) = \mathbf{d}^T H(p)\mathbf{d}$. If $\mathbf{d}$ is not a unit vector, it is easy to show that $f''_{\mathbf{d}}(p) = \mathbf{d}^T H(p)\mathbf{d}/|\mathbf{d}|^2$. Assume that f is known to satisfy the following curvature constraint:[1] for any direction $\mathbf{d}$,

$$|f''_{\mathbf{d}}(p)| \leq c_t. \tag{1}$$

How does one obtain bounds on curvature to use in generating, evaluating, or improving a mesh? The per-element curvature bounds c_t sometimes come from *a priori* error estimators, based on knowledge of the function to be interpolated. Sometimes they are provided by *a posteriori* error estimators, which are estimated from a finite element solution over another mesh of the same domain. (*A posteriori* estimates often indicate directional curvature, in which case the anisotropic error bounds in Section 2.2 are stronger.) If bounds on curvature are not available, it might not be possible to bound the interpolation error, but the formulae in Table 2 may still be used to compare elements, by dropping c_t from each formula. This is equivalent to assuming that there is some unknown bound on curvature that holds everywhere.

Let's examine the bounds. (Contour plots of the upper bounds appear in Section 6.1.) The upper bound on $\|f - g\|_\infty$, the maximum interpolation error over t, is $c_t r_{\mathrm{mc}}^2/2$. This bound is tight: for any triangle or tetrahedron t with min-containment radius r_{mc}, there is a function f such that $\|f - g\|_\infty = c_t r_{\mathrm{mc}}^2/2$. This bound (and its tightness) was first derived by Waldron [49], and it applies to higher-dimensional simplicial elements as well.

Interestingly, Rajan [39] shows that for any set of vertices in any dimensionality, the Delaunay triangulation of those vertices minimizes the maximum min-containment radius (as compared with all other triangulations of the vertices). In other words, every triangulation of the same vertex set has a triangle whose min-containment radius is at least as large as the largest min-containment radius in the Delaunay triangulation. Therefore, if c_t is constant over the triangulation domain, the Delaunay triangulation minimizes the maximum bound on interpolation error over the domain. (Read carefully—minimizing the maximum *bound* on interpolation error is not the same as minimizing the maximum interpolation error. The latter would be possible only with more knowledge of the function f.)

It is interesting to compare this bound to the bounds usually given for interpolation, which implicate the maximum edge length of each element. To obtain a specified level of accuracy, a mesh is refined until no edge is larger than a specified length. However, the min-containment radius of an element gives a tighter bound on $\|f - g\|_\infty$ than the maximum edge length.

[1]For those familiar with matrix norms, note that $\|H\|_2 = \max_{|\mathbf{d}|=1} |\mathbf{d}^T H(p)\mathbf{d}|$, so the constraint can be written $\|H\|_2 \leq c_t$. An equivalent statement is that the eigenvalues of H are all in $[-c_t, c_t]$.

Unfortunately, the min-containment radius r_{mc} is expensive to compute (see Appendix A.3). The maximum edge length $\ell_{\max}$ is a much faster alternative. For a triangle, $r_{\mathrm{mc}} \leq \ell_{\max}/\sqrt{3}$, and for a tetrahedron, $r_{\mathrm{mc}} \leq \sqrt{3/8}\ell_{\max}$. Substitution yields the faster-to-compute but slightly looser bounds $\|f - g\|_\infty \leq c_t\ell_{\max}^2/6$ (for triangles), $\|f - g\|_\infty \leq 3c_t\ell_{\max}^2/16$ (for tetrahedra).

The error $f - g$ is not the only concern. In many applications, g is expected to accurately represent the gradients of f, and the error $\nabla f - \nabla g$ is just as important as, or more important than, $f - g$. For example, when the finite element method is used to find a piecewise linear approximation h to the true solution f of a second-order partial differential equation, the discretization error $f - h$ normally can be bounded only if both $f - g$ and $\nabla f - \nabla g$ can be bounded. (Note that although g and h are both piecewise linear functions, they differ because h does not usually equal f at the mesh vertices.) See Section 4.1 for further discussion. Simulations of mechanical deformation provide a second example, where the accuracy of ∇g is particularly important because ∇f (the strains) is of more interest than f (the displacements). Visualization of height fields provides a third example, as we shall see shortly.

Newly derived bounds on $\|\nabla f - \nabla g\|_\infty$ appear in Table 2. The bound for triangles is similar to a bound in an unpublished manuscript by Handscomb [26].[2] The bounds for tetrahedra and higher-dimensional simplicial elements appear to be without precedent in the literature.

The bounds reveal that $\|\nabla f - \nabla g\|_\infty$ can grow arbitrarily large as elements become arbitrarily badly shaped, unlike $\|f - g\|_\infty$. Observe that the area or volume appears in the denominator of these bounds. Imagine distorting a triangle or tetrahedron so that its area or volume approaches zero. Then ∇g may or may not approach infinity, depending on whether the numerator of the error bound also approaches zero.

First imagine an isosceles triangle with one angle near $180°$ and two tiny angles. As the large angle approaches $180°$, A approaches zero and the edge lengths do not change much, so the upper and lower error bounds grow arbitrarily large. Now imagine an isosceles triangle with one tiny angle and two angles near $90°$. As the tiny angle approaches zero, A approaches zero, but $\ell_{\min}$ and r_{in} approach zero at the same rate, so the error bounds change little. Hence, angles near $180°$ are harmful, whereas angles near zero are, by themselves, benign. The same can be said of the dihedral angles of tetrahedra.

Figure 2 visually illustrates these effects. Three triangulations, each having 200 triangles, are used to render a paraboloid. The mesh of long thin triangles at right has no angle greater than $90°$, and visually performs only slightly worse than the isotropic triangulation at left. (The slightly worse performance is because of the longer edge lengths.) However, the middle paraboloid looks like a washboard, because the triangles with large angles have very inaccurate gradients. Because of this effect, accurate gradients are critical for applications like contouring and rendering.

Figure 3 shows why this problem occurs. The triangle illustrated has values associated with its vertices that represent heights (or, say, an approximation of some physical quantity). The values of g at the endpoints of the bottom edge are 35 and 65, so the linearly interpolated value of g at the center of the edge is 50. This value is independent of the value associated with the top vertex. As the angle at the upper vertex approaches $180°$, the interpolated point (with value 50) becomes arbitrarily close to the upper vertex (with value 40). Hence, ∇g may become arbitrarily large (in its vertical component), and is clearly specious as an approximation of ∇f, even though $g = f$ at the vertices.

The same effect is seen between two edges of a sliver tetrahedron that pass near each other, also illustrated in Figure 3. A *sliver* is a tetrahedron that is nearly flat even though none of its edges is much shorter than the others. A typical sliver is formed by uniformly spacing its four vertices around the equator of a

[2]The bound on $\|\nabla f - \nabla g\|_\infty$ for triangles given here is close to Handscomb's, being slightly better for nearly equilateral triangles and slightly worse for others.

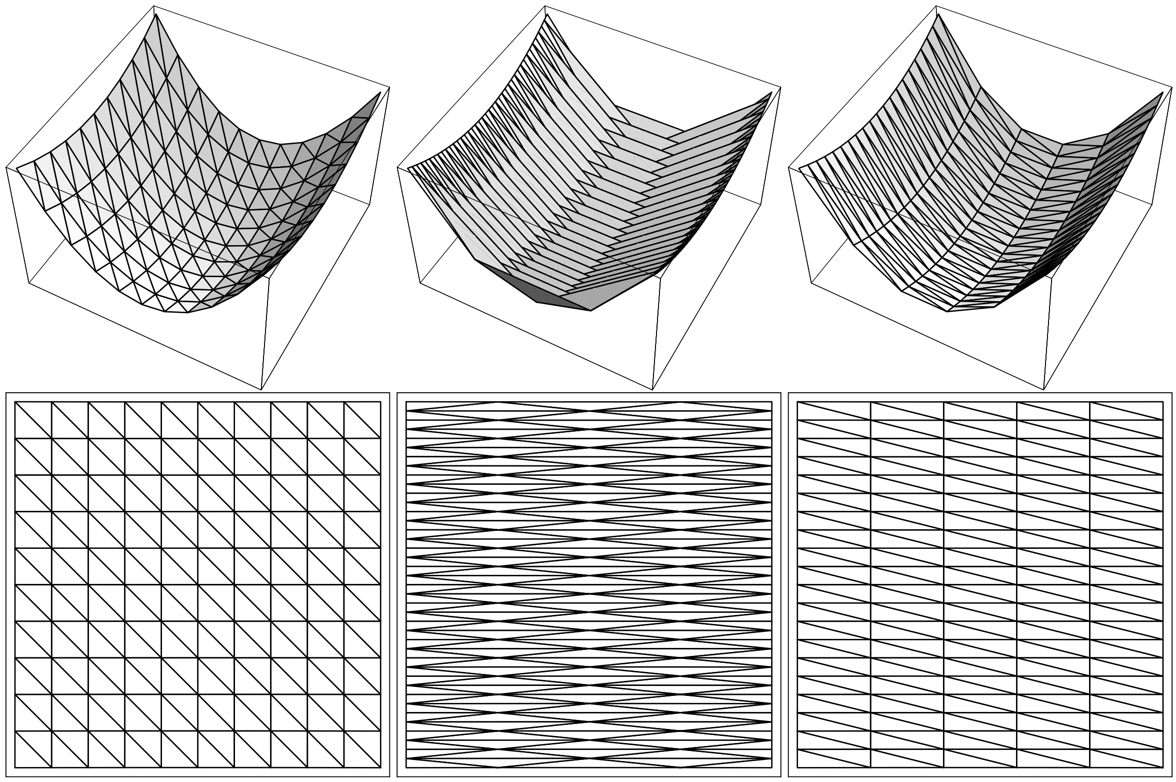


Figure 2: A visual illustration of how large angles, but not small angles, can cause the error $\nabla f - \nabla g$ to explode. In each triangulation, 200 triangles are used to render a paraboloid.

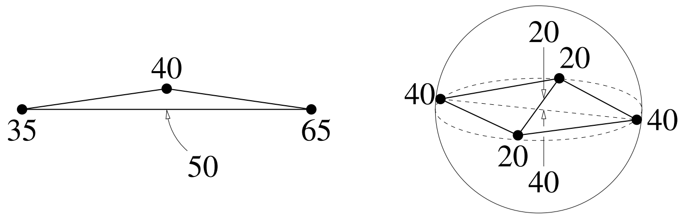


Figure 3: As the large angle of the triangle approaches $180°$, or the sliver tetrahedron becomes arbitrarily flat, the magnitude of the vertical component of ∇g becomes arbitrarily large.

sphere, then perturbing one of the vertices just off the equator so that the sliver has some (but not much) volume.

Because of this sensitivity, mesh generators usually choose the shapes of elements to control $\|\nabla f - \nabla g\|_\infty$, and not $\|f - g\|_\infty$, which can be reduced simply by using smaller elements. Section 6.1 presents quality measures that judge the shape of elements based on their fitness for interpolation.

Table 2 gives two upper bounds on $\|\nabla f - \nabla g\|_\infty$ over a triangle. The first upper bound is almost tight, to within a factor of two. The "weaker but simpler upper bound" of $3c_t r_{\mathrm{circ}}$ is not as good an indicator as the stronger upper bound, but it has the advantages of being smooth almost everywhere (and therefore more amenable to numerical optimization) and faster to compute. The weaker upper bound is tight to within a factor of three: for any triangle t, there is a function f such that $\|\nabla f - \nabla g\|_\infty = c_t r_{\mathrm{circ}}$.

These bounds are interesting because the two-dimensional Delaunay triangulation minimizes the maximum circumradius [9], just as it minimizes the maximum min-containment radius. (This property does not hold in three or more dimensions, unlike Rajan's min-containment radius result.) Hence, the two-dimensional Delaunay triangulation is good (but not optimal) for controlling the worst-case value of $\|\nabla f - \nabla g\|_\infty$.

The upper bound of $3c_t r_{\mathrm{circ}}$ and the lower bound of $c_t r_{\mathrm{circ}}$ can be expressed in three different forms (see the first note in Table 2), one of which implicates the largest angle of the triangle. The upper bound $3c_t\ell_{\max}/(2\sin\theta_{\max})$ can be loosely decomposed into a size contribution $(3/2)c_t\ell_{\max}$ and a shape contribution $1/\sin\theta_{\max}$. This seems to suggest that a triangular mesh generator should seek to minimize the maximum angle, and algorithms for that purpose are available [19, 8]. However, this inference demonstrates the folly of too eagerly separating element size from element shape: the triangulation that optimizes $\ell_{\max}/\sin\theta_{\max}$ (namely, the Delaunay triangulation) and the triangulation that optimizes $1/\sin\theta_{\max}$ are not the same, and the former is more germane to the interpolation and discretization errors. Section 6.1 discusses slightly better measures of shape quality than $1/\sin\theta_{\max}$.

Bramble and Zlámal [13] derived a well-known upper bound proportional to $c_t\ell_{\max}/\sin\theta_{\min}$. Replacing $\theta_{\min}$ with $\theta_{\max}$, as done here, obviously leads to different conclusions.

The upper bounds for tetrahedra are more difficult to interpret than the bounds for triangles. Consider the alternative form (suggested in a note in Table 2) of the weak upper bound: $\|\nabla f - \nabla g\|_\infty \leq 3c_t(\sum_{i<j}\ell_{ij}^2\ell_{kl}/\sin\theta_{kl})/(4\sum_{m=1}^{4}A_m)$. This bound suggests that the error may approach infinity as the sine of a dihedral angle θ_{kl} approaches zero. However, the error does not approach infinity if the length ℓ_{ij} of the opposite edge approaches zero at the same rate as the angle. If an angle θ_{kl} is small but the opposite edge length ℓ_{ij} is not, the tetrahedron must have a large dihedral angle as well. Small dihedral or planar angles are not problematic for interpolation unless a large angle is present too.

Figure 4 provides some insight into which tetrahedron shapes are good or bad for accurate interpolation. The "good" tetrahedra are of two types: those that are not flat, and those that can grow arbitrarily flat without having a large planar or dihedral angle. The "bad" tetrahedra have error bounds that explode, and a dihedral angle or a planar angle that approaches 180°, as they are flattened. (Unfortunately, most of the "good" tetrahedra in the figure are bad for stiffness matrix conditioning because of the small angles.)

The upper bounds for tetrahedra are not known to be asymptotically tight, but I conjecture that they are. Unfortunately, it is difficult to develop a strong lower bound that covers all tetrahedron shapes. However, the no-large-angle condition is necessary. For any tetrahedron with a dihedral angle or planar angle approaching 180°, there is a function f for which $\|\nabla f - \nabla g\|_\infty$ approaches infinity. See Section 2.4 for lower bounds that verify this claim.

2.2 Anisotropy and Interpolation

The bounds given in Section 2.1 are appropriate when the upper bound on the second directional derivative is the same (c_t) in every direction. However, some applications interpolate functions in which the largest curvature in one direction far exceeds the largest curvature in another. It is possible to obtain good accuracy with nearly equilateral elements, but often it is possible to obtain the same accuracy with many fewer elements by generating an *anisotropic* mesh. In an anisotropic mesh, long thin elements are oriented along a direction in which the curvature is small. For many applications, anisotropic elements offer the best tradeoff between accuracy and speed. In fluid flow modeling, aspect ratios of 100,000 to one are sometimes appropriate. There are applications for which it is not computationally feasible to generate or use an isotropic mesh fine enough to obtain an accurate solution, but an anisotropic mesh can offer sufficient accuracy and have few enough elements to be usable.

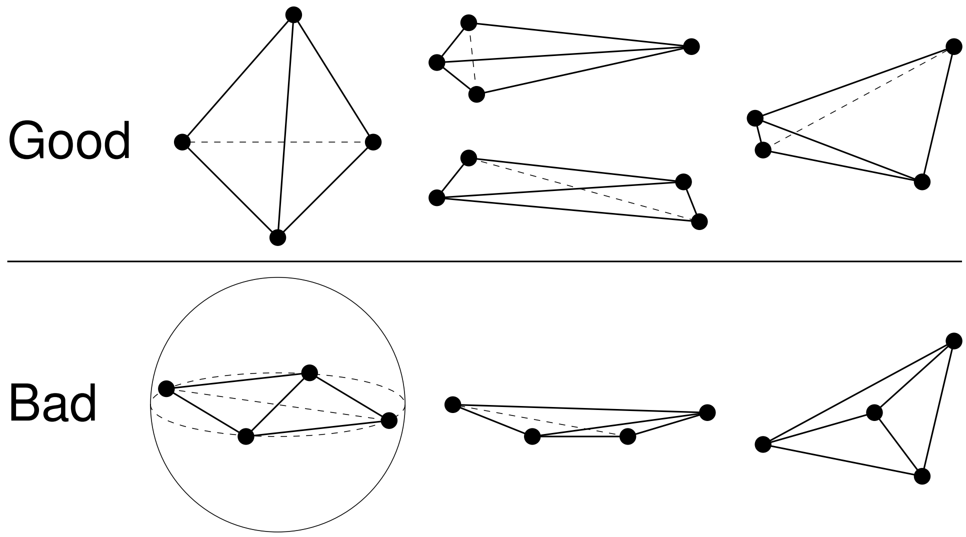


Figure 4: The top four tetrahedron shapes incur little interpolation error. The bottom three tetrahedron shapes can cause $\|\nabla f - \nabla g\|_\infty$ to be unnecessarily large, and to grow without bound if the tetrahedra are flattened.

The Babuška–Aziz result [4] applies to the anisotropic case too, but it is an asymptotic result: as the largest angle approaches $180°$, the discretization error grows without bound. This result is often interpreted to mean "large angles are bad." But this is only true if "large" is defined appropriately. There are circumstances where a triangle with a $179.9°$ angle performs excellently, and an equilateral triangle dismally [29]. This section and Section 3.2 describe some of them.

Babuška and Aziz's vindication of small angles has sometimes been taken to show that it is safe to use needle-shaped triangles to approximate anisotropic functions. However, the writers who read the result this way are confusing two issues. The truth is, both very small and very large angles can perform well in anisotropic circumstances, if they are oriented properly; and this good performance can encompass both interpolation error and stiffness matrix conditioning, even though conditioning is badly degraded by small angles in isotropic circumstances.

The basic method for judging an element, following D'Azevedo [17], is to affinely map it to an "isotropic space" in which the curvature bounds are isotropic. A skewed element in physical space that becomes equilateral when mapped to isotropic space is ideal for minimizing the interpolation error $\|f - g\|$. However, the story is more complicated for $\|\nabla f - \nabla g\|$, and not every element can be evaluated by mapping it and judging its image by the standards of the isotropic case. The error bounds yield some surprising conclusions about the shape of the ideal element, discussed in Section 2.3.

To judge or select anisotropic elements, one must characterize the directional bounds on the curvature of the function $f(p)$ being interpolated. To do this, replace the curvature bound c_t with a symmetric positive-definite matrix C_t, here called the *curvature tensor* for the element t.

The curvature tensor is best explained by constructing one. Let $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ be three orthogonal unit vectors in three-dimensional space. (For two-dimensional meshes, omit $\mathbf{v}_3$.) These vectors represent three directions along which there are three different positive upper bounds ξ_1, ξ_2, and ξ_3 on the second directional derivative of $f(p)$. Define the 3×3 matrices V and Ξ as below.

The curvature tensor is the matrix product

$$C_t = V \Xi V^T$$

$$= \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} \begin{bmatrix} \xi_1 & 0 & 0 \\ 0 & \xi_2 & 0 \\ 0 & 0 & \xi_3 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}^T \tag{2}$$
$$= \xi_1 \mathbf{v}_1 \mathbf{v}_1^T + \xi_2 \mathbf{v}_2 \mathbf{v}_2^T + \xi_3 \mathbf{v}_3 \mathbf{v}_3^T.$$

(Here, $\mathbf{v}\mathbf{v}^T$ is a 3×3 *outer product*, not to be confused with the more common inner product.)

Let $H(p)$ be the Hessian matrix of $f(p)$. The forthcoming error bounds are based on the assumption that we somehow know that $f(p)$ satisfies the following constraint. For any point $p \in t$ and any vector $\mathbf{d}$,

$$|\mathbf{d}^T H(p) \mathbf{d}| \leq \mathbf{d}^T C_t \mathbf{d}. \tag{3}$$

This inequality implies that the curvature of $f(p)$ along direction $\mathbf{v}_i$ does not exceed ξ_i.

The values $\mathbf{v}_i$ and ξ_i are the eigenvectors and eigenvalues of C_t. To see this, observe that because the $\mathbf{v}_i$ vectors are of unit length and orthogonal to each other,

$$C_t \mathbf{v}_i = \sum_{j=1}^{3} \xi_j \mathbf{v}_j (\mathbf{v}_j \cdot \mathbf{v}_i) = \xi_i \mathbf{v}_i.$$

The curvature tensor C_t can be constructed from a known set of eigenvalues and eigenvectors. Alternatively, C_t can be directly specified (for instance, based on an estimate of the Hessian of f), in which case its eigenvalues and eigenvectors can be recovered by standard algorithms [24]. Let $\xi_{\max}$ and $\xi_{\min}$ be the largest and smallest eigenvalues of C_t, and let $\mathbf{v}_{\max}$ and $\mathbf{v}_{\min}$ be the associated unit eigenvectors. For any unit vector $\mathbf{d}$, $\mathbf{d}^T C_t \mathbf{d} \leq \xi_{\max}$, so the maximum curvature of f is $\xi_{\max}$. For consistency with the isotropic bounds, I shall continue to use c_t for the maximum curvature as well, so $\xi_{\max} = c_t$.

A helpful tool for studying anisotropy is a transformation matrix E that maps a point in *physical space*—the domain over which $f(p)$ is defined—to *isotropic space*. E is the square root of $(1/\xi_{\max})C_t$. Like C_t, E can be expressed as a sum of outer products. Let

$$E = V \begin{bmatrix} \sqrt{\xi_1/\xi_{\max}} & 0 & 0 \\ 0 & \sqrt{\xi_2/\xi_{\max}} & 0 \\ 0 & 0 & \sqrt{\xi_3/\xi_{\max}} \end{bmatrix} V^T.$$

Because the $\mathbf{v}_i$ vectors are orthonormal, $V^T V = I$ and

$$E^2 = \frac{1}{\xi_{\max}} C_t = \frac{1}{c_t} C_t.$$

For any point p in physical space, let $\widehat{p} = Ep$ denote its image in isotropic space. Let $\widehat{t}$ be the image of t in isotropic space. The vertices of $\widehat{t}$ are $\widehat{v}_1$, $\widehat{v}_2$, $\widehat{v}_3$, and for a tetrahedron, $\widehat{v}_4$. Let $\kappa = \xi_{\max}/\xi_{\min}$ be the *condition number* of C_t. The transformation of t to $\widehat{t}$ can shrink an edge of t by a factor of up to $\sqrt{\kappa}$ (for an edge aligned with $\mathbf{v}_{\min}$) or as little as 1 (for an edge aligned with $\mathbf{v}_{\max}$). Figure 5 demonstrates the mapping in two dimensions for the case $\xi_1 = 1$, $\xi_2 = 4$. This transformation is intuitively useful because the interpolation errors on t are at least as small as the errors would be on $\widehat{t}$ under the isotropic bound (1), so t may be roughly judged by the size and shape of $\widehat{t}$.

Let $\widehat{f}(q) = f(E^{-1}q)$ be a function defined over the isotropic space. Observe that $\widehat{f}(\widehat{p}) \equiv f(p)$, so $\widehat{f}$ takes on the same range of values over $\widehat{t}$ as f does over t. Similarly, let $\widehat{g}(q) = g(E^{-1}q)$.

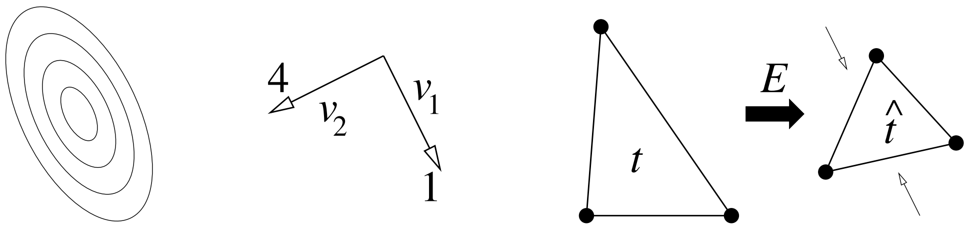


Figure 5: At left, the level sets of $p^T C_t p$, where p varies over the plane. Center, each eigenvector of C_t is drawn with its associated eigenvalue. At right, the effect of the transformation E on a triangle.

The curvature of $\widehat{f}$ has a bound that is independent of direction (hence the name "isotropic space"). To see this, observe that the second directional derivative of $\widehat{f}$ along any unit vector $\mathbf{d}$ is

$$\begin{aligned}
\widehat{f}''_{\mathbf{d}}(q) &= \left.\frac{d^2}{d\alpha^2} f(E^{-1}(q+\alpha\mathbf{d}))\right|_{\alpha=0} \\
&= (E^{-1}\mathbf{d})^T H(q)(E^{-1}\mathbf{d}) \\
&\leq \mathbf{d}^T E^{-1} C_t E^{-1}\mathbf{d} \quad \text{from (3)} \\
&= c_t|\mathbf{d}|^2 \\
&= c_t.
\end{aligned}$$

Therefore, in isotropic space, the curvature bound on $\widehat{f}(q)$ (as a function of q) is c_t in every direction, and the isotropic bounds from Table 2 apply to $\widehat{f}(q)$.

Table 3 gives upper bounds (analogous to those in Table 2) on two types of interpolation error associated with a piecewise linear approximation g of f. The table uses some new notation, defined as follows. For any edge length ℓ, let $\widehat{\ell}$ denote the length of the same edge in isotropic space. For example, $\widehat{\ell}_1$ denotes the distance between the mapped vertices $\widehat{v}_2 = Ev_2$ and $\widehat{v}_3 = Ev_3$ of a triangle $\widehat{t}$. Let $\widehat{r}_{\mathrm{mc}}$ denote the radius of the min-containment circle or sphere of $\widehat{t}$.

The error bounds for $\|f-g\|_\infty$ given in Table 3 are derived by mapping the element t to isotropic space. The isotropic bounds from Table 2 apply to $\widehat{t}$ in isotropic space, so the maximum error in $\|\widehat{f}-\widehat{g}\|_\infty$ over $\widehat{t}$ is $c_t\widehat{r}_{\mathrm{mc}}^2/2$. This error bound transfers directly to physical space. Because $f(p) \equiv \widehat{f}(\widehat{p})$ and $g(p) \equiv \widehat{g}(\widehat{p})$, $\max_{p\in t}|f(p)-g(p)| = \max_{q\in\widehat{t}}|\widehat{f}(q)-\widehat{g}(q)| \leq c_t\widehat{r}_{\mathrm{mc}}^2/2$. Both the upper and lower bounds generalize this way, so this bound, like the isotropic bound, is tight and extends to any dimension.

The min-containment circle or sphere of $\widehat{t}$ can be smaller than that of t, but not larger. Hence, the error bound for the anisotropic case is at least as good as the isotropic bound, but might be much better if t is a thin element whose longer dimension is roughly aligned with $\mathbf{v}_{\min}$. For triangles, the min-containment radius is scaled by a factor between $1/\sqrt{\kappa}$ (if t is degenerate and parallel to $\mathbf{v}_{\min}$) and 1 (if t is degenerate and parallel to $\mathbf{v}_{\max}$), as Figure 6 illustrates, and so the error bound is scaled by a factor between $1/\kappa$ and 1. Therefore, a triangle that has an aspect ratio of roughly $\sqrt{\kappa}$ and is oriented along the direction of least curvature can offer a better error bound for a fixed amount of area than an isotropic triangle. In three dimensions, the best tradeoff between accuracy and number of elements is found in a tetrahedron whose size is proportional to $1/\sqrt{\xi_i}$ along each axis $\mathbf{v}_i$. Ideally, $\widehat{t}$ is equilateral.

Stated more simply, the error $\|f-g\|_\infty$ can be estimated by shrinking t along the appropriate axes (i.e. affinely mapping t to isotropic space), and applying the isotropic error bound to the shrunken element. Unfortunately, the same is not true for $\|\nabla f - \nabla g\|_\infty$. The reason is because although $f(p) \equiv \widehat{f}(\widehat{p})$ holds, $\nabla f(p) \not\equiv \nabla\widehat{f}(\widehat{p})$. When f is mapped to isotropic space, its gradients change. Imagine f as a three-dimensional surface; when it is shrunk along an axis, its directional derivatives along that axis increase.

Table 3: Bounds on interpolation error for a single element t and a function f with anisotropic bounds on curvature. The function g is a piecewise linear approximation of f over t. All bounds assume that the directional second derivative of f in any direction $\mathbf{d}$ is bounded by the inequality $|\mathbf{d}^T H(p)\mathbf{d}| \leq \mathbf{d}^T C_t \mathbf{d}$, where $H(p)$ is the Hessian matrix of f at point p, C_t is a known positive definite matrix fixed for the element t, and c_t is the largest eigenvalue of C_t. Quantities with a caret are properties of the element $\widehat{t}$, found by applying the linear transformation E to t.

	$\|f - g\|_\infty$	$\|\nabla f - \nabla g\|_\infty$
Upper bound, triangles	$c_t \dfrac{\widehat{r}_{\mathrm{mc}}^2}{2}$	$c_t \dfrac{\frac{1}{4A}\left(\ell_1\ell_2\widehat{\ell_3}^2 + \ell_3\ell_1\widehat{\ell_2}^2 + \ell_2\ell_3\widehat{\ell_1}^2\right) + \max_{1\leq i<j\leq 3}\left(\ell_i\widehat{\ell_j} + \ell_j\widehat{\ell_i}\right)}{\ell_1 + \ell_2 + \ell_3}$
Weaker upper bound, triangles	$c_t \dfrac{\widehat{\ell}_{\max}^2}{6}$	$c_t \left(\dfrac{\ell_1\ell_2\widehat{\ell_3}^2 + \ell_3\ell_1\widehat{\ell_2}^2 + \ell_2\ell_3\widehat{\ell_1}^2}{4A(\ell_1+\ell_2+\ell_3)} + \dfrac{\widehat{\ell_1}+\widehat{\ell_2}+\widehat{\ell_3}}{2}\right)$
Note: $c_t \dfrac{\ell_1\ell_2\widehat{\ell_3}^2 + \ell_3\ell_1\widehat{\ell_2}^2 + \ell_2\ell_3\widehat{\ell_1}^2}{4A(\ell_1+\ell_2+\ell_3)} = c_t \dfrac{\ell_1\ell_2\ell_3}{4A} \cdot \dfrac{\widehat{\ell_1}^2/\ell_1 + \widehat{\ell_2}^2/\ell_2 + \widehat{\ell_3}^2/\ell_3}{\ell_1+\ell_2+\ell_3}$		
Upper bound, tetrahedra	$c_t \dfrac{\widehat{r}_{\mathrm{mc}}^2}{2}$	$c_t \dfrac{\frac{1}{6V}\sum_{1\leq i<j\leq 4} A_i A_j \widehat{\ell_{ij}}^2 + \max_i \sum_{j\neq i} A_j \widehat{\ell_{ij}}}{\sum_{m=1}^4 A_m}$
Weaker upper bound, tetrahedra	$c_t \dfrac{3\widehat{\ell}_{\max}^2}{16}$	$c_t \left(\dfrac{\sum_{1\leq i<j\leq 4} A_i A_j \widehat{\ell_{ij}}^2}{6V\sum_{m=1}^4 A_m} + \dfrac{1}{3}\sum_{i=1}^{6}\widehat{\ell_i}\right)$

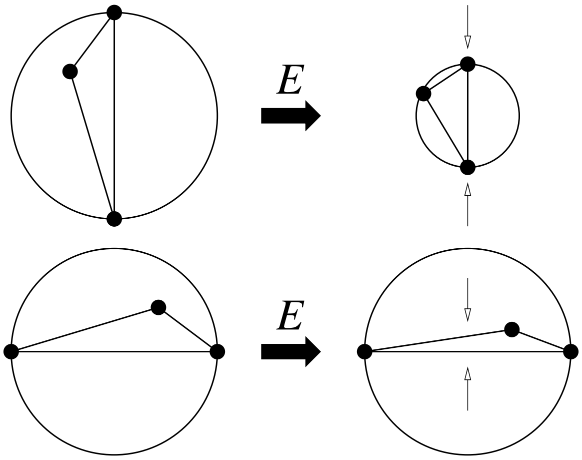


Figure 6: The transformation E scales the min-containment radius of an element by a factor ranging from a minimum of $1/\sqrt{\kappa}$ to a maximum of 1.

The bounds in Table 3 for $\|\nabla f - \nabla g\|_\infty$ are found by a separate derivation; see Section 2.4. These bounds are tricky to interpret. Roughly speaking, if the mapped element $\widehat{t}$ is close to equilateral, t is a pretty good element (but not necessarily the best), even if t has a large angle. For a triangle, this means t is long and thin, is oriented along the direction of least curvature $\mathbf{v}_{\min}$, and has an aspect ratio of $\sqrt{\kappa}$. However, Section 2.3 describes an intriguing phenomenon wherein even longer, thinner elements may offer yet better accuracy of the gradients (for a fixed number of elements).

Interestingly, the first term of the bound for triangles is the product of the first term of the isotropic bound, $c_t \ell_1 \ell_2 \ell_3/(4A)$—which punishes large angles—and an "anisotropy correctional term" $(\widehat{\ell_1}^2/\ell_1 + \widehat{\ell_2}^2/\ell_2 + \widehat{\ell_3}^2/\ell_3)/(\ell_1 + \ell_2 + \ell_3)$. The correctional term ranges from $1/\kappa$ (if all the edges of t are parallel to $\mathbf{v}_{\min}$) to 1 (if all the edges are parallel to $\mathbf{v}_{\max}$). A triangle t for which $\widehat{t}$ is equilateral does not incur much more error than an equilateral triangle (even if t has a large angle), but t occupies $\sqrt{\kappa}$ times more area, so fewer triangles are needed.

The remainder of the error bound, $c_t \max_{i,j}(\ell_i \widehat{\ell_j} + \ell_j \widehat{\ell_i})/(\ell_1 + \ell_2 + \ell_3)$, is roughly proportional to the perimeter of $\widehat{t}$ (as the weaker upper bound reflects). Therefore, among triangles t of fixed area, this term is roughly minimized if $\widehat{t}$ is equilateral.

Similarly, a tetrahedron t whose size is proportional to $1/\sqrt{\xi_i}$ along each axis $\mathbf{v}_i$, so that $\widehat{t}$ is equilateral, offers a better tradeoff between volume and error than an equilateral tetrahedron, for both $\|f - g\|_\infty$ and $\|\nabla f - \nabla g\|_\infty$.

For either a triangle or a tetrahedron t, any term of the anisotropic upper bound for $\|\nabla f - \nabla g\|_\infty$ in Table 3 may dominate asymptotically, depending on the shape of t. (This is why the anisotropic bounds cannot be simplified as nicely as the isotropic bounds.) The first term dominates if t has a large angle in isotropic space, and approaches infinity if an angle of t approaches $180°$. The remainder of the error bound dominates the first term if t is correctly oriented with an aspect ratio greater than $\sqrt{\kappa}$, but has no large angle.

Table 3 includes weaker bounds for $\|\nabla f - \nabla g\|_\infty$ that have the advantage of varying smoothly with the vertex positions, making them better suited to numerical optimization (and slightly quicker to compute). The weaker bounds for $\|f - g\|_\infty$ are not smooth, but are much quicker to compute.

2.3 Superaccuracy of Gradients over Anisotropic Elements

Surprisingly, if the circumstances are right, a triangle t whose aspect ratio is κ can offer the best bound on $\|\nabla f - \nabla g\|_\infty$ for a fixed amount of area. (Tetrahedra also exhibit this phenomenon.) However, there are four caveats.

- The Hessian H of f must vary little over t—or at least, the directions of the eigenvectors of H (the principle directions of curvature of f) must vary little. (The smaller t is, the more likely this will be true.)
- t must be very nearly parallel to the eigenvector of H associated with its largest eigenvalue.
- t must have no large angle in physical space. (It is no longer enough for $\widehat{t}$ to have no large angle in isotropic space.)
- Because t has an aspect ratio of κ, it exhibits a larger error $\|f - g\|_\infty$, and often worse conditioning, than a parallel triangle with the same area and aspect ratio $\sqrt{\kappa}$, so the result is mainly useful in cases where minimizing the error $\|\nabla f - \nabla g\|_\infty$ is more important than minimizing the error $\|f - g\|_\infty$ or controlling the element's contribution to global conditioning. (This circumstance is not unusual, as $\|\nabla f - \nabla g\|_\infty$ often dominates the discretization error.)

At first, this result might seem useful only for interpolation problems where f is known in advance, because superaccurate elements can be reliably generated only with a very accurate foreknowledge of the principle directions of curvature of f. However, PDE solvers sometimes present circumstances where these directions are easily inferred from domain boundaries. For example, consider the simulation of air flow over an airplane wing. The regions of greatest anisotropy usually appear where laminar flow passes over the surface of the wing. $\mathbf{v}_{\max}$ is perpendicular to the wing surface and the two eigenvectors parallel to the wing surface have much smaller eigenvalues. The best performance might be obtained by tetrahedra whose sizes are proportional to $1/\xi_i$ along each axis $\mathbf{v}_i$ (flat elements parallel to the surface), if the tetrahedra have no large dihedral angles.

As a precondition for superaccuracy, assume that the Hessian H is constant over an element t. (A sufficiently small variation in H will not hurt the error bound much, but this is not analyzed here.) Change one of the assumptions used in Section 2.2: C_t is no longer any matrix for which f satisfies (3). Instead, let C_t be the matrix with the same eigenvectors and eigenvalues as H, except that each negative eigenvalue $-\xi$ of H is replaced with the positive eigenvalue ξ in C_t, so C_t is positive definite. (If any eigenvalue of H is zero, substitute a tiny positive ξ in C_t.) In other words, let $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ be the unit eigenvectors of H, so that

$$H = \left[\begin{array}{ccc} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{array}\right] \left[\begin{array}{ccc} \pm\xi_1 & 0 & 0 \\ 0 & \pm\xi_2 & 0 \\ 0 & 0 & \pm\xi_3 \end{array}\right] \left[\begin{array}{ccc} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{array}\right]^T, \tag{4}$$

and define C_t with Equation (2) as usual.

With this assumption, the bounds in Table 3 improve to

$$\begin{aligned} \|\nabla f - \nabla g\|_\infty &\leq c_t \frac{\frac{1}{4A}(\ell_1\ell_2\widehat{\ell}_3^{\,2} + \ell_3\ell_1\widehat{\ell}_2^{\,2} + \ell_2\ell_3\widehat{\ell}_1^{\,2}) + \max_{1\leq i<j\leq 3}\left(\ell_i\widehat{\widehat{\ell}}_j + \ell_j\widehat{\widehat{\ell}}_i\right)}{\ell_1+\ell_2+\ell_3} \text{ for triangles;} \\ \|\nabla f - \nabla g\|_\infty &\leq c_t \frac{\frac{1}{6V}\sum_{1\leq i<j\leq 4} A_iA_j\widehat{\ell_{ij}}^{\,2} + \max_i \sum_{j\neq i} A_j\widehat{\widehat{\ell_{ij}}}}{\sum_{m=1}^4 A_m} \text{ for tetrahedra,} \end{aligned}$$

where $\widehat{\widehat{\ell}}_i = |E^2(v_j - v_k)|$ and $\widehat{\widehat{\ell_{ij}}} = |E^2(v_i - v_j)|$ are edge lengths of the element found by applying the transformation E to t twice.

Unfortunately, it is rarely possible to use these error bounds in software for finite element mesh generation or mesh improvement, because the value of H is not known accurately enough, and thus neither is the transformation E. However, a mesh generator might still be able to exploit the phenomenon of superaccuracy if the eigenvectors $\mathbf{v}_i$ of H can be accurately guessed. Somewhat less accurate approximations of the eigenvalues ξ_i will do. The key is this: for an element t with aspect ratio κ to exhibit superaccuracy, it must be nearly enough parallel to $\mathbf{v}_{\min}$ that two successive applications of the transformation E yield a small element (whose diameter is not much greater than the narrowest dimension of t).

The following thought experiment helps to clarify the upper bounds on $\|\nabla f - \nabla g\|_\infty$. Consider an equilateral triangle t with edge lengths ℓ, in which edge 1 is parallel to the x-axis (Figure 7, upper left). Suppose that the curvature of f in the y-direction is smaller than the curvature in the x-direction, so the transformation E maps a point $p = (x, y)$ to $\widehat{p} = (x, y/\sqrt{\kappa})$. Then $\widehat{\ell}_1 = \ell$, and $\widehat{\ell}_2$ and $\widehat{\ell}_3$ are at least $\ell/2$, so the correctional term is between $1/2$ and 1, and $\|\nabla f - \nabla g\|_\infty \in \Theta(\ell)$.

But the ideal triangle is tall and thin, so imagine stretching t vertically by a scaling factor s. As s increases, one of t's angles approaches $0°$, but no large angle appears. The triangle dimensions satisfy

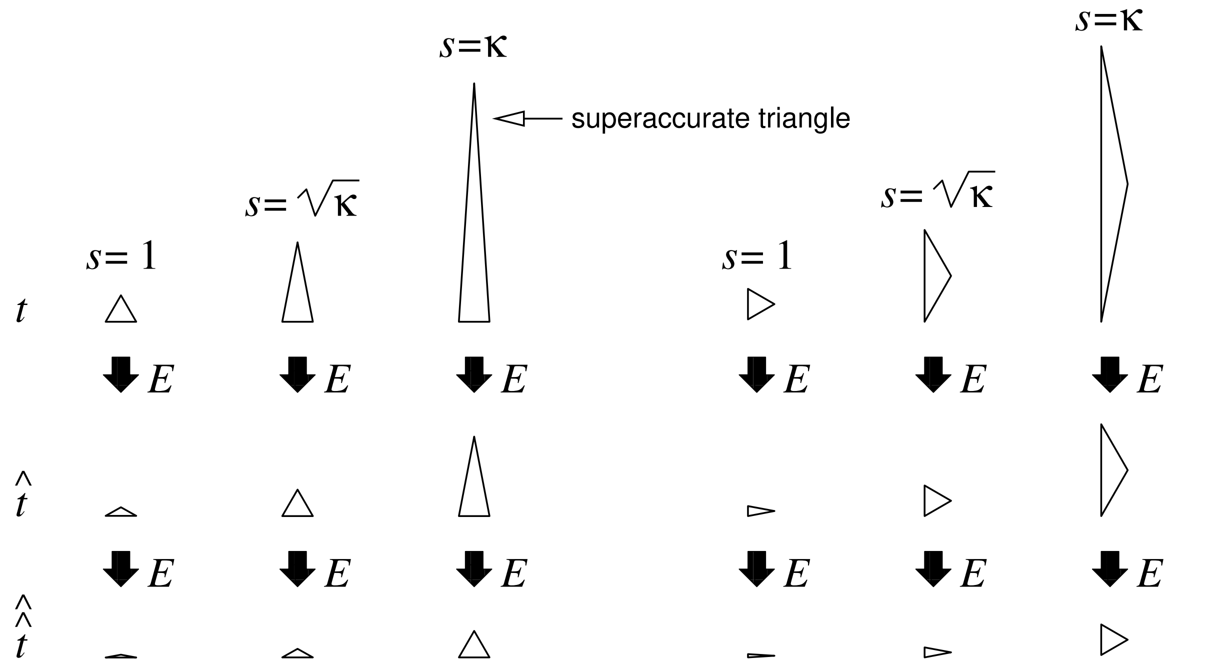


Figure 7: The error bounds for $\|\nabla f - \nabla g\|_\infty$ are in $\Theta(\ell)$ for every triangle in the top row except the upper rightmost triangle, whose bound is in $\Theta(\kappa\ell)$.

$A = (\sqrt{3}/4)s\ell^2$, $\ell_1 = \widehat{\ell}_1 = \ell$, and $\ell_2, \ell_3 \in \Theta(s\ell)$. Interestingly, $\widehat{\ell}_2$ and $\widehat{\ell}_3$ are in $\Theta(\ell)$ until s exceeds $\sqrt{\kappa}$, and $\widehat{\widehat{\ell}}_2$ and $\widehat{\widehat{\ell}}_3$ are in $\Theta(\ell)$ until s exceeds κ. When $s = \sqrt{\kappa}$, $\widehat{t}$ is an equilateral triangle. As s continues to grow, the vertical height of $\widehat{t}$ begins to dominate its horizontal width, and $\widehat{\ell}_2$ and $\widehat{\ell}_3$ grow with $\Theta(s\ell/\sqrt{\kappa})$. The terms $\ell_1\ell_2\widehat{\ell}_3^{\,2}$ and $\ell_3\ell_1\widehat{\ell}_2^{\,2}$ do not begin to dominate $\ell_2\ell_3\widehat{\ell}_1^{\,2}$ until s exceeds κ, though. Hence,

$$\|\nabla f - \nabla g\|_\infty \in \begin{cases} \Theta(\ell), & s \leq \kappa \\ \Theta(s\ell/\kappa), & s \geq \kappa. \end{cases}$$

The area A of t is proportional to s. By comparison, when a triangle is scaled isotropically (along both axes, without changing its aspect ratio) by a factor of $\sqrt{s}$, $\|\nabla f - \nabla g\|_\infty \in \Theta(\sqrt{s}\ell)$ and A is proportional to s. It follows that the aspect ratio that offers the best tradeoff between the error and t's area is roughly κ.

This is an odd result for two reasons. First, if t has an aspect ratio of κ, $\widehat{t}$ is nowhere near equilateral, having an aspect ratio of $\sqrt{\kappa}$. Second, whereas the error bound for $\|f - g\|_\infty$ finds its sweet spot when t has an aspect ratio of $\sqrt{\kappa}$, the error bound for $\|\nabla f - \nabla g\|_\infty$ finds its sweet spot when t has an aspect ratio of κ.

Let's repeat the thought experiment, but this time with edge 1 parallel to the y-axis (see the right half of Figure 7). Now, as t stretches, one of its angles approaches 180°. The only edge lengths whose asymptotic behavior changes are $\ell_1 = s\ell$ and $\widehat{\ell}_1 = s\ell/\sqrt{\kappa}$, so

$$\|\nabla f - \nabla g\|_\infty \in \begin{cases} \Theta(\ell), & s \leq \sqrt{\kappa} \\ \Theta(s^2\ell/\kappa), & s \geq \sqrt{\kappa}. \end{cases}$$

Thus, the aspect ratio that offers the best tradeoff between the error and the area of a large-angled element is roughly $\sqrt{\kappa}$. The element's accuracy deteriorates quickly if its aspect ratio increases beyond that. A triangle with a large angle cannot achieve as good a tradeoff as one without.

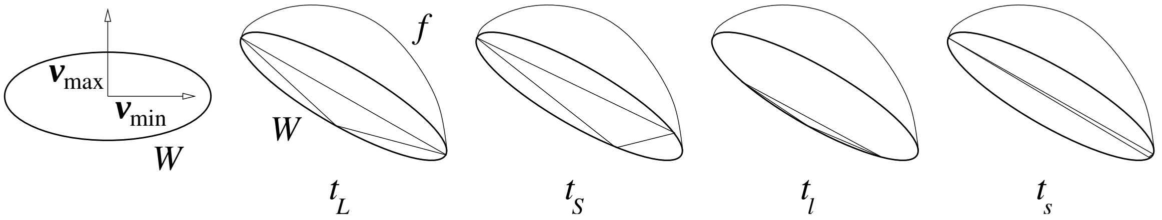


Figure 8: Oblique view of four triangles over which the function $f(p) = (\xi_{\max} - p^T C_t p)/2$ is interpolated. The ellipse W is the solution to $f(p) = 0$. $\|\nabla f - \nabla g\|_\infty$ is smaller over t_s than over the other three triangles because the vertices of t_s all lie near the axis along which the slope of f is smallest.

Here is the intuition behind why an element can be superaccurate. Let C_t be a curvature tensor with eigenvalues $\xi_{\max}$ and $\xi_{\min}$. Define the function $f(p) = (\xi_{\max} - p^T C_t p)/2$, whose Hessian matrix is $-C_t$. The solution to $f(p) = 0$ is an ellipse W whose axes have radii 1 and $\sqrt{\kappa}$, respectively. (See Figure 8, left.) The maximum value of $|\nabla f(p)|$ on W is $\xi_{\max}$, which is obtained at the intersection of W with its short axis ($\mathbf{v}_{\max}$). The minimum value of $|\nabla f(p)|$ on W is $\xi_{\max}/\sqrt{\kappa}$, obtained where W intersects its long axis ($\mathbf{v}_{\min}$).

Consider a triangle t whose three vertices lie on W. Piecewise linear interpolation sets $g(p) \equiv 0$ over t, so $\|\nabla f - \nabla g\|_\infty = \|\nabla f\|_\infty$. Figure 8 illustrates four triangles whose vertices lie on W, all taken from the thought experiments above. The triangles t_L and t_S have aspect ratios of roughly $\sqrt{\kappa}$, but t_L has an angle near 180° and t_S has none over 90°. Both triangles have a vertex v far enough from the long axis of the ellipse so that $|\nabla f(v)| \sim \xi_{\max}$. The triangles t_l and t_s also have one large angle and none, respectively, but they both have aspect ratios of roughly κ, and so are much thinner than t_L and t_S and cover less area by a factor of $\sqrt{\kappa}$. Because the vertices of t_s are so close to the long axis of W, the maximum value of $|\nabla f|$ over t_s is $\sim \xi_{\max}/\sqrt{\kappa}$. Therefore, the error bound for t_s is a factor of $\sqrt{\kappa}$ smaller than the error bounds for t_S and t_L, which more than justifies its reduced area. By contrast, t_l has a vertex v with an angle near 180°, so v must be far from the long axis of W and $|\nabla f(v)| \sim \xi_{\max}$. Thus t_l offers no improvement in accuracy over t_L despite its reduced area.

The reason a superaccurate t must have no large angle is so that both vertices of the shortest edge of t will lie near the long axis of W. The reason t must be nearly parallel to $\mathbf{v}_{\min}$ is so that the pointed end of t will lie near the long axis of W.

Let's summarize. If H is nearly constant over each triangle, the best compromise between the error $\|\nabla f - \nabla g\|_\infty$ and the number of elements is obtained by very precisely oriented triangles that have aspect ratio κ and no large angles (near 180°) in physical space. Properly oriented triangles with aspect ratio $\sqrt{\kappa}$ (with or without a large angle) are second best, so long as they do not have a large angle when mapped to isotropic space. However, it takes $\sim \sqrt{\kappa}$ more such triangles than ideal triangles to obtain the same error bound. Unless κ is near one, isotropic triangles are poor, because it takes $\sim \kappa$ more isotropic triangles than ideal ones to obtain the same bound. Anisotropic triangles oriented in the wrong direction can be even worse.

However, triangles with aspect ratio $\sqrt{\kappa}$ are better than triangles with aspect ratio κ for minimizing $\|f - g\|_\infty$, and often for stiffness matrix conditioning and discretization error too, as Sections 3.2 and 4.2 demonstrate. The ideal aspect ratio depends on which of the four criteria is more important for a particular application. Moreover, the superaccuracy of triangles with aspect ratio κ is fragile, and disappears if any of several requirements are not met. The thinner triangles are difficult to generate automatically because of the requirement that their angles not come too close to 180°. Some advancing front mesh generators can achieve this feat in critical regions; for instance, near airplane wings. In arbitrary domains, however,

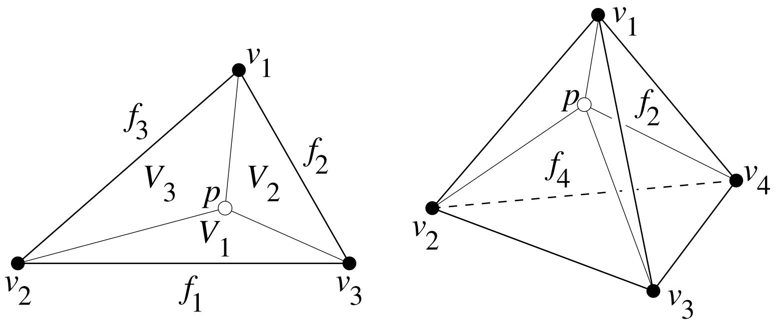


Figure 9: A point p splits a triangle into three, or a tetrahedron into four. The barycentric coordinates of p are based on the proportion of the total area or volume V_i taken by each sub-triangle or sub-tetrahedron.

generating triangles with aspect ratio $\sqrt{\kappa}$ that look well shaped in isotropic space, but may have large angles in physical space, is a more realistic goal.

The error bounds for tetrahedra have a similar interpretation to the bounds for triangles. A good compromise between the error $\|\nabla f - \nabla g\|_\infty$ and the number of elements is obtained by choosing tetrahedra that have no large angles in isotropic space, and whose size along each axis $\mathbf{v}_i$ is proportional to $1/\sqrt{\xi_i}$ in physical space. If all the requirements for superaccuracy are met, even better results can be obtained with even thinner tetrahedra whose size along each axis $\mathbf{v}_i$ is proportional to $1/\xi_i$, but only if they have no large dihedral angles in physical space. For example, if two of the eigenvalues are $\xi_{\max}$ and one is $\xi_{\min}$, needle-shaped tetrahedra that are aligned with $\mathbf{v}_{\min}$ and have one small face roughly orthogonal to $\mathbf{v}_{\min}$ are best. If one of the eigenvalues is $\xi_{\max}$ and two are $\xi_{\min}$, flat, pancake-shaped tetrahedra with one short edge roughly parallel to $\mathbf{v}_{\max}$ are best.

2.4 Derivation of the Error Bounds

This section derives the bounds given in Tables 2 and 3. Readers not interested in proofs can safely skip this section.

Piecewise linear interpolation is most easily expressed in terms of *barycentric coordinates*, so I review them first. Barycentric coordinates are a linear coordinate system based on a single triangle, tetrahedron, or (for arbitrary dimensionality d) simplex t. Let $v_1, v_2, \ldots, v_{d+1}$ be the vertices of t. The barycentric coordinates $\omega_1, \omega_2, \ldots, \omega_{d+1}$ of a point p simply express p as an affine combination of the vertices of t,

$$p = \sum_{i=1}^{d+1} \omega_i v_i, \text{ where } \sum_{i=1}^{d+1} \omega_i = 1. \tag{5}$$

Given a point p, a natural way to understand and compute its barycentric coordinates is by using the point p to divide t into $d+1$ simplices, as illustrated in Figure 9. Let V be the measure (i.e. area, volume, or equivalent for dimension d) of t. Let t_i be the simplex formed from t by replacing vertex v_i with p, and let $V_i(p)$ be its measure. Let f_i be the $(d-1)$-dimensional face of t opposite v_i; f_i is also a face of t_i.

If p is in t, then $t_1, t_2, \ldots, t_{d+1}$ are the simplices obtained by slicing up t into $d+1$ pieces meeting at p, and clearly $V = \sum_{i=1}^{d+1} V_i(p)$. If p is not in t, the same identity holds so long as $V_i(p)$ is defined to be negative if t_i is inverted—that is, if p and v_i are on opposite sides of f_i. The barycentric coordinates of p are

$$\omega_i(p) = \frac{V_i(p)}{V}.$$

Assuming V is positive, the values $\omega_i(p)$ are all nonnegative if and only if $p \in t$. The value $\omega_i(p)$ is zero if p lies in f_i. For any vertex v_i, $\omega_i(v_i) = 1$ and the other barycentric coordinates of v_i are zero.

Barycentric coordinates make linear interpolation easy. Given the values of $f(p)$ at the vertices of t, the value of the interpolated function $g(p)$, for any point $p \in t$, is

$$g(p) = \sum_{i=1}^{d+1} \omega_i(p) f(v_i).$$

Four observations about the gradients of barycentric coordinates help to derive bounds on interpolation error and (in Section 3.1) matrix conditioning. First, consider what the gradients look like. The altitude a_i of a vertex v_i is the shortest distance from v_i to the line that includes f_i (if t is a triangle), or from v_i to the plane that includes f_i (if t is a tetrahedron). A standard identity is $V = A_i a_i/d$, where A_i is the measure of f_i (e.g. the length ℓ_i if t is a triangle, an area if t is a tetrahedron, a volume if t is a four-dimensional simplex, etc.), and a_i is the signed altitude of v_i above f_i. (In triangles, for example, "area is half the base times height.") Likewise, $V_i = A_i \alpha_i(p)/d$, where $\alpha_i(p)$ is the altitude of p above f_i. Therefore,

$$\begin{aligned} \omega_i(p) &= \frac{V_i(p)}{V} \\ &= \frac{\alpha_i(p)}{a_i}, \text{ and therefore} \\ |\nabla \omega_i(p)| &= \frac{1}{a_i} |\nabla \alpha_i(p)| \\ &= \frac{1}{a_i}. \end{aligned} \tag{6}$$

$\nabla \omega_i(p)$ is a vector with length $1/a_i$ directed into t orthogonal to f_i. Because $\nabla \omega_i(p)$ is a constant vector (and does not vary with p), I will often abbreviate it to $\nabla \omega_i$.

Second, the gradients of the barycentric coordinates sum to the zero vector.

$$\sum_{i=1}^{d+1} \nabla \omega_i = \nabla \sum_{i=1}^{d+1} \omega_i(p) = \nabla 1 = \mathbf{0}. \tag{7}$$

Third, observe that for any vector $\mathbf{d}$,

$$\mathbf{d} \cdot p = \sum_{i=1}^{d+1} \omega_i(p) \mathbf{d} \cdot v_i.$$

Taking the gradient with respect to p,

$$\begin{aligned} \nabla(\mathbf{d} \cdot p) &= \sum_{i=1}^{d+1} (\mathbf{d} \cdot v_i) \nabla \omega_i \\ \mathbf{d} &= \sum_{i=1}^{d+1} (v_i \cdot \mathbf{d}) \nabla \omega_i. \end{aligned} \tag{8}$$

The fourth observation is specific to triangles. Let $\langle i, j, k \rangle$ be any permutation of $\{1, 2, 3\}$. Then

$$\begin{aligned}
\nabla\omega_i \cdot \nabla\omega_j &= \left(|\nabla\omega_i + \nabla\omega_j|^2 - |\nabla\omega_i|^2 - |\nabla\omega_j|^2\right)/2 \\
&= \left(|-\nabla\omega_k|^2 - |\nabla\omega_i|^2 - |\nabla\omega_j|^2\right)/2 \quad \text{from (7)} \\
&= \frac{1}{2a_k^2} - \frac{1}{2a_i^2} - \frac{1}{2a_j^2} \quad \text{from (6)} \\
&= \frac{\ell_k^2 - \ell_i^2 - \ell_j^2}{8A^2}. \qquad (9)
\end{aligned}$$

The derivation given here of the bounds on interpolation error initially follows Chapter 4 of Johnson [30], but a long departure is taken here to tighten the bounds. Let $e(p) = f(p) - g(p)$ be the error in $g(p)$ as an approximation of $f(p)$. Let's look at the properties of $e(p)$, restricted to a single triangle or tetrahedron t. First, $e(p)$ is zero at each vertex of t. Second, because $g(p)$ is linear over t, $e(p)$ has the same second and higher derivatives as $f(p)$. Therefore, the curvature of $e(p)$ is bounded by c_t.

The errors at two points p and q can be related to each other by integrating e along the line segment pq. Parameterize the line integral using $u(j) = (1-j)p + jq$, with $0 \le j \le 1$.

$$\begin{aligned}
e(q) &= e(p) + \int_0^1 (q-p)\cdot\nabla e(u(j))\,dj \\
&= e(p) + (q-p)\cdot\nabla e(p) + \int_0^1\int_0^j (q-p)^T H(u(k))(q-p)\,dk\,dj \\
&= e(p) + (q-p)\cdot\nabla e(p) + \frac{1}{2}(q-p)^T\mathcal{H}(q-p), \qquad (10)
\end{aligned}$$

where $\mathcal{H} = 2\int_0^1\int_0^j H(u(k))\,dk\,dj$ is a matrix defined by integrating each entry of $H(u)$ independently.[3]

If p and q lie in t (which will always be the case below), then by Inequality (1),

$$\begin{aligned}
|(q-p)^T\mathcal{H}(q-p)| &\le 2\int_0^1\int_0^j \left|(q-p)^T H(u(k))(q-p)\right|\,dk\,dj \\
&\le 2c_t\int_0^1\int_0^j |q-p|^2\,dk\,dj \\
&= c_t|q-p|^2. \qquad (11)
\end{aligned}$$

How large can the absolute error $|e(p)|$ be at a single point p in t? Substituting v_i for q in Inequality (10) yields

$$0 = e(p) + (v_i - p)\cdot\nabla e(p) + \frac{1}{2}(v_i - p)^T\mathcal{H}_i(v_i - p).$$

A subscript is appended to $\mathcal{H}$ because there is a different $\mathcal{H}_i$ for each vertex v_i of t. One would like to use this expression to derive a bound on $|e(p)|$, but the term $(v_i - p)\cdot\nabla e(p)$ is not easily bounded. Fortunately, the following trick will circumvent this obstacle. For each vertex of t, define a function

$$e_i(p) = -(v_i - p)\cdot\nabla e(p) - \frac{1}{2}(v_i - p)^T\mathcal{H}_i(v_i - p). \qquad (12)$$

[3] If f has C^2 continuity, this argument can be simplified. Using the mean value theorem, the Taylor expansion of e about p is

$$e(q) = e(p) + (q-p)\cdot\nabla e(p) + \frac{1}{2}(q-p)^T H(u)(q-p),$$

where u is some point on the line segment pq. However, the proof in the body applies even if f has only C^1 continuity.

Clearly, $e_i(p) = e(p)$ for each i, but only at a specified point p (because each matrix $\mathcal{H}_i$ is determined by the value of p).

The trick is to express $e(p)$ as a weighted sum of the functions $e_i(p)$, where the weights are the barycentric coordinates. With this choice of weights, the $(v_i - p) \cdot \nabla e(p)$ terms cancel each other out. Note that because p is in t, the barycentric coordinates are all nonnegative. As the barycentric coordinate sum to one,

$$\begin{aligned}
e(p) &= \sum_i \omega_i(p) e(p) \\
&= \sum_i \omega_i(p) e_i(p) \\
&= -\left(\sum_i \omega_i(p) v_i - p \sum_i \omega_i(p)\right) \cdot \nabla e(p) - \frac{1}{2} \sum_i \omega_i(p)(v_i - p)^T \mathcal{H}_i (v_i - p) \quad \text{from (12)} \\
&= -(p - p) \cdot \nabla e(p) - \frac{1}{2} \sum_i \omega_i(p)(v_i - p)^T \mathcal{H}_i (v_i - p) \quad \text{from (5), and therefore} \\
|e(p)| &\leq \frac{c_t}{2} \sum_i \omega_i(p) |v_i - p|^2 \quad \text{from (11).} \qquad (13)
\end{aligned}$$

The bound (13) is valid at any point p in t. Let $b(p)$ be the right-hand side of (13). Observe that $b(p) = 0$ when p is a vertex of t. To find the maximum value of $b(p)$, set $\nabla b(p) = \mathbf{0}$.

$$\begin{aligned}
\nabla b(p) &= \frac{c_t}{2} \sum_i \left(|v_i - p|^2 \nabla \omega_i + 2\omega_i(p)(p - v_i) \right) \\
&= \frac{c_t}{2} \sum_i |v_i - p|^2 \nabla \omega_i \quad \text{from (5).}
\end{aligned}$$

Identity (7) tells us that $\sum_{i=1}^{d+1} \nabla \omega_i = \mathbf{0}$, but if t is nondegenerate, any d of the $\nabla \omega_i$ are linearly independent. Therefore, $\nabla b(p) = \mathbf{0}$ only when all the $|v_i - p|$ are equal, which holds true only if p is the center of the circumscribing circle or sphere of t.

Let O_{circ} and r_{circ} be the circumcenter and circumradius of t, respectively. The error bound $b(p)$ is maximized at $p = O_{\text{circ}}$, with the value $c_t r_{\text{circ}}^2/2$. However, if O_{circ} does not lie in t, a better bound is available.

Expanding the right-hand side of (13) yields

$$\begin{aligned}
b(p) &= \frac{c_t}{2} |p|^2 \sum_i \omega_i(p) - c_t p \cdot \sum_i \omega_i(p) v_i + \frac{c_t}{2} \sum_i \omega_i(p) |v_i|^2 \\
&= -\frac{c_t}{2} |p|^2 + \frac{c_t}{2} \sum_i \omega_i(p) |v_i|^2 \quad \text{from (5),}
\end{aligned}$$

and thus the bound is quadratic in p. There is only one quadratic function that obtains a maximum value of $c_t r_{\text{circ}}^2/2$ at O_{circ} and a value of zero at each vertex v_i, namely

$$b(p) = c_t (r_{\text{circ}}^2 - |p - O_{\text{circ}}|^2)/2, \qquad p \in t.$$

This function is illustrated in Figure 10. The function $e(p) = b(p)$ obtains the maximum possible error $b(p)$ at every point in t, and satisfies the constraint (1) because $b(p)$ has curvature $-c_t$ everywhere, so $b(p)$ is a

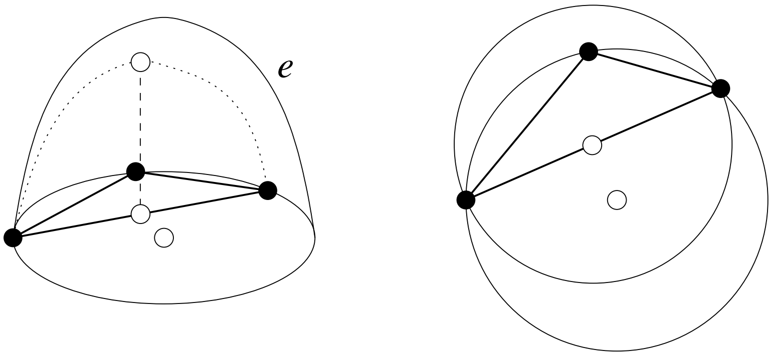


Figure 10: Left: two lower bounds for triangles are established when $e(p)$ is a paraboloid centered at the triangle's circumcenter. Within the triangle, the maximum value of $e(p)$ occurs at the center of the min-containment circle, and the maximum magnitude of $\nabla e(p)$ occurs at the vertices. Right: top view of the triangle, its circumcircle, and its min-containment circle.

pointwise tight error bound and can be used to calculate a tight error bound in any norm. For the L_∞ norm, $b(p)$ is maximized when p is the point in t closest to O_{circ}. What point is that?

Let O_{mc} be the center of the min-containment circle or sphere of t, and let r_{mc} be its radius. O_{mc} always lies in t. Lemma 1, which is postponed to the end of this section, shows that the point in t nearest O_{circ} is O_{mc}, with the error bound $b(O_{\mathrm{mc}}) = c_t r_{\mathrm{mc}}^2/2$. Therefore, $c_t r_{\mathrm{mc}}^2/2$ is a tight bound on $\|f - g\|_\infty$ over t, as Waldron [49] first showed.

Next, consider the error in the interpolated gradient, $\nabla e(p) = \nabla f(p) - \nabla g(p)$. To bound $\nabla e(p)$, the derivation uses another trick based on barycentric coordinates. Once again, choose a linear combination of the $e_i(p)$ functions (12), this time so that the $e(p)$ terms cancel each other out. From Identity (7) we have

$$
\begin{aligned}
\mathbf{0} &= e(p) \sum_i \nabla \omega_i \\
&= \sum_i e_i(p) \nabla \omega_i \\
&= -\sum_i [(v_i - p) \cdot \nabla e(p)] \nabla \omega_i - \frac{1}{2} \sum_i \left[(v_i - p)^T \mathcal{H}_i (v_i - p) \right] \nabla \omega_i \qquad \text{from (12)} \\
&= p \cdot \nabla e(p) \sum_i \nabla \omega_i - \sum_i [v_i \cdot \nabla e(p)] \nabla \omega_i - \frac{1}{2} \sum_i \left[(v_i - p)^T \mathcal{H}_i (v_i - p) \right] \nabla \omega_i \\
&= -\nabla e(p) - \frac{1}{2} \sum_i \left[(v_i - p)^T \mathcal{H}_i (v_i - p) \right] \nabla \omega_i \qquad \text{from (7) and (8).} \qquad (14) \\
|\nabla e(p)| &\le \frac{c_t}{2} \sum_i |v_i - p|^2 |\nabla \omega_i| \qquad \text{from (11) and the triangle inequality} \\
&= \frac{c_t}{2} \sum_i \frac{|v_i - p|^2}{a_i} \qquad \text{from (6).} \qquad (15)
\end{aligned}
$$

At this point in the derivation, most sources (for instance, see Johnson [30]) weaken the bound (15) to $(c_t/2)\ell_{\max}^2 \sum 1/a_i = (c_t/2)\ell_{\max}^2/r_{\mathrm{in}}$. The trouble with this bound is that it imputes an unduly large error to a triangle that has a small angle but no large angle. The following argument tightens the classical bounds and vindicates small angles too.

I begin with an easy but intuitive bound for triangles, then sharpen the analysis and generalize it to any dimension. The idea is to bound $|\nabla e(p)|$ at a specific point p, then extend the bound to the whole element t. To derive the easy bound, evaluate the expression (15) at the vertex where the triangle's shortest edge meets its median-length edge—this choice of vertex yields a particularly small bound. Call this vertex $v_{\max}$, because it is opposite the longest edge of t. The altitudes are $a_i = 2A/\ell_i$, where A is the area of the triangle. By substituting $v_{\max}$ for p, the bound (15) reduces to

$$|\nabla e(v_{\max})| \leq c_t \frac{\ell_{\text{med}}\ell_{\text{min}}(\ell_{\text{med}} + \ell_{\text{min}})}{4A}.$$

What about the other points in t? Any point in t is no further than ℓ_{med} from $v_{\max}$. The maximum rate of change of $|\nabla e(p)|$ with respect to p is the maximum curvature c_t. Therefore,

$$|\nabla e(p)| \leq c_t \frac{\ell_{\text{med}}\ell_{\text{min}}(\ell_{\text{med}} + \ell_{\text{min}})}{4A} + c_t \ell_{\text{med}} \quad \text{for all } p \in t.$$

This bound is already enough to pardon small angles. The factor of $\ell_{\min}$ in the numerator is the key difference between this bound and the classical bound $(c_t/2)\ell_{\max}^2/r_{\text{in}}$. A triangle with a tiny angle has a tiny area A, but the tiny value of $\ell_{\min}$ compensates, so the error bound does not explode.

To obtain a tighter bound for triangles and a bound for higher-dimensional simplicial elements, evaluate the expression (15) at the point where it is minimized. Basic calculus shows that this point is

$$\begin{aligned} O_{\text{in}} &= \frac{1}{\sum_i 1/a_i} \sum_i \frac{1}{a_i} v_i \\ &= \frac{1}{\sum_i A_i} \sum_i A_i v_i. \end{aligned}$$

This formula holds for a simplex t of any dimension. Recall that A_i is the measure of the face of t opposite vertex i. As chance would have it, O_{in} is the *incenter* (center of the inscribed circle or sphere) of t.

Let ℓ_{ij} be the length of the edge connecting vertices v_i and v_j. For any i, let $\ell_{ii} = 0$. Recall that V is the measure of t, and $a_i = dV/A_i$. Substituting $p = O_{\text{in}}$ into the bound (15) gives

$$\begin{aligned} |\nabla e(O_{\text{in}})| &\leq \frac{c_t}{2} \sum_{i=1}^{d+1} \frac{1}{a_i} |v_i - O_{\text{in}}|^2 \\ &= \frac{c_t}{2dV} \sum_{i=1}^{d+1} A_i \frac{|\sum_j A_j (v_i - v_j)|^2}{(\sum_m A_m)^2} \\ &= \frac{c_t}{2dV} \sum_{i=1}^{d+1} A_i \frac{\sum_{j,k} A_j A_k (v_i - v_j) \cdot (v_i - v_k)}{(\sum_m A_m)^2} \\ &= \frac{c_t}{2dV} \frac{\sum_{i,j,k} A_i A_j A_k (|v_i|^2 - v_i \cdot v_j - v_i \cdot v_k + v_j \cdot v_k)}{(\sum_m A_m)^2} \\ &= \frac{c_t}{2dV} \frac{\sum_{i,j,k} \frac{1}{2} A_i A_j A_k (|v_i|^2 + |v_j|^2 - 2 v_i \cdot v_j)}{(\sum_m A_m)^2} \\ &= \frac{c_t}{2dV} \frac{\sum_{i,j,k} \frac{1}{2} A_i A_j A_k |v_i - v_j|^2}{(\sum_m A_m)^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{c_t}{2dV} \frac{(\sum_k A_k)(\sum_{i,j} \frac{1}{2} A_i A_j \ell_{ij}^2)}{(\sum_m A_m)^2} \\
&= \frac{c_t}{2dV} \frac{\sum_{i<j} A_i A_j \ell_{ij}^2}{\sum_m A_m}.
\end{aligned}
$$

Now consider the value of $|\nabla e|$ at other points in t. The distance between a vertex of t and O_{in} is

$$
\begin{aligned}
|v_i - O_{\text{in}}| &= \frac{|\sum_j A_j (v_i - v_j)|}{\sum_m A_m} \\
&\leq \frac{\sum_{j \neq i} A_j \ell_{ij}}{\sum_m A_m}.
\end{aligned}
$$

Because the maximum rate of change of $|\nabla e(p)|$ is c_t, for any p in t,

$$
\begin{aligned}
|\nabla e(p)| &\leq |\nabla e(O_{\text{in}})| + c_t \max_i |v_i - O_{\text{in}}| \\
&\leq \frac{c_t}{2dV} \frac{\sum_{i<j} A_i A_j \ell_{ij}^2}{\sum_m A_m} + c_t \max_i \frac{\sum_{j \neq i} A_j \ell_{ij}}{\sum_m A_m},
\end{aligned}
$$

yielding the upper bound for tetrahedra in Table 2.

A simpler and smoother but weaker bound follows from the fact that for any face (with measure A_i) and edge (with length ℓ_{ij}) of a simplex that intersect at one vertex, the inequality $dV \leq A_i \ell_{ij}$ holds, with equality if and only if the edge and face are perpendicular to each other.

$$
\begin{aligned}
|\nabla e(p)| &\leq \frac{c_t}{2dV} \frac{\sum_{i<j} A_i A_j \ell_{ij}^2}{\sum_m A_m} + \frac{c_t}{dV} \max_i \frac{\sum_{j \neq i} A_i A_j \ell_{ij}^2}{\sum_m A_m} \\
&\leq \frac{3c_t}{2dV} \frac{\sum_{i<j} A_i A_j \ell_{ij}^2}{\sum_m A_m}.
\end{aligned}
$$

However, if the edge and face are in fact nearly parallel, this simplification can unnecessarily lose a factor of up to three (as it does for a triangle with an angle near 180°).

For triangles, these bounds can be simplified, because A_i is another name for ℓ_i and ℓ_{ij} is another name for ℓ_k, where k is distinct from i and j. Hence,

$$
\begin{aligned}
|\nabla e(p)| &\leq \frac{c_t}{4A} \frac{\sum_{i<j} \ell_i \ell_j \ell_k^2}{\ell_1 + \ell_2 + \ell_3} + c_t \max_{j,k \neq j} \frac{2\ell_j \ell_k}{\ell_1 + \ell_2 + \ell_3} \\
&= \frac{c_t}{4A} \ell_1 \ell_2 \ell_3 + \frac{4c_t}{4A} \ell_{\max} \ell_{\text{med}} r_{\text{in}},
\end{aligned}
$$

yielding the upper bound for triangles in Table 2. The simpler but weaker bound above likewise simplifies to $3c_t \ell_1 \ell_2 \ell_3 / (4A)$ for triangles.

Next, consider the lower bounds on $\|\nabla f - \nabla g\|_\infty$ in Table 2. These bounds are shown by exhibiting, for any fixed element t, a function f that obtains the lower bound. The bound for a triangle is the maximum of three different choices of f, illustrated in Figure 11. Which of these choices is largest depends on the shape of the triangle.

A lower bound of $c_t r_{\text{circ}}$ is established by the parabolic function $e(p) = c_t(r_{\text{circ}}^2 - |p - O_{\text{circ}}|^2)/2$, for which $|\nabla e(v_i)| = c_t r_{\text{circ}}$ at each vertex v_i of t. See Figure 10. The circumradius of a triangle is given by

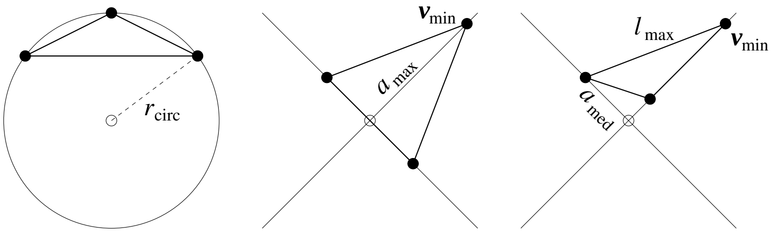


Figure 11: Three lower bounds for $\|\nabla f - \nabla g\|_\infty$ on a triangle. Thin lines represent zeros of the function $c_t(r_{\text{circ}}^2 - x^2 - y^2)/2$ (left) or $c_t(x^2 - y^2)/2$ (center and right). The vertex furthest from the origin lies at a distance of r_{circ}, $a_{\max}$, or $(\ell_{\max}^2 - a_{\text{med}}^2)^{1/2}$, respectively.

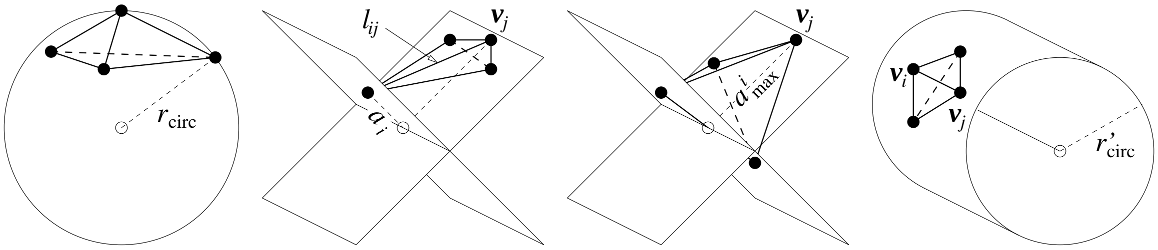


Figure 12: Four lower bounds for $\|\nabla f - \nabla g\|_\infty$ on a tetrahedron. Surfaces represent zeros of the function $c_t(r_{\text{circ}}^2 - x^2 - y^2 - z^2)/2$ (left), $c_t(x^2 - y^2)/2$ (middle two), or $c_t(r_{\text{circ}}'^2 - x^2 - y^2)/2$ (right). The furthest vertex from the z-axis lies at a distance of r_{circ}, $\max_{i,j}(\ell_{ij}^2 - a_i^2)^{1/2}$, $a_{\max}^i$, or r'_{circ}, respectively.

the formula $r_{\text{circ}} = \ell_{\max}\ell_{\text{med}}\ell_{\min}/(4A)$, so this lower bound establishes that both upper bounds are tight to within a factor of three. (The stronger upper bound is tight to within a factor of two. The proof, which takes advantage of the following two lower bounds as well, is omitted.)

For some triangles, a stronger lower bound is established by the hyperbolic function $e(p) = c_t(x^2 - y^2)/2$. Position the triangle (or the coordinate system) so that each vertex lies on one of the lines $y = x$ or $y = -x$, thus $e(v_i) = 0$ for each vertex v_i. Finding a lower bound becomes a game of placing one of the triangle's vertices as far from the origin as possible. A lower bound of $c_t a_{\max}$ follows by placing $v_{\min}$, the vertex opposite the shortest edge, on $y = x$, and the shortest edge itself on $y = -x$. Then $|\nabla e(v_{\min})| = c_t a_{\max}$. A lower bound of $c_t(\ell_{\max}^2 - a_{\text{med}}^2)^{1/2}$ follows by placing the median-length edge on $y = x$ and the opposite vertex v_{med} on $y = -x$. Then $|\nabla e(v_{\min})| = c_t(\ell_{\max}^2 - a_{\text{med}}^2)^{1/2}$.

The bound $c_t a_{\max}$ is the strongest of the three lower bounds for a triangle with no obtuse angle, including any equilateral triangle. The other two bounds dominate for triangles with obtuse angles. The bound $c_t r_{\text{circ}}$ asymptotically dominates the other two, as it approaches infinity when an angle approaches $180°$. These are not the best lower bounds possible. For some triangles, reducing the curvature of $e(p)$ along one axis can yield a stronger lower bound. I do not pursue these possibilities here.

Four lower bounds for tetrahedra are illustrated in Figure 12. The lower bound of $c_t r_{\text{circ}}$ applies to tetrahedra and higher-dimensional simplicial elements as well as triangles. The circumradius of a tetrahedron is at least as great as the circumradius of any of its faces, so faces with planar angles near $180°$ cause large errors. For some tetrahedra, though, $c_t r_{\text{circ}}$ is a weak bound. Sliver tetrahedra, for example, can have small circumspheres and nicely shaped faces (as Figure 3 shows) yet be very bad for interpolation.

The other two lower bounds for triangles also generalize to tetrahedra, with a twist. Again, consider the function $e(p) = c_t(x^2 - y^2)/2$. Finding a lower bound is a game of placing the tetrahedron's vertices on the two planes $x = \pm y$ so that one vertex lies as far as possible from the z-axis. A lower bound of

$c_t \max_{i,j}(\ell_{ij}^2 - a_i^2)^{1/2}$ follows by placing face i on the plane $y = x$ and vertex v_i on the plane $y = -x$, with edge $v_i v_j$ aligned parallel to the xy-plane. The bound is achieved at vertex v_j. The altitudes of a tetrahedron also yield lower bounds, but a better lower bound is found by choosing the largest altitude of a triangular face of the tetrahedron. Let $a_{\max}^i$ be the maximum altitude of face i, and suppose this altitude meets vertex v_j. A lower bound of $c_t a_{\max}^i$ follows by placing v_j on $y = x$ and the other two vertices of face i on the z-axis. Finally, rotate the tetrahedron around the line $x = y, z = 0$ until the fourth vertex of the tetrahedron lies on $y = x$. Then $|\nabla e(v_j)| = c_t a_{\max}^i$.

There is a large lower bound available for any tetrahedron t with a large dihedral angle, such as a sliver. Let edge $v_i v_j$ be the edge of t with the largest dihedral angle θ_{ij}, and let v_k and v_l be the other two vertices. Rotate t so that edge $v_i v_j$ is parallel to the z-axis. The projection of t onto the xy-plane is a triangle, t', which has a planar angle of θ_{ij}. Let r'_{circ} be the circumradius of t'. Consider the function $e(p) = c_t(r'^2_{\text{circ}} - x^2 - y^2)/2$, whose zeros form a cylinder, and translate t so all four of its vertices lie on this cylinder. At any vertex v of t, $|\nabla e(v)| = c_t r'_{\text{circ}} = A_k A_l |(v_i - v_j) \times (v_k - v_l)|/(3V\ell_{ij}^2)$. If θ_{ij} approaches 180°, r'_{circ} approaches infinity, thus proving that large dihedral angles can cause the same explosive errors as large planar angles.

In summary, for any tetrahedron t there is a function f satisfying the curvature constraint (1) such that

$$\|\nabla f - \nabla g\|_\infty \geq c_t \max\left\{ r_{\text{circ}}, \max_i a_{\max}^i, \max_{i,j\neq i} \sqrt{\ell_{ij}^2 - a_i^2}, \max \frac{A_k A_l |(v_i - v_j) \times (v_k - v_l)|}{3V\ell_{ij}^2} \right\}.$$

Next, consider the anisotropic bounds on $\nabla f - \nabla g$. The derivation is similar to the isotropic case, but departs with Equation (14), from which we have

$$\nabla e(p) = -\frac{1}{2} \sum_i \left[(v_i - p)^T \mathcal{H}_i (v_i - p) \right] \nabla \omega_i.$$

The curvature bound (3) implies that

$$\begin{aligned}
|(v_i - p)^T \mathcal{H}_i (v_i - p)| &\leq (v_i - p)^T C_t (v_i - p) \\
&= c_t (v_i - p)^T E^2 (v_i - p) \\
&= c_t |E(v_i - p)|^2, \text{ and therefore} \\
|\nabla e(p)| &\leq \frac{c_t}{2} \sum_i |E(v_i - p)|^2 |\nabla \omega_i| \\
&= \frac{c_t}{2} \sum_i \frac{|E(v_i - p)|^2}{a_i}. \qquad (16)
\end{aligned}$$

The bound (16) is similar to (15), but the points in the numerator are transformed to isotropic space. Both bounds are minimized at $p = O_{\text{in}}$. By analogy to the isotropic derivation, at O_{in} the bound (16) satisfies

$$\begin{aligned}
|\nabla e(O_{\text{in}})| &\leq \frac{c_t}{2dV} \sum_{i=1}^{d+1} A_i \frac{|\sum_j A_j (\widehat{v}_i - \widehat{v}_j)|^2}{(\sum_m A_m)^2} \\
&= \frac{c_t}{2dV} \frac{\sum_{i<j} A_i A_j \widehat{\ell_{ij}}^2}{\sum_m A_m}.
\end{aligned}$$

To bound $|\nabla e(p)|$ at any other point p in t, integrate ∇e along the line segment $O_{\text{in}}p$, parameterized by $u(j) = (1-j)O_{\text{in}} + jp$ (for $0 \leq j \leq 1$). Because e is smooth,

$$\begin{aligned}\nabla e(p) &= \nabla e(O_{\text{in}}) + \int_0^1 H(u(j))(p - O_{\text{in}})\,dj, \qquad \text{and therefore}\\ |\nabla e(p)| &\leq |\nabla e(O_{\text{in}})| + \int_0^1 |H(u(j))(p - O_{\text{in}})|\,dj.\end{aligned}$$

The only difficulty is bounding the term $|H(u)(p - O_{\text{in}})|$.

From Inequality (3) and the substitution $\widehat{\mathbf{d}} = E\mathbf{d}$ it follows that for any point $u \in t$ and any vector $\widehat{\mathbf{d}}$,

$$\begin{aligned}|\widehat{\mathbf{d}}^T E^{-1} H(u) E^{-1} \widehat{\mathbf{d}}| &\leq \widehat{\mathbf{d}}^T E^{-1} C_t E^{-1} \widehat{\mathbf{d}}\\ &= c_t |\widehat{\mathbf{d}}|^2.\end{aligned}$$

Therefore, each eigenvalue of $E^{-1}H(u)E^{-1}$ has a magnitude no greater than c_t. Because the matrices E and $E^{-1}H(u)E^{-1}$ are symmetric (and therefore have orthogonal eigenvectors) with eigenvalues no larger than 1 and c_t respectively, for any vector $\mathbf{d}$,

$$\begin{aligned}|H(u)\mathbf{d}| &= |EE^{-1}H(u)E^{-1}E\mathbf{d}|\\ &\leq |E^{-1}H(u)E^{-1}E\mathbf{d}|\\ &\leq c_t|E\mathbf{d}|. \qquad (17)\end{aligned}$$

The error in the gradient is thus bounded by the inequality

$$|\nabla e(p)| \leq |\nabla e(O_{\text{in}})| + c_t|E(p - O_{\text{in}})|.$$

This bound is maximized when p is one of the vertices of t, so for any point $p \in t$,

$$\begin{aligned}|\nabla e(p)| &\leq |\nabla e(O_{\text{in}})| + c_t \max_i |E(v_i - O_{\text{in}})|\\ &= |\nabla e(O_{\text{in}})| + c_t \max_i \frac{|\sum_j A_j E(v_i - v_j)|}{\sum_m A_m}\\ &\leq \frac{c_t}{2dV} \frac{\sum_{i<j} A_i A_j \widehat{\ell_{ij}}^2}{\sum_m A_m} + c_t \max_i \frac{\sum_{j \neq i} A_j \widehat{\ell_{ij}}}{\sum_m A_m}, \qquad (18)\end{aligned}$$

which yields the bounds on $\|\nabla f - \nabla g\|_\infty$ in Table 3. The weaker bound for triangles in Table 3 follows because $2\ell_i \leq \ell_1 + \ell_2 + \ell_3$ for any edge i of t. The weaker bound for tetrahedra follows because $3\widehat{\ell_{ij}} \leq \sum_{k=1}^6 \widehat{\ell_k}$ for any edge ij of $\widehat{t}$.

To verify the claims of superaccuracy in Section 2.3, consider the special case where H is constant over the triangle t, and choose $C_t = V\Xi V^T$ where V and ξ_i are defined by Equation (4). Then

$$\begin{aligned}|H\mathbf{d}| &= \left| V \begin{bmatrix} \pm\xi_1 & 0 & 0\\ 0 & \pm\xi_2 & 0\\ 0 & 0 & \pm\xi_3 \end{bmatrix} V^T \mathbf{d} \right|\\ &= |V\Xi V^T \mathbf{d}| \qquad \text{because } V \text{ is orthonormal}\\ &= |C_t \mathbf{d}|\\ &= c_t|E^2\mathbf{d}|,\end{aligned}$$

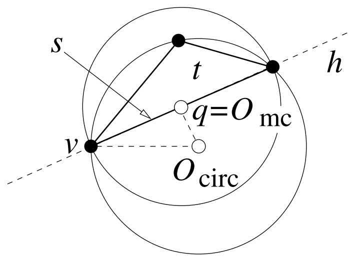


Figure 13: The center of the min-containment circle is the point in t nearest the circumcenter of t.

so Inequality (18) is strengthened to

$$|\nabla e(p)| \leq \frac{c_t}{2dV} \frac{\sum_{i<j} A_i A_j \widehat{\ell_{ij}}^2}{\sum_m A_m} + c_t \max_i \frac{\sum_{j \neq i} A_j \widehat{\widehat{\ell_{ij}}}}{\sum_m A_m} \quad \text{for all } p \in t.$$

There is one loose end to tie up, namely the proof that the center of the min-containment sphere is the point in t nearest the circumcenter.

Lemma 1 *Let O_{circ} and r_{circ} be the circumcenter and circumradius of a d-dimensional simplex t. Let q be the point in t nearest O_{circ}. Then the center and radius of the min-containment sphere of t are q and $(r_{\text{circ}}^2 - |O_{\text{circ}} - q|^2)^{1/2}$, respectively.*

Proof: Let s be the face of t (of any dimension) whose relative interior contains q. The face s is not a vertex, because the vertices of t lie on the circumsphere, and some portion of t lies inside the circumsphere. Because q is the point in s nearest O_{circ}, and because q is in the relative interior of s, the line segment $O_{\text{circ}}q$ is orthogonal to s. (This is true even if $O_{\text{circ}} \in t$, in which case $O_{\text{circ}} - q = \mathbf{0}$.) This fact, plus the fact that O_{circ} is equidistant from all the vertices of t, implies that q is equidistant from all the vertices of s (as Figure 13 demonstrates) and is thus the circumcenter of s. Let r be the distance between q and any vertex of s. Because q lies in s, there is no containing sphere of s (or t) with radius less than r, because there is no direction q can move without increasing its distance from one of the vertices of s.

It remains only to show that the sphere with center q and radius r encloses t. If $q = O_{\text{circ}}$, it follows immediately. Otherwise, let h be the hyperplane through q orthogonal to $O_{\text{circ}}q$. Observe that h contains s. No point in t is on the same side of h as O_{circ}; if there were such a point w, there would be a point of t (between w and q) closer to O_{circ} than q. Because t lies entirely in the portion of the circle on the other side of h, every point of t is within a distance of r from q. Therefore, q is the center of the min-containment sphere of t, and r is its radius.

Let v be any vertex of s. Pythagoras' Law on $\triangle O_{\text{circ}}qv$ (see Figure 13) yields $r_{\text{circ}}^2 = r^2 + |O_{\text{circ}} - q|^2$, and therefore $r = (r_{\text{circ}}^2 - |O_{\text{circ}} - q|^2)^{1/2}$. ■

For an algebraic proof of Lemma 1, see Rajan [39], Lemma 3.

The techniques described in this section can be used to bound other measures of interpolation error besides $\|f - g\|_\infty$ and $\|\nabla f - \nabla g\|_\infty$. Practitioners are often interested not just in the maximum pointwise error, but in the L_2 norm or the H^1 norm of the error over the domain. In many cases, a bound on pointwise errors in e and ∇e can be used to bound these and other measures of error. See Section 4.1 and Chapter 4 of Johnson [30] for examples.

2.5 Other Approaches to Error Estimation

Research on error estimates divides loosely into two categories. Some results, like Waldron's bound on $f - g$ and Handscomb's bound on $\|\nabla f - \nabla g\|_\infty$, are of the same character as those derived in this article: error bounds with constants as tight as the researcher can derive. However, the majority of work in the area (including the aforementioned papers of Bramble and Zlámal [13] and Babuška and Aziz [4]) is built on functional analysis and embedding theorems. The best results are much broader in scope than the work presented here—they cover piecewise polynomial functions of almost any degree, and they cover a variety of different norms. But these results are asymptotic in nature, and ignore the constants associated with the error bounds. The premise of this article is that small constants and nearly-tight bounds are valuable, because quality measures based on precise error bounds are better able to choose among differently shaped elements of similar quality, or to trade off element size against element shape; and because tighter upper bounds make it possible to guarantee a fixed bound on error with fewer elements.

A notable entry in the "tight constants" camp by Berzins [10, 11] proposes anisotropic error indicators for triangles and tetrahedra, which estimate the interpolation error by approximating the true local solution by a (possibly anisotropic) quadratic function. The two main distinctions between the present work and Berzins' are that Berzin's indicators are approximations (not true upper and lower bounds), and they are for *a posteriori* use—they require an estimate of the Hessian H of the solution within each element, not just an upper bound. By contrast, the upper bounds given in Section 2 can be used as either *a priori* or *a posteriori* error estimates. Berzins' tetrahedral indicator is not easily converted to *a priori* use, because it is not clear how to determine what value of H will give the worst error for a given tetrahedron shape. Because Berzins' indicators are approximations, they accept tetrahedra that would usually be considered poor, like sliver tetrahedra, in circumstances where they happen to perform well. For instance, slivers interpolate the function $x^2 + y^2 + z^2$ and its gradients quite well. In this case, Berzins correctly estimates that the error is small, but the error bounds in Section 2.1 estimate that it is large. This difference suggests that Berzins' indicators may have advantages for tailoring a mesh to one specific solution, whereas the error bounds and quality measures given here are more appropriate for generating a mesh that will be used for time-dependent problems or to solve several elliptic problems with different solutions.

Apel's habilitation [2, 3] includes an excellent summary of results from the functional analysis camp, and extends them. A result that illustrates the generality of this work is that for any $m \geq 0$, $l > m$, $k \geq l-1$, $p, q \in [1, \infty]$, with $p > 2$ if $l = 1$ (where m, l, and k are integers and p and q are real numbers), a degree k polynomial Lagrangian interpolant g of f has error

$$\|f - g\|_{W^{m,q}} < CA^{1/q-1/p}\ell_{\max}^{l-m}|f|_{W^{l,p}},$$

where the $W^{m,q}$-norm $\|e\|_{W^{m,q}}$ of a function e is related to the L_q-norms of the derivatives of e of order up to m (see Apel for details), and the $W^{l,p}$-seminorm $|f|_{W^{m,q}}$ of a function f is related to the L_p-norms of the derivatives of f of order precisely m, each integrated over one element. If $m = 1$, $l = 2$, $k = 1$, and $p = q = \infty$, then this inequality implies bounds on both $\|f - g\|_\infty$ and $\|\nabla f - \nabla g\|_\infty$ proportional to the largest second derivative of f—much like the bounds in Section 2, except that nobody knows what the value of C is.

C is a constant (whose value is generally not derived) that depends on m, l, k, p, and q, but not on f. C also depends on the shape of the triangle, and was once believed to grow proportionally to $(\ell_{\max}/r_{\text{in}})^m$, but Synge [48], Babuška and Aziz [4], Jamet [28], Křížek [34], Apel [2, 3], and others have shown that C is bounded if the maximum angle of the triangle is bounded.

The inequality holds for tetrahedra as well, with the area A replaced by the volume V and a few extra restrictions on the choice of parameters. Again, Apel is the best and most general source for these results, for which he builds upon earlier work by Jamet [28] and Křížek [35]. Apel also offers anisotropic versions of these results.

The functional analysis results are general and impressive. They offer some guidance on how to choose element shapes to best interpolate a function f. However, this guidance is crude, and they do not offer guidance on when an element needs to be refined, or how to choose among the many competing quality measures in the literature. There is an entire industry of numerical analysts deriving error estimators for a wide variety of PDEs, but because of their asymptotic nature, they have had only a mild impact on mesh generation and mesh improvement algorithms. Mesh generators must make distinctions between elements with greater precision than functional analysis currently allows. These distinctions can be made reliably with the help of error bounds that are tight to within small constant factors.

3 Element Size, Element Shape, and Stiffness Matrix Conditioning

This section describes the mathematical relationship between the shapes and sizes of elements and the condition numbers of the stiffness matrices used in the finite element method. It assumes that the reader is familiar with the finite element method; no introduction is given here, but many good texts are available, including Johnson [30], Strang and Fix [47], and Becker, Carey, and Oden [7]. The main points are that small angles (but not large angles in the absence of small ones) can cause poor conditioning, that the relationship can be quantified in a way useful for comparing differently shaped elements, and that PDEs with anisotropic coefficients are best served by anisotropic elements.

The systems of linear equations constructed by finite element discretization are solved using either iterative methods or direct methods. The speed of iterative methods, such as the Jacobi and conjugate gradient methods, depends in part on the conditioning of the global stiffness matrix: a larger condition number implies slower performance. Direct solvers rarely vary as much in their running time, but the solutions they produce can be inaccurate due to roundoff error in floating-point computations, and the size of the roundoff error is roughly proportional to the condition number of the stiffness matrix. As a rule of thumb, Gaussian elimination loses one decimal digit of accuracy for every digit in the integer part of the condition number. These errors can be countered by using higher-precision floating-point numbers [5, 44].

For some applications that use direct solvers, the degree of accuracy required might be small enough, or the floating-point precision great enough, that a poorly conditioned stiffness matrix is not an impediment. Usually, though, conditioning should be kept under control. Sections 3.1 and 3.2 describe the influence of element size and shape on conditioning for isotropic and anisotropic PDEs, respectively.

Time-dependent PDEs are usually solved with the help of standard integration methods for computing approximate solutions of ordinary differential equations. The stability of explicit time integration methods is linked to the largest eigenvalue of a matrix related to the stiffness matrix. This eigenvalue is related to the sizes and shapes of the elements. The connection is discussed in Section 3.3.

3.1 Bounds on the Extreme Eigenvalues

The finite element method is different for every partial differential equation, and unfortunately, so is the relationship between element shape and matrix conditioning. As a concrete example, I will study the Poisson equation,

$$-\nabla^2 f(p) = \eta(p),$$

where $\eta(p)$ is a known function of p, and the goal is to find an approximation $h(p)$ of the unknown function $f(p)$ for some specified boundary conditions.

In the Galerkin formulation of the finite element method, the piecewise approximation h is composed from local piecewise basis functions, which are in turn composed from *shape functions*. Each shape function is defined on just one element. If h is piecewise linear, then the shape functions are our familiar friends the barycentric coordinates $\omega_i(p)$.

For each element t, the finite element method constructs a $(d+1) \times (d+1)$ *element stiffness matrix* K_t, where d is the dimension. The element stiffness matrices are *assembled* into an $n \times n$ *global stiffness matrix* K, where n is (for Poisson's equation on linear elements) the number of mesh vertices. The assembly process sums the entries of each element stiffness matrix into the entries of the global stiffness matrix. The difficulty of solving the linear system of equations associated with K typically grows with K's condition number $\kappa = \lambda^K_{\max}/\lambda^K_{\min}$, where $\lambda^K_{\max}$ and $\lambda^K_{\min}$ are the largest and smallest eigenvalues of K.

$\lambda^K_{\min}$ is closely tied to properties of the physical system being modeled, and to the sizes of the elements. Fried [23] offers a lower bound for $\lambda^K_{\min}$ that is proportional to the area or volume of the smallest element, and an upper bound proportional to the largest element. Fortunately, $\lambda^K_{\min}$ is not strongly influenced by element shape, but badly shaped elements often do have tiny areas or volumes.

In contrast, $\lambda^K_{\max}$ can be made arbitrarily large by a single badly shaped element. $\lambda^K_{\max}$ is related to the largest eigenvalues of the element stiffness matrices as follows. For each element t, let $\lambda^t_{\max}$ be the largest eigenvalue of its element stiffness matrix. Let m be the maximum number of elements meeting at a single vertex. Fried shows that

$$\max_t \lambda^t_{\max} \leq \lambda^K_{\max} \leq m \max_t \lambda^t_{\max},$$

so κ is roughly proportional to the largest eigenvalue among the element stiffness matrices.

Let's examine some element stiffness matrices and their eigenvalues. Recall that each $\nabla\omega_i$ is a constant vector orthogonal to face i of t, and its length is $1/a_i$, where a_i is the altitude of vertex i. The element stiffness matrix for a linear triangle is

$$\begin{aligned} K_t &= A \begin{bmatrix} \nabla\omega_1 \cdot \nabla\omega_1 & \nabla\omega_1 \cdot \nabla\omega_2 & \nabla\omega_1 \cdot \nabla\omega_3 \\ \nabla\omega_2 \cdot \nabla\omega_1 & \nabla\omega_2 \cdot \nabla\omega_2 & \nabla\omega_2 \cdot \nabla\omega_3 \\ \nabla\omega_3 \cdot \nabla\omega_1 & \nabla\omega_3 \cdot \nabla\omega_2 & \nabla\omega_3 \cdot \nabla\omega_3 \end{bmatrix} \\ &= \frac{1}{8A} \begin{bmatrix} 2\ell_1^2 & \ell_3^2 - \ell_1^2 - \ell_2^2 & \ell_2^2 - \ell_1^2 - \ell_3^2 \\ \ell_3^2 - \ell_1^2 - \ell_2^2 & 2\ell_2^2 & \ell_1^2 - \ell_2^2 - \ell_3^2 \\ \ell_2^2 - \ell_1^2 - \ell_3^2 & \ell_1^2 - \ell_2^2 - \ell_3^2 & 2\ell_3^2 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \cot\theta_2 + \cot\theta_3 & -\cot\theta_3 & -\cot\theta_2 \\ -\cot\theta_3 & \cot\theta_1 + \cot\theta_3 & -\cot\theta_1 \\ -\cot\theta_2 & -\cot\theta_1 & \cot\theta_1 + \cot\theta_2 \end{bmatrix}. \end{aligned}$$

The second form of this matrix follows from the first by Identities (6) and (9). The sum of the eigenvalues of K_t equals the sum of the diagonal entries, and both the eigenvalues and diagonal entries are nonnegative, so $\lambda^t_{\max}$ is large if and only if at least one of the diagonal entries is large. If one of the angles approaches $0°$, its cotangent approaches infinity, and so does $\lambda^t_{\max}$. Therefore, small angles can ruin matrix conditioning. Of course, if one of the angles approaches $180°$, the other two angles approach $0°$.

If the appearance of cotangents in the matrix seems mysterious, consider this more direct argument for why small angles are deleterious. An element with an angle near $0°$ has at least two altitudes that are very small (compared to the length of t's longest edge), and therefore two of the vectors $\nabla\omega_i$ are very long. Therefore, K_t has at least two large entries on its diagonal and at least one large eigenvalue.

The eigenvalues of K_t are the roots of its characteristic polynomial $p(\lambda)$. For triangles,

$$p(\lambda) = \lambda^3 - \frac{\ell_1^2 + \ell_2^2 + \ell_3^2}{4A}\lambda^2 + \frac{3}{4}\lambda.$$

The roots of this polynomial are $\lambda = 0$ and

$$\lambda = \frac{\ell_1^2 + \ell_2^2 + \ell_3^2 \pm \sqrt{(\ell_1^2 + \ell_2^2 + \ell_3^2)^2 - 48A^2}}{8A}.$$

The largest root $\lambda_{\max}^t$ is a scale-invariant indicator of the quality of the triangle's shape. (*Scale-invariant* means that if t is scaled uniformly without any change to its shape, $\lambda_{\max}^t$ does not change.) This eigenvalue is used as a quality measure in Section 6, and a contour plot of $1/\lambda_{\max}^t$ may be found there too. Note that the gradient of $\lambda_{\max}^t$ with respect to the vertex positions is singular for any equilateral triangle, which can be a nuisance when numerical optimization methods are used to move the vertices and improve the mesh quality. If a simpler or smoother indicator is desired, the radical can be dropped, but the simplified bound is only tight to within a factor of two.

$$\frac{\ell_1^2 + \ell_2^2 + \ell_3^2}{8A} \leq \lambda_{\max}^t \leq \frac{\ell_1^2 + \ell_2^2 + \ell_3^2}{4A}.$$

Suppose the mesh has no badly shaped triangles—for every triangle t, $\lambda_{\max}^t$ is bounded below some small constant. In this case, $\lambda_{\max}^K$ is also bounded below a small constant. Because the lower bound on the smallest global eigenvalue $\lambda_{\min}^K$ is proportional to the area $A_{\min}$ of the smallest triangle, $\kappa \in \mathcal{O}(1/A_{\min})$. If the triangles are of uniform size, $\kappa \propto 1/\ell^2$ where ℓ is the typical edge length. Since the area of the domain is fixed, $\kappa \propto n$ where n is the number of mesh vertices. (The number of vertices and elements is typically dictated by the need to limit the discretization error, and therefore the interpolation error.) Highly nonuniform meshes also have $\kappa \in \mathcal{O}(1/A_{\min})$, but do not generally share the good fortune that $\kappa \in \mathcal{O}(n)$. This serves as a reminder that local refinement of meshes to reduce interpolation and discretization errors can lead to other sorts of trouble.

Putti and Cordes [38] observe that the global stiffness matrix K is guaranteed to be an M-matrix if the following three conditions hold: the mesh is a Delaunay triangulation, no edge of the domain boundary separates a triangle from its circumcenter, and the global stiffness matrix K associated with Poisson's equation is not singular (for which it suffices to set a single Dirichlet boundary condition). The observation follows from inspection of the element stiffness matrix K_t. Suppose v_1v_2, the edge of t opposite θ_3, lies on the boundary of the mesh and is not shared with another element. The entry $-\cot\theta_3$ that represents the edge v_1v_2 in the global stiffness matrix K is nonpositive if and only if $\theta_3 \leq 90°$. This condition holds if and only if the circumcenter of t is not separated from t by v_1v_2. Alternatively, suppose the edge v_1v_2 is shared by another element t'. Let θ_3 and ϕ_3 be the angles opposite v_1v_2 in t and t' respectively. The entry that represents v_1v_2 in K is $-\cot\theta_3 - \cot\phi_3$, which is nonpositive if and only if $\theta_3 + \phi_3 \leq 180°$, which holds if and only if the edge v_1v_2 is locally Delaunay. The last condition is satisfied for every interior edge by the Delaunay triangulation. Unfortunately, the M-matrix guarantee does not hold in three dimensions.

If t is a linear tetrahedron, the element stiffness matrix is

$$K_t = V \begin{bmatrix} \nabla\omega_1 \cdot \nabla\omega_1 & \nabla\omega_1 \cdot \nabla\omega_2 & \nabla\omega_1 \cdot \nabla\omega_3 & \nabla\omega_1 \cdot \nabla\omega_4 \\ \nabla\omega_2 \cdot \nabla\omega_1 & \nabla\omega_2 \cdot \nabla\omega_2 & \nabla\omega_2 \cdot \nabla\omega_3 & \nabla\omega_2 \cdot \nabla\omega_4 \\ \nabla\omega_3 \cdot \nabla\omega_1 & \nabla\omega_3 \cdot \nabla\omega_2 & \nabla\omega_3 \cdot \nabla\omega_3 & \nabla\omega_3 \cdot \nabla\omega_4 \\ \nabla\omega_4 \cdot \nabla\omega_1 & \nabla\omega_4 \cdot \nabla\omega_2 & \nabla\omega_4 \cdot \nabla\omega_3 & \nabla\omega_4 \cdot \nabla\omega_4 \end{bmatrix}$$

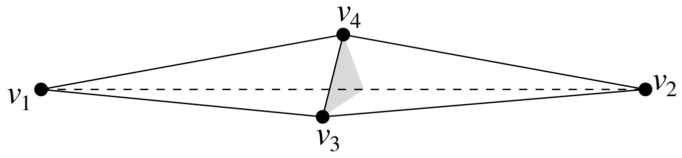


Figure 14: A tetrahedron with a large dihedral angle but no small dihedral angle. The shaded cross-section is an equilateral triangle. Every dihedral angle is at least 60°. As vertices v_3 and v_4 move toward the midpoint of edge v_1v_2, the dihedral angle θ_{34} approaches 180°, but the largest eigenvalue of the element stiffness matrix does not explode.

$$= \frac{1}{6}\begin{bmatrix} \sum_{1\neq i<j} \ell_{ij}\cot\theta_{ij} & -\ell_{34}\cot\theta_{34} & -\ell_{24}\cot\theta_{24} & -\ell_{23}\cot\theta_{23} \\ -\ell_{34}\cot\theta_{34} & \sum_{2\neq i<j\neq 2} \ell_{ij}\cot\theta_{ij} & -\ell_{14}\cot\theta_{14} & -\ell_{13}\cot\theta_{13} \\ -\ell_{24}\cot\theta_{24} & -\ell_{14}\cot\theta_{14} & \sum_{3\neq i<j\neq 3} \ell_{ij}\cot\theta_{ij} & -\ell_{12}\cot\theta_{12} \\ -\ell_{23}\cot\theta_{23} & -\ell_{13}\cot\theta_{13} & -\ell_{12}\cot\theta_{12} & \sum_{i<j\neq 4} \ell_{ij}\cot\theta_{ij} \end{bmatrix}.$$

If one of the dihedral angles approaches $0°$, its cotangent approaches infinity, and so does $\lambda^t_{\max}$. Unlike with triangles, it is possible for one dihedral angle of a tetrahedron to be arbitrarily close to $180°$ without any dihedral angle of the tetrahedron being small, as Figure 14 illustrates. Surprisingly, such a tetrahedron does not induce a large eigenvalue in K_t. If the length ℓ_{12} of the edge v_1v_2 is held fixed while v_3 and v_4 slide toward the center of v_1v_2, the angle θ_{34} approaches $180°$, but $\lambda^t_{\max}$ does not approach infinity—it approaches a constant proportional to ℓ_{12}. The reasoning is as follows. An angle approaching $0°$ has a cotangent approaching infinity, but an angle approaching $180°$ has a cotangent approaching negative infinity. Each entry on the diagonal of K_t is nonnegative and has the form $\sum_{i,j} \ell_{ij}\cot\theta_{ij}$. Therefore, if t has no dihedral angle near $0°$, the diagonal entries of K_t are bounded and thus so is $\lambda^t_{\max}$. It matters not that t has planar angles near $0°$.

Tedious manipulation reveals that the characteristic polynomial of K_t is

$$p(\lambda) = \lambda^4 - \frac{\sum_{i=1}^4 A_i^2}{9V}\lambda^3 + \frac{\sum_{1\leq j<k\leq 4}\ell_{jk}^2}{36}\lambda^2 - \frac{V}{9}\lambda.$$

There does not seem to be a simple expression for the roots of this polynomial (except the smallest root $\lambda = 0$), but they can be found numerically or by the cubic equation. However, these are expensive computations. Furthermore, the gradient of $\lambda^t_{\max}$ with respect to the vertex positions is singular for an equilateral tetrahedron. For these reasons, a simpler and smoother measure of the conditioning of K_t is useful.

An estimate of $\lambda^t_{\max}$ follows from the fact that K_t is known to be positive indefinite, so all its eigenvalues are nonnegative. The second coefficient of the characteristic polynomial is the (negated) sum of the eigenvalues, one of which is known to be zero, so

$$\frac{\sum_{i=1}^4 A_i^2}{27V} \leq \lambda^t_{\max} \leq \frac{\sum_{i=1}^4 A_i^2}{9V},$$

giving upper and lower bounds tight to within a factor of three.

$\lambda^t_{\max}$ and $\lambda^K_{\max}$ are not scale-invariant (as they are for triangles). If t is scaled uniformly, $\lambda^t_{\max}$ grows linearly with $\ell_{\max}$. Thus, the largest tetrahedron in a mesh may determine the largest eigenvalue of the global stiffness matrix, and the shapes of the largest tetrahedra are more important than the shapes of the smaller ones. However, this must not be misinterpreted to imply that refining tetrahedra is always a good way to improve the condition number, because $\lambda^K_{\min}$ is proportional to the volumes of the tetrahedra. A better recommendation is to use tetrahedra that have good shapes and are as uniform as they can be without compromising speed or interpolation accuracy. To judge tetrahedron shapes, Section 6 discusses how to define scale-invariant quality measures related to $\lambda^t_{\max}$.

If the mesh has no badly shaped tetrahedra, the largest global eigenvalue $\lambda^K_{\max}$ is proportional to the length $\ell_{\max}$ of the longest edge in the entire mesh. The lower bound on $\lambda^K_{\min}$ is proportional to the volume $V_{\min}$ of the smallest tetrahedron, so $\kappa \in \mathcal{O}(\ell_{\max}/V_{\min})$. If the tetrahedra are of uniform size, $\kappa \propto 1/\ell^2$, just like in the two-dimensional case. Hence, $\kappa \propto n^{2/3}$. However, nonuniform meshes and meshes with poorly shaped tetrahedra can have much worse conditioning than $\mathcal{O}(n^{2/3})$.

3.2 Anisotropy and Conditioning

As with interpolation, there are applications in which anisotropic elements are better suited for obtaining a well conditioned stiffness matrix than isotropic ones. Unlike with interpolation, it is not the form of the interpolated function that determines how thin elements should be, or how they should be oriented. Rather, it is the anisotropy in the partial differential equation itself.

For example, consider the anisotropic variant of the Poisson equation,

$$-\nabla \cdot B\nabla f(p) = \eta(p), \tag{19}$$

where B is a known symmetric positive definite $d \times d$ matrix of coefficients, and the goal is to approximate the unknown function $f(p)$ (given some specified boundary conditions). In the two-dimensional case, this equation expands to

$$-\left(B_{11}\frac{\partial^2}{\partial x^2} + 2B_{12}\frac{\partial^2}{\partial x \partial y} + B_{22}\frac{\partial^2}{\partial y^2}\right) f(p) = \eta(p).$$

The matrix B can be understood and constructed in the same way as the matrix C_t in Section 2.2. However, as we shall see, B has a stronger kinship to C_t^{-1} than to C_t. For example, the eigenvector for which B has the largest eigenvalue is likely to be the direction in which the solution $f(p)$ has the least curvature (although $\eta(p)$ has an influence as well).

Let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ denote the unit eigenvectors of B, and let $\xi_1, \ldots, \xi_d$ be the corresponding eigenvalues. Thus,

$$B = \xi_1 \mathbf{v}_1 \mathbf{v}_1^T + \xi_2 \mathbf{v}_2 \mathbf{v}_2^T + \xi_3 \mathbf{v}_3 \mathbf{v}_3^T.$$

As in Section 2.2, a transformation matrix F maps points in physical space to isotropic space. Here, however, "isotropic space" is the space in which the coefficients of the partial differential equation are isotropic, and not necessarily the solution. The ideal element is equilateral in isotropic space. Let

$$F = \frac{1}{\sqrt{\xi_1}}\mathbf{v}_1 \mathbf{v}_1^T + \frac{1}{\sqrt{\xi_2}}\mathbf{v}_2 \mathbf{v}_2^T + \frac{1}{\sqrt{\xi_3}}\mathbf{v}_3 \mathbf{v}_3^T.$$

Because the $\mathbf{v}_i$ vectors are orthonormal,

$$F^2 = B^{-1},$$

so F is essentially an inverse square root of B. Let $\widetilde{p} = Fp$ denote the image of a point p in isotropic space, and let $\widetilde{t}$ denote the image of an element t. Define $\widetilde{f}(q) = f(F^{-1}q)$, so that $\widetilde{f}(\widetilde{p}) \equiv f(p)$. Let ∇_p denote the gradient taken with respect to a point p. By a straightforward change of coordinates, the partial differential equation (19) in isotropic space is

$$-\nabla_q^2 \widetilde{f}(q) = \widetilde{\eta}(q).$$

Returning to the anisotropic formulation of the Poisson equation, the element stiffness matrix for a triangle t is

$$K_t = A \begin{bmatrix} (F^{-1}\nabla\omega_1)\cdot(F^{-1}\nabla\omega_1) & (F^{-1}\nabla\omega_1)\cdot(F^{-1}\nabla\omega_2) & (F^{-1}\nabla\omega_1)\cdot(F^{-1}\nabla\omega_3) \\ (F^{-1}\nabla\omega_2)\cdot(F^{-1}\nabla\omega_1) & (F^{-1}\nabla\omega_2)\cdot(F^{-1}\nabla\omega_2) & (F^{-1}\nabla\omega_2)\cdot(F^{-1}\nabla\omega_3) \\ (F^{-1}\nabla\omega_3)\cdot(F^{-1}\nabla\omega_1) & (F^{-1}\nabla\omega_3)\cdot(F^{-1}\nabla\omega_2) & (F^{-1}\nabla\omega_3)\cdot(F^{-1}\nabla\omega_3) \end{bmatrix},$$

and the stiffness matrix for a tetrahedron follows by analogy. By transforming the element to isotropic space, one can express the element stiffness matrix in its isotropic form and use the bounds on $\lambda_{\max}^t$ established in Section 3.1.

Let $\widetilde{V}$ denote the volume of $\widetilde{t}$. (The following arguments can be adapted to two-dimensional cases by replacing volume with area without any other change.) The mapping F scales the volume of t (or any volume) so that

$$\widetilde{V} = |F|V,$$

where $|F| = (\prod_{i=1}^{d} \xi_i)^{-1/2}$ is the determinant of F.

Let's examine the relationship between the shape functions $F^{-1}\nabla\omega_i$ of t and the shape functions of $\widetilde{t}$. Let p be a point in t. Its barycentric coordinates can be found by the formula

$$\omega_i(p) = \frac{V_i(p)}{V},$$

where V is the volume of t and $V_i(p)$ is the volume of the tetrahedron formed by replacing vertex v_i of t with p. Likewise, let $\widetilde{V}_i(\widetilde{p})$ be the volume of the tetrahedron formed by replacing vertex $\widetilde{v}_i$ of $\widetilde{t}$ with $\widetilde{p}$. Then

$$\begin{aligned} \nabla_p\,\omega_i(p) &= \frac{1}{V}\nabla_p V_i(p) \\ &= \frac{1}{|F|V}\nabla_p \widetilde{V}_i(Fp) \\ &= \frac{1}{\widetilde{V}} F\nabla_{\widetilde{p}}\widetilde{V}_i(\widetilde{p}), \text{ and therefore} \\ F^{-1}\nabla_p\,\omega_i(p) &= \nabla_{\widetilde{p}}\,\widetilde{\omega}_i(\widetilde{p}). \end{aligned}$$

Therefore, K_t is identical to the element stiffness matrix for $\widetilde{t}$ under the isotropic Poisson equation, and the characteristic polynomial and eigenvalues given in Section 3.1 apply directly to $\widetilde{t}$. This means that long, thin, properly oriented elements in physical space may be optimal for minimizing $\lambda_{\max}^K$ and the condition number κ. It also means that equilateral elements may sharply increase $\lambda_{\max}^K$ for anisotropic PDEs.

3.3 Eigenvalues and Explicit Time Integration

When time-dependent PDEs are solved with finite element methods, and an explicit time integration method is used, the stability of the time integration method depends critically on the largest eigenvalue of an eigensystem related to the time integration method. This section summarizes the connection. See Chapter 5 of Carey and Oden [14] for a more detailed treatment.

Time-dependent PDEs are usually treated by the use of a finite element method to discretize the spatial dimensions only, yielding a system of time-dependent coupled ordinary differential equations (ODEs). The solution to this system of equations is approximated by means of any of a variety of standard numerical methods for integrating ODEs. Numerical time integration methods discretize time by dividing it into discrete *time steps*. Most time integration methods can be categorized as *implicit* or *explicit*. When implicit methods (such as centered differencing or backward differencing) are chosen, it is usually because they have the advantage of *unconditional stability*: numerical errors, which are introduced by the approximation or by floating-point roundoff error, are not amplified as time passes. By contrast, explicit methods (such as forward differencing with lumping) are stable only if each time step is sufficiently small. The largest permissible time step is related to the stiffness matrix.

To give an example of a parabolic PDE, consider a diffusion equation of the form

$$\frac{\partial f(p,t)}{\partial t} - \nabla \cdot B \nabla f(p,t) + \zeta(p) f(p,t) = \eta(p,t).$$

The goal is to approximate the unknown function $f(p,t)$, given suitable boundary conditions and initial condition $f(p,0)$. Time integration by forward differencing, combined with spatial discretization by the finite element method, yields an approximate discrete solution whose time steps are computed by the recurrence

$$M\mathbf{f}_{i+1} = (M - \Delta t\, K)\mathbf{f}_i + \Delta t\, \mathbf{n}_i,$$

where K is the stiffness matrix, M is called the *mass matrix*, Δt is the length of each time step, $\mathbf{f}_i$ is an n-entry vector that gives the approximate value of $f(v, i\Delta t)$ for each vertex v of the n-vertex mesh at time $i\Delta t$, and $\mathbf{n}_i$ is a vector related to the function $\eta(p, i\Delta t)$. The initial condition $f(p,0)$ dictates the value of $\mathbf{f}_0$, and the recurrence determines the values of $\mathbf{f}_i$, $i = 1, 2, 3, \ldots$.

To analyze the stability of the recurrence, suppose some vector $\mathbf{f}_k$ is replaced in the computation by $\mathbf{f}_k^* = \mathbf{f}_k + \mathbf{e}_k$, where $\mathbf{e}_k$ is an error term. The time integration algorithm henceforth computes the recurrence $\mathbf{f}_{i+1}^* = (I - \Delta t\, M^{-1}K)\mathbf{f}_i^* + \Delta t\, M^{-1}\mathbf{n}_i$. By subtracting the original recurrence from this one, one obtains $\mathbf{e}_{i+1} = (I - \Delta t\, M^{-1}K)\mathbf{e}_i$, so

$$\mathbf{e}_j = (I - \Delta t\, M^{-1}K)^{j-i}\mathbf{e}_i.$$

In practice, errors are introduced during every time step (not just during step i) by the discrete integration method, floating-point roundoff error, and perhaps other causes. However, an analysis of an error introduced during just one time step leads to the same conclusions about stability. The magnitude of any such error can either attenuate or increase with time, depending on the spectrum of $I - \Delta t\, M^{-1}K$.

If the eigenvalues of $M^{-1}K$ are between zero and $2/\Delta t$, then the eigenvalues of $I - \Delta t\, M^{-1}K$ are between -1 and 1, and the error $\mathbf{e}_j$ converges to zero as j increases. If $M^{-1}K$ has an eigenvalue larger than $2/\Delta t$, the error (usually) diverges. Hence, forward differencing is stable if

$$\Delta t < \frac{2}{\lambda_{\max}(M^{-1}K)}.$$

The larger the maximum eigenvalue of $M^{-1}K$, the smaller the time steps must be; and the smaller the time steps, the longer the computation.

What effect do linear elements have on the eigenvalues of $M^{-1}K$? The answer requires a little understanding of the mass matrix M. The recurrence requires solving an equation of the form $M\mathbf{f}_{i+1} = \mathbf{g}_i$ for $\mathbf{f}_{i+1}$. The main advantage of forward differencing is that M can be made diagonal by an approximation technique called *lumping*, so that $M^{-1}\mathbf{g}_i$ can be computed quickly. This time integration method is said to be explicit because there is no need to solve a full-fledged system of linear equations. In the following, assume that M is lumped (and therefore diagonal), though the conclusion does not differ much if M is not lumped.

Each diagonal entry m_i of M represents the mass associated with some vertex i of the mesh. Typically, m_i is proportional to the sum of the areas or volumes of the elements that adjoin vertex i. Therefore, every m_i is positive. The eigenvalues of $M^{-1}K$ are the same as the eigenvalues of the symmetric positive definite matrix $M^{-1/2}KM^{-1/2}$, where $M^{-1/2}$ is the diagonal matrix whose diagonal entries are $m_i^{-1/2}$; hence, the eigenvalues are real and positive.

Consider a linear triangle t with element stiffness matrix K_t. Let j, k, and l be the indices of the vertices of t in the mesh, and let

$$J_t = \begin{bmatrix} m_j^{-1/2} & 0 & 0 \\ 0 & m_k^{-1/2} & 0 \\ 0 & 0 & m_l^{-1/2} \end{bmatrix} K_t \begin{bmatrix} m_j^{-1/2} & 0 & 0 \\ 0 & m_k^{-1/2} & 0 \\ 0 & 0 & m_l^{-1/2} \end{bmatrix}.$$

For a linear tetrahedron, add one more row and column to each matrix in the obvious way. The matrix $M^{-1/2}KM^{-1/2}$ can be assembled from the matrices J_t (where t varies over every element in the mesh) in the same way the global stiffness matrix K is assembled from the element stiffness matrices K_t. By analogy to Fried's result reported in Section 3.1,

$$\max_t \lambda_{\max}(J_t) \leq \lambda_{\max}(M^{-1}K) \leq m \max_t \lambda_{\max}(J_t),$$

where m is the maximum number of elements meeting at a single vertex. Therefore, the maximum eigenvalue of $M^{-1}K$ is roughly proportional to the maximum eigenvalue of J_t for the worst element t.

For a typical element t, the values of m_j, m_k, and m_l differ by only a small constant (whether t is poorly shaped or not). Therefore, the maximum eigenvalue of J_t grows without bound as some angle of t (some dihedral angle if t is a tetrahedron) becomes arbitrarily close to $0°$, just like the maximum eigenvalue of K_t. However, the size of an element has a very different influence on J_t than on K_t. Recall that in two dimensions, $\lambda_{\max}(K_t)$ is independent of element size (for a fixed element shape), and in three dimensions, $\lambda_{\max}(K_t)$ is directly proportional to the longest edge $\ell_{\max}$ of t. But m_j, m_k, etc. are proportional to the areas or volumes of the local elements. For well shaped elements of any dimensionality, $\lambda_{\max}(J_t)$ is directly proportional to $\ell_{\max}^{-2}$.

These observations lead to these conclusions. For a uniform mesh of well shaped elements, the largest permissible time step Δt is roughly proportional to ℓ^2, where ℓ is the typical edge length. Hence, smaller elements require smaller time steps and more computation. In a nonuniform mesh, the smallest elements usually dictate the size of the time step. Elements with small angles can force the time integrator to take even smaller time steps, and the shapes of the smallest elements of a mesh are more important than the shapes of the largest elements. (By contrast, recall that badly shaped tetrahedra are more likely to ruin the conditioning of K if they are among the largest tetrahedra in the mesh.)

The forgoing comments apply to forward differencing on parabolic PDEs. By comparison, centered differencing and backward differencing are unconditionally stable (for any choice of Δt, whatever the eigenvalues of $M^{-1}K$ may be). However, even with centered differencing, if $M^{-1}K$ has a very large eigenvalue, some components of the error attenuate very slowly. Backward differencing does not share this problem, but (like forward differencing) it is less accurate than centered differencing. (See Carey and Oden for details.) Centered and backward differencing are called implicit methods because they require an expensive step that solves a system of linear equations, which cannot be avoided by lumping the mass matrix. Therefore, forward differencing will remain popular despite the constraint it places on the size of the time step. For both forward and centered differencing, it is well worth the effort to avoid poorly shaped elements and control $\lambda_{\max}(M^{-1}K)$.

Now consider a hyperbolic PDE, namely a wave equation of the form

$$\frac{\partial^2 f(p,t)}{\partial t^2} - \nabla \cdot B\nabla f(p,t) + \zeta(p) f(p,t) = \eta(p,t).$$

Time integration by centered differencing, combined with spatial discretization by the finite element method, yields the recurrence

$$M\mathbf{f}_{i+1} = (2M - (\Delta t)^2 K)\mathbf{f}_i - M\mathbf{f}_{i-1} + (\Delta t)^2 \mathbf{n}_i.$$

Observe that this recurrence uses the values of $\mathbf{f}$ from the previous two time steps to compute the current value. If M is lumped, the recurrence is fast to compute, but is only conditionally stable. This recurrence is stable if

$$\Delta t < \frac{2}{\sqrt{\lambda_{\max}(M^{-1}K)}}.$$

Therefore, the same conclusions apply to the hyperbolic case as to the parabolic case, except that the largest permissible time step Δt is roughly proportional to ℓ, where ℓ is the typical edge length of the smallest elements in the mesh. As with the diffusion equation, the stability constraint can be eliminated by choosing an unconditionally stable time integrator (e.g. a different version of centered differencing), at the cost of an expensive linear system solution computation for each time step. Again, see Carey and Oden for details.

4 Discretization Error

The finite element method attempts to find a piecewise approximation $h(p)$ to the unknown solution $f(p)$ of a partial differential equation. The *discretization error* is $f - h$, perhaps measured in some norm. (There are as many discretization errors as there are norms.) It is named thus because it is the consequence of replacing a continuous differential equation with a discrete system of equations. Ideally, as the mesh is refined and the size of the largest element approaches zero, the approximation h should converge to f, and thus the discretization error should converge to zero. However, this convergence only occurs if the interpolation errors $f - g$ and $\nabla f - \nabla g$ both approach zero, and $\nabla f - \nabla g$ only approaches zero if the largest angle is bounded away from 180° (though the bound might be very close to 180°). If the largest angle is allowed to grow arbitrarily close to 180° as the mesh grows finer, convergence might not occur.

Although discretization errors are closely linked to interpolation errors, the nature of the connection depends on the choice of PDE. The relationship is explored in Section 4.1, and the subtle ways that PDEs with anisotropic coefficients change the relationship are discussed in Section 4.2. In some unfortunate cases, the ideal elements for controlling interpolation error, discretization error, and matrix conditioning might be three different elements, each with a different aspect ratio or orientation. This possibility is discussed in Section 4.3.

4.1 Discretization Error and Isotropic PDEs

Although discretization error is linked to interpolation error, the error bounds for interpolation do not apply directly to the discretization error. They are linked through the partial differential equation whose solution is being approximated. Although the finite element solution h is a piecewise approximation to f, h is rarely equal to the naïve linear interpolant g discussed in Section 2.1, because in general $f(v) \neq h(v)$ for most mesh vertices v. The finite element method sometimes finds an approximate solution h that approximates f *more* accurately than g.

In some formulations (i.e. Rayleigh-Ritz formulations of self-adjoint PDEs), the finite element method finds the approximation h that minimizes $\|f - h\|$ measured over the entire mesh in some *energy norm* related to the PDE. For example, consider the equation $-\nabla^2 f(p) + f(p) = \eta(p)$ (where the function $\eta(p)$ is known). The finite element method finds the piecewise linear approximation $h(p)$ that minimizes

$$\|f - h\|_{H^1(\Omega)} = \left(\int_\Omega \left((f-h)^2 + |\nabla f - \nabla h|^2 \right) d\Omega \right)^{1/2},$$

where the domain Ω is covered by a mesh T. Because h is optimal, the discretization error can be bounded by taking advantage of bounds on the interpolation error.

$$\begin{aligned} \|f - h\|_{H^1(\Omega)} &\leq \|f - g\|_{H^1(\Omega)} \\ &\leq \left(\sum_{t \in T} V_t \left(\|f - g\|^2_{\infty(t)} + \|\nabla f - \nabla g\|^2_{\infty(t)} \right) \right)^{1/2}, \end{aligned}$$

where V_t is the volume of element t. Substitution of the bounds on $\|f - g\|_\infty$ and $\|\nabla f - \nabla g\|_\infty$ derived in Section 2.1 or 2.2 into this formula gives a numerical bound on the discretization error associated with the mesh T. Other second-order self-adjoint elliptic PDEs can be approximated with similar error bounds. Therefore, the discretization error of the finite element solution of such a PDE can be bounded globally.

However, there is no guarantee that $h(p)$ or $\nabla h(p)$ is accurate at any particular point. Local or pointwise bounds on discretization error are difficult to show. Whereas interpolation errors can be localized to individual elements, discretization errors usually cannot. In practice, a bad element is often responsible for large discretization errors in neighboring elements.

The relative importance of $\|f - g\|$ versus $\|\nabla f - \nabla g\|$ to the discretization error depends on the coefficients of the PDE and the sizes of the elements. A mesh generator should take these numbers into account before deciding whether to try to exploit superaccurate elements. Because $\|f - g\|$ decreases proportionally to $\ell^2_{\max}$, but $\|\nabla f - \nabla g\|$ only decreases proportionally to $\ell_{\max}$, the latter measure (and the shape of the elements) becomes increasingly important as the mesh is refined, making superaccuracy increasingly attractive.

Although the discretization error $\|f - h\|$ (measured in the energy norm) is generally smaller than the interpolation error $\|f - g\|$, they usually differ by only a small constant factor, so large angles and large elements can be counted on to worsen the discretization error.

Of course, the norm in which the discretization error is minimized is not usually the norm that a user cares about most. Nevertheless, it is in an analyst's best interests to choose the elements that best control the discretization error in the PDE's energy norm, because the approximation h will not be more accurate in any norm than it is in the energy norm. This means that the PDE itself can affect the choice of elements, as we will see next.

4.2 Discretization Error and Anisotropic PDEs

When a partial differential equation has anisotropic coefficients, the discretization error is less tightly related to the error in the gradient, and more tightly related to the error in the skewed gradient $F^{-1}\nabla g$ (with F

defined as in Section 3.2). Therefore, the shape of an element should be chosen not for its ability to control $\|\nabla f - \nabla g\|_\infty$, but rather for its ability to control $\|F^{-1}(\nabla f - \nabla g)\|_\infty$. The ideal aspect ratio and orientation of a superaccurate element for $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ may be different from that of a superaccurate element for $\|\nabla f - \nabla g\|_\infty$.

Recall from Section 3.2 the anisotropic Poisson equation $-\nabla \cdot B\nabla f(p) = \eta(p)$, and recall that $F^2 = B^{-1}$. The finite element method finds the piecewise linear approximation $h(p)$ that minimizes the H^1-seminorm of the error in the skewed gradient,

$$|F^{-1}(f-h)|_{H^1(\Omega)} = \left(\int_\Omega |F^{-1}(\nabla f - \nabla h)|^2 \, d\Omega\right)^{1/2}.$$

Therefore, the discretization error (measured in this seminorm) is bounded.

$$\begin{aligned} |F^{-1}(f-h)|_{H^1(\Omega)} &\leq |F^{-1}(f-g)|_{H^1(\Omega)} \\ &\leq \left(\sum_{t \in T} V_t \|F^{-1}(\nabla f - \nabla g)\|^2_{\infty(t)}\right)^{1/2}. \end{aligned}$$

Reducing the error in the skewed gradient, $\|F^{-1}(\nabla f - \nabla g)\|$, controls the discretization error more effectively than reducing the error in the gradient, $\|\nabla f - \nabla g\|$. If F is not an isotropic transformation, the ideal element shape for minimizing one is not ideal for the other.

Fortunately, error bounds for $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ are easily derived. Say that the element $\widetilde{t}$ produced by the transformation F lives in "PDE isotropic space," and the element $\widehat{t}$ produced by the transformation E lives in "solution isotropic space." The error $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ is equal to the gradient error $\|\nabla \widetilde{f} - \nabla \widetilde{g}\|_\infty$ over $\widetilde{t}$ in PDE isotropic space, where $\widetilde{f}(q) = f(F^{-1}q)$ and $\widetilde{g}(q) = g(F^{-1}q)$, so one may (more or less) look up the error bounds in Table 3. However, the bounds must be interpreted carefully: they apply to the element $\widetilde{t}$ and function $\widetilde{f}$ (not to t and f), and the curvature constraint on $\widetilde{f}$ differs from the curvature constraint on f.

Suppose the Hessian H of f satisfies the curvature constraint (3) at all points in an element t. Let $\widetilde{H}(\widetilde{p})$ be the Hessian of $\widetilde{f}(\widetilde{p})$, where the derivatives are taken with respect to $\widetilde{p}$. Then $\widetilde{H}(\widetilde{p}) = F^{-1}H(p)F^{-1}$, so $\widetilde{H}$ satisfies the constraint

$$|\mathbf{d}^T \widetilde{H}(\widetilde{p})\mathbf{d}| \leq \mathbf{d}^T F^{-1} C_t F^{-1} \mathbf{d} \tag{20}$$

at all points $\widetilde{p}$ in $\widetilde{t}$.

Inequality (20) has the same form as Inequality (3), so the bounds on $\|\nabla \widetilde{f} - \nabla \widetilde{g}\|_\infty$ in Table 3 hold on the element $\widetilde{t}$, with C_t replaced by $F^{-1}C_tF^{-1}$. Let $\widetilde{c_t}$ be the largest eigenvalue of $F^{-1}C_tF^{-1}$. Recasting the bounds in terms of the element t yields

$$\begin{aligned} \|F^{-1}(\nabla f - \nabla g)\|_\infty &\leq \frac{\frac{\widetilde{c_t}}{4\widetilde{A}} \sum_{1 \leq i < j \leq 3} \widetilde{\ell_i}\widetilde{\ell_j}\widehat{\ell_k}^2 + \sqrt{c_t \widetilde{c_t}} \max_{1 \leq i < j \leq 3} \left(\widetilde{\ell_i}\widehat{\ell_j} + \widetilde{\ell_j}\widehat{\ell_i}\right)}{\widetilde{\ell_1} + \widetilde{\ell_2} + \widetilde{\ell_3}} \text{ for triangles;} \\ \|F^{-1}(\nabla f - \nabla g)\|_\infty &\leq \frac{\frac{\widetilde{c_t}}{6\widetilde{V}} \sum_{1 \leq i < j \leq 4} \widetilde{A_i}\widetilde{A_j}\widehat{\ell_{ij}}^2 + \sqrt{c_t \widetilde{c_t}} \max_i \sum_{j \neq i} \widetilde{A_j}\widehat{\ell_{ij}}}{\sum_{m=1}^4 \widetilde{A_m}} \text{ for tetrahedra.} \end{aligned}$$

(The mysterious term $\sqrt{c_t \widetilde{c_t}}$ arises because the inequality (17) in the original derivation is replaced by $|\widetilde{H}(u)\widetilde{d}| \leq \sqrt{c_t \widetilde{c_t}}|Ed|$.)

Recall that h is only as accurate in any norm as it is in the error norm, so the ideal finite element mesh for this PDE and similar PDEs should be tailored to control these error bounds (in preference to the bounds on the error in the gradients). By analogy to the bounds for $\|\nabla f - \nabla g\|_\infty$, an element t offers a good tradeoff between size and the error $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ if $\widehat{t}$ (in solution isotropic space) is nearly equilateral. The differences between the elements preferred by these bounds for $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ and the bounds in Table 3 for $\|\nabla f - \nabla g\|_\infty$ are subtle, and these subtleties are best expressed by using the error bounds in meshing software.

However, in the realm of superaccuracy, the ideal superaccurate element for controlling $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ is often quite different from the ideal superaccurate element for controlling $\|\nabla f - \nabla g\|_\infty$. Elements can be superaccurate for $\|F^{-1}(\nabla f - \nabla g)\|_\infty$ if $\widetilde{t} \neq \widehat{t}$ (that is, $E \neq F$).

Suppose that H is constant over an element t. Express $F^{-1}HF^{-1}$ in term of its eigenvectors and eigenvalues in the diagonalization

$$F^{-1}HF^{-1} = \left[\begin{array}{ccc} \widetilde{\mathbf{v}_1} & \widetilde{\mathbf{v}_2} & \widetilde{\mathbf{v}_3} \end{array}\right] \left[\begin{array}{ccc} \pm\widetilde{\xi_1} & 0 & 0 \\ 0 & \pm\widetilde{\xi_2} & 0 \\ 0 & 0 & \pm\widetilde{\xi_3} \end{array}\right] \left[\begin{array}{ccc} \widetilde{\mathbf{v}_1} & \widetilde{\mathbf{v}_2} & \widetilde{\mathbf{v}_3} \end{array}\right]^T,$$

and let C_t be the symmetric positive definite matrix that satisfies

$$F^{-1}C_tF^{-1} = \left[\begin{array}{ccc} \widetilde{\mathbf{v}_1} & \widetilde{\mathbf{v}_2} & \widetilde{\mathbf{v}_3} \end{array}\right] \left[\begin{array}{ccc} \widetilde{\xi_1} & 0 & 0 \\ 0 & \widetilde{\xi_2} & 0 \\ 0 & 0 & \widetilde{\xi_3} \end{array}\right] \left[\begin{array}{ccc} \widetilde{\mathbf{v}_1} & \widetilde{\mathbf{v}_2} & \widetilde{\mathbf{v}_3} \end{array}\right]^T.$$

Recall that E, $\widehat{t}$, and "solution isotropic space" are defined in terms of C_t. With this choice of C_t, the error bounds improve to

$$\begin{aligned}
\|F^{-1}(\nabla f - \nabla g)\|_\infty &\leq \frac{\frac{\widetilde{c_t}}{4\widetilde{A}} \sum_{1 \leq i < j \leq 3} \widetilde{\ell_i}\widetilde{\ell_j}\widehat{\ell_k}^2 + c_t \max_{1 \leq i < j \leq 3}\left(\widetilde{\ell_i}\overline{\ell_j} + \widetilde{\ell_j}\overline{\ell_i}\right)}{\widetilde{\ell_1} + \widetilde{\ell_2} + \widetilde{\ell_3}} \text{ for triangles;} \\
\|F^{-1}(\nabla f - \nabla g)\|_\infty &\leq \frac{\frac{\widetilde{c_t}}{6\widetilde{V}} \sum_{1 \leq i < j \leq 4} \widetilde{A_i}\widetilde{A_j}\widehat{\ell_{ij}}^2 + c_t \max_i \sum_{j \neq i} \widetilde{A_j}\overline{\ell_{ij}}}{\sum_{m=1}^{4} \widetilde{A_m}} \text{ for tetrahedra,}
\end{aligned}$$

where $\overline{\ell_i} = |F^{-1}E^2(v_j - v_k)|$ and $\overline{\ell_{ij}} = |F^{-1}E^2(v_i - v_j)|$ are edge lengths of the element $\overline{t}$ found by applying the transformation $F^{-1}E^2$ to t.

What is the ideal superaccurate element for $\|F^{-1}(\nabla f - \nabla g)\|_\infty$? An element t such that $\overline{t}$ is roughly equilateral. From the point of view of PDE isotropic space, $\widetilde{t}$ is aligned with the eigenvectors of $F^{-1}C_tF^{-1}$ and has an aspect ratio equal to the condition number of $F^{-1}C_tF^{-1}$. This is the same alignment, but the square of the aspect ratio, as a safe element that is roughly equilateral in solution isotropic space.

4.3 Do the Demands for Anisotropy Agree With Each Other?

The solution isotropic space in which equilateral elements are good for interpolation and the PDE isotropic space in which equilateral elements are good for matrix conditioning often coincide. For instance, the function $f(p) = x^2 - 64y^2$ is a solution to the equation $-64\partial^2/\partial x^2 - \partial^2/\partial y^2 = 0$. In this case, the mappings E (determined by f) and F (determined by the differential equation) agree, and so does the mapping $F^{-1}E^2$ associated with the discretization error of this PDE.

$$E = F = F^{-1}E^2 = \begin{bmatrix} 1/8 & 0 \\ 0 & 1 \end{bmatrix}.$$

Hence, thin triangles with an aspect ratio of roughly eight, oriented parallel to the x-axis, ideally suit the needs of interpolation accuracy, stiffness matrix conditioning, and discretization accuracy. (Longer, thinner elements could yield superaccurate interpolation of the gradients—if f were a known function—but not the skewed gradients.)

However, a strong choice of the force function $\eta(p)$, combined with (un)suitable boundary conditions, may yield a different solution and change the shape of the ideal elements for interpolation, while leaving the stiffness matrix unchanged. For instance, the function $f(p) = 64x^2 + y^2$ is a solution to the equation $-64\partial^2/\partial x^2 - \partial^2/\partial y^2 = -8194$. In this case,

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 1/8 \end{bmatrix}, F = \begin{bmatrix} 1/8 & 0 \\ 0 & 1 \end{bmatrix}, F^{-1}E^2 = \begin{bmatrix} 8 & 0 \\ 0 & 1/64 \end{bmatrix}.$$

For this equation, the ideal triangles for matrix conditioning are still oriented parallel to the x-axis, but $f(p)$ is best interpolated by triangles oriented parallel to the y-axis, with aspect ratios of roughly 8. The discretization error, governed by $\|F^{-1}(\nabla f - \nabla g)\|$, is best controlled by superaccurate triangles that are oriented parallel to the y-axis, but have aspect ratios around 512. The gradient error $\|\nabla f - \nabla g\|_\infty$ can also benefit from superaccuracy, but its ideal triangles have aspect ratios around 64.

This example requires somewhat contrived boundary conditions to make this solution possible. Mismatches between PDE isotropic space and solution isotropic space occur more easily in three dimensions than in two. The Poisson equation $\nabla^2 f = 0$ in two dimensions is a declaration that the curvature along one principle axis is the negation of the curvature along the other axis. The anisotropic Poisson equation declares that $E \propto F$. In three dimensions, however, Poisson's equation declares only that the three principle curvatures sum to zero. Three-dimensional isotropic PDEs easily yield anisotropic solution curvature, and E and F are less strongly coupled in three-dimensional anisotropic PDEs.

When a compromise must be made between interpolation accuracy, matrix conditioning, and discretization error, users should choose the nature of the compromise based on which consideration is most troubling under the circumstances. One point deserves emphasis. There is an essential difference between anisotropy for interpolation and anisotropy for conditioning: in a setting where anisotropic elements are ideal, it is possible to compensate for the effects of badly shaped elements on the interpolation and discretization errors by making them smaller. It is not possible to compensate for their effects on matrix conditioning.

5 One Bad Element

How tolerant can a triangulation be? Can one bad element spoil the lot?

For interpolation, the effects of a poorly shaped triangle or tetrahedron are entirely local. The large pointwise errors do not extend beyond the bad element. For most applications other than numerical methods for solving partial differential equations, the story ends here.

In those numerical methods, however, discretization error behaves differently than interpolation error. As Section 4 explains, the finite element method produces a solution h for which the error $f - h$ is bounded in a global manner, but not in a pointwise manner. The influence of one bad element is usually felt at locations near the element. In the extreme case where $\|\nabla f - \nabla h\|$ is very large (because some angle is extremely close to $180°$), h may be strongly in error over the entire mesh.

Stiffness matrix conditioning has a story to tell too. As we have seen, one element stiffness matrix with a large maximum eigenvalue raises the maximum eigenvalue of the global stiffness matrix K just as high. However, a mesh with only one or two bad elements will typically impose only a few large eigenvalues; the rest of the spectrum of K may lie within a limited range.

Some iterative solvers for systems of linear equations can take advantage of this, and some cannot. Primitive methods like the Jacobi Method or steepest descent behave poorly whenever the condition number of K is large, even if only one bad eigenvalue is responsible. By contrast, some Krylov subspace methods perform well in these circumstances. For example, the conjugate gradient method [27, 43] (which only works with symmetric stiffness matrices, and therefore only with self-adjoint differential equations like the Poisson equation) can circumvent each bad eigenvalue with one extra iteration, after which it performs as well as it would if the bad eigenvalue were not in the spectrum. GMRES [41], a Krylov subspace method for nonsymmetric systems, is less able to take advantage of this effect because it can usually only be run for a fixed number of iterations (say, 50) on a large-scale system before it must be restarted (to avoid running out of memory). The effect of a few bad elements depends strongly on the method used to solve the linear equations.

6 Quality Measures

Ideally, an algorithm for mesh generation or mesh improvement would directly optimize the fidelity of the interpolated surface over the mesh, or the accuracy of the approximate solution of a system of partial differential equations. However, these criteria are difficult and expensive to measure. At any rate, a single mesh is typically used to interpolate several different surfaces, or to solve several different numerical problems.

Instead, mesh generation and improvement algorithms usually select a single, easily-computed quality measure to evaluate the individual elements they create. For instance, a program might try to maximize the minimum angle. Measures based on interpolation error bounds and stiffness matrix conditioning appear in Section 6.1. Section 6.2 offers advice on when and how to use quality measures and error bounds. The mesh generation literature offers many other element quality measures, which are surveyed in Section 6.3.

6.1 Quality Measures for Interpolation and Matrix Conditioning

Table 4 tabulates several quality measures for evaluating triangular and tetrahedral elements. These measures are related to the fitness of the elements for interpolation and stiffness matrix formation. For each measure, the higher the value of the quality measure, the better the element. All the measures except those in the first row are positive for properly oriented elements, zero (or undefined) for degenerate elements (triangles whose vertices are collinear, or tetrahedra whose vertices are coplanar), and negative for inverted elements. Some of the measures are sensitive to both size and shape; the scale-invariant measures are sensitive to shape only.

In an ideal mesh, the sizes of the elements are controlled primarily by the need to bound $\|f - g\|$ and $\|\nabla f - \nabla g\|$, for which purpose the error bounds are intended. Each element should be small enough that each of these errors is below some application-determined bound—but no smaller, because an application's running time is tied to the number of elements.

Table 4: Quality measures related to interpolation error or stiffness matrix conditioning for a single element. $\lambda_{\max}$ is the largest eigenvalue of an element stiffness matrix (see Section 3.1), and is computed numerically from the characteristic polynomial or by the cubic equation. The constant c_t is the maximum eigenvalue of C_t. See Section 1 for explanations of other notation.

	Triangles	Tetrahedra
Interpolation quality measures, based on $\|f-g\|_\infty$		
Size and shape (mostly size)	$\dfrac{1}{c_t r_{\mathrm{mc}}^2}$	$\dfrac{1}{c_t r_{\mathrm{mc}}^2}$
Scale-invariant (rarely useful)	$\dfrac{A}{r_{\mathrm{mc}}^2}$	$\dfrac{V}{r_{\mathrm{mc}}^3}$
Interpolation quality measures, based on $\|\nabla f-\nabla g\|_\infty$		
Size and shape	$\dfrac{A}{c_t \ell_{\max}\ell_{\mathrm{med}}(\ell_{\min}+4\|r_{\mathrm{in}}\|)}$	$\dfrac{V\sum_{m=1}^4 A_m}{c_t(\sum_{1\leq i<j\leq 4} A_iA_j\ell_{ij}^2+6\|V\|\max_i\sum_{j\neq i}A_j\ell_{ij})}$
Size and shape (smooth)	$\dfrac{A}{c_t\ell_1\ell_2\ell_3}$	$\dfrac{V\sum_{m=1}^4 A_m}{c_t\sum_{1\leq i<j\leq 4} A_iA_j\ell_{ij}^2}$
Scale-invariant	$\dfrac{A}{(\ell_{\max}\ell_{\mathrm{med}}(\ell_{\min}+4\|r_{\mathrm{in}}\|))^{2/3}}$	$V\left(\dfrac{\sum_{m=1}^4 A_m}{\sum_{1\leq i<j\leq 4} A_iA_j\ell_{ij}^2+6\|V\|\max_i\sum_{j\neq i}A_j\ell_{ij}}\right)^{\frac{3}{4}}$
Scale-invariant (smooth)	$\dfrac{A}{(\ell_1\ell_2\ell_3)^{2/3}}$	$V\left(\dfrac{\sum_{m=1}^4 A_m}{\sum_{1\leq i<j\leq 4} A_iA_j\ell_{ij}^2}\right)^{\frac{3}{4}}$
Conditioning quality measures		
Size and shape		$\mathrm{sign}(V)\dfrac{1}{\|\lambda_{\max}\|}$ or zero if $V=0$
Size and shape (smooth)		$\dfrac{V}{A_{\mathrm{rms}}^2}$
Scale-invariant	$\dfrac{A}{3\ell_{\mathrm{rms}}^2+\sqrt{(3\ell_{\mathrm{rms}}^2)^2-48A^2}}$	$\dfrac{V}{\|V\lambda_{\max}\|^{3/4}}$ or zero if $V=0$
Scale-invariant (smooth)	$\dfrac{A}{\ell_{\mathrm{rms}}^2}$	$\dfrac{V}{A_{\mathrm{rms}}^{3/2}}$
If a measure is not used for numerical optimization, use its second, third, or fourth power if it avoids the need to compute roots (but be aware of the lost sign).		

The shapes of elements are usually controlled by the need to bound $\|\nabla f-\nabla g\|$, or $\|F^{-1}(\nabla f-\nabla g)\|$ for anisotropic PDEs, and the largest eigenvalue of the element stiffness matrix. For these purposes either the size-and-shape quality measures or the scale-invariant quality measures might be most useful, depending on the circumstances. For applications that have no stiffness matrix and do not care about accurate gradients (the latter being unusual), the shapes of elements are controlled by the need to bound $\|f-g\|$. Even so, the scale-invariant measures related to $\|f-g\|_\infty$ are rarely worth computing.

The size-and-shape quality measures for interpolation are just the reciprocals of the error bounds. (Constants have been dropped because the measures are used only to compare elements.) Maximizing an element's measure is equivalent to minimizing its error. The reciprocals of the error bounds, rather than the error bounds themselves, are preferable for several reasons: the reciprocals are not infinite for degenerate elements; they vary continuously from negative for inverted elements to positive for correctly oriented elements; and they have better behaved derivatives (with respect to the position of each vertex), a helpful property for optimization-based mesh smoothing methods. (They also have nicer contour plots.) These reasons are elaborated in Section 6.3. Some of the quality measures vary smoothly with the vertex positions, and some do not. The smooth measures simplify optimization-based smoothing, but they are based on weaker bounds, so they are less accurate indicators than the nonsmooth measures.

For some purposes, scale-invariant measures of quality are more appropriate. Size-and-shape measures give no consideration to the number of elements needed to solve a problem, but the number should be controlled, because the computation time for an application is at least linearly proportional to the number of elements. The number of elements needed to cover a domain is inversely proportional to their average area or volume. How can we measure an element's ability to offer low error and high volume? The first impulse might be to express the ratio of error to area or volume, but the error $\|\nabla f - \nabla g\|$ varies according to the square root of area or the cube root of volume, so a ratio is not appropriate.

A better idea is to use a measure that compares an element's error bound with other elements of the same area or volume. The method I advocate here is to scale an element t uniformly until its area or volume is one, then evaluate its quality using the size-and-shape quality measure. This two-step procedure can be replaced by a single formula that yields exactly the same result. To find this formula, begin with the size-and-shape quality measure. Multiply every length by a scaling factor s, every area by s^2, and every volume by s^3. Then set $s = A^{-1/2}$ if t is a triangle, or $s = V^{-1/3}$ if t is a tetrahedron, thereby scaling the element so its area or volume is one.

This procedure converts the measure $A/(c_t \ell_1 \ell_2 \ell_3)$ to the scale-invariant measure $A^{3/2}/(\ell_1 \ell_2 \ell_3)$. (The constant c_t is dropped because it is irrelevant for shape comparisons.) This measure is imperfect for two reasons: it is undefined if A is negative, and its gradient (with respect to the position of any vertex) is zero for degenerate elements, which can be crippling if the measure is used as an objective function for optimization-based smoothing. (See Section 6.3 for discussion.) This problem is fixed by raising the quality measure to a power of $2/3$ to ensure that the numerator is A, yielding the quality measure $A/(\ell_1 \ell_2 \ell_3)^{2/3}$. (For the tetrahedral measures, use a power of $3/4$ to ensure that the numerator is V.) The justification for doing this is that raising the quality measure to a power does not change which element is preferred in any comparison. The same treatment generates all the scale-invariant measures in Table 4. (For the scale-invariant conditioning measure, note that the expression $V \lambda_{\max}$ does not generally approach either zero or infinity as a tetrahedron approaches degeneracy.) When the measures are *not* used for numerical optimization, raise them to a power that makes them fast to compute; e.g. $A^3/(\ell_1 \ell_2 \ell_3)^2$.

For matrix conditioning, triangles have only a scale-invariant measure, but a size-and-shape measure for tetrahedra is offered here, reflecting the fact that larger tetrahedra are more likely to dominate $\lambda^K_{\max}$. The measure should be used carefully though, because unlike with interpolation error, the effect of an element's size on conditioning cannot be judged in isolation. The refinement of a small tetrahedron can reduce $\lambda^K_{\min}$.

Figure 15 illustrates the six quality measures related to interpolation over a triangle. In these contour plots, two vertices of a triangle are fixed at the coordinates $(0, 0)$ and $(1, 0)$, and the third vertex varies freely. The contours illustrate the quality of the triangle as a function of the position of the third vertex: the lighter the region, the higher the quality. Observe that the scale-invariant measures penalize small angles more strongly than the size-and-shape measures, because triangles with small angles consume computation time without covering much area.

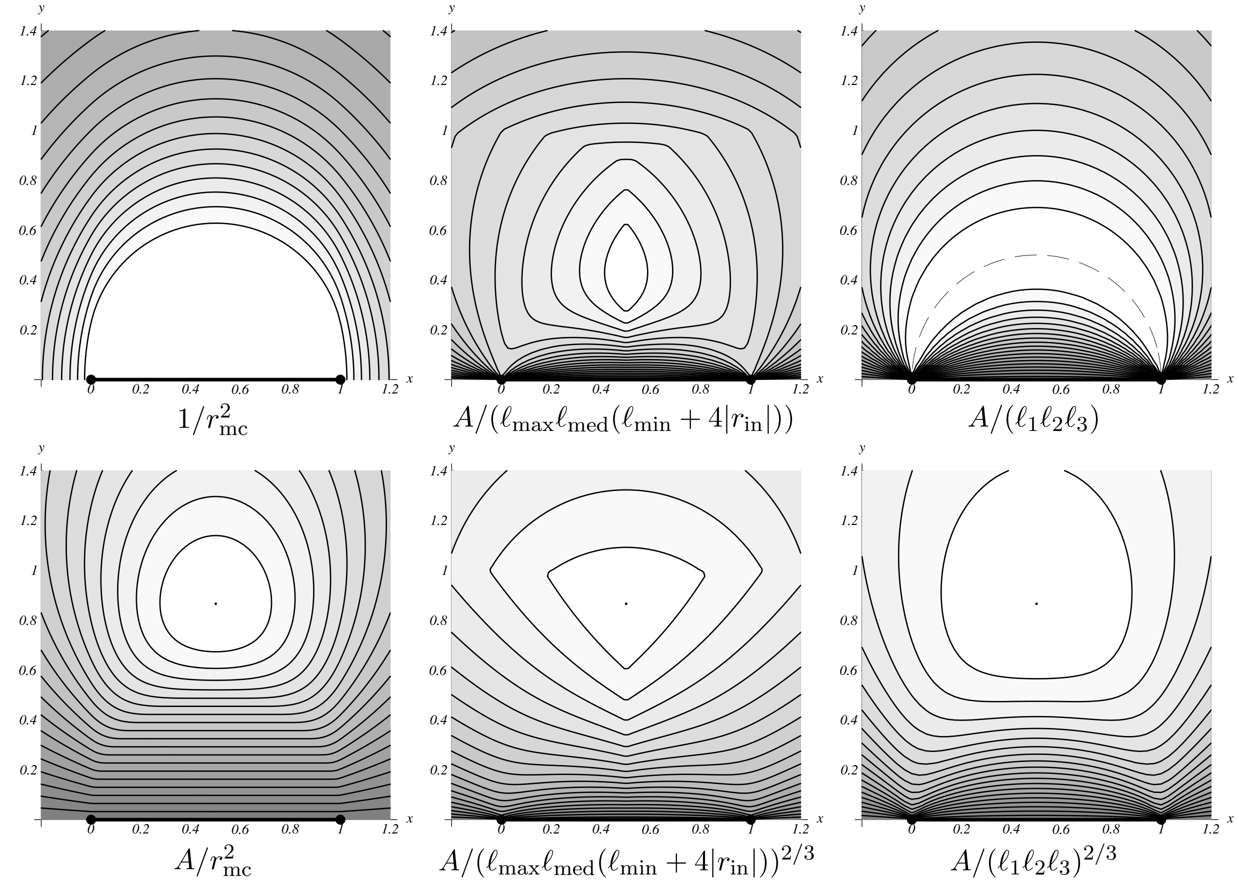


Figure 15: Six interpolation-based quality measures from Table 4 for a triangle with vertices $(0,0)$, $(1,0)$, and (x,y). The top three measures are size-and-shape measures (reciprocals of the error bounds); the bottom three measures are scale-invariant. The left measures reflect $\|f-g\|_\infty$, and the others reflect $\|\nabla f-\nabla g\|_\infty$, with the rightmost measures being smooth (except where two vertices coincide) but less accurate.

Figure 16 illustrates the six quality measures related to interpolation over a tetrahedron. Three vertices of a tetrahedron are fixed at the coordinates $(0,0,0)$, $(\sqrt{3}/2, 1/2, 0)$, and $(\sqrt{3}/2, -1/2, 0)$, and the fourth vertex varies freely along the x- and z-axes. Each plot depicts a cross-section of space, $y=0$, as Figure 17 shows.

Figure 18 illustrates three quality measures (one for triangles, two for tetrahedra) associated with the maximum eigenvalue of the element stiffness matrix.

Table 5 is a list of quality measures for elements in conditions where anisotropic elements are preferred. Note that in the scale-invariant measures for triangles, it does not matter whether A, $\widehat{A}$, or $\widetilde{A}$ is used in the numerator, because they only differ by a fixed constant (namely $|E|$ or $|F|$). Choose whichever is most convenient to calculate. Likewise, in the scale-invariant measures for tetrahedra, V, $\widehat{V}$, and $\widetilde{V}$ are interchangeable.

6.2 How to Use Error Bounds and Quality Measures

Should you use an error bound, a size-and-shape measure, or a scale-invariant measure? The answer depends on how the bound or measure is used. This section gives suggested answers for mesh refinement,

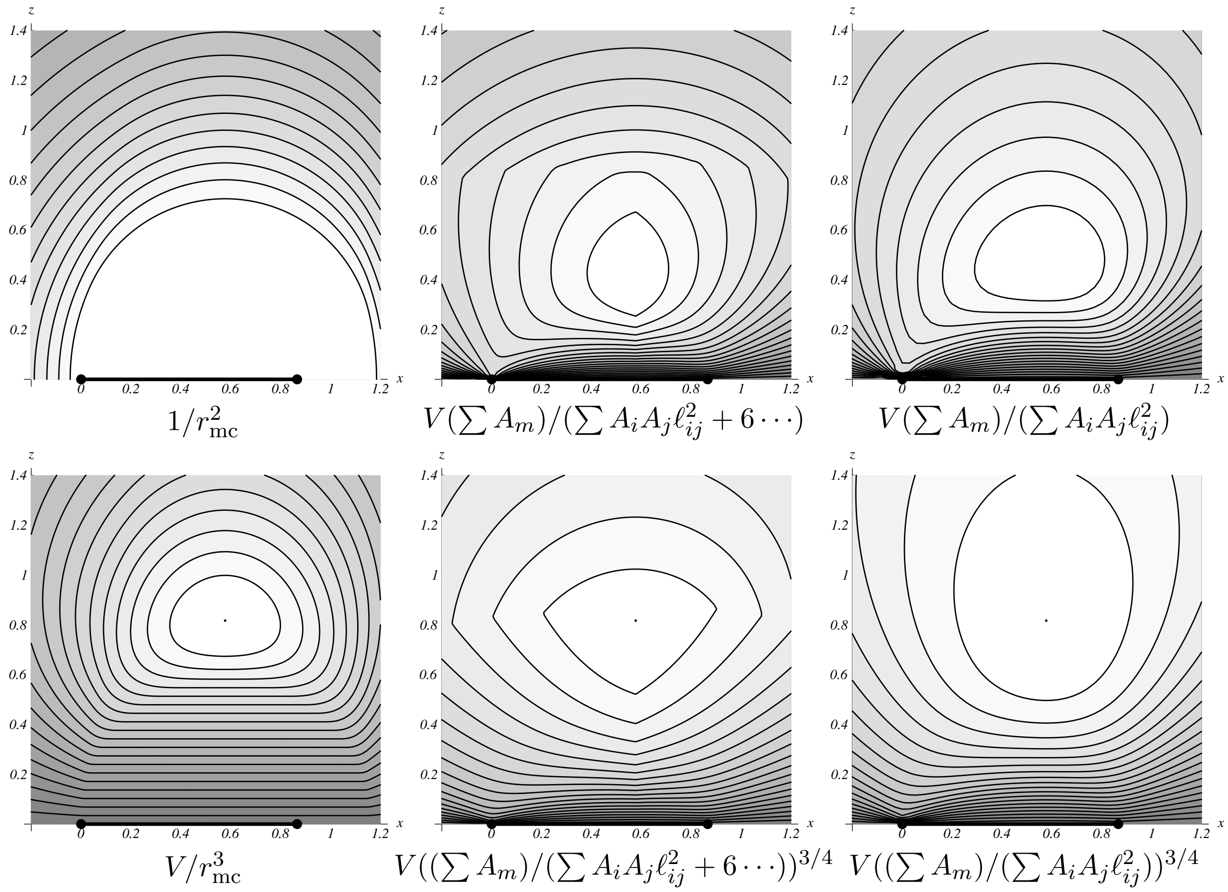


Figure 16: Six interpolation-based quality measures from Table 4 for a tetrahedron with vertices $(0, 0, 0)$, $(\sqrt{3}/2, 1/2, 0)$, $(\sqrt{3}/2, -1/2, 0)$, and $(x, 0, z)$. The top three measures are size-and-shape measures (reciprocals of the error bounds); the bottom three measures are scale-invariant. The left measures reflect $\|f - g\|_\infty$, and the others reflect $\|\nabla f - \nabla g\|_\infty$, with the rightmost measures being smooth (except where two vertices coincide) but less accurate.

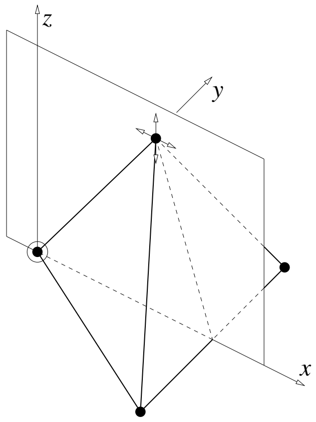


Figure 17: The tetrahedron configuration and cross-section of space used to plot the tetrahedron quality measures in Figures 16, 18, 20, and 21.

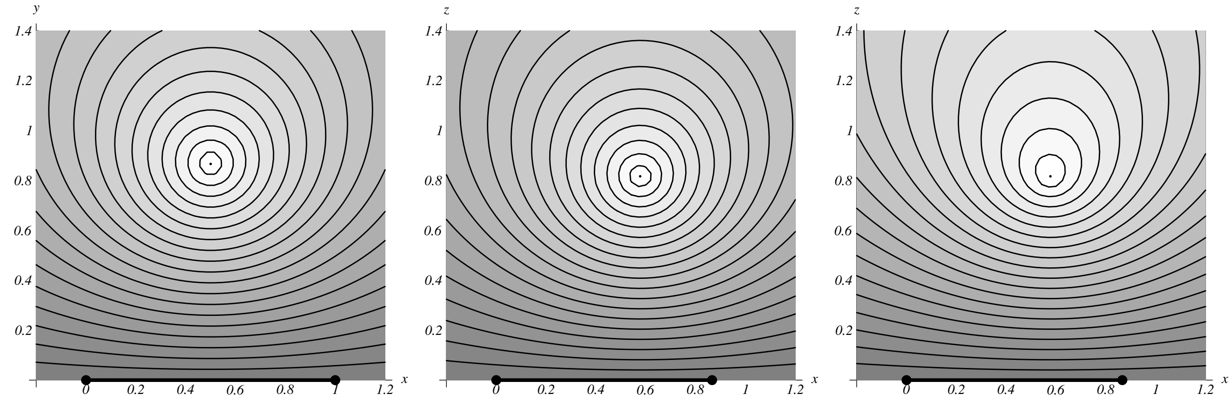


Figure 18: Three matrix conditioning-based quality measures. Left: Level sets of a scale-invariant measure for a triangle with vertices $(0,0)$, $(1,0)$, and (x,y). Center: Size-and-shape measure for a tetrahedron with vertices $(0,0,0)$, $(\sqrt{3}/2,1/2,0)$, $(\sqrt{3}/2,-1/2,0)$, and $(x,0,z)$. Right: Scale-invariant measure for a tetrahedron.

mesh smoothing, topological transformations, and point placement. For mesh refinement and topological transformations, the error bounds are most useful when an application can establish pointwise upper bounds on the sizes of the interpolation errors it is willing to accept. (For mesh smoothing, such bounds are unnecessary.)

One theme of this article is that it is usually more appropriate to use a size-and-shape measure than to treat the sizes and shapes of elements separately, and it can lead to different conclusions. For example, given a point set in the plane, an algorithm of Edelsbrunner, Tan, and Waupotitsch [19] computes the triangulation of the point set that minimizes the maximum angle in $\mathcal{O}(n^2 \log n)$ time. Superficially, the Babuška and Aziz result [4] seems to suggest that this triangulation is a nearly ideal choice for controlling the interpolation error $\|\nabla f - \nabla g\|$. However, the tighter analysis in Section 2 shows that smaller elements can tolerate angles closer to 180° than larger elements, and the circumradius of a triangular element is a better gauge of the error $\|\nabla f - \nabla g\|$ than the element's largest angle. Hence, the two-dimensional Delaunay triangulation, which can be computed in $\mathcal{O}(n \log n)$ time [42, 25], is usually a better triangulation for interpolation.

Mesh refinement. In mesh refinement algorithms, including Delaunay refinement, an element is refined if it is too large or badly shaped. There is no need to try to wrap up element quality into a single measure; instead, an element can be required to pass separate tests for interpolation error and stiffness matrix conditioning.

To control interpolation accuracy, the error bounds are most appropriate. There are two such bounds—one related to the absolute interpolation error, and one related to the error in the gradient (or the skewed gradient). The application that uses the mesh should set a pointwise maximum for one or both of these errors. A mesh refinement program can compare each element against both error bounds, and refine any element that fails either test.

If maximum allowable errors are not specified, or if estimates for c_t are not available, another approach is to choose an upper limit on the number of elements, and repeatedly refine the element with the smallest size-and-shape measure. This has the effect of keeping interpolation error bounds as uniform across elements as possible. (If c_t is unknown, simply drop it from the measure.)

An advantage of error bounds and size-and-shape measures over scale-invariant measures for mesh refinement is that the restrictions on shape are gradually relaxed as the element sizes decrease, so overrefinement is less likely.

Table 5: Anisotropic quality measures. Quantities with a caret are properties of the element $\widehat{t}$, found by applying the linear transformation E (described in Section 2.2) to t. Quantities with a tilde are properties of the element $\widetilde{t}$, found by applying the linear transformation F (described in Section 3.2) to t. The constants c_t and $\widetilde{c}_t$ are the maximum eigenvalues of C_t and $F^{-1}C_tF^{-1}$, respectively.

	Triangles	Tetrahedra
Interpolation quality measures, based on $\lVert f-g\rVert_\infty$		
Size and shape (mostly size)	$\dfrac{1}{c_t\widehat{r}_{\mathrm{mc}}^2}$	$\dfrac{1}{c_t\widehat{r}_{\mathrm{mc}}^2}$
Scale-invariant (rarely useful)	$\dfrac{\widehat{A}}{\widehat{r}_{\mathrm{mc}}^2}$	$\dfrac{\widehat{V}}{\widehat{r}_{\mathrm{mc}}^3}$
Interpolation quality measures, based on $\lVert F^{-1}\nabla f - F^{-1}\nabla g\rVert_\infty$ (for measures based on $\lVert \nabla f - \nabla g\rVert_\infty$, just remove all the tildes)		
Size and shape (smooth)	$\left(\dfrac{\widetilde{c}_t\sum_{1\le i<j\le 3}\widetilde{\ell}_i\widetilde{\ell}_j\widehat{\ell}_k^{\,2}}{2\lvert\widetilde{A}\rvert(\widetilde{\ell}_1+\widetilde{\ell}_2+\widetilde{\ell}_3)}+\sqrt{c_t\widetilde{c}_t}\sum_{i=1}^{3}\widehat{\ell}_i\right)^{-1}$	$\left(\dfrac{\widetilde{c}_t\sum_{1\le i<j\le 4}\widetilde{A}_i\widetilde{A}_j\widehat{\ell}_{ij}^{\,2}}{\lvert\widetilde{V}\rvert\sum_{m=1}^{4}\widetilde{A}_m}+2\sqrt{c_t\widetilde{c}_t}\sum_{i=1}^{6}\widehat{\ell}_i\right)^{-1}$
Scale-invariant (smooth)	$\widetilde{A}\left(\dfrac{\sum_{i<j}\widetilde{\ell}_i\widetilde{\ell}_j\widehat{\ell}_k^{\,2}}{\widetilde{\ell}_1+\widetilde{\ell}_2+\widetilde{\ell}_3}+2\lvert\widetilde{A}\rvert\sqrt{\dfrac{c_t}{\widetilde{c}_t}}\sum_{i=1}^{3}\widehat{\ell}_i\right)^{-\frac{2}{3}}$	$\widetilde{V}\left(\dfrac{\sum_{i<j}\widetilde{A}_i\widetilde{A}_j\widehat{\ell}_{ij}^{\,2}}{\sum_{m=1}^{4}\widetilde{A}_m}+2\lvert\widetilde{V}\rvert\sqrt{\dfrac{c_t}{\widetilde{c}_t}}\sum_{i=1}^{6}\widehat{\ell}_i\right)^{-\frac{3}{4}}$
where k is distinct from i and j in each term of the summation.		
Conditioning quality measures		
Size and shape		$\operatorname{sign}(\widetilde{V})\dfrac{1}{\lvert\widetilde{\lambda}_{\max}\rvert}$ or zero if $\widetilde{V}=0$
Size and shape (smooth)		$\dfrac{\widetilde{V}}{\widetilde{A}_{\mathrm{rms}}^2}$
Scale-invariant	$\dfrac{\widetilde{A}}{3\widetilde{\ell}_{\mathrm{rms}}^2+\sqrt{(3\widetilde{\ell}_{\mathrm{rms}}^2)^2-48\widetilde{A}^2}}$	$\dfrac{\widetilde{V}}{\lvert\widetilde{V}\widetilde{\lambda}_{\max}\rvert^{3/4}}$ or zero if $\widetilde{V}=0$
Scale-invariant (smooth)	$\dfrac{\widetilde{A}}{\widetilde{\ell}_{\mathrm{rms}}^2}$	$\dfrac{\widetilde{V}}{\widetilde{A}_{\mathrm{rms}}^{3/2}}$
If a measure is not used for numerical optimization, use its second, third, or fourth power if it avoids the need to compute roots (but be aware of the lost sign).		

To address matrix conditioning, one must differentiate between two types of mesh refinement: hierarchical refinement [31], in which elements are sliced into smaller elements without ever removing an existing face, and non-hierarchical methods like Delaunay refinement [16, 40, 45, 46], wherein refinement is accompanied by flipping to improve element quality. Hierarchical refinement cannot remove a small angle or increase the volume of the smallest element, so it is ineffective for improving conditioning.

Non-hierarchical refinement can improve $\lambda_{\max}^K$ by eliminating small angles, and might occasionally improve $\lambda_{\min}^K$ by replacing elements with tiny volumes (like slivers) with larger ones. However, scale-invariant measures for matrix conditioning can be dangerous in the context of refinement, because the creation of smaller elements can occasionally decrease $\lambda^K_{\min}$ and worsen the conditioning of the stiffness matrix. Refining an element to achieve a slight improvement in its smallest angle can be a false economy. One option is to use the measure to compare an element with the elements that will appear if the original element is refined. If the shape of the worst new element is better than the original by a margin large enough to justify the smaller elements, go ahead and refine the original element. If it is not possible to determine the new elements in advance, set a weak bound on the minimum acceptable element quality, so that an element is refined only if it is likely to be replaced by much better shaped elements.

Optimization-based mesh smoothing. As I have mentioned, the error bounds are poorly behaved objective functions for numerical optimization, and the quality measures behave much better. The size-and-shape measures are suitable for smoothing because they make appropriate tradeoffs between the size and shape of an element. However, because some of the size-and-shape measures do not penalize small angles harshly, an optimization-based smoother must take extra care not to create degenerate or inverted elements.

There is at least one common circumstance in which the scale-invariant measures might do better. Suppose the input to the smoother is a graded mesh generated by some other program that had access to information about the ideal sizes of elements, but the mesh smoother does not have that information. (This information might include the value of c_t for each triangle and a function that specifies the maximum allowable interpolation error at each point in the domain.) In this circumstance, the size-and-shape measures will try to make the mesh more uniform, whereas the scale-invariant measures will better preserve the original sizes of the elements.

Unfortunately, optimization-based smoothing can only optimize one objective function. For applications in which matrix conditioning is important, there is a natural tension between the needs of interpolation and discretization errors and the needs of matrix conditioning. This tension can be resolved by using a "combined" measure that is a weighted harmonic mean of two scale-invariant measures. For example, given an interpolation-based measure Q_1 and a conditioning-based measure Q_2, define a "combined" measure Q where

$$\frac{1}{Q} = \frac{w}{Q_1} + \frac{1-w}{Q_2}$$

and the constant weights w and $1 - w$ are chosen according to how much influence each measure should have. An advantage of the harmonic mean is that if one of the original measures assigns a low score (near zero) to an element, its opinion dominates.

The tension could instead be resolved by using a conditioning measure alone, but be aware that for tetrahedra, the scale-invariant conditioning measure is only mildly opposed to a tetrahedron that has an angle near 180° and no small angle (recall Figure 14), and the size-and-shape measure is not opposed to it at all. Such a tetrahedron has a terrible effect on interpolation but not on conditioning.

There is less tension between the quality measures associated with $\|\nabla f - \nabla g\|_\infty$ and with $\|f - g\|_\infty$. This tension can be resolved by treating only the former (especially if a scale-invariant measure is preferred), or by using the weighted harmonic mean of two size-and-shape measures where the weights are determined by the form of the discretization error (i.e. the energy norm).

Topological transformations. The comments on smoothing apply to topological transformations as well.

Although topological transformations and smoothing are both mesh improvement methods, they differ in that transformations can change the number of elements. The size-and-shape measures tend to prefer transformations that increase the number of elements, so they run the risk of overrefining the mesh. This pitfall is avoided by the use of an error threshold, probably the same threshold used for mesh refinement. Specifically, a topological transformation that increases the number of elements is performed only if it eliminates an element whose interpolation error or eigenvalue bound exceeds the threshold, and all the new elements are better. If such a threshold is not available, an alternative is to use scale-invariant measures.

Vertex placement in advancing front methods. An advancing front mesh generator should try to place the largest possible element whose bounds on interpolation error meet the prescribed thresholds, and which perhaps meets a threshold on a conditioning measure as well. Thus, placing a new vertex involves solving a small optimization problem.

6.3 A Brief Survey of Quality Measures

This section summarizes and depicts many of the scale-invariant quality measures for triangles and tetrahedra seen in the mesh generation literature. The measures in the literature are usually *ad hoc* in nature. To my knowledge, the new measures presented in Section 6.1 are the first attempt to derive scale-invariant quality measures directly from interpolation errors or stiffness matrix conditioning. (But see the end of this section for a related effort by Knupp [32].)

Imagine an element in which one vertex v is free to move while the other vertices are fixed, and let $q(v)$ be the quality of the element (according to some measure) as a function of v. As a crude guide, here are seven properties one might wish the quality measure to have under certain circumstances.

1. All degenerate elements have a quality of zero.
2. The measure is scale-invariant: two elements of different sizes and identical shapes have the same quality. (Section 6.2 advocates non-scale-invariant measures for most purposes. However, nearly all quality measures in the literature are scale-invariant.)
3. The measure is normalized to achieve a maximum value of one. (Most measures discussed here achieve a value of one only for an equilateral triangle or tetrahedron, but $\ell_{\max}/r_{\mathrm{circ}}$ is the exception.)
4. All inverted elements have a negative quality.
5. The gradient of $q(v)$ (with respect to v) is nonzero for most degenerate elements.
6. $q(v)$ is a smooth function of v (except where v coincides with another vertex).
7. $q(v)$ is *quasiconvex* over the domain of non-inverted elements. For the present purposes, this means that for any constant $c \geq 0$, the point set $\{v : q(v) \geq c\}$ is convex (at least for values of v where the element is not inverted).

Field [20] defines a *fair measure* to be a quality measure with the first three properties. These three properties make quality measures easier to compare and combine. Most of the quality measures in the literature are fair measures.

The fourth property is a convenience: an inverted element can be distinguished by the sign of its quality. Nonnegative measures sometimes necessitate an extra step to check the sign of an element's area or volume. (The area or volume is a subexpression of every quality measure, so little additional computation time is needed.)

The last three properties matter only for numerical optimization algorithms for mesh smoothing. These algorithms use the gradient $\nabla q(v)$ to choose search directions. Some measures, including the popular *radius ratio* $r_{\mathrm{in}}/r_{\mathrm{circ}}$ and other nonnegative measures, have gradients that are zero for all degenerate elements, and are very small for nearly degenerate elements. This implies that optimization-based smoothing will be least effective for the elements that most need it. (See below for a discussion of how to fix this.)

The smoothness of $q(v)$ is important because it dictates the type of optimization algorithm used for smoothing. Some quality measures are not smooth functions of the vertex positions. For example, suppose $q(v)$ maps the position of v to the minimum angle of the element. The gradient of q (with respect to v) changes abruptly where the identity of the minimum angle changes, so q is nonsmooth. Nonsmooth quality measures need a numerical optimizer designed specifically for nonsmooth functions [21, 22]. Optimizers for smooth functions are generally faster and easier to implement. However, nonsmooth optimizers can maximize the quality of the worst element or angle. Note that even the smoothest quality measures are not smooth where all the vertices of an element are coincident, or where all the vertices of a tetrahedron are collinear, but these deviations from smoothness rarely cause difficulties.

The final property, quasiconvexity, is interesting because some mesh smoothing algorithms attempt to move v to the point that maximizes the objective function $\min_i q_i(v)$, where $q_i(v)$ is the quality of element i, and i ranges over all elements that have v for a vertex. If the quality measure used to judge each element is quasiconvex, then every local maximum of $\min_i q_i(v)$ is a global maximum. In practice, the existence of local maxima that are not globally optimal is rarely bothersome. However, there are combinatorial smoothing algorithms based on generalized linear programming for which the quasiconvexity of the objective function, and the uniqueness of the solution, are necessary for the correctness of the algorithm [1].

Table 6 is an incomplete list of quality measures for triangles taken from the mesh generation literature. (See Table 1 for explanations of the notation.) All of these are fair measures except the last two, which are included here as warnings. Tables 7 and 8 list quality measures for tetrahedra. All are fair measures except the last four in Table 8. The "equivalent expressions" given in each table are generally the best way to compute the quality measures.

The level curves of each triangle measure are plotted in Figure 19. In each plot, two of the vertices are fixed at coordinates $(0, 0)$ and $(1, 0)$, and the third vertex varies freely. The level curves of each tetrahedron measure are plotted in Figures 20 and 21. Each plot depicts the cross-section $y = 0$ as illustrated in Figure 17. Three of the vertices are fixed at coordinates $(0, 0, 0)$, $(\sqrt{3}/2, 1/2, 0)$, and $(\sqrt{3}/2, -1/2, 0)$, and the fourth vertex varies freely.

Interestingly, every fair triangle measure surveyed here can be expressed in a form that includes A in its numerator, and every fair tetrahedron measure includes V in its numerator, although it sometimes takes some algebraic manipulation to reveal the connection. I did not find any fair measure in the literature that breaks the rule. The differences in the published quality measures (except the bad ones) are largely in their denominators, and in the power to which A or V is raised.

The first eight triangle measures in Table 6, and all the tetrahedron measures in Table 7 are reasonably good. The last two triangle measures in Table 6 and the last four tetrahedron measures in Table 8 are not fair measures because they are not zero for all degenerate elements, as Figures 19 and 21 illustrate.

Some of the other measures have the flaw that their gradients are zero for degenerate elements. These measures can often be spotted by the factor of A^2, A^3, V^2, V^3, or V^4 in their numerators, but they can be fixed. Given any fair measure q, q^c is also a fair measure if $c > 0$. For the purpose of comparing elements (that are not inverted), the measures q and q^c are equivalent—if q rates one element better (closer to 1) than another, so does q^c. However, their gradients are different, and one is usually better behaved as an objective function for numerical optimization algorithms than the other. To be considered well behaved, a measure should have a nonzero but finite gradient for most degenerate elements.

Nearly all the quality measures considered in this section can be expressed in the form A^c/X or V^c/X, where X is nonzero even for most degenerate elements. If $c > 1$, the measure has a gradient of zero when

Table 6: Quality measures for triangles. All are fair measures but the last two.

Measure	Equivalent expression	Comments
$\dfrac{4(12+7\sqrt{3})^{1/3}}{3}$	$\dfrac{A}{(\ell_{\max}\ell_{\text{med}}(\ell_{\min}+4\lvert r_{\text{in}}\rvert))^{2/3}}$	Implicated in the errors of the gradients of an interpolated piecewise linear function. A nonsmooth function of the vertex positions.
$\dfrac{4}{\sqrt{3}}\dfrac{A}{(\ell_1\ell_2\ell_3)^{2/3}}$		A smooth (but less precise) alternative to the measure above.
$4\sqrt{3}\dfrac{A}{3\ell_{\text{rms}}^2+\sqrt{(3\ell_{\text{rms}}^2)^2-48A^2}}$		Implicated in the condition number of the stiffness matrix for Poisson's equation. Level sets are circles. Not quite smooth (infinite gradient for an equilateral triangle).
$\dfrac{4}{\sqrt{3}}\dfrac{A}{\ell_{\text{rms}}^2}$	$=4\sqrt{3}\dfrac{A}{\ell_1^2+\ell_2^2+\ell_3^2}$	Suggested by Bhatia and Lawrence [12]. A smooth alternative to the measure above. Related to the condition number of an element transformation matrix [32]. Level sets are circles.
$\dfrac{3}{\pi}\theta_{\min}$	$=\dfrac{3}{\pi}\arcsin\dfrac{2A}{\ell_{\max}\ell_{\text{med}}}$	With $\theta_{\min}$ measured in radians. Nonsmooth.
$\dfrac{2}{\sqrt{3}}\sin\theta_{\min}$	$=\dfrac{1}{\sqrt{3}}\dfrac{\ell_{\min}}{r_{\text{circ}}}$ $=\dfrac{4}{\sqrt{3}}\dfrac{A}{\ell_{\max}\ell_{\text{med}}}$	This measure is naturally improved by Delaunay refinement algorithms for mesh generation [16, 40, 46]. Nonsmooth.
$\dfrac{2}{\sqrt{3}}\dfrac{a_{\min}}{\ell_{\max}}$	$=\dfrac{4}{\sqrt{3}}\dfrac{A}{\ell_{\max}^2}$	The *aspect ratio*, or ratio between the minimum and maximum dimensions of the triangle. Nonsmooth.
$2\sqrt{3}\dfrac{r_{\text{in}}}{\ell_{\max}}$	$=4\sqrt{3}\dfrac{A}{\ell_{\max}(\ell_1+\ell_2+\ell_3)}$	The measure associated with interpolation error in the usual (weaker) bounds given by approximation theory. Nonsmooth.
$2\dfrac{r_{\text{in}}}{r_{\text{circ}}}$	$=16\dfrac{A^2}{\ell_1\ell_2\ell_3(\ell_1+\ell_2+\ell_3)}$	The *radius ratio*. Not good for numerical optimization, but its square root is. See the commentary in the text. Nonnegative. Smooth.
$\dfrac{4\sqrt{3}}{9}\dfrac{A}{r_{\text{circ}}^2}$	$=\dfrac{64\sqrt{3}}{9}\dfrac{A^3}{\ell_1^2\ell_2^2\ell_3^2}$	Not good for numerical optimization, but its cube root (second entry from top) is. See the commentary in the text. Smooth.
$\dfrac{\ell_{\min}}{\ell_{\max}}$		A poor quality measure that can be as large as 0.5 for a triangle with a 180° angle. Nonnegative. Nonsmooth.
$\dfrac{1}{2}\dfrac{\ell_{\max}}{r_{\text{circ}}}$	$=2\dfrac{A}{\ell_{\text{med}}\ell_{\min}}$	Prefers right triangles, not equilateral ones (see Figure 19). Does not penalize small angles at all. Nonsmooth.

Table 7: Fair quality measures for tetrahedra. Continued in Table 8.

Measure	Equivalent expression	Comments
$\frac{9\sqrt{3}}{8}\frac{V}{r_{\mathrm{mc}}^3}$	See Appendix A.3	Implicated in the errors of an interpolated piecewise linear function. Rarely appropriate. Nonsmooth.
$\frac{3(486\sqrt{3}+594\sqrt{2})^{1/4}\,V(\sum_{m=1}^4 A_m)^{3/4}}{2(\sum_{1\le i<j\le 4} A_iA_j\ell_{ij}^2+6\lvert V\rvert\max_i\sum_{j\ne i}A_j\ell_{ij})^{3/4}}$		Implicated in the errors of the gradients of an interpolated piecewise linear function. Nonsmooth.
$\frac{3^{17/8}}{2^{3/4}}V\left(\frac{\sum_{m=1}^4 A_m}{\sum_{1\le i<j\le 4}A_iA_j\ell_{ij}^2}\right)^{3/4}$		A smooth (but less precise) alternative to the measure above.
$\frac{1}{\sqrt{3}}\frac{V}{\lvert V\lambda_{\max}\rvert^{3/4}}$		Implicated in the condition number of the stiffness matrix for Poisson's equation. $\lambda_{\max}$ is a root of a cubic polynomial (Section 3.1). Not quite smooth (infinite gradient for an equilateral tetrahedron).
$\frac{3^{7/4}}{2\sqrt{2}}\frac{V}{A_{\mathrm{rms}}^{3/2}}$	$=3^{7/4}\frac{V}{(\sum_{i=1}^4 A_i^2)^{3/4}}$	A smooth, easily computed alternative to the measure above.
$\frac{3\sqrt{6}}{2}\frac{V}{\ell_{\mathrm{rms}}A_{\mathrm{rms}}}$		Suggested by Knupp [32]. Related to the Frobenius condition number of an element transformation matrix. See the text for details. Smooth.
$6\sqrt{2}\frac{V}{\ell_{\mathrm{rms}}^3}$		Suggested by Parthasarathy, Graichen, and Hathaway [37]. Smooth.
$6\sqrt{2}\frac{V}{\ell_{\max}^3}$		Nonsmooth.
$\frac{\sqrt{3}}{\sqrt{2}}\frac{a_{\min}}{\ell_{\max}}$	$=\frac{3\sqrt{6}}{2}\frac{V}{\ell_{\max}A_{\max}}$	Nonsmooth.
$\sqrt{2}\frac{h_{\min}}{\ell_{\max}}$	$=6\sqrt{2}\frac{V}{\ell_{\max}\max_{\mathbf{w},\mathbf{x}}\lvert\mathbf{w}\times\mathbf{x}\rvert}$	The aspect ratio. $\mathbf{w}$ and $\mathbf{x}$ are chosen from among the edge vectors of the tetrahedron. Nonsmooth.
$2\sqrt{6}\frac{r_{\mathrm{in}}}{\ell_{\max}}$	$=6\sqrt{6}\frac{V}{\ell_{\max}\sum_{i=1}^4 A_i}$	The measure associated with interpolation error in the usual (weaker) bounds given by approximation theory. Suggested for mesh generation by Baker [6]. Nonsmooth.
$\frac{9}{\sqrt{6}}\min_i\sin(\phi_i/2)$	See Liu and Joe [36]	Suggested by Liu and Joe [36]. Nonsmooth.
$\frac{\theta_{\min}}{\arcsin(2\sqrt{2}/3)}$		Nonsmooth. The dihedral angles of an equilateral tetrahedron are about $70.529°$.
$\frac{3\sqrt{2}}{4}\min_{i,j}\sin\theta_{ij}$	$=\frac{9\sqrt{2}}{8}V\min_{1\le i<j\le 4}\frac{\ell_{ij}}{A_kA_l}$	With i, j, k, and l distinct. Studied by Freitag and Ollivier-Gooch [22]. Nonsmooth.

Table 8: Quality measures for tetrahedra (continued from Table 7). The first three are fair measures; the last four are not. A formula for Z is given in Appendix A.2.

Measure	Equivalent expression	Comments
$3\dfrac{r_{\text{in}}}{r_{\text{circ}}}$	$= 108\dfrac{V^2}{Z\sum_{i=1}^{4} A_i}$	The radius ratio. Suggested by Cavendish, Field, and Frey [15]. Not good for numerical optimization, but its square root is. Nonnegative. Smooth.
$\dfrac{9\sqrt{3}}{8}\dfrac{V}{r_{\text{circ}}^3}$	$= 1944\sqrt{3}\dfrac{V^4}{Z^3}$	Not good for numerical optimization, but its fourth root is. Nonnegative. Smooth.
$2187\dfrac{V^4}{(\sum_{i=1}^{4} A_i^2)^3}$		Suggested by de Cougny, Shephard, and Georges [18]. Not good for numerical optimization, but its fourth root (fifth entry in Table 7) is. Nonnegative. Smooth.
$\dfrac{\sqrt{6}}{4}\dfrac{\ell_{\min}}{r_{\text{circ}}}$	$= 3\sqrt{6}\dfrac{V\ell_{\min}}{Z}$	A mediocre measure that does not penalize slivers harshly enough. Can be as large as $\sqrt{3}/2$ for a flat sliver. However, it is the measure naturally improved by Delaunay refinement algorithms for mesh generation [45].
$\dfrac{\ell_{\min}}{\ell_{\max}}$		A poor measure that does not penalize slivers or large solid angles harshly enough. Can be as large as $1/\sqrt{2}$ for a flat sliver. Nonnegative.
$4\dfrac{A_{\min}}{\sum_{i=1}^{4} A_i}$		A poor measure that does not penalize slivers at all. Can be as large as 1 for a flat sliver and $2/3$ for a flat tetrahedron with a large solid angle. Nonnegative.
$\dfrac{1}{2}\dfrac{\ell_{\max}}{r_{\text{circ}}}$	$= 6\dfrac{V\ell_{\max}}{Z}$	Perhaps the worst quality measure ever proposed in the literature. Prefers slivers to equilateral tetrahedra.

A or V is zero, and if $c < 1$, the measure has an infinite gradient when A or V is zero. For numerical optimization, measures for which $c = 1$ are the best behaved.

The measures for which $c \neq 1$ can be improved by replacing them with $A/X^{1/c}$ or $V/X^{1/c}$. Compare the measures $r_{\text{in}}/r_{\text{circ}}$ (the radius ratio), $\sqrt{r_{\text{in}}/r_{\text{circ}}}$, and $\sqrt{r_{\text{circ}}/r_{\text{in}}}$ in Figure 21. A glance reveals that the level sets (contours) of these three functions have exactly the same shape, but the spacing of the level sets is quite different. The measure $r_{\text{in}}/r_{\text{circ}}$ has a gradient of zero for degenerate elements. By contrast, $\sqrt{r_{\text{in}}/r_{\text{circ}}}$ has a positive finite gradient, and is more easily optimized by gradient descent methods. The inverse function $\sqrt{r_{\text{circ}}/r_{\text{in}}}$ has an infinite gradient for a degenerate element, and so is strongly prejudiced against degenerate elements. However, the infinite gradient can threaten the stability of some numerical optimization algorithms, and may be fatal if the initial mesh has degenerate or inverted elements.

It is easy to tell which quality measures are quasiconvex by looking at their contour plots. For example, Figure 19 shows that the conditioning measure A/ℓ_{rms}^2 and the aspect ratio are quasiconvex, whereas the interpolation measure $A/(\ell_1\ell_2\ell_3)^{2/3}$ and the radius ratio are not.

Knupp [32, 33] sheds an interesting light on two of the measures listed in the tables. He advocates quality measures based on the condition numbers of element transformation matrices. Given a triangular or

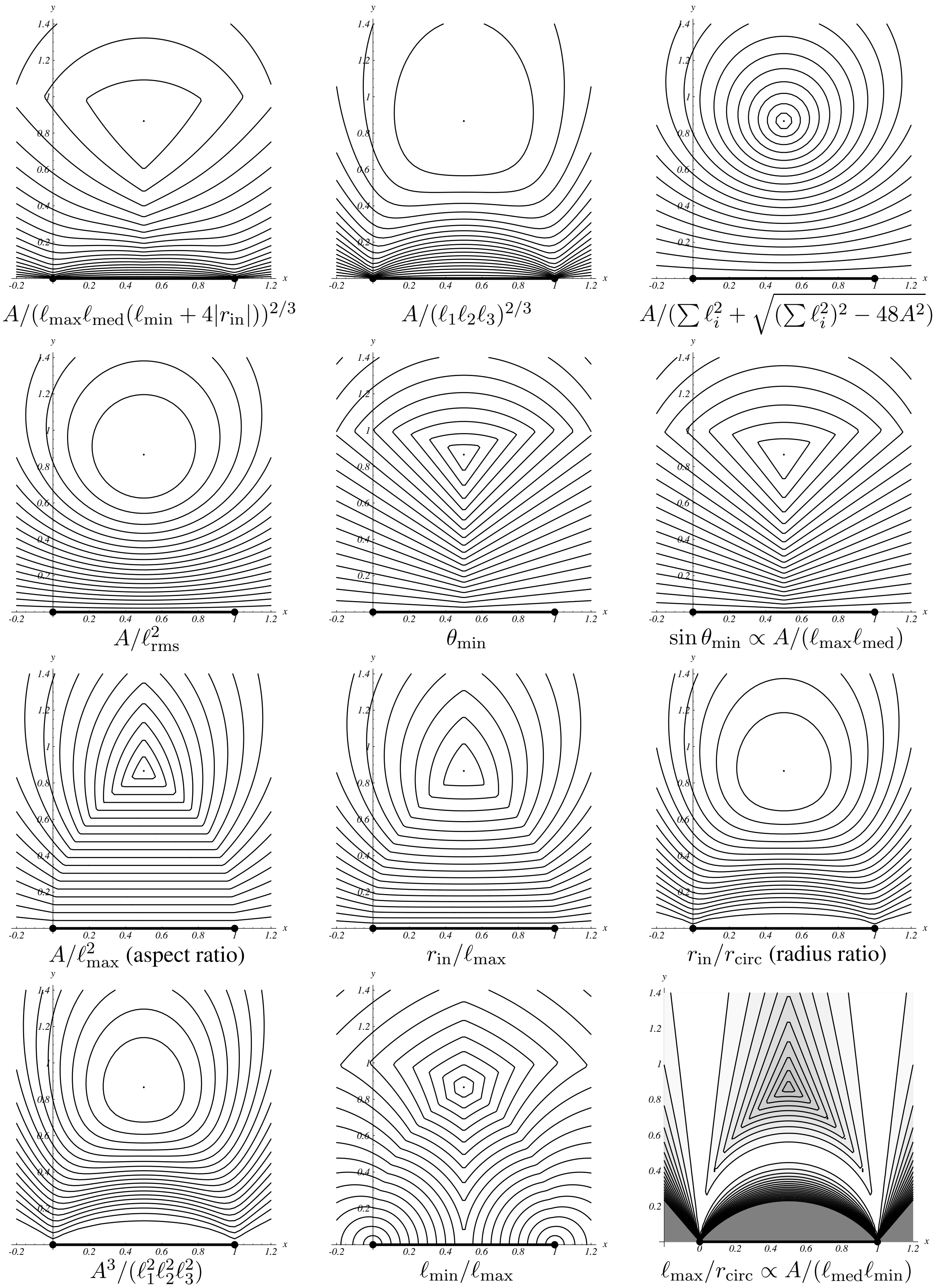


Figure 19: Triangle quality measures. Coefficients (see Table 6) are omitted for brevity.

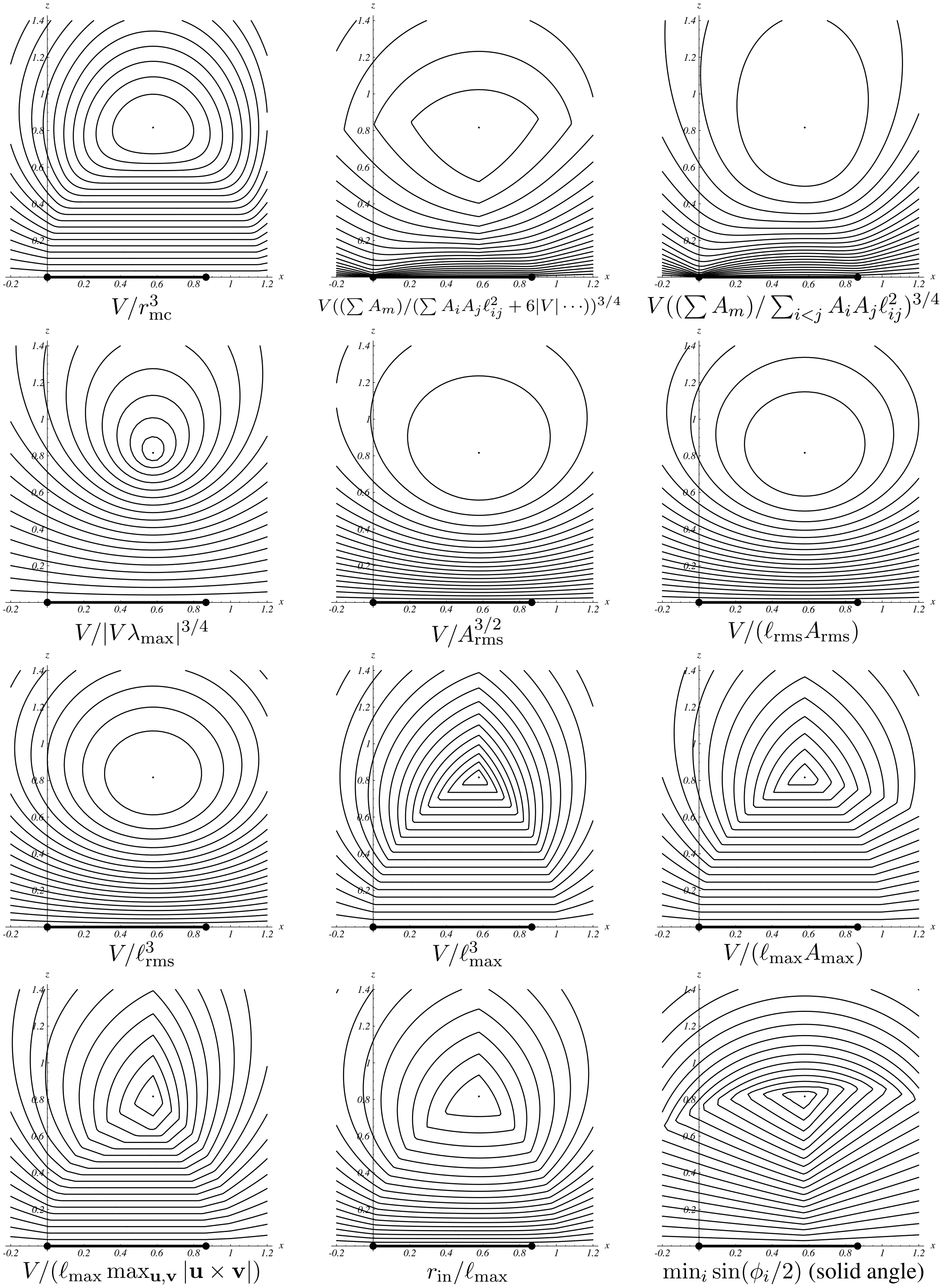
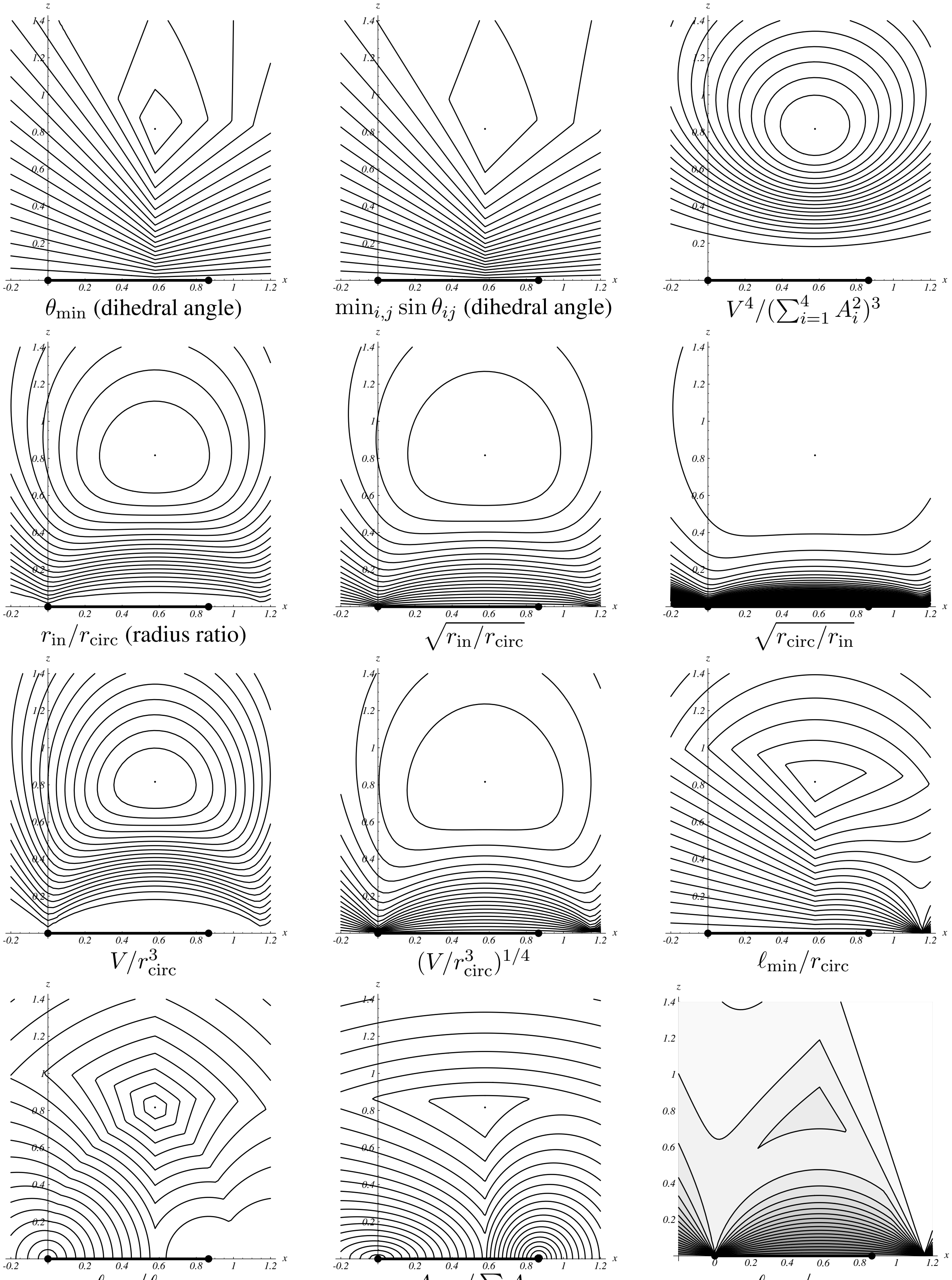


Figure 20: Tetrahedron quality measures. Coefficients (see Table 7) are omitted for brevity. Continued in Figure 21.



$\theta_{\min}$ (dihedral angle) — $\min_{i,j} \sin\theta_{ij}$ (dihedral angle) — $V^4/(\sum_{i=1}^{4} A_i^2)^3$

$r_{\text{in}}/r_{\text{circ}}$ (radius ratio) — $\sqrt{r_{\text{in}}/r_{\text{circ}}}$ — $\sqrt{r_{\text{circ}}/r_{\text{in}}}$

V/r_{circ}^3 — $(V/r_{\text{circ}}^3)^{1/4}$ — $\ell_{\min}/r_{\text{circ}}$

$\ell_{\min}/\ell_{\max}$ — $A_{\min}/\sum A_i$ — $\ell_{\max}/r_{\text{circ}}$

Figure 21: Tetrahedron quality measures (continued from Figure 20). Coefficients (see Table 8) are omitted for brevity.

tetrahedral element, assume that one of its vertices lies at the origin, and let M be a $d \times d$ matrix whose column vectors represent the edges incident on the origin vertex. Let W be a matrix that is constructed the same way, but represents an equilateral element. The transformation matrix $T = MW^{-1}$ maps the equilateral element to the real element. $\kappa(T) = \|T\|_F \|T^{-1}\|_F$ is the *Frobenius condition number* of T, where $\|T\|_F = \sqrt{\sum_{i,j} T_{ij}^2}$ denotes the Frobenius matrix norm.

With this definition, $3/\kappa(T)$ is a fair quality measure. Its strength is that it is an indicator of the "distance" of T from the nearest singular matrix. Specifically,

$$\frac{1}{\kappa(T)} = \min_{D, \text{where } D+T \text{ is singular}} \frac{\|D\|_F}{\|T\|_F}.$$

The measure $3/\kappa(T)$ is expressed in more familiar terms in Tables 6 and 7. For non-inverted triangles, it is $4A/(\sqrt{3}\ell_{\mathrm{rms}}^2)$. For non-inverted tetrahedra, it is $3\sqrt{6}V/(2\ell_{\mathrm{rms}} A_{\mathrm{rms}})$.

Knupp's measure and the measure V/ℓ_{rms}^3 approximate the conditioning measure $V/|V\lambda_{\max}|^{3/4}$ reasonably well, except that they are harder on tetrahedra with large dihedral angles like the one in Figure 14. As all-purpose measures for tetrahedra, they are strong contenders.

7 Conclusions

The tight and nearly-tight error bounds and eigenvalue estimates now available for linear elements make it possible to judge them more accurately than before. One can easily draft a long list of directions in which this work could be extended to the benefit of practitioners in finite element analysis and mesh generation, but most of the extensions will be mathematically challenging.

First, can we find nearly-tight error bounds and quality measures for triangular and tetrahedral elements of higher degree—especially for quadratic Lagrangian interpolation? (The usual starting assumption for quadratic interpolation is that the third derivatives—not the second—are bounded.) Second, can we find nearly-tight bounds on interpolation error and element stiffness matrix eigenvalues for bilinear quadrilateral elements and trilinear hexahedral elements? Third, some partial differential equations have much more complex behavior than Poisson's equation. Do PDEs associated with fluid dynamics or other phenomena lead to conclusions about element shape and anisotropy radically different than the conclusions offered here for Poisson's equation? Nonlinear PDEs are particularly difficult to understand, because their stiffness matrices vary with time.

Fourth, the bounds on interpolation error given here presuppose that there is an upper bound on the second derivatives of the function being interpolated. But in some circumstances, the Hessian H is unbounded at one or more points. For example, many boundary value problems have solutions whose derivatives have singularities. The functional analysis techniques mentioned in Section 2.5 have been successfully used to establish asymptotic error bounds for many singular cases, but nearly-tight bounds remain elusive. For example, the error bound

$$\|\nabla f - \nabla g\|_{L_2(t)} \leq C\ell_{\max} \left\| \frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x \partial y} + \frac{\partial^2 f}{\partial y^2} \right\|_{L_2(t)}$$

holds for some constant C that depends on the shape of the element t. (Note that the right-hand side may be finite even though $\partial^2 f/\partial x^2$ or another second derivative is infinite at some points.) The precise relationship between C (at its tightest) and the shape of t is an open question, but Handscomb [26] discusses how to determine a small C numerically for any particular t (using a finite element method). Are the quality measures related to $\|\nabla f - \nabla g\|_\infty$ presented in Section 6.1 appropriate in the presence of singularities? Handscomb's results suggest that they are not an ideal fit. Do singularities change the comparative effectiveness of different element shapes?

Section 3.3 discusses the constraints on the time steps used in explicit time integration methods. Another pitfall of time integration, which is more difficult to characterize mathematically, is that a sudden transition from large to small elements may cause spurious partial reflections of waves [29]. In some circumstances, a "good" mesh is one in which elements everywhere grade from small to large sizes gradually.

To date, most element quality measures in the mesh generation literature have been somewhat *ad hoc* in nature. I hope this article will start a tradition of deriving element quality measures from "first principles"—based on strong bounds on interpolation error, matrix condition numbers, discretization error, and other application needs.

Acknowledgments

My warmest thanks, respect, and appreciation go to Omar Ghattas for giving me a thorough introduction to this subject, to Dafna Talmor for teaching me a geometric view of the element stiffness matrices for the isotropic and anisotropic Poisson's equation, to Günter Rote for setting me on the right path to find good lower bounds on the errors in interpolated gradients, and to a helpful gentleman at the SIAM 50th Anniversary Meeting whose name I fear I've forgotten, for pointing out the interaction between anisotropic PDEs and discretization error.

Most of this work was carried out at Far Leaves Teas and at Masse's Pastries in Berkeley.

A Formulae

This appendix presents formulae for calculating some of the geometric quantities used in the error bounds and quality measures presented in this paper. They are included here for two reasons: some of these formulae are hard to find, and the formulae that do appear in reference books often incur excessive floating-point roundoff error. Special attention is given here to numerical accuracy.

In these formulae, the norm $|\cdot|$ denotes the Euclidean length of a vector, and the operator $\times$ denotes the vector cross product. For simplicity, the vertices are here named differently than elsewhere in the article. Formulae for triangles (in two or three dimensions) govern a triangle with vertices a, b, and c, and employ the vectors $\mathbf{r} = a - c$ and $\mathbf{s} = b - c$. Formulae for tetrahedra govern a tetrahedron with vertices a, b, c, and d, and employ the vectors $\mathbf{t} = a - d$, $\mathbf{u} = b - d$, and $\mathbf{v} = c - d$. These and other notations are illustrated in Figure 22.

A.1 Triangle Area and Tetrahedron Volume

The signed area of the triangle with vertices a, b, and c is

$$A = \frac{1}{2}\begin{vmatrix} a_x - c_x & b_x - c_x \\ a_y - c_y & b_y - c_y \end{vmatrix} = \frac{1}{2}\begin{vmatrix} \mathbf{r} & \mathbf{s} \end{vmatrix}.$$

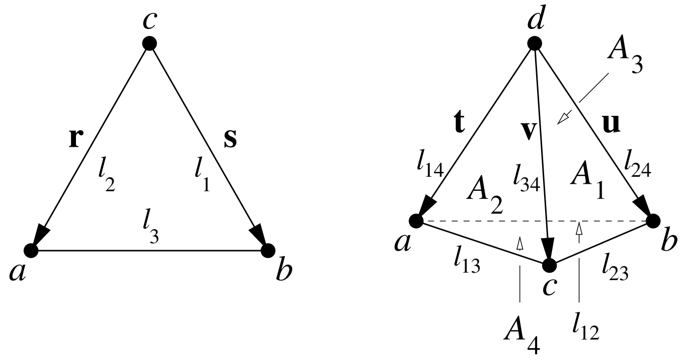


Figure 22: A triangle and a tetrahedron, both having positive orientation.

A is positive if the points occur in counterclockwise order, negative if they occur in clockwise order, and zero if they are collinear.

The following formula computes the unsigned area of a triangular face in E^3 whose vertices are a, b, and c.

$$\begin{aligned}
A_f &= \frac{|\mathbf{r} \times \mathbf{s}|}{2} \\
&= \frac{|(a-c) \times (b-c)|}{2} \\
&= \frac{\sqrt{\begin{vmatrix} a_y - c_y & b_y - c_y \\ a_z - c_z & b_z - c_z \end{vmatrix}^2 + \begin{vmatrix} a_z - c_z & b_z - c_z \\ a_x - c_x & b_x - c_x \end{vmatrix}^2 + \begin{vmatrix} a_x - c_x & b_x - c_x \\ a_y - c_y & b_y - c_y \end{vmatrix}^2}}{2}
\end{aligned}$$

The signed volume of the tetrahedron with vertices a, b, c, and d is

$$V = \frac{1}{6}\begin{vmatrix} a_x - d_x & b_x - d_x & c_x - d_x \\ a_y - d_y & b_y - d_y & c_y - d_y \\ a_z - d_z & b_z - d_z & c_z - d_z \end{vmatrix} = \frac{1}{6}\begin{vmatrix} \mathbf{t} & \mathbf{u} & \mathbf{v} \end{vmatrix}.$$

V is positive if the points occur in the orientation illustrated in Figure 22, negative if they occur in the mirror-image orientation, and zero if the four points are coplanar. You can apply a *right-hand rule*: orient your right hand with fingers curled to follow the circular sequence bcd. If your thumb points toward a, V is positive.

A.2 Circumradii of Triangles and Tetrahedra

The following expression computes the radius r_{circ} of a circle in the plane that passes through the three points a, b, and c (also known as the *circumscribing circle*, or *circumcircle*, of the triangle abc).

$$r_{\text{circ}} = \frac{|b-c||c-a||a-b|}{2\begin{vmatrix} a_x - c_x & b_x - c_x \\ a_y - c_y & b_y - c_y \end{vmatrix}} = \frac{\ell_1 \ell_2 \ell_3}{4A}$$

where ℓ_1, ℓ_2, and ℓ_3 are the edge lengths of the triangle. Because square root operations are expensive and introduce error, the numerator above is best computed as $\sqrt{|b-c|^2|c-a|^2|a-b|^2}$.

This expression is numerically unstable when the denominator is close to zero (i.e. for triangles that are nearly degenerate). For quality measures in whose denominator r_{circ} appears, compute the reciprocal $1/r_{\text{circ}}$ to avoid the possibility of division by zero.

The following expression computes the circumradius of a two-dimensional triangle in E^3.

$$\begin{aligned} r_{\text{circ}} &= \frac{\left|\left[|a-c|^2(b-c) - |b-c|^2(a-c)\right] \times [(a-c)\times(b-c)]\right|}{2|(a-c)\times(b-c)|^2} \\ &= \frac{\left|\left[|\mathbf{r}|^2\mathbf{s} - |\mathbf{s}|^2\mathbf{r}\right] \times (\mathbf{r}\times\mathbf{s})\right|}{8A_f^2}. \end{aligned}$$

The following expression computes the radius of a sphere that passes through the four points a, b, c, and d (also known as the *circumscribing sphere*, or *circumsphere*, of the tetrahedron $abcd$).

$$\begin{aligned} r_{\text{circ}} &= \frac{\left||a-d|^2(b-d)\times(c-d) + |b-d|^2(c-d)\times(a-d) + |c-d|^2(a-d)\times(b-d)\right|}{2\begin{vmatrix} a_x-d_x & b_x-d_x & c_x-d_x \\ a_y-d_y & b_y-d_y & c_y-d_y \\ a_z-d_z & b_z-d_z & c_z-d_z \end{vmatrix}} \\ &= \frac{Z}{12V}, \end{aligned}$$

where $Z = ||\mathbf{t}|^2\mathbf{u}\times\mathbf{v} + |\mathbf{u}|^2\mathbf{v}\times\mathbf{t} + |\mathbf{v}|^2\mathbf{t}\times\mathbf{u}|$. (This is the same value of Z that appears in Table 8.)

This expression, like the two-dimensional formula, is numerically unstable when the denominator is close to zero (i.e. for tetrahedra that are nearly degenerate).

A.3 Min-Containment Radii of Triangles and Tetrahedra

Let O_{mc} and r_{mc} denote the center and radius of the smallest circle or sphere that encloses a triangle or tetrahedron.

To find the min-containment radius of a triangle, first determine whether the triangle has an angle of 90° or greater. This is quickly accomplished by checking the signs of the dot products of each pair of edge vectors. If one of the angles is 90° or greater, then O_{mc} is the midpoint of the opposite edge and r_{mc} is half its length (and there is no need to test the remaining angles). If all three angles are acute, the min-containment circle is the circumcircle and the min-containment radius is the circumradius.

To find a tetrahedron's min-containment sphere, first compute its circumcenter O_{circ}.

$$\begin{aligned} O_{\text{circ}} &= d + \frac{|a-d|^2(b-d)\times(c-d) + |b-d|^2(c-d)\times(a-d) + |c-d|^2(a-d)\times(b-d)}{2\begin{vmatrix} a_x-d_x & b_x-d_x & c_x-d_x \\ a_y-d_y & b_y-d_y & c_y-d_y \\ a_z-d_z & b_z-d_z & c_z-d_z \end{vmatrix}} \\ &= d + \frac{|\mathbf{t}|^2\mathbf{u}\times\mathbf{v} + |\mathbf{u}|^2\mathbf{v}\times\mathbf{t} + |\mathbf{v}|^2\mathbf{t}\times\mathbf{u}}{12V}. \end{aligned}$$

Next, test whether O_{circ} is in the tetrahedron $abcd$ by checking it against each triangular face. If the signed volumes of tetrahedra $O_{\text{circ}}bcd$, $O_{\text{circ}}adc$, $O_{\text{circ}}dab$, and $O_{\text{circ}}cba$ are all nonnegative, then $O_{\text{mc}} = O_{\text{circ}}$ and $r_{\text{mc}} = r_{\text{circ}}$. If exactly one of the volumes is negative, then r_{mc} is the min-containment radius of the corresponding triangular face. If two of the volumes are negative, then r_{mc} is the greater of the min-containment radii of two corresponding triangular faces.

References

[1] Nina Amenta, Marshall Bern, and David Eppstein. *Optimal Point Placement for Mesh Smoothing*. Proceedings of the Eighth Annual Symposium on Discrete Algorithms (New Orleans, Louisiana), pages 528–537. Association for Computing Machinery, January 1997.

[2] Thomas Apel. *Anisotropic Finite Elements: Local Estimates and Applications*. Technical Report SFB393/99-03, Technische Universität Chemnitz, January 1999. Habilitation.

[3] ———. *Anisotropic Finite Elements: Local Estimates and Applications*. B. G. Teubner Verlag, Stuttgart, 1999.

[4] Ivo Babuška and A. K. Aziz. *On the Angle Condition in the Finite Element Method*. SIAM Journal on Numerical Analysis **13**(2):214–226, April 1976.

[5] David H. Bailey. *A Portable High Performance Multiprecision Package*. Technical Report RNR-90-022, NASA Ames Research Center, Moffett Field, California, May 1993.

[6] T. J. Baker. *Element Quality in Tetrahedral Meshes*. Proceedings of the Seventh International Conference on Finite Element Methods in Flow Problems (Huntsville, Alabama), pages 1018–1024, April 1989.

[7] Eric B. Becker, Graham F. Carey, and John Tinsley Oden. *Finite Elements: An Introduction*. Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

[8] Marshall Bern, Herbert Edelsbrunner, David Eppstein, Scott Mitchell, and Tiow Seng Tan. *Edge-Insertion for Optimal Triangulations*. Discrete & Computational Geometry **10**:47–65, 1993.

[9] Marshall Bern and David Eppstein. *Mesh Generation and Optimal Triangulation*. Computing in Euclidean Geometry (Ding-Zhu Du and Frank Hwang, editors), Lecture Notes Series on Computing, volume 1, pages 23–90. World Scientific, Singapore, 1992.

[10] Martin Berzins. *A Solution-Based Triangular and Tetrahedral Mesh Quality Indicator*. SIAM Journal on Scientific Computing **19**(6):2051–2060, November 1998.

[11] ———. *A Solution Based H^1 Norm Triangular Mesh Quality Indicator*. Grid Generation and Adaptive Mathematics: Proceedings of the IMA Workshop on Parallel and Adaptive Methods (Marshall Bern, Joseph E. Flaherty, and Mitchell Luskin, editors), IMA Volumes in Mathematics and its Applications, volume 113, pages 77–89. Springer-Verlag, Berlin, 1999.

[12] R. P. Bhatia and K. L. Lawrence. *Two-Dimensional Finite Element Mesh Generation Based on Stripwise Automatic Triangulation*. Computers and Structures **36**:309–319, 1990.

[13] James H. Bramble and Miloš Zlámal. *Triangular Elements in the Finite Element Method*. Mathematics of Computation **24**(112):809–820, October 1970.

[14] Graham F. Carey and John Tinsley Oden. *Finite Elements: A Second Course*. Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[15] James C. Cavendish, David A. Field, and William H. Frey. *An Approach to Automatic Three-Dimensional Finite Element Mesh Generation*. International Journal for Numerical Methods in Engineering **21**(2):329–347, February 1985.

[16] L. Paul Chew. *Guaranteed-Quality Triangular Meshes*. Technical Report TR-89-983, Department of Computer Science, Cornell University, 1989.

[17] E. F. D'Azevedo. *Are Bilinear Quadrilaterals Better than Linear Triangles?* Technical Report ORNL/TM-12388, Computer Science and Mathematics Division, Oak Ridge National Laboratories, Oak Ridge, Tennessee, August 1993.

[18] Hugues L. de Cougny, Mark S. Shephard, and Marcel K. Georges. *Explicit Node Point Smoothing Within Octree*. Technical Report 10-1990, Scientific Computation Research Center, Rensselaer Polytechnic Institute, Troy, New York, 1990.

[19] Herbert Edelsbrunner, Tiow Seng Tan, and Roman Waupotitsch. *A Polynomial Time Algorithm for the Minmax Angle Triangulation*. SIAM Journal on Scientific and Statistical Computing **13**:994–1008, 1992.

[20] David A. Field. *Qualitative Measures for Initial Meshes*. International Journal for Numerical Methods in Engineering **47**:887–906, 2000.

[21] Lori A. Freitag, Mark Jones, and Paul Plassman. *An Efficient Parallel Algorithm for Mesh Smoothing*. Fourth International Meshing Roundtable (Albuquerque, New Mexico), pages 47–58. Sandia National Laboratories, October 1995.

[22] Lori A. Freitag and Carl Ollivier-Gooch. *Tetrahedral Mesh Improvement Using Swapping and Smoothing*. International Journal for Numerical Methods in Engineering **40**:3979–4002, 1997.

[23] Isaac Fried. *Condition of Finite Element Matrices Generated from Nonuniform Meshes*. AIAA Journal **10**(2):219–221, February 1972.

[24] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1989.

[25] Leonidas J. Guibas and Jorge Stolfi. *Primitives for the Manipulation of General Subdivisions and the Computation of Voronoi Diagrams*. ACM Transactions on Graphics **4**(2):74–123, April 1985.

[26] David C. Handscomb. *Errors of Linear Interpolation on a Triangle*. Manuscript, Oxford University Computing Laboratory, 1995.

[27] Magnus R. Hestenes and Eduard Stiefel. *Methods of Conjugate Gradients for Solving Linear Systems*. Journal of Research of the National Bureau of Standards **49**:409–436, 1952.

[28] P. Jamet. *Estimations d'Erreur pour des Élements Finit Drots Presque Dégénérés*. R. A. I. R. O. Anal. Numérique **10**:43–61, 1976.

[29] K. E. Jansen, M. S. Shephard, and M. W. Beall. *On Anisotropic Mesh Generation and Quality Control in Complex Flow Problems*. Tenth International Meshing Roundtable (Newport Beach, California), pages 341–349. Sandia National Laboratories, October 2001.

[30] Claes Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, New York, 1987.

[31] Mark T. Jones and Paul E. Plassmann. *Adaptive Refinement of Unstructured Finite-Element Meshes*. Finite Elements in Analysis and Design **25**:41–60, 1997.

[32] Patrick M. Knupp. *Matrix Norms & the Condition Number: A General Framework to Improve Mesh Quality via Node-Movement*. Eighth International Meshing Roundtable (Lake Tahoe, California), pages 13–22, October 1999.

[33] ______. *Achieving Finite Element Mesh Quality via Optimization of the Jacobian Matrix Norm and Associated Quantities. Part II: A Framework for Volume Mesh Optimization*. International Journal for Numerical Methods in Engineering **48**:1165–1185, 2000.

[34] Michal Křížek. *On Semiregular Families of Triangulations and Linear Interpolation*. Appl. Math. **36**:223–232, 1991.

[35] ______. *On the Maximum Angle Condition for Linear Tetrahedral Elements*. SIAM Journal on Numerical Analysis **29**(2):513–520, April 1992.

[36] Anwei Liu and Barry Joe. *Relationship between Tetrahedron Shape Measures*. BIT **34**:268–287, 1994.

[37] V. N. Parthasarathy, C. M. Graichen, and A. F. Hathaway. *Fast Evaluation & Improvement of Tetrahedral 3-D Grid Quality*. Manuscript, 1991.

[38] Mario Putti and Christian Cordes. *Finite Element Approximation of the Diffusion Operator on Tetrahedra*. SIAM Journal on Scientific Computing **19**(4):1154–1168, July 1998.

[39] V. T. Rajan. *Optimality of the Delaunay Triangulation in $\mathbb{R}^d$*. Proceedings of the Seventh Annual Symposium on Computational Geometry, pages 357–363, 1991.

[40] Jim Ruppert. *A Delaunay Refinement Algorithm for Quality 2-Dimensional Mesh Generation*. Journal of Algorithms **18**(3):548–585, May 1995.

[41] Youcef Saad and Martin H. Schultz. *GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*. SIAM Journal on Scientific and Statistical Computing **7**(3):856–869, July 1986.

[42] Michael I. Shamos and Dan Hoey. *Closest-Point Problems*. 16th Annual Symposium on Foundations of Computer Science (Berkeley, California), pages 151–162. IEEE Computer Society Press, October 1975.

[43] Jonathan Richard Shewchuk. *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*. Technical Report CMU-CS-94-125, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, March 1994.

[44] ______. *Adaptive Precision Floating-Point Arithmetic and Fast Robust Geometric Predicates*. Discrete & Computational Geometry **18**(3):305–363, October 1997.

[45] ______. *Tetrahedral Mesh Generation by Delaunay Refinement*. Proceedings of the Fourteenth Annual Symposium on Computational Geometry (Minneapolis, Minnesota), pages 86–95. Association for Computing Machinery, June 1998.

[46] ______. *Delaunay Refinement Algorithms for Triangular Mesh Generation*. Computational Geometry: Theory and Applications **22**(1–3):21–74, May 2002.

[47] Gilbert Strang and George J. Fix. *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

[48] J. L. Synge. *The Hypercircle in Mathematical Physics*. Cambridge University Press, New York, 1957.

[49] Shayne Waldron. *The Error in Linear Interpolation at the Vertices of a Simplex*. SIAM Journal on Numerical Analysis **35**(3):1191–1200, 1998.