Polynomial approximations and enriched finite element method, with applications.

Frederico Nudo

Polynomial approximations and enriched finite element method, with applications

Ph.D. Thesis Federico Nudo

Supervisors:
Prof. Francesco Dell'Accio
Prof. Allal Guessab

President of the jury:
Domingo Barrera

Student:
Dott. Federico Nudo

Contents

Sommario

Un problema ricorrente nell'ambito della scienza computazionale è la determinazione di un approssimante, in un intervallo fissato della retta reale, di una funzione di cui si conoscono solamente le valutazioni in un insieme finito di punti, detti nodi. Un approccio classico per risolvere questo problema si basa sull'interpolazione polinomiale. Di particolare interesse applicativo è il caso in cui i punti seguono una distribuzione equispaziata. In queste ipotesi ha luogo il fenomeno di Runge, che consiste in un aumento dell'errore di interpolazione in prossimità degli estremi dell'intervallo. Nel 2009, J. Boyd e F. Xu dimostrarono che il fenomeno di Runge poteva essere eliminato, interpolando la funzione solo su un sottoinsieme proprio dei punti dati, costituito dai nodi *più vicini* ai nodi di Chebyshev-Lobatto, i cosidetti nodi *mock–Chebyshev*. Tuttavia, per sua natura, questa strategia comporta il non utilizzo di quasi tutti i dati disponibili. Con l'obiettivo di migliorare l'accuratezza del metodo proposto da Boyd e Xu, utilizzando al contempo tutti i dati a disposizione, S. De Marchi, F. Dell'Accio e M. Mazza proposero una nuova tecnica, chiamata *approssimazione dei minimi quadrati mock-Chebyshev vincolata*. In questa tenica il ruolo del polinomio nodale, necessario per garantire l'interpolazione nei nodi mock-Chebyshev, è fondamentale. La sua generalizzazione al caso bivariato necessita, però, di approcci alternativi. La procedura recentemente introdotta da F. Dell'Accio, F. Di Tommaso e F. Nudo, basata sull'uso dei moltiplicatori di Lagrange, consente anche di definire l'approssimante dei minimi quadrati mock-Chebyshev vincolato su una griglia uniforme di punti. Questa nuova tecnica, equivalente alla tecnica univariata precedentemente introdotta in termini analitici, risulta anche più accurata in termini numerici. La prima parte della tesi è dedicata allo studio di questa nuova tecnica e alla sua applicazione a problemi di quadratura e differenziazione numerica.

La seconda parte di questa tesi si focalizza sullo sviluppo di un framework unificato e generale per l'arricchimento degli elementi finiti triangolari lineari in due dimensioni e simpliciali lineari in più dimensioni. Il metodo degli elementi finiti rappresenta una soluzione ampiamente adottata per risolvere numericamente equazioni alle derivate parziali presenti nei contesti di ingegneria e modellistica matematica [55]. La sua popolarità è attribuibile, in parte, alla sua versatilità nel gestire varie forme geometriche. Tuttavia, le approssimazioni prodotte da questo metodo, spesso non si rivelano efficaci nel risolvere problemi che presentano delle singolarità. Per superare questo ostacolo, sono stati proposti diversi approcci, tra cui uno dei più noti è l'*arricchimento* dello spazio di approssimazione dell'elemento finito mediante l'aggiunta di adeguate funzioni di arricchimento. Uno degli elementi finiti più semplici è l'elemento finito triangolare lineare standard. Quest'ultimo è largamente utilizzato nelle applicazioni. In questa tesi, introduciamo un arricchimento polinomiale dell'elemento finito triangolare lineare standard e utilizziamo questo nuovo elemento finito per introdurre un miglioramento dell'operatore triangolare di Shepard. In seguito, introduciamo una nuova classe di elementi finiti arricchendo l'elemento triangolare lineare standard con funzioni di arricchimento non necessariamente polinomiali, soddisfacenti la condizione di annullamento nei vertici. Successivamente, generalizziamo i risultati presentati nel caso bidimensionale, al caso dell'elemento finito simpliciale lineare standard utilizzando anche funzioni di arricchimento che non soddisfano la condizione di annullamento nei vertici. Infine, applichiamo queste nuove strategie di arricchimento per estendere l'arricchimento dell'elemento finito simpliciale lineare vettoriale sviluppato da Bernardi e Raugel.

Résumé

Un problème très courant en science computationnelle est la détermination d'une approximation, dans un intervalle fixe, d'une fonction dont les évaluations ne sont connues que sur un ensemble fini de points. Une approche courante pour résoudre ce problème repose sur l'interpolation polynomiale. Un cas d'un grand intérêt pratique est celui où ces points suivent une distribution équidistante dans l'intervalle considéré. Dans ces hypothèses, un problème lié à l'interpolation polynomiale est le phénomène de Runge, caractérisé par une augmentation de l'erreur d'interpolation près des extrémités de l'intervalle. En 2009, J. Boyd et F. Xu ont démontré que le phénomène de Runge pouvait être éliminé en interpolant la fonction que sur un sous-ensemble approprié formé par les noeuds les plus proches des noeuds de Chebyshev-Lobatto, communément appelés noeuds de *mock–Chebyshev*.

Cependant, cette stratégie implique de ne pas utiliser presque toutes les données disponibles. Afin d'améliorer la précision de la méthode proposée par Boyd et Xu, tout en tirant pleinement parti des données disponibles, S. De Marchi, F. Dell'Accio et M. Mazza ont introduit une nouvelle technique appelée *constrained mock-Chebyshev least squares approximation*. Dans cette méthode, le rôle du polynôme nodal, est crucial. Son extension au cas bivarié nécessite cependant des approches alternatives. La procédure développée par F. Dell'Accio, F. Di Tommaso et F. Nudo, utilisant la méthode des multiplicateurs de Lagrange, permet également de définir l'approximation des moindres carrés de mock-Chebyshev sur une grille uniforme de points. Cette technique innovante, équivalente à la méthode univariée précédemment introduite en termes analytiques, se révèle également plus précise en termes numériques. La première partie de la thèse est consacrée à l'étude de cette nouvelle technique et à son application à des problèmes de quadrature et de différenciation numérique.

Dans la deuxième partie de cette thèse, nous nous concentrons sur le développement d'un cadre unifié et général pour l'enrichissement de l'élément fini linéaire triangulaire standard en deux dimensions et de l'élément fini linéaire simplicial standard en dimensions supérieures. La méthode des éléments finis est une approche largement adoptée pour résoudre numériquement les équations aux dérivées partielles qui se posent en ingénierie et en modélisation mathématique [55]. Sa popularité est attribuable en partie à sa polyvalence pour traiter diverses formes géométriques. Cependant, les approximations produites par cette méthode s'avèrent souvent inefficaces pour résoudre des problèmes présentant des singularités. Pour surmonter ce problème, diverses approches ont été proposées, l'une des plus célèbres reposant sur l'*enrichissement* de l'espace d'approximation des éléments finis en ajoutant des fonctions d'enrichissement appropriées. Un des éléments finis le plus simple est l'élément fini triangulaire linéaire standard, largement utilisé dans les applications. Dans cette thèse, nous introduisons un enrichissement polynomial de l'élément fini triangulaire linéaire standard et utilisons ce nouvel élément fini pour introduire une amélioration de l'opérateur triangulaire de Shepard. Ensuite, nous introduisons une nouvelle classe d'éléments finis en enrichissant l'élément triangulaire linéaire standard avec des fonctions d'enrichissement qui ne sont pas nécessairement polynomiales, mais qui satisfont la condition d'annulation aux sommets du triangle.

Nous généralisons les résultats présentés dans le cas bidimensionnel au cas de l'élément fini simplicial linéaire standard, en utilisant également des fonctions d'enrichissement qui ne satisfont pas la condition d'annulation aux sommets du simplexe.

Enfin, nous appliquons ces nouvelles stratégies d'enrichissement pour définir une version plus générale de l'enrichissement de l'élément fini linéaire vectoriel simplicial développé par Bernardi et Raugel.

Abstract

A very common problem in computational science is the determination of an approximation, in a fixed interval, of a function whose evaluations are known only on a finite set of points. A common approach to solving this problem relies on polynomial interpolation, which consists of determining a polynomial that coincides with the function at the given points. A case of great practical interest is the case where these points follow an equispaced distribution within the considered interval. In these hypotheses, a problem related to polynomial interpolation is the Runge phenomenon, which consists in increasing the magnitude of the interpolation error close to the ends of the interval. In 2009, J. Boyd and F. Xu demonstrated that the Runge phenomenon could be eliminated by interpolating the function only on a proper subset formed by nodes *closest* to the Chebyshev-Lobatto nodes, the so called *mock–Chebyshev* nodes. However, this strategy involves not using almost all available data. In order to improve the accuracy of the method proposed by Boyd and Xu, while making full use of the available data, S. De Marchi, F. Dell'Accio, and M. Mazza introduced a new technique known as the *constrained mock-Chebyshev least squares approximation*. In this method, the role of the nodal polynomial, essential for ensuring interpolation at mock-Chebyshev nodes, is crucial. Its extension to the bivariate case, however, requires alternative approaches. The recently developed procedure by F. Dell'Accio, F. Di Tommaso, and F. Nudo, employing the Lagrange multipliers method, also enables the definition of the constrained mock-Chebyshev least squares approximation on a uniform grid of points. This innovative technique, equivalent to the previously introduced univariate method in analytical terms, also proves to be more accurate in numerical terms. The first part of the thesis is dedicated to the study of this new technique and its application to numerical quadrature and differentiation problems.

In the second part of this thesis, we focus on the development of a unified and general framework for the enrichment of the standard triangular linear finite element in two dimensions and the standard simplicial linear finite element in higher dimensions. The finite element method is a widely adopted approach for numerically solving partial differential equations arising in engineering and mathematical modeling [55]. Its popularity is partly attributed to its versatility in handling various geometric shapes. However, the approximations produced by this method often prove ineffective in solving problems with singularities. To overcome this issue, various approaches have been proposed, with one of the most famous relying on the *enrichment* of the finite element approximation space by adding suitable enrichment functions. One of the simplest finite elements is the standard linear triangular element, widely used in applications. In this thesis, we introduce a polynomial enrichment of the standard triangular linear finite element and use this new finite element to introduce an improvement of the triangular Shepard operator. Subsequently, we introduce a new class of finite elements by enriching the standard triangular linear finite element with enrichment functions that are not necessarily polynomials, which satisfy the vanishing condition at the vertices of the triangle.

Later on, we generalize the results presented in the two dimensional case to the case of the standard simplicial linear finite element, also using enrichment functions that do not satisfy the vanishing condition at the vertices of the simplex.

Finally, we apply these new enrichment strategies to extend the enrichment of the simplicial vector linear finite element developed by Bernardi and Raugel.

Introduzione

Un problema molto frequente nella scienza computazionale è la determinazione di un approssimante, in un intervallo fissato $[a, b]$, di una funzione f di cui si conoscono solamente le valutazioni su un insieme di $n+1$ punti X_n, $n \in \mathbb{N}$. A meno di traslazioni e omotetie, possiamo supporre di lavorare nell'intervallo $[-1, 1]$. Un approccio standard per risolvere questo problema è utilizzare l'interpolazione polinomiale, cioè determinare un polinomio $P_n[f]$, di grado al più n, che coincida con la funzione f nei punti dell'insieme X_n. Un caso di grande interesse pratico è il caso in cui l'insieme X_n coincida con l'insieme dei punti equispaziati nell'intervallo $[-1, 1]$. In queste ipotesi, un problema relativo all'interpolazione polinomiale è il fenomeno di Runge. Questo consiste nell'aumento di ampiezza dell'errore di interpolazione in prossimità degli estremi dell'intervallo considerato $[-1, 1]$. Fu scoperto nei primi anni del 900 da Carl David Tolmé Runge mentre studiava l'andamento degli errori di interpolazione polinomiale per approssimare alcune funzioni [89]. Per ovviare a questo problema, negli anni sono state proposte diverse tecniche, vedi per esempio [10, 5, 24, 26]. In particolare, nel 2009, John P. Boyd e Fei Xu nell'articolo [10], dimostrarono che il fenomeno di Runge può essere completamente eliminato se si decide di interpolare la funzione f solamente su un sottoinsieme proprio dell'insieme X_n, costituito da $m+1 = \mathcal{O}(\sqrt{n})+1$ nodi che *sono vicini* ai nodi di Chebyshev-Lobatto di ordine $m+1$, i cosidetti nodi *mock–Chebyshev*. Utilizzando questa strategia, però, molti dati che si hanno a disposizione non vengono utilizzati, e dunque motivati da questo, Stefano De Marchi, Francesco Dell'Accio e Mariarosa Mazza, in [24], proposero una nuova tecnica, con l'intento di migliorare l'accuratezza dell'approssimante introdotto in [10]. Questa tecnica viene detta tecnica dell'approssimazione *dei minimi quadrati mock-Chebyshev vincolata*. Più precisamente, l'idea proposta in [24] consiste nell'approssimare la funzione f con un polinomio di grado $r > m$ che si ottiene interpolando f sull'insieme dei nodi mock-Chebyshev e sfruttando i nodi rimanenti per migliorare l'accuratezza dell'approssimazione attraverso una regressione simultanea. In questa tenica, come vedremo nel Capitolo 1, il ruolo del polinomio nodale è fondamentale. A causa della mancanza di generalizzazioni del polinomio nodale nella teoria dell'interpolazione bivariata, c'è quindi la necessità di trovare approcci alternativi che consentano generalizzazioni al caso bivariato dell'approssimazione dei minimi quadrati mock-Chebyshev vincolata. Una nuova procedura, introdotta da Francesco Dell'Accio, Filomena Di Tommaso e Federico Nudo nel 2022 [39], permette di definire, sotto le stesse condizioni, lo stesso approssimante introdotto in [24], utilizzando il metodo dei moltiplicatori di Lagrange. Questa tecnica, indipendente dal polinomio nodale, è stata sviluppata da Joseph-Louis Lagrange per risolvere problemi di minimi quadrati vincolati. Una volta fissata una base dello spazio dei polinomi di grado r, l'approssimante viene calcolato risolvendo delle equazioni lineari, che generalizzano le equazioni normali e che costituiscono il sistema lineare KKT (questo nome è dovuto ai tre ricercatori William Karush, Harold Kuhn, and Albert Tucker, che furono i primi ad introdurlo). Nel Capitolo 2 vengono proposte due diverse strategie per ottenere questa generalizzazione:

- la prima, basata sull'interpolazione prodotto tensore su una griglia cartesiana di nodi mock-Chebyshev,
- la seconda, basata sull'interpolazione di grado totale su un insieme di nodi che *sono vicini* ai ben noti nodi di Padova [9].

Nelle stesse ipotesi di cui sopra, un altro problema di grande interesse pratico è la determinazione di formule di quadratura accurate su nodi equispaziati. Questo è stato oggetto di studio di diversi autori negli ultimi anni, vedi per esempio [43, 58, 56, 68, 60, 71, 77]. In particolare, l'approccio proposto da Hassan Majidian in [77] consiste nell'utilizzare rispettivamente le formule di quadratura di Gauss-Christoffel in combinazione con interpolanti polinomiali locali della funzione f sui punti considerati.

Motivati da questo, nel Capitolo 3, introduciamo una formula di quadratura stabile ed accurata su nodi equispaziati che utilizza la formula di quadratura di Gauss–Christoffel e l'approssimazione dei minimi quadrati mock-Chebyshev vincolata. Inoltre, poiché l'accuratezza della formula di quadratura dipende dalla classe di differenziabilità della funzione di cui vogliamo approssimare l'integrale, in questo capitolo sviluppiamo anche un algoritmo adattivo per calcolare il *grado di regressione ottimale* il quale corrisponde alla formula di quadratura più accurata. In [40] utilizziamo l'approssimazione dei minimi quadrati mock-Chebyshev vincolata per introdurre una formula di quadratura di tipo prodotto. Questo tipo di formule di quadratura vengono utilizzate per approssimare integrali definiti su intervalli finiti di funzioni integrande che presentano una "patologia", ad esempio funzioni "quasi" singolari, funzioni altamente oscillanti, funzioni debolmente singolari. Nel Capitolo 4 studiamo ulteriormente alcune proprietà dell'operatore dei minimi quadrati mock-Chebyshev vincolato, e forniamo nuovi risultati e applicazioni. In particolare, introduciamo rappresentazioni puntuali esplicite dell'errore e delle sue derivate, assumendo che la funzione f sia sufficientemente regolare in $[-1, 1]$. Nelle stesse ipotesi, come applicazione, introduciamo un nuovo metodo di derivazione basato su una griglia di punti equispaziati per approssimare le derivate successive della funzione f con le derivate dell'approssimante dei minimi quadrati mock-Chebyshev vincolato.

La seconda parte di questa tesi è dedicata allo studio degli arricchimenti di alcuni elementi finiti standard. Un elemento finito è una terna $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$, dove

- K_d è un politopo in $\mathbb{R}^d$,
- $\mathbb{F}_{K_d}$ è uno spazio vettoriale di dimensione finita n constituito da funzioni a valori reali definite su K_d, anche dette *trial functions*,
- $\Sigma_{K_d} = \{L_j : j = 1, \dots, n\}$ è un insieme di funzionali lineari, linearmente indipendenti sullo spazio vettoriale $\mathbb{F}_{K_d}$, detti anche *gradi di libertà*,

tale che lo spazio di approssimazione $\mathbb{F}_{K_d}$ sia Σ_{K_d}-unisolvente, cioè se $f \in \mathbb{F}_{K_d}$ e

$$L_j(f) = 0, \qquad j = 1, \dots, n,$$

allora $f = 0$ [55].

Il metodo degli elementi finiti è un metodo molto utilizzato per risolvere numericamente equazioni alle derivate parziali su un dominio $D \subset \mathbb{R}^d$ [13, 19], $d \geq 1$, che sorgono in ingegneria e modellistica matematica. Uno dei motivi di questa innegabile popolarità è la sua versatilità nell'affrontare diversi tipi di geometrie. L'idea di questo metodo è quella di partizionare il dominio $\bar{D}$ in politopi e, per ognuno di essi determinare un'approssimazione locale della soluzione del problema differenziale considerato mediante una funzione appartenente a $\mathbb{F}_{K_d}$. L'approssimazione globale della soluzione sarà una funzione definita a tratti dalle approssimazioni locali. Nel caso in cui l'approssimazione globale presenti discontinuità sui bordi dei sottodomini, l'elemento finito si dirà essere non conforme, altrimenti esso si dirà essere conforme. Negli elementi finiti standard lo spazio $\mathbb{F}_{K_d}$ è generalmente uno spazio di funzioni polinomiali. Tuttavia, le approssimazioni prodotte da questi elementi finiti non sono efficaci per risolvere problemi che presentano delle singolarità. In modo da superare questo problema sono stati proposti diversi approcci. Uno degli approcci più famosi consiste nell'*arricchire* lo spazio di approssimazione $\mathbb{F}_{K_d}$ con funzioni di arricchimento appropriate [92, 45, 64]. In particolare, supponiamo di avere l'elemento finito $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$, un insieme di funzioni di arricchimento $e_1, \dots, e_N$ e un insieme di funzionali lineari

$$\Sigma_{K_d}^{\text{enr}} = \{L_j : j = 1, \dots, n + N\}$$

tale che lo spazio di approssimazione $\mathbb{F}_{K_d}$ sia $\Sigma_{K_d}^{\text{enr}}$-unisolvente. Consideriamo lo spazio arricchito

$$\mathbb{F}_{K_d}^{\text{enr}} = \mathbb{F}_{K_d} \oplus \{e_1, \dots, e_N\}.$$

Per arricchire l'elemento finito $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$ bisogna rispondere alla seguente domanda:

Come scegliere opportunamente le funzioni di arricchimento $e_1, \dots, e_N$, in modo tale che la terna $(K_d, \mathbb{F}_{K_d}^{\text{enr}}, \Sigma_{K_d}^{\text{enr}})$ sia un elemento finito?

Uno degli elementi finiti più semplici è l'elemento finito triangolare lineare standard, definito nello spazio euclideo bidimensionale [19]. È definito come la tripla

$$\mathcal{P}_1(S_2) = (S_2, \mathbb{P}_1(S_2), \Sigma_{S_2}^{\text{lin}}),$$

dove S_2 è un triangolo non degenere con vertici $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\mathbb{P}_1(S_2)$ è lo spazio dei polinomi lineari bivariati e $\Sigma_{S_2}^{\text{lin}}$ è l'insieme delle valutazioni funzionali nei vertici di S_2. L'elemento finito triangolare lineare standard è largamente utilizzato nelle applicazioni, tuttavia non sempre produce risultati soddisfacenti a causa del basso ordine di approssimazione delle relative trial functions. Per migliorare l'accuratezza dell'approssimazione, l'elemento finito triangolare lineare standard $\mathcal{P}_1(S_2)$ può essere arricchito con funzioni di arricchimento particolari (per una panoramica della letteratura pertinente si veda, ad esempio, [2, 64, 3, 4]). Più precisamente, nella seconda parte della tesi, ci concentriamo sullo sviluppo di un framework unificato e generale per l'arricchimento degli elementi finiti triangolari lineari in $\mathbb{R}^2$ e simpliciali lineari in $\mathbb{R}^d$. Come abbiamo già detto, un punto cruciale in tale approccio è determinare le condizioni sulle funzioni di arricchimento affinché esse possano generare un elemento finito. Motivati dai recenti lavori sugli arricchimenti degli elementi finiti [2, 4], nel Capitolo 5 introduciamo un arricchimento polinomiale dell'elemento finito triangolare lineare standard e utilizziamo questo nuovo elemento finito per proporre un miglioramento dell'operatore di approssimazione triangolare di Shepard (vedi [47, 101, 22, 31, 33] per altri approcci). Nel Capitolo 6 introduciamo una nuova classe di elementi finiti non conformi arricchendo l'elemento triangolare lineare standard con funzioni di arricchimento continue, linearmente indipendenti, non necessariamente polinomiali $\{e_i \,:\, i = 0, 1, 2\}$, soddisfacenti la condizione di annullamento nei vertici, cioè $e_i(\boldsymbol{v}_j) = 0$, $i, j = 0, 1, 2$. Inoltre, determiniamo una condizione sulle funzioni di arricchimento, necessaria e sufficiente, affinché esse possano generare un elemento finito. Mostriamo che l'errore di approssimazione può essere decomposto in due parti: la prima relativa all'elemento triangolare lineare standard mentre la seconda dipendente dalle funzioni di arricchimento. Questa decomposizione ci permette di ricavare dei bound per l'errore in norma L^∞ e in norma L^1. Questi bound sono proporzionali rispettivamente alla seconda e alla quarta potenza del raggio del circocentro del triangolo corrispondente. Nel Capitolo 7 generalizziamo i risultati presenti nel Capitolo 6 al caso dell'elemento finito simpliciale lineare standard. Questo è definito come

$$\mathcal{P}_1(S_d) = (S_d, \mathbb{P}_1(S_d), \Sigma_{S_d}^{\text{lin}})$$

dove S_d è un simplesso non degenere in $\mathbb{R}^d$ con vertici $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$, $\mathbb{P}_1(S_d)$ è lo spazio dei polinomi lineari in $\mathbb{R}^d$ e $\Sigma_{S_d}^{\text{lin}}$ è l'insieme dei funzionali valutazione nei vertici di S_d. In analogia con quanto fatto nel caso bivariato, arricchiamo l'elemento finito simpliciale lineare standard $\mathcal{P}_1(S_d)$ con $d+1$ funzioni continue linearmente indipendenti $\{e_i \,:\, i = 0, \ldots, d\}$, soddisfacenti la condizione di annullamento nei vertici di S_d, cioè $e_i(\boldsymbol{v}_j) = 0$, $i, j = 0, \ldots, d$. Nel Capitolo 8 forniamo una strategia generale per arricchire l'elemento finito simpliciale lineare standard $\mathcal{P}_1(S_d)$ con funzioni di arricchimento generiche, cioè senza imporre condizioni restrittive su quest'ultime, come il loro annullamento nei vertici. Un elemento finito comunemente utilizzato nelle applicazioni è l'elemento finito simpliciale lineare vettoriale. Questo è definito come

$$\boldsymbol{\mathcal{P}}_1(S_d) = (S_d, \mathbb{P}_1(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{lin}})$$

dove $\mathbb{P}_1(S_d)$ è il prodotto diretto, d volte, dello spazio vettoriale $\mathbb{P}_1(S_d)$ con se stesso e

$$\boldsymbol{\Sigma}_{S_d}^{\text{lin}} = \{\boldsymbol{L}_j \,:\, j = 0, \ldots, d\},$$

con $\boldsymbol{L}_j$ definito come

$$\boldsymbol{L}_j(\boldsymbol{f}) = \boldsymbol{f}(\boldsymbol{v}_j) = [f_1(\boldsymbol{v}_j), \ldots, f_d(\boldsymbol{v}_j)]^T, \quad \boldsymbol{f} = [f_1, \ldots, f_d]^T, \quad j = 0, \ldots, d.$$

Questo elemento finito è comunemente usato per risolvere numericamente le equazioni di Stokes stazionarie. Tuttavia, è un fatto noto che l'elemento finito $\boldsymbol{\mathcal{P}}_1(S_d)$ diventa inefficente quando viene applicato a problemi più complicati. Un arricchimento polinomiale, che supera i suddetti inconvenienti, è stato proposto e sviluppato da Bernardi e Raugel in [7]. Questo elemento, molto utilizzato in contesti pratici, può essere considerato una versione avanzata e generalizzata dell'elemento finito simpliciale lineare vettoriale. Tuttavia, quando si trattano problemi con singolarità, questo arricchimento risulta essere inefficace.

Dunque, in linea con le ricerche precedenti, nel Capitolo 9, presentiamo una strategia generale per arricchire l'elemento finito simpliciale lineare vettoriale mediante funzioni di arricchimento non polinomiali. Questo elemento finito arricchito può essere considerato come un'estensione dell'elemento di Bernardi e Raugel.

Introduction

Un problème très courant en sciences computationnelles est la détermination d'une approximation, dans un intervalle fixé $[a, b]$, d'une fonction f dont nous ne connaissons que les évaluations sur un ensemble de $n+1$ points X_n, $n \in \mathbb{N}$. Nous pouvons supposer travailler dans l'intervalle de référence $[-1, 1]$, sauf en cas de transformations linéaires. Une approche standard pour résoudre ce problème est l'interpolation polynomiale, qui consiste à déterminer un polynôme $P_n[f]$, de degré n, qui coïncide avec f en X_n. Un cas d'un grand intérêt pratique est lorsque l'ensemble X_n coïncide avec l'ensemble de points équidistants dans l'intervalle $[-1, 1]$. Dans ces conditions, un problème lié à l'interpolation polynomiale est le phénomène de Runge. Il se manifeste par une augmentation de l'erreur d'interpolation à proximité des extrémités de l'intervalle considéré, c'est-à-dire dans $[-1, 1]$. Ce phénomène a été découvert au début du XXe siècle par Carl David Tolmé Runge lors de ses études sur le comportement des erreurs produites par l'interpolation polynomiale pour l'approximation de certaines fonctions [89]. Pour résoudre ce problème, ces dernières années, plusieurs techniques ont été proposées, comme par exemple [10, 5, 24, 26]. En particulier, en 2009, John P. Boyd et Fei Xu dans [10] ont démontré que le phénomène de Runge peut être surmonté si l'on interpole la fonction f uniquement sur un sous-ensemble propre de l'ensemble X_n, constitué de $m+1 = \mathcal{O}(\sqrt{n})+1$ noeuds appelés les noeuds mock-Chebyshev, qui sont proches des $m+1$ noeuds de Chebyshev-Lobatto. Cependant, en utilisant cette stratégie, de nombreuses données ne sont pas utilisées. C'est pourquoi Stefano De Marchi, Francesco Dell'Accio et Mariarosa Mazza, dans [24], ont proposé une nouvelle technique pour améliorer la précision de l'approximation introduite dans [10]. Cette technique est appelée *constrained mock-Chebyshev least squares approximation*. Elle consiste à approcher la fonction f avec un polynôme de degré $r > m$, obtenu en interpolant f sur l'ensemble des points mock-Chebyshev, tout en utilisant les noeuds restants pour améliorer la précision de l'approximation grâce à une régression simultanée. Pour cette technique, comme nous le verrons au Chapitre 1, le rôle du polynôme nodal est crucial. En raison de l'absence de polynômes nodaux dans le cas bivarié, il est nécessaire de trouver des approches alternatives pour généraliser l'approximation constrained mock-Chebyshev least squares au cas bivarié. Une nouvelle technique, introduite par Francesco Dell'Accio, Filomena Di Tommaso et Federico Nudo en 2022 [39], permet de définir la même approximation introduite dans [24], en utilisant la méthode des multiplicateurs de Lagrange. Cette méthode, indépendante des polynômes nodaux, a été développée par Joseph-Louis Lagrange pour résoudre les problèmes des moindres carrés contraints. En fixant une base de l'espace polynomial de degré r, l'approximation est calculée en résolvant des équations linéaires qui généralisent les équations normales et qui constituent le système linéaire KKT (nommé ainsi d'après les trois chercheurs William Karush, Harold Kuhn et Albert Tucker, qui l'ont présenté en premier).

Dans le Chapitre 2, deux stratégies différentes sont proposées pour obtenir cette généralisation:

- La première repose sur l'interpolation du produit tensoriel sur une grille cartésienne de noeuds mock-Chebyshev.
- La seconde repose sur l'interpolation du degré total sur un ensemble de noeuds qui *sont proches* des noeuds de Padoue [9].

Un autre problème d'un grand intérêt pratique est la détermination de formules de quadrature précises sur des noeuds équidistants. Ce sujet a été étudié par plusieurs auteurs ces dernières années, comme on peut le voir dans [43, 58, 56, 68, 60, 71, 77]. En particulier, l'approche proposée par Hassan Majidian dans [77] consiste à utiliser les formules de quadrature de Gauss-Christoffel en combinaison avec des interpolants polynomiaux locaux de la fonction à intégrer f. En suivant cette idée, dans le Chapitre 3, nous introduisons une formule de quadrature stable et précise sur des noeuds équidistants en utilisant la formule

de quadrature de Gauss-Christoffel et l'approximation des moindres carrés de mock-Chebyshev. Étant donné que la précision de ces formules de quadrature dépend du degré de régression de l'approximation des moindres carrés de mock-Chebyshev et du degré de régularité de la fonction f, nous développons un algorithme adaptatif pour déterminer le degré optimal de régression correspondant à la formule de quadrature la plus précise. En tant qu'application supplémentaire, dans [40], nous utilisons l'approximation des moindres carrés de mock-Chebyshev pour introduire une nouvelle règle d'intégration de produits. Ce type de formule est utilisé pour l'approximation des intégrales définies sur des intervalles finis de fonctions intégrales, qui ne sont pas assez *régulières*, telles que des fonctions presque singulières, des fonctions très oscillantes ou des fonctions faiblement singulières.

Dans le Chapitre 4, nous approfondissons l'étude de certaines propriétés de l'opérateur d'approximation des moindres carrés de mock-Chebyshev et présentons de nouveaux résultats et applications. En particulier, nous introduisons des représentations explicites de l'erreur et de ses dérivées, en supposant que f est suffisamment régulière sur l'intervalle $[-1, 1]$. Dans la même hypothèse, en tant qu'application, nous présentons une méthode pour estimer les dérivées successives de f en n'importe quel point $x \in [-1, 1]$, en se basant sur l'approximation des moindres carrés de mock-Chebyshev et utilisons les représentations d'erreurs précédemment introduites pour fournir des estimations pour ces approximations.

La deuxième partie de cette thèse est dédiée à l'étude de l'enrichissement de certains éléments finis linéaires standard. Un élément fini est un triplet $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$, où

- K_d est un polytope dans $\mathbb{R}^d$,
- $\mathbb{F}_{K_d}$ est un espace vectoriel de dimension n composé de fonctions à valeurs réelles définies sur K_d, également appelées *fonctions d'essai ou fonctions test*,
- $\Sigma_{K_d} = \{L_j \,:\, j = 1, \ldots, n\}$ est un ensemble de fonctionnelles linéaires indépendantes de l'espace vectoriel $\mathbb{F}_{K_d}$, aussi appelé *degrés de liberté*,

de manière à ce que $\mathbb{F}_{K_d}$ soit Σ_{K_d}-unisolvant, c'est-à-dire que si $f \in \mathbb{F}_{K_d}$ et

$$L_j(f) = 0, \qquad j = 1, \ldots, n,$$

alors $f = 0$ [55].

La méthode des éléments finis est une méthode très populaire pour résoudre numériquement des équations aux dérivées partielles sur un domaine $D \subset \mathbb{R}^d$ [13, 19], avec $d \geq 1$. Elle est couramment utilisée en ingénierie et en modélisation mathématique. Une des raisons indiscutables de cette popularité réside dans sa polyvalence pour traiter différents types de géométries. Dans la méthode des éléments finis, le domaine $\bar{D}$ est subdivisé en polytopes, et pour chacun d'entre eux, une approximation locale appartenant à $\mathbb{F}_{K_d}$ est calculée pour estimer la solution de l'équation aux dérivées partielles. L'approximation globale sera une fonction définie par les approximations locales. Si l'approximation globale présente des discontinuités aux limites des sous-domaines, l'élément fini est qualifié de non conforme, sinon il est dit conforme. En général, pour les éléments finis linéaires standards, l'espace d'approximation $\mathbb{F}_{K_d}$ est constitué de fonctions polynomiales. Cependant, les approximations produites par ces éléments finis ne sont pas efficaces pour résoudre des problèmes impliquant des singularités. Pour surmonter ce problème, plusieurs approches ont été proposées. L'une des approches les plus célèbres consiste à *enrichir* l'espace d'approximation $\mathbb{F}_{K_d}$ avec des fonctions d'enrichissement appropriées [92, 45, 64]. Plus précisément, nous considérons l'élément fini $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$, un ensemble de fonctions d'enrichissement $e_1, \ldots, e_N$ et un ensemble de fonctionnelles linéaires

$$\Sigma_{K_d}^{\text{enr}} = \{L_j \,:\, j = 1, \ldots, n + N\}$$

de telle manière que l'espace d'approximation $\mathbb{F}_{K_d}$ soit $\Sigma_{K_d}^{\text{enr}}$-unisolvent. Nous définissons l'espace enrichi

$$\mathbb{F}_{K_d}^{\text{enr}} = \mathbb{F}_{K_d} \oplus \{e_1, \ldots, e_N\}.$$

Pour enrichir l'élément fini $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$ il faut répondre à la question suivante:

Comment choisir correctement les fonctions d'enrichissement $e_1, \ldots, e_N$, de telle sorte que le triplet $(K_d, \mathbb{F}^{\text{enr}}_{K_d}, \Sigma^{\text{enr}}_{K_d})$ soit un élément fini?

L'un des éléments finis le plus simple est l'élément fini linéaire triangulaire standard, défini dans l'espace euclidien bidimensionnel [19]. Il est défini comme le triple

$$\mathcal{P}_1(S_2) = (S_2, \mathbb{P}_1(S_2), \Sigma^{\text{lin}}_{S_2}),$$

où S_2 est un triangle non dégénéré (les trois sommets ne sont pas alignés) et avec des sommets $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\mathbb{P}_1(S_2)$ est l'espace de tous les polynômes linéaires bivariés et $\Sigma^{\text{lin}}_{S_2}$ est l'ensemble des fonctionnelles linéaires associant à chaque fonction son évaluation aux sommets de S_2. L'élément fini triangulaire linéaire standard est largement utilisé dans les applications, cependant, il ne produit pas toujours des résultats satisfaisants en raison du faible ordre d'approximation des fonctions d'essai. Pour améliorer la précision de l'approximation, l'élément $\mathcal{P}_1(S_2)$ peut être enrichi de fonctions d'enrichissement spéciales (pour un aperçu de la littérature pertinente, voir, par exemple, [2, 64, 3, 4]). Plus précisément, dans la deuxième partie de la thèse, nous nous concentrons sur le développement d'un cadre unifié et général pour enrichir les éléments finis linéaires triangulaires standards en $\mathbb{R}^2$ et les éléments finis linéaires simpliciaux standards en $\mathbb{R}^d$. Comme mentionné précédemment, un aspect crucial de cette approche consiste à déterminer les conditions requises pour les fonctions d'enrichissement de manière à ce qu'elles génèrent un élément fini. Motivés par des travaux récents sur l'enrichissement des éléments finis [2, 4], dans le Chapitre 5, nous introduisons un enrichissement polynomial de l'élément fini linéaire triangulaire standard et utilisons ce nouvel élément fini pour améliorer l'opérateur triangulaire Shepard (voir [47, 101, 22, 31, 33] pour d'autres approches). Dans le Chapitre 6, nous introduisons une nouvelle classe d'éléments finis non conformes en enrichissant l'élément fini linéaire triangulaire standard avec des fonctions d'enrichissement continues, linéairement indépendantes, qui ne sont pas nécessairement des polynômes $\{e_i : i = 0, 1, 2\}$, tout en satisfaisant la condition d'annulation aux sommets du triangle, c'est-à-dire $e_i(\boldsymbol{v}_j) = 0$, $i, j = 0, 1, 2$. De plus, nous déterminons une condition nécessaire et suffisante pour les fonctions d'enrichissement de manière à générer un élément fini. Nous montrons que l'erreur d'approximation peut être décomposée en deux parties: la première est liée à l'élément fini triangulaire linéaire standard, tandis que la seconde dépend des fonctions d'enrichissement. Cette décomposition nous permet d'obtenir des bornes de l'erreur à la fois pour la norme L^∞ et aussi pour la norme L^1. Ces limites sont proportionnelles aux deuxième et quatrième puissances du rayon du cercle circonscrit au triangle correspondant, respectivement. Dans le Chapitre 7, nous généralisons les résultats présentés dans le Chapitre 6 au cas de l'élément fini linéaire simplicial standard. Il est défini comme suit:

$$\mathcal{P}_1(S_d) = (S_d, \mathbb{P}_1(S_d), \Sigma^{\text{lin}}_{S_d})$$

où S_d est un simplexe en $\mathbb{R}^d$ ayant un volume non nul et des sommets $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$, $\mathbb{P}_1(S_d)$ est l'espace de tous les polynômes linéaires en $\mathbb{R}^d$, et $\Sigma^{\text{lin}}_{S_d}$ est l'ensemble des fonctionnelles linéaires associant à chaque fonction son évaluation aux sommets de S_d. Conformément au Chapitre 6, nous enrichissons l'élément fini linéaire simplicial standard $\mathcal{P}_1(S_d)$ avec $d+1$ fonctions continues linéairement indépendantes $\{e_i : i = 0, \ldots, d\}$, satisfaisant la condition d'annulation aux sommets du S_d, c'est-à-dire $e_i(\boldsymbol{v}_j) = 0$, $i, j = 0, \ldots, d$. Dans le Chapitre 8, nous fournissons une stratégie générale pour enrichir l'élément fini linéaire simplicial standard $\mathcal{P}_1(S_d)$ sans imposer de conditions restrictives aux fonctions d'enrichissement, telles que leur annulation aux sommets de S_d. Un élément fini couramment utilisé dans les applications est l'élément fini linéaire vectoriel simplicial. Il est défini comme suit:

$$\boldsymbol{\mathcal{P}}_1(S_d) = (S_d, \boldsymbol{\mathbb{P}}_1(S_d), \boldsymbol{\Sigma}^{\text{lin}}_{S_d})$$

où $\boldsymbol{\mathbb{P}}_1(S_d)$ est le produit direct, d fois, de l'espace vectoriel $\mathbb{P}_1(S_d)$ avec lui-même et

$$\boldsymbol{\Sigma}^{\text{lin}}_{S_d} = \{\boldsymbol{L}_j : j = 0, \ldots, d\},$$

avec $\boldsymbol{L}_j$ définies comme

$$\boldsymbol{L}_j(\boldsymbol{f}) = \boldsymbol{f}(\boldsymbol{v}_j) = [f_1(\boldsymbol{v}_j), \ldots, f_d(\boldsymbol{v}_j)]^T, \quad \boldsymbol{f} = [f_1, \ldots, f_d]^T, \quad j = 0, \ldots, d.$$

Cet élément fini est couramment utilisé pour résoudre numériquement les équations de Navier-Stokes stationnaires. Il est connu, cependant, de souffrir de graves lacunes dans l'application à des situations plus compliquées. Un élément fini enrichi, qui surmonte les inconvénients susmentionnés, a été proposé et développé par Bernardi et Raugel [7]. Il peut être considéré comme une version avancée et généralisée de l'élément fini linéaire vectoriel simplicial classique, et il a été utilisé dans un large éventail de domaines de calcul d'ingénierie pratique. Il utilise des polynômes comme fonctions d'enrichissement. Toutefois, le problème de dépendance linéaire insoluble se pose toujours lorsque ce type de fonctions d'enrichissement est utilisé. Conformément aux recherches précédentes, dans le Chapitre 9, nous présentons une stratégie générale pour enrichir l'élément fini linéaire vectoriel simplicial par des fonctions d'enrichissement non polynomiales. Cet élément fini enrichi est défini par rapport à tout simplexe, et peut être considéré comme une extension de l'élément de Bernardi et Raugel.

Introduction

A very common problem in computational sciences is the determination of an approximation, in a fixed interval $[a, b]$, of a function f whose evaluations are known only on a set of $n+1$ points X_n, $n \in \mathbb{N}$. We can suppose to work in the interval $[-1, 1]$, up to linear transformations. A standard approach to solve this problem is through the polynomial interpolation which consists in determining a polynomial $P_n[f]$, of degree n, which coincides with f at the points of the set X_n. A case of great practical interest is when the set X_n coincides with the set of equispaced points in the interval $[-1, 1]$. In these hypotheses, a problem related to polynomial interpolation is the Runge phenomenon, which consists in increasing the magnitude of the interpolation error close to the ends of the interval $[-1, 1]$. It was discovered in the early 1900s by Carl David Tolmé Runge while he was studying the behavior of the error produced by the polynomial interpolation to approximate some functions [89]. To overcome this problem, in recent years, several techniques have been proposed, see for example [10, 5, 24, 26]. In particular, in 2009, John P. Boyd and Fei Xu in [10], have proved that the Runge phenomenon can be overcome if we interpolate the function f only on a proper subset of the set X_n, constituted by $m+1 = \mathcal{O}(\sqrt{n})+1$ nodes which *are close* to the Chebyshev-Lobatto nodes of order $m+1$, the so called *mock–Chebyshev* nodes. However, by using this strategy, many data are not used, and therefore, motivated by this, Stefano De Marchi, Francesco Dell'Accio and Mariarosa Mazza, in [24], proposed a new technique, with the aim of improving the accuracy of the approximation introduced in [10]. This technique is called *constrained mock-Chebyshev least squares approximation*. It consists in approximating the function f with a polynomial of degree $r > m$ which is obtained by interpolating f on the set of mock-Chebyshev nodes and using the remaining nodes to improve the accuracy of the approximation through a simultaneous regression. For this technique, as we will see in Chapter 1, the role of the nodal polynomial is crucial. Due to the lack of the nodal polynomial in the bivariate case, there is a need to find alternative approaches that allow to generalize the constrained mock-Chebyshev least squares approximation to the bivariate case. A new technique, introduced by Francesco Dell'Accio, Filomena Di Tommaso and Federico Nudo in 2022 [39], allows to define, under the same hypotheses, the same approximation introduced in [24], using the Lagrange multipliers method. This method, independent of the nodal polynomial, was developed by Joseph-Louis Lagrange to solve constrained least squares problems. By fixing a basis of the polynomial space of degree r, the approximation is computed by solving linear equations, which generalize the normal equations and which constitute the linear system KKT (this name is due to the three researchers William Karush, Harold Kuhn and Albert Tucker, who were the first to introduce it). In Chapter 2 two different strategies are proposed to obtain this generalization:

- the first one, based on tensor product interpolation on a Cartesian grid of mock-Chebyshev nodes,
- the second one, based on the interpolation of total degree on a set of nodes that *are close* to the well-known Padua nodes [9].

Another problem of great practical interest is the determination of accurate quadrature formulas on equispaced nodes. This has been studied by several authors in recent years, see for example [43, 58, 56, 68, 60, 71, 77]. In particular, the approach proposed by Hassan Majidian in [77] consists in using the Gauss-Christoffel quadrature formulas in combination with local polynomial interpolants of the integrand function f. By following this idea, in Chapter 3, we introduce a stable and accurate quadrature formula on equispaced nodes using the Gauss–Christoffel quadrature formula and the constrained mock-Chebyshev least squares approximation. Since the accuracy of these quadrature formulas varies with the degree of

regression of the constrained mock-Chebyshev least squares approximation, depending on the degree of smoothness of the function f, we develop an adaptive algorithm for determining the optimal degree of regression which corresponds to the more accurate quadrature formula. As an additional application, in [40], we use the constrained mock-Chebyshev least squares approximation to introduce a new product integration rule. This type of formula is used to approximate integrals defined over finite intervals of integrand functions which exhibit some kind of "pathology", for example, "almost" singular functions, highly oscillating functions, weakly singular functions. In Chapter 4 we further study some properties of the constrained mock-Chebyshev least squares operator, and we present new results and applications. In particular, we introduce explicit representations of the error and its derivatives, by assuming f sufficiently smooth in $[-1,1]$. In the same hypothesis, as an application, we present a method for approximating the successive derivatives of f at any point $x \in [-1,1]$, based on the constrained mock-Chebyshev least squares operator and use the previously introduced error representations to provide estimates for these approximations.

The second part of this thesis is devoted to the study of the enrichment of some standard linear finite elements. A finite element is a triple $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$, where

- K_d is a polytope in $\mathbb{R}^d$,
- $\mathbb{F}_{K_d}$ is a vector space of dimension n formed by real-valued functions defined on K_d, also called *trial functions*,
- $\Sigma_{K_d} = \{L_j : j = 1, \ldots, n\}$ is a set of linearly independent linear functionals, from the vector space $\mathbb{F}_{K_d}$, also called *degrees of freedom*,

such that $\mathbb{F}_{K_d}$ is Σ_{K_d}-unisolvent, i.e. if $f \in \mathbb{F}_{K_d}$ and

$$L_j(f) = 0, \qquad j = 1, \ldots, n,$$

then $f = 0$ [55].

The finite element method is a very popular method for numerically solving partial differential equations on a domain $D \subset \mathbb{R}^d$ [13, 19], $d \geq 1$, which arises in engineering and mathematical modeling. One of the reasons for this undeniable popularity is its versatility to deal with different types of geometries. In the finite element method the domain $\bar{D}$ is partitioned into polytopes and, for each of them, a local approximation belonging to $\mathbb{F}_{K_d}$ is computed to approximate the solution of the partial differential equation. The global approximation will be a piecewise function defined by the local approximations. If the global approximation has discontinuities at the boundary of the subdomains, the finite element is said to be nonconforming, otherwise is said to be conforming. Generally, for the standard linear finite elements the approximation space $\mathbb{F}_{K_d}$ is a space of polynomial functions. However, the approximations produced by these finite elements are not effective for solving problems involving singularities. In order to overcome this problem, several approaches have been proposed. One of the most famous approaches is to *enrich* the approximation space $\mathbb{F}_{K_d}$ with appropriate enrichment functions [92, 45, 64]. More precisely, we consider the finite element $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$, a set of enrichment functions $e_1, \ldots, e_N$ and a set of linear functionals

$$\Sigma_{K_d}^{\text{enr}} = \{L_j : j = 1, \ldots, n+N\}$$

such that the approximation space $\mathbb{F}_{K_d}$ is $\Sigma_{K_d}^{\text{enr}}$-unisolvent. We define the enriched space

$$\mathbb{F}_{K_d}^{\text{enr}} = \mathbb{F}_{K_d} \oplus \{e_1, \ldots, e_N\}.$$

In order to enrich the finite element $(K_d, \mathbb{F}_{K_d}, \Sigma_{K_d})$ the following question must be answered:

How to properly choose the enrichment functions $e_1, \ldots, e_N$, so that the triple $(K_d, \mathbb{F}_{K_d}^{\text{enr}}, \Sigma_{K_d}^{\text{enr}})$ is a finite element?

One of the simplest finite elements is the standard triangular linear finite element, defined in two-dimensional Euclidean space [19]. It is defined as the triple

$$\mathcal{P}_1(S_2) = (S_2, \mathbb{P}_1(S_2), \Sigma_{S_2}^{\text{lin}}),$$

where S_2 is a non-degenerate triangle with vertices $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\mathbb{P}_1(S_2)$ is the space of all bivariate linear polynomials and $\Sigma_{S_2}^{\text{lin}}$ is the set of point evaluation functionals at the vertices of S_2. The standard triangular linear finite element is widely used in the applications, however, it does not always produce satisfactory results due to the low order of approximation of the related trial functions. To improve the accuracy of the approximation, the element $\mathcal{P}_1(S_2)$ can be enriched with special enrichment functions (for an overview of the relevant literature see, e.g., [2, 64, 3, 4]). More precisely, in the second part of the thesis, we focus on the development of a unified and general framework for the enrichment of standard triangular linear finite elements in $\mathbb{R}^2$ and standard simplicial linear finite elements in $\mathbb{R}^d$. As we have already said, a crucial point in this approach is to determine the conditions on the enrichment functions so that they generate a finite element. Motivated by recent works on the enrichments of the finite element [2, 4], in Chapter 5 we introduce a polynomial enrichment of the standard triangular linear finite element and use this new finite element to introduce an improvement of the triangular Shepard operator (see [47, 101, 22, 31, 33] for other approaches). In Chapter 6 we introduce a new class of nonconforming finite elements by enriching the standard triangular linear finite element with enrichment continuous functions, linearly independent, not necessarily polynomials $\{e_i : i = 0, 1, 2\}$, satisfying the vanishing condition at the vertices of the triangle, i.e. $e_i(\boldsymbol{v}_j) = 0$, $i, j = 0, 1, 2$. Moreover, we determine a necessary and sufficient condition on the enrichment functions so that they generate a finite element. We show that the approximation error can be decomposed into two parts: the first one is related to the standard triangular linear finite element while the second one depends on the enrichment functions. This decomposition allows us to obtain bounds for the error in both L^∞-norm and L^1-norm. These bounds are proportional to the second and fourth powers of the radius of the circumcircle of the corresponding triangle, respectively. In Chapter 7 we generalize the results presented in Chapter 6 to the case of the standard simplicial linear finite element. It is defined as

$$\mathcal{P}_1(S_d) = (S_d, \mathbb{P}_1(S_d), \Sigma_{S_d}^{\text{lin}}),$$

where S_d is a non-degenerate simplex in $\mathbb{R}^d$ with vertices $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$, $\mathbb{P}_1(S_d)$ is the space of all linear polynomials in $\mathbb{R}^d$ and $\Sigma_{S_d}^{\text{lin}}$ is the set of point evaluation functionals at the vertices of S_d. In line with Chapter 6, we enrich the standard simplicial linear finite element $\mathcal{P}_1(S_d)$ with $d+1$ linearly independent continuous functions $\{e_i : i = 0, \ldots, d\}$, satisfying the vanishing condition at the vertices of S_d, i.e. $e_i(\boldsymbol{v}_j) = 0$, $i, j = 0, \ldots, d$. In Chapter 8 we provide a general strategy for enriching the standard simplicial linear finite element $\mathcal{P}_1(S_d)$ without imposing restrictive conditions on the enrichment functions, like their vanishing at the vertices of S_d. A finite element commonly used in the applications is the simplicial vector linear finite element. It is defined as

$$\boldsymbol{\mathcal{P}}_1(S_d) = (S_d, \boldsymbol{\mathbb{P}}_1(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{lin}}),$$

where $\boldsymbol{\mathbb{P}}_1(S_d)$ is the direct product, d times, of the vector space $\mathbb{P}_1(S_d)$ with itself and

$$\boldsymbol{\Sigma}_{S_d}^{\text{lin}} = \{\boldsymbol{L}_j : j = 0, \ldots, d\},$$

with $\boldsymbol{L}_j$ defined as

$$\boldsymbol{L}_j(\boldsymbol{f}) = \boldsymbol{f}(\boldsymbol{v}_j) = [f_1(\boldsymbol{v}_j), \ldots, f_d(\boldsymbol{v}_j)]^T, \quad \boldsymbol{f} = [f_1, \ldots, f_d]^T, \quad j = 0, \ldots, d.$$

This finite element is commonly used for numerically solving the stationary Stokes equations. It is known, however, to suffer from severe shortcomings in application to more complicated situations. An enriched finite element, that overcomes the aforementioned drawbacks, was proposed and developed by Bernardi and Raugel [7]. It can be regarded as an advanced and generalized version of the conventional simplicial vector linear finite element, and it has been employed in a wide range of practical engineering computation fields. It uses polynomials as enrichment functions. However, the intractable linear dependence issue is always encountered when this type of enrichment functions is employed. In line with previous researches, in Chapter 9, we present a general strategy for enriching the simplicial vector linear finite element by nonpolynomial enrichment functions. This enriched finite element is defined with respect to any simplex, and can be regarded as an extension of Bernardi and Raugel element.

Chapter 1

Constrained mock-Chebyshev least squares approximation

The constrained mock-Chebyshev least squares approximation is an approximation method based on an equispaced grid of points. Like other polynomial or rational approximation methods, it was recently introduced in order to defeat the Runge phenomenon that occurs when using polynomial interpolation on large sets of equally spaced points. The idea is to improve the mock-Chebyshev subset interpolation, where the considered function f is interpolated only on a proper subset of the uniform grid, formed by nodes that mimic the behavior of Chebyshev–Lobatto nodes. In the mock-Chebyshev subset interpolation all remaining nodes are discarded, while in the constrained mock-Chebyshev least squares approximation they are used in a simultaneous regression, with the aim to further improving the accuracy of the approximation provided by the mock-Chebyshev subset interpolation. In this introductory chapter, we recall the main important properties of the constrained mock-Chebyshev least squares approximation, introduced in [24], which will be useful in throughout the thesis.

1.0.1 The constrained mock-Chebyshev least squares approximation: some computational issues

Let f be a continuous function in $[-1,1]$ and we suppose that its evaluations are known on the grid of $n+1$ equispaced nodes in $[-1,1]$, that is

$$X_n = \left\{ x_i = -1 + \frac{2}{n} i \, : \, i = 0, \dots, n \right\}. \tag{1.1}$$

The idea that underlies the mock-Chebyshev subset interpolation is to interpolate f only on a proper subset of X_n, formed by nodes which best mimic the behavior of the well-known Chebyshev-Lobatto nodes of a suitable order $m+1$. If we carefully choose m, the convergence of the interpolation process on such a subset of nodes, for n which tends to infinity, will be preserved [80]. To understand how to properly choose m, let us remember that the $m+1$ Chebyshev–Lobatto nodes are defined as

$$x_j^{CL} = -\cos\left(\frac{\pi}{m} j\right), \qquad j = 0, \dots, m.$$

By expanding x_1^{CL} in a Taylor series centered in zero, we get

$$x_1^{CL} = -1 + \frac{\pi^2}{2m^2} + \mathcal{O}\left(\frac{1}{m^4}\right) < -1 + \frac{\pi^2}{2m^2}, \tag{1.2}$$

and since $x_0^{CL} = -1$, we have

$$x_1^{CL} - x_0^{CL} = \mathcal{O}\left(\frac{1}{m^2}\right).$$

This means that the $m+1$ Chebyshev-Lobatto nodes are distributed in $[-1,1]$ with a density that is roughly quadratic in m. Then for n proportional to m^2, we can select among the given nodes a subset which mimic a sufficiently large Chebyshev–Lobatto grid. Let c be the constant of proportionality. A way to calculate it is to impose that the second node of the Chebyshev–Lobatto grid is as close as possible to the second node of the equispaced set X_n

$$-\cos\left(\frac{\pi}{m}\right) \approx -1+\frac{2}{n}.$$

By (1.2) we fix the largest integer m such that

$$-1+\frac{1}{n} < -1+\frac{\pi^2}{2m^2},$$

that is

$$m = \left\lfloor \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rfloor.$$

Therefore, for this value of m, x_1 is the point of X_n closest to x_1^{CL}. This choice of $c < \pi/\sqrt{2}$ avoids the fact that the endpoints -1 and 1 can be selected more than once. For analytic functions the polynomial interpolation on Chebyshev nodes converges geometrically and stably. We denote by

$$X'_m = \left\{ x'_j \,:\, \left|x'_j - x_j^{CL}\right| = \min_{x_i \in X_n} \left|x_i - x_j^{CL}\right|,\ j=0,\dots,m \right\} \tag{1.3}$$

the mock-Chebyshev subset of X_n of order $m+1$ [10, 81, 69, 70]. The mock-Chebyshev subset interpolation

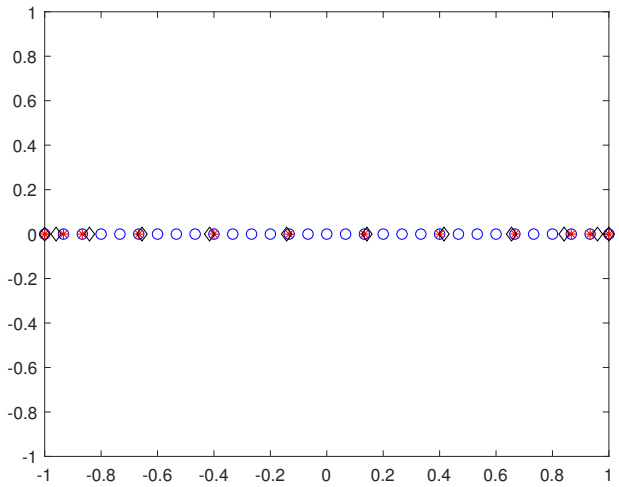


Figure 1.1: Plot of equispaced nodes in $[-1,1]$ (o), Chebyshev nodes ($\diamond$) and mock-Chebyshev nodes ($\star$) for $n+1=31$ with $m+1=12$.

is a stable procedure, but its rate of convergence is subgeometric [24]. In [81] it has been shown that on equispaced nodes no stable method can converge geometrically. In performing the mock-Chebyshev subset interpolation we know the evaluations of f on the whole set X_n, but actually we only use the information corresponding to the elements of X'_m. The idea of the constrained mock-Chebyshev least squares approximation consists to use the other $n-m$ nodes $X''_{n-m} = \{x''_k \,:\, k=1,\dots,n-m\}$ to improve the accuracy of the approximation through a simultaneous regression.

1.0.2 The nodal polynomial method

Given an analytic function f in the interval $[-1,1]$ and an integer $r \in \mathbb{N}$ such that $m < r \leq n$, the *constrained mock-Chebyshev least squares problem* consists in finding the best approximation $\hat{P}_{r,n}[f]$,

with respect to the ℓ^2-norm on X_n, to the function f from the closed convex space $\hat{\mathbb{P}}_r(\mathbb{R})$ formed by the polynomials of degree less than or equal to r interpolating f on the mock-Chebyshev nodes. We will write this problem in short form as follows

$$\text{find } \hat{P}_{r,n}[f] \in \hat{\mathbb{P}}_r(\mathbb{R}) \text{ s.t. } \|f - \hat{P}_{r,n}[f]\|_{2,X_n}^2 = \min_{P_r \in \hat{\mathbb{P}}_r(\mathbb{R})} \|f - P_r\|_{2,X_n}^2 . \tag{1.4}$$

Let $P_m[f] \in \mathbb{P}_m(\mathbb{R})$ be the interpolation polynomial on the mock-Chebyshev nodes X'_m and let ω_m be the corresponding nodal polynomial, that is

$$\omega_m(x) = \prod_{i=0}^{m} (x - x'_i). \tag{1.5}$$

Theorem 1.0.1. *The constrained mock-Chebyshev least squares problem* (1.4) *has a unique solution.*

Proof. It is not difficult to verify that a generic polynomial $P_r \in \hat{\mathbb{P}}_r(\mathbb{R})$ can be written as

$$P_r(x) = P_m[f](x) + Q_s(x)\omega_m(x),$$

where $\omega_m(x)$ is defined in (1.5) and $Q_s(x)$ is an arbitrary polynomial of degree $s = r - m - 1$. The problem (1.4) becomes

$$\min_{Q_s \in \mathbb{P}_s(\mathbb{R})} \|f - (P_m[f] + Q_s\omega_m)\|_{2,X_n}^2 = \min_{Q_s \in \mathbb{P}_s(\mathbb{R})} \sum_{k=1}^{n-m} \left(f(x''_k) - P_m[f](x''_k) - Q_s(x''_k)\omega_m(x''_k)\right)^2$$

$$= \min_{Q_s \in \mathbb{P}_s(\mathbb{R})} \sum_{k=1}^{n-m} \left(\frac{f(x''_k) - P_m[f](x''_k)}{\omega_m(x''_k)} - Q_s(x''_k) \right)^2 \omega_m^2(x''_k).$$

We introduce the following discrete weighted ℓ^2-norm

$$\|u\|_{2,\omega_m^2} = \left(\sum_{k=1}^{n-m} u^2(x''_k)\omega_m^2(x''_k) \right)^{\frac{1}{2}} \tag{1.6}$$

and the following function

$$\hat{f}(x) = \frac{f(x) - P_m[f](x)}{\omega_m(x)}, \qquad x \in [-1, 1]. \tag{1.7}$$

Then the solution of the problem (1.4) is

$$\hat{P}_{r,n}[f](x) = P_m[f](x) + \hat{Q}_s[\hat{f}](x)\omega_m(x), \tag{1.8}$$

where

$$\left\| \hat{f} - \hat{Q}_s[\hat{f}] \right\|_{2,\omega_m^2}^2 = \min_{Q_s \in \mathbb{P}_s(\mathbb{R})} \left\| \hat{f} - Q_s \right\|_{2,\omega_m^2}^2 \tag{1.9}$$

which has a unique solution. □

The name *nodal polynomial method* is due to the fact that we apply a classical least squares method to the analytic function $\hat{f}$ obtained from the function f through the nodal polynomial as in equation (1.7). The degree of $\hat{P}_{r,n}[f]$ depends on the degree of the polynomial of simultaneous regression $\hat{Q}_s[\hat{f}]$. When r increases up to n, the approximation provided by $\hat{P}_{r,n}[f]$ can get worse, since the constrained mock-Chebyshev least squares approximation approaches the interpolation polynomial on X_n. The problem of determining a degree r which gives, in the uniform norm, the better accuracy of $\hat{P}_{r,n}[f]$ with respect to $P_m[f]$, has been tackled in [24], basing on a general result by L. Reichel [85] on the polynomial approximation in the uniform norm by the discrete least squares method. This result implies that, for an equispaced set of q internal nodes of $[-1, 1]$,

$$z_k = -1 + \frac{2k-1}{q}, \quad k = 1, \ldots, q, \tag{1.10}$$

the degree p of the least squares polynomial should be selected as the greatest p so that there is a subset of cardinality $p+1$ of the equispaced set (1.10) which is close to the $p+1$ Chebyshev grid

$$X_p^C = \left\{ x_k^C = \cos\left(\frac{2k-1}{2p+2}\pi\right) \,:\, k = 1, \ldots, p+1 \right\}.$$

In other words, p should be selected in the mock-Chebyshev sense [24]. In the case of simultaneous regression (1.9) the nodes used in the least squares approximation are those of X''_{n-m} and therefore they are not equally spaced. Despite X''_{n-m} is not an equispaced grid, in [24] it is proven that, for n sufficiently large, it is possible to approximate an equispaced grid of $q = \lfloor \frac{n}{6} \rfloor$ internal nodes of $[-1, 1]$ with nodes of X''_{n-m}. We denote this grid by $\tilde{X}''_{n-m}$, and by

$$X'''_p = \{x'''_0, \ldots, x'''_p\} \tag{1.11}$$

the mock-Chebyshev subset of $\tilde{X}''_{n-m}$. The choice for the degree of simultaneous regression which gives good approximation in the uniform norm has been also determined in [24]

$$p = \left\lfloor \frac{\pi}{\sqrt{2}}\sqrt{q} \right\rfloor = \left\lfloor \pi\sqrt{\frac{n}{12}} \right\rfloor.$$

Therefore, the degree r of the polynomial $\hat{P}_{r,n}[f]$, which gives more accurate approximation to f is

$$r = m + p + 1 = \left\lfloor \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rfloor + \left\lfloor \pi\sqrt{\frac{n}{12}} \right\rfloor + 1. \tag{1.12}$$

Proposition 1.0.2. *The following upper and lower bound for r holds*

$$\left(1 + \frac{1}{\sqrt{6}}\right) m - 1 < r < \left(1 + \frac{1}{\sqrt{6}}\right)(m+1) + 1. \tag{1.13}$$

Proof. Firstly, we note that

$$\begin{aligned} r &= \left\lfloor \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rfloor + \left\lfloor \pi\sqrt{\frac{n}{12}} \right\rfloor + 1 \leq \left\lfloor \left(1 + \frac{1}{\sqrt{6}}\right) \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rfloor + 1 \\ &< \left(1 + \frac{1}{\sqrt{6}}\right) \frac{\pi}{\sqrt{2}}\sqrt{n} + 1 < \left(1 + \frac{1}{\sqrt{6}}\right) \left\lceil \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rceil + 1 \\ &= \left(1 + \frac{1}{\sqrt{6}}\right)(m+1) + 1. \end{aligned}$$

Similarly, the following inequalities prove the left-hand side of (1.13)

$$\begin{aligned} r &= \left\lfloor \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rfloor + \left\lfloor \pi\sqrt{\frac{n}{12}} \right\rfloor + 1 = \left\lceil \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rceil + \left\lceil \pi\sqrt{\frac{n}{12}} \right\rceil - 1 \\ &\geq \left\lceil \left(1 + \frac{1}{\sqrt{6}}\right) \frac{\pi}{\sqrt{2}}\sqrt{n} \right\rceil - 1 > \left(1 + \frac{1}{\sqrt{6}}\right) \frac{\pi}{\sqrt{2}}\sqrt{n} - 1 \\ &> \left(1 + \frac{1}{\sqrt{6}}\right) m - 1. \end{aligned}$$

□

Chapter 2

Generalizations of the constrained mock-Chebyshev least squares in two variables: Tensor product vs total degree polynomial interpolation

The main goal of this chapter is to extend the univariate constrained mock-Chebyshev least squares approximation to the bivariate case through the Lagrange multipliers method. This is done in two different ways, the first one based on the tensor product interpolation and the second one based on the mock-Padua points, that is the set of $(m+1)(m+2)/2$ nodes extracted from a uniform grid of points in the square $[-1,1]^2$ that mimic the behavior of the well-known Padua points [8]. The result presented in this chapter can be found in [39].

2.1 Constrained mock-Chebyshev least squares approximation through the Lagrange multipliers method

The *Lagrange multipliers method* was developed by the mathematician Joseph-Louis Lagrange and it is very helpful in solving the constrained least squares problems. The method requires the choice of a basis $\mathcal{B}_r = \{u_j(x) : j = 0, \dots, r\}$ of the polynomial space $\mathbb{P}_r(\mathbb{R})$. Let f be a continuous function in $[-1,1]$ and let X_n and X'_m be the sets introduced in (1.1) and (1.3), respectively. We denote by V the interpolation matrix at the nodes of the equispaced grid X_n relative to $\mathcal{B}_r$, that is

$$V = [u_j(x_i)]_{\substack{i=0,\dots n \\ j=0,\dots,r}}$$

and $\boldsymbol{b} = [f(x_0), \dots, f(x_n)]^T$. Without loss of generality we assume that the first $m+1$ points of X_n are those ones of X'_m and that $\mathcal{B}_m = \{u_j(x) : j = 0, \dots, m\}$ spans the polynomial space $\mathbb{P}_m(\mathbb{R})$. Let $C = [\boldsymbol{c}_i^T]_{i=0,\dots,m}$ be the matrix formed by the first $m+1$ rows, $\boldsymbol{c}_0^T, \dots, \boldsymbol{c}_m^T$, of V and $\boldsymbol{d} = [d_0, \dots, d_m]^T$ the column vector formed by the first $m+1$ components of $\boldsymbol{b}$. The solution $\hat{P}_{r,n}[f]$ of problem (1.4) in the basis $\mathcal{B}_r$ is

$$\hat{P}_{r,n}[f](x) = \sum_{i=0}^{r} \hat{a}_i u_i(x),$$

where the vector of coefficients $\hat{\boldsymbol{a}} = [\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r]^T$ satisfies

$$C\hat{\boldsymbol{a}} = \boldsymbol{d} \quad \text{and} \quad \|V\hat{\boldsymbol{a}} - \boldsymbol{b}\|_2^2 = \min_{\boldsymbol{a} \in \mathbb{R}^{r+1}} \|V\boldsymbol{a} - \boldsymbol{b}\|_2^2 .$$

Usually previous constrained least squares problem is written in compact form as follows

$$\text{find } \hat{\boldsymbol{a}} \in \mathbb{R}^{r+1} \text{ s.t. } \|V\hat{\boldsymbol{a}} - \boldsymbol{b}\|_2^2 = \min_{\substack{\boldsymbol{a}\in\mathbb{R}^{r+1} \\ C\boldsymbol{a}=\boldsymbol{d}}} \|V\boldsymbol{a} - \boldsymbol{b}\|_2^2. \tag{2.1}$$

Since the nodes of X_n are pairwise distinct, the interpolation matrix V has maximum rank and therefore the problem (2.1) has a unique solution [11, Ch. 16]. This solution can be computed by the method of Lagrange multipliers, with Lagrangian function

$$L(\boldsymbol{a}, \boldsymbol{z}) = \|V\boldsymbol{a} - \boldsymbol{b}\|_2^2 + z_0(\boldsymbol{c}_0^T\boldsymbol{a} - d_0) + \cdots + z_m(\boldsymbol{c}_m^T\boldsymbol{a} - d_m),$$

where $\boldsymbol{z} = [z_0, \ldots, z_m]^T$ is the vector of Lagrange multipliers. As well-known, if $\hat{\boldsymbol{a}}$ is a solution of problem (2.1) then there exist a vector $\hat{\boldsymbol{z}} = [\hat{z_0}, \ldots, \hat{z}_m]^T$ such that

$$\frac{\partial L}{\partial a_i}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{z}}) = 0 \quad \text{and} \quad \frac{\partial L}{\partial z_j}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{z}}) = 0, \qquad i = 0, \ldots, r, \quad j = 0, \ldots, m,$$

or, equivalently,

$$\begin{bmatrix} 2V^TV & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} 2V^T\boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix}. \tag{2.2}$$

The equations (2.2), which are an extension of the normal equations for a least squares problem with no constraints, are called *KKT linear equations* and the $(r+m+2)\times(r+m+2)$ coefficient matrix is called *KKT matrix*, in honor of W. Karush, H. Kuhn and A. Tucker (see [11, Ch. 16] for more details).

2.1.1 Numerical experiments

We numerically compare the approximation produced by the nodal polynomial method with that one produced by the Lagrange multipliers method. In line with the numerical experiments presented in [24] we use 1001 equispaced nodes in $[-1, 1]$ as sample points and 71 mock-Chebyshev nodes, that is $n = 1000$ and $m = 70$. We perform the constrained mock-Chebyshev least squares approximation through the nodal polynomial method and through the Lagrange multipliers method, by choosing the following polynomial bases

- $\mathcal{B}_r^P = \{x^i \,:\, i = 0, \ldots, r\}$,
- $\mathcal{B}_r^{C,1} = \{T_i(x) \,:\, i = 0, \ldots, r\}$, where $T_0(x) = 1,\ T_1(x) = x, \quad T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x),\ i \geq 2$,
- $\mathcal{B}_{r,m}^L = \left\{\ell_i(x), \omega_m(x)x^k \,:\, i = 0, \ldots, m, \quad k = 0, \ldots, r-m-1\right\}$, where $\ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{m} \frac{x - x'_j}{x'_i - x'_j}$ and $\omega_m(x) = \prod_{j=0}^{m}(x - x'_j)$

and compute the condition number of the relative KKT matrices for various degrees $p = r - m - 1$ of simultaneous regression. We notice that $\mathcal{B}_r^P$ is the classical monomial basis while $\mathcal{B}_r^{C,1}$ is the Chebyshev polynomial basis of the first kind. We choose the basis $\mathcal{B}_{r,m}^L$ since it is used in (1.8); for this basis the interpolation matrix C is the identity matrix of order $m+1$. We compute the mean approximation error (MAE) for the test functions

$$f_1(x) = \frac{1}{1 + 25x^2}, \quad f_2(x) = \frac{1}{x^4 + \left(\frac{2}{50}\right)^2},$$

at 10001 equispaced points in $[-1, 1]$. The results are shown in Figure 2.1, where left and center plots are related to the MAE, while the right plot shows the trends of the condition numbers of the KKT matrices and of the normal equation matrix of the nodal polynomial method. Notice that the MAE of the nodal polynomial method and those of the Lagrange multipliers method, with respect to the bases $\mathcal{B}_r^{C,1}$, $\mathcal{B}_{r,m}^L$, are of the same order of magnitude till a degree of simultaneous regression of about 30. For greater degrees

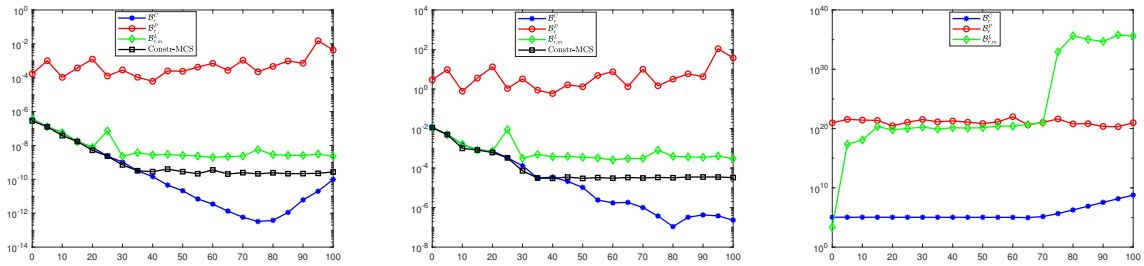


Figure 2.1: Mean approximation error for f_1 (left), for f_2 (center) compared with those obtained by the nodal polynomial method Constr-MCS and condition number of the KKT matrices (right) relative to the bases $\mathcal{B}_r^P$, $\mathcal{B}_r^{C,1}$ and $\mathcal{B}_{r,m}^L$ and of the normal equation matrix of the nodal polynomial method.

of simultaneous regression both the nodal polynomial method and the Lagrange multipliers method, with respect to the basis $\mathcal{B}_{r,m}^L$, do not improve nor get worse the MAE, which oscillate around the previous reached precision. On the other hand, the MAE of the Lagrange multipliers method, with respect to the basis $\mathcal{B}_r^{C,1}$, decreases till a degree of simultaneous regression of about 80 after which it starts to become worse, since regression *tends* to interpolation. Finally, the Lagrange multipliers method, with respect to the basis $\mathcal{B}_r^P$, is not comparable with the ones obtained with the other basis in approximation accuracy. The approximation accuracies of the various methods reflect the behaviour of the condition numbers of the related matrices.

2.2 Tensor product vs total degree interpolation

2.2.1 Constrained mock-Chebyshev least squares tensor product interpolation

A natural way to extend the univariate constrained mock-Chebyshev least squares approximation to the bivariate case is through the tensor product interpolation [18, Ch. 7]. To this aim, let f be an analytic function in the square $[-1,1]^2$, we set $\boldsymbol{n}_{x,y} = (n_x, n_y) \in \mathbb{N} \times \mathbb{N}$ and let us consider the uniform grid of nodes

$$X_{n_x} \times Y_{n_y} = \left\{ \left(-1 + \frac{2}{n_x} i, -1 + \frac{2}{n_y} j \right) \,:\, i = 0, \dots, n_x,\, j = 0, \dots, n_y \right\}. \tag{2.3}$$

In line with the notations of Chapter 1, we denote by

$$m_x = \left\lfloor \pi \sqrt{\frac{n_x}{2}} \right\rfloor, \quad m_y = \left\lfloor \pi \sqrt{\frac{n_y}{2}} \right\rfloor,$$

$\boldsymbol{m}_{x,y} = (m_x, m_y)$ and we consider the Cartesian grid of mock-Chebyshev nodes

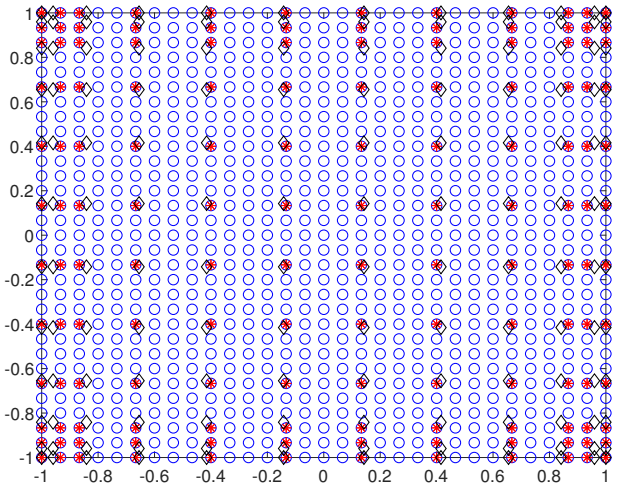
$$X'_{m_x} \times Y'_{m_y} = \{(x'_i, y'_j) \,:\, i = 0, \dots, m_x,\, j = 0, \dots, m_y\}. \tag{2.4}$$

Let $\mathbb{P}_{r_x}(\mathbb{R}) \otimes \mathbb{P}_{r_y}(\mathbb{R})$ be the tensor product of the polynomial spaces $\mathbb{P}_{r_x}(\mathbb{R})$, $\mathbb{P}_{r_y}(\mathbb{R})$ in the variables x, y, respectively, with basis $\mathcal{B}_{r_x} \otimes \mathcal{B}_{r_y} = \{u_i(x)v_j(y) \,:\, i = 0, \dots, r_x,\, j = 0, \dots, r_y\}$. Let $\boldsymbol{r}_{x,y} = (r_x, r_y)$ and $\hat{\mathbb{P}}_{\boldsymbol{r}_{x,y}}(\mathbb{R}^2)$ be the closed convex space of all polynomials in $\mathbb{P}_{r_x}(\mathbb{R}) \otimes \mathbb{P}_{r_y}(\mathbb{R})$ interpolating f on $X'_{m_x} \times Y'_{m_y}$. The constrained mock-Chebyshev least squares tensor product interpolation problem is the following

$$\text{find } \hat{P}_{\boldsymbol{r}_{x,y}, \boldsymbol{n}_{x,y}}[f] \in \hat{\mathbb{P}}_{\boldsymbol{r}_{x,y}}(\mathbb{R}^2) \text{ s.t. } \left\| f - \hat{P}_{\boldsymbol{r}_{x,y}, \boldsymbol{n}_{x,y}}[f] \right\|_2^2 = \min_{P_{\boldsymbol{r}_{x,y}} \in \hat{\mathbb{P}}_{\boldsymbol{r}_{x,y}}(\mathbb{R}^2)} \left\| f - P_{\boldsymbol{r}_{x,y}} \right\|_2^2. \tag{2.5}$$

Since each polynomial $U(x)$, expressed in the basis $\mathcal{B}_{r,m}^L$, interpolates its first $m+1$ coefficients at the mock-Chebyshev nodes, then each polynomial

$$L(x,y) = \sum_{i=0}^{r_x} \sum_{j=0}^{r_y} a_{ij} u_i(x) u_j(y),$$



Plot of uniform grid of nodes (o), Cartesian grid of Chebyshev–Lobatto nodes ($\diamond$) and Cartesian grid of mock-Chebyshev nodes ($\star$) for $n+1=31$ with $m+1=12$.

expressed in the basis $\mathcal{B}^L_{r_x,m_x} \otimes \mathcal{B}^L_{r_y,m_y}$, interpolates its coefficients a_{ij}, $i=0,\dots,m_x$, $j=0,\dots,m_y$ [18, Ch. 7, Lemma 4]. Therefore, the polynomial $\hat{P}_{\boldsymbol{r}_{x,y},\boldsymbol{n}_{x,y}}[f]$, in the basis $\mathcal{B}^L_{r_x,m_x} \otimes \mathcal{B}^L_{r_y,m_y}$, has the form

$$\hat{P}_{\boldsymbol{r}_{x,y},\boldsymbol{n}_{x,y}}[f] = \hat{P}_{\boldsymbol{m}_{x,y}}[f] + \omega_{m_x}(x)Q_1(x,y) + \omega_{m_y}(y)Q_2(x,y) + \omega_{m_x}(x)\omega_{m_y}(y)Q_3(x,y), \tag{2.6}$$

where $P_{\boldsymbol{m}_{x,y}}[f]$ is the interpolation polynomial at $X_{m_x} \times Y_{m_y}$, $Q_1 \in \mathbb{P}_{r_x-m_x}(\mathbb{R}) \otimes \mathbb{P}_{r_y}(\mathbb{R})$, $Q_2 \in \mathbb{P}_{r_x}(\mathbb{R}) \otimes \mathbb{P}_{r_y-m_y}(\mathbb{R})$, $Q_3 \in \mathbb{P}_{r_x-m_x}(\mathbb{R}) \otimes \mathbb{P}_{r_y-m_y}(\mathbb{R})$. The possibility to approach the problem (2.5) by the nodal polynomial method, recalled in Chapter 1, is precluded by the expression (2.6) of its solution, which does not allow the introduction of a discrete norm like (1.6) nor a function like (1.7). On the contrary, the Lagrange multipliers method, recalled in Section 2.1, can be used in analogy with the univariate case, with the settings and requirements there specified. In particular, we assume that the $(n_x+1)(n_y+1)$ nodes of the Cartesian grid (2.3) are reorganized into a sequence and that the first $(m_x+1)(m_y+1)$ nodes of this sequence are those ones of the Cartesian grid (2.4). Similarly, we assume that the $(r_x+1)(r_y+1)$ elements of the basis $\mathcal{B}_{r_x} \otimes \mathcal{B}_{r_y}$ are reorganized into a sequence and that the first $(m_x+1)(m_y+1)$ elements of this sequence spans the polynomial space $\mathbb{P}_{m_x}(\mathbb{R}) \otimes \mathbb{P}_{m_y}(\mathbb{R})$. The maximum rank of the corresponding interpolation matrix $V \in \mathbb{R}^{(n_x+1)(n_y+1)\times(r_x+1)(r_y+1)}$ is guaranteed by the unisolvence of the tensor product interpolation problem on the Cartesian grid (2.3) by polynomials of $\mathbb{P}_{n_x}(\mathbb{R}) \otimes \mathbb{P}_{n_y}(\mathbb{R})$ [18, Ch. 7].

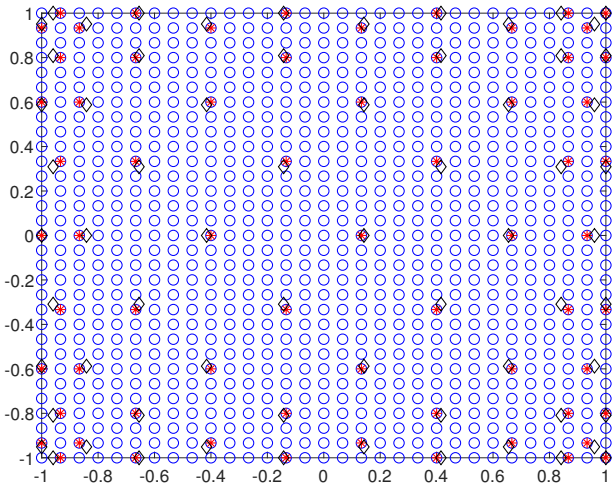
2.2.2 Constrained mock-Padua least squares approximation

The polynomial interpolation of total degree in two variables offers a further possibility to extend the idea of the constrained mock-Chebyshev least squares approximation to the bivariate case. In this case, in fact, the Padua points constitute an optimal unisolvent set of points for total degree polynomial interpolation with minimal growth of their Lebesgue constant [8]. These points are defined in the square $[-1,1]^2$ by sampling suitable *generating curves* $\gamma_s(t)$, $s=1,\dots,4$, which are slightly different from each other and allow to get four different families of points [9, 14]. In the following we consider the first family of Padua points of degree m, $m \in \mathbb{N}$, and denote this set by Pad_m by simply calling them Padua points of degree m. In line with the case of mock-Chebyshev points, it is possible to define the mock-Padua points of degree m as the set of $(m+1)(m+2)/2$ points extracted from a square uniform grid (2.3) that mimic the behavior of the Padua points of the same degree. To do this, we set $n=n_x=n_y$ and $m=\left\lfloor \pi\sqrt{\frac{n}{2}} \right\rfloor$. We denote by $X'_m = \{x'_i : i=0,\dots,m\}$ the mock-Chebyshev subset of X_n of order $m+1$ and with $Y'_{m+1} = \{y'_j : j=0,\dots,m+1\}$ the mock-Chebyshev subset of Y_n of order $m+2$. We use the formula for computing the Padua points from two sets of Chebyshev-Lobatto nodes of order m and $m+1$ given

in [8] to obtain the set of mock-Padua points [69] of degree m from the mock-Chebyshev subsets X'_m and Y'_{m+1} as follows

$$Pad'_m = \{(x'_i, y'_j) \, : \, i = 0, \ldots, m, \quad j = 1, \ldots, \lfloor m/2 \rfloor + 1 + \delta_k\},$$

where $\delta_k = 0$ if m is even or m is odd but i is even, $\delta_k = 1$ if m is odd and i is odd.



Plot of uniform grid of nodes (o), Padua nodes ($\diamond$) and mock-Padua nodes ($\star$) for $n + 1 = 31$ with $m + 1 = 11$.

Given an analytic function f in the square $[-1, 1]^2$ and an integer $r \in \mathbb{N}$ such that $m < r \leq n$, the *constrained mock-Padua least squares problem* consists in finding the best approximation $\hat{P}_{r,n}[f]$, with respect to the ℓ^2-norm on $X_n \times Y_n$, to the function f from the closed convex space $\hat{\mathbb{P}}_r(\mathbb{R}^2)$ of all polynomials of total degree less than or equal to r interpolating f on the mock-Padua subset of nodes. A basis $\tilde{\mathcal{B}}_r = \{u_i(x)u_j(y) \, : \, 0 \leq i + j \leq r\}$ for the polynomial space $\mathbb{P}_r(\mathbb{R}^2)$ can be obtained from any basis $\mathcal{B}_r = \{u_i(x) \, : \, i = 0, \ldots, r\}$ of the polynomial space $\mathbb{P}_r(\mathbb{R})$ [18, Ch. 4]. Similarly with the case of tensor product interpolation, we assume that the $(n + 1)^2$ nodes of the square Cartesian grid (2.3) are reorganized into a sequence and that the first $(m + 1)(m + 2)/2$ nodes of this sequence are those ones of Pad'_m; we assume also that the $(r + 1)(r + 2)/2$ elements of the basis $\tilde{\mathcal{B}}_r$ are reorganized into a sequence and that the first $(m + 1)(m + 2)/2$ elements of this sequence, forming the set $\tilde{\mathcal{B}}_m$, spans the polynomial space $\mathbb{P}_m(\mathbb{R}^2)$. Let V be the interpolation matrix at the nodes of $X_n \times Y_n$ relative to the basis $\tilde{\mathcal{B}}_r$ and C the matrix formed by the first $(m + 1)(m + 2)/2$ rows of V. The rank of the matrix C is maximum since $\tilde{\mathcal{B}}_m$ interpolates on Pad'_m [96] while the rank of V is maximum since $\tilde{\mathcal{B}}_r$ can be completed to the basis $\mathcal{B}_n \otimes \mathcal{B}_n$ interpolating on $X_n \times Y_n$.

2.2.3 Numerical experiments

We consider the uniform grid of 101×101 nodes in $[-1, 1]^2$ and we compute the condition numbers of the KKT matrices relative to the bases $\mathcal{B}_r^{C,1} \otimes \mathcal{B}_r^{C,1}, \mathcal{B}_{r,m}^{L} \otimes \mathcal{B}_{r,m}^{L}$ and $\tilde{\mathcal{B}}_r^{C}$ for different degrees $r - m - 1$ of simultaneous regression. We compute the MAE for the test functions

$$f_3(x, y) = \frac{1}{1 + 25(x^2 + y^2)}, \quad f_4(x, y) = \frac{1}{x^2 + y^2 - 2.5},$$

at the uniform grid of 132×132 points in $[-1, 1]^2$. The results are shown in Figure 2.2, where left and center plots are related to the MAE, while the right plot shows the trends of the condition numbers of the KKT matrices. We note that the basis $\mathcal{B}_r^{C,1} \otimes \mathcal{B}_r^{C,1}$ and $\mathcal{B}_{r,m}^{L} \otimes \mathcal{B}_{r,m}^{L}$ interpolate on a set of nodes whose cardinality is greater than the cardinality of the set Pad'_m used by the basis $\tilde{\mathcal{B}}_r^{C}$ as nodes set for

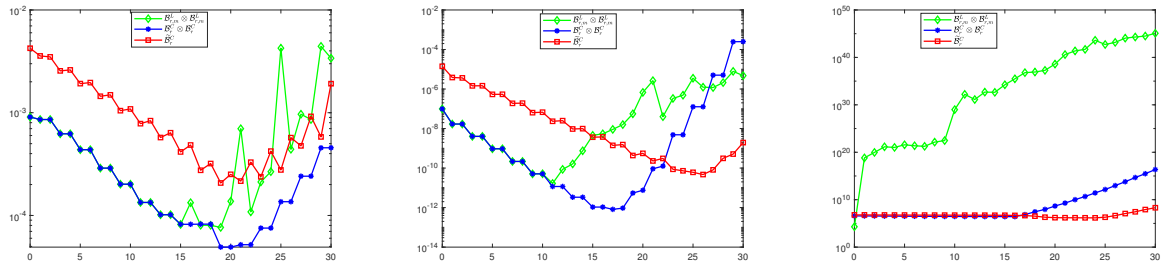


Figure 2.2: Mean approximation error for f_3 (left), for f_4 (center) and condition number of the KKT matrices (right) relative to the bases $\mathcal{B}_r^{C,1} \otimes \mathcal{B}_r^{C,1}$, $\mathcal{B}_{r,m}^{L} \otimes \mathcal{B}_{r,m}^{L}$, $\tilde{\mathcal{B}}_r^{C}$.

the interpolation. This reflect on the different accuracies of approximation reached by the three bases on their respective interpolation set. In line with the univariate case, the MAE of the Lagrange multipliers method, with respect to the bases $\mathcal{B}_r^{C,1} \otimes \mathcal{B}_r^{C,1}$, $\mathcal{B}_{r,m}^{L} \otimes \mathcal{B}_{r,m}^{L}$, are of the same order of magnitude till a certain degree of simultaneous regression which varies with the test function (15 and 10 in the case of function f_3 and f_4, respectively). From these degrees on, the MAE relative to the basis $\mathcal{B}_r^{C,1} \otimes \mathcal{B}_r^{C,1}$ continues to decrease, while that one relative to the basis $\mathcal{B}_{r,m}^{L} \otimes \mathcal{B}_{r,m}^{L}$ starts to become worse. The Lagrange multipliers method, with respect to the basis $\tilde{\mathcal{B}}_r^{C}$, has a similar decreasing behaviour although it is less accurate for the aforesaid reason. The approximation accuracies of the various methods reflect the behaviour of the condition numbers of the related matrices.

Chapter 3

Constrained mock-Chebyshev least squares quadrature

Let f be a continuous function in $[-1,1]$ and let $w \in L^1(-1,1)$ be a positive (or nonnegative) weight function in $[-1,1]$. A recurring problem in applied mathematics is to approximate the weighted integral

$$I(f) = \int_{-1}^{1} f(x)w(x)dx \tag{3.1}$$

by a weighted quadrature formula of the type

$$I(f) = \sum_{k=1}^{m} w_k f(\xi_k) + E_m(f) \tag{3.2}$$

based on the nodes $-1 \leq \xi_1 < \cdots < \xi_m \leq 1$ and weights $w_1, \ldots, w_m \in \mathbb{R}$. Let $\mathbb{P}_s(\mathbb{R})$ be the space of polynomials of degree less than or equal to s. The largest positive integer s such that

$$E_m(P_s) = 0 \text{ for all } P_s \in \mathbb{P}_s(\mathbb{R})$$

is called *degree of exactness* of the quadrature formula (3.2). If the function f is known or computable on the whole interval $[-1,1]$, to approximate the integral (3.1) it is convenient to choose a suitable Gaussian quadrature formula, depending on the weight function w. For example, if $w(x) = 1$, we can use the Gauss–Legendre formula, while if

$$w(x) = \frac{1}{\sqrt{1-x^2}} \qquad \text{or} \qquad w(x) = (1-x)^\alpha (1+x)^\beta,\ \alpha, \beta > -1,$$

we can use the Gauss–Chebyshev formula or Gauss–Jacobi formula, respectively, of maximum degree of exactness $2m-1$. In many practical applications, however, the function f is not known at each point of the interval $[-1,1]$ but only at a finite number of nodes, often equispaced. In these cases, composite trapezoidal or composite Simpson rules, of degree of exactness 1, 3, respectively, are widely used, since all Newton–Cotes rules of higher order (greater than 7 for $w(x) = 1$) have weights which differ in sign and become rapidly unstable [68]. The problem of finding accurate quadrature rules based on equidistant points has gained the attention of some authors in more recent years (see, f.e., [68, 71, 77] and references therein). The approach by Huybrechs [68] is based on the idea of determining the quadrature weights in a least squares sense, by assuming that the number n of the equispaced nodes is greater than the degree of exactness d of the quadrature rule. The stability of the quadrature formula is guaranteed by the fact that, for $n \gg d$, the minimization process leads to positive weights. This idea was firstly proposed in 1970 by Wilson [99, 100] and very recently used by Glaubitz to construct stable high-order cubature formulas for experimental data [49]. Other approaches to get quadrature formulas from equispaced nodes have been presented in [71, 77, 25] and are based on the idea that it is possible to obtain quadrature

rules by using Gauss–Christoffel formulas in combination with local polynomial interpolants or global rational interpolants, respectively. The key point is to substitute the exact values $f(\xi_k)$ in (3.2) with the values, at ξ_k, of an approximating function. In this case the accuracy, the degree of exactness and the stability of the quadrature formulas are related to the accuracy, the degree of exactness and the stability of the approximating functions, respectively. In this chapter we get stable and accurate quadrature formulas on equispaced nodes, with high degree of exactness, by using Gaussian–Christoffel formulas and a mixed interpolation regression process, the so-called constrained mock-Chebyshev least squares approximation [24, 39], in combination. The result presented in this chapter can be found in [38, 29].

3.1 Constrained mock-Chebyshev least squares quadrature

In this Section, we use Gauss–Christoffel quadrature formulas and the constrained mock-Chebyshev least squares approximation, in combination, in order to get accurate and stable quadrature formulas with degree of exactness s, such that $m < s \leq 2m-1$, $m = \lfloor \pi\sqrt{n}/\sqrt{2} \rfloor$. In the Gauss–Christoffel quadrature of order m [48, Ch. 3]

$$\int_{-1}^{1} f(x)w(x)dx = \sum_{i=1}^{m} w_i f(\xi_i) + E_m(f), \tag{3.3}$$

the nodes $\xi_1, \ldots, \xi_m$ are the zeros of the m degree orthogonal polynomial $\pi_m(\cdot; w)$ belonging to the weight function w, thus

$$\pi_m(\xi_k; w) = 0, \qquad k = 1, 2, \ldots, m,$$

$$w_k = \int_{-1}^{1} \frac{\pi_m(t, w)}{(t-\xi_k)\pi_n'(\xi_k, w)} w(t)dt.$$

For $s \in \mathbb{N}$ s.t. $m < s \leq 2m-1$, we define the constrained mock-Chebyshev least squares quadrature method as follows

$$\int_{-1}^{1} f(x)w(x)dx = \hat{I}_{s,n}(f) + \hat{E}_{s,n}(f), \tag{3.4}$$

where

$$\hat{I}_{s,n}(f) = \sum_{i=1}^{m} w_i \hat{P}_{s,n}[f](\xi_i). \tag{3.5}$$

Despite formula (3.5) does not contain any explicit reference to the nodes of the set X_n, it is clear that it can be rewritten in terms of the evaluations of the integrand function f at all nodes of X_n. The crucial observations to get the explicit expression of the quadrature method (3.5) as a weighted sum of the sampled function values rely in the following results.

Theorem 3.1.1. *Let $s \in \mathbb{N}$ such that $m < s \leq 2m-1$ and let $\mathcal{B}_s = \{u_i : i = 0, \ldots, s\}$ be a basis of $\mathbb{P}_s(\mathbb{R})$. Then the operator*

$$\hat{P}_{s,n} : f \quad \mapsto \quad \hat{P}_{s,n}[f](x) = \sum_{i=0}^{s} \hat{a}_i u_i(x), \qquad x \in [-1, 1]$$

is a linear operator.

Proof. The linearity of the operator $\hat{P}_{s,n}$ follows from the fact that the vector of coefficients $\hat{a} = [\hat{a}_0, \ldots, \hat{a}_s]^T$ is the solution of the linear system (2.2). □

Theorem 3.1.2. *Let $s \in \mathbb{N}$ such that $m < s \leq 2m-1$ and let f be a continuous function in $[-1, 1]$, then*

$$\hat{P}_{s,n}[f] = \hat{P}_{s,n}\left[P_n[f]\right],$$

where

$$P_n[f](x) = \sum_{j=0}^{n} f(x_j)\ell_j(x), \qquad x \in [-1, 1]$$

and

$$\ell_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}, \qquad j = 0, \ldots, n, \qquad x \in [-1, 1].$$

Proof. It is sufficient to note that $f(x_i) = P_n[f](x_i)$ for each $x_i \in X_n$, $i = 0, \ldots, n$. □

Theorem 3.1.3. *The quadrature method* (3.5) *is a quadrature formula, that is*

$$\hat{I}_{s,n}(f) = \sum_{j=0}^{n} \hat{w}_j f(x_j), \tag{3.6}$$

where

$$\hat{w}_j = \sum_{i=1}^{m} w_i \hat{P}_{s,n}[\ell_j](\xi_i), \qquad j = 0, \ldots, n.$$

Proof. By using Theorem 3.1.1 and Theorem 3.1.2, we get

$$\hat{P}_{s,n}[f] = \hat{P}_{s,n}\left[\sum_{j=0}^{n} f(x_j)\ell_j\right] = \sum_{j=0}^{n} f(x_j)\hat{P}_{s,n}[\ell_j]. \tag{3.7}$$

By substituting (3.7) in (3.5), we obtain

$$\begin{aligned}
\hat{I}_{s,n}(f) = \sum_{i=1}^{m} w_i \hat{P}_{s,n}[f](\xi_i) &= \sum_{i=1}^{m} w_i \left(\sum_{j=0}^{n} f(x_j)\hat{P}_{s,n}[\ell_j](\xi_i) \right) \\
&= \sum_{j=0}^{n} \left(\sum_{i=1}^{m} w_i \hat{P}_{s,n}[\ell_j](\xi_i) \right) f(x_j) \\
&= \sum_{j=0}^{n} \hat{w}_j f(x_j).
\end{aligned}$$

□

As soon as the $PA = LU$ factorization of the KKT matrix is obtained, each weight $\hat{w}_j$, $j = 0, \ldots, n$, can be computed by solving two triangular systems through forward and back substitutions. These are stable or backward stable processes, in practice [95, Ch. 17,20,21,22]. The stability of the quadrature rule (3.6) can be measured by the ℓ^1-norm of its weights [68]

$$\kappa(n) = \sum_{j=0}^{n} |\hat{w}_j|. \tag{3.8}$$

In fact, by assuming that $\tilde{f}$ is a perturbation of the function f such that $\left\| \tilde{f} - f \right\|_\infty \leq \epsilon$, we get

$$\left| \sum_{j=0}^{n} \hat{w}_j \tilde{f}(x_j) - \sum_{j=0}^{n} \hat{w}_j f(x_j) \right| \leq \sum_{j=0}^{n} \left| \hat{w}_j \left(\tilde{f}(x_j) - f(x_j) \right) \right| \leq \sum_{j=0}^{n} |\hat{w}_j| \, \epsilon = \epsilon\kappa(n). \tag{3.9}$$

If all weights are positive, then $\kappa(n) = I(1)$ and the quadrature formula is stable. In all other cases $\kappa(n) > I(1)$ and the stability depends on the magnitude of $\kappa(n)$. In Table 3.1, we explicitly compute $\kappa(n)$ for different values of n, ranging from $n = 100$ to $n = 100000$, in the case of $w(x) = 1$. These values decrease from 2.1961 ($n = 100$) to 2.0079 ($n = 100000$). Therefore we can conclude, at least in the cases there specified, the stability of the constrained mock-Chebyshev least squares quadrature. This result is not surprising, since the quadrature formula (3.4) is obtained by combining two stable processes, namely the constrained mock-Chebyshev least squares approximation and the Gaussian–Christoffel quadrature.

n	100	500	1000	5000	10000	50000	100000
$\kappa(n)$	2.1961	2.0948	2.0773	2.0323	2.0252	2.0124	2.0079

Table 3.1: Computation of $\kappa(n)$ for different values of n, ranging from $n = 100$ to $n = 100000$, in the case of $w(x) = 1$.

The constrained mock-Chebyshev least squares quadrature has an high degree of exactness, which, however, can not overcome the degree of exactness of the Gaussian–Christoffel quadrature formula on m nodes, as shown in the following theorem.

Theorem 3.1.4. *Let $s \in \mathbb{N}$ such that $m < s \leq 2m-1$. Then the quadrature formula (3.4) has degree of exactness s.*

Proof. Let P_s be a polynomial of degree s. By the uniqueness of the constrained mock-Chebyshev least squares approximation [24, 39]

$$P_s(x) = \hat{P}_{s,n}\left[P_s\right](x), \qquad x \in [-1,1].$$

Therefore

$$\hat{E}_{s,n}\left(P_s\right) = E_m\left(P_s\right) = 0,$$

since the quadrature formula (3.3) has degree of exactness $2m-1$. □

Note that the quadrature formulas (3.4) provide different approximations of the exact integral $I(f)$ for different values of s, $m < s \leq 2m-1$. Clearly, as much $\hat{P}_{s,n}[f]$ well approximate, in the uniform norm, the integrand function f, as much accurate the quadrature formula (3.4) results. In fact we have

Theorem 3.1.5. *Let $s \in \mathbb{N}$ such that $m < s \leq 2m-1$, then*

$$|\hat{E}_{s,n}(f)| \leq \left\|\hat{R}_{s,n}[f]\right\|_\infty \int_{-1}^{1} w(x)dx, \tag{3.10}$$

where

$$\hat{R}_{s,n}[f] = f - \hat{P}_{s,n}[f].$$

Proof. By Theorem 3.1.4, we have

$$\left|\int_{-1}^{1} f(x)w(x)dx - \sum_{i=1}^{m} w_i \hat{P}_{s,n}[f](\xi_i)\right| = \left|\int_{-1}^{1} f(x)w(x)dx - \int_{-1}^{1} \hat{P}_{s,n}[f](x)w(x)dx\right|$$

$$\leq \int_{-1}^{1} \left|f(x) - \hat{P}_{s,n}[f](x)\right| w(x)dx \leq \left\|f - \hat{P}_{s,n}[f]\right\|_\infty \int_{-1}^{1} w(x)dx.$$

□

From now on, we consider the value of r given in (1.12). We notice that, for this value of r, the bound (1.13) implies that $m < r \leq 2m-1$, therefore both previous results hold. Moreover, the bound for $\left\|\hat{R}_{r,n}[f]\right\|_\infty$ developed in [24] can be used to estimate the error (3.10), if some information about f and its derivatives are known. Since we assume that the function f is known only at the nodes of an equispaced grid, we propose an alternative method to estimate the error $\hat{E}_{r,n}(f)$, which avoids the knowledge of the analytic expression of the integrand function f. To this aim, by assuming $\hat{I}_{r,n}(f) \neq 0$, we set

$$\tilde{E}^{rel,k}_{r,n}(f) = \max\left\{\frac{\left|\hat{I}_{s+1,n}(f) - \hat{I}_{s,n}(f)\right|}{\left|\hat{I}_{r,n}(f)\right|} \,:\, s \in \{r-k,\ldots,r+k\}\right\}, \qquad 0 < k < p. \tag{3.11}$$

The setting (3.11) arises from the analysis of the trend of the approximations $\hat{I}_{s,n}(f)$. These approximations, for a small number of subsequent values of s, can be excessively close each other. The numerical

results, provided in the next section, show that the value $k = 3$ provides quite good estimate of the exact relative error

$$\hat{E}^{rel}_{r,n}(f) = \frac{\left|\hat{E}_{r,n}(f)\right|}{|I(f)|}. \tag{3.12}$$

An algorithm for computing the constrained mock-Chebyshev least squares quadrature can be organized as follows.

Algorithm 1 Constrained mock-Chebyshev least squares quadrature

Input: $X_n = [x_0, \ldots, x_n]^T, f = [f_0, \ldots, f_n]^T, k$
Output: $\tilde{I}_{r,n}(f)$, $\tilde{E}^{rel,k}_{r,n}(f)$
1: Compute m, r
2: Compute the mock-Chebyshev subset X'_m
3: Compute $X''_{n-m} = X_n \setminus X'_m$
4: Set $X_n = [X'_m, X''_{n-m}]$
5: Compute the Gauss-Christoffel nodes and weights of order m
6: **for** $s = r - k : r + k$ **do**
7: Compute $\hat{P}_{s,n}[f]$
8: Compute $\hat{I}_{s,n}(f)$
9: **end for**
10: Compute $\tilde{E}^{rel,k}_{r,n}(f)$

3.1.1 Numerical experiments

In this Section we compute approximations of the integral (3.1) by using the quadrature formula (3.4) with r given in (1.12), where the polynomial $\hat{P}_{r,n}[f]$ is expressed in the Chebyshev polynomial basis of the first kind. To this aim, we consider the grid of 1001 equispaced nodes in $[-1, 1]$, that is $n = 1000$, from which we get $m = 70$, $p = 28$ and then $r = 99$. In order to demonstrate the effectiveness of the estimate of the exact relative error $\hat{E}^{rel}_{r,n}(f)$ through $\tilde{E}^{rel,k}_{r,n}(f)$ $(k = 3)$, we assume that the vector of the functional data $f = [f_0, \ldots, f_n]^T$ results from the evaluation of the following functions

$$f_1(x) = \frac{1}{1 + 8x^2}, \quad f_2(x) = \frac{1}{1 + 25x^2}, \quad f_3(x) = \cos(20x), \quad f_4(x) = 1 + x^{120},$$

on the points of the equispaced grid $X_n = [x_0, \ldots, x_n]^T$. Moreover we use the following weight functions

	name	weight function $w(x)$
1	Gauss–Legendre	1
2	Gauss–Chebyshev	$(1 - x^2)^{-1/2}$
3	Gauss–Jacobi	$(1 - x)^{\alpha}(1 + x)^{\beta}$, $\alpha, \beta > -1$

In Tables 3.2 - 3.4, we can appreciate the accuracy of approximations provided by the constrained mock-Chebyshev least squares quadrature formula (3.5) which, we emphasize, uses the functional data only at the equispaced nodes. In the second column we list the exact relative errors (3.12), while in the third column we list their estimate through formula (3.11), which, as before, uses the functional data only at the equispaced nodes.

Finally, we numerically test the stability of the proposed method by analyzing the sensitivity of the constrained mock-Chebyshev least squares quadrature formula to random perturbations of the function values. More precisely, in the case of the classical Runge function $f_2(x)$, in Table 3.5 we compare left and right side of the inequality (3.9) by setting $\epsilon = 10^{-8}$, $n = 1000$ and by using different weight functions.

	$\hat{E}^{rel}_{r,n}(f)$	$\tilde{E}^{rel,3}_{r,n}(f)$
$f_1(x)$	3.8265e-16	6.3775e-16
$f_2(x)$	1.3949e-10	1.4032e-10
$f_3(x)$	1.8241e-15	1.2161e-15
$f_4(x)$	0	2.2022e-16

Table 3.2: Absolute relative errors using the constrained mock-Chebyshev least squares quadrature formula (3.4) with Gauss–Legendre weights and their estimates using formula (3.11) with $k = 3$.

	$\hat{E}^{rel}_{r,n}(f)$	$\tilde{E}^{rel,3}_{r,n}(f)$
$f_1(x)$	8.6935e-15	2.5444e-15
$f_2(x)$	1.5671e-09	3.5801e-09
$f_3(x)$	7.0880e-14	8.4633e-16
$f_4(x)$	5.6402e-14	2.6356e-16

Table 3.3: Absolute relative errors using the constrained mock-Chebyshev least squares quadrature formula (3.4) with Gauss–Chebyshev weights and their estimates using formula (3.11) with $k = 3$.

	$\hat{E}^{rel}_{r,n}(f)$	$\tilde{E}^{rel,3}_{r,n}(f)$
$f_1(x)$	4.2407e-16	8.4815e-16
$f_2(x)$	9.4603e-11	3.8728e-11
$f_3(x)$	3.6023e-14	9.7492e-15
$f_4(x)$	2.8238e-16	2.8238e-16

Table 3.4: Absolute relative errors using the constrained mock-Chebyshev least squares quadrature formula (3.4) with Gauss–Jacobi weights with $\alpha = \beta = 1/2$ and their estimates using formula (3.11) with $k = 3$.

$w(x)$	$\left\lvert \hat{I}_{r,n}[\tilde{f}_2] - \hat{I}_{r,n}[f_2] \right\rvert$	$\epsilon\kappa(n)$
1	3.00e-09	9.82e-08
$(1-x^2)^{-1/2}$	2.90e-09	6.97e-08
$(1-x)^{\frac{1}{2}}(1+x)^{\frac{1}{2}}$	2.19e-09	1.75e-08

Table 3.5: Sensitivity of the constrained mock-Chebyshev least squares quadrature formula (3.4) to random perturbations of the function values, in the case of $f_2(x) = 1/(1+25x^2)$, $n = 1000$ and by using different weight functions.

3.2 An adaptive algorithm for determining the optimal degree of regression in constrained mock-Chebyshev least squares quadrature

The accuracy of the quadrature formulas (3.5) varies with s depending on the degree of smoothness of the function f. In this Section, we develop an adaptive algorithm for determining, given a function f, the optimal degree of regression which corresponds to the quadrature formula with higher accuracy. In other words, given the family of quadrature formulas

$$\left\{ \hat{I}_{s,n}(f) = \sum_{i=1}^{m} w_i \hat{P}_{s,n}[f](\xi_i), \qquad s = m, \ldots, 2m-1 \right\}, \tag{3.13}$$

we determine the *optimal* degree $r^\star_{opt}$ of the constrained mock-Chebyshev least squares approximation which produces a more accurate quadrature formula $\hat{I}_{r^\star_{opt},n}(f)$.

3.3 Computing accurate quadrature formulas with high degree of exactness from equispaced nodes

The main goal of this section is the determination of a procedure for the choice of the *optimal* value of s which guarantees the *best approximation accuracy* of the quadrature formula $\hat{I}_{s,n}(f)$, measured through the *exact relative error*

$$\hat{E}^{rel}_{s,n}(f) = \frac{\left|\hat{E}_{s,n}(f)\right|}{|I(f)|},$$

where we assume $|I(f)| > 0$. We denote this value by $r^\star_{opt} = r^\star_{opt}(f)$. To this aim, we analyze the trend of *approximate relative errors*

$$\tilde{E}^{rel}_{s,n}(f) = \frac{\left|\hat{I}_{s+1,n}(f) - \hat{I}_{s,n}(f)\right|}{\left|\hat{I}_{s,n}(f)\right|}, \qquad m \le s \le 2m-2, \tag{3.14}$$

computed by using quadrature formulas of subsequent degrees up to the maximum degree of exactness $2m-1$. At first sight, it might be thought to choose $r^\star_{opt}$ as the value of $s \in \{m, m+1, \ldots, 2m-2\}$ which minimizes the approximate relative error $\tilde{E}^{rel}_{s,n}(f)$. Unfortunately, in general this choice could be misleading since, even for starting values of s, it could occur that two successive approximations $\hat{I}_{s,n}(f)$ and $\hat{I}_{s+1,n}(f)$ are so close to each other that the approximate relative error $\tilde{E}^{rel}_{s,n}(f)$ is very small, for example less than a tolerance *tol*, despite the exact relative error $\hat{E}^{rel}_{s,n}(f)$ is not, being much greater than *tol*. An example of this situation is well illustrated in Figure 3.1, where the sequence of approximate relative errors assumes values less than 10^{-14} despite all exact relative errors are not less than 10^{-6}. The approximate relative errors $\tilde{E}^{rel}_{s,n}(f)$ less than *tol* are then outliers and therefore they have to be discarded in the process of the determination of $r^\star_{opt}$. Instead of fixing a tolerance *tol* a priori, we distinguish outliers from valid values of relative errors $\tilde{E}^{rel}_{s,n}(f)$ by analyzing the sequence of consecutive triples

$$t_s = \left\{ \tilde{E}^{rel}_{s,n}(f), \tilde{E}^{rel}_{s+1,n}(f), \tilde{E}^{rel}_{s+2,n}(f) \right\}, \qquad s = m, \ldots, 2m-3. \tag{3.15}$$

We call the triple t_s monotonic if and only if

$$\tilde{E}^{rel}_{s,n}(f) \ge \tilde{E}^{rel}_{s+1,n}(f) \ge \tilde{E}^{rel}_{s+2,n}(f) \qquad \text{or} \qquad \tilde{E}^{rel}_{s,n}(f) \le \tilde{E}^{rel}_{s+1,n}(f) \le \tilde{E}^{rel}_{s+2,n}(f),$$

otherwise we call the triple t_s non monotonic. Note that a triple t_s is monotonic if and only if

$$d_s d_{s+1} \ge 0,$$

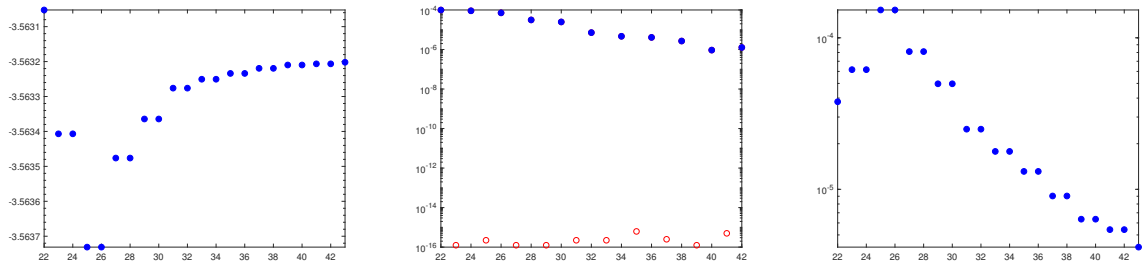


Figure 3.1: Sequence of approximations $\hat{I}_{s,n}(f)$ (left) and sequence of approximate relative errors $\tilde{E}^{rel}_{s,n}(f)$ (center) versus sequence of exact relative errors $\hat{E}^{rel}_{s,n}(f)$ (right) with $f(x) = \frac{1}{x^2-1.1}$, $w(x) = 1$, $n = 100$ and $m = 22$.

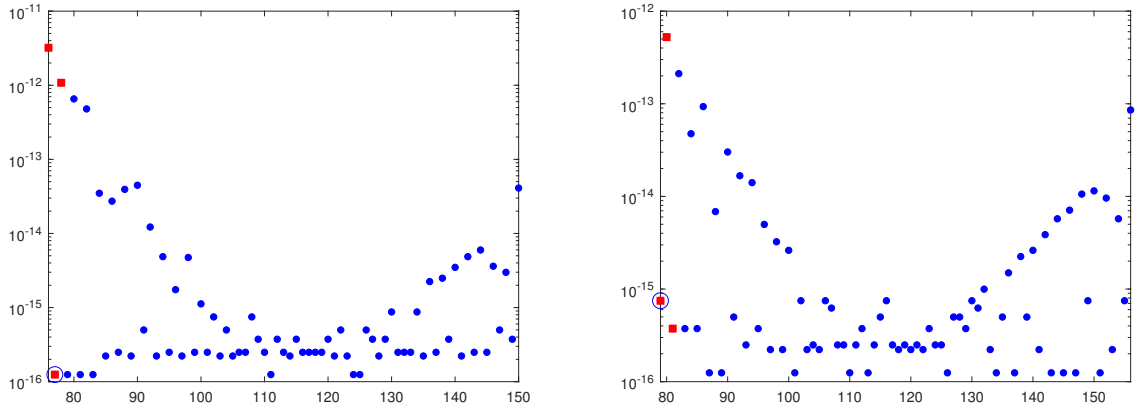


Figure 3.2: Approximate relative errors, with respect to the Gauss–Legendre scheme ($w(x) = 1$), for the function $f(x) = \frac{1}{x^2-1.1}$ (●) with $n = 1200$, $m = 76$ (left) and $n = 1280$, $m = 79$ (right) with related first non monotonic triples (■) and corresponding outliers (○).

where we set

$$d_s = \log_{10}\left(\tilde{E}^{rel}_{s+1,n}(f)\right) - \log_{10}\left(\tilde{E}^{rel}_{s,n}(f)\right), \qquad s = m, \ldots, 2m-2.$$

We fix a suitable constant $\delta > 0$ (for example $\delta = 0.5$) and we search the outliers among the elements of non monotonic triples by assuming that:

- if $d_s < -\delta$ and $d_{s+1} > \delta$, then $\tilde{E}^{rel}_{s+1,n}(f)$ is outlier (see Figure 3.2 (left));
- if $d_s > \delta$ and $d_{s+1} < -\delta$, then $\tilde{E}^{rel}_{s,n}(f)$ is outlier (see Figure 3.2 (right)).

The process of determining new outliers ends when $s = 2m-2$. The tolerance *tol* is determined adaptively by initializing it with the epsilon machine *eps* and by updating it by the rule

$$tol = \max\{tol, \tilde{E}^{rel}_{s,n}(f)\} \tag{3.16}$$

as soon as we find a new outlier $\tilde{E}^{rel}_{s,n}(f)$. To avoid under-estimation of the exact relative error we take into account the possible presence of outliers, by assuming as outliers all approximate relative errors $\tilde{E}^{rel}_{s,n}(f)$ less than or equal to the computed tolerance *tol*. The Algorithm 2 computes the tolerance *tol*. The Algorithm 3 detects the outliers and remove them, providing in output the increasing sequence $\{s_1, \ldots, s_p\} \subset \{m, \ldots, 2m-2\}$ of degrees s such that $\tilde{E}^{rel}_{s,n}(f) > tol$. In Figure 3.3 we display the effect of the Algorithm 3, in detecting outliers (left) and removing them (right).

Algorithm 2 Tolerance determination

Input: $\tilde{E}^{rel}_{m,n}(f), \ldots, \tilde{E}^{rel}_{2m-2,n}(f)$
Output: tol
  $tol \leftarrow eps$
  for $i = m, \ldots, 2m-3$ **do**
    $d_i \leftarrow \log_{10}(\tilde{E}^{rel}_{i+1,n}(f)) - \log_{10}(\tilde{E}^{rel}_{i,n}(f))$
  end for
  while $j \leq m-2$ **do**
    if $d_j \leq -\delta$ and $d_{j+1} \geq \delta$ **then**
      $tol = \max\{tol, \tilde{E}^{rel}_{j+1,n}(f)\}$
      $j \leftarrow j+2$
    else if $d_j \geq \delta$ and $d_{j+1} \leq -\delta$ **then**
      $tol = \max\{tol, \tilde{E}^{rel}_{j,n}(f)\}$
      $j \leftarrow j+1$
    else
      $j \leftarrow j+2$
    end if
  end while

Algorithm 3 Outlier detection

Input: $\tilde{E}^{rel}_{m,n}(f), \ldots, \tilde{E}^{rel}_{2m-2,n}(f)$
Output: $s_1, \ldots, s_p$
  $j \leftarrow 1$
  for $i = m, \ldots, 2m-2$ **do**
    if $\tilde{E}^{rel}_{i,n}(f) > tol$ **then**
      $s_j \leftarrow i$
      $j \leftarrow j+1$
    end if
  end for

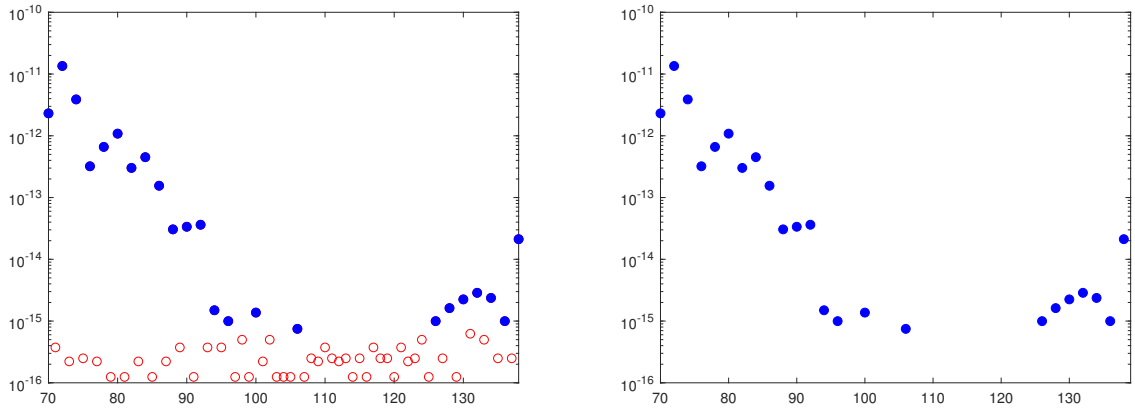


Figure 3.3: Significant approximate relative errors (•) and outliers (○) for the function $f(x) = \frac{1}{x^2-1.1}$ $n = 1000$, $m = 70$ relative to the Gauss–Legendre scheme ($w(x) = 1$).

From now on we assume that all outliers in the sequence $\{(s, \tilde{E}^{rel}_{s,n})\}_{s=m}^{2m-2}$ have been removed. We denote by $\{(s_j, \tilde{E}^{rel}_{s_j,n})\}_{j=1}^{p}$, $m \le s_1 < \cdots < s_p \le 2m-2$ the subset of significant data. We consider the sequence of intervals $\{\mathcal{I}_j\}_{j=0}^{p}$, defined as follows

$$\mathcal{I}_j = \begin{cases} [m, s_1), & \text{if } j = 0, \\ (s_j, s_{j+1}), & \text{if } j = 1, \ldots, p-1, \\ (s_p, 2m-2], & \text{if } j = p. \end{cases} \tag{3.17}$$

By definition, all outliers belong to

$$\mathcal{I} = \bigcup_{j=0}^{p} \mathcal{I}_j.$$

Let $q \ge 0$ be the number of intervals $\mathcal{I}_j$ in (3.17) containing at least one outlier.

Case $q > 0$. We denote by $\mathcal{I}_{j_k}$, $0 \le j_1 < j_2 < \cdots < j_q \le p$ the intervals containing at least one outlier and by $\mathcal{N}_{j_k}$ the number of outliers in $\mathcal{I}_{j_k}$. We set

$$\mu = \frac{1}{q}\sum_{k=1}^{q} \mathcal{N}_{j_k}, \quad \sigma = \sqrt{\frac{1}{q}\sum_{k=1}^{q} (\mathcal{N}_{j_k} - \mu)^2}.$$

As well-known, the standard deviation σ tells us the typical amount by which the values $\{\mathcal{N}_{j_\ell}\}$ deviate from their average value μ. We define

$$\mathcal{R} = \{ r_{j_\ell} \,:\, \ell = 1, \ldots, q \wedge \mathcal{N}_{j_\ell} > \mu + \sigma \},$$

and we set

$$r^{\star}_{opt} = \begin{cases} \min \mathcal{R} & \text{if } \mathcal{R} \neq \emptyset, \\ 2m-2 & \text{if } \mathcal{R} = \emptyset. \end{cases}$$

By the nature of the constrained mock-Chebyshev least squares approximation [24, 39], the case $r^{\star}_{opt} < 2m-2$ frequently occurs as soon as the exact relative error $\hat{E}^{rel}_{r^{\star}_{opt},n}(f)$ is near to the machine precision already for values of $s \ll 2m-2$. In such cases, by increasing the degree of the regression $s \ge r^{\star}_{opt}$, it is also possible that the exact relative error $\hat{E}^{rel}_{r,n}(f)$ became worse. In fact, as s approaches to n, the polynomial $\hat{P}_{s,n}[f](x)$ *tends* to the polynomial interpolant on the set of nodes X_n, $\hat{P}_{n,n}[f](x)$, which in its turn, can suffer from the Runge phenomenon. If $\mathcal{R} \neq \emptyset$, from the definition of $r^{\star}_{opt}$, we expect a not increasing trend of significant data $\tilde{E}^{rel}_{s_1,n}(f), \ldots, \tilde{E}^{rel}_{r^{\star}_{opt},n}(f)$ if $r^{\star}_{opt} > s_1$ or a non decreasing trend of the significant data $\tilde{E}^{rel}_{r^{\star}_{opt},n}(f), \ldots, \tilde{E}^{rel}_{s_p,n}(f)$ if $r^{\star}_{opt} = s_1$. If $\mathcal{R} = \emptyset$ nothing can be said on the trend of the significant data.

Case $q = 0$. We set $r^{\star}_{opt} = 2m-2$. In this case nothing can be said on the trend of the significant data.

We are now able to determine a value $r^{\star}_{opt}$ of the degree of regression which produces more accurate quadrature formulas. The accuracy of $\hat{I}_{r^{\star}_{opt},n}(f)$ will depend on the quality of the mock-Chebyshev constrained least squares approximation to the function f. We distinguish the following cases:

- if $r^{\star}_{opt} = s_1$, we use a linear regression $l(s)$ to model the trend of the data

  $$\left\{ \log_{10}\left(\tilde{E}^{rel}_{r^{\star}_{opt},n}(f)\right), \ldots, \log_{10}\left(\tilde{E}^{rel}_{s_p,n}(f)\right) \right\}$$

  and we set

  $$r^{\star}_{opt} = s_k,$$

  where

  $$\log_{10}\left(\tilde{E}^{rel}_{s_k,n}\right) = \min_{\alpha \in \{1,\ldots,p\}} \left\{ \log_{10}\left(\tilde{E}^{rel}_{s_\alpha,n}\right) : \log_{10}\left(\tilde{E}^{rel}_{s_\alpha,n}\right) - l(s_\alpha) \ge 0 \right\};$$

- if $r^\star_{opt} = s_j$ with $j = 2, \ldots, p-1$, then if $\log_{10}\left(\tilde{E}^{rel}_{s_j,n}\right) - \log_{10}\left(\tilde{E}^{rel}_{s_{j+1},n}\right) > \delta$, we set $r^\star_{opt} = s_{j+1}$, else we use a linear regression $l(s)$ to model the trend of the data

$$\left\{\log_{10}\left(\tilde{E}^{rel}_{s_1,n}(f)\right), \ldots, \log_{10}\left(\tilde{E}^{rel}_{r^\star_{opt},n}(f)\right)\right\}$$

and we set

$$r^\star_{opt} = s_k,$$

where

$$\log_{10}\left(\tilde{E}^{rel}_{s_k,n}\right) = \min_{\alpha\in\{1,\ldots,j\}} \left\{\log_{10}\left(\tilde{E}^{rel}_{s_\alpha,n}\right) : \log_{10}\left(\tilde{E}^{rel}_{s_\alpha,n}\right) - l(s_\alpha) \geq 0\right\};$$

- if $r^\star_{opt} = s_p$, we use a linear regression $l(s)$ to model the trend of the data

$$\left\{\log_{10}\left(\tilde{E}^{rel}_{s_1,n}(f)\right), \ldots, \log_{10}\left(\tilde{E}^{rel}_{r^\star_{opt},n}(f)\right)\right\}$$

and we set

$$r^\star_{opt} = s_k,$$

where

$$\log_{10}\left(\tilde{E}^{rel}_{s_k,n}\right) = \min_{\alpha\in\{1,\ldots,j\}} \left\{\log_{10}\left(\tilde{E}^{rel}_{s_\alpha,n}\right) : \log_{10}\left(\tilde{E}^{rel}_{s_\alpha,n}\right) - l(s_\alpha) \geq 0\right\}.$$

Algorithm 4 Adaptive algorithm for determining a quadrature formulas with high degree of exactness and accuracy from equispaced nodes

Input: $X_n = [x_0, \ldots, x_n]^T, f = [f_0, \ldots, f_n]^T$
Output: $\hat{I}_{r^\star_{opt},n}(f)$
1: Compute m
2: Compute the mock-Chebyshev subset X'_m
3: Compute $X''_{n-m} = X_n \setminus X'_m$
4: Set $X_n = [X'_m, X''_{n-m}]$
5: Compute the Gauss–Christoffel nodes and weights of order m
6: **for** $s = m : 2m-1$ **do**
7: Compute $\hat{P}_{s,n}[f]$
8: Compute $\hat{I}_{s,n}(f)$
9: **end for**
10: Compute the approximate relative errors $\tilde{E}^{rel}_{m,n}(f), \ldots, \tilde{E}^{rel}_{2m-2,n}(f)$
11: Run Algorithm 2
12: Run Algorithm 3
13: Compute $r^\star_{opt}$

3.3.1 Computational cost

We determine the computational cost of the Algorithm 4 described above. The computational cost of m is negligible. By using the procedure proposed in [10], the selection of the mock-Chebyshev subset from the uniform grid X_n requires about $O(nm)$ flops. The reordering of the set X_n involves a searching algorithm for the computation of the set X''_{n-m}, whose computational cost is $\mathcal{O}(m\log(n))$ flops. Gauss-Christoffel nodes and weights can be computed through the Chebfun package [66]. The function `legpts` to compute Legendre nodes and weights, used in the numerical experiments, requires $\mathcal{O}\left(\frac{m(\log m)^2}{\log(\log m)}\right)$ flops. The computation of the coefficients of the polynomial $\hat{P}_{r,n}[f]$ through the Lagrange multipliers method [39] needs the reordering in point 3 and the computation of the Chebyshev polynomial basis, whose total cost is $\mathcal{O}(n^2)$ flops [11]. Since we must compute the polynomial $\hat{P}_{s,n}[f]$ for each $s \in \{m, \ldots, 2m-1\}$, the computational cost is $\mathcal{O}(mn^2)$ flops. The tolerance is computed through the Algorithm 2 and the detection

of the outliers is made by using the Algorithm 3, which both require $\mathcal{O}(m)$ flops. Finally, we determine the degree of regression $r^\star_{opt}$ which produces accurate quadrature formulas by using the procedure described in the Section 3.3, which requires $\mathcal{O}(n)$ flops. Since $m = \mathcal{O}(\sqrt{n})$, the computational cost of the procedure is $\mathcal{O}(n^{5/2})$ flops.

3.4 Computing accurate cubature formulas with high degree of exactness from regular grids of nodes

Let f be a continuous function in the square $[-1,1]^2$. In line with the notations of Chapter 2, we set $\boldsymbol{n}_{x,y} = (n_x, n_y) \in \mathbb{N}\times\mathbb{N}$ and consider the uniform grid of $(n_x+1)\times(n_y+1)$ equispaced points $X_{n_x}\times Y_{n_y}$ in the square $[-1,1]^2$. In analogy with the univariate case, we set

$$m_x = \left\lfloor \pi\sqrt{\frac{n_x}{2}} \right\rfloor, \quad m_y = \left\lfloor \pi\sqrt{\frac{n_y}{2}} \right\rfloor,$$

$\boldsymbol{m}_{x,y} = (m_x, m_y)$ and we fix $\boldsymbol{s}_{x,y} = (s_x, s_y) \in \mathbb{N}\times\mathbb{N}$ such that $m_x \le s_x \le n_x$ and $m_y \le s_y \le n_y$. We denote by

$$\hat{P}_{\boldsymbol{s}_{x,y},\boldsymbol{n}_{x,y}}[f] = \hat{P}_{(s_x,s_y),(n_x,n_y)}[f], \tag{3.18}$$

the tensor product extension of the polynominal $\hat{P}_{s,n}[f]$ and we consider the quadrature formulas

$$\begin{aligned} I(f) &= \int_{-1}^{1}\int_{-1}^{1} w(s,t)f(s,t)dsdt \approx \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} w_i\kappa_j f(\xi_i,\eta_j) \\ &\approx \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} w_i\kappa_j \hat{P}_{\boldsymbol{s}_{x,y},\boldsymbol{n}_{x,y}}[f](\xi_i,\eta_j) =: \hat{I}_{\boldsymbol{s}_{x,y},\boldsymbol{n}_{x,y}}(f), \end{aligned} \tag{3.19}$$

where $\xi_1,\dots,\xi_{m_x}$ and $\eta_1,\dots,\eta_{m_y}$ are nodes of a Gaussian quadrature formula with weights $w_1,\dots,w_{m_x}$ and $\kappa_1,\dots,\kappa_{m_y}$ of order m_x and m_y, respectively. We consider the family of quadrature formulas

$$\hat{I}_{\boldsymbol{s},\boldsymbol{n}}(f), \quad \boldsymbol{s} = (s,s) \in \mathbb{N}\times\mathbb{N}, \quad \max\{m_x,m_y\} \le s \le \min\{2m_x-1, 2m_y-1\} \tag{3.20}$$

and the approximate relative errors

$$\tilde{E}^{rel}_{\boldsymbol{s},\boldsymbol{n}}(f) = \frac{\left|\hat{I}_{\boldsymbol{s}+\boldsymbol{1},\boldsymbol{n}}(f) - \hat{I}_{\boldsymbol{s},\boldsymbol{n}}(f)\right|}{\left|\hat{I}_{\boldsymbol{s},\boldsymbol{n}}(f)\right|}, \quad \max\{m_x,m_y\} \le s \le \min\{2m_x-1, 2m_y-1\}. \tag{3.21}$$

In analogy to the univariate case, by using the Algorithm 4, it is possible to determine a value $r^\star_{opt}$ of the degree of regression which produces accurate quadrature formulas.

3.4.1 Numerical experiments

We compute the approximation of the integral (3.1) by using the quadrature formula (3.13) with $s = r^\star_{opt}$, where the polynomial $\hat{P}_{r^\star_{opt},n}[f]$ is expressed in the Chebyshev polynomial basis of the first kind $\mathcal{B}^C_{r^\star_{opt}}$. To this aim, we consider the grid of 1001 equispaced nodes in $[-1,1]$, that is $n = 1000$, $m = 70$. The experiments are performed on the following functions

$$f_1(x) = \frac{1}{1+8x^2}, \qquad f_2(x) = \frac{1}{1+25x^2}, \qquad f_3(x) = \frac{1}{((t+1)^4 + (2/50)^2)},$$

$$f_4(x) = e^{-x^2}, \qquad f_5(x) = \frac{1}{x^4 + (\frac{\sqrt{26}}{5} - 1)x^2 + (\frac{13}{50})^2}, \qquad f_6(x) = \frac{1}{t+1.01},$$

by using the weight function $w(x) = 1$.

	$E_T(f)$	$E_{CS}(f)$	$E_M(f)$	$\hat{E}^{rel}_{r,n}(f)$	$\hat{E}^{rel}_{r^\star_{opt},n}(f)$	$\tilde{E}^{rel}_{r^\star_{opt},n}(f)$
$f_1(x)$	1.33e-04	8.88e-10	7.84e-13	2.55e-16	0	2.55e-16
$f_2(x)$	8.97e-08	1.51e-14	1.02e-12	1.39e-10	4.13e-12	1.98e-12
$f_3(x)$	3.00e-10	4.09e-16	1.06e-10	3.75e-13	1.59e-14	6.34e-15
$f_4(x)$	3.28e-07	8.77e-15	4.44e-16	5.94e-16	5.94e-16	2.97e-16
$f_5(x)$	1.44e-07	5.41e-14	1.24e-13	2.24e-16	7.86e-16	8.99e-16
$f_6(x)$	6.26e-04	6.14e-07	1.52e-06	1.67e-07	8.81e-09	9.26e-10

Table 3.6: Comparisons among the relative errors in trapezoidal composite rule ($E_T(f)$), Cavalieri–Simpson rule ($E_{CS}(f)$), quadrature formula proposed in [77] ($E_M(f)$), quadrature formula through the constrained mock-Chebyshev interpolant with optimal degree relative to the Gauss–Legendre scheme ($\hat{E}^{rel}_{r,n}(f)$) and the approximate relative error obtained through Algorithm 4 ($\hat{E}^{rel}_{r^\star_{opt},n}$).

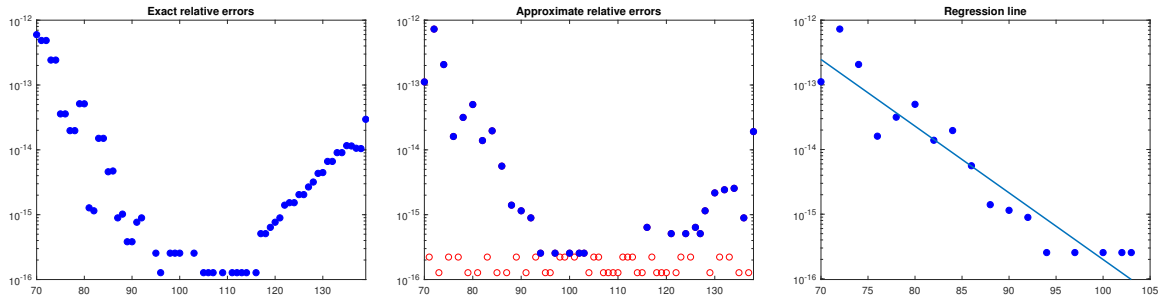


Figure 3.4: From left to right. Exact relative errors, approximate relative errors (•) with discarded outliers (○) and regression line of the significant data for the function $f_1(x)$ relative to the Gauss–Legendre scheme.

In Table 3.6, from left to right, we compare the relative errors obtained by applying the trapezoidal composite rule ($E_T(f)$), the Cavalieri–Simpson composite rule ($E_{CS}(f)$), the quadrature formula proposed in [77] with $s = 6$ ($E_M(f)$), the constrained mock-Chebyshev least squares quadrature formula proposed in [38] ($\hat{E}^{rel}_{r,n}(f)$) and the proposed here quadrature formula ($\hat{E}^{rel}_{r^\star_{opt},n}(f)$). To appreciate the accuracy of the estimate of the exact relative error, obtained through the Algorithm 4, in the last column we report also $\tilde{E}^{rel}_{r^\star_{opt},n}(f)$. To show the efficacy of the Algorithm 4 in computing the optimal regression degree $r^\star_{opt}$, in Figures 3.4 - 3.9 we display the sequences of exact relative errors, approximate relative errors (with discarded outliers, in red) and the regression line of the significant data, for all test functions.

Now, we consider the grid of 151×151 equispaced nodes in $[-1, 1]^2$, that is $n_x = n_y = 150$, $m_x = m_y = 27$. The experiments are perfomed by using the Gauss–Legendre weight on the well-known Franke's function

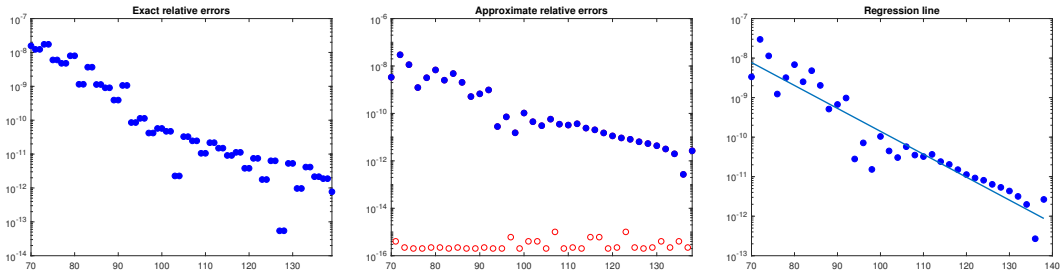


Figure 3.5: From left to right. Exact relative errors, approximate relative errors (•) with discarded outliers (○) and regression line of the significant data for the function $f_2(x)$ relative to the Gauss–Legendre scheme.

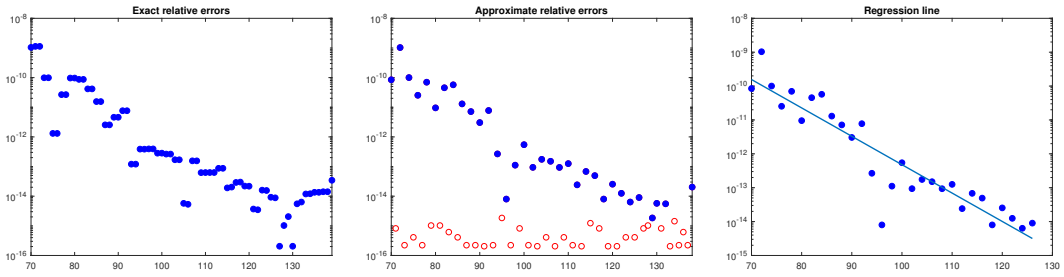


Figure 3.6: From left to right. Exact relative errors, approximate relative errors (•) with discarded outliers (○) and regression line of the significant data for the function $f_3(x)$ relative to the Gauss–Legendre scheme.

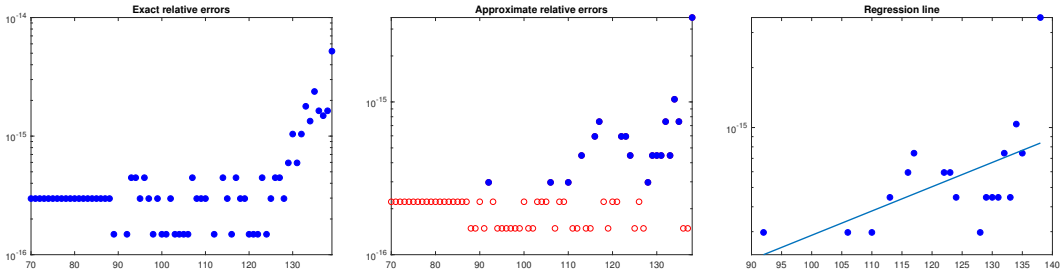


Figure 3.7: From left to right. Exact relative errors, approximate relative errors (•) with discarded outliers (○) and regression line of the significant data for the function $f_4(x)$ relative to the Gauss–Legendre scheme.

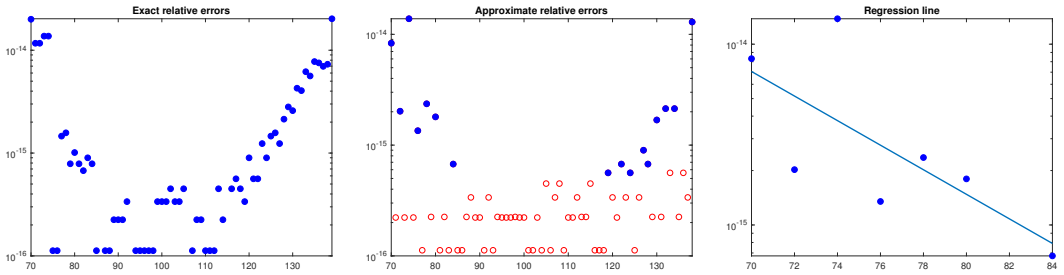


Figure 3.8: From left to right. Exact relative errors, approximate relative errors (•) with discarded outliers (○) and regression line of the significant data for the function $f_5(x)$ relative to the Gauss–Legendre scheme.

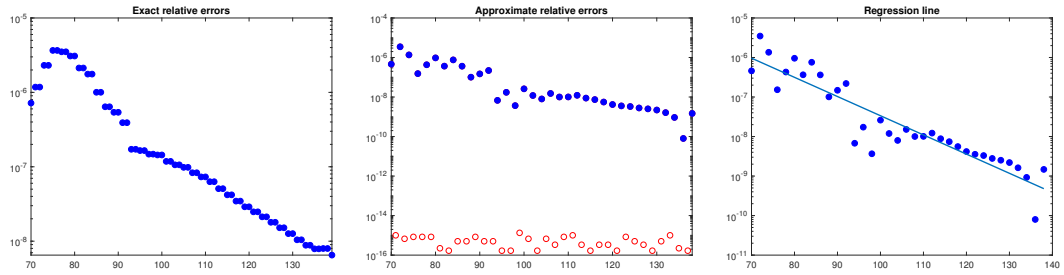


Figure 3.9: From left to right. Exact relative errors, approximate relative errors (•) with discarded outliers (○) and regression line of the significant data for the function $f_6(x)$ relative to the Gauss–Legendre scheme.

	$E_T(f)$	$E_{CS}(f)$	$\hat{E}^{rel}_{\boldsymbol{r}^\star_{opt},\boldsymbol{n}}(f)$	$\tilde{E}^{rel}_{\boldsymbol{r}^\star_{opt},\boldsymbol{n}}(f)$
$f(x,y)$	2.30e-05	1.71e-10	4.69e-12	2.46e-11

Table 3.7: Comparisons among the relative errors in trapezoidal composite rule ($E_T(f)$), Cavalieri–Simpson rule ($E_{CS}(f)$), cubature formulas through the constrained mock-Chebyshev tensor product interpolant with optimal degree ($\hat{E}^{rel}_{\boldsymbol{r}^\star_{opt},\boldsymbol{n}}(f)$) and approximate relative error obtained through Algorithm 4 ($\tilde{E}^{rel}_{\boldsymbol{r}^\star_{opt},\boldsymbol{n}}(f)$) relative to the Gauss–Legendre scheme.

$$
\begin{aligned}
f(x,y) = 0.75\exp\Big(-\frac{(9(x+1)/2-2)^2}{4} - \frac{(9(y+1)/2-2)^2}{4}\Big)\\
+0.75\exp\Big(-\frac{(9(x+1)/2+1)^2}{49} - \frac{(9(y+1)/2+1)}{10}\Big)\\
+0.5\exp\Big(-\frac{(9(x+1)/2-7)^2}{4} - \frac{(9(y+1)/2-3)^2}{4}\Big)\\
-0.2\exp\Big(-(9(x+1)/2-4)^2 - (9(y+1)/2-7)^2\Big),
\end{aligned}
$$

where the polynomial $\hat{P}_{\boldsymbol{s}_{x,y},\boldsymbol{n}_{x,y}}[f]$ is expressed in the basis $\mathcal{B}^C_{s_x} \otimes \mathcal{B}^C_{s_y}$. In Table 3.7, from left to right, we compare the relative errors obtained by applying the tensor product trapezoidal composite rule ($E_T(f)$), the tensor product Cavalieri–Simpson composite rule ($E_{CS}(f)$) and the proposed here quadrature formula ($\hat{E}^{rel}_{\boldsymbol{r}^\star_{opt},\boldsymbol{n}}(f)$). To appreciate the accuracy of the estimate of the exact relative error, obtained through the Algorithm 4, in the last column we report also $\tilde{E}^{rel}_{\boldsymbol{r}^\star_{opt},\boldsymbol{n}}(f)$.

Chapter 4

Polynomial approximation of derivatives by the constrained mock-Chebyshev least squares operator

The goal of this chapter is two-fold. We discuss some theoretical aspects of the constrained mock-Chebyshev least squares operator and present new results. In particular, we introduce explicit representations of the error and its derivatives. Moreover, for a sufficiently smooth function f in $[-1,1]$, we present a method for approximating the successive derivatives of f at a point $x \in [-1,1]$, based on the constrained mock-Chebyshev least squares operator and provide estimates for these approximations.

Let $X_n = \{x_0, \ldots, x_n\}$ be the set of $n+1$ equispaced nodes in $[-1,1]$. The constrained mock-Chebyshev least squares linear operator is defined as follows

$$\begin{aligned} \hat{P}_{r,n} : C([-1,1]) &\to C([-1,1]) \\ f(x) &\mapsto \hat{P}_{r,n}[f](x) = \sum_{i=0}^{r} \hat{a}_i u_i(x), \qquad x \in [-1,1], \end{aligned} \tag{4.1}$$

where, as we have seen in Chapter 2, the vector $\hat{\boldsymbol{a}} = [\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_r]^T$ is the solution of the KKT linear equations

$$\begin{bmatrix} 2V^TV & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} 2V^T\boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix}. \tag{4.2}$$

This operator is well defined since, in Chapter 2, we have proved that the KKT matrix

$$M = \begin{bmatrix} 2V^TV & C^T \\ C & 0 \end{bmatrix} \tag{4.3}$$

is nonsingular. We recall that, the operator $\hat{P}_{r,n}$ satisfies the following properties:

i) $\hat{P}_{r,n}$ is a linear operator, that is

$$\hat{P}_{r,n}[\lambda f + \mu g] = \lambda \hat{P}_{r,n}[f] + \mu \hat{P}_{r,n}[g], \qquad f, g \in C([-1,1]), \qquad \lambda, \mu \in \mathbb{R}; \tag{4.4}$$

ii) the range of $\hat{P}_{r,n}$ is $\mathbb{P}_r(\mathbb{R})$;

iii) $\hat{P}_{r,n}$ reproduces polynomials of degree $\leq r$, that is

$$\hat{P}_{r,n}[P_r] = P_r, \quad \text{for each } P_r \in \mathbb{P}_r(\mathbb{R}); \tag{4.5}$$

iv) $\hat{P}_{r,n}$ is idempotent, that is

$$\hat{P}_{r,n}^2 = \hat{P}_{r,n};$$

v) $\hat{P}_{r,n}[f]$ is completely determined by the evaluations of f on the grid X_n, in particular

$$\hat{P}_{r,n}[f] = \hat{P}_{r,n}\left[P_n[f]\right], \quad \text{for each } f \in C([-1,1]), \tag{4.6}$$

where

$$P_n[f](x) = \sum_{i=0}^{n} f(x_i)\ell_i(x), \qquad x \in [-1,1],$$

and

$$\ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \qquad i = 0,\ldots,n, \qquad x \in [-1,1];$$

vi) $\hat{P}_{r,n}$ interpolates the function f at the mock-Chebyshev subset of nodes $X'_m = \{x'_i : i = 0,\ldots,m\}$, introduced in (1.3), that is

$$\hat{P}_{r,n}[f](x'_i) = f(x'_i), \qquad i = 0,\ldots,m.$$

In this chapter we further study some properties of the constrained mock-Chebyshev least squares operator, by providing new results and applications. More precisely, in Theorem 4.1.1 we give a new bound for the infinity norm of the operator $\hat{P}_{r,n}$. This result allows us to give, in Theorem 4.1.3, estimation for the approximation error

$$\hat{R}_{r,n}[f](x) = f(x) - \hat{P}_{r,n}[f](x), \quad x \in [-1,1], \tag{4.7}$$

in uniform norm, through the error of the best uniform approximation by polynomials of degree less than or equal to r. Further, by assuming $f \in C([-1,1])$, in Theorem 4.1.5 we provide new explicit representations of the pointwise error (4.7) which takes into account the peculiarity of $\hat{P}_{r,n}[f]$ of being a mixed interpolation-regression polynomial. As a consequence, by assuming $f \in C^{r+1}[-1,1]$, in Theorem 4.1.7 we use the Peano Kernel Theorem [23, 67] to obtain explicit representations and pointwise bounds of the successive derivatives of the error (4.7), that is

$$\hat{R}^{(\nu)}_{r,n}[f](x) = f^{(\nu)}(x) - \hat{P}^{(\nu)}_{r,n}[f](x), \quad x \in [-1,1], \quad \nu = 1,\ldots,r, \tag{4.8}$$

and the Markov Theorem [79] to obtain, as a corollary, bounds in uniform norm of the error (4.8). As an application, we introduce a new differentiation method, based on an equispaced grid of points, for approximating the successive derivatives of a sufficiently smooth function f in $[-1,1]$ through the derivatives of the constrained mock-Chebyshev least squares operator. Furthermore, we prove an iterative relationship between the coefficients of $\hat{P}^{(\nu)}_{r,n}$ and those of $\hat{P}^{(\nu-1)}_{r,n}$, $\nu = 1,\ldots,r$, when they are expressed in the Chebyshev polynomial basis of the first kind.

4.1 Constrained mock-Chebyshev least squares linear operator: theoretical aspects

Now, we discuss some theoretical aspects of the representation of the error (4.7), its derivatives and related bounds. To this aim, some preliminary results are needed. First of all, we give a bound for the norm of the operator $\hat{P}_{r,n}$,

$$\left\|\hat{P}_{r,n}\right\| = \sup_{\substack{f\in C([-1,1]) \\ \|f\|_\infty \leq 1}} \left\|\hat{P}_{r,n}[f]\right\|_\infty. \tag{4.9}$$

Theorem 4.1.1. *The norm of the constrained mock-Chebyshev least squares operator satisfies*

$$\left\|\hat{P}_{r,n}\right\| \leq C\left(2(r+1)\kappa(M) + (m+1)\left\|M^{-1}\right\|_1\right), \tag{4.10}$$

where

$$C = \max_{j=0,\ldots,r} \|u_j\|_\infty, \qquad \kappa(M) = \|M\|_1 \|M^{-1}\|_1. \tag{4.11}$$

n	100	500	1000	5000	10000	50000	100000
$\kappa(M)$	7.8e+03	8.4e+04	2.4e+05	2.9e+06	8.3e+06	1.0e+08	2.9e+08
$\left\|M^{-1}\right\|_1$	21.80	52.90	78.20	186.75	275.55	661.36	965.75

Table 4.1: Values of the condition number $\kappa(M)$ and of the norm $\left\|M^{-1}\right\|_1$ in correspondence of some values of n ranging from $n = 100$ to $n = 100000$, by using the Chebyshev polynomial basis of the first kind.

Proof. Let $f \in C([-1,1])$ satisfy $\|f\|_\infty \leq 1$. By using the triangular inequality, we get from (4.1)

$$\left|\hat{P}_{r,n}[f](x)\right| = \left|\sum_{i=0}^{r} \hat{a}_i u_i(x)\right| \leq \sum_{i=0}^{r} |\hat{a}_i| \, |u_i(x)| \,, \quad x \in [-1,1], \tag{4.12}$$

and then, by passing to the supremum with respect to $x \in [-1,1]$ and by the setting (4.11),

$$\left\|\hat{P}_{r,n}[f]\right\|_\infty \leq C \sum_{i=0}^{r} |\hat{a}_i| = C \, \|\hat{\boldsymbol{a}}\|_1 \,. \tag{4.13}$$

Since the KKT matrix (4.3) is nonsingular, by equation (4.2) we have

$$\begin{aligned}
\|\hat{\boldsymbol{a}}\|_1 \leq \left\| \begin{bmatrix} \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{z}} \end{bmatrix} \right\|_1 &\leq \|M^{-1}\|_1 \left\| \begin{bmatrix} 2V^T \boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix} \right\|_1 \\
&= \|M^{-1}\|_1 \left(2 \sum_{j=0}^{r} \left| \sum_{i=0}^{n} u_j(x_i) f(x_i) \right| + \sum_{j=0}^{m} |f(x_j)| \right) \\
&\leq \|M^{-1}\|_1 \left(2 \sum_{j=0}^{r} \sum_{i=0}^{n} |u_j(x_i)| + m + 1 \right) \\
&\leq \|M^{-1}\|_1 \left(2 \|M\|_1 (r+1) + m + 1 \right),
\end{aligned}$$

and therefore

$$\|\hat{\boldsymbol{a}}\|_1 \leq 2(r+1)\kappa(M) + (m+1) \left\|M^{-1}\right\|_1 . \tag{4.14}$$

By using the bound (4.14) in (4.13), we get

$$\left\|\hat{P}_{r,n}[f]\right\|_\infty \leq C \left(2(r+1)\kappa(M) + (m+1) \left\|M^{-1}\right\|_1 \right), \tag{4.15}$$

and then

$$\left\|\hat{P}_{r,n}\right\| \leq C \left(2(r+1)\kappa(M) + (m+1) \left\|M^{-1}\right\|_1 \right)$$

since the right-hand side of (4.15) does not depend on f. □

Remark 4.1.2. *In the case of the Chebyshev polynomial basis of the first kind [88, Ch. 1]*

$$\mathcal{B}_r^{C,1} = \{T_0(x), \ldots, T_r(x)\}, \quad T_k(x) = \cos(k \arccos(x)), \quad x \in [-1,1], \quad k = 0, \ldots, r, \tag{4.16}$$

we have $\|T_k\|_\infty = 1$ *for each* $k = 0, \ldots, r$. *In this case equation* (4.10) *becomes*

$$\left\|\hat{P}_{r,n}\right\| \leq B_n, \tag{4.17}$$

where

$$B_n = 2(r+1)\kappa(M) + (m+1)\left\|M^{-1}\right\|_1. \tag{4.18}$$

In order to appreciate the quality of the bound (4.17), in Figure 4.1 we plot the behavior of B_n in correspondence of some values of n ranging from 100 to 100000. The plot shows a linear relation between the logarithm of n and the logarithm of the bound B_n. We computed the coefficients of this relation through a linear regression and after standard computations, we found

$$B_n \approx e^{3.66} n^{2.03}.$$

Figure 4.1 contains the approximations of the bound B_n computed through the regression line, as well. In Table 4.1 we show the values of $\kappa(M)$, $\left\|M^{-1}\right\|_1$ in correspondence of some values of n ranging from 100 to 100000.

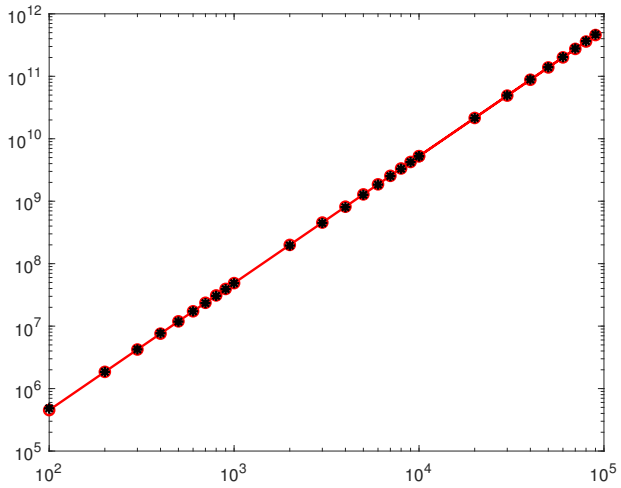


Figure 4.1: Log-log plot of the values B_n (in black stars, $\star$) of the bound (4.17) of the norm $\left\|\hat{P}_{r,n}\right\|$ relative to sets of $n+1$ equispaced nodes, with n ranging from 10^2 to 10^5. The linear relation between the values on the x-axis and those on the y-axis is evident and confirmed by the closeness of their approximations (in red circles, o) computed through a linear regression.

Since the constrained mock-Chebyshev least squares operator is a projection on the polynomial space $\mathbb{P}_r(\mathbb{R})$ [17, Ch. 6], it is possible to give a standard estimation for the uniform norm of the approximation error

$$\left\|\hat{R}_{r,n}[f]\right\|_\infty = \left\|f - \hat{P}_{r,n}[f]\right\|_\infty, \quad f \in C([-1,1]), \tag{4.19}$$

through the error of best uniform approximation by polynomials of $\mathbb{P}_r(\mathbb{R})$

$$E_r^\star(f) = \left\|f - P_r^\star[f]\right\|_\infty, \tag{4.20}$$

also known as minmax error. With this aim, for computational convenience and to short the notation, we suppose to work with the Chebyshev polynomial basis of the first kind $\mathcal{B}_r^{C,1}$.

Theorem 4.1.3. *Let $f \in C([-1,1])$, then*

$$\left\|\hat{R}_{r,n}[f]\right\|_\infty \leq (1 + B_n)\, E_r^\star(f). \tag{4.21}$$

Proof. Let $P_r^\star[f] \in \mathbb{P}_r(\mathbb{R})$ be the polynomial of best uniform approximation to f. By properties ***i)***-***iii)*** we easily find

$$\begin{aligned}\left\|f-\hat{P}_{r,n}[f]\right\|_\infty &= \left\|f-P_r^\star[f]+P_r^\star[f]-\hat{P}_{r,n}[f]\right\|_\infty \\ &= \left\|f-P_r^\star[f]-\hat{P}_{r,n}[f-P_r^\star[f]]\right\|_\infty \\ &\leq \left(1+\left\|\hat{P}_{r,n}\right\|\right)\|f-P_r^\star[f]\|_\infty .\end{aligned}$$

The thesis then follows from Theorem 4.1.1 after setting (4.18). □

The following result is a direct consequence of Theorem 4.1.3 and Jackson Theorem [17, Ch. 4].

Corollary 4.1.4. *Let $f \in C^k([-1,1])$, $k=0,\dots,r$. Then we have*

$$\left\|\hat{R}_{r,n}[f]\right\|_\infty \leq (1+B_n)\,\omega_f\left(\frac{\pi}{r+1}\right), \quad k=0, \tag{4.22}$$

$$\left\|\hat{R}_{r,n}[f]\right\|_\infty \leq \left(\frac{\pi}{2}\right)^k (1+B_n)\,\frac{\left\|f^{(k)}\right\|_\infty}{(r+1)r\cdots(r-k+2)}, \quad 0<k\leq r, \tag{4.23}$$

where $\omega_f(\cdot)$ is the modulus of continuity of the function f [17].

Let $X_p''' = \{x_i''' : i=0,\dots,p\}$ be the set introduced in (1.11). In order to give a pointwise representation of the error (4.7) for all $x \in [-1,1]$, we denote by $P_r[f]$ the Lagrange interpolation polynomial of the function f at the node set $X_m' \cup X_p'''$, that is

$$P_r[f](x) = \sum_{i=0}^{m} \ell_{i,m}(x) f(x_i') + \sum_{j=0}^{p} \ell_{j,p}(x) f(x_j'''), \quad x \in [-1,1], \tag{4.24}$$

where

$$\ell_{i,m}(x) = \prod_{\substack{k=0\\k\neq i}}^{m} \frac{x-x_k'}{x_i'-x_k'} \prod_{s=0}^{p} \frac{x-x_s'''}{x_i'-x_s'''}, \qquad \ell_{j,p}(x) = \prod_{k=0}^{m} \frac{x-x_k'}{x_j'''-x_k'} \prod_{\substack{s=0\\s\neq j}}^{p} \frac{x-x_s'''}{x_j'''-x_s'''},$$

and by

$$R_r[f](x) = f(x) - P_r[f](x), \quad x \in [-1,1], \tag{4.25}$$

the error of Lagrange interpolation.

Theorem 4.1.5. *Let $f \in C([-1,1])$. Then for the approximation error $\hat{R}_{r,n}[f](x)$, introduced in (4.7), we have*

$$\hat{R}_{r,n}[f](x) = R_r[f](x) + \sum_{j=0}^{p} \ell_{j,p}(x)\hat{R}_{r,n}[f](x_j'''), \quad x \in [-1,1]. \tag{4.26}$$

Proof. By Property ***ii)*** $\hat{P}_{r,n}[f] \in \mathbb{P}_r(\mathbb{R})$ and by Property ***vi)*** $\hat{P}_{r,n}[f](x_i') = f(x_i')$, $i=0,\dots,m$. Then by the uniqueness of the Lagrange interpolation polynomial, we get

$$\begin{aligned}\hat{P}_{r,n}[f](x) = P_r[\hat{P}_{r,n}[f]](x) &= \sum_{i=0}^{m} \ell_{i,m}(x)\hat{P}_{r,n}[f](x_i') + \sum_{j=0}^{p} \ell_{j,p}(x)\hat{P}_{r,n}[f](x_j''') \\ &= \sum_{i=0}^{m} \ell_{i,m}(x) f(x_i') + \sum_{j=0}^{p} \ell_{j,p}(x)\hat{P}_{r,n}[f](x_j''') \\ &= \sum_{i=0}^{m} \ell_{i,m}(x) f(x_i') + \sum_{j=0}^{p} \ell_{j,p}(x) f(x_j''') + \sum_{j=0}^{p} \ell_{j,p}(x)\left(\hat{P}_{r,n}[f](x_j''') - f(x_j''')\right) \\ &= P_r[f](x) + \sum_{j=0}^{p} \ell_{j,p}(x)\left(\hat{P}_{r,n}[f](x_j''') - f(x_j''')\right).\end{aligned}$$

Therefore

$$\begin{aligned} f(x) - \hat{P}_{r,n}[f](x) &= f(x) - P_r[f](x) + \sum_{j=0}^{p} \ell_{j,p}(x) \left(f(x_j''') - \hat{P}_{r,n}[f](x_j''') \right) \\ &= R_r[f](x) + \sum_{j=0}^{p} \ell_{j,p}(x) \hat{R}_{r,n}[f](x_j'''). \end{aligned}$$

□

Corollary 4.1.6. *For any $x \in [-1,1]$, we have*

$$\left| \left| \hat{R}_{r,n}[f](x) \right| - \left| \sum_{j=0}^{p} \ell_{j,p}(x) \hat{R}_{r,n}[f](x_j''') \right| \right| \leq (1 + \Lambda_r)\, E_r^\star(f), \tag{4.27}$$

where

$$\Lambda_r = \max_{x \in [-1,1]} \left(\sum_{i=0}^{m} |\ell_{i,m}(x)| + \sum_{j=0}^{p} |\ell_{j,p}(x)| \right) \tag{4.28}$$

is the Lebesgue constant of the polynomial interpolation operator P_r and $E_r^\star(f)$ is defined as in (4.20).

Proof. The result follows by bounding the error of interpolation $R_r[f]$ in standard way by using the Lebesgue constant and the minmax error, see [48, Ch. 2] and the reverse triangle inequality of the absolute value. □

Inequality (4.27) of Corollary 4.1.6 suggests that we can estimate $\left| \hat{R}_{r,n}[f](x) \right|$ with $\left| \sum_{j=0}^{p} \ell_{j,p}(x) \hat{R}_{r,n}[f](x_j''') \right|$ with an accuracy that depends on Λ_r and $E_r^\star(f)$. In Figure 4.2 we represent the behavior of the Lebesgue constants Λ_r computed for the values of $n = 100 : 2 : 6000$ and the graph of the function $n \to \gamma + \delta\sqrt{n}$, with $\gamma = 5$ and $\delta = \frac{1}{2} - 0.025$ which seems to follow the growth of Λ_r on average, in this interval. Since

$$\left| \sum_{j=0}^{p} \ell_{j,p}(x) \hat{R}_{r,n}[f](x_j''') \right| \leq \Lambda_r \max_{j=0,\ldots,p} \left| \hat{R}_{r,n}[f](x_j''') \right|, \quad x \in [-1,1], \tag{4.29}$$

we can use the right member of inequality (4.29) as an estimator of a bound of $\left| \hat{R}_{r,n}[f](x) \right|$, $x \in [-1,1]$, which can be easily computed by the available data. The numerical examples provided in Section 4.3 shows the goodness of the estimator (4.29).

Let $f \in C^{r+1}([-1,1])$. In this case, the Peano kernel Theorem [23, 67] allows us to represent the remainder $R_r[f]$ of Lagrange interpolation on the node set $X_m' \cup X_p'''$ in integral form

$$R_r[f](x) = \int_{-1}^{1} K(x,t) \frac{f^{(r+1)}(t)}{r!} dt, \quad x \in [-1,1], \tag{4.30}$$

where

$$K(x,t) = (x-t)_+^r - P_r[(x-t)_+^r], \quad x,t \in [-1,1],$$

and

$$(x-t)_+^r = \begin{cases} (x-t)^r, & \text{if } x - t \geq 0, \\ 0, & \text{if } x - t < 0. \end{cases}$$

We can differentiate both members of (4.30) ν times, $\nu = 1, \ldots, r$, with respect to x, in order to obtain pointwise representations of the successive derivatives of the error of Lagrange interpolation

$$R_r^{(\nu)}[f](x) = \int_{-1}^{1} \frac{\partial^\nu K(x,t)}{\partial x^\nu} \frac{f^{(r+1)}(t)}{r!} dt. \tag{4.31}$$

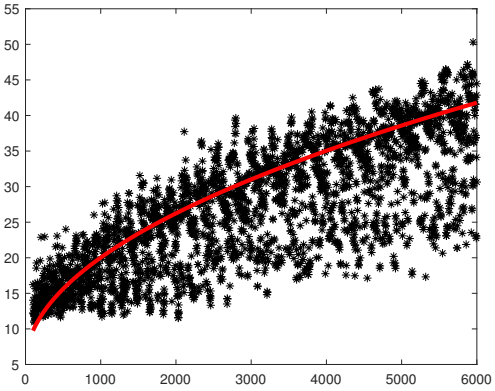


Figure 4.2: Plot of the values of Λ_r computed in correspondence of the values of $n = 100 : 2 : 6000$ (in black stars, $\star$), together with the graph of the functions $n \to \gamma + \delta\sqrt{n}$, with $\gamma = 5$ and $\delta = \frac{1}{2} - 0.025$.

By using these representations, pointwise and uniform bounds for $R_r^{(\nu)}[f](x)$ are obtained [67]. In particular, in [67] it is proven that

$$\left|R_r^{(\nu)}[f](x)\right| \leq \int_{-1}^{1} \left|\frac{\partial^\nu K(x,t)}{\partial x^\nu}\right| dt \left\|\frac{f^{(r+1)}}{r!}\right\|_\infty, \quad x \in [-1,1] \tag{4.32}$$

and also that

$$\left\|R_r^{(\nu)}\right\|_\infty \leq \left\|\omega_r^{(\nu)}\right\|_\infty \frac{\left\|f^{(r+1)}\right\|_\infty}{\nu!(r+1-\nu)!}, \quad \nu = 1,\ldots,r, \tag{4.33}$$

where

$$\omega_r(x) = \prod_{k=0}^{m}(x - x'_k)\prod_{s=0}^{p}(x - x'''_s), \quad x \in [-1,1]$$

is the nodal polynomial associated with the grid $X'_m \cup X'''_p$. Then, for the derivative of the remainder of the constrained mock-Chebyshev least squares approximation, the following bounds hold.

Theorem 4.1.7. *Let $f \in C^{r+1}([-1,1])$ and $\nu = 0,\ldots,r$, then*

$$\left|\hat{R}_{r,n}^{(\nu)}[f](x)\right| \leq \int_{-1}^{1} \left|\frac{\partial^\nu K(x,t)}{\partial x^\nu}\right| dt \left\|\frac{f^{(r+1)}}{r!}\right\|_\infty + \sum_{j=0}^{p} \left|\ell_{j,p}^{(\nu)}(x)\right| \left|\hat{P}_{r,n}[f](x_j''') - f(x_j''')\right|, \quad x \in [-1,1]. \tag{4.34}$$

Proof. The proof follows by differentiating both members of (4.26) in Theorem 4.1.5 and using the bound (4.32). □

Theorem 4.1.8. *Let $f \in C^{r+1}([-1,1])$ and $\nu = 1,\ldots,r$, then*

$$\begin{aligned}
\left\|\hat{R}_{r,n}^{(\nu)}[f]\right\|_\infty &= \left\|f^{(\nu)} - \hat{P}_{r,n}^{(\nu)}[f]\right\|_\infty \\
&\leq \frac{\prod_{j=0}^{\nu-1}\left((r+1)^2 - j^2\right)}{\prod_{j=0}^{\nu-1}(2j+1)} \|\omega_r\|_\infty \frac{\left\|f^{(r+1)}\right\|_\infty}{\nu!(r+1-\nu)!} \\
&\quad + \frac{\prod_{j=0}^{\nu-1}\left(r^2 - j^2\right)}{\prod_{j=0}^{\nu-1}(2j+1)} \sum_{k=0}^{p} \|\ell_{k,p}\|_\infty \left|\hat{P}_{r,n}[f](x_k''') - f(x_k''')\right|.
\end{aligned} \tag{4.35}$$

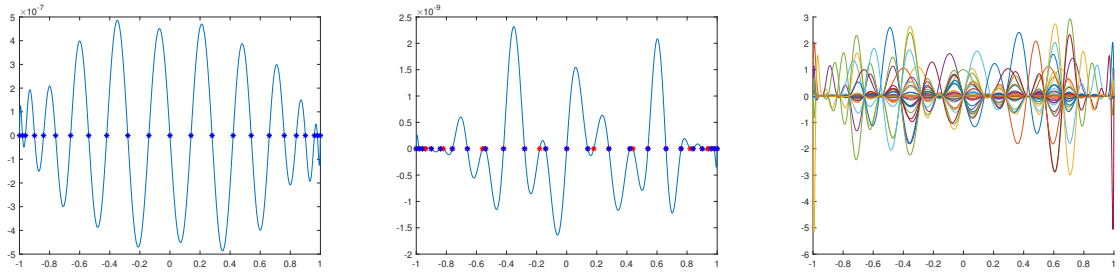


Figure 4.3: The nodal polynomial at the mock-Chebyshev nodes X'_m (left), the nodal polynomial at the node set $X'_m \cup X'''_p$(center) and the Lagrange fundamental polynomials at the node set $X'_m \cup X'''_p$(right) for $n = 100$ (and then $m = 22$, $p = 9$, $r = 32$).

Proof. For each $\nu = 1, \dots, r$, by Theorem 4.1.5, we have

$$\hat{R}^{(\nu)}_{r,n}[f](x) = R^{(\nu)}_r[f](x) + \sum_{j=0}^{p} \ell^{(\nu)}_{j,p}(x) \left(\hat{P}_{r,n}[f](x'''_j) - f(x'''_j) \right),$$

then, by using the triangular inequality

$$\left\| \hat{R}^{(\nu)}_{r,n}[f] \right\|_\infty \leq \left\| R^{(\nu)}_r[f] \right\|_\infty + \sum_{j=0}^{p} \left\| \ell^{(\nu)}_{j,p} \right\|_\infty \left| \hat{R}_{r,n}[f](x'''_j) \right|. \tag{4.36}$$

We use inequality (4.33) to bound $\left\| \hat{R}^{(\nu)}_{r,n}[f] \right\|_\infty$ and the Markov's inequality [79] to bound both $\left\| \omega^{(\nu)}_r \right\|_\infty$ and $\left\| \ell^{(\nu)}_{j,p} \right\|_\infty$. The thesis follows. □

Remark 4.1.9. *As the classical Cauchy's bound for the error of Lagrange interpolation, the bound* (4.34) *suffers from the presence of the infinity norm of the* $(r+1)$*-th derivative of the function* f*, which can increase very rapidly as* r *increases, even if* $P_r[f](x)$ *converges to* $f(x)$ *for any* $x \in [-1, 1]$ *(see, e.g., the case of the Runge function in Figure 4.10, which satisfies* $\left\| f^{(r+1)} \right\|_\infty \geq 5^{r+1}(r+1)!$*). When the magnitude of the derivatives of* f *grows moderately, the bound* (4.34) *is reliable, as shown by the examples provided in Table 4.3 and Table 4.4 for the functions* $\sin(50x)$ *and* e^{5x}*, respectively. In that examples the right-hand side member of inequality* (4.34) *is denoted by* $\widetilde{B}^{(\nu)}_{r,n}[f]$*.*

Remark 4.1.10. *In Figures 4.3, 4.4 and 4.5 the nodal polynomial* ω_m *at the mock-Chebyshev nodes* X'_m*, the nodal polynomial* ω_r *at the node set* $X'_m \cup X'''_p$ *and the Lagrange fundamental polynomials at the node set* $X'_m \cup X'''_p$ *are graphically represented for* $n = 100, 1000, 10000$*. It is worth noting that* $\|\omega_r\|_\infty \approx \dfrac{1}{2^r} = \|\mathring{T}_{r+1}\|_\infty$*, where* $\mathring{T}_{r+1}(x) = \dfrac{1}{2^r} T_{r+1}(x)$ *is the monic Chebyshev polynomial of degree* $r+1$ *[48, Ch. 2].*

4.2 Numerical differentiation through constrained mock-Chebyshev least squares operator

We introduce a numerical differentiation formula based on the constrained mock-Chebyshev least squares operator. Let $f \in C^1([-1, 1])$ be a differentiable function whose first derivative is continuous on the interval $(-1, 1)$. It is worth emphasizing that we suppose to know exclusively f on the set X_n of $n+1$ equispaced nodes. We apply the constrained mock-Chebyshev least squares operator to f and compute the polynomial

$$\hat{P}_{r,n}[f](x) = \sum_{i=0}^{r} \hat{a}_i u_i(x), \qquad x \in [-1, 1]. \tag{4.37}$$

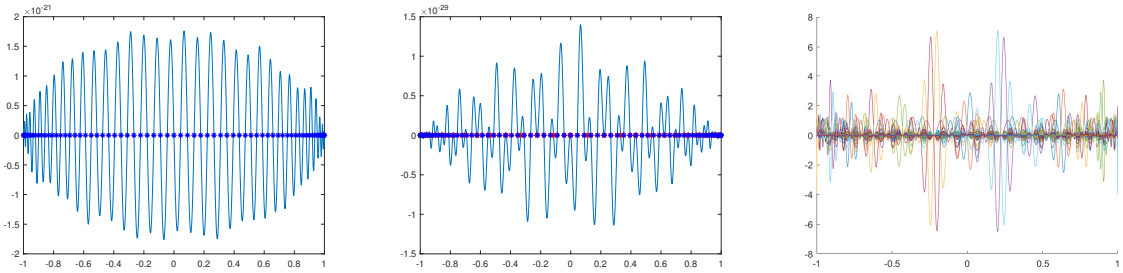


Figure 4.4: The nodal polynomial at the mock-Chebyshev nodes X'_m (left), the nodal polynomial at the node set $X'_m \cup X'''_p$(center) and the Lagrange fundamental polynomials at the node set $X'_m \cup X'''_p$(right) for $n = 1000$, (and then $m = 70$, $p = 28$, $r = 99$).

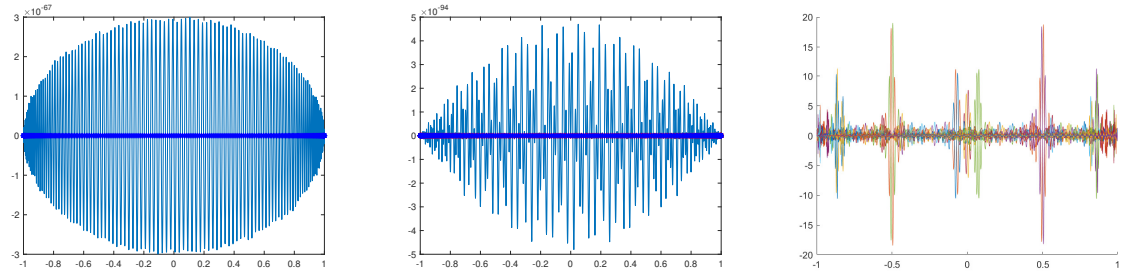


Figure 4.5: The nodal polynomial at the mock-Chebyshev nodes X'_m (left), the nodal polynomial at the node set $X'_m \cup X'''_p$(center) and the Lagrange fundamental polynomials at the node set $X'_m \cup X'''_p$(right) for $n = 10000$, (and then $m = 222$, $p = 90$, $r = 313$).

By differentiating (4.37), we obtain

$$\hat{P}'_{r,n}[f](x) = \sum_{i=0}^{r} \hat{a}_i u'_i(x), \qquad x \in [-1,1], \tag{4.38}$$

and, since $\hat{P}'_{r,n}[f] \in \mathbb{P}_{r-1}(\mathbb{R})$, from (4.5), we get

$$\hat{P}'_{r,n}[f] = \hat{P}_{r,n}\left[\hat{P}'_{r,n}[f]\right] = \sum_{i=0}^{r} \hat{a}'_i u_i(x). \tag{4.39}$$

Remark 4.2.1. *We notice that the coefficients vector $\hat{a}' = [\hat{a}'_0, \dots, \hat{a}'_r]^T$ is uniquely determined, since the KKT matrix* (4.3) *has nonzero determinant, see [39].*

Relations between the vectors of coefficients $\hat{a} = [\hat{a}_0, \dots, \hat{a}_r]^T$ and $\hat{a}' = [\hat{a}'_0, \dots, \hat{a}'_r]^T$ depend on the chosen polynomial basis $\{u_1(x), \dots, u_r(x)\}$. Since we assume to work with the Chebyshev polynomial basis of the first kind $\mathcal{B}_r^{C,1}$, in the following we make this relation explicit in this particular case. To this aim we recall some useful identities between the Chebyshev polynomials of the first kind $T_k(x)$, $k = 0, \dots, r$, and the Chebyshev polynomial of the second kind $U_k(x)$, $k = 0, \dots, r$. These polynomials are defined by [87]

$$U_k(x) = \frac{\sin\left((k+1)\arccos(x)\right)}{\sin\left(\arccos(x)\right)}, \qquad x \in [-1,1], \qquad k = 0, \dots, r,$$

and satisfy the following relations

$$U_0(x) = T_0(x) = 1, \quad U_1(x) = 2T_1(x), \quad U_k(x) - U_{k-2}(x) = 2T_k(x), \qquad k \geq 2. \tag{4.40}$$

Moreover

$$T'_k(x) = kU_{k-1}(x), \quad U'_k(x) = \frac{(k+1)T_{k+1}(x) - xU_k(x)}{x^2 - 1}. \tag{4.41}$$

Theorem 4.2.2. *Let $f \in C([-1,1])$, by expressing the polynomial $\hat{P}_{r,n}[f]$ and its first derivative $\hat{P}'_{r,n}[f]$ in the basis $\mathcal{B}_r^{C,1}$ as in equations* (4.37) *and* (4.39)*, respectively, we have*

$$\hat{a}'_0 = \sum_{j=0}^{\lfloor \frac{r}{2} \rfloor} (2j+1)\hat{a}_{2j+1}, \qquad \hat{a}'_i = 2 \sum_{j=0}^{\lfloor \frac{r-i}{2} \rfloor} (i+2j+1)\hat{a}_{i+2j+1}, \qquad i = 1, \dots, r. \tag{4.42}$$

Proof. By (4.37), the polynomial $\hat{P}_{r,n}[f]$ in the basis $\mathcal{B}_r^{C,1}$ has the form

$$\hat{P}_{r,n}[f](x) = \sum_{j=0}^{r} \hat{a}_j T_j(x), \qquad x \in [-1,1].$$

By using the identities (4.38) and (4.41), we get

$$\hat{P}'_{r,n}[f](x) = \sum_{j=0}^{r} \hat{a}_j T'_j(x) = \sum_{j=1}^{r} j\hat{a}_j U_{j-1}(x), \qquad x \in [-1,1]. \tag{4.43}$$

By setting $\hat{a}_{r+1} = 0$, after a change the dummy index, the polynomial $\hat{P}'_{r,n}[f]$ can be written as

$$\hat{P}'_{r,n}[f](x) = \sum_{j=0}^{r} (j+1)\hat{a}_{j+1} U_j(x), \qquad x \in [-1,1]. \tag{4.44}$$

The result follows from the identity (4.40). □

From the above results, one can deduce that there are two different strategies in order to compute the analytic expression (4.39) of the polynomial $\hat{P}'_{r,n}[f]$ in the Chebyshev polynomial basis of the first kind.

S_1) Evaluate $U_k(x)$, $k = 0, \ldots, r$, on the set X_n. Compute the values of $\hat{P}'_{r,n}[f]$ on the equispaced nodes using (4.38). Solve the KKT linear equations (4.2) in order to compute $\hat{P}_{r,n}\left[\hat{P}'_{r,n}[f]\right]$ from the values of $\hat{P}'_{r,n}[f]$ on the equispaced nodes.

S_2) Use the equation (4.42) in order to compute the analytic expression of $\hat{P}'_{r,n}[f]$.

We notice that, although the strategy S_2 is more direct with respect to the strategy S_1, it can be applied only in the case of the Chebyshev polynomial basis of the first kind. In the next Section, we will show that the two strategies are equivalent in terms of accuracy of results.

We emphasize, that formula (4.39) provides a global polynomial approximation of the first derivative of the function f. Clearly, it is possible to repeat both procedures to approximate the derivative of order k, for $k \geq 1$, by supposing that $f \in C^k([-1,1])$. In this regard, the following Theorem holds.

Theorem 4.2.3. *Let $f \in C([-1,1])$. We express the polynomial $\hat{P}_{r,n}[f]$ and its successive derivatives in the basis $\mathcal{B}_r^{C,1}$, that is*

$$\hat{P}^{(\nu)}_{r,n}[f](x) = \sum_{i=0}^{r} \hat{a}_i^{(\nu)} T_i(x), \qquad x \in [-1,1], \qquad \nu = 1, \ldots, r.$$

For each $\nu \geq 1$, we get

$$\hat{a}_0^{(\nu)} = \sum_{j=0}^{\lfloor \frac{r}{2} \rfloor} (2j+1)\hat{a}_{2j+1}^{(\nu-1)}, \qquad \hat{a}_i^{(\nu)} = 2 \sum_{j=0}^{\lfloor \frac{r-i}{2} \rfloor} (i+2j+1)\hat{a}_{i+2j+1}^{(\nu-1)}, \qquad i = 1, \ldots, r. \tag{4.45}$$

Proof. The proof follows the same argument of Theorem 4.2.2. It is therefore omitted here. □

4.3 Numerical experiments

In this Section, we numerically prove the accuracy of the proposed approximation methods by several examples. The numerical experiments are performed using `MatLab` software. In particular, the command `diff` is used in order to compute the exact successive derivatives of all considered functions and the Chebfun package is used in order to compute the Chebyshev polynomial basis of the first kind [65].

We perform three different types of numerical tests. In the first test, we consider the function

$$f_1(x) = xe^{-2x} + \sin(3x)$$

used in [74] in order to test general explicit finite difference formulas with arbitrary order accuracy for approximating first and higher derivatives. These formulas are applicable to unequally or equally spaced data. In line with the experiments presented in [74], we consider a set of $n+1 = 67$ equispaced points in the interval $[-1,1]$, in order to have a stepsize $h = 0.03$. We compute the errors

$$e_{mean} = \frac{1}{N}\sum_{i=1}^{N} e_i, \qquad e_{max} = \max_{i=1,\ldots,N} e_i, \tag{4.46}$$

obtained in approximating the first four order derivatives of the function f_1 by the constrained mock-Chebyshev least squares operator on the uniform grid of 67 points in $[-1,1]$, computed by following the strategy S_2. In equation (4.46) e_i is the absolute value of the approximation error at the i-th point of this grid. The numerical results are reported in Table 4.2. The approximation accuracies are comparable or even better with respect to those one reported in [74] for the case of the finite difference formula at 11 equally spaced points with stepsize $h = 0.03$ in the interval $[0, 0.3]$. To better appreciate the behavior of the approximation errors in the whole interval $[-1,1]$, in Figure 4.6 we plot the absolute values of the

order	0	1	2	3	4
e_{mean}	1.24e-15	7.59-14	9.02-12	9.92e-10	8.57e-08
e_{max}	1.77e-14	4.43e-12	7.46e-10	7.67e-08	5.78e-06

Table 4.2: Mean and max approximation errors obtained in approximating the function $f_1(x) = xe^{-2x} + \sin(3x)$ and its first four order derivatives on the grid of 67 equispaced points in $[-1, 1]$, by using the constrained mock-Chebyshev least squares operator with $n + 1 = 67$.

	$\hat{R}_{r,n}^{(\nu)}[f_6]$	$\widetilde{B}_{r,n}^{(\nu)}[f_6]$
$\nu = 0$	1.95e-14	6.99e-13
$\nu = 1$	2.55e-11	6.86e-09
$\nu = 2$	6.72e-08	2.38e-05
$\nu = 3$	1.14e-04	1.35e-01

Table 4.3: Comparison between the maximum approximation error produced by the constrained mock-Chebyshev least squares operator computed at a uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_6(x) = \sin(50x)$ and its first four derivatives, and its relative bounds $\widetilde{B}_{r,n}^{(\nu)}[f_6]$.

pointwise errors computed on the equispaced grid of $N = 201$ points for the first four order derivatives of the function f_1. The plots are displayed in a lexicographic order, by increasing the order of derivatives. The red dash-dotted line is the approximation error related to the application of the strategy S_1 while the black dashed line is the approximation error related to the application of the strategy S_2. From the Figure, it is evident that the application of the two strategies S_1 and S_2 gives practically the same results.

In the second type of test, we consider the following functions [74, 72]

$$f_1(x) = xe^{-2x} + \sin(3x), \quad f_2(x) = e^{-50(x-0.4)^2} + \sinh(x),$$
$$f_3(x) = \frac{1}{1+8x^2}, \quad f_4(x) = \frac{1}{1+25x^2}, \quad f_5(x) = \frac{\sin(8(x+1))}{(x+1.1)^{3/2}},$$

and we analyze the trend of the mean approximation errors and the maximum approximation errors (4.46) obtained in approximating the first four order derivatives of f_i, $i = 1, \ldots, 5$, by using the constrained mock-Chebyshev least squares operator on uniform grids of different stepsize. In particular, we consider sets of $n + 1$ equispaced nodes with $n = 50k$, $k = 1, \ldots, 80$ and compute the errors on the uniform grid of $N = 10104$ equispaced points of the interval $[-1, 1]$.

The results of the tests are shown in Figures 4.7 - 4.11. All these examples show, with clear evidence, that once the maximum precision is reached for $\hat{P}_{r,n}[f]$, the increase in the number of nodes does not lead to more accurate approximation for the derivatives, on the contrary, the increase of the condition number of the matrices involved in the strategies S_1 and S_2 causes worsening of results.

Finally, we consider the functions

$$f_6(x) = \sin(50x), \quad f_7(x) = e^{5x}$$

and in Tables 4.3 and 4.4 we compare the maximum approximation error $\max_{x \in X_N} \left| \hat{R}_{r,n}^{(\nu)}[\cdot](x) \right|$ with $n = 1000$, $\nu = 0, 1, 2, 3, 4$, computed at the grid X_N of $N = 10104$ equispaced points of the interval $[-1, 1]$, with the bound $\widetilde{B}_{r,n}^{(\nu)}[f]$ introduced in (4.34). When the magnitude of the derivatives of f grows moderately, the bound (4.34) is reliable, as shown by the numerical results.

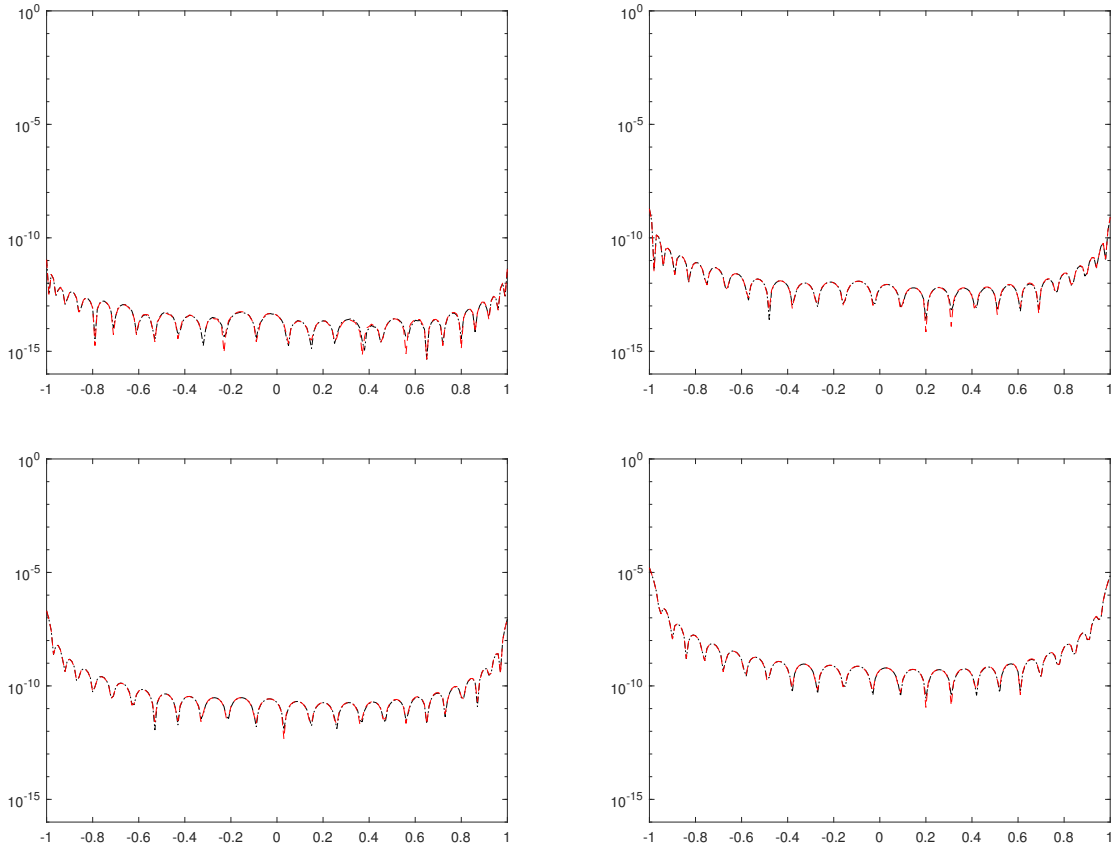


Figure 4.6: Behavior of the approximation errors, in absolute value, of the first four order derivatives of the function $f_1(x) = xe^{-2x} + \sin(3x)$ in the whole interval $[-1, 1]$, by using the constrained mock-Chebyshev least squares operator with $n + 1 = 67$. The absolute values of the pointwise errors are computed on the equispaced grid of $N = 201$ points. The plots are displayed in a lexicographic order, by increasing the order of derivatives. From the plots, it is evident that the application of the two strategies S_1 (red dash-dotted line) and S_2 (black dashed line) gives practically the same results.

	$\hat{R}_{r,n}^{(\nu)}[f_7]$	$\widetilde{B}_{r,n}^{(\nu)}[f_7]$
$\nu = 0$	2.27e-13	3.88e-12
$\nu = 1$	3.94e-10	3.80e-08
$\nu = 2$	9.97e-07	1.24e-04
$\nu = 3$	1.46e-03	2.44e-01

Table 4.4: Comparison between the maximum approximation error produced by the constrained mock-Chebyshev least squares operator computed at a uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_7(x) = e^{5x}$ and its first four derivatives, and its relative bounds $\widetilde{B}_{r,n}^{(\nu)}[f_7]$.

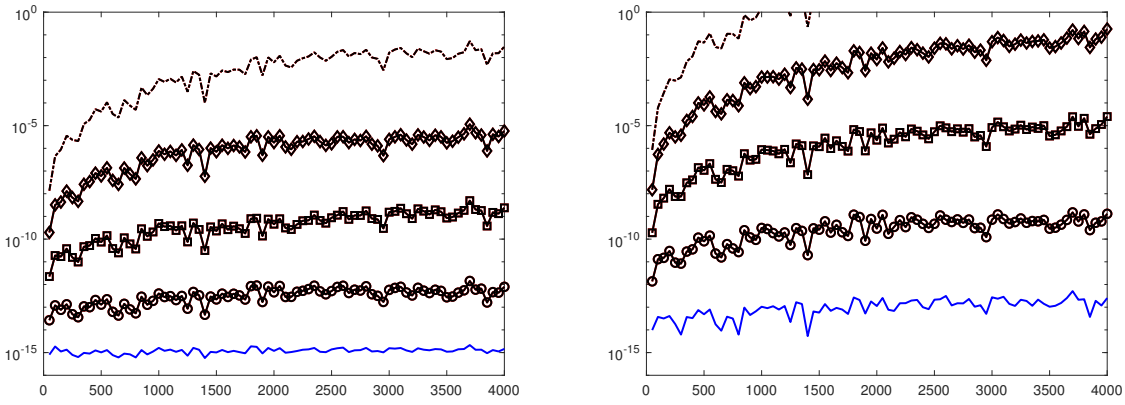


Figure 4.7: Mean approximation error (left) and Maximum approximation error (right) at the uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_1(x) = xe^{-2x} + \sin(3x)$ (in blue) and its first four derivatives, with the increasing order from the bottom to the top, when approximated by using the constrained mock-Chebyshev least squares operator with $n = 50k$, $k = 1, \ldots, 80$. From the plots, it is evident that the application of the two strategies S_1 (red) and S_2 (black) gives practically the same results.

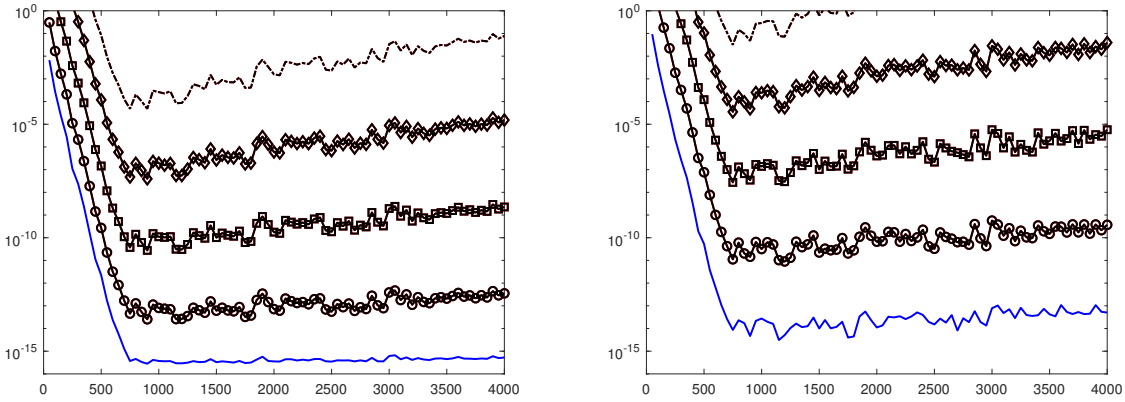


Figure 4.8: Mean approximation error (left) and Maximum approximation error (right) at the uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_2(x) = e^{-50(x-0.4)^2} + \sinh(x)$ (in blue) and its first four derivatives, with the increasing order from the bottom to the top, when approximated by using the constrained mock-Chebyshev least squares operator with $n = 50k$, $k = 1, \ldots, 80$. From the plots, it is evident that the application of the two strategies S_1 (red) and S_2 (black) gives practically the same results.

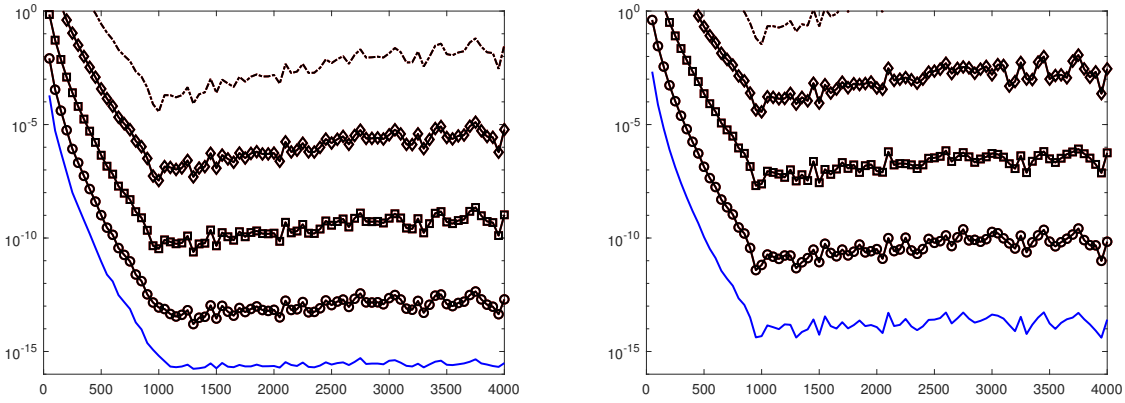


Figure 4.9: Mean approximation error (left) and Maximum approximation error (right) at the uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_3(x) = \frac{1}{1+8x^2}$ (in blue) and its first four derivatives, with the increasing order from the bottom to the top, when approximated by using the constrained mock-Chebyshev least squares operator with $n = 50k$, $k = 1, \ldots, 80$. From the plots, it is evident that the application of the two strategies S_1 (red) and S_2 (black) gives practically the same results.

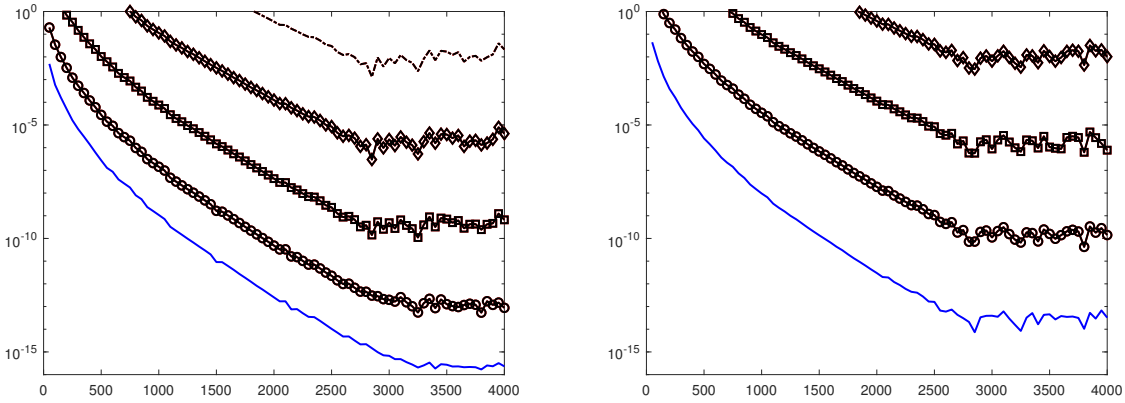


Figure 4.10: Mean approximation error (left) and Maximum approximation error (right) at the uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_4(x) = \frac{1}{1+25x^2}$ (in blue) and its first four derivatives, with the increasing order from the bottom to the top, when approximated by using the constrained mock-Chebyshev least squares operator with $n = 50k$, $k = 1, \ldots, 80$. From the plots, it is evident that the application of the two strategies S_1 (red) and S_2 (black) gives practically the same results.

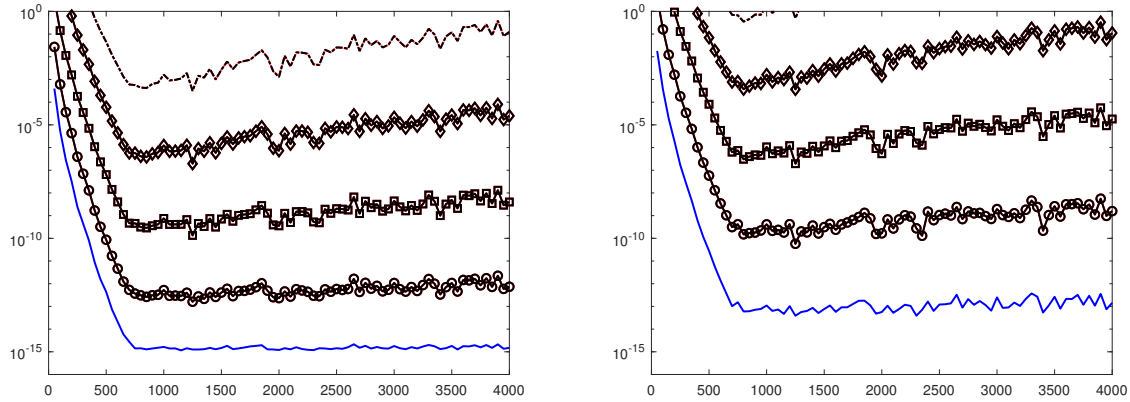


Figure 4.11: Mean approximation error (left) and Maximum approximation error (right) at the uniform grid of $N = 10104$ equispaced points in the interval $[-1, 1]$ relative to the function $f_5(x) = \frac{\sin(8(x+1))}{(x+1.1)^{3/2}}$ (in blue) and its first four derivatives, with the increasing order from the bottom to the top, when approximated by using the constrained mock-Chebyshev least squares operator with $n = 50k$, $k = 1, \ldots, 80$. From the plots, it is evident that the application of the two strategies S_1 (red) and S_2 (black) gives practically the same results.

Chapter 5

On the improvement of the triangular Shepard method by non conformal polynomial elements

Most of classical numerical methods for approximating a multivariate function (or integrals of it) use function values at sample points. However, as shown in [6, 61, 62, 46, 64, 63], in many practical applications, the available data are not restricted to function or derivative evaluations, but also contain several integrals over certain hyperplane sections or, more generally, over simple smooth surfaces in $\mathbb{R}^d$. In such cases, generalizations of the existing theory and algorithms of approximation are required, which are based on the *enriched* set of data. The motivations for discussing such a fundamental issue arise in a variety of cases, since, in several applications, the data obtained in measurements contain the mean values of a function over some line or surface. This type of data is inherent to computer tomography and is widely used in geology, radiology, medicine etc. In this chapter, we focus on the problem of reconstruction of a function from functional data and line integrals, in the setting of two-dimensional scattered data. Scattered data approximation deals with the problem of reconstructing an unknown function from data based on points which have no structure or order between their relative locations. The most famous operator for scattered data interpolation is the Shepard's one, introduced by D. Shepard in 1968 and based on a weighted average of values at the data points [90]. Several variations of the original Shepard's operator have been proposed with the aim of increasing its accuracy of approximation, to improve its efficiency or even to solve specific interpolation problems: the studies carried out in this direction are well known (see, e.g. [50, 44, 101, 22, 33] or the recent survey [27]). Here, in particular, we introduce a modification of the Shepard operator based on a new enrichment of the standard triangular linear finite element, which uses line integrals and quadratic polynomials. Dealing with triangular elements, a good strategy to improve the Shepard method is to follow the Little idea [75] of considering a triangulation of the data location and blending standard triangular linear finite elements with Shepard's like basis functions based on triangles. The introduction of the notion of compact triangulations in [36] and the use of a searching technique to detect and select the nearest neighbor points in the interpolation scheme [16], has allowed the *triangular Shepard method* [75] to be considered a fast method with quadratic approximation order and good accuracy of approximation. In this chapter, we improve the triangular Shepard method by using the proposed enrichment of the standard triangular linear finite element, in line with previous papers [31, 37].

5.1 Polynomial enrichment of the standard triangular linear finite element

The finite element method is commonly used to solve partial differential equations [19]. This method provides a discrete solution on the finite element space, usually formed by piecewise polynomials, to

approximate the exact solution of the considered differential problem. Recall that the finite element is said to be conforming if the space of finite elements is a subspace of the space of solutions of the problem. More precisely, for an elliptic boundary value problem of order $2r$, the conforming finite element space is a subspace of C^{r-1}. It means that the *shape function* in this conforming finite element space is continuous together with its derivatives up to the order $r-1$ [93]. Otherwise, the finite element is said to be nonconforming. We develop an approximation operator based on triangular elements, which use not only the values of the function at a certain set of points, but also the values of some line integrals.

5.1.1 Triangular linear element

We assume that the data are related to a reference triangle $S_2 \subset \Omega \subset \mathbb{R}^2$ contained in a compact domain Ω, with vertices $\boldsymbol{v}_i = (x_i, y_i)$, $i = 0, 1, 2$ and nonzero signed area

$$A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{2} \begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \end{vmatrix}.$$

The barycentric coordinates $\lambda_0(\boldsymbol{x}), \lambda_1(\boldsymbol{x}), \lambda_2(\boldsymbol{x})$ of the point $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$ with respect to the reference triangle S_2 are defined by the area-ratios

$$\lambda_0(\boldsymbol{x}) = \frac{A(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2)}{A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2)}, \quad \lambda_1(\boldsymbol{x}) = \frac{A(\boldsymbol{v}_0, \boldsymbol{x}, \boldsymbol{v}_2)}{A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2)}, \quad \lambda_2(\boldsymbol{x}) = \frac{A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{x})}{A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2)}. \tag{5.1}$$

For $f \in C(S_2)$, we set

$$L_j(f) = f(\boldsymbol{v}_j), \qquad j = 0, 1, 2. \tag{5.2}$$

The standard triangular linear finite element is the triple

$$\mathcal{P}_1(S_2) = (S_2, \mathbb{P}_1(S_2), \Sigma_{S_2}^{\text{lin}}), \tag{5.3}$$

where

$$\mathbb{P}_1(S_2) = \text{span}\{\lambda_0, \lambda_1, \lambda_2\}$$

is the space of bivariate linear polynomials and

$$\Sigma_{S_2}^{\text{lin}} = \{L_0(f),\, L_1(f),\, L_2(f)\}$$

is the set of point evaluation functionals at the vertices of the triangle S_2, called *degrees of freedom*.

Theorem 5.1.1. *The linear approximation operator based on the standard triangular linear finite element* $\mathcal{P}_1(S_2)$*, defined in* (5.3)

$$\begin{aligned} \Pi^{\text{lin}} : C(S_2) &\to \mathbb{P}_1(S_2) \\ f &\mapsto \sum_{j=0}^{2} L_j(f)\lambda_j \end{aligned}$$

reproduces linear polynomials and satisfies the interpolation conditions

$$L_j\left(\Pi^{\text{lin}}[f]\right) = L_j(f), \qquad j = 0, 1, 2.$$

Proof. The proof follows from the Lagrange property of the barycentric coordinates, that is $\lambda_i(\boldsymbol{v}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta operator. □

5.1.2 Quadratic polynomial enrichment

For each $i = 0, 1, 2$, we denote by Γ_i the edge of S_2 opposite to the vertex $\boldsymbol{v}_i$ and by $|\Gamma_i|$ its euclidean length, that is

$$|\Gamma_0| = \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2, \quad |\Gamma_1| = \|\boldsymbol{v}_0 - \boldsymbol{v}_2\|_2, \quad |\Gamma_2| = \|\boldsymbol{v}_0 - \boldsymbol{v}_1\|_2,$$

where $\|\cdot\|_2$ is the L^2-norm in $\mathbb{R}^2$. Our goal is to develop an approximation operator based on the triangle S_2, which uses not only the evaluation functionals $L_j(f)$, $j = 0, 1, 2$, but also the values of the line integrals

$$\frac{1}{|\Gamma_j|}\int_{\Gamma_j} f(\boldsymbol{x})d\sigma(\boldsymbol{x}), \qquad j = 0, 1, 2,$$

that we assume are given, where the integral is computed with respect to the Lebesgue measure on Γ_j, $j = 0, 1, 2$. The idea is to enrich the standard triangular linear finite element $\mathcal{P}_1(S_2)$, by using the above introduced functionals and quadratic polynomial functions. With this aim, we set

$$\mathbb{P}_2(S_2) = \mathbb{P}_1(S_2) \oplus \operatorname{span}\{\lambda_0\lambda_1, \lambda_0\lambda_2, \lambda_1\lambda_2\}$$

and

$$\Sigma_{S_2}^{\mathrm{enr}} = \{L_j, I_j \,:\, j = 0, 1, 2\},$$

where

$$I_j(f) = \frac{1}{|\Gamma_j|}\int_{\Gamma_j} f(\boldsymbol{x})d\sigma(\boldsymbol{x}), \qquad j = 0, 1, 2. \tag{5.4}$$

To show that the triple

$$(S_2, \mathbb{P}_2(S_2), \Sigma_{S_2}^{\mathrm{enr}})$$

is a finite element we have to prove that $\mathbb{P}_2(S_2)$ is $\Sigma_{S_2}^{\mathrm{enr}}$-unisolvent, or, equivalently that the set of linear functionals $\Sigma_{S_2}^{\mathrm{enr}}$ is linearly independent in the dual space $\mathbb{P}_2(S_2)^\star$. To this aim we need the following technical Lemma, which results from a direct application of the Simpson's rule to the line integrals $I_j(p)$, $j = 0, 1, 2$, $p \in \mathbb{P}_2(S_2)$. We denote by

$$\boldsymbol{v}_{01} = \frac{\boldsymbol{v}_0 + \boldsymbol{v}_1}{2} \in \Gamma_2, \quad \boldsymbol{v}_{02} = \frac{\boldsymbol{v}_0 + \boldsymbol{v}_2}{2} \in \Gamma_1, \quad \boldsymbol{v}_{12} = \frac{\boldsymbol{v}_1 + \boldsymbol{v}_2}{2} \in \Gamma_0,$$

the midpoints of the sides of S_2.

Lemma 5.1.2. *Let $p \in \mathbb{P}_2(S_2)$. Then we have*

$$\begin{aligned} I_0(p) &= \frac{1}{6}\left(p(\boldsymbol{v}_1) + 4p(\boldsymbol{v}_{12}) + p(\boldsymbol{v}_2)\right), \\ I_1(p) &= \frac{1}{6}\left(p(\boldsymbol{v}_0) + 4p(\boldsymbol{v}_{02}) + p(\boldsymbol{v}_2)\right), \\ I_2(p) &= \frac{1}{6}\left(p(\boldsymbol{v}_0) + 4p(\boldsymbol{v}_{01}) + p(\boldsymbol{v}_1)\right). \end{aligned}$$

Proof. Let us prove the first identity. We parametrize the edge Γ_0 by

$$t \to t\boldsymbol{v}_2 + (1-t)\boldsymbol{v}_1, \qquad t \in [0, 1].$$

Since the restriction of p to the edge Γ_0 is a quadratic polynomial in t, the Simpson's rule provides exact results for the integral, and then

$$I_0(p) = \frac{1}{|\Gamma_0|}\int_{\Gamma_0} p(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \int_0^1 p(t\boldsymbol{v}_2 + (1-t)\boldsymbol{v}_1)dt = \frac{1}{6}\left(p(\boldsymbol{v}_1) + 4p(\boldsymbol{v}_{12}) + p(\boldsymbol{v}_2)\right).$$

The same argument can be used to prove the other identities. □

Remark 5.1.3. *Note that, if $p \in \mathbb{P}_2(S_2)$ and $L_j(p) = I_j(p) = 0$, $j = 0, 1, 2$, then, by Lemma 5.1.2, we get*

$$p(\boldsymbol{v}_{12}) = p(\boldsymbol{v}_{02}) = p(\boldsymbol{v}_{01}) = 0,$$

that is, p vanishes at the midpoints of the sides of S_2.

Since p vanishes at the vertices of S_2 and at the midpoints of the sides of S_2, the unisolvence of the set $\Sigma_{S_2}^{\mathrm{enr}}$ in the polynomial space $\mathbb{P}_2(S_2)$ follows by a classical result of multivariate polynomial interpolation [18, Ch. 10]. However, we give an alternative proof of this result, which can also be adapted in the case of more general polynomial or nonpolynomial enrichments of the triple $(S_2, \mathbb{P}_1(S_2), \Sigma_{S_2}^{\mathrm{lin}})$ by using the same linear functionals and vector spaces

$$\mathbb{P}_1^{\mathrm{enr}}(S_2) = \mathbb{P}_1(S_2) \oplus \mathrm{span}\left\{l_0\lambda_1\lambda_2, l_1\lambda_0\lambda_2, l_2\lambda_0\lambda_1\right\},$$
$$\mathbb{P}_1^{\mathrm{enr}}(S_2) = \mathbb{P}_1(S_2) \oplus \mathrm{span}\left\{l_0\lambda_1^{\alpha_0-1}\lambda_2^{\beta_0-1}, l_1\lambda_0^{\alpha_1-1}\lambda_2^{\beta_1-1}, l_2\lambda_0^{\alpha_2-1}\lambda_1^{\beta_2-1}\right\},\ \alpha_i, \beta_i > 1,\ i = 0, 1, 2,$$

where l_i, $i = 0, 1, 2$, are linear polynomials satisfying the nonvanishing conditions at the special points

$$\begin{aligned}
\widetilde{\boldsymbol{v}}_{12} &= \frac{\alpha_0}{\alpha_0+\beta_0}\boldsymbol{v}_1 + \frac{\beta_0}{\alpha_0+\beta_0}\boldsymbol{v}_2,\\
\widetilde{\boldsymbol{v}}_{02} &= \frac{\alpha_1}{\alpha_1+\beta_1}\boldsymbol{v}_0 + \frac{\beta_1}{\alpha_1+\beta_1}\boldsymbol{v}_2,\\
\widetilde{\boldsymbol{v}}_{01} &= \frac{\alpha_2}{\alpha_2+\beta_2}\boldsymbol{v}_0 + \frac{\beta_2}{\alpha_2+\beta_2}\boldsymbol{v}_1,
\end{aligned}$$

of the side of S_2. These enrichments will be studied in the next chapter.

Theorem 5.1.4. *The triple $(S_2, \mathbb{P}_2(S_2), \Sigma_{S_2}^{\mathrm{enr}})$ is a finite element.*

Proof. We have to show that $\mathbb{P}_2(S_2)$ is $\Sigma_{S_2}^{\mathrm{enr}}$-unisolvent, i.e., if $p \in \mathbb{P}_2(S_2)$ and

$$L_j(p) = 0, \qquad j = 0, 1, 2, \tag{5.5}$$

$$I_j(p) = 0, \qquad j = 0, 1, 2, \tag{5.6}$$

then $p = 0$ [23, Ch. 2]. Let $p \in \mathbb{P}_2(S_2)$ and assume that (5.5) and (5.6) hold. By definition p can be represented as

$$p = \alpha_0\lambda_0 + \alpha_1\lambda_1 + \alpha_2\lambda_2 + \beta_0\lambda_1\lambda_2 + \beta_1\lambda_0\lambda_2 + \beta_2\lambda_0\lambda_1,$$

for some constants α_i and β_i, $i = 0, 1, 2$. The barycentric coordinates satisfy Lagrange property, that is $L_j(\lambda_i) = \lambda_i(\boldsymbol{v}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta operator, and then $\lambda_i\lambda_j$ vanishes at the vertices $\boldsymbol{v}_k$ for each $i, j, k = 0, 1, 2$, $i \neq j$. Then we get $\alpha_0 = \alpha_1 = \alpha_2 = 0$ and p reduces to

$$p = \beta_0\lambda_1\lambda_2 + \beta_1\lambda_0\lambda_2 + \beta_2\lambda_0\lambda_1.$$

On the other hand, $I_0(p) = 0$ and $L_1(p) = L_2(p) = 0$ imply $p(\boldsymbol{v}_{12}) = 0$. Now, since $\lambda_1(\boldsymbol{v}_{12}) = \lambda_2(\boldsymbol{v}_{12}) = \frac{1}{2}$ and $\lambda_0(\boldsymbol{v}_{12}) = 0$, we get

$$0 = \frac{4}{6}p(\boldsymbol{v}_{12}) = \frac{\beta_0}{6}.$$

Thus $\beta_0 = 0$. A similar argument can be used to show that $\beta_1 = \beta_2 = 0$, therefore $p = 0$ and the proof of the theorem is completed. □

Since we have shown the linear independence of the functionals L_j, I_j, $j = 0, 1, 2$ in the dual space $\mathbb{P}_2(S_2)^\star$, there exists a related biorthonormal set of polynomials $\{\varphi_j, \psi_j : j = 0, 1, 2\}$, which span $\mathbb{P}_2(S_2)$ and satisfy

$$L_j(\varphi_i) = \delta_{ij}, \quad I_j(\varphi_i) = 0, \qquad i, j = 0, 1, 2, \tag{5.7}$$

$$L_j(\psi_i) = 0, \quad I_j(\psi_i) = \delta_{ij}, \qquad i, j = 0, 1, 2. \tag{5.8}$$

Theorem 5.1.5. *The basis functions $\{\varphi_j, \psi_j : j = 0, 1, 2\}$ of $\mathbb{P}_2(S_2)$ associated to the finite element $(S_2, \mathbb{P}_2(S_2), \Sigma_{S_2}^{\mathrm{enr}})$, which satisfy* (5.7) *and* (5.8) *have the following expressions*

$$\varphi_0 = \lambda_0(1 - 3\lambda_1 - 3\lambda_2), \quad \varphi_1 = \lambda_1(1 - 3\lambda_0 - 3\lambda_2), \quad \varphi_2 = \lambda_2(1 - 3\lambda_0 - 3\lambda_1), \tag{5.9}$$

$$\psi_0 = 6\lambda_1\lambda_2, \quad \psi_1 = 6\lambda_0\lambda_2, \quad \psi_2 = 6\lambda_0\lambda_1. \tag{5.10}$$

Proof. Let us prove the first of identities (5.9) for the element $\varphi_0 \in \mathbb{P}_2(S_2)$. It can be represented as

$$\varphi_0 = \alpha_0\lambda_0 + \alpha_1\lambda_1 + \alpha_2\lambda_2 + \beta_0\lambda_1\lambda_2 + \beta_1\lambda_0\lambda_2 + \beta_2\lambda_0\lambda_1 \tag{5.11}$$

for some constants α_i and β_i, $i = 0, 1, 2$. Since $L_0(\varphi_0) = 1$, $L_1(\varphi_0) = L_2(\varphi_0) = 0$, the Lagrange property of the barycentric coordinates implies $\alpha_0 = 1$, $\alpha_1 = \alpha_2 = 0$, therefore

$$\varphi_0 = \lambda_0 + \beta_0\lambda_1\lambda_2 + \beta_1\lambda_0\lambda_2 + \beta_2\lambda_0\lambda_1.$$

Moreover, since $\lambda_0 = 0$ on Γ_0, from $I_0(\varphi_0) = 0$, by applying the Simpson's rule, we get

$$0 = \frac{1}{|\Gamma_0|}\int_{\Gamma_0} \beta_0\lambda_1(\boldsymbol{x})\lambda_2(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \frac{4}{6}\beta_0\lambda_1(\boldsymbol{v}_{12})\lambda_2(\boldsymbol{v}_{12}) = \frac{\beta_0}{6},$$

which readily gives $\beta_0 = 0$. Similarly, since $\lambda_1 = 0$ on Γ_1, from $I_1(\varphi_0) = 0$ we get

$$0 = \frac{1}{|\Gamma_1|}\int_{\Gamma_1} (\lambda_0(\boldsymbol{x}) + \beta_1\lambda_0(\boldsymbol{x})\lambda_2(\boldsymbol{x}))d\sigma(\boldsymbol{x}) = \frac{1}{6}\left(1 + 4\left(\frac{1}{2} + \beta_1\lambda_0(\boldsymbol{v}_{02})\lambda_2(\boldsymbol{v}_{02})\right)\right) = \frac{1}{6}(3 + \beta_1),$$

that implies $\beta_1 = -3$. Similarly, since $\lambda_2 = 0$ on Γ_2, from $I_2(\varphi_0) = 0$ we get $\beta_2 = -3$.

Let us now prove the first of identities (5.10). Since $L_j(\psi_0) = 0$, $j = 0, 1, 2$, the Lagrange property of the barycentric coordinates implies that $\psi_0 \in \operatorname{span}\{\lambda_1\lambda_2, \lambda_0\lambda_2, \lambda_0\lambda_1\}$ and then it can be written as

$$\psi_0 = \beta_0\lambda_1\lambda_2 + \beta_1\lambda_0\lambda_2 + \beta_2\lambda_0\lambda_1,$$

for some constant β_i, $i = 0, 1, 2$. Moreover, from $I_0(\psi_0) = 1$ we get

$$1 = \frac{1}{|\Gamma_0|}\int_{\Gamma_0} \beta_0\lambda_1(\boldsymbol{x})\lambda_2(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \frac{4}{6}\beta_0\lambda_1(\boldsymbol{v}_{12})\lambda_2(\boldsymbol{v}_{12}) = \frac{\beta_0}{6}.$$

Hence, $\beta_0 = 6$. From $I_1(\psi_0) = I_2(\psi_0) = 0$, we also get $\beta_1 = \beta_2 = 0$.

Using symmetry arguments, we can obtain the expressions for the other functions in equations (5.9) and (5.10). □

Remark 5.1.6. *By setting*

$$e_0 = \lambda_1\lambda_2, \quad e_1 = \lambda_0\lambda_2, \quad e_2 = \lambda_0\lambda_1,$$

the basis functions associated to the enriched element $(S_2, \mathbb{P}_2(S_2), \Sigma^{\mathrm{enr}}_{S_2})$ *can be rewritten as follows*

$$\varphi_j = \lambda_j - \frac{1}{2}\sum_{\substack{k=0\\k\neq j}}^{2}\psi_k, \quad \psi_j = 6e_j, \qquad j = 0, 1, 2. \tag{5.12}$$

Theorem 5.1.7. *The quadratic approximation operator based on the finite element* $(S_2, \mathbb{P}_2(S_2), \Sigma^{\mathrm{enr}}_{S_2})$

$$\begin{aligned} \Pi^{\mathrm{enr}} : C(S_2) &\to \mathbb{P}_2(S_2) \\ f &\mapsto \sum_{j=0}^{2} L_j(f)\varphi_j + \sum_{j=0}^{2} I_j(f)\psi_j \end{aligned}$$

reproduces quadratic polynomials and satisfies the interpolation conditions

$$L_j\left(\Pi^{\mathrm{enr}}[f]\right) = L_j(f), \quad I_j\left(\Pi^{\mathrm{enr}}[f]\right) = I_j(f), \qquad j = 0, 1, 2.$$

Proof. The proof follows by conditions (5.7) and (5.8). □

Remark 5.1.8. *We note that the operator* Π^{enr} *depends on the triangle* S_2*. In order to highlight this dependence, when necessary we will write* $\Pi^{\mathrm{enr}}[f, S_2]$ *instead of* $\Pi^{\mathrm{enr}}[f]$ *(see Section 5.3).*

5.2 Error bound

We are interested in evaluating or estimating the approximation error

$$E^{\text{enr}}[f] = f - \Pi^{\text{enr}}[f] \tag{5.13}$$

under certain hypotheses on the differentiability class of the function f. In the following representation we make use of a generalization of the classical trapezoidal formula to the case of line integrals. We define the following linear operators

$$\mathcal{L}_k = \frac{1}{2}\sum_{\substack{j=0\\ j\neq k}}^{2} L_j, \quad \mathcal{E}_k^{\text{tra}} = \mathcal{L}_k - I_k, \qquad k = 0, 1, 2.$$

The following Theorem shows that the error (5.13) can be decomposed in two parts: the first one is related to the linear triangular element while the second one depends on the enrichment functions e_i, $i = 0, 1, 2$.

Theorem 5.2.1. *For all $f \in C(\Omega)$, the approximation error is given by*

$$E^{\text{enr}}[f] = E^{\text{lin}}[f] + E^{\text{tra}}[f],$$

where

$$E^{\text{lin}}[f] = f - \sum_{j=0}^{2} L_j(f)\lambda_j \quad \textit{and} \quad E^{\text{tra}}[f] = 6\sum_{k=0}^{2} \mathcal{E}_k^{\text{tra}}(f)e_k.$$

Proof. It follows easily from (5.12), by interchanging the order of the double sum, that

$$\begin{aligned}
\sum_{j=0}^{2} L_j(f)\varphi_j &= \sum_{j=0}^{2} L_j(f)\left(\lambda_j - \sum_{k=0}^{2} 3e_k\left(1-\delta_{jk}\right)\right)\\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} 3e_k \sum_{j=0}^{2}\left(1-\delta_{jk}\right) L_j(f)\\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} 3e_k \sum_{\substack{j=0\\ j\neq k}}^{2} L_j(f)\\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} 6\mathcal{L}_k(f)e_k.
\end{aligned}$$

Finally, we get

$$\begin{aligned}
E^{\text{enr}}[f] = f - \Pi^{\text{enr}}[f] &= f - \sum_{j=0}^{2} L_j(f)\varphi_j - \sum_{j=0}^{2} I_j(f)\psi_j\\
&= f - \sum_{j=0}^{2} L_j(f)\lambda_j + \sum_{k=0}^{2} 6\mathcal{L}_k(f)e_k - \sum_{j=0}^{2} I_j(f)\psi_j\\
&= E^{\text{lin}}[f] + 6\sum_{k=0}^{2}\left(\mathcal{L}_k(f) - I_k(f)\right)e_k\\
&= E^{\text{lin}}[f] + E^{\text{tra}}[f]
\end{aligned}$$

as desired. □

Now we derive another expression for the remainder of the quadratic operator, which will be helpful in the next Section, to determine the approximation order of the enriched triangular Shepard method. We denote by $C^{2,1}(\Omega)$ the space of functions $f \in C^2(\Omega)$ with partial derivatives $\frac{\partial f}{\partial x^{2-j}\partial y^j}$, $j = 0, 1, 2$, Lipschitz-continuous in Ω. For each function $f \in C^{2,1}(\Omega)$ we define

$$|f|_{2,1} = \sup \left\{ \frac{\left| \frac{\partial f}{\partial x^{2-j}\partial y^j}(\boldsymbol{x}) - \frac{\partial f}{\partial x^{2-j}\partial y^j}(\boldsymbol{y}) \right|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2} : \boldsymbol{x}, \boldsymbol{y} \in \Omega \text{ and } \boldsymbol{x} \neq \boldsymbol{y} \right\}.$$

We are interested in evaluating or estimating the approximation error $E^{\text{enr}}[f]$ for $f \in C^{2,1}(\Omega)$. Let $T_2[f, \boldsymbol{x}_B]$ be the Taylor polynomial of order 2 for f at the barycenter $\boldsymbol{x}_B = (x_B, y_B)$ of S_2. Then, we get

$$f = T_2[f, \boldsymbol{x}_B] + R_{T_2}[f], \tag{5.14}$$

where

$$R_{T_2}[f](\boldsymbol{x}) = \frac{1}{3!} \int_0^1 D^{(3)}_{\boldsymbol{x}-\boldsymbol{x}_B} f(\boldsymbol{x}_B + s(\boldsymbol{x} - \boldsymbol{x}_B))(1-s)^2 ds, \qquad \boldsymbol{x} = (x, y) \in \Omega, \tag{5.15}$$

and $D_{\boldsymbol{x}-\boldsymbol{x}_B}$ is the directional derivative along the vector $\boldsymbol{x} - \boldsymbol{x}_B$. Since the interpolation operator Π^{enr} based on the finite element $(S_2, \mathbb{P}_2(S_2), \Sigma^{\text{enr}}_{S_2})$, reproduces polynomials of degree ≤ 2, by applying it to both sides of (5.14), we get

$$\Pi^{\text{enr}}[f] = T_2[f, \boldsymbol{x}_B] + \Pi^{\text{enr}}[R_{T_2}[f]],$$

and then

$$E^{\text{enr}}[f] = f - \Pi^{\text{enr}}[f] = f - T_2[f, \boldsymbol{x}_B] - \Pi^{\text{enr}}[R_{T_2}[f]] = R_{T_2}[f] - \Pi^{\text{enr}}[R_{T_2}[f]]. \tag{5.16}$$

In order to prove the next Theorem, some preliminary Lemmas are needed. We set

$$h = \max_{i=0,1,2} |\Gamma_i| \qquad \text{and} \qquad S = \frac{1}{|A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2)|}.$$

Lemma 5.2.2. *For any $\boldsymbol{x} \in \Omega$, the barycentric coordinates are bounded by*

$$|\lambda_j(\boldsymbol{x})| \leq Sh \left(\|\boldsymbol{x} - \boldsymbol{x}_B\|_2 + h \right), \qquad j = 0, 1, 2.$$

Proof. Without loss of generality, let us consider the case $j = 0$. By equation (5.1) we get

$$\lambda_0(\boldsymbol{x}) = \frac{1}{2A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2)} \begin{vmatrix} 1 & 1 & 1 \\ x & x_1 & x_2 \\ y & y_1 & y_2 \end{vmatrix} = \frac{(y_1 - y_2)(x - x_1) - (y - y_1)(x_1 - x_2)}{2A(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2)},$$

then

$$|\lambda_0(\boldsymbol{x})| \leq \frac{S}{2} \left(|y_1 - y_2||x - x_1| + |y - y_1||x_1 - x_2| \right) \leq \frac{Sh}{2} \left(|x - x_1| + |y - y_1| \right).$$

Consequently

$$\begin{aligned} |\lambda_0(\boldsymbol{x})| &\leq \frac{Sh}{2} \left(|x - x_1 + x_B - x_B| + |y - y_1 + y_B - y_B| \right) \\ &\leq \frac{Sh}{2} \left(|x - x_B| + |x_B - x_1| + |y - y_B| + |y_B - y_1| \right) \\ &\leq \frac{Sh}{2} \left(\|\boldsymbol{x} - \boldsymbol{x}_B\|_2 + h + \|\boldsymbol{x} - \boldsymbol{x}_B\|_2 + h \right) \leq Sh \left(\|\boldsymbol{x} - \boldsymbol{x}_B\|_2 + h \right). \end{aligned}$$

The proof is completed by noting that the same argument can be done for $j = 1$ and $j = 2$. □

Lemma 5.2.3. *Let $f \in C^{2,1}(\Omega)$. Then we have*

$$|L_j(R_{T_2}[f])| \leq \frac{4}{9} h^3 |f|_{2,1}, \qquad j = 0, 1, 2 \tag{5.17}$$

and

$$|I_j(R_{T_2}[f])| \leq \frac{4}{9} h^3 |f|_{2,1}, \qquad j = 0, 1, 2. \tag{5.18}$$

Proof. By [32, Lemma 12], for each $s \in [0,1]$, we get

$$\left| D^{(3)}_{\boldsymbol{v}_j - \boldsymbol{x}_B} f(\boldsymbol{x}_B + s(\boldsymbol{v}_j - \boldsymbol{x}_B)) \right| \leq 8 \left\| \boldsymbol{v}_j - \boldsymbol{x}_B \right\|_2^3 |f|_{2,1} \leq 8h^3 |f|_{2,1}, \qquad j = 0,1,2.$$

Consequently, from (5.2) and (5.15), by using the triangular inequality, we obtain

$$|L_j(R_{T_2}[f])| = \frac{1}{3!} \left| \int_0^1 D^{(3)}_{\boldsymbol{v}_j - \boldsymbol{x}_B} f(\boldsymbol{x}_B + s(\boldsymbol{v}_j - \boldsymbol{x}_B))(1-s)^2 ds \right| \leq \frac{4}{9} h^3 |f|_{2,1}, \qquad j = 0,1,2.$$

Analogously, from (5.4) and (5.15) we get

$$\begin{aligned} |I_j(R_{T_2}[f])| &= \frac{1}{|\Gamma_j|} \left| \int_{\Gamma_j} \frac{1}{3!} \int_0^1 D^{(3)}_{\boldsymbol{x} - \boldsymbol{x}_B} f(\boldsymbol{x}_B + s(\boldsymbol{x} - \boldsymbol{x}_B))(1-s)^2 ds d\sigma(\boldsymbol{x}) \right| \\ &\leq \frac{1}{|\Gamma_j|} \int_{\Gamma_j} \frac{8}{3!} h^3 |f|_{2,1} \int_0^1 (1-s)^2 ds d\sigma(\boldsymbol{x}) \\ &\leq \frac{4}{9} h^3 |f|_{2,1}, \qquad j = 0,1,2. \end{aligned}$$

□

Theorem 5.2.4. *Let $f \in C^{2,1}(\Omega)$. Then, for any $\boldsymbol{x} \in \Omega$, we get*

$$|E^{\text{enr}}[f](\boldsymbol{x})| \leq 4|f|_{2,1} \left(\frac{\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2^3}{9} + \frac{1}{3} \sum_{k=1}^{2} 12^{k-1} h^{3-k} C^k (\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2 + h)^k \right), \tag{5.19}$$

where $C = Sh^2$.

Proof. By applying the operator Π^{enr} to the remainder term $R_{T_2}[f]$ and by rearranging, we get

$$\begin{aligned} \Pi^{\text{enr}}[R_{T_2}[f]] &= L_0(R_{T_2}[f])\lambda_0 + L_1(R_{T_2}[f])\lambda_1 + L_2(R_{T_2}[f])\lambda_2 + \\ &+ \left(6I_0(R_{T_2}[f]) - 3L_1(R_{T_2}[f]) - 3L_2(R_{T_2}[f])\right)\lambda_1\lambda_2 + \\ &+ \left(6I_1(R_{T_2}[f]) - 3L_0(R_{T_2}[f]) - 3L_2(R_{T_2}[f])\right)\lambda_0\lambda_2 + \\ &+ \left(6I_2(R_{T_2}[f]) - 3L_0(R_{T_2}[f]) - 3L_1(R_{T_2}[f])\right)\lambda_0\lambda_1. \end{aligned}$$

By Lemma 5.2.2, Lemma 5.2.3 and by the triangular inequality, we have

$$|\Pi^{\text{enr}}[R_{T_2}[f]](\boldsymbol{x})| \leq \frac{4}{3} h^2 |f|_{2,1} C \left(\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2 + h \right) + 16h|f|_{2,1} C^2 \left(\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2 + h \right)^2. \tag{5.20}$$

By (5.20) and by bounding the Taylor remainder in standard way, we finally get

$$\begin{aligned} & |E^{\text{enr}}[f](\boldsymbol{x})| \leq |R_{T_2}[f](\boldsymbol{x})| + |\Pi^{\text{enr}}[R_{T_2}[f]](\boldsymbol{x})| \\ \leq\ & \frac{4}{9} \left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2^3 |f|_{2,1} + \frac{4}{3} h^2 |f|_{2,1} C (\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2 + h) + 16h|f|_{2,1} C^2 (\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2 + h)^2 \\ \leq\ & 4|f|_{2,1} \left(\frac{\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2^3}{9} + \frac{1}{3} \sum_{k=1}^{2} 12^{k-1} h^{3-k} C^k (\left\| \boldsymbol{x} - \boldsymbol{x}_B \right\|_2 + h)^k \right). \end{aligned}$$

□

5.3 Enriched triangular Shepard method

Let $X_n = \{\boldsymbol{x}_i : i = 1 \ldots, n\}$ be a set of n scattered data in a compact domain $\Omega \subset \mathbb{R}^2$ and let $\mathcal{T} = \{t_j : j = 1, \ldots, m\}$ be a triangulation of X_n, where t_j is the triangle with vertices $\boldsymbol{x}_{j_k}$, $k = 1,2,3$. The

triangular Shepard basis functions are defined as follows [36]

$$B_{\mu,j}(\boldsymbol{x}) = \frac{\prod_{k=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_k}\|_2^{-\mu}}{\sum_{k=1}^{m} \prod_{l=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{k_l}\|_2^{-\mu}}, \qquad j = 1, \ldots, m, \qquad \mu > 0.$$

They satisfy the following properties:

$$B_{\mu,j}(\boldsymbol{x}) \geq 0, \tag{5.21}$$

$$\sum_{j=1}^{m} B_{\mu,j}(\boldsymbol{x}) = 1, \tag{5.22}$$

$$B_{\mu,j}(\boldsymbol{x}_i) = 0, \tag{5.23}$$

for each $j = 1, \ldots, m$ and $\boldsymbol{x}_i$ which is not a vertex of t_j. For more details, see [36, 27].

Let $f : \Omega \to \mathbb{R}$ be a continuous function. We assume that the values $f_i = f(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, and the values of the line integrals of the function f along each segment $[\boldsymbol{x}_i, \boldsymbol{x}_j]$, $i \neq j$ are given. The *enriched triangular Shepard operator* is defined as follows

$$K_\mu^{\mathrm{enr}}(\boldsymbol{x}) = \sum_{j=1}^{m} B_{\mu,j}(\boldsymbol{x}) \Pi^{\mathrm{enr}}[f, t_j](\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{5.24}$$

where $\Pi^{\mathrm{enr}}[f, t_j](\boldsymbol{x})$ is the quadratic approximation operator based on the triangular element $(t_j, \mathbb{P}_2(t_j), \Sigma_{t_j}^{\mathrm{enr}})$.

In order to determine the approximation order of the enriched triangular Shepard operator, we need the following notations. We denote by $\|\cdot\|_\infty$ the maximum norm of $\mathbb{R}^2$ and by

$$R_r(\boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{R}^2 \,:\, \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \leq r\}$$

the ball centered in $\boldsymbol{y}$ with radius $r \geq 0$. Let $V(t)$ be the set of vertices of the triangle $t \in \mathcal{T}$. We define

$$h' = \inf\{r > 0 \,:\, \forall \boldsymbol{x} \in \Omega \, \exists t \in \mathcal{T} : R_r(\boldsymbol{x}) \cap V(t) = \emptyset\}, \tag{5.25}$$

$$h'' = \inf\{r > 0 \,:\, \forall t \in \mathcal{T} \, \exists \boldsymbol{x} \in \Omega : t \subset R_r(\boldsymbol{x})\} \tag{5.26}$$

and

$$h = \max\{h', h''\}.$$

It is worth noting that a small value of h implies a rather uniform triangle distribution and excludes the presence of large triangles. Finally, we set

$$M = \sup_{x \in \Omega} \#\{t \in \mathcal{T} \,:\, R_h(\boldsymbol{x}) \cap V(t) \neq \emptyset\}, \tag{5.27}$$

where $\#\{\cdot\}$ is the cardinality operator. In line with [31] it is possible to prove the following theorem.

Theorem 5.3.1. *Let Ω be a compact, convex domain containing X_n, $f \in C^{2,1}(\Omega)$ and $\mu > \frac{5}{3}$. Then*

$$|f(\boldsymbol{x}) - K_\mu^{\mathrm{enr}}(\boldsymbol{x})| \leq CM|f|_{2,1}h^3, \qquad \boldsymbol{x} \in \Omega,$$

with C a positive constant which depends only on $\mathcal{T}$ and μ.

Proof. For any $\boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2$, we denote by

$$Q_r(\boldsymbol{y}) = \{\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2 \,:\, y_k - r < x_k \leq y_k + r, \, k = 1, 2\}$$

the axis-aligned half open square of center $\boldsymbol{y} = (y_1, y_2)$ and side length $2r$ and we consider the covering of Ω by pairwise disjoint sets $\{U_j(\boldsymbol{x})\}_{j \in \mathbb{N}_0}$, where

$$U_j(\boldsymbol{x}) = \bigcup_{\substack{\nu \in \mathbb{Z}^2 \\ \|\nu\|_\infty = j}} Q_h(\boldsymbol{x} + 2h\nu).$$

More precisely, $U_0(\boldsymbol{x})$ is the half open square of center $\boldsymbol{x}$ and side length $2h$ while, for $j > 0$, $U_j(\boldsymbol{x})$ is the half-open annulus with center $\boldsymbol{x}$, radius $2hj$ and thickness $2h$. Since the set Ω is compact, there exists $N \in \mathbb{N}_0$, independent on $\boldsymbol{x}$ and of order $O(1/h)$, such that

$$\Omega \subset \bigcup_{j=0}^{N} U_j(\boldsymbol{x}). \tag{5.28}$$

Note that, by the definition (5.27), we get

$$\#\{t \in \mathcal{T} \,:\, V(t) \cap U_j \neq \emptyset\} \leq 8jM, \qquad j = 1, \ldots, N. \tag{5.29}$$

For any $t \in \mathcal{T}$ with at least one vertex in U_j, only one of the following cases is possible:

$$\begin{aligned} V(t) \cap U_{j-1} \neq \emptyset \quad &\Longrightarrow \quad (2j-3)h \leq \|\boldsymbol{x} - \boldsymbol{v}\|_\infty \leq (2j+1)\,h, \ \forall \boldsymbol{v} \in V(t), \\ V(t) \subset U_j \quad &\Longrightarrow \quad (2j-1)h \leq \|\boldsymbol{x} - \boldsymbol{v}\|_\infty \leq (2j+1)\,h, \ \forall \boldsymbol{v} \in V(t), \\ V(t) \cap U_{j+1} \neq \emptyset \quad &\Longrightarrow \quad (2j-1)h \leq \|\boldsymbol{x} - \boldsymbol{v}\|_\infty \leq (2j+3)\,h, \ \forall \boldsymbol{v} \in V(t). \end{aligned} \tag{5.30}$$

Now we denote by T_0 the set of all triangles with at least one vertex in U_0. It follows from the definition of h' (5.25) and of M (5.27) that

$$0 < \#(T_0) \leq M.$$

Since one vertex of $t_j \in T_0$ lies in U_0 and the remaining ones lies in $U_0 \cup U_1$, we get

$$\prod_{i=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_i}\|_\infty \leq 9h^3. \tag{5.31}$$

For $k = 1, \ldots, N$, we denote by T_k the set of the triangles with at least one vertex in U_k and no vertex in U_{k-1}. By definition (5.29), the cardinality of this set is bounded as follows

$$\#(T_k) \leq 8kM$$

and each $t_j \in T_k$ satisfies

$$((2k-1)\,h)^3 \leq \prod_{i=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_i}\|_\infty \leq ((2k+3)\,h)^3\,.$$

Further we note that

$$\bigcup_{k=0}^{N} T_k = \mathcal{T} \quad \text{and} \quad \bigcap_{k=0}^{N} T_k = \emptyset.$$

Finally, we set

$$e(\boldsymbol{x}) = |f(\boldsymbol{x}) - K_\mu^{\mathrm{enr}}(\boldsymbol{x})|, \qquad \boldsymbol{x} \in \Omega.$$

By (5.24) and by the properties (5.21) and (5.22) of the triangular Shepard basis function $B_{\mu,j}$, we get

$$e(\boldsymbol{x}) = \left| \sum_{j=1}^{m} B_{\mu,j}(\boldsymbol{x}) f(\boldsymbol{x}) - \sum_{j=1}^{m} B_{\mu,j}(\boldsymbol{x}) \Pi^{\mathrm{enr}}[f, t_j](\boldsymbol{x}) \right| \leq \sum_{j=1}^{m} |f(\boldsymbol{x}) - \Pi^{\mathrm{enr}}[f, t_j](\boldsymbol{x})| B_{\mu,j}(\boldsymbol{x}).$$

With reference to Theorem 5.2.4, we denote by $\boldsymbol{x}_{B_j}$ the barycenter of the triangle t_j and we set

$$h_j = \max\left\{ \|\boldsymbol{x}_{j_1} - \boldsymbol{x}_{j_2}\|_2\,, \|\boldsymbol{x}_{j_2} - \boldsymbol{x}_{j_3}\|_2\,, \|\boldsymbol{x}_{j_3} - \boldsymbol{x}_{j_1}\|_2 \right\}, \qquad S_j = \frac{1}{|A(\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, \boldsymbol{x}_{j_3})|}$$

and $C_j = S_j h_j^2$, $j = 1, \ldots, m$. Then we have

$$
\begin{aligned}
e(\boldsymbol{x}) &\leq 4|f|_{2,1} \sum_{j=1}^{m} \left(\frac{\left\|\boldsymbol{x} - \boldsymbol{x}_{B_j}\right\|_2^3}{9} + \frac{1}{3} \sum_{k=1}^{2} 12^{k-1} h_j^{3-k} C_j^k \left(\left\|\boldsymbol{x} - \boldsymbol{x}_{B_j}\right\|_2 + h_j \right)^k \right) \times \\
&\times \frac{\prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_2^{-\mu}}{\sum_{k=1}^{m} \prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{k_\ell}\|_2^{-\mu}} \\
&\leq 4|f|_{2,1} C' \sum_{j=1}^{m} \left(\frac{\left\|\boldsymbol{x} - \boldsymbol{x}_{B_j}\right\|_\infty^3}{9} + \frac{1}{3} \sum_{k=1}^{2} 12^{k-1} h_j^{3-k} C_j^k (\left\|\boldsymbol{x} - \boldsymbol{x}_{B_j}\right\|_\infty + h_j)^k \right) \times \\
&\times \frac{\prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty^{-\mu}}{\sum_{k=1}^{m} \prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{k_\ell}\|_\infty^{-\mu}},
\end{aligned}
$$

where $C' = \sqrt{2}^{3m\mu}$ is the constant that appears by bounding the Euclidean norm with the maximum norm.

We denote by $t_i \in \mathcal{T}$ the triangle satisfying

$$
\prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{i_\ell}\|_\infty = \min_{j=1\ldots,m} \prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty .
$$

Since $T_0 \neq \emptyset$, from the equation (5.31) we get

$$
\prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{i_\ell}\|_\infty \leq 9h^3.
$$

Consequently, if $t_j \in T_0$, then

$$
\prod_{\ell=1}^{3} \frac{\|\boldsymbol{x} - \boldsymbol{x}_{i_\ell}\|_\infty}{\|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty} \leq 1
$$

otherwise, if $t_j \in T_k$ with $k \neq 0$, then

$$
\prod_{\ell=1}^{3} \frac{\|\boldsymbol{x} - \boldsymbol{x}_{i_\ell}\|_\infty}{\|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty} \leq \frac{9h^3}{\left((2k-1)\,h\right)^3} = \frac{9}{(2k-1)^3}.
$$

From these inequalities we deduce that, if $t_j \in T_0$

$$
\frac{\prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty^{-\mu}}{\sum_{k=1}^{m} \prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{k_\ell}\|_\infty^{-\mu}} \leq \prod_{\ell=1}^{3} \frac{\|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty^{-\mu}}{\|\boldsymbol{x} - \boldsymbol{x}_{i_\ell}\|_\infty^{-\mu}} \leq 1,
$$

otherwise, if $t_j \in T_k$ with $k \neq 0$,

$$
\frac{\prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty^{-\mu}}{\sum_{k=1}^{m} \prod_{\ell=1}^{3} \|\boldsymbol{x} - \boldsymbol{x}_{k_\ell}\|_\infty^{-\mu}} \leq \prod_{\ell=1}^{3} \frac{\|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty^{-\mu}}{\|\boldsymbol{x} - \boldsymbol{x}_{i_\ell}\|_\infty^{-\mu}} \leq \frac{9^\mu}{(2k-1)^{3\mu}}.
$$

By definition of T_k, if $t_j \in T_k$ then $V(t_j) \subset U_{k-1} \cup U_k \cup U_{k+1}$ (here we assume that $U_{-1} = \emptyset$). Consequently

$$
\left\|\boldsymbol{x} - \boldsymbol{x}_{B_j}\right\|_\infty \leq \max_{\ell=1,2,3} \|\boldsymbol{x} - \boldsymbol{x}_{j_\ell}\|_\infty \leq (2k+3)h, \qquad k = 0, \ldots, N.
$$

Finally, by taking into account that $h_j \leq \sqrt{8}h < 3h$, we get

$$e(\boldsymbol{x}) \leq 4|f|_{2,1}C' \sum_{j=1}^{m} \left(\frac{\left\|\boldsymbol{x}-\boldsymbol{x}_{B_j}\right\|_\infty^3}{9} + \frac{1}{3}\sum_{k=1}^{2} 12^{k-1} h_j^{3-k} C_j^k \left(\left\|\boldsymbol{x}-\boldsymbol{x}_{B_j}\right\|_\infty + h_j\right)^k \right) \frac{\prod_{\ell=1}^{3} \|\boldsymbol{x}-\boldsymbol{x}_{j_\ell}\|_\infty^{-\mu}}{\sum_{k=1}^{m}\prod_{\ell=1}^{3} \|\boldsymbol{x}-\boldsymbol{x}_{k_\ell}\|_\infty^{-\mu}}$$

$$\leq |f|_{2,1}C' \left[\sum_{t_j \in T_0} \left(12h^3 + \frac{4}{3}h_j^2 C_j(3h+h_j) + 16h_j C_j^2 (3h+h_j)^2 \right) + \right.$$

$$\left. + \sum_{k=1}^{N} \sum_{t_j \in T_k} \left(\frac{4}{9}(2k+3)^3 h^3 + \frac{4}{3}h_j^2 C_j((2k+3)h+h_j) + 16h_j C_j^2((2k+3)h+h_j)^2 \right) \frac{9^\mu}{(2k-1)^{3\mu}} \right]$$

$$\leq |f|_{2,1}C' \left(\sum_{t_j \in T_0} \left(12 + 16C'' \left(\frac{1}{3} + 16C'' \right) \right) \right.$$

$$\left. + \sum_{k=1}^{N} \sum_{t_j \in T_k} \frac{\frac{4}{9}\cdot 9^\mu (2k+3)^3 + 4\cdot 9^\mu C''(2k+4)\left(\frac{1}{3} + 4C''(2k+4)\right)}{(2k-1)^{3\mu}} \right) h^3,$$

where $C'' = \max\limits_{j=1,\ldots,m} C_j$. By using (5.29), we get

$$e(\boldsymbol{x}) \leq |f|_{2,1} M C' \left(12 + 16C'' \left(\tfrac{1}{3} + 16C'' \right) \right.$$
$$\left. + 9^\mu \sum_{k=1}^{N} 8k \frac{\frac{4}{9}(2k+3)^3 + 4C''(2k+4)\left(\frac{1}{3} + 4C''(2k+4)\right)}{(2k-1)^{3\mu}} \right) h^3.$$

Since the series

$$\sum_{k=1}^{\infty} \frac{k(2k+4)}{(2k-1)^{3\mu}}, \quad \sum_{k=1}^{\infty} \frac{k(2k+4)^2}{(2k-1)^{3\mu}}, \quad \sum_{k=1}^{\infty} \frac{k(2k+3)^3}{(2k-1)^{3\mu}}$$

converge for $\mu > \frac{5}{3}$, then we can conclude that the approximation order of K_μ^{enr} is $\mathcal{O}(h^3)$. □

Remark 5.3.2. *In Theorem 5.3.1 we prove that the approximation order of the enriched triangular Shepard operator* (5.24) *is at least cubic. This result is in line with the general theorem on the approximation order of the multinode Shepard operator [28, Thm. 3.1].*

5.3.1 Numerical experiments

In this Section, we numerically test the accuracy of the enriched triangular Shepard method introduced in the previous section. To this aim, we perform several experiments by using the 10 test functions f_1 - f_{10} defined in [86] (see Figure 5.2) and two different Delaunay triangulations (see Figure 5.1). These triangulations are obtained through the Shewchuk's triangle program [91] by prescribing $N = 108$ and $N = 324$ uniformly distributed nodes on the boundary of the square $[0,1]^2$ and by constructing a conforming Delaunay triangulation with no angle smaller than 20° and no triangle area greater than $4\sqrt{3}/N^2$ by inserting Steiner points.

Numerical results, obtained by taking into account the efficient algorithm for the computation of triangular Shepard interpolation method [16], are reported in Tables 5.1 and 5.2. In these Tables we compare the approximation accuracy produced by the triangular Shepard operator with that produced by the enriched triangular Shepard operator by reporting the corresponding maximum approximation error e_{max}, mean approximation error e_{mean} and root mean square approximation error e_{MS} defined as

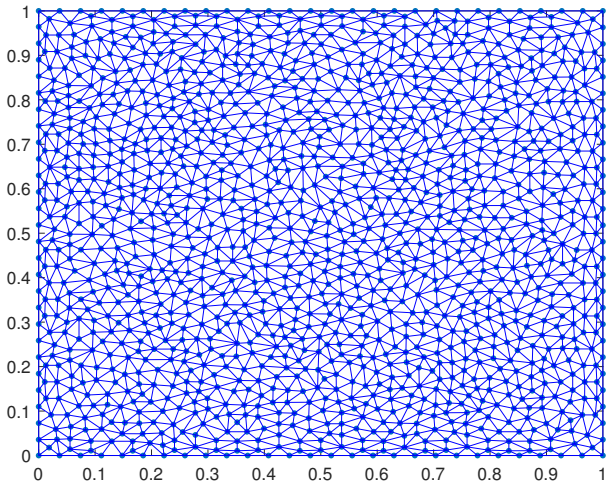


Figure 5.1: Conforming Delaunay triangulation with $N = 108$ uniformly distributed boundary nodes, no angle smaller than 20° and no triangle area greater than $4\sqrt{3}/N^2$.

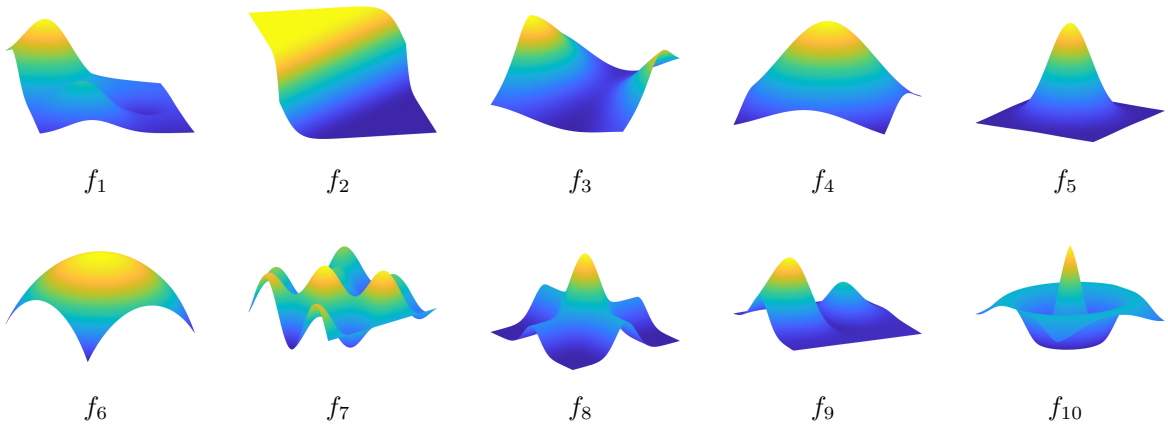


Figure 5.2: Test functions for numerical experiments defined in [86]

follows

$$e_{max} = \max_{i=1,\ldots,n_e} r_i, \quad e_{mean} = \frac{1}{n_e}\sum_{i=1}^{n_e} r_i, \quad e_{MS} = \sqrt{\frac{\sum_{i=1}^{n_e} r_i^2}{n_e}},$$

where r_i is the absolute approximation error at the $n_e = 100 \times 100$ points of a regular grid of $[0,1]^2$.

As we can note, the enriched triangular Shepard operator realizes a better approximation if compared with the classical triangular Shepard operator.

		Triangular Shepard operator	Enriched triangular Shepard operator
f_1	e_{max}	2.2464e-03	1.0588e-03
	e_{mean}	2.0062e-04	4.2208e-05
	e_{MS}	3.1644e-04	7.6905e-05
f_2	e_{max}	1.2415e-03	4.5919e-04
	e_{mean}	6.6836e-05	2.4912e-05
	e_{MS}	1.4454e-04	5.4291e-05
f_3	e_{max}	2.6579e-04	3.3678e-05
	e_{mean}	3.3458e-05	2.7390e-06
	e_{MS}	4.5864e-05	3.9321e-06
f_4	e_{max}	2.1272e-04	1.8476e-05
	e_{mean}	1.8259e-05	9.6884e-07
	e_{MS}	2.6987e-05	1.5773e-06
f_5	e_{max}	8.8847e-04	3.0575e-04
	e_{mean}	6.6021e-05	9.7233e-06
	e_{MS}	1.1416e-04	2.0691e-05
f_6	e_{max}	3.6121e-04	7.0155e-05
	e_{mean}	4.6343e-05	4.1339e-06
	e_{MS}	6.6921e-05	6.9302e-06
f_7	e_{max}	2.4980e-02	4.8345e-03
	e_{mean}	2.8090e-03	6.5907e-04
	e_{MS}	4.0365e-03	9.0295e-04
f_8	e_{max}	1.7054e-02	3.9667e-03
	e_{mean}	7.4217e-04	2.1144e-04
	e_{MS}	1.3314e-03	3.6111e-04
f_9	e_{max}	1.2282e+00	3.5011e-01
	e_{mean}	5.2157e-02	1.1609e-02
	e_{MS}	9.6110e-02	2.6151e-02
f_{10}	e_{max}	4.2243e-02	1.3485e-02
	e_{mean}	6.9346e-04	1.6602e-04
	e_{MS}	1.3407e-03	4.1463e-04

Table 5.1: Comparison between the triangular Shepard operator and the enriched triangular Shepard operator applied to the 10 test functions, see Figure 5.2, using a Delaunay triangulation of the interpolation nodes with $N = 108$ uniformly distributed nodes on the boundary of $[0, 1]^2$.

		Triangular Shepard operator	Enriched triangular Shepard operator
f_1	e_{max}	2.5273e-04	4.5717e-05
	e_{mean}	2.1024e-05	1.2243e-06
	e_{MS}	3.3547e-05	2.2525e-06
f_2	e_{max}	1.6270e-04	1.5714e-05
	e_{mean}	6.2165e-06	6.8689e-07
	e_{MS}	1.3851e-05	1.5326e-06
f_3	e_{max}	2.6081e-05	7.1061e-07
	e_{mean}	3.6096e-06	8.5854e-08
	e_{MS}	5.0213e-06	1.1894e-07
f_4	e_{max}	2.8911e-05	1.0709e-06
	e_{mean}	2.0016e-06	3.3043e-08
	e_{MS}	2.9416e-06	6.0762e-08
f_5	e_{max}	1.0407e-04	7.3289e-06
	e_{mean}	6.9815e-06	2.7880e-07
	e_{MS}	1.2153e-05	5.3943e-07
f_6	e_{max}	5.3319e-05	1.6370e-06
	e_{mean}	5.0698e-06	1.2403e-07
	e_{MS}	7.4439e-06	2.0099e-07
f_7	e_{max}	4.0309e-03	2.1767e-04
	e_{mean}	2.9033e-04	1.9395e-05
	e_{MS}	4.3809e-04	2.6318e-05
f_8	e_{max}	1.4253e-03	1.0759e-04
	e_{mean}	7.2410e-05	5.4961e-06
	e_{MS}	1.3607e-04	9.3920e-06
f_9	e_{max}	1.0972e-01	6.3777e-03
	e_{mean}	5.6108e-03	3.2246e-04
	e_{MS}	1.0418e-02	6.5976e-04
f_{10}	e_{max}	9.7463e-03	4.1772e-03
	e_{mean}	7.2077e-05	5.1653e-06
	e_{MS}	1.5993e-04	4.4371e-05

Table 5.2: Comparison between the triangular Shepard operator and the enriched triangular Shepard operator applied to the 10 test functions, see Figure 5.2, using a Delaunay triangulation of the interpolation nodes with $N = 324$ uniformly distributed nodes on the boundary of $[0, 1]^2$.

Chapter 6

A unified enrichment approach of the standard triangular linear finite element

The aim of this chapter is to unify the ideas and to extend to a more general setting the work done in Chapter 5 for a polynomial enrichment of the standard triangular linear finite element using line integrals and quadratic polynomials. More precisely, we introduce a new class of nonconforming finite elements by enriching the class of linear polynomial functions with additional functions which are not necessarily polynomials. We provide a simple condition on the enrichment functions, which is both necessary and sufficient, that guarantees the existence of a family of such enriched elements. Several sets of admissible enrichment functions that satisfy the admissibility condition are also provided, together with the explicit expression of the related approximation error. Our main result shows that the approximation error can be decomposed into two parts: the first one is related to the standard triangular linear finite element while the second one depends on the enrichment functions. This representation of the approximation error allows us to derive error bounds in both L^∞-norm and L^1-norm, with explicit constants, for continuously differentiable functions with Lipschitz continuous gradients. These bounds are proportional to the second and the fourth power of the circumcircle radius of the triangle, respectively. We also provide explicit expressions of these bounds in terms of the circumcircle diameter and the sum of squares of the triangle edge lengths. The result presented in this chapter can be found in [30].

6.1 The general problem for polynomial enrichment of the standard triangular linear finite element

Let $S_2 \subset \mathbb{R}^2$ be a triangle with nonzero signed area with vertices $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and barycentric coordinates λ_0, λ_1 λ_2. In the following, we denote by $\langle \cdot, \cdot \rangle$ the scalar product of the Euclidean space $\mathbb{R}^2$. By using the same notations of the previous chapter, we start by generalizing the polynomial enrichment of the standard triangular linear finite element $\mathcal{P}_1(S_2) = (S_2, \mathbb{P}_1(S_2), \Sigma_{S_2}^{\mathrm{lin}})$ described in Chapter 5 and based on the set of linear functionals

$$\Sigma_{S_2}^{\mathrm{enr}} = \{L_j, I_j \,:\, j = 0, 1, 2\}\,, \tag{6.1}$$

where

$$L_j(f) = f(\boldsymbol{v}_j), \qquad j = 0, 1, 2, \tag{6.2}$$

$$I_j(f) = \frac{1}{|\Gamma_j|} \int_{\Gamma_j} f(\boldsymbol{x}) d\sigma(\boldsymbol{x}), \qquad j = 0, 1, 2, \tag{6.3}$$

to the case of polynomials of degree greater than 2. To this aim, we consider three linear polynomials l_0, l_1, l_2, satisfying

$$l_0(\boldsymbol{v}_{01}) = l_1(\boldsymbol{v}_{12}) = l_2(\boldsymbol{v}_{02}) = 1, \tag{6.4}$$

where $\boldsymbol{v}_{01}, \boldsymbol{v}_{12}$ and $\boldsymbol{v}_{02}$ are the midpoints of the sides of S_2, that is

$$\boldsymbol{v}_{01} = \frac{\boldsymbol{v}_0 + \boldsymbol{v}_1}{2} \in \Gamma_2, \quad \boldsymbol{v}_{12} = \frac{\boldsymbol{v}_1 + \boldsymbol{v}_2}{2} \in \Gamma_0, \quad \boldsymbol{v}_{02} = \frac{\boldsymbol{v}_0 + \boldsymbol{v}_2}{2} \in \Gamma_1.$$

By using the same notations of the previous chapter, we introduce the enriched space $\mathbb{P}_1^{\text{enr}}(S_2)$ as follows

$$\mathbb{P}_1^{\text{enr}}(S_2) = \mathbb{P}_1(S_2) \oplus \text{span}\{l_0\lambda_1\lambda_2, l_1\lambda_0\lambda_2, l_2\lambda_0\lambda_1\} \tag{6.5}$$

and we consider the triple

$$AF3 = (S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}}). \tag{6.6}$$

The following technical lemma will be useful in establishing the proof of Theorem 6.1.3 about the unisolvence of the element $AF3$.

Lemma 6.1.1. *Let $p \in \mathbb{P}_1^{\text{enr}}(S_2)$. Then we have*

$$I_0(p) = \frac{1}{6}\left(p(\boldsymbol{v}_1) + 4p(\boldsymbol{v}_{12}) + p(\boldsymbol{v}_2)\right), \tag{6.7}$$

$$I_1(p) = \frac{1}{6}\left(p(\boldsymbol{v}_0) + 4p(\boldsymbol{v}_{02}) + p(\boldsymbol{v}_2)\right), \tag{6.8}$$

$$I_2(p) = \frac{1}{6}\left(p(\boldsymbol{v}_0) + 4p(\boldsymbol{v}_{01}) + p(\boldsymbol{v}_1)\right). \tag{6.9}$$

Proof. Since, Simpson's rule provides exact results for polynomials up to and including 3rd degree, the proof follows the same argument of Lemma 5.1.2. It is therefore omitted here. □

Remark 6.1.2. *Note that, if $p \in \mathbb{P}_1^{\text{enr}}(S_2)$ and $L_j(p) = I_j(p) = 0$, $j = 0, 1, 2$, then, by Lemma 6.1.1, we get*

$$p(\boldsymbol{v}_{12}) = p(\boldsymbol{v}_{02}) = p(\boldsymbol{v}_{01}) = 0, \tag{6.10}$$

that is, p vanishes at the midpoints of the sides of S_2.

Theorem 6.1.3. *The triple $AF3 = (S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}})$ is a finite element.*

Proof. We have to show that $\mathbb{P}_1^{\text{enr}}(S_2)$ is $\Sigma_{S_2}^{\text{enr}}$-unisolvent, i.e., if $p \in \mathbb{P}_1^{\text{enr}}(S_2)$ and

$$L_j(p) = 0, \qquad j = 0, 1, 2, \tag{6.11}$$

$$I_j(p) = 0, \qquad j = 0, 1, 2, \tag{6.12}$$

then $p = 0$ [23, Ch. 2]. Let $p \in \mathbb{P}_1^{\text{enr}}(S_2)$ and assume that (6.11) and (6.12) hold. By definition, p can be represented as

$$p = \alpha_0\lambda_0 + \alpha_1\lambda_1 + \alpha_2\lambda_2 + \beta_0 l_0\lambda_1\lambda_2 + \beta_1 l_1\lambda_0\lambda_2 + \beta_2 l_2\lambda_0\lambda_1 \tag{6.13}$$

for some constants $\alpha_i, \beta_i \in \mathbb{R}$, $i = 0, 1, 2$. Since the barycentric coordinates satisfy Lagrange property, that is $L_j(\lambda_i) = \lambda_i(\boldsymbol{v}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta symbol, then (6.11) implies $\alpha_0 = \alpha_1 = \alpha_2 = 0$ and p reduces to

$$p = \beta_0 l_0\lambda_1\lambda_2 + \beta_1 l_1\lambda_0\lambda_2 + \beta_2 l_2\lambda_0\lambda_1.$$

On the other hand, $I_0(p) = 0$ and $L_1(p) = L_2(p) = 0$ imply, by (6.7), $p(\boldsymbol{v}_{12}) = 0$. Now, since $\lambda_1(\boldsymbol{v}_{12}) = \lambda_2(\boldsymbol{v}_{12}) = \frac{1}{2}$ and $\lambda_0(\boldsymbol{v}_{12}) = 0$, we get

$$0 = \frac{4}{6}p(\boldsymbol{v}_{12}) = \frac{\beta_0}{6}l_0(\boldsymbol{v}_{12}) = \frac{\beta_0}{6},$$

where in the last equality, we used the conditions (6.4). Thus $\beta_0 = 0$. A similar argument can be used to show that $\beta_1 = \beta_2 = 0$, therefore $p = 0$ and the proof of the theorem is completed. □

Remark 6.1.4. *We notice that if we take $l_0 = l_1 = l_2 = 1$, then the enriched space $\mathbb{P}_1^{\text{enr}}(S_2)$ becomes the standard space of quadratic polynomials $\mathbb{P}_2(S_2)$ and, in this sense, the element $AF3$ generalizes the element $(S_2, \mathbb{P}_2(S_2), \Sigma_{S_2}^{\text{enr}})$ introduced in Chapter 5.*

Remark 6.1.5. *The enriched space $\mathbb{P}_1^{\text{enr}}(S_2)$ satisfies the following properties:*

i) it is a subspace of $\mathbb{P}_3(S_2)$, the space of cubic polynomials, and hence the restriction of any $p \in \mathbb{P}_1^{\text{enr}}(S_2)$ to each side of S_2 is a cubic polynomial in one variable;

ii) it contains the set of linear polynomials;

iii) the nonlinear terms in the expression (6.13) *of* $p \in \mathbb{P}_1^{\mathrm{enr}}(S_2)$ *vanish at the vertices of* S_2.

Theorem 6.1.3 can be stated equivalently by saying that the functionals of $\Sigma_{S_2}^{\mathrm{enr}}$ are linearly independent in the dual space $\mathbb{P}_1^{\mathrm{enr}}(S_2)^\star$. Then, there exists a basis $\{\varphi_j, \psi_j \,:\, j = 0,1,2\}$ of $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ which satisfies

$$L_j(\varphi_i) = \delta_{ij}, \quad I_j(\varphi_i) = 0, \qquad i,j = 0,1,2, \tag{6.14}$$

$$L_j(\psi_i) = 0, \quad I_j(\psi_i) = \delta_{ij}, \qquad i,j = 0,1,2. \tag{6.15}$$

Theorem 6.1.6. *The basis functions* $\{\varphi_j, \psi_j \,:\, j = 0,1,2\}$ *of* $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ *associated to the finite element* $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$, *which satisfy* (6.14) *and* (6.15) *have the following expressions*

$$\varphi_0 = \lambda_0(1 - 3l_2\lambda_1 - 3l_1\lambda_2), \quad \varphi_1 = \lambda_1(1 - 3l_2\lambda_0 - 3l_0\lambda_2), \quad \varphi_2 = \lambda_2(1 - 3l_1\lambda_0 - 3l_0\lambda_1), \tag{6.16}$$

$$\psi_0 = 6l_0\lambda_1\lambda_2, \quad \psi_1 = 6l_1\lambda_0\lambda_2, \quad \psi_2 = 6l_2\lambda_0\lambda_1. \tag{6.17}$$

Proof. Let us prove the first of identities (6.16). The element $\varphi_0 \in \mathbb{P}_1^{\mathrm{enr}}(S_2)$ can be represented as

$$\varphi_0 = \alpha_0\lambda_0 + \alpha_1\lambda_1 + \alpha_2\lambda_2 + \beta_0 l_0\lambda_1\lambda_2 + \beta_1 l_1\lambda_0\lambda_2 + \beta_2 l_2\lambda_0\lambda_1, \tag{6.18}$$

for some constants $\alpha_i, \beta_i \in \mathbb{R}$ $i = 0,1,2$. Since $L_0(\varphi_0) = 1$, $L_1(\varphi_0) = L_2(\varphi_0) = 0$, the Lagrange property of the barycentric coordinates implies $\alpha_0 = 1$, $\alpha_1 = \alpha_2 = 0$ and therefore

$$\varphi_0 = \lambda_0 + \beta_0 l_0\lambda_1\lambda_2 + \beta_1 l_1\lambda_0\lambda_2 + \beta_2 l_2\lambda_0\lambda_1.$$

Moreover, since $\lambda_0 = 0$ on Γ_0, from $I_0(\varphi_0) = 0$, by applying the Simpson's rule, we get

$$0 = \frac{1}{|\Gamma_0|}\int_{\Gamma_0} \beta_0 l_0(\boldsymbol{x})\lambda_1(\boldsymbol{x})\lambda_2(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \frac{4}{6}\beta_0 l_0(\boldsymbol{v}_{12})\lambda_1(\boldsymbol{v}_{12})\lambda_2(\boldsymbol{v}_{12}) = \frac{\beta_0}{6},$$

which readily gives $\beta_0 = 0$. Similarly, since $\lambda_1 = 0$ on Γ_1, from $I_1(\varphi_0) = 0$ we get

$$\begin{aligned} 0 &= \frac{1}{|\Gamma_1|}\int_{\Gamma_1} (\lambda_0(\boldsymbol{x}) + \beta_1 l_1(\boldsymbol{x})\lambda_0(\boldsymbol{x})\lambda_2(\boldsymbol{x}))d\sigma(\boldsymbol{x}) \\ &= \frac{1}{6}\left(1 + 4\left(\frac{1}{2} + \beta_1 l_1(\boldsymbol{v}_{02})\lambda_0(\boldsymbol{v}_{02})\lambda_2(\boldsymbol{v}_{02})\right)\right) \\ &= \frac{1}{6}(3 + \beta_1), \end{aligned}$$

that implies $\beta_1 = -3$. Finally, since $\lambda_2 = 0$ on Γ_2, from $I_2(\varphi_0) = 0$ we get $\beta_2 = -3$, and then the first of (6.16) is proved. We now show that the first of identities (6.17) holds. Since $L_i(\psi_0) = 0$, $i = 0,1,2$, the Lagrange property of the barycentric coordinates implies that $\psi_0 \in \operatorname{span}\{l_0\lambda_1\lambda_2, l_1\lambda_0\lambda_2, l_2\lambda_0\lambda_1\}$ and then it can be written as

$$\psi_0 = \beta_0 l_0\lambda_1\lambda_2 + \beta_1 l_1\lambda_0\lambda_2 + \beta_2 l_2\lambda_0\lambda_1,$$

for some constants β_i, $i = 0,1,2$. Moreover, from $I_0(\psi_0) = 1$ we get

$$1 = \frac{1}{|\Gamma_0|}\int_{\Gamma_0} \beta_0 l_0(\boldsymbol{x})\lambda_1(\boldsymbol{x})\lambda_2(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \frac{4}{6}\beta_0 l_0(\boldsymbol{v}_{12})\lambda_1(\boldsymbol{v}_{12})\lambda_2(\boldsymbol{v}_{12}) = \frac{\beta_0}{6},$$

hence $\beta_0 = 6$. From $I_1(\psi_0) = I_2(\psi_0) = 0$, we also get $\beta_1 = \beta_2 = 0$ and then the first of (6.17) is proved. We can obtain the expressions for the other functions by using symmetry arguments. □

Remark 6.1.7. *It is easily seen that conditions* (6.4), *which are sufficient for the existence of the enrichment* (6.6) *of the element* $(S_2, \mathbb{P}_1(S_2), \Sigma_{S_2})$, *can be replaced by the more general ones*

$$l_0(\boldsymbol{v}_{01}) \neq 0, \quad l_1(\boldsymbol{v}_{12}) \neq 0, \quad l_2(\boldsymbol{v}_{02}) \neq 0. \tag{6.19}$$

As it has become clear during the discussion, general conditions (6.19) *are also necessary.*

More generally, let us consider three interior points $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2$, of the sides $\Gamma_0, \Gamma_1, \Gamma_2$ of the triangle S_2, respectively, and three linear polynomials l_0, l_1, l_2. We call these points (and polynomials) admissible if they generate a unisolvent enriched element, possibly with different basis functions. In analogy with (6.19), we introduce the *nonvanishing conditions*

$$l_0(\boldsymbol{x}_0) \neq 0, \quad l_1(\boldsymbol{x}_1) \neq 0, \quad l_2(\boldsymbol{x}_2) \neq 0. \tag{6.20}$$

The question arises whether conditions (6.20) are sufficient (and necessary) to generate a unisolvent enriched element. As we will see below, the answer to previous question is positive, but the differentiability class of the enrichment functions will depend on the position of the point $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2$. To prove this result, we use appropriate Gauss quadrature rules on one point, instead of Simpson's rule, used in the case of the midpoints of the sides of S_2. The idea is rather simple and dates back to paper [57]. By assumption, there exist real numbers $\alpha_i, \beta_i > 1$, $i = 0, 1, 2$, such that

$$\begin{aligned} \boldsymbol{x}_0 &= \frac{\alpha_0}{\alpha_0+\beta_0}\boldsymbol{v}_1 + \frac{\beta_0}{\alpha_0+\beta_0}\boldsymbol{v}_2, \\ \boldsymbol{x}_1 &= \frac{\alpha_1}{\alpha_1+\beta_1}\boldsymbol{v}_0 + \frac{\beta_1}{\alpha_1+\beta_1}\boldsymbol{v}_2, \\ \boldsymbol{x}_2 &= \frac{\alpha_2}{\alpha_2+\beta_2}\boldsymbol{v}_0 + \frac{\beta_2}{\alpha_2+\beta_2}\boldsymbol{v}_1. \end{aligned} \tag{6.21}$$

For the sake of simplicity, we can assume that

$$l_0(\boldsymbol{x}_0) = l_1(\boldsymbol{x}_1) = l_2(\boldsymbol{x}_2) = 1. \tag{6.22}$$

We introduce the following more general enriched space

$$\mathbb{P}_1^{\mathrm{enr}}(S_2) = \mathbb{P}_1(S_2) \oplus \operatorname{span}\left\{ l_0\lambda_1^{\alpha_0-1}\lambda_2^{\beta_0-1}, l_1\lambda_0^{\alpha_1-1}\lambda_2^{\beta_1-1}, l_2\lambda_0^{\alpha_2-1}\lambda_1^{\beta_2-1} \right\}, \tag{6.23}$$

which includes the space (6.5) as particular case. In order to test whether or not $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ is $\Sigma_{S_2}^{\mathrm{enr}}$-unisolvent and to compute the basis of the enriched space, we recall the classical Euler beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt, \qquad \alpha, \beta > 0, \tag{6.24}$$

which satisfies the key property

$$B(\alpha+1, \beta) = \frac{\alpha}{\alpha+\beta}B(\alpha, \beta), \qquad \alpha, \beta > 0. \tag{6.25}$$

The beta function is connected to the gamma function through the equation [1]

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{6.26}$$

and hence for positive integers α and β, we have

$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}. \tag{6.27}$$

The following Proposition is crucial to prove that the triple $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$ is a finite element.

Proposition 6.1.8. *Under nonvanishing conditions* (6.22)*, the enrichment functions of* $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ *satisfy the following delta properties*

$$\begin{aligned} \frac{1}{|\Gamma_j|}\int_{\Gamma_j} l_0(\boldsymbol{x})\lambda_1^{\alpha_0-1}(\boldsymbol{x})\lambda_2^{\beta_0-1}(\boldsymbol{x})\, d\sigma(\boldsymbol{x}) &= \delta_{0j}B(\alpha_0, \beta_0), \\ \frac{1}{|\Gamma_j|}\int_{\Gamma_j} l_1(\boldsymbol{x})\lambda_0^{\alpha_1-1}(\boldsymbol{x})\lambda_2^{\beta_1-1}(\boldsymbol{x})\, d\sigma(\boldsymbol{x}) &= \delta_{1j}B(\alpha_1, \beta_1), \\ \frac{1}{|\Gamma_j|}\int_{\Gamma_j} l_2(\boldsymbol{x})\lambda_0^{\alpha_2-1}(\boldsymbol{x})\lambda_1^{\beta_2-1}(\boldsymbol{x})\, d\sigma(\boldsymbol{x}) &= \delta_{2j}B(\alpha_2, \beta_2), \end{aligned} \tag{6.28}$$

for each $j = 0, 1, 2$.

Proof. Let us prove the first of identities (6.28). Since $\alpha_0 > 1$ and $\beta_0 > 1$ then $\lambda_1^{\alpha_0-1}\lambda_2^{\beta_0-1}$ vanishes on Γ_1 and Γ_2, therefore the first of identities (6.28) holds for $j = 1, 2$. For $j = 0$, using the fact that λ_1 and λ_2 are affine functions, we get

$$\begin{aligned}\frac{1}{|\Gamma_0|}\int_{\Gamma_0} l_0(\boldsymbol{x})\lambda_1^{\alpha_0-1}(\boldsymbol{x})\lambda_2^{\beta_0-1}(\boldsymbol{x})\,d\sigma(\boldsymbol{x}) &= \int_0^1 \left(l_0\lambda_1^{\alpha_0-1}\lambda_2^{\beta_0-1}\right)(t\boldsymbol{v}_1+(1-t)\boldsymbol{v}_2)\,dt\\ &= \int_0^1 l_0(t\boldsymbol{v}_1+(1-t)\boldsymbol{v}_2)t^{\alpha_0-1}(1-t)^{\beta_0-1}\,dt.\end{aligned}$$

For the 1-point Gauss quadrature in the interval $[0, 1]$ associated with the weight function $w_{\alpha_0,\beta_0}(t) = t^{\alpha_0-1}(1-t)^{\beta_0-1}$, the node is located at the point $\frac{\alpha_0}{\alpha_0+\beta_0}$, while the corresponding weight is equal to $B(\alpha_0, \beta_0)$ [84, Sect. 3.1]. Indeed, it suffices to determine the orthogonal polynomial $q(t)$ of degree 1 relative to the weight function w_{α_0,β_0} on the interval $[0, 1]$ and using (6.24) and (6.25), we get

$$q(t) = t - \frac{\alpha_0}{\alpha_0+\beta_0}. \tag{6.29}$$

Since we assumed by (6.22) that $l_0(\boldsymbol{x}_0) = 1$, the result then follows from the exactness of the 1-point Gauss–Jacobi quadrature for polynomials of degree 1. □

The following theorem extends the Theorem 6.1.3 to the case of the general configuration of three points (6.21).

Theorem 6.1.9. *Let l_0, l_1, l_2 be linear polynomials satisfying the nonvanishing conditions* (6.22) *at the points $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2$. Then the triple $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$ is a finite element.*

Proof. As the dimension of $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ is equal to the cardinality of $\Sigma_{S_2}^{\mathrm{enr}}$, it suffices to show that $f \in \mathbb{P}_1^{\mathrm{enr}}(S_2)$ is identically zero if all the degrees of freedom (6.2) and (6.3) vanish when applied to f. The proof follows the same argument of Theorem 6.1.3 by using the general identities given in Proposition 6.1.8. It is therefore omitted here. □

Remark 6.1.10. *The nonvanishing conditions* (6.20) *are also necessary. Indeed, let us assume that conditions* (6.20) *do not hold, and without loss of generality, we can assume that $l_0(\boldsymbol{x}_0) = 0$. The function $e_0 = l_0\lambda_1^{\alpha_0-1}\lambda_2^{\beta_0-1} \in \mathbb{P}_1^{\mathrm{enr}}(S_2)$, satisfies $e_0(\boldsymbol{v}_i) = 0$, $i = 0, 1, 2$ and*

$$\frac{1}{|\Gamma_1|}\int_{\Gamma_1} e_0(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \frac{1}{|\Gamma_2|}\int_{\Gamma_2} e_0(\boldsymbol{x})d\sigma(\boldsymbol{x}) = 0,$$

since e_0 vanishes both on Γ_1 and Γ_2. We also have

$$\frac{1}{|\Gamma_0|}\int_{\Gamma_0} e_0(\boldsymbol{x})d\sigma(\boldsymbol{x}) = \int_0^1 t^{\alpha_0-1}(1-t)^{\beta_0-1}l_0(t\boldsymbol{v}_1+(1-t)\boldsymbol{v}_2)\,dt = B(\alpha_0,\beta_0)l_0(\boldsymbol{x}_0) = 0.$$

Then $L_j(e_0) = 0$ and $I_j(e_0) = 0$, $j = 0, 1, 2$, and therefore $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ is not $\Sigma_{S_2}^{\mathrm{enr}}$-unisolvent since $e_0 \neq 0$.

By assuming that the nonvanishing conditions (6.20) hold, we introduce the following notations

$$e_0 = l_0\lambda_1^{\alpha_0-1}\lambda_2^{\beta_0-1}, \quad e_1 = l_1\lambda_0^{\alpha_1-1}\lambda_2^{\beta_1-1}, \quad e_2 = l_2\lambda_0^{\alpha_2-1}\lambda_1^{\beta_2-1}, \tag{6.30}$$

$$\gamma_i = B(\alpha_i, \beta_i), \qquad i = 0, 1, 2.$$

Using the delta properties of the enriched terms given in Proposition 6.1.8, we can provide simple but elegant expressions for the functions $\{\varphi_j, \psi_j : j = 0, 1, 2\}$ of $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ associated to the finite element $AF3$, which satisfy (6.14) and (6.15). The proof of the following theorem follows the same argument of Theorem 6.1.6 and it is omitted.

Theorem 6.1.11. *The basis functions* $\{\varphi_j, \psi_j : j = 0,1,2\}$ *of* $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ *associated to the unisolvent element* $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$*, which satisfy* (6.14) *and* (6.15) *have the following expressions*

$$\varphi_j = \lambda_j - \frac{1}{2}\sum_{\substack{k=0\\k\neq j}}^{2} \psi_k, \quad j = 0,1,2, \tag{6.31}$$

$$\psi_j = \frac{e_j}{\gamma_j}, \qquad j = 0,1,2. \tag{6.32}$$

6.2 An explicit error representation

For the enriched space $\mathbb{P}_1^{\mathrm{enr}}(S_2)$, we introduce the approximation operator based on the finite element $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$

$$\begin{aligned} \Pi^{\mathrm{enr}} : C(S_2) &\to \mathbb{P}_1^{\mathrm{enr}}(S_2) \\ f &\mapsto \sum_{j=0}^{2} L_j(f)\varphi_j + \sum_{j=0}^{2} I_j(f)\psi_j, \end{aligned} \tag{6.33}$$

where φ_j, ψ_j, $j = 0,1,2$, are the basis functions defined in Theorem 6.1.11. We are interested in evaluating or estimating the approximation error

$$E^{\mathrm{enr}}[f] = f - \Pi^{\mathrm{enr}}[f]. \tag{6.34}$$

The following result shows that the error (6.34) can be decomposed in two parts: the first one is related to the standard triangular linear finite element $\mathcal{P}_1(S_2)$ while the second one depends on the enrichment functions e_i, $i = 0,1,2$. To short the notation, as we did in the previous chapter, in this representation we make use of a generalization of the classical trapezoidal formula to the case of line integrals. More precisely, for each $k = 0,1,2$, we set

$$\mathcal{L}_k = \frac{1}{2}\sum_{\substack{j=0\\j\neq k}}^{2} L_j \tag{6.35}$$

and

$$\mathcal{E}_k^{\mathrm{tra}} = \mathcal{L}_k - I_k, \tag{6.36}$$

where L_j and I_j, $j = 0,1,2$ are defined in (6.2) and (6.3), respectively.

Proposition 6.2.1. *Let us assume that the nonvanishing conditions* (6.22) *hold and let* e_i, $i = 0,1,2$, *be the enrichment functions defined in* (6.30)*. Then, for any* $f \in C(S_2)$*, the approximation error at any point* $\boldsymbol{x} \in S_2$ *is given by*

$$E^{\mathrm{enr}}[f](\boldsymbol{x}) = E^{\mathrm{lin}}[f](\boldsymbol{x}) + E^{\mathrm{tra}}[f](\boldsymbol{x}), \tag{6.37}$$

where

$$E^{\mathrm{lin}}[f](\boldsymbol{x}) = f(\boldsymbol{x}) - \sum_{j=0}^{2} L_j(f)\lambda_j(\boldsymbol{x}), \tag{6.38}$$

and

$$E^{\mathrm{tra}}[f](\boldsymbol{x}) = \sum_{k=0}^{2} \frac{e_k(\boldsymbol{x})}{\gamma_k}\mathcal{E}_k^{\mathrm{tra}}(f). \tag{6.39}$$

Proof. By (6.34) and (6.33)

$$E^{\mathrm{enr}}[f] = f - \sum_{j=0}^{2} L_j(f)\varphi_j - \sum_{j=0}^{2} I_j(f)\psi_j.$$

By (6.31) and by changing the order of the summation, we get

$$\begin{aligned}
\sum_{j=0}^{2} L_j(f)\varphi_j &= \sum_{j=0}^{2} L_j(f)\left(\lambda_j - \sum_{k=0}^{2} \frac{e_k}{2\gamma_k}\left(1-\delta_{jk}\right)\right) \\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} \frac{e_k}{2\gamma_k} \sum_{j=0}^{2} \left(1-\delta_{jk}\right) L_j(f) \\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} \frac{e_k}{2\gamma_k} \sum_{\substack{j=0 \\ j\neq k}}^{2} L_j(f) \\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} \frac{e_k}{\gamma_k} \mathcal{L}_k(f).
\end{aligned}$$

Therefore, for all $\boldsymbol{x} \in S_2$, we get

$$\begin{aligned}
E^{\mathrm{enr}}[f](\boldsymbol{x}) &= f(\boldsymbol{x}) - \sum_{j=0}^{2} L_j(f)\varphi_j(\boldsymbol{x}) - \sum_{j=0}^{2} I_j(f)\psi_j(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) - \sum_{j=0}^{2} L_j(f)\lambda_j(\boldsymbol{x}) + \sum_{k=0}^{2} \frac{e_k(\boldsymbol{x})}{\gamma_k} \mathcal{L}_k(f) - \sum_{j=0}^{2} I_j(f)\psi_j(\boldsymbol{x}) \\
&= E^{\mathrm{lin}}[f](\boldsymbol{x}) + \sum_{k=0}^{2} \frac{e_k(\boldsymbol{x})}{\gamma_k} \left(\mathcal{L}_k(f) - I_k(f)\right),
\end{aligned}$$

as desired. □

6.3 Nonpolynomial enrichment of the standard triangular linear finite element

In this Section, we introduce a more general enrichment of the standard triangular linear finite element $\mathcal{P}_1(S_2)$ based on three linearly independent continuous enrichment functions e_0, e_1, e_2. As before, we assume that these functions satisfy the vanishing conditions at the vertices

$$e_0(\boldsymbol{v}_i) = e_1(\boldsymbol{v}_i) = e_2(\boldsymbol{v}_i) = 0, \qquad i = 0, 1, 2. \tag{6.40}$$

We consider the triple

$$GF3 = (S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}}), \tag{6.41}$$

where

$$\mathbb{P}_1^{\mathrm{enr}}(S_2) = \mathbb{P}_1(S_2) \oplus \mathrm{span}\left\{e_0, e_1, e_2\right\}, \tag{6.42}$$

and $\Sigma_{S_2}^{\mathrm{enr}}$ is defined as in (6.1). The motivation for the introduction of this new enrichment lies in the possibility to capture, through the new enrichment functions, features of the function to be approximated that cannot be accurately captured by previously considered basis. It is worthwhile to note that, in the present situation, we cannot apply previously developed approaches, based on the use of Simpson's rule or Gauss quadrature rule on one point. The following theorem gives necessary and sufficient conditions on the enrichment functions e_0, e_1, e_2, such that the triple $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$ is a finite element or, equivalently, so that $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ is $\Sigma_{S_2}^{\mathrm{enr}}$-unisolvent.

Theorem 6.3.1. *Let*

$$N = \begin{bmatrix} I_0(e_0) & I_0(e_1) & I_0(e_2) \\ I_1(e_0) & I_1(e_1) & I_1(e_2) \\ I_2(e_0) & I_2(e_1) & I_2(e_2) \end{bmatrix}, \tag{6.43}$$

then the triple $(S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}})$ *is a finite element if and only if*

$$\det(N) \neq 0. \tag{6.44}$$

Proof. Let us assume that $\det(N) \neq 0$ and we prove that $\mathbb{P}_1^{\text{enr}}(S_2)$ is $\Sigma_{S_2}^{\text{enr}}$-unisolvent. Let $f \in \mathbb{P}_1^{\text{enr}}(S_2)$ be a function satisfying

$$L_j(f) = f(\boldsymbol{v}_j) = 0, \qquad j = 0, 1, 2, \tag{6.45}$$

$$I_j(f) = \frac{1}{|\Gamma_j|} \int_{\Gamma_j} f(\boldsymbol{x}) d\sigma(\boldsymbol{x}) = 0, \qquad j = 0, 1, 2. \tag{6.46}$$

By (6.42), f can be decomposed into the sum of a linear polynomial $p \in \mathbb{P}_1(S_2)$ and an enriched part, that is

$$f = p + \beta_0 e_0 + \beta_1 e_1 + \beta_2 e_2, \qquad \beta_i \in \mathbb{R}, \qquad i = 0, 1, 2.$$

The equations (6.45), by the vanishing conditions (6.40) of the enrichment functions, imply that

$$p(\boldsymbol{v}_i) = 0, \qquad i = 0, 1, 2.$$

Therefore, since p is linear, $p = 0$ and f coincides with its enriched part

$$f = \beta_0 e_0 + \beta_1 e_1 + \beta_2 e_2.$$

By using the linearity of the functionals I_j, $j = 0, 1, 2$, equations (6.46) can be represented in matrix form as

$$\begin{bmatrix} I_0(e_0) & I_0(e_1) & I_0(e_2) \\ I_1(e_0) & I_1(e_1) & I_1(e_2) \\ I_2(e_0) & I_2(e_1) & I_2(e_2) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.47}$$

Since the determinant of N is nonzero, this system has the unique solution $\beta_0 = \beta_1 = \beta_2 = 0$. Hence $f = 0$.
In order to prove the reverse implication, let us assume that $\det(N) = 0$ and we prove that $\mathbb{P}_1^{\text{enr}}(S_2)$ is not $\Sigma_{S_2}^{\text{enr}}$-unisolvent. Since $\det(N) = 0$, there exist three real numbers $\gamma_0, \gamma_1, \gamma_2$, not all zero, for which the function

$$e = \sum_{i=0}^{2} \gamma_i e_i$$

satisfies

$$I_j(e) = 0, \qquad j = 0, 1, 2.$$

Moreover the vanishing conditions (6.40) imply that

$$L_j(e) = 0, \qquad j = 0, 1, 2,$$

therefore, we can exhibit a linear combination of the basis functions of $\mathbb{P}_1^{\text{enr}}(S_2)$ with coefficients not all zero in which all degrees of freedom vanish. Then $\mathbb{P}_1^{\text{enr}}(S_2)$ is not $\Sigma_{S_2}^{\text{enr}}$-unisolvent. □

Definition 6.3.2. *Let* e_0, e_1, e_2 *be linearly independent continuous enrichment functions satisfying the vanishing conditions* (6.40). *They are said admissible enrichment functions if we can enrich* $\mathcal{P}_1(S_2)$ *to the finite element* $(S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}})$.

There exists a large class of admissible enrichment functions, as the following example shows.

Example 6.3.3. *Let us consider the three functions*

$$\begin{aligned} e_0 &= (1-\lambda_0)^{\alpha_0 - 1} \lambda_1^{\beta_0 - 1} \lambda_2^{\gamma_0 - 1}, \\ e_1 &= (1-\lambda_1)^{\alpha_1 - 1} \lambda_0^{\beta_1 - 1} \lambda_2^{\gamma_1 - 1}, \\ e_2 &= (1-\lambda_2)^{\alpha_2 - 1} \lambda_0^{\beta_2 - 1} \lambda_1^{\gamma_2 - 1}, \end{aligned} \tag{6.48}$$

with $\alpha_i, \beta_i, \gamma_i > 1$, $i = 0, 1, 2$. *It is easy to see that* e_0, e_1, e_2 *satisfy conditions* (6.40). *Moreover, by using* (6.28) *we get*

$$I_j(e_i) = \frac{1}{|\Gamma_j|} \int_{\Gamma_j} e_i(\boldsymbol{x})\, d\sigma(\boldsymbol{x}) = \delta_{ij} B(\beta_i, \gamma_i), \qquad i, j = 0, 1, 2 \tag{6.49}$$

and then the matrix N *defined in* (6.43) *is a diagonal matrix with determinant different from zero. This imply that the functions* e_i, $i = 0, 1, 2$, *are linearly independent and therefore, by Theorem 6.3.1 we can enrich* $\mathcal{P}_1(S_2)$ *to the unisolvent element* $(S_2, \mathbb{P}_1^{\mathrm{enr}}(S_2), \Sigma_{S_2}^{\mathrm{enr}})$ *by using* (6.48) *as enrichment functions.*

In the following, we assume that the matrix N is nonsingular and we denote its inverse by

$$N^{-1} = [\boldsymbol{c}_0\, \boldsymbol{c}_1\, \boldsymbol{c}_2], \tag{6.50}$$

where $\boldsymbol{c}_i \in \mathbb{R}^3$, $i = 0, 1, 2$, are column vectors. A direct consequence of Theorem 6.3.1 is the linear independence of the functionals of $\Sigma_{S_2}^{\mathrm{enr}}$ in the dual space $\mathbb{P}_1^{\mathrm{enr}}(S_2)^\star$ [23, Ch 2]. Then, there exists a basis $\{\varphi_j, \psi_j : j = 0, 1, 2\}$ of $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ associated to the finite element $GF3$, which satisfy (6.14) and (6.15).

Theorem 6.3.4. *The basis functions* $\{\varphi_j, \psi_j : j = 0, 1, 2\}$ *of* $\mathbb{P}_1^{\mathrm{enr}}(S_2)$ *associated to the finite element* $GF3$, *which satisfy* (6.14) *and* (6.15) *have the following expressions*

$$\varphi_j = \lambda_j - \frac{1}{2} \sum_{\substack{k=0 \\ k \neq j}}^{2} \psi_k, \qquad j = 0, 1, 2, \tag{6.51}$$

$$\psi_j = \langle \boldsymbol{e}, \boldsymbol{c}_j \rangle\,, \qquad j = 0, 1, 2, \tag{6.52}$$

where

$$\boldsymbol{e} = [e_0,\, e_1,\, e_2]^T. \tag{6.53}$$

Proof. Without loss of generality, we prove (6.51) for the case $j = 0$. We set

$$I_j(\boldsymbol{e}) = [I_j(e_0), I_j(e_1), I_j(e_2)]^T, \qquad j = 0, 1, 2$$

hence, since $NN^{-1} = I$, we easily get

$$\langle I_j(\boldsymbol{e}), \boldsymbol{c}_i \rangle = \delta_{ij}. \tag{6.54}$$

Moreover, by Lemma 6.1.1, we get

$$I_j(\lambda_i) = \frac{1}{2}(1 - \delta_{ij}). \tag{6.55}$$

As an element of $\mathbb{P}_1^{\mathrm{enr}}(S_2)$, φ_0 can be represented as

$$\varphi_0 = p + \beta_0 e_0 + \beta_1 e_1 + \beta_2 e_2 = p + \langle \boldsymbol{e}, \boldsymbol{\beta} \rangle\,, \tag{6.56}$$

where $p \in \mathbb{P}_1(S_2)$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T \in \mathbb{R}^3$. By using (6.14) and the vanishing conditions (6.40) we have

$$\lambda_0(\boldsymbol{v}_j) = \delta_{0j} = L_j(\varphi_0) = p(\boldsymbol{v}_j), \qquad j = 0, 1, 2,$$

so that $p = \lambda_0$, since they are linear polynomials. Therefore (6.56) becomes

$$\varphi_0 = \lambda_0 + \langle \boldsymbol{e}, \boldsymbol{\beta} \rangle\,, \tag{6.57}$$

and by applying I_j, $j = 0, 1, 2$, to both members of (6.57), by (6.14) and (6.55), we get

$$0 = \frac{1}{2}(1 - \delta_{0j}) + \langle I_j(\boldsymbol{e}), \boldsymbol{\beta} \rangle\,, \qquad j = 0, 1, 2,$$

or, in matrix form,

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + N\boldsymbol{\beta}.$$

Consequently

$$\boldsymbol{\beta} = -\frac{1}{2} N^{-1} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = -\frac{1}{2} (\boldsymbol{c}_1 + \boldsymbol{c}_2) ,$$

which, substituted in (6.57), gives the following expression for φ_0

$$\varphi_0 = \lambda_0 - \frac{1}{2} \sum_{k=1}^{2} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle . \tag{6.58}$$

In order to prove (6.51) for $j = 0$, it remains to prove (6.52). To this aim, without loss of generality, we show the validity of (6.52) for $j = 0$. We proceed in analogy to the previous case and then we set

$$\psi_0 = q + \langle \boldsymbol{e}, \boldsymbol{\gamma} \rangle ,$$

where $q \in \mathbb{P}_1(S_2)$ and $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3]^T \in \mathbb{R}^3$. Since $\psi_0(\boldsymbol{v}_j) = 0$ for $j = 0, 1, 2$, this function can be expressed as

$$\psi_0 = \langle \boldsymbol{e}, \boldsymbol{\gamma} \rangle . \tag{6.59}$$

By applying I_j, $j = 0, 1, 2$, to both members of (6.59), by (6.15), we get

$$\delta_{j0} = \langle I_j(\boldsymbol{e}), \boldsymbol{\gamma} \rangle , \qquad j = 0, 1, 2,$$

or, in matrix form,

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = N\boldsymbol{\gamma}.$$

Consequently

$$\boldsymbol{\gamma} = N^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \boldsymbol{c}_0,$$

which, substituted in (6.59), gives the required expression (6.52) for ψ_0. Similarly, we can prove (6.52) for $j = 1, 2$ and consequently (6.51) for $j = 0$ is proved. The expression of the other functions can be obtained using symmetry arguments. □

6.4 Error estimates

6.4.1 An explicit error representation

In analogy to the case of polynomial enrichment described in Section 6.1, we are interested in evaluating or estimating the error

$$E^{\mathrm{enr}}[f] = f - \Pi^{\mathrm{enr}}[f] \tag{6.60}$$

of the approximation operator based on the finite element $GF3$ (6.41)

$$\Pi^{\mathrm{enr}}[f] = \sum_{j=0}^{2} L_j(f)\varphi_j + \sum_{j=0}^{2} I_j(f)\psi_j, \tag{6.61}$$

where the basis functions φ_j, ψ_j, $j = 0, 1, 2$ are now given as in (6.51) and (6.52). As before, we start by proving a decomposition of the error $E^{\mathrm{enr}}[f]$ as a sum of the error of the standard linear triangular element plus an additional term which depends both on the enrichment functions e_i, $i = 0, 1, 2$ and the error (6.36) of the generalization of the classical trapezoidal formula to the case of line integrals (6.35). This representation will play an important role in the derivation of explicit bounds for the error in L^1-norm.

Proposition 6.4.1. *Let $e_i \in C(S_2)$, $i = 0, 1, 2$ be admissible enrichment functions. Then, for any $f \in C(S_2)$, the approximation error $E^{\mathrm{enr}}[f]$ at any point $\boldsymbol{x} \in S_2$ is given by*

$$E^{\mathrm{enr}}[f](\boldsymbol{x}) = E^{\mathrm{lin}}[f](\boldsymbol{x}) + \sum_{k=0}^{2} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle \, \mathcal{E}_k^{\mathrm{tra}}(f), \tag{6.62}$$

where $E^{\mathrm{lin}}[f](\boldsymbol{x})$, $\mathcal{E}_k^{\mathrm{tra}}(f)$, $\boldsymbol{e}(\boldsymbol{x})$ and $\boldsymbol{c}_k$, $k = 0, 1, 2$ are defined as in (6.38), (6.36), (6.53) *and* (6.50), *respectively.*

Proof. By (6.60) and (6.61)

$$E^{\mathrm{enr}}[f] = f - \sum_{j=0}^{2} L_j(f)\varphi_j - \sum_{j=0}^{2} I_j(f)\psi_j.$$

By (6.51) and by changing the order of the summation, we get

$$\begin{aligned}
\sum_{j=0}^{2} L_j(f)\varphi_j &= \sum_{j=0}^{2} L_j(f) \left(\lambda_j - \frac{1}{2} \sum_{\substack{k=0 \\ k \neq j}}^{2} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \right) \\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{j=0}^{2} L_j(f) \frac{1}{2} \sum_{k=0}^{2} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \, (1 - \delta_{jk}) \\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \, \frac{1}{2} \sum_{j=0}^{2} L_j(f)(1 - \delta_{jk}) \\
&= \sum_{j=0}^{2} L_j(f)\lambda_j - \sum_{k=0}^{2} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \, \frac{1}{2} \sum_{\substack{j=0 \\ j \neq k}}^{2} L_j(f).
\end{aligned}$$

Consequently, for each $\boldsymbol{x} \in S_2$, we get

$$\begin{aligned}
E^{\mathrm{enr}}[f](\boldsymbol{x}) &= f(\boldsymbol{x}) - \sum_{j=0}^{2} L_j(f)\varphi_i(\boldsymbol{x}) - \sum_{j=0}^{2} I_j(f)\psi_j(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) - \sum_{j=0}^{2} L_j(f)\lambda_j(\boldsymbol{x}) - \sum_{k=0}^{2} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle \, \frac{1}{2} \sum_{\substack{j=0 \\ j \neq k}}^{2} L_j(f) - \sum_{j=0}^{2} I_j(f) \, \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_j \rangle \\
&= E^{\mathrm{lin}}[f](\boldsymbol{x}) + \sum_{k=0}^{2} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle \left(\frac{1}{2} \sum_{\substack{j=0 \\ j \neq k}}^{2} L_j(f) - I_k(f) \right) \\
&= E^{\mathrm{lin}}[f](\boldsymbol{x}) + \sum_{k=0}^{2} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle \, (\mathcal{L}_k(f) - I_k(f)) \,,
\end{aligned} \tag{6.63}$$

as required. □

Remark 6.4.2. *It may be interesting to compare the proposed approximation operator Π^{enr} and the interpolation operator based on the standard triangular linear finite element $\mathcal{P}_1(S_2)$, defined in* (5.3)

$$\Pi^{\mathrm{lin}}[f](\boldsymbol{x}) = \sum_{j=0}^{2} L_j(f)\lambda_j(\boldsymbol{x}). \tag{6.64}$$

Using the representation of the error (6.63)*, the operator* Π^{enr} *can be formulated in an subtractive form more convenient for practical computation*

$$\Pi^{\text{enr}}[f](\boldsymbol{x}) = \Pi^{\text{lin}}[f](\boldsymbol{x}) - E^{\text{tra}}[f](\boldsymbol{x}), \tag{6.65}$$

where

$$E^{\text{tra}}[f](\boldsymbol{x}) = \sum_{k=0}^{2} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle \left(\mathcal{L}_k(f) - I_k(f) \right).$$

Indeed, as shown in equation (6.65)*, the operator* Π^{enr} *may be computed by simply subtracting the approximation operator* E^{tra} *from the operator* Π^{lin}*, so that the two contributions can be evaluated separately.*

6.4.2 Error bounds

The decomposition (6.62) is the key result to get the estimate of the error $E^{\text{enr}}[f]$ in the case of a particular class of functions with continuous gradient. As usually, we say that f is continuously differentiable on S_2 if it is continuously differentiable on an open set containing S_2. Other useful terminology and notations are clarified in the following.

Definition 6.4.3. *A differentiable function* f *is said to have a Lipschitz continuous gradient on* S_2*, if there exists a constant* $\rho > 0$ *such that*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \rho \|\boldsymbol{x} - \boldsymbol{y}\|_2, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in S_2, \tag{6.66}$$

where $\|\cdot\|_2$ *is the* L^2*-norm in* $\mathbb{R}^2$*.*

By $C^{1,1}(S_2)$ we denote the subclass of all functions f which are continuously differentiable with Lipschitz continuous gradient on S_2. We call the smallest possible ρ such that (6.66) holds *Lipschitz constant* for ∇f and we denote it by $L(\nabla f)$.

The following result (see [53, Thm. 2.3]) will be useful in the following.

Theorem 6.4.4. *Let* $A : C^1(S_2) \to C(S_2)$ *be a linear operator. The following statements are equivalent:*

(i) for any convex function $g \in C^1(S_2)$*, we have*

$$g(\boldsymbol{x}) \leq A[g](\boldsymbol{x}), \qquad \boldsymbol{x} \in S_2; \tag{6.67}$$

(ii) for any $f \in C^{1,1}(S_2)$*, we have*

$$|f(\boldsymbol{x}) - A[f](\boldsymbol{x})| \leq \frac{L(\nabla f)}{2} \left(A\left[\|\cdot\|_2^2 \right](\boldsymbol{x}) - \|\boldsymbol{x}\|_2^2 \right), \qquad \boldsymbol{x} \in S_2. \tag{6.68}$$

Equality is attained for all functions of the form

$$f(\boldsymbol{x}) = a(\boldsymbol{x}) + c \|\boldsymbol{x}\|_2^2,$$

where $c \in \mathbb{R}$ *and* $a(\boldsymbol{x})$ *is any affine function.*

Remark 6.4.5. *We notice that the results of Theorem 6.4.4 hold true, with the needed changes, in the case of standard simplex in* $\mathbb{R}^d$*,* $d \in \mathbb{N}$ *(see [53, Thm. 2.3]).*

Since each $\boldsymbol{x} \in S_2$ can be expressed as $\boldsymbol{x} = \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \boldsymbol{v}_i$, then, for any convex function f on S_2, we have

$$f(\boldsymbol{x}) \leq \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) f(\boldsymbol{v}_i) =: \Pi^{\text{lin}}[f](\boldsymbol{x}), \qquad \boldsymbol{x} \in S_2, \tag{6.69}$$

that is, the linear interpolation operator Π^{lin} based on the standard triangular linear finite element $\mathcal{P}_1(S_2)$, defined in (5.3) satisfies condition (6.67), i.e. it approximates every convex function from above [52, 54]. Then, the following result holds.

Theorem 6.4.6. *For any $f \in C^{1,1}(S_2)$, we have*

$$\left|E^{\text{lin}}[f](\boldsymbol{x})\right| = \left|f - \Pi^{\text{lin}}[f]\right| \leq \frac{L(\nabla f)}{2} \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x} - \boldsymbol{v}_i\right\|_2^2, \qquad \boldsymbol{x} \in S_2.$$

Equality is attained for all functions of the form

$$f(\boldsymbol{x}) = a(\boldsymbol{x}) + c\left\|\boldsymbol{x}\right\|_2^2,$$

where $c \in \mathbb{R}$ and $a(\boldsymbol{x})$ is any affine function.

Proof. From (6.69) we can apply Theorem 6.4.4 to the linear operator Π^{lin} so that (6.68) becomes

$$\left|f(\boldsymbol{x}) - \Pi^{\text{lin}}[f](\boldsymbol{x})\right| \leq \frac{L(\nabla f)}{2} \left(\sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{x}\right\|_2^2\right), \qquad \boldsymbol{x} \in S_2.$$

It remains to show that

$$\sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{x}\right\|_2^2 = \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x} - \boldsymbol{v}_i\right\|_2^2.$$

Indeed, we have

$$\left\|\boldsymbol{x} - \boldsymbol{v}_i\right\|_2^2 = \left\|\boldsymbol{x}\right\|_2^2 - 2\left\langle \boldsymbol{x}, \boldsymbol{v}_i \right\rangle + \left\|\boldsymbol{v}_i\right\|_2^2, \qquad i = 0, 1, 2.$$

By multiplying each of the above equalities by $\lambda_i(\boldsymbol{x})$ and summing over all $i = 0, 1, 2$, we immediately get

$$\sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x} - \boldsymbol{v}_i\right\|_2^2 = \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x}\right\|_2^2 - 2\left\langle \boldsymbol{x}, \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \boldsymbol{v}_i \right\rangle + \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{v}_i\right\|_2^2.$$

The desired result now follows from the partition of unity and the linear precision properties of the barycentric coordinates. The last statement of the theorem follows directly by Theorem 6.4.4. □

With reference to the formula (6.62), it remains to bound the error $\mathcal{E}_k^{\text{tra}}(f)$ defined in (6.36).

Theorem 6.4.7. *For any $f \in C^{1,1}(S_2)$, we have*

$$\left|\mathcal{L}_k(f) - \frac{1}{|\Gamma_k|} \int_{\Gamma_k} f(\boldsymbol{x}) d\sigma(\boldsymbol{x})\right| \leq \frac{L(\nabla f)}{12} \left|\Gamma_k\right|^2, \qquad k = 0, 1, 2. \tag{6.70}$$

Equality in (6.70) *is attained for all functions of the form*

$$f(\boldsymbol{x}) = a(\boldsymbol{x}) + c\left\|\boldsymbol{x}\right\|_2^2, \tag{6.71}$$

where $c \in \mathbb{R}$ and $a(\boldsymbol{x})$ is any affine function.

Proof. We prove (6.70) in the particular case $k = 0$ since the remaining cases can be proved by analogy. Let us denote by $\widetilde{f}$ the map

$$\widetilde{f}(t) = f((1-t)\boldsymbol{v}_1 + t\boldsymbol{v}_2), \qquad t \in [0, 1] \tag{6.72}$$

and by

$$\widetilde{\Pi}^{\text{lin}}[\widetilde{f}](t) = (1-t)\widetilde{f}(0) + t\widetilde{f}(1) \tag{6.73}$$

its linear interpolant at the end points of the interval $[0, 1]$. Therefore, we have

$$\begin{aligned}
\mathcal{E}_0^{\text{tra}}(f) &= \frac{1}{2}\left(f(\boldsymbol{v}_1) + f(\boldsymbol{v}_2)\right) - \frac{1}{|\Gamma_0|} \int_{\Gamma_0} f(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \\
&= \frac{1}{2}\left(\widetilde{f}(0) + \widetilde{f}(1)\right) - \int_0^1 \widetilde{f}(t)\, dt \\
&= \int_0^1 \left(\widetilde{\Pi}^{\text{lin}}[\widetilde{f}](t) - \widetilde{f}(t)\right) dt.
\end{aligned}$$

Consequently

$$\left|\mathcal{E}_0^{\mathrm{tra}}(f)\right| \leq \int_0^1 \left|\widetilde{\Pi}^{\mathrm{lin}}[\widetilde{f}](t) - \widetilde{f}(t)\right| dt. \tag{6.74}$$

Since $\widetilde{\Pi}^{\mathrm{lin}}$ approximates from above any convex function on $[0,1]$, from Remark 6.4.5 we have

$$\left|\widetilde{\Pi}^{\mathrm{lin}}[\widetilde{f}](t) - \widetilde{f}(t)\right| \leq \frac{L(\widetilde{f}')}{2}\left(\widetilde{\Pi}^{\mathrm{lin}}[|\cdot|^2](t) - |t|^2\right) = \frac{L(\widetilde{f}')}{2}\left(t - t^2\right),$$

and therefore, from (6.74) we get

$$\left|\mathcal{E}_0^{\mathrm{tra}}(f)\right| \leq \frac{L(\widetilde{f}')}{12}. \tag{6.75}$$

Moreover, for each $s,t \in [0,1]$, we have

$$\begin{aligned}
\left|\widetilde{f}'(s) - \widetilde{f}'(t)\right| &= \left|\langle \nabla f(s\boldsymbol{v}_1 + (1-s)\boldsymbol{v}_2) - \nabla f(t\boldsymbol{v}_1 + (1-t)\boldsymbol{v}_2), \boldsymbol{v}_1 - \boldsymbol{v}_2\rangle\right| \\
&\leq \left\|\nabla f(s\boldsymbol{v}_1 + (1-s)\boldsymbol{v}_2) - \nabla f(t\boldsymbol{v}_1 + (1-t)\boldsymbol{v}_2)\right\|_2 \left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2 \\
&\leq L(\nabla f)\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2 |s-t|,
\end{aligned} \tag{6.76}$$

and then, by definition of Lipschitz constant, we get

$$L(\widetilde{f}') \leq L(\nabla f)\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2 = L(\nabla f)|\Gamma_0|^2 \tag{6.77}$$

which concludes the proof of the inequality (6.70). Let assume that the function f has the form (6.71) with $c = 0$. Therefore the function $\widetilde{f}$ defined in (6.72) is affine and then $\widetilde{\Pi}^{\mathrm{lin}}[\widetilde{f}] = \widetilde{f}$ and $L(\widetilde{f}') = 0$. Consequently, (6.70) holds with equality in this case. Now let assume that $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2$. In this case the function in (6.72)

$$\widetilde{f}(t) = \left\|(1-t)\boldsymbol{v}_1 + t\boldsymbol{v}_2\right\|_2^2, \qquad t \in [0,1],$$

is a univariate quadratic polynomial satisfying

$$\widetilde{f}''(t) = 2\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2. \tag{6.78}$$

We expand $\widetilde{f}(0)$ and $\widetilde{f}(1)$ in Taylor series centered in t and we get, from (6.78),

$$\widetilde{f}(0) = \widetilde{f}(t) - t\widetilde{f}'(t) + t^2\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2, \tag{6.79}$$

$$\widetilde{f}(1) = \widetilde{f}(t) + (1-t)\widetilde{f}'(t) + (1-t)^2\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2. \tag{6.80}$$

Multiplying (6.79) by $1-t$, (6.80) by t and summing yields

$$(1-t)\widetilde{f}(0) + t\widetilde{f}(1) = \widetilde{f}(t) + \left((1-t)t^2 + t(1-t)^2\right)\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2, \tag{6.81}$$

or equivalently

$$\widetilde{\Pi}^{\mathrm{lin}}[\widetilde{f}](t) - \widetilde{f}(t) = \left((1-t)t^2 + t(1-t)^2\right)\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2. \tag{6.82}$$

Hence, by integrating (6.82), we get from (6.74)

$$\begin{aligned}
\mathcal{E}_0^{\mathrm{tra}}\left(\|\cdot\|_2^2\right) &= \left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2 \int_0^1 \left((1-t)t^2 + t(1-t)^2\right) dt \\
&= \left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2 \left(B(3,2) + B(2,3)\right) \\
&= \frac{1}{6}\left\|\boldsymbol{v}_1 - \boldsymbol{v}_2\right\|_2^2,
\end{aligned}$$

which is exactly the term on the right-hand side in (6.70), since $L\left(\nabla\left(\|\cdot\|_2^2\right)\right) = 2$. Finally, if f has the general form (6.71), then the equality (6.70) easily follows. □

Combining Proposition 6.4.1, Theorems 6.4.6 and 6.4.7, we arrive at the main error estimate.

Theorem 6.4.8. *For any $f \in C^{1,1}(S_2)$, the following explicit error estimate holds*

$$\left|f(\boldsymbol{x}) - \Pi^{\mathrm{enr}}[f](\boldsymbol{x})\right| \leq \frac{L(\nabla f)}{2}\left(\sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\left\|\boldsymbol{x} - \boldsymbol{v}_i\right\|_2^2 + \frac{1}{6}\sum_{i=0}^{2}|\Gamma_i|^2\left|\langle \boldsymbol{c}_i, \boldsymbol{e}(\boldsymbol{x})\rangle\right|\right), \qquad \boldsymbol{x} \in S_2. \tag{6.83}$$

6.4.3 The L^∞ error estimate

The bound of the pointwise error given in Theorem 6.4.8 allows us to get a bound in L^∞-norm

$$\|f\|_\infty = \max_{\boldsymbol{x}\in S_2} |f(\boldsymbol{x})|$$

of the error $E^{\text{enr}}[f]$ defined in (6.60). As shown in the following result, this bound is proportional to the square of the radius of the circumcircle of the triangle S_2, by a constant factor which depends on f and on the enrichment functions e_i, $i = 0, 1, 2$.

Corollary 6.4.9. *For any $f \in C^{1,1}(S_2)$, the following explicit error estimate holds*

$$\|f - \Pi^{\text{enr}}[f]\|_\infty \leq \frac{L(\nabla f)}{2}\left(1 + \frac{3}{2}\max_{i=0,1,2}\|\langle \boldsymbol{c}_i, \boldsymbol{e}\rangle\|_\infty\right) R^2, \tag{6.84}$$

where $\boldsymbol{c}_i$, $i = 0, 1, 2$ and $\boldsymbol{e}$ are defined in (6.50) *and* (6.53)*, respectively, and R is the circumradius of S_2.*

Proof. We bound the two terms inside the bracket appearing on the right-hand side of equation (6.83) separately. With regard to the first term, we show that

$$\sup_{\boldsymbol{x}\in S_2}\sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{x} - \boldsymbol{v}_i\|_2^2 \leq R^2. \tag{6.85}$$

Indeed, if $\boldsymbol{c}$ is the circumcenter of S_2, by using the linear precision and partition of unity properties of barycentric coordinates, we have

$$\begin{aligned}
\sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{x} - \boldsymbol{v}_i\|_2^2 &= \sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{x} - \boldsymbol{c} - (\boldsymbol{v}_i - \boldsymbol{c})\|_2^2 \\
&= \sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\left(\|\boldsymbol{x} - \boldsymbol{c}\|_2^2 - 2\,\langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{v}_i - \boldsymbol{c}\rangle + \|\boldsymbol{v}_i - \boldsymbol{c}\|_2^2\right) \\
&= \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 - 2\left\langle \boldsymbol{x} - \boldsymbol{c}, \sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\boldsymbol{v}_i - \boldsymbol{c}\right\rangle + \sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i - \boldsymbol{c}\|_2^2 \\
&= \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 - 2\,\langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{x} - \boldsymbol{c}\rangle + R^2 \\
&= R^2 - \|\boldsymbol{x} - \boldsymbol{c}\|_2^2.
\end{aligned}$$

Therefore, for each $\boldsymbol{x} \in S_2$

$$\sum_{i=0}^{2}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{x} - \boldsymbol{v}_i\|_2^2 \leq R^2 - \min_{\boldsymbol{x}\in S_2}\|\boldsymbol{x} - \boldsymbol{c}\|_2^2 \leq R^2, \qquad \boldsymbol{x} \in S_2,$$

and so (6.85) is valid. With regard to the second term inside the bracket, we use Leibniz's inequality to bound the sum of the squares of edge lengths of the triangle in terms of its circumradius [82], i.e.

$$\sum_{i=0}^{2}|\Gamma_i|^2 \leq 9R^2. \tag{6.86}$$

Combining the last estimate with (6.85) gives us the desired inequality (6.84). □

Remark 6.4.10. *We notice that, for $\boldsymbol{x} = \boldsymbol{c}$ we get*

$$\sum_{i=0}^{2}\lambda_i(\boldsymbol{c})\,\|\boldsymbol{c} - \boldsymbol{v}_i\|_2^2 = R^2,$$

then inequality (6.86) *holds with equality if and only if the circumcenter $\boldsymbol{c}$ belongs to S_2. This property characterizes acute triangles, i.e. triangles having all angles less than $\pi/2$.*

Remark 6.4.11. *Let X_n be a set of discrete points in a general position, e.g. a set of scattered points, and assume that we need to triangulate X_n. Inequality* (6.84) *suggests us the use of a triangulation $\mathcal{T}$ which minimizes the maximum of the squares of the circumradii over all triangles of $\mathcal{T}$. The Delaunay triangulation has this property.*

Remark 6.4.12. *For any triangle S_2, it is possible to obtain a more precise L^∞-error bound in terms of the square of the circumradius minus a nonnegative quantity. Indeed, denoting by $\boldsymbol{c}$, $\boldsymbol{b}$ and R the circumcenter, the barycenter and the circumradius of S_2, respectively, then the following bound holds, see [78, Thm. 2.4.4]*

$$\frac{1}{9}\sum_{i=0}^{2}|\Gamma_i|^2 = R^2 - \|\boldsymbol{b}-\boldsymbol{c}\|_2^2 \leq R^2 - \min_{\boldsymbol{x}\in S_2}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2 .$$

The last step follows because $\boldsymbol{b}$ belongs to S_2. Therefore, in Corollary 6.4.9, R^2 can be replaced by the smaller value $R^2 - \min_{\boldsymbol{x}\in S_2}\|\boldsymbol{c}-\boldsymbol{x}\|_2^2$.

Remark 6.4.13. *Let Γ_2 be the longest side of S_2 and let $\boldsymbol{v}_{01}$ be its midpoint. The line segment joining $\boldsymbol{c}$ with $\boldsymbol{v}_{01}$ is a perpendicular bisector of Γ_2. Denoting by*

$$h = \sup_{\boldsymbol{v},\boldsymbol{w}\in S_2}\|\boldsymbol{v}-\boldsymbol{w}\|_2$$

the diameter of S_2, by the Pythagorean Theorem we get

$$\left(\frac{h}{2}\right)^2 + \min_{\boldsymbol{x}\in S_2}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2 = R^2,$$

and then

$$R^2 - \min_{\boldsymbol{x}\in S_2}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2 = \frac{1}{4}h^2.$$

From Remark 6.4.12 and Remark 6.4.13, the Corollary 6.4.9 becomes

Corollary 6.4.14. *For any $f \in C^{1,1}(S_2)$, the following explicit error estimate holds*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_\infty \leq \frac{L(\nabla f)}{8}\left(1 + \frac{3}{2}\max_{i=0,1,2}\|\langle \boldsymbol{c}_i, \boldsymbol{e}\rangle\|_\infty\right)h^2, \tag{6.87}$$

where h is the diameter of S_2.

6.4.4 The L^1 error estimate

The bound of the pointwise error given in Theorem 6.4.8 allows us to get a bound in L^1-norm

$$\|f\|_1 = \int_{S_2}|f(\boldsymbol{x})|\,d\boldsymbol{x}$$

of the error $E^{\mathrm{enr}}[f]$ defined in (6.60).

Theorem 6.4.15. *For any $f \in C^{1,1}(S_2)$, the following explicit error estimate holds*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_1 \leq \frac{L(\nabla f)}{24}\,(1 + e_{max})\,|S_2|\,\omega(S_2), \tag{6.88}$$

where $|S_2|$ is the area of the triangle S_2,

$$e_{max} = \max_{i=0,1,2}\frac{2}{|S_2|}\,\|\langle \boldsymbol{c}_i, \boldsymbol{e}\rangle\|_1 \tag{6.89}$$

and

$$\omega(S_2) = \sum_{i=0}^{2}|\Gamma_i|^2 .$$

Proof. We start by proving the identity

$$\int_{S_2} \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x}-\boldsymbol{v}_i\right\|_2^2 \, d\boldsymbol{x} = \frac{|S_2|}{12} \sum_{i=0}^{2} |\Gamma_i|^2 \,. \tag{6.90}$$

Since $\sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x}-\boldsymbol{v}_i\right\|_2^2$ is a quadratic polynomial which vanishes at all vertices of S_2, by [59, Thm. 5.1], we get

$$\int_{S_2} \sum_{i=0}^{2} \lambda_i(\boldsymbol{x}) \left\|\boldsymbol{x}-\boldsymbol{v}_i\right\|_2^2 \, d\boldsymbol{x} = \frac{|S_2|}{4} \sum_{i=0}^{2} \left\|\boldsymbol{b}-\boldsymbol{v}_i\right\|_2^2 , \tag{6.91}$$

where $\boldsymbol{b}$ is the barycenter of S_2. Moreover, since

$$\begin{aligned} \left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 &= \left\|\boldsymbol{v}_i\right\|_2^2 + \left\|\boldsymbol{v}_j\right\|_2^2 - 2\left\langle \boldsymbol{v}_i, \boldsymbol{v}_j \right\rangle , \qquad i,j=0,1,2, \\ \left\|\boldsymbol{b}-\boldsymbol{v}_j\right\|_2^2 &= \left\|\boldsymbol{b}\right\|_2^2 + \left\|\boldsymbol{v}_j\right\|_2^2 - 2\left\langle \boldsymbol{b}, \boldsymbol{v}_j \right\rangle , \qquad j=0,1,2, \end{aligned}$$

$$\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 - \left\|\boldsymbol{b}-\boldsymbol{v}_j\right\|_2^2 = \left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{b}\right\|_2^2 + 2\left\langle \boldsymbol{b}-\boldsymbol{v}_i, \boldsymbol{v}_j \right\rangle , \qquad i,j=0,1,2,$$

and then, by summing over all $j=0,1,2$, both members of the last equality, we immediately get

$$\begin{aligned} \sum_{j=0}^{2} \left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 - \sum_{j=0}^{2} \left\|\boldsymbol{b}-\boldsymbol{v}_j\right\|_2^2 &= 3\left(\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{b}\right\|_2^2\right) + 6\left\langle \boldsymbol{b}-\boldsymbol{v}_i, \boldsymbol{b} \right\rangle \\ &= 3\left(\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{b}\right\|_2^2 + 2\left\langle \boldsymbol{b}-\boldsymbol{v}_i, \boldsymbol{b} \right\rangle\right) \\ &= 3\left\|\boldsymbol{b}-\boldsymbol{v}_i\right\|_2^2 , \qquad i=0,1,2, \end{aligned}$$

or equivalently,

$$\sum_{j=0}^{2} \left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 = \sum_{j=0}^{2} \left\|\boldsymbol{b}-\boldsymbol{v}_j\right\|_2^2 + 3\left\|\boldsymbol{b}-\boldsymbol{v}_i\right\|_2^2 , \qquad i=0,1,2.$$

Now we sum both members of the above equality over all $i=0,1,2$ and we get

$$\sum_{i=0}^{2} \sum_{j=0}^{2} \left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 = 6 \sum_{i=0}^{2} \left\|\boldsymbol{b}-\boldsymbol{v}_i\right\|_2^2 ,$$

or equivalently

$$\sum_{i=0}^{2} \left\|\boldsymbol{b}-\boldsymbol{v}_i\right\|_2^2 = \frac{1}{6} \sum_{i=0}^{2} \sum_{j=0}^{2} \left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 = \frac{1}{3} \sum_{i=0}^{2} |\Gamma_i|^2 \,. \tag{6.92}$$

The identity (6.90) follows by substituting (6.92) in (6.91). Finally, by integrating (6.83) and by using (6.90), we get (6.88) and then the thesis. □

In the following, we further analyze the case where the enrichment functions are

$$e_0 = \lambda_1^{\alpha_0-1} \lambda_2^{\beta_0-1}, \qquad e_1 = \lambda_0^{\alpha_1-1} \lambda_2^{\beta_1-1}, \quad e_2 = \lambda_0^{\alpha_2-1} \lambda_1^{\beta_2-1}, \tag{6.93}$$

with $\alpha_i, \beta_i > 1$. In particular, we determine the *best* parameters α_i, β_i, in order to minimize the approximation error (6.88).

Theorem 6.4.16. *Let e_i, $i=0,1,2$, be the enrichment functions defined in* (6.93)*. Then, for any $f \in C^{1,1}(S_2)$, we get*

$$\left\|f - \Pi^{\mathrm{enr}}[f]\right\|_1 \leq \frac{L(\nabla f)}{24} \left(1+\frac{4}{\mu}\right) |S_2| \, \omega(S_2), \tag{6.94}$$

where $\mu = \min\limits_{i=0,1,2} (\alpha_i + \beta_i)$.

Proof. Let us compute e_{max} defined in (6.89), relative to the enrichment functions (6.93). In this case, from (6.28), the matrix N defined in (6.43) is a diagonal matrix and

$$N^{-1} = [\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{c}_2] = \begin{bmatrix} \frac{1}{B(\alpha_0,\beta_0)} & 0 & 0 \\ 0 & \frac{1}{B(\alpha_1,\beta_1)} & 0 \\ 0 & 0 & \frac{1}{B(\alpha_2,\beta_2)} \end{bmatrix}.$$

Consequently we get

$$\begin{aligned} \frac{2}{|S_2|}\int_{S_2} |\langle \boldsymbol{c}_0, \boldsymbol{e}(\boldsymbol{x})\rangle|\, d\boldsymbol{x} &= \frac{2}{|S_2|\, B(\alpha_0,\beta_0)}\int_{S_2} \lambda_1^{\alpha_0-1}(\boldsymbol{x})\lambda_2^{\beta_0-1}(\boldsymbol{x})\, d\boldsymbol{x}, \\ \frac{2}{|S_2|}\int_{S_2} |\langle \boldsymbol{c}_1, \boldsymbol{e}(\boldsymbol{x})\rangle|\, d\boldsymbol{x} &= \frac{2}{|S_2|\, B(\alpha_1,\beta_1)}\int_{S_2} \lambda_0^{\alpha_1-1}(\boldsymbol{x})\lambda_2^{\beta_1-1}(\boldsymbol{x})\, d\boldsymbol{x}, \\ \frac{2}{|S_2|}\int_{S_2} |\langle \boldsymbol{c}_2, \boldsymbol{e}(\boldsymbol{x})\rangle|\, d\boldsymbol{x} &= \frac{2}{|S_2|\, B(\alpha_2,\beta_2)}\int_{S_2} \lambda_0^{\alpha_2-1}(\boldsymbol{x})\lambda_1^{\beta_2-1}(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned} \tag{6.95}$$

In order to compute the integrals at the right side of (6.95) we denote by

$$\widehat{S}_2 = \left\{\widehat{\boldsymbol{x}} = (\widehat{x},\widehat{y}) \in \mathbb{R}^2,\, \widehat{x} \geq 0,\, \widehat{y} \geq 0,\, \widehat{x}+\widehat{y} \leq 1\right\}, \tag{6.96}$$

the triangle of vertices $\widehat{\boldsymbol{v}}_0 = (0,0)$, $\widehat{\boldsymbol{v}}_1 = (1,0)$ and $\widehat{\boldsymbol{v}}_2 = (0,1)$ and by $F : \widehat{S}_2 \to S_2$ the affine map defined as

$$F(\widehat{\boldsymbol{x}}) = \widehat{\lambda}_0(\widehat{\boldsymbol{x}})\boldsymbol{v}_0 + \widehat{\lambda}_1(\widehat{\boldsymbol{x}})\boldsymbol{v}_1 + \widehat{\lambda}_2(\widehat{\boldsymbol{x}})\boldsymbol{v}_2,$$

where $\widehat{\lambda}_0(\widehat{\boldsymbol{x}})$, $\widehat{\lambda}_1(\widehat{\boldsymbol{x}})$, $\widehat{\lambda}_2(\widehat{\boldsymbol{x}})$ are the barycentric coordinates of the point $\widehat{\boldsymbol{x}}$ with respect to the triangle $\widehat{S}_2$. Since F preserves the ratio of areas of two triangles, by definition of barycenter coordinates we get

$$\lambda_i \circ F(\widehat{\boldsymbol{x}}) = \widehat{\lambda}_i(\widehat{\boldsymbol{x}}).$$

Therefore, by using F as change of variables, we get

$$\begin{aligned} \int_{S_2} \lambda_1^{\alpha_0-1}(\boldsymbol{x})\lambda_2^{\beta_0-1}(\boldsymbol{x})\, d\boldsymbol{x} &= 2|S_2| \int_{\widehat{S}_2} \widehat{\lambda}_1^{\alpha_0-1}(\widehat{\boldsymbol{x}})\widehat{\lambda}_2^{\beta_0-1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}}, \\ \int_{S_2} \lambda_0^{\alpha_1-1}(\boldsymbol{x})\lambda_2^{\beta_1-1}(\boldsymbol{x})\, d\boldsymbol{x} &= 2|S_2| \int_{\widehat{S}_2} \widehat{\lambda}_0^{\alpha_1-1}(\widehat{\boldsymbol{x}})\widehat{\lambda}_2^{\beta_1-1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}}, \\ \int_{S_2} \lambda_0^{\alpha_2-1}(\boldsymbol{x})\lambda_1^{\beta_2-1}(\boldsymbol{x})\, d\boldsymbol{x} &= 2|S_2| \int_{\widehat{S}_2} \widehat{\lambda}_0^{\alpha_2-1}(\widehat{\boldsymbol{x}})\widehat{\lambda}_1^{\beta_2-1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}}, \end{aligned} \tag{6.97}$$

where $2\,|S_2|$ is the Jacobian determinant of F. The integrals at right member of equalities (6.97) can be computed by using the powerful formula presented in [15]

$$\int_{\widehat{S}_2} \widehat{\lambda}_0^{\alpha}(\widehat{\boldsymbol{x}})\widehat{\lambda}_1^{\beta}(\widehat{\boldsymbol{x}})\widehat{\lambda}_2^{\gamma}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}} = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)\Gamma(\gamma+1)}{\Gamma(\alpha+\beta+\gamma+3)}, \tag{6.98}$$

valid for each $\alpha, \beta, \gamma > -1$. By combining equalities (6.95), (6.97), (6.98) and (6.26), we get (6.94). □

Now we consider the more general enrichment functions, already introduced in Example 7.1.5

$$\widetilde{e}_i = (1-\lambda_i)^{\gamma_i} e_i, \qquad i = 0,1,2, \qquad \gamma_i \geq 0, \tag{6.99}$$

where e_i, $i = 0,1,2$, are defined in (6.93). In analogy to the Theorem 6.4.16, in Theorem 6.4.18 we determine, by using the enrichment functions (6.99), the *best* parameters α_i, β_i, γ_i which minimize the approximation error (6.88). To this aim we consider the triangle $\widehat{S}_2$ with vertices $\widehat{\boldsymbol{v}}_0 = (0,0)$, $\widehat{\boldsymbol{v}}_1 = (1,0)$ and $\widehat{\boldsymbol{v}}_2 = (0,1)$ and barycentric coordinates $\widehat{\lambda}_0(\widehat{\boldsymbol{x}})$, $\widehat{\lambda}_1(\widehat{\boldsymbol{x}})$, $\widehat{\lambda}_2(\widehat{\boldsymbol{x}})$.

Lemma 6.4.17. *For each $i = 0, 1, 2$, the following equality holds*

$$\int_{\widehat{S}_2} (1 - \widehat{\lambda}_i(\widehat{\boldsymbol{x}}))^{\gamma_i} \widehat{\lambda}_0^{\alpha_0 - 1}(\widehat{\boldsymbol{x}}) \widehat{\lambda}_1^{\alpha_1 - 1}(\widehat{\boldsymbol{x}}) \widehat{\lambda}_2^{\alpha_2 - 1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}} = \frac{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma\left(\sum_{j=0, j\neq i}^{2} \alpha_j\right)} \mu_i, \tag{6.100}$$

where

$$\mu_i = \frac{\Gamma\left(\gamma_i + \sum_{j=0, j\neq i}^{2} \alpha_j\right)}{\Gamma\left(\gamma_i + \sum_{j=0}^{2} \alpha_i\right)}, \qquad \gamma_i \geq 0, \qquad \alpha_i > 1, \qquad i = 0, 1, 2.$$

Proof. It is well known that the barycentric coordinates of $\widehat{S}_2$ in terms of Cartesian coordinates are

$$\widehat{\lambda}_0(\widehat{\boldsymbol{x}}) = 1 - \widehat{x} - \widehat{y}, \quad \widehat{\lambda}_1(\widehat{\boldsymbol{x}}) = \widehat{x}, \quad \widehat{\lambda}_2(\widehat{\boldsymbol{x}}) = \widehat{y}.$$

First we prove (6.100) for $i \neq 0$. Without loss of generality we can assume $i = 1$. We note that

$$\int_{\widehat{S}_2} (1 - \widehat{x})^{\gamma_1} \widehat{x}^{\alpha_1 - 1} \widehat{y}^{\alpha_2 - 1} (1 - \widehat{x} - \widehat{y})^{\alpha_0 - 1}\, d\widehat{x} d\widehat{y} = \int_0^1 (1 - \widehat{x})^{\gamma_1} \widehat{x}^{\alpha_1 - 1} I(\widehat{x})\, d\widehat{x}, \tag{6.101}$$

where

$$I(\widehat{x}) = \int_0^{1-\widehat{x}} \widehat{y}^{\alpha_2 - 1} (1 - \widehat{x} - \widehat{y})^{\alpha_0 - 1}\, d\widehat{y}. \tag{6.102}$$

In order to compute the integral (6.102) we consider the change of variable

$$\widehat{z} = \frac{\widehat{y}}{1 - \widehat{x}},$$

and then

$$\begin{aligned} I(\widehat{x}) &= (1 - \widehat{x})^{\alpha_0 + \alpha_2 - 1} \int_0^1 \widehat{z}^{\alpha_2 - 1} (1 - \widehat{z})^{\alpha_0 - 1}\, d\widehat{z} && (6.103) \\ &= (1 - \widehat{x})^{\alpha_0 + \alpha_2 - 1} \frac{\Gamma(\alpha_0)\Gamma(\alpha_2)}{\Gamma(\alpha_0 + \alpha_2)}. && (6.104) \end{aligned}$$

By definition of beta function, we get

$$\int_0^1 \widehat{x}^{\alpha_1 - 1} (1 - \widehat{x})^{\gamma_1 + \alpha_0 + \alpha_2 - 1}\, dx = \frac{\Gamma(\alpha_1)\Gamma(\gamma_1 + \alpha_0 + \alpha_2)}{\Gamma(\alpha_0 + \alpha_1 + \alpha_2 + \gamma_1)}. \tag{6.105}$$

The result follows by combining (6.101), (6.104) and (6.105). By the same arguments we can prove formula (6.100) for $i = 2$. In order to prove (6.100) for $i = 0$, we consider the coordinate transformations

$$\begin{aligned} &\pi : \widehat{S}_2 \to \widehat{S}_2, \quad \pi(x, y) = (y, x), \\ &\theta : \widehat{S}_2 \to \widehat{S}_2, \quad \theta(x, y) = (1 - x - y, y) =: (u, v), \end{aligned}$$

and the change of variables $\pi \circ \theta$ whose Jacobian determinant is equal to 1. Then we get

$$\int_{\widehat{S}_2} (x + y)^{\gamma_0} x^{\alpha_1 - 1} y^{\alpha_2 - 1} (1 - x - y)^{\alpha_0 - 1}\, dx dy = \int_{\widehat{S}_2} (1 - u)^{\gamma_0} G(u, v)\, du dv,$$

where $G(u, v) = u^{\alpha_0 - 1} v^{\alpha_2 - 1} (1 - u - v)^{\alpha_1 - 1}$. The result now follows from the case $i = 1$. □

Theorem 6.4.18. *Let $\widetilde{e}_i$, $i = 0, 1, 2$, be the enrichment functions defined in* (6.99)*. Then, for any $f \in C^{1,1}(S_2)$, we get*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_1 \leq \frac{L(\nabla f)}{24} \left(1 + \frac{4}{\mu}\right) |S_2|\, \omega(S_2), \tag{6.106}$$

where $\mu = \min_{i=0,1,2} (\alpha_i + \beta_i + \gamma_i)$.

Proof. Let us compute e_{max} defined in (6.89), relative to the enrichment functions (6.99). By (6.49), the matrix N defined in (6.43) is a diagonal matrix and

$$N^{-1} = [\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{c}_2] = \begin{bmatrix} \frac{1}{B(\alpha_0,\beta_0)} & 0 & 0 \\ 0 & \frac{1}{B(\alpha_1,\beta_1)} & 0 \\ 0 & 0 & \frac{1}{B(\alpha_2,\beta_2)} \end{bmatrix}.$$

Consequently we get

$$\begin{aligned}
\frac{2}{|S_2|}\int_{S_2} |\langle \boldsymbol{c}_0, \boldsymbol{e}(\boldsymbol{x})\rangle|\, d\boldsymbol{x} &= \frac{2}{|S_2|\, B(\alpha_0,\beta_0)}\int_{S_2} (1-\lambda_0(\boldsymbol{x}))^{\gamma_0}\lambda_1^{\alpha_0-1}(\boldsymbol{x})\lambda_2^{\beta_0-1}(\boldsymbol{x})\, d\boldsymbol{x},\\
\frac{2}{|S_2|}\int_{S_2} |\langle \boldsymbol{c}_1, \boldsymbol{e}(\boldsymbol{x})\rangle|\, d\boldsymbol{x} &= \frac{2}{|S_2|\, B(\alpha_1,\beta_1)}\int_{S_2} (1-\lambda_1(\boldsymbol{x}))^{\gamma_1}\lambda_0^{\alpha_1-1}(\boldsymbol{x})\lambda_2^{\beta_1-1}(\boldsymbol{x})\, d\boldsymbol{x},\\
\frac{2}{|S_2|}\int_{S_2} |\langle \boldsymbol{c}_2, \boldsymbol{e}(\boldsymbol{x})\rangle|\, d\boldsymbol{x} &= \frac{2}{|S_2|\, B(\alpha_2,\beta_2)}\int_{S_2} (1-\lambda_2(\boldsymbol{x}))^{\gamma_2}\lambda_0^{\alpha_2-1}(\boldsymbol{x})\lambda_1^{\beta_2-1}(\boldsymbol{x})\, d\boldsymbol{x}.
\end{aligned}$$

By using the same strategy of the proof of Theorem 6.4.16, we get

$$\begin{aligned}
\int_{S_2} (1-\lambda_0(\boldsymbol{x}))^{\gamma_0}\lambda_1^{\alpha_0-1}(\boldsymbol{x})\lambda_2^{\beta_0-1}(\boldsymbol{x})\, d\boldsymbol{x} &= 2|S_2|\int_{\widehat{S}_2} (1-\widehat{\lambda}_0(\widehat{\boldsymbol{x}}))^{\gamma_0}\widehat{\lambda}_1^{\alpha_0-1}(\widehat{\boldsymbol{x}})\widehat{\lambda}_2^{\beta_0-1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}},\\
\int_{S_2} (1-\lambda_1(\boldsymbol{x}))^{\gamma_1}\lambda_0^{\alpha_1-1}(\boldsymbol{x})\lambda_2^{\beta_1-1}(\boldsymbol{x})\, d\boldsymbol{x} &= 2|S_2|\int_{\widehat{S}_2} (1-\widehat{\lambda}_1(\widehat{\boldsymbol{x}}))^{\gamma_1}\widehat{\lambda}_0^{\alpha_1-1}(\widehat{\boldsymbol{x}})\widehat{\lambda}_2^{\beta_1-1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}},\\
\int_{S_2} (1-\lambda_2(\boldsymbol{x}))^{\gamma_2}\lambda_0^{\alpha_2-1}(\boldsymbol{x})\lambda_1^{\beta_2-1}(\boldsymbol{x})\, d\boldsymbol{x} &= 2|S_2|\int_{\widehat{S}_2} (1-\widehat{\lambda}_2(\widehat{\boldsymbol{x}}))^{\gamma_2}\widehat{\lambda}_0^{\alpha_2-1}(\widehat{\boldsymbol{x}})\widehat{\lambda}_1^{\beta_2-1}(\widehat{\boldsymbol{x}})\, d\widehat{\boldsymbol{x}}.
\end{aligned}$$

The result follows from Lemma 6.4.17. □

The following result shows that the bound (6.106) is proportional to the fourth power of the circumradius of the triangle S_2.

Corollary 6.4.19. *Let $\widetilde{e}_i$, $i = 0,1,2$, be the enrichment functions defined in* (6.99)*. Then, for any $f \in C^{1,1}(S_2)$, we get*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_1 \leq \frac{9\sqrt{3}L(\nabla f)}{32}\left(1+\frac{4}{\mu}\right)R^4,$$

where R is the circumradius of S_2 and $\mu = \min_{i=0,1,2}(\alpha_i+\beta_i+\gamma_i)$.

Proof. By Theorem 6.4.18 and Leibniz's inequality, it suffices to bound the area of S_2 in terms of its circumradius. To this end, we use Weitzenböck's inequality [82], which states that for any triangle S_2 of sides Γ_0,Γ_1, Γ_2, the following inequality holds

$$4\sqrt{3}\,|S_2| \leq \sum_{i=0}^{2} |\Gamma_i|^2.$$

With the aid of this estimate and Leibniz's inequality, it is easily seen that

$$|S_2| \leq \frac{9}{4\sqrt{3}}R^2. \tag{6.107}$$

The result follows by (6.86), (6.107) and (6.106). □

By Remark 6.4.12, it is possible to obtain a more precise L^1-error bound in terms of the diameter of S_2.

Corollary 6.4.20. *Let e_i, $i = 0,1,2$, be the enrichment functions defined in* (6.99)*. Then, for any $f \in C^{1,1}(S_2)$, we have*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_1 \leq \frac{9\sqrt{3}L(\nabla f)}{512}\left(1+\frac{4}{\mu}\right)h^4,$$

where h is the diameter of S_2 and $\mu = \min_{i=0,1,2}(\alpha_i+\beta_i+\gamma_i)$.

6.5 Practical consideration

Let $X_n = \{\boldsymbol{x}_i : i = 1\ldots, n\}$ be a set of n scattered data in $\mathbb{R}^2$ and let $\mathcal{T}_m = \{t_j : j = 1, \ldots, m\}$ be a triangulation of X_n. We denote by $\mathcal{C} = \text{conv}(X_n)$ the convex hull of X_n. We define the global approximation operator $\Pi^{\text{enr}}_{\mathcal{T}_m}$ by setting, for any $f \in C^{1,1}(\mathcal{C})$ and $\boldsymbol{x} \in \mathcal{C}$,

$$\Pi^{\text{enr}}_{\mathcal{T}_m}[f](\boldsymbol{x}) = \Pi^{\text{enr}}[f, t_j](\boldsymbol{x}), \quad \text{if} \qquad \boldsymbol{x} \in t_j, \qquad j = 1, \ldots, m,$$

where $\Pi^{\text{enr}}[f, t_j]$ is the approximation operator defined in (6.61), based on the triangular element $(t_j, \mathbb{P}^{\text{enr}}_1(t_j), \Sigma^{\text{enr}}_{t_j})$. In this case, Theorem 6.4.16 gives the following global L^1-error bound

$$\begin{aligned} \left\| f - \Pi^{\text{enr}}_{\mathcal{T}_m}[f] \right\|_1 &= \int_{\mathcal{C}} \left| f(\boldsymbol{x}) - \Pi^{\text{enr}}_{\mathcal{T}_m}[f](\boldsymbol{x}) \right| \, d\boldsymbol{x} = \sum_{j=1}^{m} \int_{t_j} \left| f(\boldsymbol{x}) - \Pi^{\text{enr}}[f, t_j](\boldsymbol{x}) \right| \, d\boldsymbol{x} \\ &\leq \frac{L(\nabla f)}{24} \left(1 + \frac{4}{\mu}\right) \sum_{j=1}^{m} |t_j| \, \omega(t_j). \end{aligned} \tag{6.108}$$

For any triangulation $\mathcal{T}_m$, we denote by

$$E_1(\mathcal{T}_m) = \sum_{j=1}^{m} |t_j| \, \omega(t_j).$$

The global error bound (6.108) is proportional to E_1, by a constant factor which is independent on the triangulation $\mathcal{T}_m$. In analogy, by using the results of Corollaries 6.4.9 and 6.4.19 we denote by

$$E_2(\mathcal{T}_m) = \sum_{j=1}^{m} R_j^2,$$

$$E_3(\mathcal{T}_m) = \sum_{j=1}^{m} R_j^4,$$

where R_j is the circumradius of the triangle $t_j \in \mathcal{T}_m$. By using the optimality results of Delaunay triangulation, which can be found in [41] and [53], it is possible to prove the following result.

Theorem 6.5.1. *Let X_n be a set of n scattered data in $\mathbb{R}^2$ and let $\mathcal{T}_m$ be a triangulation of X_n. Then $E_i(\mathcal{T}_m)$, $i = 1, 2, 3$ achieve their minimum if and only if $\mathcal{T}_m$ is the Delaunay triangulation.*

6.6 Numerical experiments

In this Section, we test the accuracy of the approximation produced by the finite element $(S_2, \mathbb{P}^{\text{enr}}_1(S_2), \Sigma^{\text{enr}}_{S_2})$ obtained by enriching the standard triangular linear finite element $\mathcal{P}_1(S_2)$ with the sets of enrichment functions introduced in Example 6.3.3. For each experiment, we use a regular grid of $(n+1) \times (n+1)$ equispaced points, with $n = 2^k$, $k = 2, \ldots, 6$ and the relative Delaunay triangulation, see Figure 6.1.

We consider the following test functions

$$f_1(x, y) = \sqrt{x^2 + y^2 + 1}, \quad f_2(x, y) = \sin(\pi x y),$$

and the following sets of enrichment functions

$$\mathcal{E}_1 = \left\{ e_i = (1 - \lambda_i) \prod_{\substack{j=0 \\ j \neq i}}^{2} \lambda_j \, : \, i = 0, 1, 2 \right\}, \qquad \mathcal{E}_2 = \left\{ e_i = \sqrt{(1 - \lambda_i)} \prod_{\substack{j=0 \\ j \neq i}}^{2} \lambda_j \, : \, i = 0, 1, 2 \right\}$$

already introduced in Example 6.3.3. For each of these, we compare the accuracy of approximation, in L^1-norm, produced by the standard triangular linear finite element with that one produced by the enriched

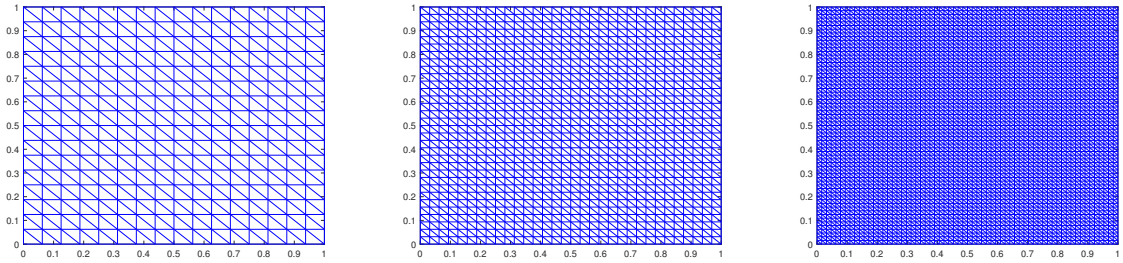


Figure 6.1: Delaunay triangulation of a regular grid of $(n+1) \times (n+1)$ equispaced points, with $n = 16$ (left), $n = 32$ (center), $n = 64$ (right).

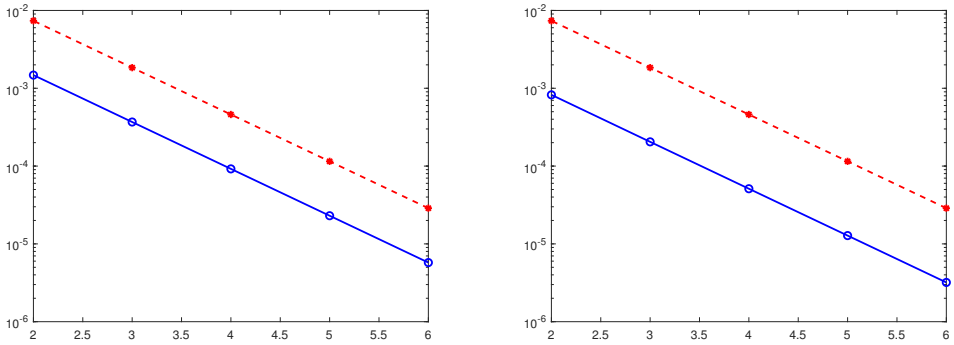


Figure 6.2: Semilog plot of the trend of the errors, in L^1-norm, produced by approximating the function $f_1(x, y)$ working with Delaunay triangulations of the unit square $[0, 1]^2$, the standard triangular linear finite element (red dashed line), and the enriched finite element $(S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}})$ (blue line). The Delaunay triangulations are realized by using regular grids of $(n+1) \times (n+1)$ equispaced nodes with $n = 2^k$, $k = 2, \ldots, 6$. The enrichments of the standard triangular linear finite element are realized by using the sets of enrichment functions $\mathcal{E}_1$ and $\mathcal{E}_2$ for the left and the right picture, respectively.

finite element $(S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}})$. We perform the numerical experiments by using `MatLab` software. To compute the integral of a bivariate function over a side of the triangle S_2 and the integral of a bivariate function over S_2 we use the command `integral2`. The results are reported in Figure 6.2 and Figure 6.3. We notice that for a fixed function f, not every set of enrichment functions significantly improves the accuracy of the approximation realized by the enriched finite element. The accuracy of the approximation depends on the chosen set of enrichment functions.

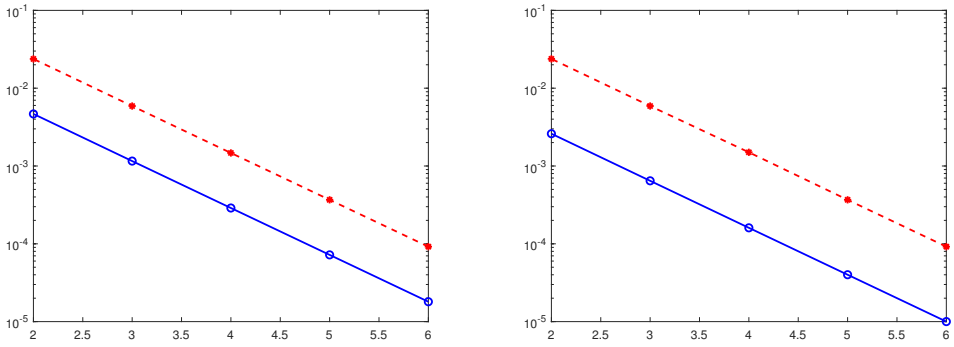


Figure 6.3: Semilog plot of the trend of the errors, in L^1-norm, produced by approximating the function $f_2(x, y)$ working with Delaunay triangulations of the unit square $[0, 1]^2$, the standard triangular linear finite element (red dashed line), and the enriched finite element $(S_2, \mathbb{P}_1^{\text{enr}}(S_2), \Sigma_{S_2}^{\text{enr}})$ (blue line). The Delaunay triangulations are realized by using regular grids of $(n + 1) \times (n + 1)$ equispaced nodes with $n = 2^k$, $k = 2, \ldots, 6$. The enrichments of the standard triangular linear finite element are realized by using the sets of enrichment functions $\mathcal{E}_1$ and $\mathcal{E}_2$ for the left and the right picture, respectively.

Chapter 7

Enrichment strategies for the standard simplicial linear finite elements

In this chapter, we generalize the results presented in Chapter 6 to the d dimensional case. In particular, we introduce a new class of finite elements by enriching the standard simplicial linear finite element in $\mathbb{R}^d$, $\mathcal{P}_1(S_d)$, with additional functions which are not necessarily polynomials. We provide necessary and sufficient conditions on the enrichment functions, which guarantee the existence of families of such enriched elements. Furthermore, we derive explicit formulas for their associated basis functions. We also show that the approximation error, obtained by using the proposed enriched elements, can be written as the error of the standard simplicial linear finite element plus a second term which depends on the enrichment functions. By using this decomposition, we derive explicit bounds in both L^∞-norm and L^1-norm. The results presented in this chapter can be found in [34].

7.1 Enrichment of the standard simplicial linear finite element

Let $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$ be affinely independent points in $\mathbb{R}^d$ and let S_d be the d-simplex in $\mathbb{R}^d$ with vertices $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$. Every point $\boldsymbol{x} \in S_d$ can be uniquely expressed as

$$\boldsymbol{x} = \sum_{i=0}^{d} \lambda_i(\boldsymbol{x}) \boldsymbol{v}_i, \tag{7.1}$$

where $\lambda_i(\boldsymbol{x})$, $i = 0, \ldots, d$, are the barycentric coordinates of the point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to S_d [97]. An m-dimensional face of S_d is any m-simplex generated by $m+1$ vertices of S_d. In particular, we denote by F_i, $i = 0, \ldots, d$, the $(d-1)$-dimensional face generated by $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{v}_{i+1}, \ldots, \boldsymbol{v}_d$ and by $|F_i|$ its area. In the discussion below we will use a well-known result, which holds for any d-simplex S_d, $d \in \mathbb{N}$ [42, Ch. 2].

Lemma 7.1.1. *Let $\alpha_0, \ldots, \alpha_d$ be positive real numbers. Then the following identity holds*

$$\frac{1}{|S_d|} \int_{S_d} \prod_{i=0}^{d} \lambda_i^{\alpha_i}(\boldsymbol{x}) \, d\boldsymbol{x} = \frac{d! \prod_{i=0}^{d} \Gamma(\alpha_i + 1)}{\Gamma(d + 1 + \sum_{i=0}^{d} \alpha_i)}, \tag{7.2}$$

where $|S_d|$ is the volume of S_d and $\Gamma(z)$ is the gamma function [1].

We notice that, for any $\boldsymbol{x} \in F_i$, we have

$$\lambda_i(\boldsymbol{x}) = 0, \tag{7.3}$$

and therefore, by equation (7.1), we get

$$\boldsymbol{x} = \sum_{\substack{j=0 \\ j \neq i}}^{d} \lambda_j(\boldsymbol{x}) \boldsymbol{v}_j, \quad \sum_{\substack{j=0 \\ j \neq i}}^{d} \lambda_j(\boldsymbol{x}) = 1, \tag{7.4}$$

that is, $\lambda_j(\boldsymbol{x})$, $j = 0, \ldots, d$, $j \neq i$, are the barycentric coordinates of $\boldsymbol{x}$ with respect to the $(d-1)$-dimensional simplex F_i. Consequently, by Lemma 7.1.1, we get

$$\frac{1}{|F_i|}\int_{F_i}\prod_{\substack{j=0\\ j\neq i}}^{d}\lambda_j^{\alpha_j}(\boldsymbol{x})\,d\sigma(\boldsymbol{x}) = \frac{(d-1)!\prod_{j=0,j\neq i}^{d}\Gamma(\alpha_j+1)}{\Gamma(d+\sum_{j=0,j\neq i}^{d}\alpha_j)}, \qquad i = 0,\ldots,d, \tag{7.5}$$

where the integral is computed with respect to the Lebesgue measure on the $(d-1)$-dimensional face F_i.

For $f \in C(S_d)$, we set

$$L_j(f) = f(\boldsymbol{v}_j), \qquad j = 0,\ldots,d. \tag{7.6}$$

The standard simplicial linear finite element is the triple

$$\mathcal{P}_1(S_d) = (S_d, \mathbb{P}_1(S_d), \Sigma_{S_d}^{\mathrm{lin}}), \tag{7.7}$$

where

$$\mathbb{P}_1(S_d) = \mathrm{span}\{\lambda_0,\ldots,\lambda_d\} \tag{7.8}$$

is the space of linear polynomials in $\mathbb{R}^d$ and

$$\Sigma_{S_d}^{\mathrm{lin}} = \{L_0(f), \ldots L_d(f)\}$$

is the set of degrees of freedom of $\mathcal{P}_1(S_d)$. The main goal of this chapter is to generalize the results presented in [30] to the case of a non-degenerate d-simplex $S_d \subset \mathbb{R}^d$. For this purpose, we consider $d+1$ linearly independent continuous functions $e_0,\ldots,e_d$ on S_d, satisfying the *vanishing conditions*

$$e_i(\boldsymbol{v}_j) = 0, \qquad i,j = 0,\ldots,d, \tag{7.9}$$

and the enrichment linear functionals

$$I_j(f) = \frac{1}{|F_j|}\int_{F_j} f(\boldsymbol{x})d\sigma(\boldsymbol{x}), \qquad j = 0,\ldots,d. \tag{7.10}$$

In analogy to Chapter 6, we denote by

$$\Sigma_{S_d}^{\mathrm{enr}} = \{L_j, I_j \,:\, j = 0,\ldots,d\},$$

and by

$$\mathbb{P}_1^{\mathrm{enr}}(S_d) = \mathbb{P}_1(S_d) \oplus \mathrm{span}\,\{e_0,\ldots,e_d\}.$$

In the next Theorem we prove a characterization result for the enrichment functions $e_0,\ldots,e_d$, so that the triple

$$(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$$

is a finite element, or equivalently so that $\mathbb{P}_1^{\mathrm{enr}}(S_d)$ is $\Sigma_{S_d}^{\mathrm{enr}}$-unisolvent.

Theorem 7.1.2. *Let*

$$N = \begin{bmatrix} I_0(e_0) & \ldots & I_0(e_d) \\ I_1(e_0) & \ldots & I_1(e_d) \\ \vdots & \vdots & \vdots \\ I_d(e_0) & \ldots & I_d(e_d) \end{bmatrix}, \tag{7.11}$$

then the triple $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$ *is a finite element if and only if*

$$\det(N) \neq 0.$$

Proof. Let us assume that $\det(N) \neq 0$ and we prove that $\mathbb{P}_1^{\text{enr}}(S_d)$ is $\Sigma_{S_d}^{\text{enr}}$-unisolvent. Let $f \in \mathbb{P}_1^{\text{enr}}(S_d)$ be a function such that

$$L_j(f) = 0, \qquad j = 0, \dots, d, \tag{7.12}$$

$$I_j(f) = 0, \qquad j = 0, \dots, d. \tag{7.13}$$

Since f belongs to $\mathbb{P}_1^{\text{enr}}(S_d)$, it can be expressed as

$$f = p + \sum_{i=0}^{d} \beta_i e_i,$$

where $p \in \mathbb{P}_1(S_d)$ and $\beta_0, \dots, \beta_d$ are real numbers. The vanishing conditions (7.9) imply that

$$L_j(f) = L_j(p) = p(\boldsymbol{v}_j) = 0, \qquad j = 0 \dots, d,$$

and then $p = 0$. Consequently f coincides with its enriched part, that is

$$f = \sum_{i=0}^{d} \beta_i e_i.$$

By using the linearity of the functionals I_j, $j = 0, \dots, d$, equations (7.13) can be represented in matrix form as

$$N\boldsymbol{\beta} = \mathbf{0}, \tag{7.14}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_d]^T$ and $\mathbf{0}$ is the zero vector in $\mathbb{R}^{d+1}$. Since, by hypothesis, the matrix N is nonsingular, the linear system (7.14) has a unique solution $\beta_0 = \beta_1 = \cdots = \beta_d = 0$ and then $f = 0$.
In order to prove the reverse implication, let us assume that $\det(N) = 0$ and we prove that $\mathbb{P}_1^{\text{enr}}(S_d)$ is not $\Sigma_{S_d}^{\text{enr}}$-unisolvent. Since $\det(N) = 0$, there exists $[\gamma_0, \dots, \gamma_d]^T \neq \mathbf{0}$ such that the function

$$e = \sum_{i=0}^{d} \gamma_i e_i$$

satisfies

$$I_j(e) = 0, \qquad j = 0, \dots, d.$$

Since the enrichment functions e_i, $i = 0, \dots, d$, satisfy the vanishing conditions (7.9), we get

$$L_j(e) = 0, \qquad j = 0, \dots, d.$$

We have found a linear combination of the basis functions of $\mathbb{P}_1^{\text{enr}}(S_d)$ with coefficients not all zero in which all degrees of freedom vanish. Then $\mathbb{P}_1^{\text{enr}}(S_d)$ is not $\Sigma_{S_d}^{\text{enr}}$-unisolvent. □

Definition 7.1.3. *Let $e_0, \dots, e_d$ be linearly independent continuous enrichment functions satisfying the vanishing conditions* (7.9)*. They are said admissible enrichment functions if we can enrich $\mathcal{P}_1(S_d)$ to the finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$.*

In the following, we assume that the matrix N is nonsingular and we denote its inverse by

$$N^{-1} = [\boldsymbol{c}_0 \dots \boldsymbol{c}_d],$$

where $\boldsymbol{c}_i \in \mathbb{R}^d$, $i = 0, \dots, d$, are column vectors. A direct consequence of Theorem 7.1.2 is the linear independence of the functionals of $\Sigma_{S_d}^{\text{enr}}$ in the dual space $\mathbb{P}_1^{\text{enr}}(S_d)^\star$ [23, Ch 2]. Then, there exists a basis $\{\varphi_j, \psi_j : j = 0, \dots, d\}$ of $\mathbb{P}_1^{\text{enr}}(S_d)$ which satisfy

$$L_j(\varphi_i) = \delta_{ij}, \quad I_j(\varphi_i) = 0, \qquad i, j = 0, \dots, d, \tag{7.15}$$

$$L_j(\psi_i) = 0, \quad I_j(\psi_i) = \delta_{ij}, \qquad i, j = 0, \dots, d. \tag{7.16}$$

Theorem 7.1.4. *The basis functions* $\{\varphi_j, \psi_j : j = 0, \dots, d\}$ *of* $\mathbb{P}_1^{\text{enr}}(S_d)$ *associated to the finite element* $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$*, which satisfy* (7.15) *and* (7.16) *have the following expressions*

$$\varphi_j = \lambda_j - \frac{1}{d} \sum_{\substack{k=0 \\ k \neq j}}^{d} \psi_k, \qquad j = 0, \dots, d, \tag{7.17}$$

$$\psi_j = \langle \boldsymbol{e}, \boldsymbol{c}_j \rangle, \qquad j = 0, \dots, d, \tag{7.18}$$

where

$$\boldsymbol{e} = [e_0, \dots, e_d]^T.$$

Proof. Without loss of generality, we prove (7.17) for the case $j = 0$. We set

$$I_j(\boldsymbol{e}) = [I_j(e_0), \dots, I_j(e_d)]^T, \qquad j = 0, \dots, d.$$

Since $NN^{-1} = I$, we easily get

$$\langle I_j(\boldsymbol{e}), \boldsymbol{c}_i \rangle = \delta_{ij}. \tag{7.19}$$

Moreover, by (7.5) we get

$$I_j(\lambda_i) = \frac{1}{d}(1 - \delta_{ij}), \qquad i, j = 0, \dots, d. \tag{7.20}$$

The basis function $\varphi_0 \in \mathbb{P}_1^{\text{enr}}(S_d)$ can be represented as

$$\varphi_0 = p_0 + \sum_{i=0}^{d} \beta_i e_i = p_0 + \langle \boldsymbol{e}, \boldsymbol{\beta} \rangle,$$

where $p_0 \in \mathbb{P}_1(S_d)$ and $\boldsymbol{\beta} = [\beta_0, \dots, \beta_d]^T \in \mathbb{R}^{d+1}$. The vanishing conditions (7.9) imply that

$$\delta_{0j} = L_j(\varphi_0) = p_0(\boldsymbol{v}_j), \qquad j = 0, \dots, d,$$

and then $p_0 = \lambda_0$. Hence

$$\varphi_0 = \lambda_0 + \langle \boldsymbol{e}, \boldsymbol{\beta} \rangle. \tag{7.21}$$

By applying the linear functionals I_j, $j = 0, \dots, d$, to both sides of (7.21) we get, by (7.15),

$$0 = I_j(\lambda_0) + \langle \boldsymbol{\beta}, I_j(\boldsymbol{e}) \rangle, \qquad j = 0, \dots, d,$$

or equivalently, in matrix form

$$\boldsymbol{I}(\lambda_0) + N\boldsymbol{\beta} = \boldsymbol{0},$$

where $\boldsymbol{I}(\lambda_0) = [I_0(\lambda_0), \dots, I_d(\lambda_0)]^T$. Then, by taking into account (7.20), we obtain

$$\boldsymbol{\beta} = -\frac{1}{d} \sum_{k=0}^{d} \boldsymbol{c}_k (1 - \delta_{0k}). \tag{7.22}$$

Finally, by substituting (7.22) in (7.21) and by using the bilinearity of the scalar product, we get

$$\varphi_0 = \lambda_0 - \frac{1}{d} \sum_{k=1}^{d} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle.$$

In order to prove (7.17) for $j = 0$, it remains to prove (7.18). To this aim, without loss of generality, we show the validity of (7.18) for $j = 0$. We proceed in analogy to the previous case and then we set

$$\psi_0 = p_0 + \sum_{i=0}^{d} \gamma_i e_i = p_0 + \langle \boldsymbol{e}, \boldsymbol{\gamma} \rangle,$$

where $p_0 \in \mathbb{P}_1(S_d)$ and $\boldsymbol{\gamma} = [\gamma_0, \ldots, \gamma_d]^T \in \mathbb{R}^{d+1}$. The vanishing conditions (7.9) imply that

$$L_j(\psi_0) = \psi_0(\boldsymbol{v}_j) = p_0(\boldsymbol{v}_j) = 0, \qquad j = 0, \ldots, d,$$

and then $p_0 = 0$. Hence

$$\psi_0 = \langle \boldsymbol{e}, \boldsymbol{\gamma} \rangle . \tag{7.23}$$

By applying the linear functionals I_j, $j = 0, \ldots, d$, to both sides of (7.23) we get, by (7.16),

$$\delta_{0j} = \langle I_j(\boldsymbol{e}), \boldsymbol{\gamma} \rangle , \qquad j = 0, \ldots, d,$$

or, equivalently, in matrix form

$$N\boldsymbol{\gamma} = \boldsymbol{u}_0,$$

where $\boldsymbol{u}_0 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{d+1}$. Hence

$$\boldsymbol{\gamma} = \boldsymbol{c}_0. \tag{7.24}$$

Finally, by substituting (7.24) in (7.23), we get the required expression for ψ_0. Similarly, we can prove (7.18) for $j = 1, \ldots, d$ and consequently (7.17) for $j = 0$ is proved. The expression of the other functions can be obtained using symmetry arguments. □

Example 7.1.5. *Let us consider the enrichment functions*

$$e_i = (1 - \lambda_i)^{\gamma_i} \prod_{k=0}^{d} \lambda_k^{\alpha_{i,k}-1}, \qquad i = 0, \ldots, d, \tag{7.25}$$

where $\gamma_i > 0$, $\alpha_{i,i} = 1$ $\alpha_{i,k} > 1$, $i, k = 0, \ldots, d$, $i \neq k$. It is easy to see that the enrichment functions e_i satisfy the vanishing conditions (7.9). Moreover, we get $I_j(e_i) = 0$ for $i \neq j$, and

$$I_i(e_i) = \frac{1}{|F_i|} \int_{F_i} (1 - \lambda_i(\boldsymbol{x}))^{\gamma_i} \prod_{k=0}^{d} \lambda_k^{\alpha_{i,k}-1}(\boldsymbol{x}) d\sigma(\boldsymbol{x}) = \frac{1}{|F_i|} \int_{F_i} \prod_{k=0}^{d} \lambda_k^{\alpha_{i,k}-1}(\boldsymbol{x})\, d\sigma(\boldsymbol{x}).$$

Therefore, by (7.5), for each $i, j = 0, \ldots, d$, we have

$$I_j(e_i) = \frac{1}{|F_j|} \int_{F_j} (1 - \lambda_i(\boldsymbol{x}))^{\gamma_i} \prod_{k=0}^{d} \lambda_k^{\alpha_{i,k}-1}(\boldsymbol{x})\, d\sigma(\boldsymbol{x}) = \frac{(d-1)! \prod_{k=0,k\neq i}^{d} \Gamma(\alpha_{i,k})}{\Gamma(\sum_{k=0,k\neq i}^{d} \alpha_{i,k})} \delta_{ij}. \tag{7.26}$$

Consequently, the matrix (7.11) is a nonsingular diagonal matrix and hence, by Theorem 7.1.2, we can enrich $(S_d, \mathbb{P}_1(S_d), \Sigma_{S_d})$ to the element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$ by using (7.25) as enrichment functions.

Theorem 7.1.6. *The linear approximation operator based on the standard simplicial linear finite element $\mathcal{P}_1(S_d)$, defined in (7.7)*

$$\begin{aligned} \Pi^{\text{lin}} : C(S_d) &\to \mathbb{P}_1(S_d) \\ f &\mapsto \sum_{j=0}^{d} L_j(f) \lambda_j, \end{aligned} \tag{7.27}$$

reproduces linear polynomials and satisfies the interpolation conditions

$$L_j\left(\Pi^{\text{lin}}[f]\right) = L_j(f), \qquad j = 0, \ldots, d. \tag{7.28}$$

Proof. The proof follows from the Lagrange property of the barycentric coordinates, that is $\lambda_i(\boldsymbol{v}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta operator. □

In the following, we denote by E^{lin} the approximation error of the operator Π^{lin}, that is

$$E^{\text{lin}}[f] = f - \Pi^{\text{lin}}[f], \qquad f \in C(S_d). \tag{7.29}$$

7.2 Error estimates

7.2.1 An explicit error representation

Let $\Pi^{\rm enr}$ be the approximation operator defined as

$$\begin{aligned} \Pi^{\rm enr} : C(S_d) &\rightarrow \mathbb{P}_1^{\rm enr}(S_d) \\ f &\mapsto \sum_{j=0}^{d} L_j(f)\varphi_j + \sum_{j=0}^{d} I_j(f)\psi_j, \end{aligned} \tag{7.30}$$

where φ_j, ψ_j, $j = 0, \ldots, d$, are the basis functions introduced in Theorem 7.1.4. The goal of this Section is to provide explicit bounds for the approximation error

$$E^{\rm enr}[f] = f - \Pi^{\rm enr}[f], \tag{7.31}$$

in both L^∞-norm and in L^1-norm. Sharp explicit error bounds in L^1-norm, for some special choices of admissible enrichment functions, are also derived. We set

$$\mathcal{L}_k = \frac{1}{d} \sum_{\substack{j=0 \\ j\neq k}}^{d} L_j, \qquad k = 0, \ldots, d, \tag{7.32}$$

$$\mathcal{E}_k^{\rm tra} = \mathcal{L}_k - I_k, \qquad k = 0, \ldots, d, \tag{7.33}$$

and, in line with Chapter 6, we prove that the approximation error (7.31) can be expressed as the error of the standard simplicial linear finite element plus a second term, which depends on the enrichment functions e_i, $i = 0, \ldots, d$. In fact, we have

Proposition 7.2.1. *Let $f \in C(S_d)$. Then, for any $\boldsymbol{x} \in S_d$, we have*

$$E^{\rm enr}[f](\boldsymbol{x}) = E^{\rm lin}[f](\boldsymbol{x}) + \sum_{k=0}^{d} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle \, \mathcal{E}_k^{\rm tra}(f), \tag{7.34}$$

where $E^{\rm lin}[f]$ is defined in (7.29).

Proof. By the definition of $\Pi^{\rm enr}[f]$ given in (7.30), the approximation error (7.31) can be written as

$$E^{\rm enr}[f] = f - \sum_{j=0}^{d} L_j(f)\varphi_j - \sum_{j=0}^{d} I_j(f)\psi_j.$$

We get, using (7.17), (7.18) and (7.32)

$$\begin{aligned} \sum_{j=0}^{d} L_j(f)\varphi_j &= \sum_{j=0}^{d} L_j(f) \left(\lambda_j - \frac{1}{d} \sum_{\substack{k=0 \\ k\neq j}}^{d} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \right) \\ &= \sum_{j=0}^{d} L_j(f)\lambda_j - \frac{1}{d} \sum_{j=0}^{d} L_j(f) \sum_{k=0}^{d} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \, (1 - \delta_{jk}) \\ &= \sum_{j=0}^{d} L_j(f)\lambda_j - \frac{1}{d} \sum_{k=0}^{d} \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \sum_{j=0}^{d} L_j(f)(1 - \delta_{jk}) \\ &= \sum_{j=0}^{d} L_j(f)\lambda_j - \sum_{k=0}^{d} \mathcal{L}_k(f) \, \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle \end{aligned}$$

and

$$\sum_{j=0}^{d} I_j(f)\psi_j = \sum_{j=0}^{d} I_j(f)\,\langle \boldsymbol{e}, \boldsymbol{c}_j\rangle\,.$$

Therefore, for each $\boldsymbol{x} \in S_d$, we have

$$\begin{aligned} E^{\mathrm{enr}}[f](\boldsymbol{x}) &= f(\boldsymbol{x}) - \sum_{j=0}^{d} L_j(f)\varphi_j(\boldsymbol{x}) - \sum_{j=0}^{d} I_j(f)\psi_j(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) - \left(\sum_{j=0}^{d} L_j(f)\lambda_j(\boldsymbol{x}) - \sum_{k=0}^{d} \mathcal{L}_k(f)\,\langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k\rangle\right) - \sum_{j=0}^{d} I_j(f)\,\langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_j\rangle \\ &= E^{\mathrm{lin}}[f](\boldsymbol{x}) + \sum_{k=0}^{d} \langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k\rangle\,(\mathcal{L}_k(f) - I_k(f))\,, \end{aligned}$$

which is the thesis. □

7.2.2 Error bounds

The decomposition (7.34) is the key result to get the estimate of the error $E^{\mathrm{enr}}[f]$ in the case of a particular class of functions with continuous gradient. As usually, we say that f is continuously differentiable on S_d if it is continuously differentiable on an open set containing S_d. Other useful terminology and notations are clarified in the following.

Definition 7.2.2. *A differentiable function f is said to have a Lipschitz continuous gradient on S_d, if there exists a constant $\rho > 0$ such that*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le \rho\,\|\boldsymbol{x} - \boldsymbol{y}\|_2\,, \qquad \boldsymbol{x}, \boldsymbol{y} \in S_d, \tag{7.35}$$

where $\|\cdot\|_2$ is the L^2-norm in $\mathbb{R}^2$.

By $C^{1,1}(S_d)$ we denote the subclass of all functions f which are continuously differentiable with Lipschitz continuous gradient on S_d. We call the smallest possible ρ such that (7.35) holds *Lipschitz constant* for ∇f and we denote it by $L(\nabla f)$.

If $\boldsymbol{x} \in S_d$ and f is a continuous convex function on S_d, from

$$\boldsymbol{x} = \sum_{i=0}^{d} \lambda_i(\boldsymbol{x})\boldsymbol{v}_i$$

it follows that

$$f(\boldsymbol{x}) \le \sum_{i=0}^{d} \lambda_i(\boldsymbol{x}) f(\boldsymbol{v}_i) = \Pi^{\mathrm{lin}}[f](\boldsymbol{x}), \tag{7.36}$$

that is, the linear interpolation operator Π^{lin} based on the standard simplicial linear finite element $\mathcal{P}_1(S_d)$ approximates every continuous convex function from above. As a consequence, the following bound holds as a particular case of the more general Theorem 6.4.4.

Theorem 7.2.3. *For any $f \in C^{1,1}(S_d)$, we have*

$$\left|E^{\mathrm{lin}}[f](\boldsymbol{x})\right| = \left|f(\boldsymbol{x}) - \Pi^{\mathrm{lin}}[f](\boldsymbol{x})\right| \le \frac{L(\nabla f)}{2}\left(\sum_{i=0}^{d} \lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2\right), \qquad \boldsymbol{x} \in S_d.$$

Equality is attained for all functions of the form

$$f(\boldsymbol{x}) = a(\boldsymbol{x}) + c\,\|\boldsymbol{x}\|_2^2\,,$$

where $c \in \mathbb{R}$ and $a(\boldsymbol{x})$ is any affine function.

With reference to the formula (7.34), it remains to bound $\mathcal{E}_k^{\mathrm{tra}}$ defined in (7.33). To this aim, we need two preliminary lemmas. The fist lemma is a well-known multidimensional integration formula [59], which we recall here for the ease of the reader, without proof. The second lemma follows from the previous one applied to a particular integrand function f and it is used to bound $\mathcal{E}_k^{\mathrm{tra}}$.

Lemma 7.2.4. *Let $\boldsymbol{b}$ be the barycenter of S_d, then*

$$\frac{1}{|S_d|}\int_{S_d} f(\boldsymbol{x})d\boldsymbol{x} = \frac{d+1}{d+2}f(\boldsymbol{b}) + \frac{1}{(d+2)(d+1)}\sum_{i=0}^{d} f(\boldsymbol{v}_i) + R_{S_d}(f), \tag{7.37}$$

with

$$R_{S_d}(f) = 0,$$

for any polynomial f in d variables of total degree at most two.

Lemma 7.2.5. *The following identity holds*

$$\frac{1}{|S_d|}\int_{S_d}\left(\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{x}\right\|_2^2\right)d\boldsymbol{x} = \frac{1}{(d+1)(d+2)}\sum_{0\le i<j\le d}\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2. \tag{7.38}$$

Proof. Since

$$f(\boldsymbol{x}) = \sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{x}\right\|_2^2$$

is a quadratic polynomial in d variables which vanishes at all vertices of S_d, by Lemma 7.2.4, we get

$$\frac{1}{|S_d|}\int_{S_d}\left(\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{x}\right\|_2^2\right)d\boldsymbol{x} = \frac{d+1}{d+2}\left(\frac{1}{d+1}\sum_{i=0}^{d}\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{b}\right\|_2^2\right),$$

where in the last equality, we used the identity

$$\lambda_i(\boldsymbol{b}) = \frac{1}{d+1}, \qquad i=0,\dots,d.$$

It remains to show that

$$\left(\frac{1}{d+1}\sum_{i=0}^{d}\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{b}\right\|_2^2\right) = \frac{1}{(d+1)^2}\sum_{0\le i<j\le d}\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2. \tag{7.39}$$

We notice that

$$\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 = \left\|\boldsymbol{v}_i\right\|_2^2 + \left\|\boldsymbol{v}_j\right\|_2^2 - 2\left\langle\boldsymbol{v}_i,\boldsymbol{v}_j\right\rangle, \qquad i,j=0,\dots,d,$$

therefore, by summing over all i,j both members of previous equality, we get

$$\begin{aligned}\sum_{i,j=0}^{d}\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 &= (d+1)\left(\sum_{i=0}^{d}\left\|\boldsymbol{v}_i\right\|_2^2 + \sum_{j=0}^{d}\left\|\boldsymbol{v}_j\right\|_2^2\right) - 2\left\langle\sum_{i=0}^{d}\boldsymbol{v}_i, \sum_{j=0}^{d}\boldsymbol{v}_j\right\rangle \\ &= 2(d+1)\sum_{i=0}^{d}\left\|\boldsymbol{v}_i\right\|_2^2 - 2(d+1)^2\left\|\boldsymbol{b}\right\|_2^2,\end{aligned}$$

and then

$$2\sum_{0\le i<j\le d}\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2^2 = 2(d+1)^2\left(\frac{1}{d+1}\sum_{i=0}^{d}\left\|\boldsymbol{v}_i\right\|_2^2 - \left\|\boldsymbol{b}\right\|_2^2\right).$$

Dividing both members of the previous equality by $2(d+1)^2$ we get equality (7.39) and then the thesis. □

Now we are able to bound $\mathcal{E}_k^{\mathrm{tra}}(f)$ defined in (7.33).

Theorem 7.2.6. *For any $f \in C^{1,1}(S_d)$, we get*

$$|\mathcal{L}_k(f) - I_k(f)| \leq \frac{L(\nabla f)}{2d(d+1)} \sum_{\substack{0 \leq j < \ell \leq d \\ j,\ell \neq k}}^{d} \|\boldsymbol{v}_j - \boldsymbol{v}_\ell\|_2^2, \qquad k = 0, \ldots, d. \tag{7.40}$$

Equality in (7.40) *is attained for all functions of the form*

$$f(\boldsymbol{x}) = a(\boldsymbol{x}) + c\,\|\boldsymbol{x}\|_2^2, \tag{7.41}$$

where $c \in \mathbb{R}$ and $a(\boldsymbol{x})$ is any affine function.

Proof. Without loss of generality, we prove (7.40) for the particular case $k = 0$. We set

$$\Pi_{F_0}^{\mathrm{lin}}[f](\boldsymbol{x}) = \sum_{j=1}^{d} \lambda_j(\boldsymbol{x}) f(\boldsymbol{v}_j), \qquad \boldsymbol{x} \in F_0. \tag{7.42}$$

Then we have

$$\mathcal{L}_0(f) = \frac{1}{d}\sum_{j=1}^{d} L_j(f) = \frac{1}{d}\sum_{j=1}^{d} f(\boldsymbol{v}_j) = \frac{1}{|F_0|}\int_{F_0} \Pi_{F_0}^{\mathrm{lin}}[f](\boldsymbol{x}) d\sigma(\boldsymbol{x}) \tag{7.43}$$

since the *vertex rule* for the face F_0, see [76, Definition 4.2]

$$\frac{1}{|F_0|}\int_{F_0} f(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \approx \frac{1}{d}\sum_{j=1}^{d} f(\boldsymbol{v}_j) \tag{7.44}$$

is exact for linear polynomials, as it can be easily proven by using equation (7.5). Therefore, we get

$$\mathcal{L}_0(f) - I_0(f) = \frac{1}{|F_0|}\int_{F_0} \left(\Pi_{F_0}^{\mathrm{lin}}[f](\boldsymbol{x}) - f(\boldsymbol{x})\right) d\sigma(\boldsymbol{x}). \tag{7.45}$$

Since $\Pi_{F_0}^{\mathrm{lin}}$ is the restriction of the interpolation operator Π^{lin}, defined in (7.36), to F_0, by Theorem 7.2.3, we have

$$\left|f(\boldsymbol{x}) - \Pi_{F_0}^{\mathrm{lin}}[f](\boldsymbol{x})\right| \leq \frac{L(\nabla f)}{2}\left(\sum_{i=1}^{d} \lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2\right), \qquad \boldsymbol{x} \in F_0. \tag{7.46}$$

Then, by (7.46) and (7.45) we get

$$\begin{aligned}
|\mathcal{L}_0(f) - I_0(f)| &= \left|\frac{1}{|F_0|}\int_{F_0} \left(\Pi_{F_0}^{\mathrm{lin}}[f](\boldsymbol{x}) - f(\boldsymbol{x})\right) d\sigma(\boldsymbol{x})\right| \\
&\leq \frac{1}{|F_0|}\int_{F_0} \left|\Pi_{F_0}^{\mathrm{lin}}[f](\boldsymbol{x}) - f(\boldsymbol{x})\right| d\sigma(\boldsymbol{x}) \\
&\leq \frac{L(\nabla f)}{2}\frac{1}{|F_0|}\int_{F_0}\left(\sum_{i=1}^{d} \lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2\right) d\sigma(\boldsymbol{x}).
\end{aligned}$$

The desired result now follows by Lemma 7.2.5. Finally, simple computations show that the equality in (7.40) is attained for all functions of the form (7.41). □

A pointwise error bound for the approximation operator Π^{enr} can be easily obtained by combining Proposition 7.2.1, Theorem 7.2.3 and Theorem 7.2.6. In fact we have

Theorem 7.2.7. *For any $f \in C^{1,1}(S_d)$, the following explicit error estimate holds for any $\boldsymbol{x} \in S_d$*

$$|E^{\mathrm{enr}}[f](\boldsymbol{x})| = |f(\boldsymbol{x}) - \Pi^{\mathrm{enr}}[f](\boldsymbol{x})| \leq \frac{L(\nabla f)}{2}\left(\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2 + u_{enr}(\boldsymbol{x})\right), \tag{7.47}$$

where

$$u_{enr}(\boldsymbol{x}) = \frac{1}{d(d+1)}\sum_{k=0}^{d}\left(|\langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k\rangle| \sum_{\substack{0\leq i<j\leq d\\ i,j\neq k}}^{d} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2\right). \tag{7.48}$$

7.2.3 The L^∞ error estimate

The pointwise error bound given in Theorem 7.2.7 is the key result to get a bound in L^∞-norm

$$\|f\|_\infty = \max_{\boldsymbol{x}\in S_d} |f(\boldsymbol{x})|$$

of the error $E^{\mathrm{enr}}[f]$ defined in (7.31), in terms of the barycenter $\boldsymbol{b}$, the circumcenter $\boldsymbol{c}$ and the circumradius R of S_d [20]. As a corollary, we estimate $\|E^{\mathrm{enr}}[f]\|_\infty$ through the diameter of S_d. With reference to (7.47), we use the following lemma to bound $\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2$ for any $\boldsymbol{x} \in S_d$, $d \geq 1$.

Lemma 7.2.8. *For any $d \geq 2$, the following identity holds*

$$\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2 = R^2 - \|\boldsymbol{x}-\boldsymbol{c}\|_2^2 \tag{7.49}$$

and then

$$\sup_{\boldsymbol{x}\in S_d}\left(\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2\right) = R^2 - \min_{\boldsymbol{x}\in S_d}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2\,. \tag{7.50}$$

Proof. By the bilinearity of the inner product and by the identity $\boldsymbol{x} = \sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\boldsymbol{v}_i$, we have

$$\begin{aligned}
\sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\,\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2 &= \sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\,\langle \boldsymbol{v}_i - \boldsymbol{c} + \boldsymbol{c}, \boldsymbol{v}_i - \boldsymbol{c} + \boldsymbol{c}\rangle - \langle \boldsymbol{x} - \boldsymbol{c} + \boldsymbol{c}, \boldsymbol{x} - \boldsymbol{c} + \boldsymbol{c}\rangle \\
&= \sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\left(\|\boldsymbol{v}_i - \boldsymbol{c}\|_2^2 + 2\,\langle \boldsymbol{v}_i, \boldsymbol{c}\rangle - \|\boldsymbol{c}\|_2^2\right) - \left(\|\boldsymbol{x}-\boldsymbol{c}\|_2^2 + 2\,\langle \boldsymbol{x}, \boldsymbol{c}\rangle - \|\boldsymbol{c}\|_2^2\right) \\
&= R^2 + 2\left\langle \sum_{i=0}^{d}\lambda_i(\boldsymbol{x})\boldsymbol{v}_i, \boldsymbol{c}\right\rangle - \|\boldsymbol{x}-\boldsymbol{c}\|_2^2 - 2\,\langle \boldsymbol{x}, \boldsymbol{c}\rangle \\
&= R^2 - \|\boldsymbol{x}-\boldsymbol{c}\|_2^2\,.
\end{aligned}$$

Equation (7.50) follows directly from (7.49). □

In order to bound the L^∞-norm of $u_{enr}(\boldsymbol{x})$, whose expression is given in (7.48), we set

$$e_{\max} = \max_{i=0,\ldots,d}\|\langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_i\rangle\|_\infty \tag{7.51}$$

and consequently

$$|u_{enr}(\boldsymbol{x})| \leq \frac{e_{\max}}{d(d+1)}\sum_{k=0}^{d}\sum_{\substack{0\leq i<j\leq d\\ i,j\neq k}}\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2, \qquad \boldsymbol{x}\in S_d.$$

By applying equation (7.38) of Lemma 7.2.5 to the face F_k and by summing over all k from 0 to d, we get

$$\frac{1}{d(d+1)} \sum_{k=0}^{d} \sum_{\substack{0 \le i<j \le d \\ i,j \neq k}} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 = \sum_{k=0}^{d} \frac{1}{|F_k|} \int_{F_k} \left(\sum_{i=0}^{d} \lambda_i(\boldsymbol{x}) \|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2 \right) d\sigma(\boldsymbol{x}). \tag{7.52}$$

The following lemma allows us to express the right-hand side member of (7.52) in terms of the barycenter $\boldsymbol{b}$, the circumcenter $\boldsymbol{c}$ and the circumradius R of S_d.

Lemma 7.2.9. *For each $d \geq 2$, the following identity holds*

$$\sum_{k=0}^{d} \frac{1}{|F_k|} \int_{F_k} \left(\sum_{i=0}^{d} \lambda_i(\boldsymbol{x}) \|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2 \right) d\sigma(\boldsymbol{x}) = \frac{d^2-1}{d} \left(R^2 - \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 \right), \tag{7.53}$$

where $\boldsymbol{b}$, $\boldsymbol{c}$ and R are the barycenter, the circumcenter and the circumradius of S_d, respectively.

Proof. We make use of the well-known Green's formula [55, Ch 7]

$$\int_{S_d} f(\boldsymbol{x}) \Delta u(\boldsymbol{x}) d\boldsymbol{x} + \int_{S_d} \langle \nabla f(\boldsymbol{x}), \nabla u(\boldsymbol{x}) \rangle \, d\boldsymbol{x} = \sum_{k=0}^{d} \int_{F_k} f(\boldsymbol{x}) \, \langle \boldsymbol{n}_k(\boldsymbol{x}), \nabla u(\boldsymbol{x}) \rangle \, d\sigma(\boldsymbol{x}), \tag{7.54}$$

where $f \in C^1(S_d)$, $u \in C^2(S_d)$ and $\boldsymbol{n}_k(\boldsymbol{x})$ is the unit normal vector to the boundary of F_k at the point $\boldsymbol{x}$. We set

$$u(\boldsymbol{x}) = \frac{\|\boldsymbol{x} - \boldsymbol{b}\|_2^2}{2},$$

then

$$\nabla u(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{b}, \qquad \Delta u(\boldsymbol{x}) = d,$$

and Green's formula (7.54) becomes

$$d \int_{S_d} f(\boldsymbol{x}) d\boldsymbol{x} + \int_{S_d} \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{b} \rangle \, d\boldsymbol{x} = \sum_{k=0}^{d} \int_{F_k} f(\boldsymbol{x}) \, \langle \boldsymbol{n}_k(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{b} \rangle \, d\sigma(\boldsymbol{x}). \tag{7.55}$$

Dividing by d both members of (7.55) and by using [55, Thm. 7.10.4], we get

$$\int_{S_d} f(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{d} \int_{S_d} \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{b} \rangle \, d\boldsymbol{x} = \frac{|S_d|}{d+1} \sum_{k=0}^{d} \frac{1}{|F_k|} \int_{F_k} f(\boldsymbol{x}) d\sigma(\boldsymbol{x}),$$

or, equivalently

$$\frac{1}{|S_d|} \int_{S_d} f(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{d\,|S_d|} \int_{S_d} \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{b} \rangle \, d\boldsymbol{x} = \frac{1}{d+1} \sum_{k=0}^{d} \frac{1}{|F_k|} \int_{F_k} f(\boldsymbol{x}) d\sigma(\boldsymbol{x}). \tag{7.56}$$

Now we set

$$f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{c}\|_2^2,$$

then

$$\nabla f(\boldsymbol{x}) = 2(\boldsymbol{x} - \boldsymbol{c}),$$

and the equation (7.56) becomes

$$\frac{1}{|S_d|} \int_{S_d} \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 \, d\boldsymbol{x} + \frac{2}{d\,|S_d|} \int_{S_d} \langle \boldsymbol{x} - \boldsymbol{c}, \boldsymbol{x} - \boldsymbol{b} \rangle \, d\boldsymbol{x} = \frac{1}{d+1} \sum_{k=0}^{d} \frac{1}{|F_k|} \int_{F_k} \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 \, d\sigma(\boldsymbol{x}). \tag{7.57}$$

By Lemma 7.2.4

$$\frac{1}{|S_d|} \int_{S_d} \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 \, d\boldsymbol{x} = \frac{d+1}{d+2} \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 + \frac{1}{d+2} R^2, \tag{7.58}$$

moreover

$$\begin{aligned}\frac{1}{|S_d|}\int_{S_d}\langle \boldsymbol{x}-\boldsymbol{c},\boldsymbol{x}-\boldsymbol{b}\rangle\,d\boldsymbol{x} &= \frac{1}{|S_d|}\int_{S_d}\langle \boldsymbol{x}-\boldsymbol{c},\boldsymbol{x}-\boldsymbol{c}+\boldsymbol{c}-\boldsymbol{b}\rangle\,d\boldsymbol{x}\\ &= \frac{1}{|S_d|}\int_{S_d}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2\,d\boldsymbol{x}-\frac{1}{|S_d|}\int_{S_d}\langle \boldsymbol{x}-\boldsymbol{c},\boldsymbol{b}-\boldsymbol{c}\rangle\,d\boldsymbol{x}\\ &= \frac{d+1}{d+2}\|\boldsymbol{b}-\boldsymbol{c}\|_2^2+\frac{1}{d+2}R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2, \end{aligned} \tag{7.59}$$

since equation (7.58) and the equality

$$\frac{1}{|S_d|}\int_{S_d}\langle \boldsymbol{x}-\boldsymbol{c},\boldsymbol{b}-\boldsymbol{c}\rangle\,d\boldsymbol{x}=\|\boldsymbol{b}-\boldsymbol{c}\|_2^2$$

which follows by the midpoint cubature formula [94, Ch 8.8]. By using (7.58) and (7.59), after easy computations, equation (7.57) becomes

$$\frac{1}{d}R^2+\frac{d-1}{d}\|\boldsymbol{b}-\boldsymbol{c}\|_2^2=\frac{1}{d+1}\sum_{k=0}^{d}\frac{1}{|F_k|}\int_{F_k}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2\,d\sigma(\boldsymbol{x})$$

or, equivalently,

$$\frac{1}{d+1}\sum_{k=0}^{d}\frac{1}{|F_k|}\int_{F_k}\left(R^2-\|\boldsymbol{x}-\boldsymbol{c}\|_2^2\right)d\sigma(\boldsymbol{x})=\frac{d-1}{d}\left(R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2\right). \tag{7.60}$$

Finally, by multiplying both sides of (7.60) by $d+1$, equation (7.53) is a direct consequence of Lemma 7.2.8. □

By combining (7.52) and (7.53) we get the following result.

Lemma 7.2.10. *For each $d\geq 2$, we have*

$$\frac{1}{d(d+1)}\sum_{k=0}^{d}\sum_{\substack{0\leq i<j\leq d\\ i,j\neq k}}\|\boldsymbol{v}_i-\boldsymbol{v}_j\|_2^2=\frac{d^2-1}{d}\left(R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2\right). \tag{7.61}$$

Finally, by Lemma 7.2.8 and Lemma 7.2.10, from Theorem 7.2.7, the following result follows.

Theorem 7.2.11. *For any $f\in C^{1,1}(S_d)$, the following explicit error estimate holds*

$$\|f-\Pi^{\mathrm{enr}}[f]\|_\infty\leq\frac{L(\nabla f)}{2}\left(R^2-\min_{\boldsymbol{x}\in S_d}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2+\frac{d^2-1}{d}\left(R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2\right)e_{\max}\right). \tag{7.62}$$

Let

$$h=\sup_{\boldsymbol{v},\boldsymbol{w}\in S_d}\|\boldsymbol{v}-\boldsymbol{w}\|_2$$

be the diameter of S_d. By Pythagorean Theorem, we get

$$R^2=\min_{\boldsymbol{x}\in S_d}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2+\frac{h^2}{4},$$

and then

$$R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2\leq R^2-\min_{\boldsymbol{x}\in S_d}\|\boldsymbol{x}-\boldsymbol{c}\|_2^2=\frac{h^2}{4}. \tag{7.63}$$

Consequently, the right-hand side of equation (7.62) can be bounded by

$$\frac{L(\nabla f)}{8}\left(1+\frac{d^2-1}{d}e_{\max}\right)h^2$$

and the following estimate holds.

Corollary 7.2.12. *For any $f\in C^{1,1}(S_d)$, the following explicit error estimate holds*

$$\|f-\Pi^{\mathrm{enr}}[f]\|_\infty\leq\frac{L(\nabla f)}{8}\left(1+\frac{d^2-1}{d}e_{\max}\right)h^2. \tag{7.64}$$

7.2.4 The L^1 error estimate

The pointwise error bound given in Theorem 7.2.7 is the key result to get a bound in L^1-norm

$$\|f\|_1 = \int_{S_d} |f(\boldsymbol{x})|\, d\boldsymbol{x}$$

of the error $E^{\mathrm{enr}}[f]$ defined in (7.31) in terms of the diameter h and the volume $|S_d|$ of S_d.

Theorem 7.2.13. *For any $f \in C^{1,1}(S_d)$, we get*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_1 \leq \frac{L(\nabla f)}{8} \left(\frac{d+1}{d+2} + \frac{d^2-1}{d} e'_{\max} \right) h^2\, |S_d|, \tag{7.65}$$

where we set

$$e'_{\max} = \frac{1}{|S_d|} \max_{i=0,\ldots,d} \|\langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_i \rangle\|_1 \,. \tag{7.66}$$

Proof. With reference to (7.47), we use Lemma 7.2.8 and Lemma 7.2.4 to get

$$\begin{aligned} \int_{S_d} \left(\sum_{i=0}^{d} \lambda_i(\boldsymbol{x}) \|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{x}\|_2^2 \right) d\boldsymbol{x} &= \int_{S_d} \left(R^2 - \|\boldsymbol{x} - \boldsymbol{c}\|_2^2 \right) d\boldsymbol{x} \\ &= \frac{d+1}{d+2} \left(R^2 - \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 \right) |S_d|. \end{aligned} \tag{7.67}$$

On the other hand, from (7.48), by applying Lemma 7.2.10, we have

$$\begin{aligned} \int_{S_d} |u_{enr}(\boldsymbol{x})|\, d\boldsymbol{x} &\leq \max_{k=0,\ldots,d} \int_{S_d} |\langle \boldsymbol{e}(\boldsymbol{x}), \boldsymbol{c}_k \rangle|\, d\boldsymbol{x} \left(\frac{1}{d(d+1)} \sum_{k=0}^{d} \sum_{\substack{0 \leq i < j \leq d \\ i,j \neq k}}^{d} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 \right) \\ &= \max_{k=0,\ldots,d} \|\langle \boldsymbol{c}_k, \boldsymbol{e} \rangle\|_1 \, \frac{d^2-1}{d} \left(R^2 - \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 \right). \end{aligned} \tag{7.68}$$

Consequently

$$\int_{S_d} |f(\boldsymbol{x}) - \Pi^{\mathrm{enr}}[f](\boldsymbol{x})|\, d\boldsymbol{x} \leq \frac{L(\nabla f)}{2} \left(\frac{d+1}{d+2} + \frac{d^2-1}{d} e'_{\max} \right) \left(R^2 - \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 \right) |S_d| \,. \tag{7.69}$$

The required result follows by recalling that $R^2 - \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 \leq \frac{h^2}{4}$. □

Now we bound $|S_d|$ in (7.65) in order to show that the L^1 error bound for $E^{\mathrm{enr}}[f]$ is proportional to the $(d+2)$-th power of the diameter of S_d.

Theorem 7.2.14. *For any $f \in C^{1,1}(S_d)$, we have*

$$\|f - \Pi^{\mathrm{enr}}[f]\|_1 \leq \frac{L(\nabla f)}{2^{d+3} d!} \sqrt{\frac{(d+1)^{d+1}}{d^d}} \left(\frac{d+1}{d+2} + \frac{d^2-1}{d} e'_{\max} \right) h^{d+2}. \tag{7.70}$$

Proof. By using (7.67) and (7.38) we get

$$(d+1)^2 \left(R^2 - \|\boldsymbol{b} - \boldsymbol{c}\|_2^2 \right) = \sum_{0 \leq i < j \leq d} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 \,. \tag{7.71}$$

By recalling the arithmetic-geometric mean inequality

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \cdots x_n}, \qquad n \in \mathbb{N}, \qquad x_1, \ldots, x_n \geq 0,$$

from equation (7.71), we get

$$(d+1)^2\left(R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2\right)\geq\frac{d(d+1)}{2}\left(\prod_{0\leq i<j\leq d}\|\boldsymbol{v}_i-\boldsymbol{v}_j\|_2^2\right)^{\frac{2}{d(d+1)}}. \tag{7.72}$$

Moreover, the following inequality holds, see [73]

$$\prod_{0\leq i<j\leq d}\|\boldsymbol{v}_i-\boldsymbol{v}_j\|_2^{\frac{2}{d+1}}\geq d!\,|S_d|\sqrt{\frac{2^d}{d+1}}. \tag{7.73}$$

From (7.72) and (7.73) we get

$$\begin{aligned}R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2 &\geq \frac{d}{2(d+1)}\left(\prod_{0\leq i<j\leq d}\|\boldsymbol{v}_i-\boldsymbol{v}_j\|_2^{\frac{2}{d+1}}\right)^{\frac{2}{d}}\\ &\geq \frac{d}{2(d+1)}\left(\frac{2^d}{d+1}\right)^{\frac{1}{d}}(d!)^{\frac{2}{d}}\,|S_d|^{\frac{2}{d}}\\ &= \frac{d}{(d+1)^{\frac{d+1}{d}}}(d!)^{\frac{2}{d}}\,|S_d|^{\frac{2}{d}},\end{aligned}$$

and then

$$|S_d|\leq\frac{1}{d!}\sqrt{\frac{(d+1)^{d+1}}{d^d}}\left(R^2-\|\boldsymbol{b}-\boldsymbol{c}\|_2^2\right)^{\frac{d}{2}}.$$

The thesis follows. □

7.2.5 A practical example

In the following, we go back on the enrichment functions

$$e_i=(1-\lambda_i)^{\gamma_i}\prod_{k=0}^{d}\lambda_k^{\alpha_{i,k}-1},\qquad i=0,\ldots,d, \tag{7.74}$$

of Example 7.1.5, where $\gamma_i>0$, $\alpha_{i,i}=1$ $\alpha_{i,k}>1$, $i,k=0,\ldots,d$, $i\neq k$. We determine the behavior of the bound of the error (7.70) in dependence of all parameters γ_i, $\alpha_{i,k}$. To this aim we consider the standard simplex in $\mathbb{R}^d$

$$\widehat{S}_d=\left\{\widehat{\boldsymbol{x}}=(\widehat{x}_1,\ldots,\widehat{x}_d)\in\mathbb{R}^d,\,\widehat{x}_i\geq 0\,:\,i=1,\ldots,d,\,\sum_{j=1}^{d}\widehat{x}_j\leq 1\right\},$$

with barycentric coordinates

$$\widehat{\lambda}_0(\widehat{\boldsymbol{x}})=1-\sum_{j=1}^{d}\widehat{x}_j,\qquad\widehat{\lambda}_i(\widehat{\boldsymbol{x}})=\widehat{x}_i,\qquad i=1,\ldots,d. \tag{7.75}$$

Lemma 7.2.15. *For each $i=0,\ldots,d$, the following equality holds*

$$\int_{\widehat{S}_d}(1-\widehat{\lambda}_i)^{\gamma_i}\prod_{k=0}^{d}\widehat{\lambda}_k^{\alpha_{i,k}-1}d\widehat{\boldsymbol{x}}=\frac{\prod_{k=0}^{d}\Gamma(\alpha_{i,k})}{\Gamma(\sum_{k=0,k\neq i}^{d}\alpha_{i,k})}\frac{\Gamma(\gamma_i+\sum_{k=0,k\neq i}^{d}\alpha_{i,k})}{\Gamma(\gamma_i+\sum_{k=0}^{d}\alpha_{i,k})}, \tag{7.76}$$

where $\gamma_i\geq 0$, and $\alpha_{i,k}\geq 1$, $i,k=0,\ldots,d$.

Proof. First we prove (7.76) for $i \neq 0$. Without loss of generality we can assume $i = 1$. By using (7.75), we get

$$\begin{aligned}
&\int_{\widehat{S}_d} (1-\widehat{\lambda}_1)^{\gamma_1} \prod_{k=0}^{d} \widehat{\lambda}_k^{\alpha_{1,k}-1} d\widehat{\boldsymbol{x}} \\
&= \int_{\widehat{S}_d} (1-\widehat{x}_1)^{\gamma_1} \widehat{x}_1^{\alpha_{1,1}-1} \cdots \widehat{x}_d^{\alpha_{1,d}-1} (1-\widehat{x}_1-\cdots-\widehat{x}_d)^{\alpha_{1,0}-1} d\widehat{\boldsymbol{x}} \\
&= \int_0^1 (1-\widehat{x}_1)^{\gamma_1} \widehat{x}_1^{\alpha_{1,1}-1} \cdots \int_0^{1-\widehat{x}_1-\cdots-\widehat{x}_{d-1}} \widehat{x}_d^{\alpha_{1,d}-1} (1-\widehat{x}_1-\cdots-\widehat{x}_d)^{\alpha_{1,0}-1} d\widehat{\boldsymbol{x}}. \qquad (7.77)
\end{aligned}$$

We set

$$w = \frac{\widehat{x}_d}{1-\widehat{x}_1-\cdots-\widehat{x}_{d-1}}$$

and then, we get

$$\begin{aligned}
&\int_0^{1-\widehat{x}_1-\cdots-\widehat{x}_{d-1}} \widehat{x}_d^{\alpha_{1,d}-1} (1-\widehat{x}_1-\cdots-\widehat{x}_d)^{\alpha_{1,0}-1} d\widehat{x}_d \\
&\quad = (1-\widehat{x}_1-\cdots-\widehat{x}_{d-1})^{\alpha_{1,0}+\alpha_{1,d}-1} \int_0^1 w^{\alpha_{1,d}-1} (1-w)^{\alpha_{1,0}-1} dw \\
&\quad = (1-\widehat{x}_1-\cdots-\widehat{x}_{d-1})^{\alpha_{1,0}+\alpha_{1,d}-1} \frac{\Gamma(\alpha_{1,d})\Gamma(\alpha_{1,0})}{\Gamma(\alpha_{1,d}+\alpha_{1,0})}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\int_{\widehat{S}_d} (1-\widehat{\lambda}_1)^{\gamma_1} \prod_{j=0}^{d} \widehat{\lambda}_j^{\alpha_{1,j}-1} d\widehat{\boldsymbol{x}} = \frac{\Gamma(\alpha_{1,d})\Gamma(\alpha_{1,0})}{\Gamma(\alpha_{1,d}+\alpha_{1,0})} \int_0^1 (1-\widehat{x}_1)^{\gamma_1} \widehat{x}_1^{\alpha_{1,1}-1} \cdots \\
\cdots \int_0^{1-\widehat{x}_1-\cdots-\widehat{x}_{d-2}} \widehat{x}_{d-1}^{\alpha_{1,d-1}-1} (1-\widehat{x}_1-\cdots-\widehat{x}_{d-1})^{\alpha_{1,0}+\alpha_{1,d}-1} d\widehat{x}_1 \ldots d\widehat{x}_{d-1}.
\end{aligned}$$

Similar substitutions can be applied to the integral

$$\int_0^{1-\widehat{x}_1-\cdots-\widehat{x}_{d-2}} \widehat{x}_{d-1}^{\alpha_{1,d-1}-1} (1-\widehat{x}_1-\cdots-\widehat{x}_{d-1})^{\alpha_{1,0}+\alpha_{1,d}-1} d\widehat{x}_{d-1}$$

and subsequent ones, in order to get, after some rearrangement, equation (7.76) for $i = 1$. Now we prove equation (7.76) for $i = 0$. To this aim, we introduce the linear transformations

$$\begin{aligned}
&\pi : \widehat{S}_d \to \widehat{S}_d, \quad \pi(\widehat{x}_1, \widehat{x}_2 \ldots, \widehat{x}_d) = (\widehat{u}_1, \widehat{u}_2 \ldots, \widehat{u}_d), \\
&\theta : \widehat{S}_d \to \widehat{S}_d, \quad \theta(\widehat{x}_1, \widehat{x}_2 \ldots, \widehat{x}_d) = (\widehat{x}_2, \widehat{x}_1 \ldots, \widehat{x}_d),
\end{aligned}$$

where

$$\widehat{u}_1 = 1 - \sum_{j=1}^{d} \widehat{x}_j, \quad \widehat{u}_j = \widehat{x}_j, \qquad j = 2, \ldots, d.$$

The Jacobian determinant of the change of variables $\theta \circ \pi$ is 1, then we get

$$\begin{aligned}
&\int_{\widehat{S}_d} (\widehat{x}_1 + \cdots + \widehat{x}_d)^{\gamma_0} (1-\widehat{x}_1-\cdots-\widehat{x}_d)^{\alpha_{0,0}-1} \widehat{x}_1^{\alpha_{0,1}-1} \ldots \widehat{x}_d^{\alpha_{0,d}-1} d\widehat{\boldsymbol{x}} \\
&\quad = \int_{\widehat{S}_d} (1-\widehat{u}_1)^{\gamma_0} \widehat{u}_1^{\alpha_{0,0}-1} \widehat{u}_2^{\alpha_{0,2}-1} \ldots (1-\widehat{u}_1-\cdots-\widehat{u}_d)^{\alpha_{0,1}-1} d\widehat{\boldsymbol{u}}.
\end{aligned}$$

The result now follows from the case $i = 1$. □

Remark 7.2.16. *We notice that, by setting $\gamma_i = 0$, $i = 0, \dots, d$, equation (7.76) reduces to the integration formula (7.2).*

Theorem 7.2.17. *Let e_i, $i = 0, \dots, d$, be the enrichment functions defined in (7.74). Then, for any $f \in C^{1,1}(S_d)$, we get*

$$\|f - \Pi^{\text{enr}}[f]\|_1 \leq \frac{L(\nabla f)}{2^{d+3} d!} \sqrt{\frac{(d+1)^{d+1}}{d^d}} \left(\frac{d+1}{d+2} + \frac{d^2-1}{\mu} \right) h^{d+2}, \tag{7.78}$$

where

$$\mu = \min_{i=0,\dots,d} \left(\gamma_i + \sum_{k=0, k\neq i}^{d} \alpha_{i,k} \right). \tag{7.79}$$

Proof. With reference to the error bound (7.65), we compute $e'_{\max}$ defined in (7.66) for the particular case of the enrichment functions defined in (7.74). In particular, we prove that

$$e'_{\max} = \frac{d}{\mu}.$$

From equation (7.26), the matrix N introduced in (7.11) is

$$N = \begin{bmatrix} E_0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & E_d \end{bmatrix},$$

where

$$E_i = I_i(e_i) = \frac{(d-1)! \prod_{k=0, k\neq i}^{d} \Gamma(\alpha_{i,k})}{\Gamma(\sum_{k=0, k\neq i}^{d} \alpha_{i,k})}, \qquad i = 0, \dots, d. \tag{7.80}$$

Therefore

$$N^{-1} = [\boldsymbol{c}_0 \dots \boldsymbol{c}_d] = \begin{bmatrix} \frac{1}{E_0} & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & \frac{1}{E_d} \end{bmatrix},$$

and consequently

$$e'_{\max} = \frac{1}{|S_d|} \max_{i=0,\dots,d} \frac{1}{|E_i|} \int_{S_d} (1 - \lambda_i(\boldsymbol{x}))^{\gamma_i} \prod_{k=0}^{d} \lambda_k^{\alpha_{i,k}-1}(\boldsymbol{x}) d\boldsymbol{x}. \tag{7.81}$$

We use the change of variables $\widehat{\boldsymbol{x}} \mapsto \sum_{i=0}^{d} \widehat{\lambda}_i(\widehat{\boldsymbol{x}}) \boldsymbol{v}_i$, $\widehat{\boldsymbol{x}} \in \widehat{S}_d$ to compute the integrals at the right-hand side of (7.81) and we get

$$e'_{\max} = \frac{1}{\left|\widehat{S}_d\right|} \max_{i=0,\dots,d} \frac{1}{|E_i|} \int_{\widehat{S}_d} (1 - \widehat{\lambda}_i(\boldsymbol{x}))^{\gamma_j} \prod_{k=0}^{d} \widehat{\lambda}_k^{\alpha_{i,k}-1}(\boldsymbol{x}) d\widehat{\boldsymbol{x}}.$$

The thesis follows by using equation (7.76) and (7.80). □

Remark 7.2.18. *We notice that when $d = 2$, the results introduced here are equivalent to the results presented in Chapter 6. Then the enrichment strategy introduced here, in this sense, generalizes the enrichment strategy introduced in Chapter 6.*

7.2.6 Global approximation operator

Let $X_n = \{\boldsymbol{x}_i : i = 1\ldots, n\}$ be a set of n scattered points in $\mathbb{R}^d$ and let $\mathcal{S}_m = \{S_j : j = 1, \ldots, m\}$ be a partition into simplices, with the points of X_n as vertices, of the convex hull of X_n, $\mathcal{C} = \text{conv}(X_n)$. For any $f \in C^{1,1}(\mathcal{C})$ and $\boldsymbol{x} \in \mathcal{C}$, we define

$$\Pi^{\text{enr}}_{\mathcal{S}_m}[f](\boldsymbol{x}) = \Pi^{\text{enr}}[f, S_j](\boldsymbol{x}), \text{ if } \boldsymbol{x} \in S_j, \qquad j = 1, \ldots, m, \tag{7.82}$$

where $\Pi^{\text{enr}}[f, S_j]$ is the approximation operator based on the triple $(S_j, \mathbb{P}^{\text{enr}}_1(S_j), \Sigma^{\text{enr}}_{S_j})$ introduced in (7.30). For the enrichment functions (7.74), Theorem 7.2.17 gives us the following global estimate for any $f \in C^{1,1}(\mathcal{C})$

$$\begin{aligned} \left\|f - \Pi^{\text{enr}}_{\mathcal{S}_m}[f]\right\|_1 &= \int_{\mathcal{C}} \left|f(\boldsymbol{x}) - \Pi^{\text{enr}}_{\mathcal{S}_m}[f](\boldsymbol{x})\right| d\boldsymbol{x} = \sum_{S_j \in \mathcal{S}_m} \int_{S_j} |f(\boldsymbol{x}) - \Pi^{\text{enr}}[f, S_j](\boldsymbol{x})| \, d\boldsymbol{x} \\ &\leq \frac{L(\nabla f)}{2^{d+3} d!} \sqrt{\frac{(d+1)^{d+1}}{d^d}} \left(\frac{d+1}{d+2} + \frac{d^2-1}{\mu}\right) \sum_{S_j \in \mathcal{S}_m} h^{d+2}_{S_j}, \end{aligned} \tag{7.83}$$

where h_{S_j} is the diameter of the simplex S_j. Denoting by

$$E_1(\mathcal{S}_m) = \sum_{S_j \in \mathcal{S}_m} h^{d+2}_{S_j},$$

we find that the global error bound (7.83) is proportional to $E_1(\mathcal{S}_m)$, by a constant factor which is independent on $\mathcal{S}_m$. By using the optimality results of Delaunay triangulation, which can be found in [41] and [53], it is possible to prove the following result.

Theorem 7.2.19. *Let X_n be a set of n distinct and non-collinear points in $\mathbb{R}^d$ and let $\mathcal{S}_m$ be any partition of the convex hull of X_n into simplices with the points of X_n as vertices. Then $E_1(\mathcal{S}_m)$ achieves its minimum if and only if $\mathcal{S}_m$ is the Delaunay triangulation in $\mathbb{R}^d$.*

7.3 Numerical experiments

In this Section, we numerically prove the accuracy of the proposed method by several examples. The numerical experiments are performed by using `MatLab` software. In particular, the command `integral2` is used in order to compute the integrals of all considered functions over all faces of the simplex S_d.

We consider $d = 3$ and the following enrichments of the standard simplicial linear finite element

- $\mathcal{E}_1 = \left\{ e_i = (1 - \lambda_i) \prod_{\substack{j=0 \\ j \neq i}}^{3} \lambda_j \, : \, i = 0, \ldots, 3 \right\}$,
- $\mathcal{E}_2 = \left\{ e_i = \prod_{\substack{j=0 \\ j \neq i}}^{3} \sin(\lambda_j) \, : \, i = 0, \ldots, 3 \right\}$,
- $\mathcal{E}_3 = \left\{ e_i = e^{-\lambda_i} \prod_{\substack{j=0 \\ j \neq i}}^{3} (1 - e^{-\lambda_j}) \, : \, i = 0, \ldots, 3 \right\}$.

We compute the approximation error in L^1-norm for the test functions

$$f_1(x, y, z) = \sin(xyz), \quad f_2(x, y, z) = \sin(\pi x) \sin(\pi y) \sin(\pi z).$$

Furthermore, we perform a direct comparison between the approximation produced by the enriched simplicial finite element and that produced by the simplicial finite element. In each experiments we

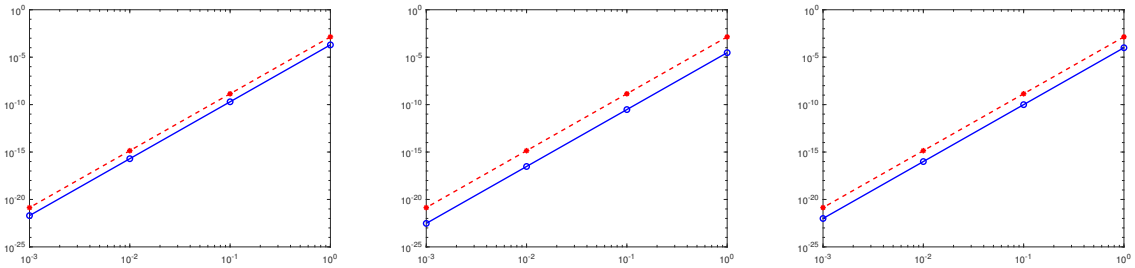


Figure 7.1: Loglog plot of the approximation errors, in L^1-norm, of the function $f_1(x, y, z) = \sin(xyz)$ obtained by using the enriched finite elements $\mathcal{E}_1$ (left), $\mathcal{E}_2$ (center), $\mathcal{E}_3$(right) (blue line) compared with that produced by the simplicial finite element (red dashed line), where the diameter of the simplex is $h = 10^{-k}$, $k = 0, 1, 2, 3$.

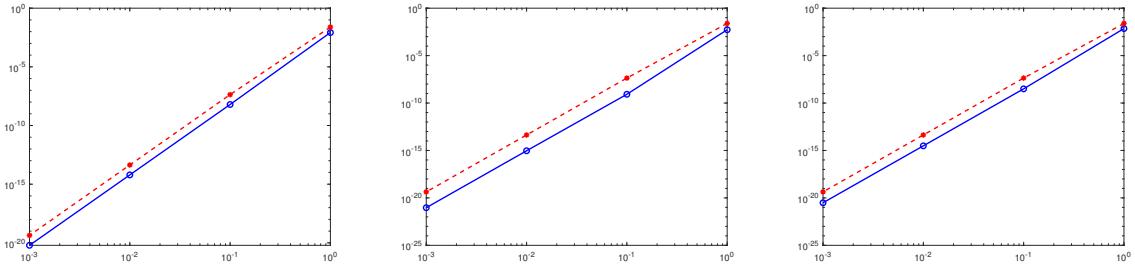


Figure 7.2: Loglog plot of the approximation errors, in L^1-norm, of the function $f_2(x, y, z) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$, obtained by using the enriched finite elements $\mathcal{E}_1$ (left), $\mathcal{E}_2$ (center), $\mathcal{E}_3$ (right) (blue line) compared with that produced by the simplicial finite element (red dashed line), where the diameter of the simplex is $h = 10^{-k}$, $k = 0, 1, 2, 3$.

consider a simplex of diameter $h = 10^{-k}$, $k = 0, 1, 2, 3$. The results of the experiments are represented in Figure 7.1 and in Figure 7.2.

For all examples, it is possible to notice an improvement in the accuracy of the approximation produced by the enriched finite element.

Chapter 8

A general class of enriched methods for the simplicial linear finite elements

Low-order elements are widely used and preferred for finite element analysis, specifically the three-node triangular and four-node tetrahedral elements, both based on linear polynomials in barycentric coordinates. They are known, however, to under-perform when nearly incompressible materials are involved. The problem may be circumvented by the use of higher degree polynomial elements, but their application become both more complex and computationally expensive. For this reason, nonpolynomial enriched finite element methods have been proposed for solving engineering problems. In line with previous chapters, the main goal of this chapter is to provide a general strategy for enriching the standard simplicial linear finite element by nonpolynomial functions, without imposing restrictive conditions on the enrichment functions, like their vanishing at the vertices. In other words, we extend the results presented in Chapter 7 to a more general case, by using generic enrichment continuous functions and generic linear functionals. We investigate the conditions on the enrichment functions under which the associated interpolation problem is solvable. The results presented in this chapter can be found in [35].

8.1 Enrichment of the standard simplicial linear finite element

Let $S_d \subset \mathbb{R}^d$ be the d-simplex in $\mathbb{R}^d$ with vertices $\boldsymbol{v}_0, \dots, \boldsymbol{v}_d$ and barycentric coordinates $\lambda_0, \dots, \lambda_d$. For $i = 0, \dots, d$, we denote by F_i the face of S_d which does not contain the vertex $\boldsymbol{v}_i$. For each $j = 0, \dots, d$, we consider the *enriched* set of linear functionals

$$\Sigma_{S_d}^{\text{enr}} = \{L_j, I_j^{\text{enr}} \,:\, j = 0, \dots, d\}, \tag{8.1}$$

where $L_j : C(S_d) \to \mathbb{R}$ is the point evaluation functionals at the vertices of the simplex S_d, that is

$$L_j(f) = f(\boldsymbol{v}_j), \tag{8.2}$$

while I_j^{enr} is an *enrichment* linear functional on $C(S_d)$. By using the same notations of the previous chapter, we aim to enrich the space

$$\mathbb{P}_1(S_d) = \text{span}\,\{1, x_1, \dots, x_d\}\,, \tag{8.3}$$

up to a space $\mathbb{P}_1^{\text{enr}}(S_d)$ with a set of appropriate continuous enrichment functions $e_0, \dots, e_d$ such that, by setting

$$\mathbb{F}^{\text{enr}} = \text{span}\,\{e_0, \dots, e_d\}\,, \tag{8.4}$$

we get

$$\mathbb{P}_1^{\text{enr}}(S_d) = \mathbb{P}_1(S_d) \oplus \mathbb{F}^{\text{enr}}. \tag{8.5}$$

Definition 8.1.1. *Let $e_0, \dots, e_d$ be linearly independent continuous enrichment functions. They are said admissible enrichment functions if we can enrich $\mathcal{P}_1(S_d)$, defined in (7.7), to the finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$.*

In order to find conditions under which $e_0, \ldots, e_d$ are admissible enrichment functions, we assume that

$$\dim(\mathbb{P}_1^{\rm enr}(S_d)) = 2d + 2. \tag{8.6}$$

A key issue in the selection of the enrichment functions, in both conforming and non-conforming finite element methods, is to assure that $\mathbb{P}_1^{\rm enr}(S_d)$ is $\Sigma_{S_d}^{\rm enr}$-unisolvent [23, Ch 2]. Consequently, given a set of enrichment functionals $\{I_j^{\rm enr} : j = 0, \ldots, d\}$, the problem arises in determining classes of sets of admissible enrichment functions $\mathbb{F}^{\rm enr}$. This problem is commonly referred as to the *local enrichment approximation* [3] and occurs widely in practical applications, such as the locally enriched finite element [98], and the surface reconstruction with enriched reproducing kernel particle approximation, via the use of additional enrichment functions.

Remark 8.1.2. *We remark that, the operator $\Pi^{\rm lin}$, defined in* (7.27)*, satisfies*

$$L_j\left(\Pi^{\rm lin}[f]\right) = L_j(f), \qquad j = 0, \ldots, d, \tag{8.7}$$

and the relative approximation error

$$E^{\rm lin}[f] = f - \Pi^{\rm lin}[f], \qquad f \in C(S_d) \tag{8.8}$$

satisfies

$$E^{\rm lin}[f] = 0, \qquad f \in \mathbb{P}_1(S_d). \tag{8.9}$$

For any $k = 0, \ldots, d$, we introduce the functional

$$\mathcal{E}_k^{\rm tra} = \sum_{j=0}^{d} I_k^{\rm enr}(\lambda_j)\, L_j - I_k^{\rm enr}, \tag{8.10}$$

which can be seen as the error in approximating the linear functional $I_k^{\rm enr}$ by a linear combination of functionals L_j, $j = 0, \ldots, d$ [48]. Then, we have the following Lemma.

Lemma 8.1.3. *Let $f \in \mathbb{P}_1(S_d)$. Then, for any $k = 0, \ldots, d$, we have*

$$\mathcal{E}_k^{\rm tra}(f) = 0. \tag{8.11}$$

Proof. Since the set of barycentric coordinates forms a basis of $\mathbb{P}_1(S_d)$, it suffices to show that for any $k = 0, \ldots, d$, we have

$$\mathcal{E}_k^{\rm tra}(\lambda_i) = 0, \qquad i = 0, \ldots, d. \tag{8.12}$$

In fact, by using the Lagrange property of barycentric coordinates

$$\lambda_i(\boldsymbol{v}_j) = \delta_{ij} \tag{8.13}$$

we get

$$\begin{aligned} \mathcal{E}_k^{\rm tra}(\lambda_i) &= \sum_{j=0}^{d} I_k^{\rm enr}(\lambda_j)\, \lambda_i(\boldsymbol{v}_j) - I_k^{\rm enr}(\lambda_i) \\ &= \sum_{j=0}^{d} I_k^{\rm enr}(\lambda_j)\, \delta_{ij} - I_k^{\rm enr}(\lambda_i) \\ &= I_k^{\rm enr}(\lambda_i) - I_k^{\rm enr}(\lambda_i) = 0. \end{aligned}$$

□

Further, for any $j = 0, \ldots, d$, we introduce the vector $\boldsymbol{g}^j \in \mathbb{R}^{2d+2}$ of components

$$g_i^j = I_j^{\rm enr}(\lambda_i), \qquad g_{i+d+1}^j = -\delta_{ij}, \qquad i = 0, \ldots, d, \tag{8.14}$$

and, for any $f \in C(S_d)$, the vector $\mathbf{L}(f)$ of components

$$L_0(f), \dots, L_d(f), \qquad L_{j+d+1}(f) = I_j^{\text{enr}}(f) \qquad j = 0, \dots, d. \tag{8.15}$$

Therefore, the functionals $\mathcal{E}_k^{\text{tra}}$ defined in (8.10) can be expressed in terms of $\boldsymbol{g}^k$ and $\mathbf{L}(f)$ as

$$\mathcal{E}_k^{\text{tra}}(f) = \left\langle \boldsymbol{g}^k, \mathbf{L}(f) \right\rangle, \tag{8.16}$$

where $\langle \cdot, \cdot \rangle$ is the usual scalar product on $\mathbb{R}^{2d+2}$.

Lemma 8.1.4. *The vectors $\boldsymbol{g}^j$, $j = 0, \dots, d$, are linearly independent.*

Proof. Let $\alpha_0, \dots, \alpha_d \in \mathbb{R}$ such that $\sum_{j=0}^{d} \alpha_j \boldsymbol{g}^j = \mathbf{0}$. Then, in particular, from (8.14) we get

$$-\sum_{j=0}^{d} \alpha_j \delta_{ij} = 0, \qquad i = 0, \dots, d,$$

i.e. $\alpha_i = 0$, $i = 0, \dots, d$, and then the vectors $\boldsymbol{g}^j$, $j = 0, \dots, d$ are linearly independent. □

In the light of the dimension of the enriched space $\mathbb{P}_1^{\text{enr}}(S_d)$, stated in condition (8.6), previous lemmas imply the following characterization result for the enrichment functions $e_0, \dots, e_d$, so that the triple

$$(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$$

is a finite element, or equivalently so that $\mathbb{P}_1^{\text{enr}}(S_d)$ is $\Sigma_{S_d}^{\text{enr}}$-unisolvent.

Theorem 8.1.5. *Let*

$$N = \begin{bmatrix} -\mathcal{E}_0^{\text{tra}}(e_0) & \dots & -\mathcal{E}_0^{\text{tra}}(e_d) \\ -\mathcal{E}_1^{\text{tra}}(e_0) & \dots & -\mathcal{E}_1^{\text{tra}}(e_d) \\ \vdots & \vdots & \vdots \\ -\mathcal{E}_d^{\text{tra}}(e_0) & \dots & -\mathcal{E}_d^{\text{tra}}(e_d) \end{bmatrix}, \tag{8.17}$$

then the triple $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$ is a finite element if and only

$$\det(N) \neq 0.$$

Proof. Let us assume that $\det(N) \neq 0$ and we prove that $\mathbb{P}_1^{\text{enr}}(S_d)$ is $\Sigma_{S_d}^{\text{enr}}$-unisolvent. Let $f \in \mathbb{P}_1^{\text{enr}}(S_d)$ such that

$$L_j(f) = 0, \quad j = 0, \dots, d, \tag{8.18}$$

$$I_j^{\text{enr}}(f) = 0, \qquad j = 0, \dots, d. \tag{8.19}$$

Since $f \in \mathbb{P}_1^{\text{enr}}(S_d)$, it can be expressed as

$$f = p + \sum_{i=0}^{d} \beta_i e_i,$$

where $p \in \mathbb{P}_1(S_d)$ and $\beta_0, \dots, \beta_d$ are real numbers. Each functional $\mathcal{E}_k^{\text{tra}}$, $k = 0, \dots, d$ is a linear combination of the functionals L_j, $j = 0, \dots, d$ and I_k^{enr} then, as $\mathcal{E}_k^{\text{tra}}(p) = 0$ if $p \in \mathbb{P}_1(S_d)$ by (8.11), from (8.18) and (8.19), we obtain

$$\begin{aligned} 0 &= \mathcal{E}_k^{\text{tra}}(f) \\ &= \mathcal{E}_k^{\text{tra}}(p) + \sum_{i=0}^{d} \beta_i \mathcal{E}_k^{\text{tra}}(e_i) \\ &= \sum_{i=0}^{d} \beta_i \mathcal{E}_k^{\text{tra}}(e_i), \qquad k = 0, \dots, d. \end{aligned} \tag{8.20}$$

Equation (8.20) can be represented in matrix form as

$$-N\boldsymbol{\beta} = \mathbf{0},$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^T$. Since, by hypothesis, the matrix N is nonsingular, we get $\beta_0 = \beta_1 = \cdots = \beta_d = 0$ and therefore $f = p$. Taking into account that f vanishes at the vertices of the simplex S_d, by (8.18), we find that f is identically zero.
In order to prove the reverse implication, let us assume to the contrary that

$$\det(N) = 0.$$

Since the determinant of any square matrix is equal to the determinant of its transpose, then there exist constants $\gamma_0, \ldots, \gamma_d$, not all zero, such that the functional

$$\mathcal{E}^{\mathrm{tra}} = \sum_{k=0}^{d} \gamma_k \mathcal{E}_k^{\mathrm{tra}},$$

vanishes at the enrichment functions $e_0, \ldots, e_d$. By the linearity of $\mathcal{E}^{\mathrm{tra}}$ and by equation (8.11) in Lemma 8.1.3, we deduce that $\mathcal{E}^{\mathrm{tra}}$ vanishes on the whole space $\mathbb{P}_1^{\mathrm{enr}}(S_d)$. Therefore, for any $f \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$, from (8.16), we have

$$\begin{aligned} 0 = \mathcal{E}^{\mathrm{tra}}(f) &= \sum_{k=0}^{d} \gamma_k \mathcal{E}_k^{\mathrm{tra}}(f) \\ &= \sum_{k=0}^{d} \gamma_k \left\langle \boldsymbol{g}^k, \boldsymbol{L}(f) \right\rangle \\ &= \left\langle \sum_{k=0}^{d} \gamma_k \boldsymbol{g}^k, \boldsymbol{L}(f) \right\rangle. \end{aligned} \tag{8.21}$$

Since, by hypothesis, $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$ is a finite element, there exist $2d+2$ linearly independent functions $f_i \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$, $i = 1, \ldots, 2d+2$, such that $\boldsymbol{L}(f_i) = \boldsymbol{u}_i$, where $\boldsymbol{u}_i$ is the i-th element of the canonical basis of $\mathbb{R}^{2d+2}$. Therefore, by equation (8.21), we get

$$\left\langle \sum_{k=0}^{d} \gamma_k \boldsymbol{g}^k, \boldsymbol{u}_i \right\rangle = 0, \qquad i = 1, \ldots, 2d+2,$$

i.e. $\sum_{k=0}^{d} \gamma_k \boldsymbol{g}^k$ is orthogonal to the basis vectors $\boldsymbol{u}_i$, $i = 1, \ldots, 2d+2$, and therefore

$$\sum_{k=0}^{d} \gamma_k \boldsymbol{g}^k = \mathbf{0}.$$

From Lemma 8.1.4, the vectors $\boldsymbol{g}^k$, $k = 0, \ldots, d$ are linearly independent in $\mathbb{R}^{2d+2}$ and then $\gamma_k = 0$, $k = 0, \ldots, d$, which is a contradiction. □

In the special case in which the enrichment functions vanish at the vertices of the simplex, by (8.10) we get

$$\mathcal{E}_k^{\mathrm{tra}}(e_i) = -I_k^{\mathrm{enr}}(e_i), \qquad i, k = 0, \ldots, d,$$

and then we can reformulate the previous theorem as follows.

Theorem 8.1.6. *Let $\Delta^{\mathrm{enr}} = \{I_0^{\mathrm{enr}}, \ldots, I_d^{\mathrm{enr}}\} \subset \Sigma_{S_d}^{\mathrm{enr}}$ be a set of enrichment linear functionals and $\mathbb{F}^{\mathrm{enr}} = \operatorname{span}\{e_0, \ldots, e_d\}$, where $e_0, \ldots, e_d$ are enrichment functions satisfying the vanishing condition at the vertices of S_d, that is*

$$e_i(\boldsymbol{v}_j) = 0, \qquad i, j = 0, \ldots, d. \tag{8.22}$$

Then $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$ is a finite element if and only if $(S_d, \mathbb{F}^{\mathrm{enr}}, \Delta^{\mathrm{enr}})$ is a finite element.

Proof. Since the enrichment functions $e_0, \dots, e_d$ are assumed to be linearly independent, the condition $\det(N) \neq 0$ of Theorem 8.1.5 reformulated in terms of the functionals I_k^{enr}, is equivalent to the statement that $(S_d, \mathbb{F}^{\text{enr}}, \Delta^{\text{enr}})$ is a finite element [23, Thm. 2.2.2]. □

In the following, we assume that the matrix N is nonsingular and we denote its inverse by

$$N^{-1} = [\boldsymbol{c}_0 \dots \boldsymbol{c}_d],$$

where $\boldsymbol{c}_i \in \mathbb{R}^d$, $i = 0, \dots, d$, are column vectors. A direct consequence of Theorem 8.1.5 is the linear independence of the functionals of $\Sigma_{S_d}^{\text{enr}}$ in the dual space $\mathbb{P}_1^{\text{enr}}(S_d)^\star$ [23, Ch 2]. Then, there exists a basis $\{\varphi_j, \psi_j : j = 0, \dots, d\}$ of $\mathbb{P}_1^{\text{enr}}(S_d)$ which satisfy

$$L_j(\varphi_i) = \delta_{ij}, \quad I_j^{\text{enr}}(\varphi_i) = 0, \qquad i, j = 0, \dots, d, \tag{8.23}$$

$$L_j(\psi_i) = 0, \quad I_j^{\text{enr}}(\psi_i) = \delta_{ij}, \qquad i, j = 0, \dots, d. \tag{8.24}$$

In the following, we derive explicit expressions for such basis functions.

Theorem 8.1.7. *The basis functions φ_j, ψ_j, $j = 0, \dots, d$ of the finite element $(S_d, \mathbb{P}_1^{\text{enr}}, \Sigma_{S_d}^{\text{enr}})$ have the following expressions*

$$\varphi_j = \lambda_j - \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_j)\psi_k, \qquad j = 0, \dots, d, \tag{8.25}$$

$$\psi_j = \left\langle E^{\text{lin}}[\boldsymbol{e}], \boldsymbol{c}_j \right\rangle, \qquad j = 0, \dots, d, \tag{8.26}$$

where

$$\boldsymbol{e} = [e_0, \dots, e_d]^T$$

and, according to (8.8), $E^{\text{lin}}[\boldsymbol{e}] = \boldsymbol{e} - \Pi^{\text{lin}}[\boldsymbol{e}]$.

Proof. Without loss of generality, we prove (8.25) for the case $j = 0$. Since φ_0 belongs to $\mathbb{P}_1^{\text{enr}}(S_d)$, by (8.5) it can be expressed as

$$\varphi_0 = p + \sum_{i=0}^{d} \beta_{0_i} e_i, \tag{8.27}$$

or, equivalently,

$$\varphi_0 = p + \langle \boldsymbol{e}, \boldsymbol{\beta}_0 \rangle, \tag{8.28}$$

where $\boldsymbol{\beta}_0 = [\beta_{0_0}, \dots, \beta_{0_d}]^T \in \mathbb{R}^{d+1}$. Since, by (8.10),

$$\mathcal{E}_k^{\text{tra}}(\varphi_0) = \sum_{j=0}^{d} I_k^{\text{enr}}(\lambda_j) L_j(\varphi_0) - I_k^{\text{enr}}(\varphi_0) \qquad k = 0, \dots, d,$$

from (8.23) we have

$$\mathcal{E}_k^{\text{tra}}(\varphi_0) = \sum_{j=0}^{d} I_k^{\text{enr}}(\lambda_j)\delta_{0j} - 0 = I_k^{\text{enr}}(\lambda_0), \qquad k = 0, \dots, d. \tag{8.29}$$

We apply $\mathcal{E}_k^{\text{tra}}$ to both members of (8.27) and by previous equation we get

$$I_k^{\text{enr}}(\lambda_0) = \mathcal{E}_k^{\text{tra}}(p) + \sum_{i=0}^{d} \beta_{0_i} \mathcal{E}_k^{\text{tra}}(e_i), \qquad k = 0, \dots, d.$$

Equation (8.11) in Lemma 8.1.3 then yields

$$I_k^{\text{enr}}(\lambda_0) = \sum_{i=0}^{d} \beta_{0_i} \mathcal{E}_k^{\text{tra}}(e_i), \qquad k = 0, \dots, d,$$

or, in matrix form,

$$\boldsymbol{I}_0^{\text{enr}} = -N\boldsymbol{\beta}_0,$$

where $\boldsymbol{I}_0^{\text{enr}} = [I_0^{\text{enr}}(\lambda_0), \dots, I_d^{\text{enr}}(\lambda_0)]^T$. Since N is not singular, we have

$$\boldsymbol{\beta}_0 = -N^{-1}\boldsymbol{I}_0^{\text{enr}} = -\sum_{k=0}^{d} \boldsymbol{c}_k I_k^{\text{enr}}(\lambda_0),$$

which we substitute in (8.28) to obtain

$$\varphi_0 = p - \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_0) \langle \boldsymbol{e}, \boldsymbol{c}_k \rangle . \tag{8.30}$$

By applying L_j, $j = 0, \dots, d$ to both sides of previous equation, by (8.23) we have

$$\delta_{0j} = p(\boldsymbol{v}_j) - \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_0) \langle \boldsymbol{e}(\boldsymbol{v}_j), \boldsymbol{c}_k \rangle , \qquad j = 0, \dots, d,$$

and then

$$p(\boldsymbol{v}_j) = \delta_{0j} + \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_0) \langle \boldsymbol{e}(\boldsymbol{v}_j), \boldsymbol{c}_k \rangle , \qquad j = 0, \dots, d.$$

By multiplying the above equality by λ_j and by summing over all $j = 0, \dots, d$ we immediately get

$$\sum_{j=0}^{d} \lambda_j p(\boldsymbol{v}_j) = \lambda_0 + \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_0) \left\langle \sum_{j=0}^{d} \boldsymbol{e}(\boldsymbol{v}_j)\lambda_j, \boldsymbol{c}_k \right\rangle . \tag{8.31}$$

Since p is a linear polynomial, by equations (8.7) and (8.9) the identity (8.31) becomes

$$p = \lambda_0 + \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_0) \left\langle \Pi^{\text{lin}}[\boldsymbol{e}], \boldsymbol{c}_k \right\rangle \tag{8.32}$$

and then, by substituting (8.32) in (8.30), we get

$$\varphi_0 = \lambda_0 + \sum_{k=0}^{d} I_k^{\text{enr}}(\lambda_0) \left\langle \Pi^{\text{lin}}[\boldsymbol{e}] - \boldsymbol{e}, \boldsymbol{c}_k \right\rangle .$$

In order to prove (8.25) for $j = 0$, it remains to prove (8.26). To this aim, without loss of generality, we show the validity of (8.26) for $j = 0$. We proceed in analogy to the previous case and then we set

$$\psi_0 = p + \sum_{i=0}^{d} \gamma_{0_i} e_i, \tag{8.33}$$

or, equivalently,

$$\psi_0 = p + \langle \boldsymbol{e}, \boldsymbol{\gamma}_0 \rangle , \tag{8.34}$$

where $\boldsymbol{\gamma}_0 = [\gamma_{0_0}, \dots, \gamma_{0_d}]^T \in \mathbb{R}^{d+1}$. Since, by (8.10),

$$\mathcal{E}_k^{\text{tra}}(\psi_0) = \sum_{j=0}^{d} I_k^{\text{enr}}(\lambda_j) L_j(\psi_0) - I_k^{\text{enr}}(\psi_0) \qquad k = 0, \dots, d,$$

from (8.24) we have

$$\mathcal{E}_k^{\text{tra}}(\psi_0) = -\delta_{0k}.$$

We apply $\mathcal{E}_k^{\mathrm{tra}}$ to both members of (8.33) and by previous equation we get

$$-\delta_{0k} = \mathcal{E}_k^{\mathrm{tra}}(p) + \sum_{i=0}^{d} \gamma_{0_i} \mathcal{E}_k^{\mathrm{tra}}(e_i), \qquad k = 0, \ldots, d.$$

Equation (8.11) in Lemma 8.1.3 then yields

$$-\delta_{0k} = \sum_{i=0}^{d} \gamma_{0_i} \mathcal{E}_k^{\mathrm{tra}}(e_i), \qquad k = 0, \ldots, d,$$

or, in matrix form,

$$\boldsymbol{u}_0 = N\boldsymbol{\gamma}_0,$$

where $\boldsymbol{u}_i$, $i = 0, \ldots, d$ is the standard basis of $\mathbb{R}^{d+1}$. Therefore, we get

$$\boldsymbol{\gamma}_0 = N^{-1}\boldsymbol{u}_0 = \boldsymbol{c}_0,$$

and by substituting into equation (8.34), we find

$$\psi_0 = p + \langle \boldsymbol{e}, \boldsymbol{c}_0 \rangle . \tag{8.35}$$

By applying L_j, $j = 0, \ldots, d$ to both sides of previous equation, by (8.24) we have

$$0 = p(\boldsymbol{v}_j) + \langle \boldsymbol{e}(\boldsymbol{v}_j), \boldsymbol{c}_0 \rangle ,$$

and then

$$p(\boldsymbol{v}_j) = -\langle \boldsymbol{e}(\boldsymbol{v}_j), \boldsymbol{c}_0 \rangle .$$

By multiplying the above equality by λ_j and by summing over all $j = 0, \ldots, d$, we have

$$\sum_{j=0}^{d} \lambda_j p(\boldsymbol{v}_j) = -\left\langle \sum_{j=0}^{d} \lambda_j \boldsymbol{e}(\boldsymbol{v}_j), \boldsymbol{c}_0 \right\rangle , \tag{8.36}$$

and then

$$p = -\left\langle \Pi^{\mathrm{lin}}[\boldsymbol{e}], \boldsymbol{c}_0 \right\rangle .$$

Finally, equation (8.35) yields

$$\psi_0 = \left\langle \boldsymbol{e} - \Pi^{\mathrm{lin}}[\boldsymbol{e}], \boldsymbol{c}_0 \right\rangle ,$$

then the equation (8.26) is proved for $j = 0$. Similarly, we can prove (8.26) for $j = 1, \ldots, d$ and consequently (8.25) for $j = 0$ is proved. The expression of the other functions can be obtained using symmetry arguments. □

Remark 8.1.8. *Let us assume that $e_0, \ldots, e_d$ satisfy the vanishing conditions* (8.22)*. Then the basis functions of the element $(S_d, \mathbb{F}^{\mathrm{enr}}, \Delta^{\mathrm{enr}})$ are*

$$\psi_j = \langle \boldsymbol{e}, \boldsymbol{c}_j \rangle , \qquad j = 0, \ldots, d.$$

8.2 Admissible enrichment functions

In this section, we collect sets of *admissible enrichment functions*, that is functions for which the triple $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$ is a finite element. The main issue when using the proposed enriched method is to ensure that the matrix N given in (8.17) is invertible. These enrichment functions constitute a very general class, which can be used for many types of applications. In the following, we consider the enrichment linear functionals

$$I_j^{\mathrm{enr}}(f) = \frac{1}{|F_j|} \int_{F_j} f(\boldsymbol{x})\, d\sigma(\boldsymbol{x}), \qquad j = 0, \ldots, d, \tag{8.37}$$

where F_j is the face of S_d which does not contain the vertex $\boldsymbol{v}_j$ and $d\sigma(\boldsymbol{x})$ is the Lebesgue measure on the face F_j.

There are two special classes of enrichment functions which are of particular interest.

8.2.1 Admissible enrichment functions of the first class

The functions of the first class can be represented as a product of n convex, increasing (or decreasing), nonnegative functions. The following lemma from [59] and the subsequent one play a crucial role in our analysis.

Lemma 8.2.1. *Let g be a convex function on S_d, then for any $j = 0, \dots, d$, we have*

$$\frac{1}{|F_j|} \int_{F_j} g(\boldsymbol{x})\, d\sigma(\boldsymbol{x}) \leq \frac{1}{d} \sum_{\substack{i=0 \\ i \neq j}}^{d} g(\boldsymbol{v}_i).$$

Equality is attained if and only if g is an affine function.

Lemma 8.2.2. *Let $n \in \mathbb{N}$ and let $f_1, \dots, f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero. Let us assume that f_1 is strictly convex. Then*

$$h_n = \prod_{\ell=1}^{n} f_\ell$$

is a strictly convex, increasing (or decreasing), nonnegative function.

Proof. The proof is by induction on n. The case $n = 1$ is trivial. We assume that

$$h_{n-1} = \prod_{\ell=1}^{n-1} f_\ell$$

is a strictly convex, increasing (or decreasing), nonnegative function. We prove that $h_n = h_{n-1} f_n$ is an increasing (or decreasing) function. Let $x, y \in [0,1]$ such that $x < y$, then

$$(h_{n-1}f_n)(x) - (h_{n-1}f_n)(y) = h_{n-1}(x)\,(f_n(x) - f_n(y)) + f_n(y)\,(h_{n-1}(x) - h_{n-1}(y))\,.$$

Therefore $h_{n-1}f_n$ is an increasing (or decreasing) function. For each $x, y \in [0,1]$ and $t \in (0,1)$, we set

$$\delta = t(h_{n-1}f_n)(x) + (1-t)(h_{n-1}f_n)(y) - (h_{n-1}f_n)(tx + (1-t)y).$$

We want to prove that $\delta > 0$. Since h_{n-1} is a strictly convex function, we get

$$\begin{aligned} (h_{n-1}f_n)(tx + (1-t)y) &= h_{n-1}(tx + (1-t)y) f_n(tx + (1-t)y) \\ &< (th_{n-1}(x) + (1-t)h_{n-1}(y))(tf_n(x) + (1-t)f_n(y)). \end{aligned} \tag{8.38}$$

After easy computations, from (8.38), we get

$$\delta > t(1-t)\,(h_{n-1}(x) - h_{n-1}(y))\,(f_n(x) - f_n(y)) > 0.$$

Then $h_{n-1}f_n$ is a strictly convex, increasing (or decreasing), nonnegative function. □

Now we can prove the following theorem.

Theorem 8.2.3. *Let $n \in \mathbb{N}$ and let $f_1, \dots, f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero and let us assume that at least one function is strictly convex. We consider the enrichment functions*

$$e_i = \prod_{\ell=1}^{n} f_\ell(\lambda_i), \qquad i = 0, \dots, d, \tag{8.39}$$

and the enrichment linear functionals defined in (8.37)*. Then we have*

$$\mathcal{E}_i^{\mathrm{tra}}(e_i) = 0, \qquad i = 0, \dots, d \tag{8.40}$$

and

$$\mathcal{E}_k^{\mathrm{tra}}(e_i) > 0, \qquad i, k = 0, \dots, d,\ i \neq k, \tag{8.41}$$

where the functionals $\mathcal{E}_k^{\mathrm{tra}}$ are defined in (8.10)*.*

Proof. Since $\lambda_i(\boldsymbol{x}) = 0$ for any $\boldsymbol{x} \in F_i$, $i = 0, \ldots, d$, by the integral formula (7.5), we get

$$\begin{aligned}
\mathcal{E}_i^{\mathrm{tra}}(e_i) &= \sum_{j=0}^{d} I_i^{\mathrm{enr}}(\lambda_j)\, e_i(\boldsymbol{v}_j) - I_i^{\mathrm{enr}}(e_i) \\
&= \sum_{\substack{j=0 \\ j\neq i}}^{d} I_i^{\mathrm{enr}}(\lambda_j)\, e_i(\boldsymbol{v}_j) - I_i^{\mathrm{enr}}(e_i) \\
&= \sum_{\substack{j=0 \\ j\neq i}}^{d} \prod_{\ell=1}^{n} \frac{f_\ell(0)}{d} - \prod_{\ell=1}^{n} f_\ell(0) \\
&= \prod_{\ell=1}^{n} f_\ell(0) - \prod_{\ell=1}^{n} f_\ell(0) = 0.
\end{aligned}$$

For $i \neq k$, similarly, we get

$$\begin{aligned}
\mathcal{E}_k^{\mathrm{tra}}(e_i) &= \sum_{j=0}^{d} I_k^{\mathrm{enr}}(\lambda_j)\, e_i(\boldsymbol{v}_j) - I_k^{\mathrm{enr}}(e_i) \\
&= \sum_{\substack{j=0 \\ j\neq k}}^{d} \frac{1}{d} e_i(\boldsymbol{v}_j) - I_k^{\mathrm{enr}}(e_i).
\end{aligned}$$

By Lemma 8.2.2 the functions e_i, $i = 0, \ldots, d$ are convex functions and then, by Lemma 8.2.1, we get

$$\mathcal{E}_k^{\mathrm{tra}}(e_i) > 0, \qquad i, k = 0, \ldots, d, \quad i \neq k.$$

□

We denote by $\widehat{\lambda}_i$, $i = 0, \ldots, d$, the barycentric coordinates of the d-dimensional standard simplex

$$\widehat{S}_d = \left\{ \widehat{\boldsymbol{x}} = (\widehat{x}_1, \ldots, \widehat{x}_d) \in \mathbb{R}^d,\, \widehat{x}_i \geq 0,\, i = 1, \ldots, d,\, \sum_{j=1}^{d} \widehat{x}_j \leq 1 \right\},$$

that is

$$\widehat{\lambda}_0(\widehat{\boldsymbol{x}}) = 1 - \sum_{j=1}^{d} \widehat{x}_j, \quad \widehat{\lambda}_i(\widehat{\boldsymbol{x}}) = \widehat{x}_i, \qquad i = 1, \ldots, d,$$

and by $\widehat{F}_i$ the face of $\widehat{S}_d$ which does not contain the vertex $\widehat{\boldsymbol{v}}_i$. Then, the following result holds.

Theorem 8.2.4. *Let $n \in \mathbb{N}$ and let $f_1, \ldots, f_n \in C([0,1])$. Then*

$$\int_{\widehat{S}_d} \prod_{\ell=1}^{n} f_\ell(\widehat{\lambda}_i(\widehat{\boldsymbol{x}}))\, d\widehat{\boldsymbol{x}} = \int_0^1 \prod_{\ell=1}^{n} f_\ell(t) \frac{(1-t)^{d-1}}{(d-1)!} dt, \qquad i = 1, \ldots, d. \tag{8.42}$$

Proof. First we prove (8.42) for $i \neq 0$. Without loss of generality we assume $i = 1$, then

$$\int_{\widehat{S}_d} \prod_{\ell=1}^{n} f_\ell(\widehat{\lambda}_1(\widehat{\boldsymbol{x}}))\, d\widehat{\boldsymbol{x}} = \int_{\widehat{S}_d} \prod_{\ell=1}^{n} f_\ell(\widehat{x}_1)\, d\widehat{\boldsymbol{x}} = \int_0^1 \prod_{\ell=1}^{n} f_\ell(t) \left(\int_{(1-t)\widehat{S}_{d-1}^1} d\widehat{x}_2 \cdots d\widehat{x}_d \right) dt, \tag{8.43}$$

where we set

$$(1-t)\widehat{S}_{d-1}^1 = \left\{ (\widehat{x}_2, \ldots, \widehat{x}_d) \in \mathbb{R}^{d-1},\, \widehat{x}_i \geq 0,\, i = 2, \ldots, d, \sum_{j=2}^{d} \widehat{x}_j \leq (1-t) \right\}.$$

In order to compute this integral, we consider the map $\zeta : \widehat{S}_{d-1} \to (1-t)\widehat{S}^1_{d-1}$ defined as

$$\zeta(\widehat{x}_1, \ldots, \widehat{x}_{d-1}) = (\zeta_1, \ldots, \zeta_{d-1}),$$

where $\zeta_i = (1-t)\widehat{x}_i$, $i = 1, \ldots, d-1$. By using ζ as change of variables, we get

$$\int_{(1-t)\widehat{S}^1_{d-1}} d\widehat{x}_2 \cdots d\widehat{x}_d = (1-t)^{d-1} \int_{\widehat{S}_{d-1}} d\widehat{x}_1 \cdots d\widehat{x}_{d-1} = (1-t)^{d-1} \left|\widehat{S}_{d-1}\right|. \tag{8.44}$$

By noting that

$$\left|\widehat{S}_{d-1}\right| = \frac{1}{(d-1)!},$$

the result follows by substituting (8.44) in (8.43). Now we prove (8.42) for $i = 0$. To this aim, we consider the linear transformations

$$\begin{aligned} \pi : \widehat{S}_d \to \widehat{S}_d, &\quad \pi(\widehat{x}_1, \widehat{x}_2 \ldots, \widehat{x}_d) = (\widehat{u}_1, \widehat{u}_2 \ldots, \widehat{u}_d), \\ \theta : \widehat{S}_d \to \widehat{S}_d, &\quad \theta(\widehat{x}_1, \widehat{x}_2 \ldots, \widehat{x}_d) = (\widehat{x}_2, \widehat{x}_1 \ldots, \widehat{x}_d), \end{aligned}$$

where

$$\widehat{u}_1 = 1 - \sum_{j=1}^{d} \widehat{x}_j, \quad \widehat{u}_j = \widehat{x}_j, \qquad j = 2, \ldots, d.$$

The Jacobian determinant of the change of variables $\theta \circ \pi$ is 1, then we get

$$\int_{\widehat{S}_d} \prod_{\ell=1}^{n} f_\ell(\widehat{\lambda}_0(\widehat{\boldsymbol{x}}))d\widehat{\boldsymbol{x}} = \int_{\widehat{S}_d} \prod_{\ell=1}^{n} f_\ell(\widehat{\lambda}_1(\widehat{\boldsymbol{x}}))d\widehat{\boldsymbol{x}}.$$

The result follows from the case $i = 1$. □

The previous result can be stated for a generic simplex as follows.

Corollary 8.2.5. *Let $n \in \mathbb{N}$ and let $f_1, \ldots, f_n \in C([0,1])$. For any $i = 0, \ldots, d$, we have*

$$\int_{S_d} \prod_{\ell=1}^{n} f_\ell(\lambda_i(\boldsymbol{x}))d\boldsymbol{x} = \frac{|S_d|}{\left|\widehat{S}_d\right|} \int_0^1 \prod_{\ell=1}^{n} f_\ell(t) \frac{(1-t)^{d-1}}{(d-1)!} dt, \tag{8.45}$$

and then the value of the integral on the right-hand side of (8.45) *does not depend on i.*

Proof. We denote by

$$\zeta : \widehat{S}_d \to S_d, \quad \zeta(\widehat{\boldsymbol{x}}) = \sum_{i=0}^{d} \widehat{\lambda}_i(\widehat{\boldsymbol{x}})\boldsymbol{v}_i.$$

The proof follows directly by Theorem 8.2.4, by using ζ^{-1} as a change of variables. □

Remark 8.2.6. *In the hypotheses of Theorem 8.2.3, for the functionals defined in* (8.10) *and the enrichment functions defined in* (8.39)*, by Lemma 7.1.1, we have*

$$\mathcal{E}^{\mathrm{tra}}_k(e_i) = \frac{1}{d} \prod_{\ell=1}^{n} f_\ell(1) + \frac{(d-1)}{d} \prod_{\ell=1}^{n} f_\ell(0) - I^{\mathrm{enr}}_k \left(\prod_{\ell=1}^{n} f_\ell(\lambda_i) \right), \qquad i, k = 0, \ldots, d, \quad i \neq k.$$

Consequently, taking into account the definition (8.37) *of the enrichment linear functionals $I^{\mathrm{enr}}_k(\cdot)$, by Corollary 8.2.5, we get*

$$\mathcal{E}^{\mathrm{tra}}_k(e_i) = \mathcal{E}^{\mathrm{tra}}_k(e_j), \qquad i, j, k = 0, \ldots, d, \quad k \neq i, j. \tag{8.46}$$

Theorem 8.2.7. *Let $n \in \mathbb{N}$ and let $f_1, \dots, f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero and let us assume that at least one function is strictly convex. Let e_i, I_j^{enr}, $i,j = 0,\dots,d$, be the enrichment functions and the enrichment linear functionals defined in* (8.39) *and* (8.37)*, respectively. Then the matrix N defined in* (8.17) *is nonsingular.*

Proof. By equation (8.46) of Remark 8.2.6 we can set

$$\mu_k = -\mathcal{E}_k^{\mathrm{tra}}(e_i), \qquad i,k = 0,\dots,d, \quad k \neq i. \tag{8.47}$$

Consequently, the matrix N defined in (8.17) is

$$N = \begin{bmatrix} 0 & \mu_0 & \cdots & \mu_0 \\ \mu_1 & 0 & \cdots & \mu_1 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_d & \mu_d & \cdots & 0 \end{bmatrix},$$

where the diagonal elements are zero in force of equation (8.40). Consequently

$$\det(N) = \det \begin{bmatrix} 0 & \mu_0 & \cdots & \mu_0 \\ \mu_1 & 0 & \cdots & \mu_1 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_d & \mu_d & \cdots & 0 \end{bmatrix} = \prod_{k=0}^{d} \mu_k \det \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}.$$

In order to evaluate the last determinant, we replace the first column by the sum of all the columns and then

$$\det(N) = d \prod_{k=0}^{d} \mu_k \det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}.$$

Finally, by subtracting the first row from the rest of the rows, we get

$$\det(N) = d \prod_{k=0}^{d} \mu_k \det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{bmatrix} = (-1)^d d \prod_{k=0}^{d} \mu_k.$$

Since, by Theorem 8.2.3, $\mu_k = -\mathcal{E}_k^{\mathrm{tra}}(e_i) < 0$, for each $i,k = 0,\dots,d$, $k \neq i$, the thesis follows. □

In the following, we give an explicit expression, in closed-form, of the basis functions associated to the finite element enriched with the functionals defined in (8.37) and the enrichment functions defined in (8.39).

Theorem 8.2.8. *Let $e_0, \dots, e_d$ and $I_0^{\mathrm{enr}}, \dots, I_d^{\mathrm{enr}}$ be the enrichment functions and the enrichment functionals defined in* (8.39) *and in* (8.37)*, respectively. Then the basis functions* (8.25), (8.26) *of the enriched finite element have the following expressions*

$$\varphi_j = \lambda_j - \frac{1}{d} \sum_{\substack{k=0 \\ k \neq j}}^{d} \psi_k, \qquad j = 0,\dots,d, \tag{8.48}$$

$$\psi_j = \frac{1}{d\mu_j} \sum_{k=0}^{d} (1 - d\delta_{jk})\, E^{\mathrm{lin}}[e_k], \qquad j = 0,\dots,d, \tag{8.49}$$

where

$$\mu_k = -\mathcal{E}_k^{\mathrm{tra}}(e_i), \qquad i,k = 0,\dots,d, \quad k \neq i. \tag{8.50}$$

Proof. With reference to equation (8.26) of Theorem 8.1.7, we compute

$$\left\langle E^{\mathrm{lin}}[\boldsymbol{e}], \boldsymbol{c}_j \right\rangle,$$

where $\boldsymbol{c}_j$ is the j-th column of the inverse of the matrix

$$N = \begin{bmatrix} 0 & \mu_0 & \cdots & \mu_0 \\ \mu_1 & 0 & \cdots & \mu_1 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_d & \mu_d & \cdots & 0 \end{bmatrix}.$$

It is easy to verify that N has the inverse matrix

$$N^{-1} = \begin{bmatrix} \frac{1-d}{d\mu_0} & \frac{1}{d\mu_1} & \cdots & \frac{1}{d\mu_d} \\ \frac{1}{d\mu_0} & \frac{1-d}{d\mu_1} & \cdots & \frac{1}{d\mu_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{d\mu_0} & \frac{1}{d\mu_1} & \cdots & \frac{1-d}{d\mu_d} \end{bmatrix}. \tag{8.51}$$

By using the Lagrange property of barycentric coordinates, we get

$$E^{\mathrm{lin}}[\boldsymbol{e}] = [E^{\mathrm{lin}}[e_0], \ldots, E^{\mathrm{lin}}[e_d]]^T, \tag{8.52}$$

where

$$E^{\mathrm{lin}}[e_k] = \prod_{\ell=1}^{n} f_\ell(\lambda_k) - \sum_{j=0}^{d} \prod_{\ell=1}^{n} f_\ell(\delta_{jk})\lambda_j, \qquad k = 0, \ldots, d.$$

By combining (8.51), (8.52), we get

$$\left\langle E^{\mathrm{lin}}[\boldsymbol{e}], \boldsymbol{c}_j \right\rangle = \frac{1}{d\mu_j} \sum_{k=0}^{d} (1 - d\delta_{jk})\, E^{\mathrm{lin}}[e_k], \qquad j = 0, \ldots, d. \tag{8.53}$$

The theorem is proved by noting that, by (7.5)

$$I_j^{\mathrm{enr}}(\lambda_i) = \frac{1 - \delta_{ij}}{d}, \qquad i, j = 0, \ldots, d.$$

□

Example 8.2.9. *Theorem 8.2.7 allows us to enrich the standard simplicial linear finite element to the finite element* $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$ *by using the following sets of enrichment functions:*

- $\mathcal{E}_1 = \left\{ e_i = \sin\left(\frac{\pi}{2}(\lambda_i + 2)\right) + 2 \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_2 = \left\{ e_i = \frac{1}{1+\lambda_i} \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_3 = \left\{ e_i = e^{\lambda_i} \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_4 = \{ e_i = \lambda_i^\alpha,\, \alpha > 1 \,:\, i = 0, \ldots, d \}$,
- $\mathcal{E}_5 = \left\{ e_i = \lambda_i^\alpha e^{\lambda_i},\, \alpha > 1 \,:\, i = 0, \ldots, d \right\}$.

It is worth reminding that the columns of the inverse of the matrix N are involved in the expression of the special basis (8.25), (8.26) of the enriched element $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$. For this reason, in the following examples 8.2.10 and 8.2.11 we determine an explicit expression of $\mu_k = -\mathcal{E}_k^{\mathrm{tra}}(e_i)$, $k = 0, \ldots, d$, $k \neq i$, for the sets of admissible enrichment functions $\mathcal{E}_3$ and $\mathcal{E}_4$, respectively.

Example 8.2.10. *We set $e_i = e^{\lambda_i}$, $i = 0, \dots, d$. By equation* (8.10)*, for each $k = 0, \dots, d$, $k \neq i$, we get*

$$\begin{aligned}\mathcal{E}_k^{\mathrm{tra}}(e^{\lambda_i}) &= \sum_{j=0}^{d} I_k^{\mathrm{enr}}(\lambda_j) L_j(e^{\lambda_i}) - I_k^{\mathrm{enr}}(e^{\lambda_i}) \\ &= \sum_{j=0}^{d} I_k^{\mathrm{enr}}(\lambda_j) e^{\lambda_i(\boldsymbol{v}_j)} - I_k^{\mathrm{enr}}(e^{\lambda_i}) \\ &= \sum_{\substack{j=0 \\ j\neq i,k}}^{d} I_k^{\mathrm{enr}}(\lambda_j) e^{\lambda_i(\boldsymbol{v}_j)} + I_k^{\mathrm{enr}}(\lambda_i) e^{\lambda_i(\boldsymbol{v}_i)} - I_k^{\mathrm{enr}}(e^{\lambda_i}),\end{aligned}$$

since $I_k^{\mathrm{enr}}(\lambda_k) = 0$ by equation (7.3)*. We use Kronecker delta property of the barycentric coordinates* (8.13) *and equation* (7.5) *to get*

$$\sum_{\substack{j=0 \\ j\neq i,k}}^{d} I_k^{\mathrm{enr}}(\lambda_j) e^{\lambda_i(\boldsymbol{v}_j)} + I_k^{\mathrm{enr}}(\lambda_i) e^{\lambda_i(\boldsymbol{v}_i)} = \frac{d-1}{d} + \frac{e}{d}.$$

Moreover, by equation (8.45)

$$I_k^{\mathrm{enr}}(e^{\lambda_i}) = \frac{1}{|F_k|}\int_{F_k} e^{\lambda_i(\boldsymbol{x})} d\sigma(\boldsymbol{x}) = \frac{1}{\left|\widehat{F}_k\right|}\int_0^1 e^t \frac{(1-t)^{d-2}}{(d-2)!} dt \qquad k = 0, \dots, d,\ k \neq i.$$

We use the $d-2$ order Taylor series expansion of the function e^x, centered at 0, evaluated at $x = 1$ and integral remainder to get, from previous equation

$$I_k^{\mathrm{enr}}(e^{\lambda_i}) = \frac{1}{\left|\widehat{F}_k\right|}\int_0^1 e^t \frac{(1-t)^{d-2}}{(d-2)!} dt = \frac{1}{\left|\widehat{F}_k\right|}\left(e - \sum_{j=0}^{d-2}\frac{1}{j!}\right), \qquad k = 0, \dots, d,\ k \neq i. \tag{8.54}$$

We set

$$R_{d-2} = e - \sum_{j=0}^{d-2}\frac{1}{j!}$$

and

$$A_k = \left|\widehat{F}_k\right|(d - 1 + e) - dR_{d-2}, \qquad k = 0, \dots, d,$$

to get, finally

$$\mu_k = \mathcal{E}_k^{\mathrm{tra}}\left(e^{\lambda_i}\right) = \frac{A_k}{d\left|\widehat{F}_k\right|}, \qquad k = 0, \dots, d,\ k \neq i. \tag{8.55}$$

Example 8.2.11. *We set $e_i = \lambda_i^\alpha$, $i = 0, \dots, d$, where α is a real number greater than one. By equation* (8.10) *and by using the Kronecker delta property of the barycentric coordinates* (8.13)*, for each $k = 0, \dots, d$, $k \neq i$, we get*

$$\begin{aligned}\mathcal{E}_k^{\mathrm{tra}}\left(\lambda_i^\alpha\right) &= \sum_{j=0}^{d} I_k^{\mathrm{enr}}(\lambda_j) L_j(\lambda_i^\alpha) - I_k^{\mathrm{enr}}(\lambda_i^\alpha) \\ &= \sum_{j=0}^{d} I_k^{\mathrm{enr}}(\lambda_j) \lambda_i^\alpha(\boldsymbol{v}_j) - I_k^{\mathrm{enr}}\left(\lambda_i^\alpha\right) \\ &= I_k^{\mathrm{enr}}(\lambda_i) - I_k^{\mathrm{enr}}(\lambda_i^\alpha).\end{aligned}$$

By equation (7.5)*, finally, we get*

$$\mathcal{E}_k^{\mathrm{tra}}(\lambda_i^\alpha) = \frac{1}{d} - \frac{\Gamma(\alpha+1)}{(d)_\alpha}, \qquad k = 0, \dots, d, \quad k \neq i,$$

where $(d)_\alpha = \frac{\Gamma(\alpha+d)}{\Gamma(d)}$ is the Pochhammer symbol.

8.2.2 Admissible enrichment functions of the second class

The enrichment functions of the second class are the product of positive powers of barycentric coordinates with properly evaluated continuous functions, as shown in the following Theorem.

Theorem 8.2.12. *Let $f_0, \ldots, f_d \in C([0,1])$ be functions such that $f_i(0) \neq 0$, $i = 0, \ldots, d$, and let $\alpha_0, \ldots, \alpha_d$ be real numbers greater than one. Let*

$$e_i = f_i(\lambda_i(\boldsymbol{x})) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{x}), \qquad i = 0, \ldots, d \tag{8.56}$$

be enrichment functions and let I_j^{enr}, $j = 0, \ldots, d$, be the enrichment linear functionals defined in (8.37)*. Then the matrix N defined in* (8.17) *is nonsingular.*

Proof. Using (7.5) and the Kronecker delta property of barycentric coordinates (8.13), for each $i = 0, \ldots, d$, we get

$$\begin{aligned}
\mathcal{E}_i^{\mathrm{tra}}(e_i) &= \sum_{j=0}^{d} I_i^{\mathrm{enr}}(\lambda_j)\, e_i(\boldsymbol{v}_j) - I_i^{\mathrm{enr}}(e_i) \\
&= \frac{1}{d} \sum_{\substack{j=0 \\ j \neq i}}^{d} e_i(\boldsymbol{v}_j) - I_i^{\mathrm{enr}}(e_i) \\
&= \frac{1}{d} \sum_{\substack{j=0 \\ j \neq i}}^{d} f_i(\lambda_i(\boldsymbol{v}_j)) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{v}_j) - \frac{1}{|F_i|} \int_{F_i} f_i(\lambda_i(\boldsymbol{x})) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \\
&= \frac{1}{d} f_i(0) \sum_{\substack{j=0 \\ j \neq i}}^{d} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{v}_j) - f_i(0) \frac{1}{|F_i|} \int_{F_i} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \\
&= -f_i(0) \frac{(d-1)! \prod_{\ell=0, \ell \neq i}^{d} \Gamma(\alpha_\ell)}{\Gamma(\sum_{\ell=0, \ell \neq i}^{d} \alpha_\ell)} \neq 0,
\end{aligned}$$

while, for each $i, k = 0, \ldots, d$, $i \neq k$, we get

$$\begin{aligned}
\mathcal{E}_k^{\mathrm{tra}}(e_i) &= \sum_{j=0}^{d} I_k^{\mathrm{enr}}(\lambda_j) e_i(\boldsymbol{v}_j) - I_k^{\mathrm{enr}}(e_i) \\
&= \frac{1}{d} \sum_{\substack{j=0 \\ j \neq k}}^{d} e_i(\boldsymbol{v}_j) - I_k^{\mathrm{enr}}(e_i) \\
&= \frac{1}{d} \sum_{\substack{j=0 \\ j \neq k}}^{d} f_i(\lambda_i(\boldsymbol{v}_j)) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{v}_j) - \frac{1}{|F_k|} \int_{F_k} f_i(\lambda_i(\boldsymbol{x})) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \\
&= \frac{1}{d} f_i(0) \sum_{\substack{j=0 \\ j \neq i,k}}^{d} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{v}_j) + \frac{1}{d} f_i(1) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1}(\boldsymbol{v}_i) = 0.
\end{aligned}$$

Consequently, the matrix N defined in (8.17) is a diagonal matrix with elements different from zero and then nonsingular. □

In the following, we give an explicit expression, in closed-form, of the basis functions associated to the finite element enriched with the functionals defined in (8.37) and the enrichment functions defined in (8.56).

Theorem 8.2.13. *Let $e_0, \dots, e_d$ and $I_0^{\mathrm{enr}}, \dots, I_d^{\mathrm{enr}}$ be the enrichment functions and the enrichment functionals defined in* (8.56) *and in* (8.37)*, respectively. Then the basis functions* (8.25), (8.26) *of the enriched finite element have the following expressions*

$$\varphi_j = \lambda_j - \frac{1}{d}\sum_{\substack{k=0\\k\neq j}}^{d} \psi_k, \qquad j = 0, \dots, d, \tag{8.57}$$

$$\psi_j = \frac{E^{\mathrm{lin}}[e_j]}{\mu_j}, \qquad j = 0, \dots, d, \tag{8.58}$$

where

$$\mu_j = -\mathcal{E}_j^{\mathrm{tra}}(e_j), \qquad j = 0, \dots, d. \tag{8.59}$$

Proof. With reference to equation (8.26) of Theorem 8.1.7, we compute

$$\left\langle E^{\mathrm{lin}}[\boldsymbol{e}], \boldsymbol{c}_j \right\rangle,$$

where $\boldsymbol{c}_j$ is the j-th column of the inverse of the matrix

$$N = \begin{bmatrix} \mu_0 & 0 & \cdots & 0 \\ 0 & \mu_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_d \end{bmatrix}.$$

It is easy to verify that N has the inverse matrix

$$N^{-1} = \begin{bmatrix} \frac{1}{\mu_0} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mu_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\mu_d} \end{bmatrix}. \tag{8.60}$$

By using the Lagrange property of barycentric coordinates, we get

$$E^{\mathrm{lin}}[\boldsymbol{e}] = [E^{\mathrm{lin}}[e_0], \dots, E^{\mathrm{lin}}[e_d]]^T, \tag{8.61}$$

where

$$E^{\mathrm{lin}}[e_j] = e_j, \qquad j = 0, \dots, d.$$

By combining (8.60), (8.61), we get

$$\left\langle E^{\mathrm{lin}}[\boldsymbol{e}], \boldsymbol{c}_j \right\rangle = \frac{E^{\mathrm{lin}}[e_j]}{\mu_j}, \qquad j = 0, \dots, d. \tag{8.62}$$

The theorem is proved by noting that, by (7.5)

$$I_j^{\mathrm{enr}}(\lambda_i) = \frac{1-\delta_{ij}}{d}, \qquad i, j = 0, \dots, d.$$

□

Example 8.2.14. *Theorem 8.2.12 allows us to enrich the standard simplicial linear finite element to the finite element $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \Sigma_{S_d}^{\mathrm{enr}})$ by using the following sets of enrichment functions:*

- $\mathcal{E}_1' = \left\{ e_i = \sin\left(\frac{\pi}{2i+2}(\lambda_i + 1)\right) \prod_{\substack{\ell=0\\\ell\neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \dots, d \right\},$

- $\mathcal{E}_2' = \left\{ e_i = \frac{1+i}{1+\lambda_i} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \ldots, d \right\},$
- $\mathcal{E}_3' = \left\{ e_i = e^{i\lambda_i} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \ldots, d \right\},$
- $\mathcal{E}_4' = \left\{ e_i = \log(i\lambda_i + 2) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \ldots, d \right\}.$

8.3 Error representations

We introduce the approximation operator

$$\begin{aligned} \Pi^{\text{enr}} : C(S_d) &\rightarrow \mathbb{P}_1^{\text{enr}}(S_d) \\ f &\mapsto \sum_{j=0}^{d} L_j(f)\varphi_j + \sum_{j=0}^{d} I_j^{\text{enr}}(f)\psi_j, \end{aligned} \tag{8.63}$$

where φ_j, ψ_j, $j = 0, \ldots, d$, are defined in (8.25) and (8.26), and the approximation error

$$E^{\text{enr}}[f] = f - \Pi^{\text{enr}}[f]. \tag{8.64}$$

We now present an elegant decomposition of the error E^{enr} in terms of the approximation error associated to the approximation operator Π^{lin}, defined in (7.27), and an additional term which depends on the enrichment functions. Indeed, we have the following Theorem.

Theorem 8.3.1. *For any $f \in C(S_d)$, the approximation error* (8.64) *can be decomposed as follows*

$$E^{\text{enr}}[f] = E^{\text{lin}}[f] - \sum_{j=0}^{d} I_j^{\text{enr}}\left(E^{\text{lin}}[f]\right)\psi_j, \tag{8.65}$$

where E^{lin} is defined in (8.8).

Proof. Since the operator Π^{enr} reproduces linear polynomials, we have

$$\Pi^{\text{enr}}\left[\Pi^{\text{lin}}[f]\right] = \Pi^{\text{lin}}[f].$$

By taking into account also the linearity of Π^{enr}, we get

$$\begin{aligned} E^{\text{enr}}[f] &= f - \Pi^{\text{enr}}[f] \\ &= f - \Pi^{\text{enr}}\left[f - \Pi^{\text{lin}}[f] + \Pi^{\text{lin}}[f]\right] \\ &= f - \Pi^{\text{enr}}\left[f - \Pi^{\text{lin}}[f]\right] - \Pi^{\text{lin}}[f] \\ &= E^{\text{lin}}[f] - \Pi^{\text{enr}}\left[E^{\text{lin}}[f]\right]. \end{aligned} \tag{8.66}$$

By (8.7) $E^{\text{lin}}[f]$ vanishes at all the vertices of S_d and then, the definition (8.63) of Π^{enr} yields

$$\Pi^{\text{enr}}\left[E^{\text{lin}}[f]\right] = \sum_{j=0}^{d} I_j^{\text{enr}}\left(E^{\text{lin}}[f]\right)\psi_j. \tag{8.67}$$

By substituting (8.67) in (8.66), we obtain (8.65). □

Now we introduce the approximation operator

$$\begin{aligned} \Pi^{\text{imp}} : C(S_d) &\rightarrow \mathbb{P}_1^{\text{enr}}(S_d) \\ f &\mapsto \Pi^{\text{imp}}[f] = \sum_{j=0}^{d} I_j^{\text{enr}}(f)\psi_j, \end{aligned}$$

and the approximation error

$$E^{\text{imp}}[f] = f - \Pi^{\text{imp}}[f]. \tag{8.68}$$

Remark 8.3.2. *According to the decomposition given in* (8.65)*, the approximation error* E^{enr} *defined in* (8.64) *can be rewritten as follows*

$$E^{\text{enr}}[f] = E^{\text{imp}}[E^{\text{lin}}[f]], \tag{8.69}$$

where E^{imp} is defined in (8.68).

Remark 8.3.3. *We notice that when the enrichment functions* e_i*,* $i = 0, \ldots, d$*, satisfy the vanishing conditions*

$$e_i(\boldsymbol{v}_j) = 0, \qquad i, j = 0, \ldots, d$$

and the enrichment linear functionals are those defined in (8.37)*, the results introduced here are equivalent to the results presented in Chapter 7. Then the enrichment strategy introduced here, in this sense, generalizes the enrichment strategy introduced Chapter 7.*

8.4 Numerical experiments

In this Section, we test the accuracy of the approximation produced by the finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$ obtained by enriching the standard simplicial linear finite element with the enrichment functionals defined in (8.37) and the sets of enrichment functions $\mathcal{E}_1$ and $\mathcal{E}_2'$, introduced in Example 8.2.9 and in Example 8.2.14, respectively. We perform the numerical experiments for $d = 2$ and for each experiment, we use a regular grid of $(n+1) \times (n+1)$ equispaced points, with $n = 2^k$, $k = 2, \ldots, 6$ and the relative Delaunay triangulation. We consider the following test functions

$$f_1(x, y) = \frac{1}{1 + x^2 + y^2}, \quad f_2(x, y) = e^{xy}, \quad f_3(x, y, z) = \sin(\pi xy),$$

and the following sets of enrichment functions

$$\mathcal{E}_1 = \left\{ e_i = \sin\left(\frac{\pi}{2}(\lambda_i + 2)\right) + 2 \,:\, i = 0, 1, 2 \right\}, \qquad \mathcal{E}_2' = \left\{ e_i = \frac{1+i}{1+\lambda_i} \prod_{\substack{k=0 \\ k \neq i}}^{2} \lambda_k \,:\, i = 0, 1, 2 \right\},$$

already introduced in Examples 8.2.9 and 8.2.14, respectively. For each of these, we compare the accuracy of approximation, in L^1-norm, produced by the standard simplicial linear finite element with that one produced by the enriched finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$. We perform the numerical experiments by using `MatLab` software. To compute the integral of a bivariate function over a face of the simplex S_d (for example, the evaluation of the enrichment functionals at the enrichment functions, needed in the construction of the matrix N) and the integral of a bivariate function over S_d (for example, the L^1-norm of the approximation error) we use the command `integral2`. The results are reported in Figures 8.1 - 8.3. We notice that for a fixed function f, not every set of enrichment functions significantly improves the accuracy of the approximation realized by the enriched finite element. The accuracy of the approximation depends on the chosen set of enrichment functions.

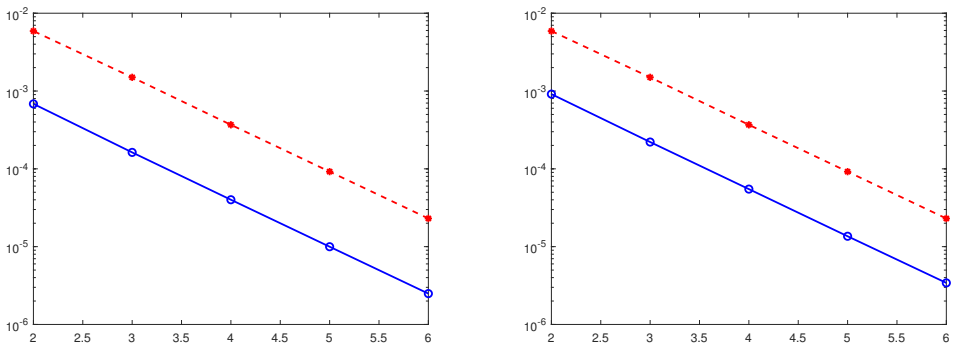


Figure 8.1: Semilog plot of the trend of the errors, in L^1-norm, produced by approximating the function $f_1(x, y)$ working with Delaunay triangulations of the unit square $[0, 1]^2$, the standard simplicial linear finite element (red dashed line), and the enriched finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$ (blue line). The Delaunay triangulations are realized by using regular grids of $(n + 1) \times (n + 1)$ equispaced nodes with $n = 2^k$, $k = 2, \ldots, 6$. The enrichments of the standard simplicial linear finite element are realized by using the functionals defined in (8.37) and the sets of enrichment functions $\mathcal{E}_1$ and $\mathcal{E}_2'$ for the left and the right picture, respectively.

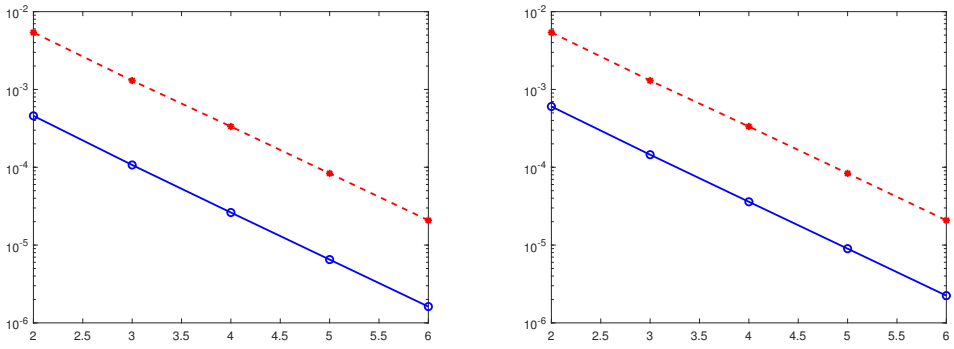


Figure 8.2: Semilog plot of the trend of the errors, in L^1-norm, produced by approximating the function $f_2(x, y)$ working with Delaunay triangulations of the unit square $[0, 1]^2$, the standard simplicial linear finite element (red dashed line), and the enriched finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$ (blue line). The Delaunay triangulations are realized by using regular grids of $(n + 1) \times (n + 1)$ equispaced nodes with $n = 2^k$, $k = 2, \ldots, 6$. The enrichments of the standard simplicial linear finite element are realized by using the functionals defined in (8.37) and the sets of enrichment functions $\mathcal{E}_1$ and $\mathcal{E}_2'$ for the left and the right picture, respectively.

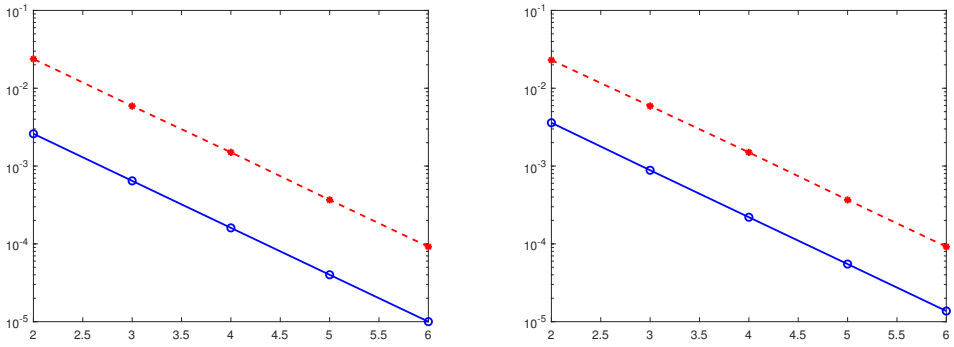


Figure 8.3: Semilog plot of the trend of the errors, in L^1-norm, produced by approximating the function $f_3(x, y)$ working with Delaunay triangulations of the unit square $[0, 1]^2$, the standard simplicial linear finite element (red dashed line), and the enriched finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \Sigma_{S_d}^{\text{enr}})$ (blue line). The Delaunay triangulations are realized by using regular grids of $(n + 1) \times (n + 1)$ equispaced nodes with $n = 2^k$, $k = 2, \ldots, 6$. The enrichments of the standard simplicial linear finite element are realized by using the functionals defined in (8.37) and the sets of enrichment functions $\mathcal{E}_1$ and $\mathcal{E}_2'$ for the left and the right picture, respectively.

Chapter 9

Improved methods for the enrichment and analysis of the simplicial vector-valued linear finite elements

The simplicial vector linear finite elements are commonly used for numerically solving the stationary Stokes equations. They are known, however, to suffer from severe shortcomings in application to more complicated situations. An enriched finite element, that overcomes the aforementioned drawbacks, was proposed and developed by Bernardi and Raugel in [7]. It can be regarded as an advanced and generalized version of the conventional simplicial vector linear finite element, and it has been employed in a wide range of practical engineering computational fields. It uses polynomials as enrichment functions. However, for some types of problems, these enrichment functions are not very efficient. In line with previous chapters, the main goal of this chapter is to present a general strategy for enriching the simplicial vector linear finite element with enrichment functions which are not necessarily polynomials. This enriched finite element can be seen as an extension of Bernardi and Raugel finite element. A key role is played by a characterization result, given in terms of the nonvanishing of a certain determinant, which provides necessary and sufficient conditions, on the enrichment functions and functionals, that guarantee the existence of families of such enriched elements. In conclusion, we present numerical tests that show the efficacy of the suggested enrichment strategy.

9.1 Bernardi–Raugel finite element

In the following, we denote by $\boldsymbol{x} \in \mathbb{R}^d$ the column vector of components $x_1, \ldots, x_d \in \mathbb{R}$, that is

$$\boldsymbol{x} = [x_1, \ldots, x_d]^T.$$

Let $S_d \subset \mathbb{R}^d$ be the d-simplex in $\mathbb{R}^d$ with vertices $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_d$ and barycentric coordinates $\lambda_0, \ldots, \lambda_d$. For $i = 0, \ldots, d$, we denote by F_i the face of S_d which does not contain the vertex $\boldsymbol{v}_i$ and by $\boldsymbol{n}_i = [n_{i_1}, \ldots, n_{i_d}]^T \in \mathbb{R}^d$ the unit outward normal to the face F_i. A finite element commonly used in the applications is the simplicial vector linear finite element. It is defined as

$$\boldsymbol{\mathcal{P}}_1(S_d) = (S_d, \mathbb{P}_1(S_d), \boldsymbol{\Sigma}^{\mathrm{lin}}_{S_d}), \tag{9.1}$$

where $\mathbb{P}_1(S_d)$ is the direct product d times of the vector space $\mathbb{P}_1(S_d)$, defined in (7.8), with itself and

$$\boldsymbol{\Sigma}^{\mathrm{lin}}_{S_d} = \{\boldsymbol{L}_j \,:\, j = 0, \ldots, d\},$$

with $\boldsymbol{L}_j$ defined as

$$\boldsymbol{L}_j(\boldsymbol{f}) = \boldsymbol{f}(\boldsymbol{v}_j) = [f_1(\boldsymbol{v}_j), \ldots, f_d(\boldsymbol{v}_j)]^T, \quad \boldsymbol{f} = [f_1, \ldots, f_d]^T, \qquad j = 0, \ldots, d.$$

We consider the enrichment functions

$$\boldsymbol{b}_i : S_d \to \mathbb{R}^d, \quad \boldsymbol{b}_i(\boldsymbol{x}) = \prod_{\substack{j=0 \\ j\neq i}}^{d} \lambda_j(\boldsymbol{x})\boldsymbol{n}_i, \qquad i = 0, \ldots, d. \tag{9.2}$$

The polynomial enrichment of the simplicial vector linear finite element proposed by Bernardi and Raugel is the triple

$$\boldsymbol{\mathcal{P}}_1^{\mathrm{BR}}(S_d) = (S_d, \mathbf{V}_1^{\mathrm{BR}}, \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}}), \tag{9.3}$$

where

$$\mathbf{V}_1^{\mathrm{BR}} = \mathbb{P}_1(S_d) \oplus \mathrm{span}\{\boldsymbol{b}_0, \ldots, \boldsymbol{b}_d\} \quad \text{and} \quad \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}} = \{\boldsymbol{L}_j, I_j \,:\, j = 0, \ldots, d\},$$

with the enrichment linear functionals

$$I_j(\boldsymbol{f}) = \frac{1}{|F_j|}\int_{F_j} \langle \boldsymbol{f}, \boldsymbol{n}_j\rangle \, d\sigma(\boldsymbol{x}), \qquad j = 0, \ldots, d. \tag{9.4}$$

Here we use letters in bold font to denote vector-valued functions and their associated spaces. In analogy, we do the same with vector-valued linear functionals and relative sets.

Observe that the number of information provided by the elements of the set $\boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}}$ can be computed as follows: the information related to the vector linear functionals $\boldsymbol{L}_j$ $j = 0, \ldots, d$, are $d(d+1)$ while those related to the enrichment linear functionals I_j, $j = 0, \ldots, d$ are $d+1$. Hence, the number of local degrees of freedom used for enrichment is

$$d(d+1) + (d+1) = (d+1)^2.$$

Moreover, as well known [7]

$$\dim(\mathbf{V}_1^{\mathrm{BR}}) = (d+1)^2.$$

The finite element (9.3) is known as Bernardi–Raugel finite element and it finds several applications in engineering, for example, it is used for the approximation of the Stokes problem, see [7].

9.2 Enrichment of the simplicial vector linear finite element

Now, we are interested in generalizing the Bernardi–Raugel finite element by extending the class of enrichment functions to a more general class of functions, which are not necessarily polynomials. To this aim, we consider $d+1$ linearly independent continuous functions $e_0, \ldots, e_d$ on S_d and the enrichment linear functionals defined in (9.4). We denote by

$$\mathbb{P}_1^{\mathrm{enr}}(S_d) = \mathbb{P}_1(S_d) \oplus \mathrm{span}\{\boldsymbol{e}_i = e_i\boldsymbol{n}_i \,:\, i = 0, \ldots, d\} \tag{9.5}$$

and we consider the triple

$$(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}}). \tag{9.6}$$

In line with previous chapters, we are tacitly assuming that the following technical condition is satisfied

$$\dim(\mathbb{P}_1^{\mathrm{enr}}(S_d)) = (d+1)^2. \tag{9.7}$$

In analogy to Chapter 8, for $k = 0, \ldots, d$, we introduce the functional

$$\mathcal{E}_k^{\mathrm{tra}} = \sum_{j=0}^{d} I_k\left(\lambda_j \boldsymbol{L}_j\right) - I_k, \tag{9.8}$$

and for $j = 0, \ldots, d$, we define the vector $\boldsymbol{g}^j \in \mathbb{R}^{2d+2}$ of components

$$g_i^j = 1, \qquad g_{i+d+1}^j = -\delta_{ij}, \qquad i = 0, \ldots, d. \tag{9.9}$$

For $k = 0, \ldots, d$, and $\boldsymbol{f} = [f_1, \ldots, f_d]^T$, $f_i \in C(S_d)$, $i = 1, \ldots, d$, we consider the vector $\boldsymbol{M}^{\{k\}}(\boldsymbol{f})$ of components

$$M_0^{\{k\}}(\boldsymbol{f}) = I_k(\lambda_0 \boldsymbol{L}_0(\boldsymbol{f})), \ldots, M_d^{\{k\}}(\boldsymbol{f}) = I_k(\lambda_d \boldsymbol{L}_d(\boldsymbol{f})), \qquad M_{j+d+1}^{\{k\}}(\boldsymbol{f}) = I_j(\boldsymbol{f}) \qquad j = 0, \ldots, d. \tag{9.10}$$

Therefore, the functionals $\mathcal{E}_k^{\mathrm{tra}}$ defined in (9.8) can be expressed in terms of $\boldsymbol{g}^k$ and $\boldsymbol{M}^{\{k\}}(\boldsymbol{f})$ as

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) = \left\langle \boldsymbol{g}^k, \boldsymbol{M}^{\{k\}}(\boldsymbol{f}) \right\rangle, \tag{9.11}$$

where $\langle \cdot, \cdot \rangle$ is the usual scalar product on $\mathbb{R}^{2d+2}$.

Remark 9.2.1. *In the particular case where $\boldsymbol{f} \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$ and $\boldsymbol{f}$ vanishes at all vertices of S_d the formula* (9.11) *becomes*

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) = \left\langle \boldsymbol{g}^k, \boldsymbol{M}^{\{k\}}(\boldsymbol{f}) \right\rangle = -I_k(\boldsymbol{f}).$$

Lemma 9.2.2. *Let $\boldsymbol{f} = [f_1, \ldots, f_d]^T \in \mathbb{P}_1(S_d)$. Then, for any $k = 0, \ldots, d$, we have*

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) = 0. \tag{9.12}$$

Proof. Since $f_i \in \mathbb{P}_1(S_d)$, we have

$$f_i = \sum_{j=0}^{d} \lambda_j f_i(\boldsymbol{v}_j), \qquad i = 1, \ldots, d,$$

and then

$$\boldsymbol{f} = \sum_{j=0}^{d} \lambda_j \boldsymbol{f}(\boldsymbol{v}_j).$$

From this equality and using the linearity of the enrichment linear functionals I_k, $k = 0, \ldots, d$, we immediately get

$$\begin{aligned} \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) &= \sum_{j=0}^{d} I_k(\lambda_j \boldsymbol{L}_j(\boldsymbol{f})) - I_k(\boldsymbol{f}) \\ &= \sum_{j=0}^{d} I_k(\lambda_j \boldsymbol{f}(\boldsymbol{v}_j)) - \sum_{j=0}^{d} I_k(\lambda_j \boldsymbol{f}(\boldsymbol{v}_j)) \\ &= 0, \qquad k = 0, \ldots, d. \end{aligned}$$

□

The previous lemma implies the following characterization result for the enrichment functions $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_d$, so that the triple

$$(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$$

is a finite element, or equivalently so that $\mathbb{P}_1^{\mathrm{enr}}(S_d)$ is $\boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}}$-unisolvent.

Theorem 9.2.3. *Let*

$$N = \begin{bmatrix} -\mathcal{E}_0^{\mathrm{tra}}(\boldsymbol{e}_0) & \ldots & -\mathcal{E}_0^{\mathrm{tra}}(\boldsymbol{e}_d) \\ -\mathcal{E}_1^{\mathrm{tra}}(\boldsymbol{e}_0) & \ldots & -\mathcal{E}_1^{\mathrm{tra}}(\boldsymbol{e}_d) \\ \vdots & \vdots & \vdots \\ -\mathcal{E}_d^{\mathrm{tra}}(\boldsymbol{e}_0) & \ldots & -\mathcal{E}_d^{\mathrm{tra}}(\boldsymbol{e}_d) \end{bmatrix}, \tag{9.13}$$

then the triple $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ is a finite element if and only

$$\det(N) \neq 0.$$

Proof. Let us assume that $\det(N) \neq 0$ and we prove that $\mathbb{P}_1^{\mathrm{enr}}(S_d)$ is $\boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}}$-unisolvent. Let $\boldsymbol{f} \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$ such that

$$\boldsymbol{L}_j(\boldsymbol{f}) = \boldsymbol{0}, \qquad j = 0, \ldots, d, \tag{9.14}$$

$$I_j(\boldsymbol{f}) = 0, \qquad j = 0, \ldots, d. \tag{9.15}$$

Since $\boldsymbol{f} \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$, it can be expressed as

$$\boldsymbol{f} = \boldsymbol{p} + \sum_{i=0}^{d} \beta_i \boldsymbol{e}_i,$$

where $\boldsymbol{p} \in \mathbb{P}_1(S_d)$ and $\beta_0, \ldots, \beta_d$ are real numbers. Since $\boldsymbol{f}$ satisfies (9.14) and (9.15), by definition (9.8), we get

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) = 0.$$

Then by (9.12) of Lemma 9.2.2, we obtain

$$\begin{aligned} 0 &= \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) \\ &= \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{p}) + \sum_{i=0}^{d} \beta_i \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i) \\ &= \sum_{i=0}^{d} \beta_i \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i), \qquad k = 0, \ldots, d. \end{aligned} \tag{9.16}$$

Equation (9.16) can be represented in matrix form as

$$-N\boldsymbol{\beta} = \boldsymbol{0},$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^T$. Since, by hypothesis, the matrix N is nonsingular, we get $\beta_0 = \beta_1 = \cdots = \beta_d = 0$ and therefore $\boldsymbol{f} = \boldsymbol{p}$. Taking into account that $\boldsymbol{f}$ vanishes at the vertices of the simplex S_d, by (9.14), we find that $\boldsymbol{f} = \boldsymbol{0}$.
In order to prove the reverse implication, let us assume to the contrary that

$$\det(N) = 0.$$

Since the determinant of N is equal to the determinant of its transpose N^T, then there exist $\gamma_0, \ldots, \gamma_d$ not all zero such that the functional

$$\mathcal{E}^{\mathrm{tra}} = \sum_{k=0}^{d} \gamma_k \mathcal{E}_k^{\mathrm{tra}}$$

vanishes at the enrichment functions $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_d$. By the linearity of $\mathcal{E}^{\mathrm{tra}}$ and by (9.12) of Lemma 9.2.2, we deduce that $\mathcal{E}^{\mathrm{tra}}$ vanishes on the whole space $\mathbb{P}_1^{\mathrm{enr}}(S_d)$. Therefore, for any $\boldsymbol{f} \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$, from (9.11), we have

$$\begin{aligned} 0 = \mathcal{E}^{\mathrm{tra}}(\boldsymbol{f}) &= \sum_{k=0}^{d} \gamma_k \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{f}) \\ &= \sum_{k=0}^{d} \gamma_k \left\langle \boldsymbol{g}^k, \boldsymbol{M}^{\{k\}}(\boldsymbol{f}) \right\rangle. \end{aligned} \tag{9.17}$$

Since, by hypothesis, $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ is a finite element, there exist $\boldsymbol{f}_i \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$, $i = 0, \ldots, d$, such that $\boldsymbol{L}_j(\boldsymbol{f}_i) = \boldsymbol{0}$ and $I_j(\boldsymbol{f}_i) = \delta_{ij}$, $j = 0, \ldots, d$. Consequently, by Remark 9.2.1, we get

$$\left\langle \boldsymbol{g}^k, \boldsymbol{M}^{\{k\}}(\boldsymbol{f}_i) \right\rangle = -I_k(\boldsymbol{f}_i) = -\delta_{ik}. \tag{9.18}$$

Finally, by substituting (9.18) in (9.17), we get

$$\begin{aligned} 0 &= \sum_{k=0}^{d} \gamma_k \left\langle \boldsymbol{g}^k, \boldsymbol{M}^{\{k\}}(\boldsymbol{f}_i) \right\rangle \\ &= -\sum_{k=0}^{d} \gamma_k \delta_{ik} = -\gamma_i, \qquad i = 0, \ldots, d, \end{aligned}$$

and then we have

$$\gamma_0 = \cdots = \gamma_d = 0,$$

which is a contradiction. □

The following remarks are an immediate consequence of Theorem 9.2.3 and Remark 9.2.1.

Remark 9.2.4. *If the enrichment functions $e_0, \ldots, e_d$ satisfy the vanishing conditions at the vertices of S_d, that is*

$$e_i(\boldsymbol{v}_j) = 0, \quad i, j = 0, \ldots, d, \tag{9.19}$$

or, equivalently,

$$\boldsymbol{e}_i(\boldsymbol{v}_j) = e_i(\boldsymbol{v}_j)\boldsymbol{n}_i = 0, \quad i, j = 0, \ldots, d,$$

then, the matrix N introduced in (9.13) *becomes*

$$N = \begin{bmatrix} I_0(\boldsymbol{e}_0) & \ldots & I_0(\boldsymbol{e}_d) \\ I_1(\boldsymbol{e}_0) & \ldots & I_1(\boldsymbol{e}_d) \\ \vdots & \vdots & \vdots \\ I_d(\boldsymbol{e}_0) & \ldots & I_d(\boldsymbol{e}_d) \end{bmatrix}. \tag{9.20}$$

Remark 9.2.5. *Let $\boldsymbol{b}_i$, $i = 0, \ldots, d$, be the enrichment functions defined in* (9.2). *These enrichment functions satisfy the vanishing conditions* (9.19) *and the conditions*

$$I_i(\boldsymbol{b}_i) \neq 0, \qquad i = 0, \ldots, d,$$

$$I_i(\boldsymbol{b}_j) = 0, \qquad i, j = 0, \ldots, d, \quad i \neq j.$$

Then, the matrix N defined in (9.13) *is a diagonal matrix with determinant different from zero. Then, by Theorem 9.2.3, we can simply prove that the triple* (9.3) *is a finite element.*

In the following, we assume that the matrix N is nonsingular and we denote its inverse by

$$N^{-1} = [\boldsymbol{c}_0 \ldots \boldsymbol{c}_d], \tag{9.21}$$

where $\boldsymbol{c}_i \in \mathbb{R}^d$, $i = 0, \ldots, d$, are column vectors. A direct consequence of Theorem 9.2.3 is the linear independence of the functionals of $\boldsymbol{\Sigma}_{S_d}^{\text{enr}}$ in the dual space $\mathbb{P}_1^{\text{enr}}(S_d)^\star$ [23, Ch 2]. Then, there exists a basis $\{\boldsymbol{\varphi}_{j\ell}, \boldsymbol{\psi}_j : j = 0, \ldots, d, \ell = 1, \ldots, d\}$ of $\mathbb{P}_1^{\text{enr}}(S_d)$ which satisfy

$$\boldsymbol{L}_j(\boldsymbol{\varphi}_{i\ell}) = \delta_{ij}\boldsymbol{u}_\ell, \quad I_j(\boldsymbol{\varphi}_{i\ell}) = 0, \qquad i, j = 0, \ldots, d, \qquad \ell = 1, \ldots, d, \tag{9.22}$$

$$\boldsymbol{L}_j(\boldsymbol{\psi}_i) = \boldsymbol{0}, \quad I_j(\boldsymbol{\psi}_i) = \delta_{ij}, \qquad i, j = 0, \ldots, d, \tag{9.23}$$

where $\boldsymbol{u}_\ell$, $\ell = 1, \ldots, d$ is the canonical basis of $\mathbb{R}^d$. In the following, we derive explicit expressions for such basis functions.

Theorem 9.2.6. *The basis functions $\{\boldsymbol{\varphi}_{j\ell}, \boldsymbol{\psi}_j : j = 0, \ldots, d, \ell = 1, \ldots, d\}$ of $\mathbb{P}_1^{\text{enr}}(S_d)$ associated to the finite element $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{enr}})$ which satisfy* (9.22) *and* (9.23) *have the following expressions*

$$\boldsymbol{\varphi}_{j\ell} = \lambda_j \boldsymbol{u}_\ell - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{\psi}_k, \qquad j = 0, \ldots, d, \qquad \ell = 1, \ldots, d, \tag{9.24}$$

$$\boldsymbol{\psi}_j = \left(\boldsymbol{E} - \sum_{k=0}^{d} \lambda_k \boldsymbol{E}(\boldsymbol{v}_k) \right) \boldsymbol{c}_j, \qquad j = 0, \ldots, d, \tag{9.25}$$

where $\boldsymbol{E}$ *is the* $d \times (d+1)$ *matrix defined by*

$$\boldsymbol{E} = [\boldsymbol{e}_0 \ldots \boldsymbol{e}_d]. \tag{9.26}$$

Proof. Since $\boldsymbol{\varphi}_{j\ell} \in \mathbb{P}_1^{\mathrm{enr}}(S_d)$, it can be expressed as

$$\boldsymbol{\varphi}_{j\ell} = \boldsymbol{p}_{j\ell} + \sum_{s=0}^{d} \beta_s \boldsymbol{e}_s, \tag{9.27}$$

or, equivalently,

$$\boldsymbol{\varphi}_{j\ell} = \boldsymbol{p}_{j\ell} + \boldsymbol{E}\boldsymbol{\beta}, \tag{9.28}$$

where $\boldsymbol{p}_{j\ell} \in \mathbb{P}_1(S_d)$, $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_d]^T \in \mathbb{R}^{d+1}$ and $\boldsymbol{E}$ is the matrix defined in (9.26). By (9.8), we have

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{\varphi}_{j\ell}) = \sum_{i=0}^{d} I_k(\lambda_i \boldsymbol{L}_i(\boldsymbol{\varphi}_{j\ell})) - I_k(\boldsymbol{\varphi}_{j\ell}), \qquad k = 0, \ldots, d,$$

and then, by (9.22)

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{\varphi}_{j\ell}) = \sum_{i=0}^{d} I_k(\lambda_i \delta_{ij} \boldsymbol{u}_\ell) = I_k(\lambda_j \boldsymbol{u}_\ell), \qquad k = 0, \ldots, d. \tag{9.29}$$

We apply $\mathcal{E}_k^{\mathrm{tra}}$ to both members of (9.27) and by previous equation we get

$$I_k(\lambda_j \boldsymbol{u}_\ell) = \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{p}_{j\ell}) + \sum_{s=0}^{d} \beta_s \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_s), \qquad k = 0, \ldots, d.$$

Then, by (9.12) of Lemma 9.2.2, we obtain

$$I_k(\lambda_j \boldsymbol{u}_\ell) = \sum_{s=0}^{d} \beta_s \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_s), \qquad k = 0, \ldots, d,$$

or, equivalently

$$\boldsymbol{I}_{j\ell} = -N\boldsymbol{\beta},$$

where $\boldsymbol{I}_{j\ell} = [I_0(\lambda_j \boldsymbol{u}_\ell), \ldots, I_d(\lambda_j \boldsymbol{u}_\ell)]^T$. Since we are assuming that the matrix N is invertible, by (9.21), we get

$$\boldsymbol{\beta} = -N^{-1}\boldsymbol{I}_{j\ell} = -\sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{c}_k,$$

and by substituting into equation (9.28), we have

$$\boldsymbol{\varphi}_{j\ell} = \boldsymbol{p}_{j\ell} - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{E}\boldsymbol{c}_k. \tag{9.30}$$

Now we apply $\boldsymbol{L}_i$, $i = 0, \ldots, d$ to both members of (9.30), by (9.22), we get

$$\boldsymbol{p}_{j\ell}(\boldsymbol{v}_i) = \delta_{ij}\boldsymbol{u}_\ell + \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{E}(\boldsymbol{v}_i)\boldsymbol{c}_k, \qquad i = 0, \ldots, d, \tag{9.31}$$

and then

$$\boldsymbol{p}_{j\ell} = \sum_{i=0}^{d} \lambda_i \boldsymbol{p}_{i\ell}(\boldsymbol{v}_i) = \lambda_j \boldsymbol{u}_\ell + \sum_{i=0}^{d}\sum_{k=0}^{d} \lambda_i I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{E}(\boldsymbol{v}_i)\boldsymbol{c}_k. \tag{9.32}$$

Finally, by substituting (9.32) in (9.30) and by changing the order of the summation, we get

$$\begin{aligned}\boldsymbol{\varphi}_{j\ell} &= \lambda_j \boldsymbol{u}_\ell + \sum_{i=0}^{d}\sum_{k=0}^{d} \lambda_i I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{E}(\boldsymbol{v}_i)\boldsymbol{c}_k - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{E}\boldsymbol{c}_k \\ &= \lambda_j \boldsymbol{u}_\ell + \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\sum_{i=0}^{d} \lambda_i \boldsymbol{E}(\boldsymbol{v}_i)\boldsymbol{c}_k - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{E}\boldsymbol{c}_k \\ &= \lambda_j \boldsymbol{u}_\ell - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\left(\boldsymbol{E} - \sum_{i=0}^{d} \lambda_i \boldsymbol{E}(\boldsymbol{v}_i)\right)\boldsymbol{c}_k.\end{aligned}$$

It remains to prove (9.25). We proceed in analogy to the previous case and then we set

$$\boldsymbol{\psi}_j = \boldsymbol{p}_j + \sum_{i=0}^{d} \gamma_i \boldsymbol{e}_i, \tag{9.33}$$

or, equivalently,

$$\boldsymbol{\psi}_j = \boldsymbol{p}_j + \boldsymbol{E}\boldsymbol{\gamma}, \tag{9.34}$$

where $\boldsymbol{p}_j \in \mathbb{P}_1(S_d)$ and $\boldsymbol{\gamma} = [\gamma_0, \ldots, \gamma_d]^T \in \mathbb{R}^{d+1}$. By (9.8), we have

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{\psi}_j) = \sum_{i=0}^{d} I_k(\lambda_i \boldsymbol{L}_i(\boldsymbol{\psi}_j)) - I_k(\boldsymbol{\psi}_j), \qquad k = 0, \ldots, d, \tag{9.35}$$

and then, by (9.23)

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{\psi}_j) = -\delta_{jk}, \qquad k = 0, \ldots, d. \tag{9.36}$$

We apply $\mathcal{E}_k^{\mathrm{tra}}$ to both members of (9.33) and by previous equation we get

$$-\delta_{jk} = \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{p}_j) + \sum_{i=0}^{d} \gamma_i \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i), \qquad k = 0, \ldots, d.$$

Then, by (9.12) of Lemma 9.2.2, we obtain

$$-\delta_{jk} = \sum_{i=0}^{d} \gamma_i \mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i), \qquad k = 0, \ldots, d,$$

or, equivalently

$$\boldsymbol{u}_j = N\boldsymbol{\gamma}.$$

Therefore we get

$$\boldsymbol{\gamma} = N^{-1}\boldsymbol{u}_j = \boldsymbol{c}_j,$$

and by substituting into equation (9.34), we have

$$\boldsymbol{\psi}_j = \boldsymbol{p}_j + \boldsymbol{E}\boldsymbol{c}_j. \tag{9.37}$$

Now we apply $\boldsymbol{L}_i$, $i = 0, \ldots, d$, to both members of (9.37), by (9.23) we get

$$\boldsymbol{p}_j(\boldsymbol{v}_i) = -\boldsymbol{E}(\boldsymbol{v}_i)\boldsymbol{c}_j,$$

and then

$$\boldsymbol{p}_j = \sum_{i=0}^{d} \lambda_i \boldsymbol{p}_j(\boldsymbol{v}_i) = -\sum_{i=0}^{d} \lambda_i \boldsymbol{E}(\boldsymbol{v}_i)\boldsymbol{c}_j. \tag{9.38}$$

By substituting (9.38) in (9.37), (9.25) is proved. □

Remark 9.2.7. *Let us assume that $e_0, \ldots, e_d$ satisfy the vanishing conditions* (9.19)*. Then, the basis functions of the finite element $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ are*

$$\boldsymbol{\varphi}_{j\ell} = \lambda_j \boldsymbol{u}_\ell - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{\psi}_k, \qquad j = 0, \ldots, d, \qquad \ell = 1, \ldots, d, \tag{9.39}$$

$$\boldsymbol{\psi}_j = \boldsymbol{E}\boldsymbol{c}_j, \qquad j = 0, \ldots, d. \tag{9.40}$$

We denote by $\boldsymbol{C}(S_d)$ the direct product d times of $C(S_d)$ with itself.

Theorem 9.2.8. *The linear approximation operator based on the simplicial vector linear finite element $\boldsymbol{\mathcal{P}}_1(S_d)$, defined in* (9.1)

$$\begin{aligned} \boldsymbol{\Pi}^{\mathrm{lin}} : \boldsymbol{C}(S_d) &\to \mathbb{P}_1(S_d) \\ \boldsymbol{f} &\mapsto \sum_{j=0}^{d} \lambda_j \boldsymbol{L}_j(\boldsymbol{f}), \end{aligned} \tag{9.41}$$

reproduces linear polynomials and satisfies the interpolation conditions

$$\boldsymbol{L}_j\left(\boldsymbol{\Pi}^{\mathrm{lin}}[\boldsymbol{f}]\right) = \boldsymbol{L}_j(\boldsymbol{f}), \qquad j = 0, \ldots, d. \tag{9.42}$$

Proof. The proof follows from the Lagrange property of the barycentric coordinates, that is $\lambda_i(\boldsymbol{v}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta operator. □

In the following, we denote by $\boldsymbol{E}^{\mathrm{lin}}$ the approximation error of the operator $\boldsymbol{\Pi}^{\mathrm{lin}}$, that is

$$\boldsymbol{E}^{\mathrm{lin}}[\boldsymbol{f}] = \boldsymbol{f} - \boldsymbol{\Pi}^{\mathrm{lin}}[\boldsymbol{f}], \qquad \boldsymbol{f} \in \boldsymbol{C}(S_d). \tag{9.43}$$

9.3 Admissible enrichment functions

In this section, we collect sets of *admissible enrichment functions*, that is functions for which the triple $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ is a finite element. The main issue when using the proposed enriched method is to ensure that the matrix N given in (9.13) is invertible. These enrichment functions constitute a very general class, which can be used for many types of applications. Before starting, we first give some important definitions.

Definition 9.3.1. *For $d \geq 1$ the dihedral angle α_{ij} between two faces of S_d, F_i and F_j is defined by means of the inner product of their outward unit normals $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$ [12]*

$$\cos(\alpha_{ij}) = -\langle \boldsymbol{n}_i, \boldsymbol{n}_j \rangle .$$

In the following, we denote by

$$\rho_{ij} = \langle \boldsymbol{n}_i, \boldsymbol{n}_j \rangle , \qquad i, j = 0, \ldots, d. \tag{9.44}$$

Definition 9.3.2. *A simplex is said to be acute if all its dihedral angles satisfy*

$$0 < \alpha_{ij} < \pi/2, \tag{9.45}$$

and then, by (9.44)*, we get*

$$-1 < \rho_{ij} < 0. \tag{9.46}$$

An acute triangulation is a triangulation into acute simplices. The problem of finding acute triangulations has many applications in computational geometry, including mesh generation, finite element analysis, see e.g. [51, Ch. 33].

Definition 9.3.3. *A flag of faces of a convex polytope K_d in $\mathbb{R}^d$ is a sequence*

$$F^{\{0\}} \subset F^{\{1\}} \subset \cdots \subset F^{\{d\}} = K_d,$$

where $F^{\{i\}}$ is a face of K_d of dimension i, $i = 0, \ldots, d$ [21].

Definition 9.3.4. *A convex polytope K_d in $\mathbb{R}^d$ is called regular if the group $G(K_d)$ of isometries of $\mathbb{R}^d$ which leave K_d invariant acts transitively on the set of flags of faces of K_d [21]. A regular simplex is a simplex that is also a regular polytope.*

It is well-known that dihedral angles of a regular simplex S_d^{reg} in $\mathbb{R}^d$ satisfy the following relation [12]

$$\alpha_{ij} = \arccos\left(\frac{1}{d}\right), \qquad i,j = 0,\dots,d,\ i \neq j. \tag{9.47}$$

9.3.1 Admissible enrichment functions of the first class

Let $n \in \mathbb{N}$ and let $f_1,\dots,f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero and let us assume that at least one function is strictly convex. We consider the enrichment functions

$$e_i = \prod_{\ell=1}^{n} f_\ell(\lambda_i), \qquad \boldsymbol{e}_i = e_i \boldsymbol{n}_i, \qquad i = 0,\dots,d. \tag{9.48}$$

Theorem 9.3.5. *Let $n \in \mathbb{N}$ and let $f_1,\dots,f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero and let us assume that at least one function is strictly convex. We assume that the simplex S_d satisfies*

$$\rho_{ij} = \rho \neq 0, \qquad i,j = 0,\dots,d-1, \qquad \rho_{id} = \eta \neq 0, \qquad i = 0,\dots,d-1, \tag{9.49}$$

where ρ_{ij}, $i,j = 0,\dots,d$, $i \neq j$ is defined in (9.44)*. Let $\boldsymbol{e}_i$, I_j, $i,j = 0,\dots,d$, be the enrichment functions and the enrichment linear functionals defined in* (9.48) *and* (9.4)*, respectively. Then, the matrix N defined in* (9.13) *is nonsingular.*

Proof. Using the linearity of I_j, $j = 0,\dots,d$, by (9.4) and (9.8), we get

$$\begin{aligned}
\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i) &= \sum_{j=0}^{d} I_k(\lambda_j \boldsymbol{L}_j(\boldsymbol{e}_i)) - I_k(\boldsymbol{e}_i) \\
&= I_k\left(\sum_{j=0}^{d} \lambda_j \boldsymbol{e}_i(\boldsymbol{v}_j) - \boldsymbol{e}_i\right) \\
&= I_k\left(\left(\sum_{j=0}^{d} \lambda_j e_i(\boldsymbol{v}_j) - e_i\right)\boldsymbol{n}_i\right) \\
&= \frac{1}{|F_k|}\int_{F_k}\left(\sum_{j=0}^{d} \lambda_j(\boldsymbol{x}) e_i(\boldsymbol{v}_j) - e_i(\boldsymbol{x})\right)\langle \boldsymbol{n}_i, \boldsymbol{n}_k\rangle \, d\sigma(\boldsymbol{x}) \\
&= \frac{1}{|F_k|}\left(\sum_{j=0}^{d}\left(\int_{F_k}\lambda_j(\boldsymbol{x})d\sigma(\boldsymbol{x})\right) e_i(\boldsymbol{v}_j) - \int_{F_k} e_i(\boldsymbol{x})d\sigma(\boldsymbol{x})\right)\langle \boldsymbol{n}_i, \boldsymbol{n}_k\rangle \\
&= -\mu_{ik}\langle \boldsymbol{n}_i, \boldsymbol{n}_k\rangle .
\end{aligned}$$

In this case, by Theorem 8.2.3, the matrix N defined in (9.13) becomes

$$N = \begin{bmatrix} 0 & \mu_0\rho & \cdots & \mu_0\eta \\ \mu_1\rho & 0 & \cdots & \mu_1\eta \\ \vdots & \vdots & \ddots & \vdots \\ \mu_d\eta & \mu_d\eta & \cdots & 0 \end{bmatrix},$$

where we write μ_k instead of μ_{ik} since, by equation (8.46) of Remark 8.2.6, this term does not depend on i. Then, we have

$$\det(N) = \det \begin{bmatrix} 0 & \rho & \cdots & \eta \\ \rho & 0 & \cdots & \eta \\ \vdots & \vdots & \ddots & \vdots \\ \eta & \eta & \cdots & 0 \end{bmatrix} \prod_{i=0}^{d} \mu_i. \tag{9.50}$$

By substituting the i-th row, R_i with $R_i - R_{i+1}$, $i = 1, \ldots, d-1$, and developing the determinant with respect the last column, we get

$$\det(G) = \det \begin{bmatrix} -\rho & \rho & \cdots & 0 \\ 0 & -\rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & \eta \\ \eta & \eta & \cdots & 0 \end{bmatrix} \prod_{i=0}^{d} \mu_i = (-1)^{2d-1}\eta^2 \det \begin{bmatrix} -\rho & \rho & \cdots & 0 \\ 0 & -\rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho \\ 1 & 1 & \cdots & 1 \end{bmatrix} \prod_{i=0}^{d} \mu_i.$$

Finally, by multiplying the last row R_{d+1} by ρ and by substituting R_{d+1} with $R_{d+1} + \sum_{j=1}^{d} jR_j$, we get

$$\det(N) = (-1)^{2d-1}\frac{\eta^2}{\rho} \det \begin{bmatrix} -\rho & \rho & 0 & \cdots & 0 \\ 0 & -\rho & \rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\rho & \rho \\ 0 & 0 & \cdots & \cdots & d\rho \end{bmatrix} \prod_{i=0}^{d} \mu_i = d(-1)^d \eta^2 \rho^{d-1} \prod_{i=0}^{d} \mu_i.$$

By Theorem 8.2.3, we get

$$\mu_i \neq 0, \qquad i = 0, \ldots, d,$$

and then, by (9.49), the result follows. □

Remark 9.3.6. *We notice that if S_d^{reg} is a regular simplex, by* (9.47) *and* (9.44)*, we get*

$$\rho_{ij} = -\frac{1}{d}, \qquad i, j = 0, \ldots, d, \quad i \neq j.$$

Then, by using the enrichment functions and the enrichment linear functionals defined in (9.48) *and* (9.4)*, respectively, by Theorem 9.3.5, the matrix N defined in* (9.13) *is nonsingular.*

In the following, we consider the most common cases used in the applications, that is $d = 2$ and $d = 3$.

Theorem 9.3.7. *Let $n \in \mathbb{N}$ and let $f_1, \ldots, f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero and let us assume that at least one function is strictly convex. We assume that the simplex S_2 is an acute triangle. Let $\boldsymbol{e}_i$, I_j, $i, j = 0, 1, 2$, be the enrichment functions and the enrichment linear functionals defined in* (9.48) *and* (9.4)*, respectively. Then, the matrix N defined in* (9.13) *is nonsingular.*

Proof. By following the same line of the proof of Theorem 9.3.5, we get

$$N = \begin{bmatrix} 0 & \mu_0\rho_{10} & \mu_0\rho_{20} \\ \mu_1\rho_{10} & 0 & \mu_1\rho_{21} \\ \mu_2\rho_{20} & \mu_2\rho_{21} & 0 \end{bmatrix},$$

where ρ_{ij}, $i, j = 0, 1, 2$, $i \neq j$ is defined in (9.44). Then, we have

$$\det(N) = \mu_0\mu_1\mu_2 \det \begin{bmatrix} 0 & \rho_{10} & \rho_{20} \\ \rho_{10} & 0 & \rho_{21} \\ \rho_{20} & \rho_{21} & 0 \end{bmatrix} = 2\mu_0\mu_1\mu_2\rho_{10}\rho_{20}\rho_{21}. \tag{9.51}$$

Since S_2 is an acute simplex, we get

$$\rho_{10}\rho_{20}\rho_{21} \neq 0.$$

The theorem is proved by noting that, by Theorem 8.2.3

$$\mu_i \neq 0, \qquad i = 0, 1, 2.$$

□

Now, we want to extend Theorem 9.3.7 to the case $d = 3$. To this aim, we assume that the simplex S_3 is an acute tetrahedron. Then

$$\rho_{ij} < 0, \qquad i, j = 0, 1, 2, 3, \quad i \neq j.$$

We denote by

$$\zeta_{ij} = -\rho_{ij} \qquad i, j = 0, 1, 2, 3, \quad i \neq j \tag{9.52}$$

and

$$\xi_{ij} = \sqrt{\zeta_{ij}} \qquad i, j = 0, 1, 2, 3, \quad i \neq j. \tag{9.53}$$

Theorem 9.3.8. *Let $n \in \mathbb{N}$ and let $f_1, \ldots, f_n \in C([0,1])$ be convex, increasing (or decreasing) and nonnegative functions different from zero and let us assume that at least one function is strictly convex. We assume that the simplex S_3 is an acute tetrahedron. Let $\boldsymbol{e}_i$, I_j, $i, j = 0, 1, 2, 3$, be the enrichment functions and the enrichment linear functionals defined in* (9.48) *and* (9.4)*, respectively. Then, the matrix N defined in* (9.13) *is nonsingular.*

Proof. By following the same line of the proof of Theorem 9.3.5, we get

$$N = \begin{bmatrix} 0 & \mu_0\rho_{10} & \mu_0\rho_{20} & \mu_0\rho_{30} \\ \mu_1\rho_{10} & 0 & \mu_1\rho_{21} & \mu_1\rho_{31} \\ \mu_2\rho_{20} & \mu_2\rho_{21} & 0 & \mu_2\rho_{32} \\ \mu_3\rho_{30} & \mu_3\rho_{31} & \mu_3\rho_{32} & 0 \end{bmatrix},$$

where ρ_{ij}, $i, j = 0, 1, 2, 3$, $i \neq j$ is defined in (9.44). Then, by (9.52) and (9.53), we have

$$\det(N) = \mu_0\mu_1\mu_2\mu_3 \det \begin{bmatrix} 0 & -\xi_{10}^2 & -\xi_{20}^2 & -\xi_{30}^2 \\ -\xi_{10}^2 & 0 & -\xi_{21}^2 & -\xi_{31}^2 \\ -\xi_{20}^2 & -\xi_{21}^2 & 0 & -\xi_{32}^2 \\ -\xi_{30}^2 & -\xi_{31}^2 & -\xi_{32}^2 & 0 \end{bmatrix} = \mu_0\mu_1\mu_2\mu_3 \det \begin{bmatrix} 0 & \xi_{10}^2 & \xi_{20}^2 & \xi_{30}^2 \\ \xi_{10}^2 & 0 & \xi_{21}^2 & \xi_{31}^2 \\ \xi_{20}^2 & \xi_{21}^2 & 0 & \xi_{32}^2 \\ \xi_{30}^2 & \xi_{31}^2 & \xi_{32}^2 & 0 \end{bmatrix}.$$

After easy computations, we get

$$\begin{aligned} \det(N) &= \mu_0\mu_1\mu_2\mu_3(\xi_{30}\xi_{21} - \xi_{20}\xi_{31} - \xi_{10}\xi_{32})(\xi_{30}\xi_{21} + \xi_{20}\xi_{31} - \xi_{10}\xi_{32}) \times \\ &\quad (\xi_{30}\xi_{21} - \xi_{20}\xi_{31} + \xi_{10}\xi_{32})(\xi_{30}\xi_{21} + \xi_{20}\xi_{31} + \xi_{10}\xi_{32}) \\ &= \mu_0\mu_1\mu_2\mu_3(a - b - c)(a + b - c)(a - b + c)(a + b + c), \end{aligned}$$

where we set

$$a = \xi_{30}\xi_{21}, \qquad b = \xi_{20}\xi_{31}, \qquad c = \xi_{10}\xi_{32}.$$

By (9.53), we have

$$a > 0, \qquad b > 0, \qquad c > 0.$$

Then, by Heron's formula [83], we get

$$\det(N) = -16\mu_0\mu_1\mu_2\mu_3 |T|^2,$$

where T is the triangle with sides of length a, b, c. The theorem is proved by noting that, by Theorem 8.2.3

$$\mu_i \neq 0, \qquad i = 0, 1, 2, 3.$$

□

Example 9.3.9. *Previous theorems allow us to enrich the simplicial vector linear finite element to the finite element* $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ *by using the following sets of enrichment functions:*

- $\mathcal{E}_1 = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \sin\left(\frac{\pi}{2}\left(\lambda_i + 2\right)\right) + 2 \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_2 = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i\, e_i = \frac{1}{1+\lambda_i} \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_3 = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = e^{\lambda_i} \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_4 = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \lambda_i^{\alpha},\ \alpha > 1 \,:\, i = 0, \ldots, d \right\}$,
- $\mathcal{E}_5 = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \lambda_i^{\alpha} e^{\lambda_i},\ \alpha > 1 \,:\, i = 0, \ldots, d \right\}$.

9.3.2 Admissible enrichment functions of the second class

Let $f_0, \ldots, f_d \in C([0,1])$ be continuous functions such that $f_i(0) \neq 0$, $i = 0, \ldots, d$, and let $\alpha_0, \ldots, \alpha_d$ be real numbers greater than one. We consider the enrichment functions

$$e_i = f_i(\lambda_i(\boldsymbol{x})) \prod_{\substack{k=0 \\ k \neq i}}^{d} \lambda_k^{\alpha_k - 1}(\boldsymbol{x}), \qquad \boldsymbol{e}_i = e_i \boldsymbol{n}_i, \qquad i = 0, \ldots, d. \tag{9.54}$$

Theorem 9.3.10. *Let* $f_0, \ldots, f_d \in C([0,1])$ *be continuous functions such that* $f_i(0) \neq 0$, $i = 0, \ldots, d$, *and let* $\alpha_0, \ldots, \alpha_d$ *be real numbers greater than one. Let* $\boldsymbol{e}_i$, I_j, $i, j = 0, \ldots, d$, *be the enrichment functions and the enrichment linear functionals defined in* (9.54) *and* (9.4), *respectively. Then the matrix* N *defined in* (9.13) *is nonsingular.*

Proof. We prove that

$$\mathcal{E}_i^{\mathrm{tra}}(\boldsymbol{e}_i) \neq 0, \qquad i = 0, \ldots, d,$$

and

$$\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i) = 0, \qquad i, k = 0, \ldots, d, \quad i \neq k.$$

Using the linearity of I_k, $k = 0, \ldots, d$, by (9.4) and (9.8), we get

$$\begin{aligned}
\mathcal{E}_k^{\mathrm{tra}}(\boldsymbol{e}_i) &= \sum_{j=0}^{d} I_k(\lambda_j \boldsymbol{L}_j(\boldsymbol{e}_i)) - I_k(\boldsymbol{e}_i) \\
&= I_k\left(\sum_{j=0}^{d} \lambda_j \boldsymbol{e}_i(\boldsymbol{v}_j) - \boldsymbol{e}_i \right) \\
&= I_k\left(\left(\sum_{j=0}^{d} \lambda_j e_i(\boldsymbol{v}_j) - e_i \right) \boldsymbol{n}_i \right) \\
&= \frac{1}{|F_k|} \int_{F_k} \left(\sum_{j=0}^{d} \lambda_j(\boldsymbol{x}) e_i(\boldsymbol{v}_j) - e_i(\boldsymbol{x}) \right) \langle \boldsymbol{n}_i, \boldsymbol{n}_k \rangle \, d\sigma(\boldsymbol{x}) \\
&= \frac{1}{|F_k|} \left(\sum_{j=0}^{d} \left(\int_{F_k} \lambda_j(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \right) e_i(\boldsymbol{v}_j) - \int_{F_k} e_i(\boldsymbol{x}) d\sigma(\boldsymbol{x}) \right) \langle \boldsymbol{n}_i, \boldsymbol{n}_k \rangle .
\end{aligned}$$

The result follows by Theorem 8.2.12. □

In the next theorem, we give an explicit expression, in closed-form, of the basis functions associated to the finite element enriched with the functionals defined in (9.4) and the enrichment functions defined in (9.54).

Theorem 9.3.11. *Let $f_0, \dots, f_d \in C([0,1])$ be continuous functions such that $f_i(0) \neq 0$, $i = 0, \dots, d$, and let $\alpha_0, \dots, \alpha_d$ be real numbers greater than one. Let $\boldsymbol{e}_i$, I_j, $i, j = 0, \dots, d$, be the enrichment functions and the enrichment linear functionals defined in* (9.54) *and* (9.4)*, respectively. Then, the basis functions* (9.24) *and* (9.25) *of the finite element $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ have the following expressions*

$$\boldsymbol{\varphi}_{j\ell} = \lambda_j \boldsymbol{u}_\ell - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell)\boldsymbol{\psi}_k, \qquad j = 0, \dots, d, \quad \ell = 1, \dots, d, \tag{9.55}$$

$$\boldsymbol{\psi}_j = \frac{\boldsymbol{e}_j}{\mu_j}, \qquad j = 0, \dots, d, \tag{9.56}$$

where

$$\mu_j = -\mathcal{E}_j^{\mathrm{tra}}(\boldsymbol{e}_j), \qquad j = 0, \dots, d. \tag{9.57}$$

Proof. Since the enrichment functions (9.54) satisfy the vanishing conditions (9.19), they can be written as (9.39) and (9.40). Then we compute the j-th column $\boldsymbol{c}_j$ of the inverse of the matrix N defined in (9.13). By Theorem 9.3.10, we get

$$N = \begin{bmatrix} \mu_0 & 0 & \cdots & 0 \\ 0 & \mu_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_d \end{bmatrix}.$$

Therefore

$$N^{-1} = \begin{bmatrix} \frac{1}{\mu_0} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mu_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\mu_d} \end{bmatrix}, \tag{9.58}$$

and then we get

$$\boldsymbol{E}\boldsymbol{c}_j = \frac{\boldsymbol{e}_j}{\mu_j}. \tag{9.59}$$

The theorem is proved by combining (9.40) and (9.59). □

Example 9.3.12. *Theorem 9.3.10 allows us to enrich the simplicial vector linear finite element to the finite element $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ by using the following sets of enrichment functions:*

- $\mathcal{E}_1' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \cos(\lambda_i) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \dots, d \right\},$
- $\mathcal{E}_2' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \frac{1+i}{1+\lambda_i} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \dots, d \right\},$
- $\mathcal{E}_3' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \sin\left(\frac{\pi}{2i+2}(\lambda_i + 1)\right) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \dots, d \right\},$
- $\mathcal{E}_4' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = e^{i\lambda_i} \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \dots, d \right\},$
- $\mathcal{E}_5' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \log(i\lambda_i + 2) \prod_{\substack{\ell=0 \\ \ell \neq i}}^{d} \lambda_\ell^{\alpha_\ell - 1},\ \alpha_\ell > 1 \,:\, i, \ell = 0, \dots, d \right\}.$

9.4 Error estimates

9.4.1 An explicit error representation

Let $\boldsymbol{\Pi}^{\mathrm{enr}}$ be the approximation operator defined as

$$\begin{aligned} \boldsymbol{\Pi}^{\mathrm{enr}} : \boldsymbol{C}(S_d) &\to \mathbb{P}_1^{\mathrm{enr}}(S_d) \\ \boldsymbol{f} &\mapsto \sum_{\ell=1}^{d}\sum_{j=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, \boldsymbol{\varphi}_{j\ell} + \sum_{j=0}^{d} I_j(\boldsymbol{f}) \boldsymbol{\psi}_j, \end{aligned} \tag{9.60}$$

where $\boldsymbol{\varphi}_{j\ell}, \boldsymbol{\psi}_j$, $j = 0, \ldots, d$, $\ell = 1, \ldots, d$, are the basis functions introduced in Theorem 9.2.6. We provide explicit representation for the approximation error

$$\boldsymbol{E}^{\mathrm{enr}}[\boldsymbol{f}] = \boldsymbol{f} - \boldsymbol{\Pi}^{\mathrm{enr}}[\boldsymbol{f}]. \tag{9.61}$$

We now present a decomposition of the error $\boldsymbol{E}^{\mathrm{enr}}$ in terms of the approximation error associated to the approximation operator $\boldsymbol{\Pi}^{\mathrm{lin}}$, defined in (9.41), and an additional term which depends on the enrichment functions. We set

$$\boldsymbol{\mathcal{L}}_k = \frac{1}{d} \sum_{\substack{j=0 \\ j \neq k}}^{d} \boldsymbol{L}_j, \qquad k = 0, \ldots, d, \tag{9.62}$$

we prove that the approximation error (9.61) can be expressed as the error of the simplicial vector linear finite element plus a second term, which depends on the enrichment functions $\boldsymbol{e}_i$, $i = 0, \ldots, d$.

Theorem 9.4.1. *Let $\boldsymbol{f} \in \boldsymbol{C}(S_d)$. Then, for any $\boldsymbol{x} \in S_d$, we have*

$$\boldsymbol{E}^{\mathrm{enr}}[\boldsymbol{f}](\boldsymbol{x}) = \boldsymbol{E}^{\mathrm{lin}}[\boldsymbol{f}](\boldsymbol{x}) + \sum_{k=0}^{d} \left(\sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} - I_k(\boldsymbol{f}) \right) \left(\boldsymbol{E}(\boldsymbol{x}) - \sum_{i=0}^{d} \lambda_i(\boldsymbol{x}) \boldsymbol{E}(\boldsymbol{v}_i) \right) \boldsymbol{c}_k. \tag{9.63}$$

Proof. By (9.60), the approximation error (9.61) can be written as

$$\boldsymbol{E}^{\mathrm{enr}}[\boldsymbol{f}] = \boldsymbol{f} - \sum_{\ell=1}^{d}\sum_{j=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, \boldsymbol{\varphi}_{j\ell} - \sum_{j=0}^{d} I_j(\boldsymbol{f}) \boldsymbol{\psi}_j.$$

By (9.24), by applying Lemma 7.1.1 and by changing the order of the summation, we get

$$\begin{aligned} \sum_{\ell=1}^{d}\sum_{j=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, \boldsymbol{\varphi}_{j\ell} &= \sum_{\ell=1}^{d}\sum_{j=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \left(\lambda_j \boldsymbol{u}_\ell - \sum_{k=0}^{d} I_k(\lambda_j \boldsymbol{u}_\ell) \boldsymbol{\psi}_k \right) \\ &= \sum_{\ell=1}^{d}\sum_{j=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, \lambda_j \boldsymbol{u}_\ell - \sum_{\ell=1}^{d}\sum_{j=0}^{d}\sum_{k=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, I_k(\lambda_j \boldsymbol{u}_\ell) \boldsymbol{\psi}_k \\ &= \sum_{\ell=1}^{d} \left\langle \sum_{j=0}^{d} \lambda_j \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \right\rangle \boldsymbol{u}_\ell - \sum_{\ell=1}^{d}\sum_{j=0}^{d}\sum_{k=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, \frac{1}{d}(1 - \delta_{kj}) n_{k_\ell} \boldsymbol{\psi}_k \\ &= \sum_{\ell=1}^{d} \langle \boldsymbol{\Pi}^{\mathrm{lin}}[\boldsymbol{f}], \boldsymbol{u}_\ell \rangle \, \boldsymbol{u}_\ell - \sum_{\ell=1}^{d}\sum_{k=0}^{d}\sum_{\substack{j=0 \\ j \neq k}}^{d} \left\langle \frac{1}{d} \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \right\rangle n_{k_\ell} \boldsymbol{\psi}_k \\ &= \boldsymbol{\Pi}^{\mathrm{lin}}[\boldsymbol{f}] - \sum_{k=0}^{d}\sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} \boldsymbol{\psi}_k. \end{aligned}$$

Therefore, for each $\boldsymbol{x} \in S_d$, we have

$$\begin{aligned}
\boldsymbol{E}^{\mathrm{enr}}[\boldsymbol{f}](\boldsymbol{x}) &= \boldsymbol{f}(\boldsymbol{x}) - \sum_{\ell=1}^{d}\sum_{j=0}^{d} \langle \boldsymbol{L}_j(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, \boldsymbol{\varphi}_{j\ell}(\boldsymbol{x}) - \sum_{j=0}^{d} I_j(\boldsymbol{f}) \boldsymbol{\psi}_j(\boldsymbol{x}) \\
&= \boldsymbol{f}(\boldsymbol{x}) - \left(\boldsymbol{\Pi}^{\mathrm{lin}}[\boldsymbol{f}] - \sum_{k=0}^{d}\sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} \boldsymbol{\psi}_k \right) - \sum_{j=0}^{d} I_j(\boldsymbol{f}) \boldsymbol{\psi}_j(\boldsymbol{x}) \\
&= \boldsymbol{E}^{\mathrm{lin}}[\boldsymbol{f}](\boldsymbol{x}) + \sum_{k=0}^{d} \left(\sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} - I_k(\boldsymbol{f}) \right) \boldsymbol{\psi}_k(\boldsymbol{x}).
\end{aligned}$$

The thesis follows by (9.25). □

Definition 9.4.2. *A continuously differentiable function $f_\ell \in C^1(S_d)$ is said to have a Lipschitz continuous gradient on S_d, if there exists a constant $\rho > 0$ such that*

$$\|\nabla f_\ell(\boldsymbol{x}) - \nabla f_\ell(\boldsymbol{y})\|_2 \leq \rho \, \|\boldsymbol{x} - \boldsymbol{y}\|_2 , \qquad \boldsymbol{x}, \boldsymbol{y} \in S_d. \tag{9.64}$$

We call the smallest possible ρ such that (9.64) holds *Lipschitz constant* for ∇f_ℓ and we denote it by $L(\nabla f_\ell)$. We denote by $C^{1,1}(S_d)$ the class of all functions f_ℓ which are continuously differentiable with Lipschitz continuous gradient on S_d and by $\boldsymbol{C}^{1,1}(S_d)$ the direct product d times of the vector space $C^{1,1}(S_d)$ with itself. If $\boldsymbol{x} \in S_d$ and f_ℓ is a continuous convex function on S_d, from

$$\boldsymbol{x} = \sum_{j=0}^{d} \lambda_j(\boldsymbol{x}) \boldsymbol{v}_j$$

it follows that

$$f_\ell(\boldsymbol{x}) \leq \sum_{j=0}^{d} \lambda_j(\boldsymbol{x}) f_\ell(\boldsymbol{v}_j) = \Pi^{\mathrm{lin}}[\boldsymbol{f}]_\ell(\boldsymbol{x}), \tag{9.65}$$

and then, as a consequence of the more general Theorem [53, Theorem 2.3], the following bound holds.

Theorem 9.4.3. *For any $\boldsymbol{f} \in \boldsymbol{C}^{1,1}(S_d)$, we have*

$$\left| E^{\mathrm{lin}}[\boldsymbol{f}]_\ell(\boldsymbol{x}) \right| = \left| f_\ell(\boldsymbol{x}) - \Pi^{\mathrm{lin}}[f_\ell](\boldsymbol{x}) \right| \leq \frac{L(\nabla f_\ell)}{2} \left(\sum_{j=0}^{d} \lambda_j(\boldsymbol{x}) \, \|\boldsymbol{v}_j\|_2^2 - \|\boldsymbol{x}\|_2^2 \right), \qquad \boldsymbol{x} \in S_d, \qquad \ell = 1, \ldots, d.$$

Equality is attained for all functions of the form

$$f_\ell(\boldsymbol{x}) = a_\ell(\boldsymbol{x}) + c_\ell \, \|\boldsymbol{x}\|_2^2 , \qquad \ell = 1, \ldots, d,$$

where $c_\ell \in \mathbb{R}$ and $a_\ell(\boldsymbol{x})$ is any affine function.

9.4.2 The L^1 error estimate

In the following, we give a bound of the approximation error (9.63) given in Theorem 9.4.1 in L^1-norm

$$\|\boldsymbol{f}\|_1 = \int_{S_d} \sum_{\ell=1}^{d} |f_\ell(\boldsymbol{x})| \, d\boldsymbol{x}, \qquad \boldsymbol{f} = [f_1, \ldots, f_d] \in \boldsymbol{C}^{1,1}(S_d).$$

Therefore, by using the triangular inequality, we get

$$\|\boldsymbol{f}\|_1 \leq \sum_{\ell=1}^{d} \|f_\ell\|_1 , \tag{9.66}$$

and consequently, it is sufficient to determine a bound of each component of the error (9.63) in L^1-norm. By using Theorem 7.2.13, the following result is proved.

Lemma 9.4.4. *For any $\boldsymbol{f} \in \boldsymbol{C}^{1,1}(S_d)$, we have*

$$\left\| E^{\mathrm{lin}}[\boldsymbol{f}]_\ell \right\|_1 \leq \frac{L(\nabla f_\ell)}{8} \frac{d+1}{d+2} \left| S_d \right| h^2, \tag{9.67}$$

where

$$h = \sup_{\boldsymbol{v},\boldsymbol{w} \in S_d} \left\| \boldsymbol{v} - \boldsymbol{w} \right\|_2$$

is the diameter of the simplex S_d. By (9.66)*, we get*

$$\left\| \boldsymbol{E}^{\mathrm{lin}}[\boldsymbol{f}] \right\|_1 \leq \max_{\ell=1,\ldots,d} \frac{L(\nabla f_\ell)}{8} \frac{d(d+1)}{d+2} \left| S_d \right| h^2. \tag{9.68}$$

Theorem 9.4.5. *For any $\boldsymbol{f} \in \boldsymbol{C}^{1,1}(S_d)$, we get*

$$\left| \sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} - I_k(\boldsymbol{f}) \right| \leq \max_{\ell=1,\ldots,d} \frac{L(\nabla f_\ell)}{2(d+1)} \left(\sum_{\substack{0 \leq j < \ell \leq d \\ j,\ell \neq k}}^{d} \left\| \boldsymbol{v}_j - \boldsymbol{v}_\ell \right\|_2^2 \right), \qquad k = 0, \ldots, d. \tag{9.69}$$

Proof. Since the *vertex rule* for the face F_k, is exact for linear polynomials, see [76], by setting

$$\mathit{\Pi}_{F_k}^{\mathrm{lin}}[f_\ell](\boldsymbol{x}) = \sum_{\substack{i=0 \\ i \neq k}}^{d} \lambda_i(\boldsymbol{x}) f_\ell(\boldsymbol{v}_i), \qquad \boldsymbol{x} \in F_k, \qquad \ell = 1, \ldots, d, \tag{9.70}$$

we get

$$\mathcal{L}_k(\boldsymbol{f})_\ell = \frac{1}{d} \sum_{\substack{i=0 \\ i \neq k}}^{d} L_i(f_\ell) = \frac{1}{d} \sum_{\substack{i=0 \\ i \neq k}}^{d} f_\ell(\boldsymbol{v}_i) = \frac{1}{|F_k|} \int_{F_k} \mathit{\Pi}_{F_k}^{\mathrm{lin}}[f_\ell](\boldsymbol{x}) d\sigma(\boldsymbol{x}). \tag{9.71}$$

Therefore

$$\sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} - I_k(\boldsymbol{f}) = \sum_{\ell=1}^{d} \mathcal{L}_k(\boldsymbol{f})_\ell n_{k_\ell} - I_k(\boldsymbol{f}) = \sum_{\ell=1}^{d} \frac{1}{|F_k|} \int_{F_k} \left(\mathit{\Pi}_{F_k}^{\mathrm{lin}}[f_\ell](\boldsymbol{x}) - f_\ell(\boldsymbol{x}) \right) n_{k_\ell} d\sigma(\boldsymbol{x}).$$

Then

$$\begin{aligned} \left| \sum_{\ell=1}^{d} \langle \boldsymbol{\mathcal{L}}_k(\boldsymbol{f}), \boldsymbol{u}_\ell \rangle \, n_{k_\ell} - I_k(\boldsymbol{f}) \right| &\leq \sum_{\ell=1}^{d} \frac{1}{|F_k|} \int_{F_k} \left| \left(\mathit{\Pi}_{F_k}^{\mathrm{lin}}[f_\ell](\boldsymbol{x}) - f_\ell(\boldsymbol{x}) \right) n_{k_\ell} \right| d\sigma(\boldsymbol{x}) \\ &\leq \sum_{\ell=1}^{d} \frac{1}{|F_k|} \int_{F_k} \left| \mathit{\Pi}_{F_k}^{\mathrm{lin}}[f_\ell](\boldsymbol{x}) - f_\ell(\boldsymbol{x}) \right| d\sigma(\boldsymbol{x}), \end{aligned}$$

since $\boldsymbol{n}_k$ is the unit outward normal to the face F_k. By Theorem 7.2.6, the result follows. □

Theorem 9.4.6. *For any $\boldsymbol{f} \in \boldsymbol{C}^{1,1}(S_d)$, we get*

$$\left\| \boldsymbol{E}^{\mathrm{enr}}[\boldsymbol{f}] \right\|_1 \leq \max_{\ell=1,\ldots,d} L(\nabla f_\ell) \frac{d+1}{d! 2^{d+3}} \sqrt{\frac{(d+1)^{d+1}}{d^d}} \left(\frac{d}{d+2} + (d-1)\nu \right) h^{d+2},$$

where we set

$$\nu = \frac{1}{|S_d|} \max_{j=0,\ldots,d} \left\| \boldsymbol{\psi}_j \right\|_1.$$

Proof. By combining Lemma 9.4.4, Theorem 9.4.5 and Lemma 7.2.10, we get

$$\begin{aligned}\|\boldsymbol{E}^{\mathrm{enr}}[\boldsymbol{f}]\|_1 &\leq \max_{\ell=1,\dots,d} \frac{L(\nabla f_\ell)}{2} \left(\frac{1}{4}\frac{d(d+1)}{d+2}h^2 + \nu \frac{1}{d+1} \sum_{k=0}^{d} \sum_{\substack{0\leq j<\ell\leq d \\ j,\ell\neq k}}^{d} \|\boldsymbol{v}_j - \boldsymbol{v}_\ell\|_2^2 \right) |S_d| \\ &= \max_{\ell=1,\dots,d} \frac{L(\nabla f_\ell)}{8} \left(\frac{d(d+1)}{d+2} + (d^2-1)\nu \right) |S_d|\, h^2.\end{aligned}$$

Hence, the desired result follows by using the estimate

$$|S_d| \leq \frac{1}{2^d d!} \sqrt{\frac{(d+1)^{d+1}}{d^d}} h^d,$$

that we have established in Theorem 7.2.14. □

9.5 Numerical results

In this Section, we numerically demonstrate the effectiveness of the proposed enrichment strategy by using several examples. We compare the accuracy of the approximation, computed in L^1 norm, produced by the simplicial vector linear finite element $\boldsymbol{\mathcal{P}}_1(S_d)$ with that produced by $(S_d, \mathbb{P}_1^{\mathrm{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\mathrm{enr}})$ obtained by enriching $\boldsymbol{\mathcal{P}}_1(S_d)$ with the linear functionals (9.4) and the admissible enrichment functions of the sets

$$\mathcal{E}_1' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \cos(\lambda_i) \prod_{\substack{\ell=0 \\ \ell\neq i}}^{2} \lambda_\ell, \quad i = 0,1,2 \right\},$$

$$\mathcal{E}_2' = \left\{ \boldsymbol{e}_i = e_i \boldsymbol{n}_i,\ e_i = \tfrac{1+i}{1+\lambda_i} \prod_{\substack{\ell=0 \\ \ell\neq i}}^{2} \lambda_\ell, \quad i = 0,1,2 \right\},$$

of the Example 9.3.12 for $d = 2$. To this aim, we consider the following functions

$$\boldsymbol{f}_1(x,y) = \left[\sin(\pi(x+y)^2), x^2+y^2+25\right]^T,$$

$$\boldsymbol{f}_2(x,y) = \left[\frac{1}{x+y+3}, e^{x^5+y^5}\right]^T,$$

$$\boldsymbol{f}_3(x,y) = \left[\cos(x^3+y^3), (x^2+xy)\cos(x^3+y^3)\right]^T,$$

$$\boldsymbol{f}_4(x,y) = \left[\frac{1}{x^2+y^2+3}, x-y\right]^T.$$

For each experiment, we use an acute triangulation with $N = 1722$ triangles (see Figure 9.1) and a Delaunay triangulation with $N = 2650$ triangles (see Figure 9.2).

The numerical tests are realized by using `MatLab` software. The results are reported in Table 9.1 and Table 9.2 for the acute triangulation illustrated in Figure 9.1, and in Table 9.3 and Table 9.4 for the Delaunay triangulation illustrated in Figure 9.2. It is worth noting that, when dealing with a given function $\boldsymbol{f}$, the enhancement in the accuracy of the approximation achieved by the enriched finite element is not universally significant across all sets of enrichment functions. Instead, the precision of the approximation hinges on the particular set of enrichment functions selected.

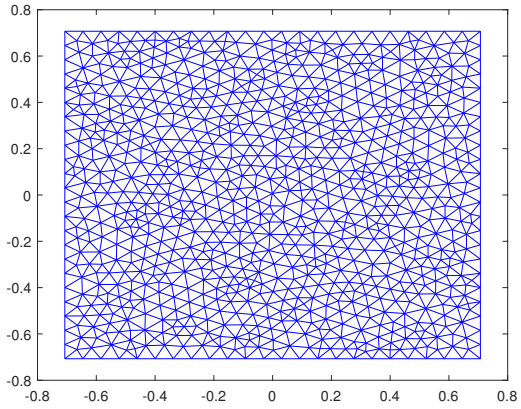


Figure 9.1: Acute triangulation with $N = 1722$ triangles.

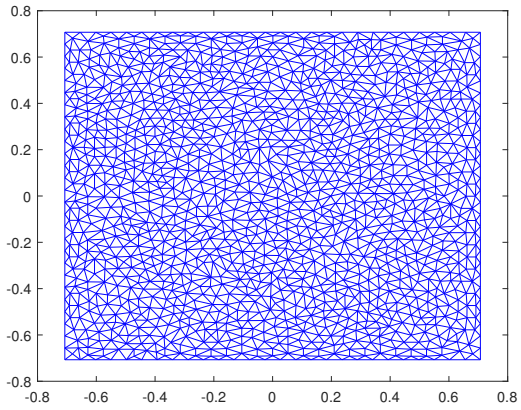


Figure 9.2: Delaunay triangulation with $N = 2650$ triangles.

	Simplicial vector linear finite element	Enriched vector linear finite element
$\boldsymbol{f}_1(x)$	1.0007e-02	6.9083e-03
$\boldsymbol{f}_2(x)$	1.2140e-03	8.1065e-04
$\boldsymbol{f}_3(x)$	1.2659e-03	5.6754e-04
$\boldsymbol{f}_4(x)$	1.1148e-04	7.6070e-05

Table 9.1: Comparison between the approximation errors computed in the L^1-norm, produced by approximating the test functions $\boldsymbol{f}_1$ - $\boldsymbol{f}_4$ through the simplicial vector linear finite element $\boldsymbol{\mathcal{P}}_1(S_d)$ and through $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{enr}})$ obtained by enriching $\boldsymbol{\mathcal{P}}_1(S_d)$ with the set of admissible enrichment functions $\mathcal{E}_1'$ using the triangulation shown in Figure 9.1.

	Simplicial vector linear finite element	Enriched vector linear finite element
$\boldsymbol{f}_1(x)$	1.0007e-02	7.2206e-03
$\boldsymbol{f}_2(x)$	1.2140e-03	8.5929e-04
$\boldsymbol{f}_3(x)$	1.2659e-03	6.3240e-04
$\boldsymbol{f}_4(x)$	1.1148e-04	8.1520e-05

Table 9.2: Comparison between the approximation errors computed in the L^1-norm, produced by approximating the test functions $\boldsymbol{f}_1$ - $\boldsymbol{f}_4$ through the simplicial vector linear finite element $\boldsymbol{\mathcal{P}}_1(S_d)$ and through $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{enr}})$ obtained by enriching $\boldsymbol{\mathcal{P}}_1(S_d)$ with the set of admissible enrichment functions $\mathcal{E}_2'$ using the triangulation shown in Figure 9.1.

	Simplicial vector linear finite element	Enriched vector linear finite element
$\boldsymbol{f}_1(x)$	6.7177e-03	4.5864e-03
$\boldsymbol{f}_2(x)$	7.3163e-04	4.9944e-04
$\boldsymbol{f}_3(x)$	8.2252e-04	3.5822e-04
$\boldsymbol{f}_4(x)$	7.8497e-05	5.5181e-05

Table 9.3: Comparison between the approximation errors computed in the L^1-norm, produced by approximating the test functions $\boldsymbol{f}_1$ - $\boldsymbol{f}_4$ through the simplicial vector linear finite element $\boldsymbol{\mathcal{P}}_1(S_d)$ and through $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{enr}})$ obtained by enriching $\boldsymbol{\mathcal{P}}_1(S_d)$ with the set of admissible enrichment functions $\mathcal{E}_1'$ using the triangulation shown in Figure 9.2.

	Simplicial vector linear finite element	Enriched vector linear finite element
$\boldsymbol{f}_1(x)$	6.7177e-03	4.8035e-03
$\boldsymbol{f}_2(x)$	7.3163e-04	5.2542e-04
$\boldsymbol{f}_3(x)$	8.2252e-04	4.004e-04
$\boldsymbol{f}_4(x)$	7.8497e-05	5.8658e-05

Table 9.4: Comparison between the approximation errors computed in the L^1-norm, produced by approximating the test functions $\boldsymbol{f}_1$ - $\boldsymbol{f}_4$ through the simplicial vector linear finite element $\boldsymbol{\mathcal{P}}_1(S_d)$ and through $(S_d, \mathbb{P}_1^{\text{enr}}(S_d), \boldsymbol{\Sigma}_{S_d}^{\text{enr}})$ obtained by enriching $\boldsymbol{\mathcal{P}}_1(S_d)$ with the set of admissible enrichment functions $\mathcal{E}_2'$ using the triangulation shown in Figure 9.2.

Conclusions and Future Works

The first part of this thesis concerned the study of the constrained mock-Chebyshev least squares approximation. In Chapter 2 we have generalized the univariate constrained mock-Chebyshev approximation to the bivariate case through the constrained mock-Chebyshev least squares tensor product approximation and the constrained mock-Padua least squares approximation. We have realized these approximations by using different basis functions and noticed that better accuracies are reached by the tensor product Chebyshev basis. Working with data sampled on a fixed regular grid, this basis has a KKT matrix with well behaved condition number, by increasing the degree of simultaneous regression, and a greater number of basis elements if compared with the total degree Chebyshev basis. In Chapter 3 we have used the constrained mock-Chebyshev least squares approximation to obtain stable quadrature formulas with high degree of exactness and accuracy from equispaced nodes. Since the accuracy of these quadrature formulas varies with the degree of the constrained mock-Chebyshev least squares approximation, depending on the degree of smoothness of the function f, we have developed an adaptive algorithm for determining the optimal degree which corresponds to the more accurate quadrature formula. In Chapter 4 we have analyzed new theoretical aspects of the constrained mock-Chebyshev least squares operator. We have introduced pointwise explicit representations of the error and its derivatives. By using the constrained mock-Chebyshev least squares operator, we have presented a method for approximating the successive derivatives of f at any point $x \in [-1, 1]$ and we have provided estimates for these approximations. This formula provides a global polynomial approximation of the successive derivatives of the function f. In the second part of the thesis, we focused on the development of a unified and general framework for the enrichment of standard triangular linear finite elements in $\mathbb{R}^2$ and of the standard simplicial linear finite elements in $\mathbb{R}^d$. In Chapter 5, we have introduced and studied a new nonconforming finite element based on an enrichment of the standard triangular linear finite element, which uses line integrals and quadratic polynomials. Starting from this new element, we have proposed an improvement of the triangular Shepard operator, for the reconstruction of a function from two-dimensional scattered data, when line integrals are given, in addition to functional values. In Chapter 6 we have extended to a more general setting the results presented in Chapter 5. More precisely, we have introduced a new class of nonconforming finite elements by enriching the class of linear polynomial functions with additional functions which are not necessarily polynomials. We have provided a simple condition on the enrichment functions, which is both necessary and sufficient, that guarantees the existence of a family of such enriched elements. Several sets of admissible enrichment functions that satisfy the admissibility condition have been also provided, together with the explicit expression of the related approximation error. Our main result has shown that the approximation error can be decomposed into two parts: the first one is related to the linear triangular element while the second one depends on the enrichment functions. This representation of the approximation error has allowed us to derive sharp error bounds in both L^∞-norm and L^1-norm, with explicit constants, for continuously differentiable functions with Lipschitz continuous gradients. These bounds have been proportional to the second and the fourth power of the circumcircle radius of the triangle, respectively. We have also provided explicit expressions of these bounds in terms of the circumcircle diameter and the sum of squares of the triangle edge lengths. In Chapter 7 we have generalized the results presented in Chapter 6 to the case of the standard simplicial linear finite element in $\mathbb{R}^d$. More precisely, we have introduced a new class of finite elements by enriching the standard simplicial linear finite element in $\mathbb{R}^d$ with additional functions (not necessarily polynomials) satisfying the vanishing condition at all vertices. Chapter 8 was supplementary to Chapter 7, and its main goal has been to provide a general strategy for enriching the standard simplicial linear finite element without imposing restrictive conditions

on the enrichment functions, like their vanishing at the vertices. In particular, we have extended the results presented in Chapter 7 to a more general case, by using generic enrichment functions and generic linear functionals. In line with previous researches, in Chapter 9, we have present a general strategy for enriching the simplicial vector linear finite element by nonpolynomial enrichment functions. This enriched finite element can be regarded as an extension of Bernardi and Raugel finite element.

Future works will develop in two directions. The first one will concern the constrained mock-Chebyshev least squares approximation. This approximation operator can be extended by considering other possible Jacobi zeros, as for instance Legendre zeros with the additional points ± 1, by taking into account that in this case still optimal Lebesgue constants are achieved. Other exciting research problems are related to the possibility of the generalization of the constrained mock-Chebyshev least squares approximation to the case of functions sampled at the nodes of the n-th subdivision of a triangle. The second one will focus on the study of the enrichment of finite elements. In particular, we can apply the enrichment strategies introduced in this thesis to approximate the solution of partial differential equations.

Acknowledgments

To my parents who supported me during these years.

Special thanks to my supervisors Professor Francesco Dell'Accio and Professor Allal Guessab, for their immense patience, for their valuable advice and for the knowledge transmitted during the thesis writing process.

I would like to thank my supervisor Prof. Francesco Dell'Accio who is the first who believed and still believes in me.

I would like to thank Prof. Allal Guessab for allowing me to visit the Université de Pau et des Pays de l'Adour for twelve months. It was a great experience that I will never forget.

I would like to thank Prof. Domingo Barrera for inviting me to the Universidad de Granada for three months.

I would like to thank Prof. Francisco Marcellán for inviting me to the Universidad Carlos III de Madrid for one month.

I would like to thank Prof. Filomena Di Tommaso for her suggestions.

I would like to thank my friends Antonio Bitonti, Francesca Coscarelli, Mario Carbone, Davide De Luca, Simone Lucanto, Mario Iazzolino, Salvatore Staine, Francesco De Luca and Giulia Rizzuti.

I would like to thank my French family:

- Ambasciata members: Momo, Stefano, Barbara, Sally, Giulio, Martina, Valeria, Alice, Miriam and Federica
- Pelofanta, Martina, Imanol, Jesus, Andrés, Rodrigo, Markel and Aritz.

for supporting me like a real family and making my stay in France amazing.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1965.

[2] B. Achchab, K. Bouihat, A. Guessab, and G. Schmeisser. A general approach to the construction of nonconforming finite elements on convex polytopes. *Applied Mathematics and Computation*, 268:916–923, 2015.

[3] M. Bachar and A. Guessab. A Simple Necessary and Sufficient Condition for the Enrichment of the Crouzeix-Raviart Element. *Applicable Analysis and Discrete Mathematics*, 10:378–393, 2016.

[4] M. Bachar and A. Guessab. Characterization of the Existence of an Enriched Linear Finite Element Approximation Using Biorthogonal Systems. *Results in Mathematics*, 70:401–413, 2016.

[5] D. Barrera, A. Guessab, M. J. Ibáñez, and O. Nouisser. Construction techniques for multivariate modified quasi-interpolants with high approximation order. *Computers & Mathematics with Applications*, 65:29–41, 2013.

[6] J. Beckmann, H. N. Mhaskar, and J. Prestin. Local numerical integration on the sphere. *GEM-International Journal on Geomathematics*, 5:143–162, 2014.

[7] C. Bernardi and G. Raugel. Analysis of some finite elements for the Stokes problem. *Mathematics of Computation*, 44:71–79, 1985.

[8] L. Bos, M. Caliari, S. De Marchi, M. Vianello, and Y. Xu. Bivariate Lagrange interpolation at the Padua points: the generating curve approach. *Journal of Approximation Theory*, 143:15–25, 2006.

[9] L. Bos, S. De Marchi, M. Vianello, and Y. Xu. Bivariate Lagrange interpolation at the Padua points: the ideal theory approach. *Numerische Mathematik*, 108:43–57, 2007.

[10] J. P. Boyd and F. Xu. Divergence (Runge phenomenon) for least-squares polynomial approximation on an equispaced grid and mock–Chebyshev subset interpolation. *Applied Mathematics and Computation*, 210:158–168, 2009.

[11] S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, 2018.

[12] J. Brandts, S. Korotov, M. Křížek, and J. Šolc. On nonobtuse simplicial partitions. *SIAM review*, 51:317–335, 2009.

[13] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Springer-Verlag, 2008.

[14] M. Caliari, S. De Marchi, and M. Vianello. Algorithm 886: Padua2D-Lagrange interpolation at Padua points on bivariate domains. *ACM Transactions on Mathematical Software*, 35:1–11, 2008.

[15] B. C. Carlson. *Special functions of applied mathematics*. Academic Press, 1977.

[16] R. Cavoretto, A. De Rossi, F. Dell'Accio, and F. Di Tommaso. Fast computation of triangular Shepard interpolants. *Journal of Computational and Applied Mathematics*, 354:457–470, 2019.

[17] E. W. Cheney. *Introduction to Approximation Theory*. American Mathematical Society, 1998.

[18] E. W. Cheney and W. A. Light. *A course in Approximation Theory*. American Mathematical Society, 2009.

[19] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Society for Industrial and Applied Mathematics, 2002.

[20] H. S. M. Coxeter. The circumradius of the general simplex. *The Mathematical Gazette*, 15:229–231, 1930.

[21] H. S. M. Coxeter. *Regular polytopes*. Courier Corporation, 1973.

[22] T. Cătinaş. The bivariate Shepard operator of Bernoulli type. *Calcolo*, 44:189–202, 2007.

[23] P. J. Davis. *Interpolation and Approximation*. Dover Publications, 1975.

[24] S. De Marchi, F. Dell'Accio, and M. Mazza. On the constrained mock-Chebyshev least-squares. *Journal of Computational and Applied Mathematics*, 280:94–109, 2015.

[25] S. De Marchi, G. Elefante, E. Perracchione, and D. Poggiali. Quadrature at fake nodes. *Dolomites Research Notes on Approximation*, 14:39–45, 2021.

[26] S. De Marchi, F. Marchetti, E. Perracchione, and D. Poggiali. Polynomial interpolation via mapped bases without resampling. *Journal of Computational and Applied Mathematics*, 364:112347, 2020.

[27] F. Dell'Accio and F. Di Tommaso. Scattered data interpolation by Shepard's like methods: classical results and recent advances. *Dolomites Research Notes on Approximation*, 9:32–44, 2016.

[28] F. Dell'Accio and F. Di Tommaso. Rate of convergence of multinode Shepard operators. *Dolomites Research Notes on Approximation*, 12:1–6, 2019.

[29] F. Dell'Accio, F. Di Tommaso, E. Francomano, and F. Nudo. An adaptive algorithm for determining the optimal degree of regression in constrained mock-Chebyshev least squares quadrature. *Dolomites Research Notes on Approximation*, 15:35–44, 2022.

[30] F. Dell'Accio, F. Di Tommaso, A. Guessab, and F. Nudo. A unified enrichment approach of the standard three-node triangular element. *Applied Numerical Mathematics*, 187:1–23, 2023.

[31] F. Dell'Accio, F. Di Tommaso, O. Nouisser, and B. Zerroudi. Increasing the approximation order of the triangular Shepard method. *Applied Numerical Mathematics*, 126:78–91, 2018.

[32] F. Dell'Accio and F. Di Tommaso. Complete Hermite–Birkhoff interpolation on scattered data by combined Shepard operators. *Journal of Computational and Applied Mathematics*, 300:192–206, 2016.

[33] F. Dell'Accio and F. Di Tommaso. Bivariate Shepard–Bernoulli operators. *Mathematics and Computers in Simulation*, 141:65–82, 2017.

[34] F. Dell'Accio, F. Di Tommaso, A. Guessab, and F. Nudo. Enrichment strategies for the simplicial linear finite elements. *Applied Mathematics and Computation*, 451:128023, 2023.

[35] F. Dell'Accio, F. Di Tommaso, A. Guessab, and F. Nudo. A general class of enriched methods for the simplicial linear finite elements. *Applied Mathematics and Computation*, 456:128149, 2023.

[36] F. Dell'Accio, F. Di Tommaso, and K. Hormann. On the approximation order of triangular Shepard interpolation. *IMA Journal of Numerical Analysis*, 36:359–379, 2015.

[37] F. Dell'Accio, F. Di Tommaso, O. Nouisser, and B. Zerroudi. Fast and accurate scattered Hermite interpolation by triangular Shepard operators. *Journal of Computational and Applied Mathematics*, 382:113092, 2021.

[38] F. Dell'Accio, F. Di Tommaso, and F. Nudo. Constrained mock-Chebyshev least squares quadrature. *Applied Mathematics Letters*, 134:108328, 2022.

[39] F. Dell'Accio, F. Di Tommaso, and F. Nudo. Generalizations of the constrained mock-Chebyshev least squares in two variables: Tensor product vs total degree polynomial interpolation. *Applied Mathematics Letters*, 125:107732, 2022.

[40] F. Dell'Accio, D. Mezzanotte, F. Nudo, and D. Occorsio. Product integration rules by the constrained mock-Chebyshev least squares operator. *BIT Numerical Mathematics*, 63:24, 2023.

[41] N. P. Dolbilin, H. Edelsbrunner, and O. R. Musin. On the optimality of functionals over triangulations of Delaunay sets. *Russian Mathematical Surveys*, 67:781–788, 2012.

[42] C. F. Dunkl and Y. Xu. *Orthogonal Polynomials of Several Variables.* Cambridge University Press, 2014.

[43] A. Ezzirani and A. Guessab. A fast algorithm for Gaussian type quadrature formulae with mixed boundary conditions and some lumped mass spectral approximations. *Mathematics of Computation*, 68:217–248, 1999.

[44] R. Farwig. Rate of convergence of Shepard's global interpolation formula. *Mathematics of Computation*, 46:577–590, 1986.

[45] G. J. Fix, S. Gulati, and G. I. Wakoff. On the use of singular functions with finite element approximations. *Journal of Computational Physics*, 13:209–228, 1973.

[46] G. B. Folland. How to Integrate A Polynomial Over A Sphere. *The American Mathematical Monthly*, 108:446–448, 2001.

[47] R. Franke and G. Nielson. Smooth interpolation of large sets of scattered data. *International Journal for Numerical Methods in Engineering*, 15:1691–1704, 1980.

[48] W. Gautschi. *Numerical Analysis.* Birkhäuser, 2011.

[49] J. Glaubitz. Stable high-order cubature formulas for experimental data. *Journal of Computational Physics*, 447:110693, 2021.

[50] W. J. Gordon and J. A. Wixom. Shepard's method of "metric interpolation" to bivariate and multivariate interpolation. *Mathematics of Computation*, 32:253–264, 1978.

[51] J. Guermond and A. Ern. *Finite Elements II: Galerkin Approximation, Elliptic and Mixed PDEs.* Springer, 2021.

[52] A. Guessab. Direct and converse results for generalized multivariate Jensen-type inequalities. *Journal of Nonlinear and Convex Analysis*, 13:777–797, 2012.

[53] A. Guessab. Approximations of differentiable convex functions on arbitrary convex polytopes. *Applied Mathematics and Computation*, 240:326–338, 2014.

[54] A. Guessab. Generalized barycentric coordinates and Jensen type inequalities on convex polytopes. *Journal of Nonlinear and Convex Analysis*, 17:527–547, 2016.

[55] A. Guessab. *Sharp Approximations based on Delaunay Triangulations and Voronoi Diagrams.* NSU Publishing and Printing Center., 2022.

[56] A. Guessab, O. Nouisser, and G. Schmeisser. A definiteness theory for cubature formulae of order two. *Constructive Approximation*, 24:263–288, 2006.

[57] A. Guessab and Q. I. Rahman. Quadrature Formulae and Polynomial Inequalities. *Journal of Approximation Theory*, 90:255–282, 1997.

[58] A. Guessab and G. Schmeisser. Sharp integral inequalities of the Hermite–Hadamard type. *Journal of Approximation Theory*, 115:260–288, 2002.

[59] A. Guessab and G. Schmeisser. Convexity results and sharp error estimates in approximate multivariate integration. *Mathematics of Computation*, 73:1365–1384, 2004.

[60] A. Guessab and G. Schmeisser. Construction of positive definite cubature formulae and approximation of functions via Voronoi tessellations. *Advances in Computational Mathematics*, 32:25–41, 2010.

[61] A. Guessab and B. Semisalov. A Multivariate Version of Hammer's Inequality and Its Consequences in Numerical Integration. *Results in Mathematics*, 73:1–37, 2018.

[62] A. Guessab and B. Semisalov. Numerical integration using integrals over hyperplane sections of simplices in a triangulation of a polytope. *BIT Numerical Mathematics*, 58:613–660, 2018.

[63] A. Guessab and B. Semisalov. Extended Multidimensional Integration Formulas on Polytope meshes. *SIAM Journal on Scientific Computing*, 41:A3152–A3181, 2019.

[64] A. Guessab and Y. Zaim. A unified and general framework for enriching finite element approximations. *Progress in Approximation Theory and Applicable Complex Analysis: In Memory of Q. I. Rahman*, pages 491–519, 2017.

[65] N. Hale and A. Townsend. Fast and accurate computation of Gauss–Legendre and Gauss–Jacobi quadrature nodes and weights. *SIAM Journal on Scientific Computing*, 35:A652–A674, 2013.

[66] N. Hale and A. Townsend. A fast, simple, and stable Chebyshev–Legendre transform using an asymptotic formula. *SIAM Journal on Scientific Computing*, 36:A148–A167, 2014.

[67] G. W. Howell. Derivative error bounds for Lagrange interpolation: an extension of Cauchy's bound for the error of Lagrange interpolation. *Journal of Approximation Theory*, 67:164–173, 1991.

[68] D. Huybrechs. Stable high-order quadrature rules with equidistant points. *Journal of Computational and Applied Mathematics*, 231:933–947, 2009.

[69] B. A. Ibrahimoglu. A fast algorithm for computing the mock-Chebyshev nodes. *Journal of Computational and Applied Mathematics*, 373:112336, 2020.

[70] B. A. Ibrahimoglu. A new approach for constructing mock-Chebyshev grids. *Mathematical Methods in the Applied Sciences*, 44:14766–14775, 2021.

[71] G. Klein and J. P. Berrut. Linear barycentric rational quadrature. *BIT Numerical Mathematics*, 52:407–424, 2012.

[72] G. Klein and J. P. Berrut. Linear rational finite differences from derivatives of barycentric rational interpolants. *SIAM Journal on Numerical Analysis*, 50:643–656, 2012.

[73] G. Korchmaros. Una limitazione per il volume di un simplesso n-dimensionale avente spigoli di date lunghezze. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti*, 56:876–879, 1974.

[74] J. Li. General explicit difference formulas for numerical differentiation. *Journal of Computational and Applied Mathematics*, 183:29–52, 2005.

[75] F. Little. Convex combination surfaces. In R. E. Barnhill and W. Boehm, editors, *Surfaces in Computer Aided Geometric Design*, pages 99–107. North-Holland, Amsterdam, 1983.

[76] J. N. Lyness and A. C. Genz. On Simplex Trapezoidal Rule Families. *SIAM Journal on Numerical Analysis*, 17:126–147, 1980.

[77] H. Majidian. Creating stable quadrature rules with preassigned points by interpolation. *Calcolo*, 53:217–226, 2016.

[78] R. B. Manfrino, J. A. G. Ortega, and R. V. Delgado. *Inequalities: A Mathematical Olympiad Approach*. Birkhäuser, 2009.

[79] W. Markoff and J. Grossmann. Über Polynome, die in einem gegebenen Intervalle möglichst wenig von Null abweichen. *Mathematische Annalen*, 77:213–258, 1916.

[80] F. Piazzon and M. Vianello. Small perturbations of polynomial meshes. *Applicable Analysis*, 92:1063–1073, 2013.

[81] R. B. Platte, L. N. Trefethen, and A. B. J. Kuijlaars. Impossibility of Fast Stable Approximation of Analytic Functions from Equispaced Samples. *SIAM Review*, 53:308–318, 2011.

[82] A. S. Posamentier and I. Lehmann. *The Secrets of Triangles: A Mathematical Journey*. Prometheus Books, 2012.

[83] C. H. Raifaizen. A simpler proof of Heron's formula. *Mathematics Magazine*, 44:27–28, 1971.

[84] A. Ralston and P. Rabinowitz. *A First Course in Numerical Analysis*. Dover Publications, 2001.

[85] L. Reichel. On polynomial approximation in the uniform norm by the discrete least squares method. *BIT Numerical Mathematics*, 26:349–368, 1986.

[86] R. J. Renka and R. Brown. Algorithm 792: accuracy test of ACM algorithms for interpolation of scattered data in the plane. *ACM Transactions on Mathematical Software (TOMS)*, 25:78–94, 1999.

[87] T. J. Rivlin. *Chebyshev Polynomials*. Wiley, 1974.

[88] T. J. Rivlin. Optimally Stable Lagrangian Numerical Differentiation. *SIAM Journal on Numerical Analysis*, 12:712–725, 1975.

[89] C. Runge. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Mathematik und Physik*, 46:224–243, 1901.

[90] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524, 1968.

[91] J. R. Shewchuk. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In *Applied Computational Geometry: Towards Geometric Engineering*, volume 1148 of *Lecture Notes in Computer Science*, pages 203–222. Springer-Verlag, 1996.

[92] D. Shi and L. Pei. Nonconforming quadrilateral finite element method for a class of nonlinear sine–Gordon equations. *Applied Mathematics and Computation*, 219:9447–9460, 2013.

[93] Z. Shi. Nonconforming finite element methods. *Journal of Computational and Applied Mathematics*, 149:221–225, 2002.

[94] A. H. Stroud. *Approximate Calculation of Multiple Integrals. Prentice-Hall Series in Automatic Computation*. Prentice-Hall, 1971.

[95] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.

[96] M. Vianello. Dubiner distance and stability of Lebesgue constants. *Journal of Inequalities and Special Functions*, 10:49–60, 2019.

[97] J. Warren. Barycentric coordinates for convex polytopes. *Advances in Computational Mathematics*, 6:97–108, 1996.

[98] N. Watanabe, W. Wang, J. Taron, U. Görke, and O. Kolditz. Lower-dimensional interface elements with local enrichment: application to coupled hydro-mechanical problems in discretely fractured porous media. *International Journal for Numerical Methods in Engineering*, 90:1010–1034, 2012.

[99] M. W. Wilson. Discrete least squares and quadrature formulas. *Mathematics of Computation*, 24:271–282, 1970.

[100] M. W. Wilson. Necessary and Sufficient Conditions for Equidistant Quadrature Formula. *SIAM Journal on Numerical Analysis*, 7:134–141, 1970.

[101] C. Zuppa. Error estimates for modified local Shepard's interpolation formula. *Applied Numerical Mathematics*, 49:245–259, 2004.