GBC-based serendipity functions of any order over arbitrary polygons

Master's Thesis submitted to the
Dipartimento di Scienza e Alta Tecnologia of the *Università degli Studi dell'Insubria*
in partial fulfillment of the requirements for the degree of
Laurea magistrale in matematica

and to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Computational Science

presented by
Marta Celio

under the supervision of
Prof. Matteo Semplice and Kai Hormann

March 2025

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Marta Celio
Como, 20 March 2025

Life is lived through being kind
That's the way our tale's designed

Alex Beckam and Adriana Figueroa
From "Undertale: The Musical"

Abstract

This thesis extends a construction for quadratic serendipity coordinates over polygons, by obtaining serendipity coordinates for the reproducibility of polynomials of arbitrary order. The serendipity coordinates are obtained as linear combinations of products of generalized barycentric coordinates. Interpolation nodes and functions are then defined from said serendipity coordinates, being in equal number to them: these represent the goal of this thesis, as the intent of this research is to use these results for the solution of FEM problems. For polynomials of order q, the construction is well-defined over strongly convex and concave n-gons such that $n > q$, yielding qn functions with rigorously defined interpolation nodes. The construction can be extended to the cases $q = 3, n = 3$, $q = 4, n = 3$ and $q = 4, n = 4$; yielding 10, 15 and 17 functions, respectively, with some interpolation nodes not rigorously defined and instead randomly picked on the interior of the polygon. Therefore, the method presented in this thesis can be defined for any strongly convex and concave polygon for $q \leq 4$, with interpolation nodes being rigorously defined in most cases, making it widely applicable compared to the literature. It is also more resistant to polygons representing edge cases.

Acknowledgements

There have been many people I have met, talked to, befriended, who have been an important part of my daily life while I knew them. Some I have known for longer, some for less long. If I had to make an exhaustive list, this section would turn into an essay itself. Since I have the opportunity, though, I would like to thank the most important people who are still by my side today.

Firstly, I would like to thank my thesis advisors Kai Hormann and Matteo Semplice, without whom this thesis quite literally would not have been possible. I feel proud of my work and I am grateful they gave me the means and opportunity to carry it out. I would also like to thank the system and offices of the Università degli Studi dell'Insubria and the Università della Svizzera Italiana and everyone involved, especially the thesis committee, the Decanato, Alberto Giulio Setti and Marco Donatelli, for providing the system which allowed for the birth of this thesis and helping me keep up as needed throughout the process.

Because I did not get the chance two years ago, I would also like to thank Marco Benini, the advisor for my bachelor's thesis. The topic of his course and my thesis was very fun work for me and he, too, has been extremely helpful throughout. I would like to thank the thesis committee for my bachelor's thesis, as well.

I would like to thank my family. My parents Daniela and Luigi "Gigi" are the people I love most in the world. My brother Mattia and I are close in our own way: I hope that he can continue thriving and that we can keep the special kind of trust we have for each other. My dog Chicco is the goodest boy a rascal like him could be. My grandparents Guglielmina "Mina", Orsolina "Lilli" and Francesco "Ciccillo" have always doted on me. I would be remiss not to mention that my aunt Teresita gifted me some of my clothes for my thesis defense.

I would like to thank the group of friends I made by studying at USI, especially Eleonora, Claudio, Ian and Su. They and everyone else have always been kind friends and lovely people. They have accepted to come with me to events I wanted to go to when I needed some company; and they have included me in hangouts they had organized themselves. I got to talk to them freely about deep topics and videogames alike; and I have been shown that they have always listened and cared. I get along with them really well and I hope I can continue being friends with them for much longer.

I would like to thank the friends I have made online, especially Nenn and Plush. They have given me a place to find people with the same passions and ideas as me even back when I knew so few of them in person. Talking to them is always fun and they, too, have been really kind friends and valuable individuals. Some of them, like LP, have even kept checking in on me during a time where I had distanced myself from some online spaces.

I would like to thank the friends I have made by living in my town. Most importantly Fabrizia "Fabry", with whom I've been friends ever since elementary school: we have shared

many passions together and we have been best friends for a very long time. My neighbors (especially Saskia) have also been an important part of my circle of friends growing up and I still talk to them on occasion. The same goes for other people my age living in my town, such as the ones I have met through summer and winter camps, theater and other community activities (especially Chiara, Federica, Silvia and Dario). I must also mention the people older than me who have organized those activities in the first place and who, to this day, make an effort to include me as a friend and as part of the community (especially Silvia, Cinzia and Luca).

There are so many others I could mention, such as the other two Insubria double degree students Pietro and Gabriele, the rest of my classmates at Insubria, my high school classmates (with whom I am hoping to reconnect!), my middle school classmates, other people from other online communities, my host families during my study trips abroad and so many more people I have met in brief occasions and ended up not keeping much in touch with. However, if I went on for so long, I would probably have to list almost all the people I have ever met. Suffice it to say, many people have left an impact on me, as is the case for everyone; and out of all those people, the ones I have chosen to thank in this section are the ones I hold closest to me.

Contents

Chapter 1

Introduction

This thesis aims to extend the work by Hackemack and Ragusa [2018] for the construction of GBC-based quadratic serendipity coordinates, by instead building serendipity coordinates of any order. The work on this thesis is carried out with the intent of rendering these results usable for the solution of an FEM problem.

The finite element method (FEM) involves dividing a domain into many small "elements" (polygons in our case) and finding an approximation of the solution on each element. More specifically, we would like to find a set of points to use as interpolation nodes over each element and associate them with a set of basis functions. These basis functions should ideally be interpolation functions over the nodes (each being equal to 1 over one node and to 0 over the others) and be able to reproduce polynomials up to a given degree.

The method studied here builds interpolation functions as linear combinations of products of generalized barycentric coordinates (GBCs). GBCs are a generalization of barycentric coordinates over larger polygons: similarly to barycentric coordinates over triangles, they are defined so that they can reproduce linear polynomials over their polygon of definition. Because of this, over a given polygon of definition, the functions defined as all possible products of q GBCs can reproduce polynomials of order q. However, the number of such functions rapidly increases as q does, so a smaller set of "serendipity coordinates" is built from them: these are linear combinations of products of GBCs, defined so as to keep the property of polynomial reproducibility for order q. Finally, each serendipity coordinate is associated to one interpolation node and one interpolation function, with the interpolation functions being themselves linear combinations of serendipity coordinates.

Chapter 2 of this thesis is a theoretical introduction to GBCs. It showcases their basic definition and as a few different types of GBCs, as unlike barycentric coordinates on a triangle, GBCs are not uniquely defined. The three types of GBCs showcased are Wachspress coordinates, discrete harmonic coordinates and mean value coordinates, which all belong to the family of three-point coordinates: their different properties, such as their polygons of definition, are also stated. The chapter puts a bigger emphasis on mean value coordinates especially: because they can be defined over non-convex polygons, they are used as the GBCs of reference throughout the rest of the thesis.

Chapter 3 retreads the work by Rand et al. [2014] and Hackemack and Ragusa [2018] in order to construct quadratic serendipity coordinates, specifically. First, quadratic coordinates (i.e. products between pairs of GBCs, specifically mean value coordinates) are defined and it is

proven they reproduce quadratic polynomials. Then, those coordinates are reduced to serendipity coordinates. More specifically, because serendipity coordinates are linear combinations of quadratic coordinates, the relation between the two is written as a linear algebra problem, with a matrix containing all the coefficients of the linear combinations. Then, some constraints are imposed on the coefficients which allow for the resulting serendipity coordinates to keep the polynomial reproducibility property. The study of these constraints leads to finding a linear algebra method for constructing the matrix of coefficients itself, as a matrix product.

From chapter 4 onwards, this thesis starts presenting our own original work. Chapter 4 specifically extends the work in chapter 3 to orders higher than 2. First, it obtains an analogous result regarding polynomial reproducibility. Then, it actually rewrites said result, so as to obtain it in a form that is easier to implement in code. Finally, it employs a similar algebraic approach as the previous chapter in order to find the linear combination coefficients associated with serendipity coordinates: similar constraints are imposed on the coefficients; and the matrix of coefficients is found through a more general application of the matrix product in chapter 3.

Chapter 5 concerns the derivation of the desired interpolation functions from these newly found serendipity coordinates. First, it is proven that serendipity coordinates are proportional to Bernstein basis polynomials on the edges of the polygon of definition. Then, Bernstein basis polynomials are studied individually, fixing appropriate interpolation nodes and obtaining relative Lagrange basis polynomials. Finally, the results obtained on Bernstein basis polynomials are applied to the actual interpolation functions, by appropriately picking interpolation nodes on the polygon boundary and defining interpolation functions by relying entirely on this boundary behavior. A significant portion of this chapter is an application of the theory laid out by Floater and Lai [2016] to our own serendipity coordinates.

Chapter 6 presents a few complications of our method, mainly related to the construction of the matrix of coefficients. One complication concerns the ability to span the space of polynomials: in some cases, by applying our approach directly, it would yield fewer functions than the dimension of said space, making our method fail. Another complication concerns the polygon of definition more generally: having collinear vertices, or too few vertices in general, also causes our method to fail. The chapter presents two approaches to this issue; one in which those cases are simply not considered and one which attempts to extend our method to as many of those cases as possible.

Chapter 7 compares our method to other similar methods in literature, namely the ones by Floater and Lai [2016] and Cao et al. [2022]. It shows that our method is more widely applicable in terms of polynomial order and polygon of definition. It also shows that it is more resistant to edge cases such as polygons with short edges, small angles or almost flat angles.

Finally, chapter 8 serves as a conclusion to the thesis by outlining possible future work to be carried out regarding it. The main points of future interest are a theoretical proof of the numerical complications outlined in chapter 6 and an actual application of the results of this thesis to an FEM problem.

All the code used in this thesis is available at `https://github.com/mcelio001/thesis-2025`.

Chapter 2

Generalized barycentric coordinates

2.1 Introduction to generalized barycentric coordinates

The most common method of representing a point $\boldsymbol{v}$ in a d-dimensional Euclidean space, such as the plane ($\mathbb{R}^2$) or more generally $\mathbb{R}^d$, is by assigning it a set of Cartesian coordinates. The space is given an origin and a set of axes (d lines, usually perpendicular, going through the origin); then, $\boldsymbol{v}$ is identified through a tuple of length d, in which each number represents the point's distance from the origin "along each axis". More rigorously, for each axis, one considers a hyperplane parallel to the axis going through $\boldsymbol{v}$, then notes the point of intersection between the hyperplane and the axis: the distance between this intersection and the origin is the Cartesian coordinate of $\boldsymbol{v}$ along the axis of interest.

There are other systems of coordinates that can be used to identify points, though. One such case is the family of homogeneous coordinate systems, first proposed by Möbius [1827]. In a system of that kind, points are represented by a tuple of length $d+1$. In general, a homogeneous coordinate system is characterized by the fact that coordinates proportional to each other represent the same point: in other words, if the coordinates of a point are all multiplied by the same non-zero scalar, the resulting coordinates will still represent the same point. Homogeneous coordinate systems are used a lot in fields such as computer graphics and computer vision, for their ability to easily codify things such as translations and points at infinity.

One type of homogeneous coordinates is barycentric coordinates. They are defined in reference to a simplex in the space (a triangle in a 2D plane, for example). The idea behind them is that, given a point $\boldsymbol{v}$, its barycentric coordinates can be interpreted as quantities of mass located on the veritces of the simplex, chosen such that $\boldsymbol{v}$ ends up as the center of mass (the barycenter) of the simplex. In practice, if we define $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{d+1}$ as the Cartesian coordinates of the vertices of a simplex $\triangle$, writing the barycentric coordinates of $\boldsymbol{v}$ as $(\lambda_1(\boldsymbol{v}), \cdots, \lambda_{d+1}(\boldsymbol{v}))$, the equation

$$\sum_{i=1}^{d+1} \lambda_i(\boldsymbol{v}) \boldsymbol{v}_i = \left(\sum_{i=1}^{d+1} \lambda_i(\boldsymbol{v}) \right) \boldsymbol{v} \quad \forall \boldsymbol{v} \in \triangle$$

must hold.

It is easy to see that these coordinates are homogeneous: if they are all multiplied by a scalar, the same equation still holds for the same $\boldsymbol{v}$. However, in some contexts, it is useful to treat these coordinates are uniquely determined, hence why a further condition is imposed,

in the form of asking their sum to be equal to 1. These *normalized barycentric coordinates* are therefore defined such that, given a simplex $\triangle$, the equations

$$\sum_{i=1}^{d+1} \lambda_i(\boldsymbol{v}) = 1 \quad \forall \boldsymbol{v} \in \triangle$$
$$\sum_{i=1}^{d+1} \lambda_i(\boldsymbol{v}) \boldsymbol{v}_i = \boldsymbol{v} \quad \forall \boldsymbol{v} \in \triangle$$

must hold. These two equations are given the name of *partition of unity property* and *linear reproduction property*, respectively.

For the purposes of this thesis, whenever we discuss barycentric coordinates and their generalizations, we will assume they are normalized and therefore uniquely determined for each point. Note that this allows us to see these coordinates as functions over $\triangle$, hence the notation used in these formulas.

It is possible to now extend the notion of barycentric coordinates to any polytope, instead of just simplices. Given a polytope P in a d-dimensional Euclidian space, noting with $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ its vertices, we would like to find a set of *generalized barycentric coordinates* (*GBCs*) $\lambda_1, \cdots, \lambda_n$ which satisfy the *partition of unity property* and the *linear reproduction property*:

$$\sum_{i=1}^{n} \lambda_i(\boldsymbol{v}) = 1 \quad \forall \boldsymbol{v} \in \overline{P}$$
$$\sum_{i=1}^{n} \lambda_i(\boldsymbol{v}) \boldsymbol{v}_i = \boldsymbol{v} \quad \forall \boldsymbol{v} \in \overline{P}$$

Unlike barycentric coordinates on a simplex, there are many possible choices for functions satisfying these conditions for a generic polytope. In fact, it is also useful to impose further conditions that, while not necessary, would be desirable:

- *Lagrange property*: $\lambda_i(\boldsymbol{v}_j) = \delta_{i,j}$ (with $\delta_{i,j}$ being the Kronecker delta); (2.1a)
- *Linearity on the boundary*: λ_i is linear on each facet of P; (2.1b)
- *Non-negativity*: $\lambda_i(\boldsymbol{v}) \geq 0 \ \forall \boldsymbol{v} \in \overline{P}$; (2.1c)
- *Smoothness*: $\lambda_i \in C^\infty$ (2.1d)

Note that, due to their uniqueness, barycentric coordinates on a simplex satisfy all these conditions.

2.2 A few examples of 2D generalized barycentric coordinates

For the purposes of this thesis, we will focus on coordinates over a 2D space from now on. In that respect, it can actually be shown that continuous GBCs which satisfy property (2.1c) also satisfy properties (2.1b) and (2.1a) (or at least they can be extended to GBCs that do).

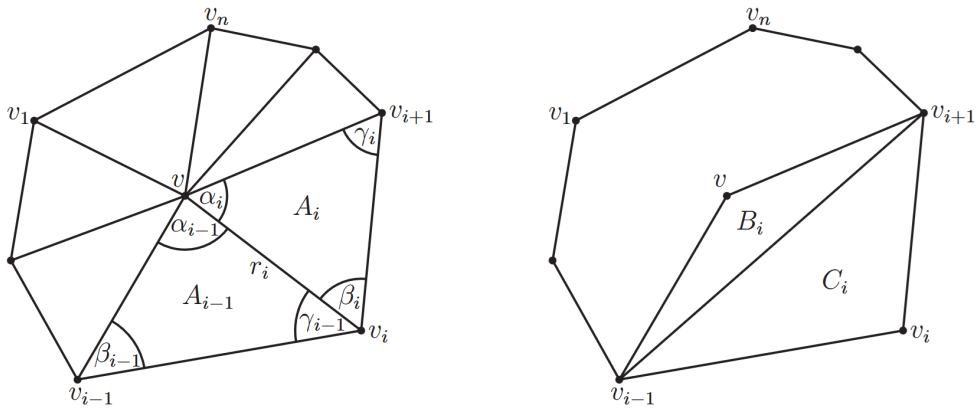


Figure 2.1. Notations for areas and angles.

For a while, much of the research on 2D GBCs focused on generalizations over convex polygons, specifically. In that regard, one of the most notable types of coordinates are *Wachspress coordinates*:

$$\lambda_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}; \quad \omega_i(\boldsymbol{v}) = \frac{\cot\gamma_{i-1} + \cot\beta_i}{||\boldsymbol{v}_i - \boldsymbol{v}||^2}; \quad i = 1, \cdots, n$$

with γ_i and β_i as in Figure 2.1.

These coordinates have been first proposed by Wachspress [1975] and later studied by Warren [1996] and Meyer et al. [2002]. When defined, they satisfy properties (2.1c) and (2.1d), which also means they satisfy (2.1a) and (2.1b). Their definition can only be applied to strongly convex polygons, as the denominator vanishes otherwise, but it can actually be can be shown that they can be extended to weakly convex polygons.

Wachspress coordinates are also *affine invariant*: if we denote $\lambda_i(\boldsymbol{v})$'s dependency on $\boldsymbol{v}_i$ by writing it as $\lambda_i(\boldsymbol{v}; \boldsymbol{v}_1, \cdots, \boldsymbol{v}_n)$, it holds that

$$\lambda_i(T\boldsymbol{v}; T\boldsymbol{v}_1, \cdots, T\boldsymbol{v}_n) = \lambda_i(\boldsymbol{v}; \boldsymbol{v}_1, \cdots, \boldsymbol{v}_n) \tag{2.2}$$

for any affine transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$.

Wachspress coordinates belong to a family of coordinates known as *three-point coordinates*: given a polytope P and defining $r_i = ||\boldsymbol{v}_i - \boldsymbol{v}||$ (taking $r_{n+1} = r_1$), it is possible to define $\lambda_{i,p}$ such that

$$\lambda_{i,p} = \frac{\omega_{i,p}}{\sum_{j=1}^{n} \omega_{j,p}}; \quad \omega_{i,p}(\boldsymbol{v}) = \frac{r_{i-1}^p A_i - r_i^p B_i + r_{i+1}^p A_{i-1}}{A_{i-1} A_i}; \quad i = 1, \cdots, n \tag{2.3}$$

with A_i and B_i as in Figure 2.1. $\lambda_{i,p}$ defined this way are GBCs, whenever they are well defined (e.g. when the denominator does not vanish). They are also *similarity invariant*: equation (2.2), which holds for any affine transformation for Wachspress coordinates, also holds for any three-point coordinates if T is specifically a similarity.

Three-point coordinates are actually a more specific version of a general definition: Floater et al. [2006] show that any set of GBCs λ_i can be expressed in the form

$$\lambda_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}; \quad \omega_i = \frac{c_{i-1} A_i - c_i B_i + c_{i+1} A_{i-1}}{A_{i-1} A_i}; \quad i = 1, \cdots, n$$

with A_i and B_i as in Figure 2.1 and with $c_i : \overline{P} \to \mathbb{R},\ i = 1, \cdots, n$ being a set of arbitrary real functions (taking $c_{n+1} = c_1$). Three-point coordinates are a useful sub-family to consider, because of their similarity invariance and because the definition of each λ_i only depends on $\boldsymbol{v}_{i-1}$, $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i+1}$.

More specifically, Wachspress coordinates are three-point coordinates for $p = 0$ in (2.3). Other values of p yield other coordinates. For instance, $p = 2$ corresponds to *discrete harmonic coordinates*:

$$\lambda_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j};\quad \omega_i(\boldsymbol{v}) = \cot\beta_{i-1} + \cot\gamma_i;\quad i = 1, \cdots, n$$

with β_i and γ_i as in Figure 2.1.

These coordinates were studied by Pinkall and Polthier [1993] and Eck et al. [1995]. They arise from the standard piecewise linear finite element approximation to the Laplace equation. Just like Wachspress coordinates, they are only well-defined on strongly convex polygons. They satisfy properties (2.1a) and (2.1b), but not necessarily (2.1c). In fact, they are only positive on the interior of a polygon if all vertices of the polygon lie on a circle. In that scenario, it can be shown that they are equal to Wachspress coordinates.

The GBCs we are most interested in are the three-point coordinates we obtain by setting $p = 1$ in (2.3), known as *mean value coordinates*:

$$\lambda_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j};\quad \omega_i(\boldsymbol{v}) = \frac{\tan(\alpha_{i-1}/2) + \tan(\alpha_i/2)}{2};\quad i = 1, \cdots, n$$

with α_i as in Figure 2.1.

2.3 Mean value coordinates

Mean value coordinates were introduced by Floater [2003] and studied by Hormann and Floater [2006]. They were originally obtained as a means to approximate harmonic functions through the use of the *mean value theorem*, from which they get their name. Indeed, for a harmonic function $u : \Omega \to \mathbb{R}$ and a disc $B = B(\boldsymbol{v}_0, r) \subseteq \Omega$ with boundary Γ, it holds that

$$u(\boldsymbol{v}_0) = \frac{1}{2\pi r} \int_\Gamma u(\boldsymbol{v}) ds \tag{2.4a}$$

(*circumferential mean value theorem*) and

$$u(\boldsymbol{v}_0) = \frac{1}{\pi r^2} \int_B u(\boldsymbol{v}) dx dy \tag{2.4b}$$

(*solid mean value theorem*).

If a triangulation $\mathcal{T}$ of Ω is defined, u can be approximated as a map $u_\mathcal{T}$ over $\mathcal{T}$. It can then be imposed that the mean value theorem (either variation) also holds for $u_\mathcal{T}$. This can be used to obtain a *convex approximation* of u by deriving the equation

$$u_\mathcal{T} = \sum_{i=1}^{n} \lambda_i u_\mathcal{T}(\boldsymbol{v}_i) \tag{2.5}$$

with λ_i being the mean value coordinates and $\boldsymbol{v}_i$ being the vertices on the boundary.

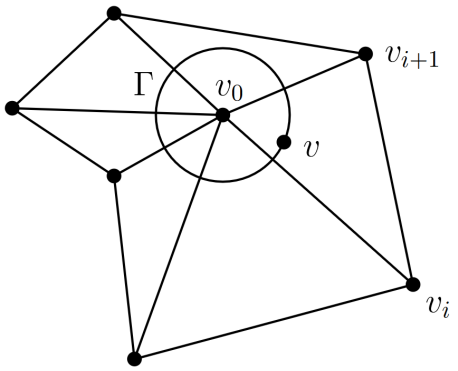


Figure 2.2. Circle for the application of the mean value theorem.

More specifically, fixing an interior triangulation vertex $\boldsymbol{v}_0$, a disc $B = B(\boldsymbol{v}_0, r)$ is taken, such that no other interior vertices are contained in B, as in Figure 2.2. The areas B_i (and curves Γ_i) are then defined as the portion of B (or Γ) contained within the triangle delimitated by $\boldsymbol{v}_0$, $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i+1}$. $u_{\mathscr{T}}(\boldsymbol{v}_0)$ is written as an integral on B (or Γ) by applying (2.4b) (or (2.4a)); then, the integral is written as the sum of integrals over every B_i (or Γ_i). It is finally possible to prove that every addend of this sum is exactly equal to $\lambda_i(\boldsymbol{v}_0)u_{\mathscr{T}}(\boldsymbol{v}_i)$.

Mean value coordinates have an advantage over the GBCs considered earlier, in that they can be easily defined over non-convex polygons. In fact, it is possible to define them not just over simple polygons, but also over sets of simple polygons, even nested ones. Not only that, but once defined, their domain can be extended to all of $\mathbb{R}^2$. They also have *affine precision*: the equality in (2.5) also holds for any affine function $\phi : \mathbb{R}^2 \to \mathbb{R}^m$.

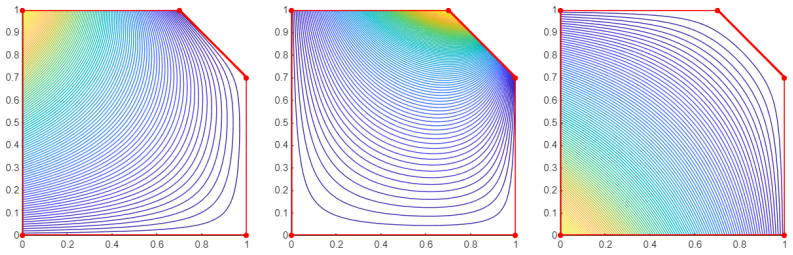
Figure 2.3 shows a comparison between Wachspress coordinates, discrete harmonic coordinates and mean value coordinates over the same polygon.

Mean value coordinates satisfy properties (2.1a), (2.1b) and (2.1d) (though they are only C^0 on the polygon vertices), but not necessarily (2.1c): they can be negative over arbitrary polygons. However, they are positive on the kernel of star-shaped polygons, which also implies they are positive on the interior of convex polygons. In fact, it can be shown that Wachspress coordinates and mean value coordinates are the only three-point coordinates which satisfy (2.1c) for convex polygons. This actually motivates the search for *five-point coordinates* with the same property: indeed, both Wachspress coordinates and mean value coordinates can be shown to have such five-point generalizations.

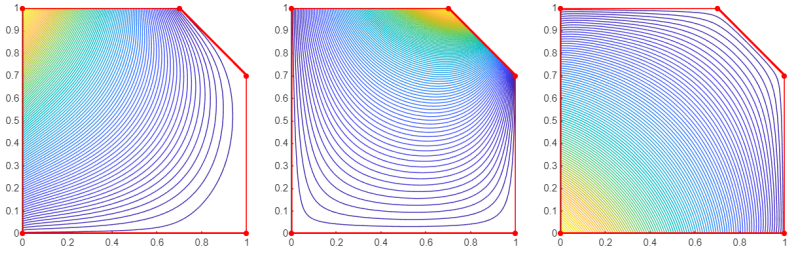
What Wachspress coordinates have over mean value coordinates is affine invariance. Mean value coordinates are actually invariant with respect to similarities, but not with respect to general affine transformations.

On the other hand, the possibility to define mean value coordinates over non-convex polygons is a lot more important for the purposes of this thesis, hence why they are our main focus. There exist other GBCs which can be defined over generic simple polygons, such as *metric coordinates* and *Gordon-Wixom coordinates*, but mean value coordintes are the ones with the simplest closed form.

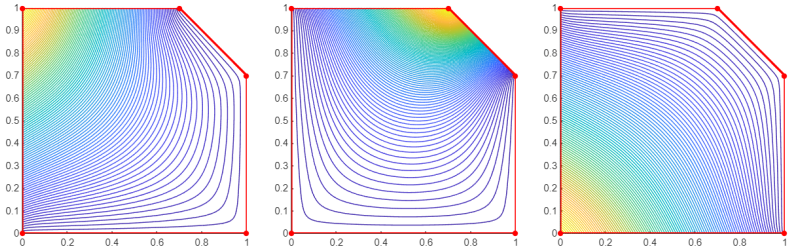
Mean value coordinates and GBCs in general have a variety of applications. A common one involves interpolating values that are given at the vertices of a set triangulation, for which



(a) Wachspress coordinates.



(b) Discrete harmonic coordinates.



(c) Mean value coordinates.

Figure 2.3. Comparison between different GBCs over the polygon P with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,0.7), \boldsymbol{v}_4 = (0.7,1), \boldsymbol{v}_5 = (0,1)$, through a contour plot with 100 countour lines.

GBCs with properties (2.1a) and (2.1b) work especially well. This can also be used for image warping, or for morphing two compatible triangulations into each other: for these purposes, mean value coordinates yield promising empirical results. For image warping specifically, mean value coordinates allow to define a warp that preserves straight lines along the edges of source polygons. The fact that mean value coordinates can be defined for arbitrary simple polygons obviously aids in both these applications, as does the fact that they are smooth everywhere except on the polygon vertices: the latter is also useful in applications like shading. Not to mention that, since the definition of mean value coordinates relies on their relationship with harmonic functions, there is a possibility they could be used to approximate harmonic maps given a fine enough triangulation.

However, for this thesis, we are more concerned with the application of mean value coordinates to the finite element method: we would like to use them in order to construct a set of basis functions of any order.

Chapter 3

Quadratic serendipity coordinates

3.1 Introduction to quadratic serendipity coordinates

The finite element method (FEM) is an approach which can be applied for the numerical solution of partial differential equations (PDEs). Being a numerical method, it operates on a discretization of the analytical domain: for example, on a partition of a 1D interval; or on a triangulation of a 2D subset of the plane. In our case, we would like to apply the method on a mesh of miscellaneous simple polygons over a 2D domain.

Every "portion" of the discretization, e.g. every polygon in our case, is called an *element*. The idea behind FEM is to approximate the solution to the PDE problem over each element, yielding a result that is, for example, piecewise linear, or piecewise quadratic. More generally, a finite element approximation of a fixed order (e.g. linear, quadratic) involves approximating the PDE solution as a function of that order on every element.

In order to do so, a set of *basis functions* is defined on each element: over the element in question, the approximate solution will be written as a linear combination of those basis functions. Because of this, it must be possible to write every polynomial of the desired order as a linear combination of those basis functions; or in other words, the basis functions must be able to *reproduce* every polynomial of that order. Basis functions for an FEM generally also need to satisfy a variation of (2.1a) over a set of *interpolation nodes* on the element of definition, but we will ignore this for now and expand upon it in a later chapter.

We note the following:

Remark 3.1.1. *If a set of functions $\phi_i : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$ can reproduce any monomial $M = M(x, y)$ of degree up to q, i.e. if they are such that*

$$\sum_{i=1}^{n} \alpha_i \phi_i(x, y) = M(x, y) \quad \forall (x, y) \in \Omega$$

for appropriate choices of α_i and for every monomial $M = M(x, y)$ of degree up to q, then the functions ϕ_i can reproduce any polynomial of degree up to q.

This is trivial from the definition of linear combination.

We recall that GBCs satisfy the *partition of unity property* and the *linear reproduction property*:

$$\sum_{i=1}^{n} \lambda_i(\boldsymbol{v}) = 1 \quad \forall \boldsymbol{v} \in \overline{P}$$
$$\sum_{i=1}^{n} \lambda_i(\boldsymbol{v}) \boldsymbol{v}_i = \boldsymbol{v} \quad \forall \boldsymbol{v} \in \overline{P}$$

We can write $\boldsymbol{v} = (x, y)$, with x and y being the Cartesian coordinates of $\boldsymbol{v}$. We do the same with $\boldsymbol{v}_i = (x_i, y_i)$, $i = 1, \cdots, n$. This allows us to rewrite the two properties as

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i(x, y) &= 1 \quad \forall (x, y) \in \overline{P} \\ \sum_{i=1}^{n} \lambda_i(x, y) x_i &= x \quad \forall (x, y) \in \overline{P} \\ \sum_{i=1}^{n} \lambda_i(x, y) y_i &= y \quad \forall (x, y) \in \overline{P} \end{aligned} \tag{3.1}$$

Through Remark 3.1.1, this shows that any set of GBCs can reproduce linear polynomials over their polygon of definition $\overline{P}$. This makes them candidates for basis functions for an FEM of linear order.

However, we are also interested in how GBCs can be used in order to find basis functions for an FEM of higher order. For now, we focus on the quadratic order. Let $\lambda_1, \cdots, \lambda_n$ be a set of GBCs (mean value coordinates in our case) over a polygon $\overline{P}$. We then state the following:

Theorem 3.1.2. *The set of functions of the form μ_{ab}, with $\mu_{ab} = \lambda_a \lambda_b$, is such that*

$$\begin{aligned} \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) = 1; \quad \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) x_a = x; \\ \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) y_a = y; \quad \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) x_a x_b = x^2; \\ \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) x_a y_b = xy; \quad \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) y_a y_b = y^2 \end{aligned}$$

for every $(x, y) \in \overline{P}$.

Proof. Let $(x, y) \in \overline{P}$. Let us consider the case

$$\sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) x_a y_b = xy$$

It holds that

$$\begin{aligned} \sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab}(x, y) x_a y_b &= \sum_{a=1}^{n}\sum_{b=1}^{n} \lambda_a(x, y) \lambda_b(x, y) x_a y_b \\ &= \left(\sum_{a=1}^{n} \lambda_a(x, y) x_a \right) \left(\sum_{b=1}^{n} \lambda_b(x, y) y_b \right) = xy \qquad (\star) \end{aligned}$$

where $(\star)$ is by (3.1). All other cases can be proven in a similar way. □

To make the writing more compact, we introduce a new notation. Given a monomial $M = M(x,y)$ of degree at most 2, we define $\hat{M}(a,b),\ 1 \le a,b \le n$ such that

$$\begin{aligned} \hat{M}(a,b) = 1 \text{ if } M(x,y) = 1;\ \ \hat{M}(a,b) = x_a \text{ if } M(x,y) = x; \\ \hat{M}(a,b) = y_a \text{ if } M(x,y) = y;\ \ \hat{M}(a,b) = x_a x_b \text{ if } M(x,y) = x^2; \\ \hat{M}(a,b) = x_a y_b \text{ if } M(x,y) = xy;\ \ \hat{M}(a,b) = y_a y_b \text{ if } M(x,y) = y^2 \end{aligned} \tag{3.2}$$

Thus, Theorem 3.1.2 can be more succintly rewritten as

Corollary 3.1.3. *The set of functions of the form μ_{ab}, with $\mu_{ab} = \lambda_a \lambda_b$, is such that*

$$\sum_{a=1}^{n}\sum_{b=1}^{n} \hat{M}(a,b)\mu_{ab}(x,y) = M(x,y)$$

for every monomial $M = M(x,y)$ of degree up to 2.

By Remark 3.1.1, it therefore holds that

Corollary 3.1.4. *The set of functions of the form μ_{ab}, with $\mu_{ab} = \lambda_a \lambda_b$, can reproduce polynomials of degree up to 2.*

We call these functions the *quadratic coordinates* over $\overline{P}$ (defined with respect to mean value coordinates). Figure 3.1 shows a few examples of quadratic coordinates over polygons.

Since $\mu_{ab} = \mu_{ba}$, there are $n + \binom{n}{2} = \frac{n(n+1)}{2}$ total distinct quadratic coordinates (n of the form μ_{aa} and $\binom{n}{2}$ of the form $\mu_{ab},\ a \neq b$). It follows from Corollary 3.1.4 that these functions could already be used as candidates for basis functions for a quadratic FEM. However, the number of these functions scales with a factor of $O(n^2)$: it would be more convenient to reduce the number of functions to work with.

This is the idea behind the construction of quadratic *serendipity coordinates*, pioneered by Rand et al. [2014]. They are a smaller set of functions, written as linear combinations of coordinate functions, such that they keep the same properties of polynomial reproducibility. We will explain the approach employed by Hackemack and Ragusa [2018] in order to construct quadratic serendipity coordinates, then we will extend that approach to the construction of serendipity coordinates based on coordinates of any order.

3.2 Constructing quadratic serendipity coordinates

Let us introduce the notation

$$a\,b \tag{3.3}$$

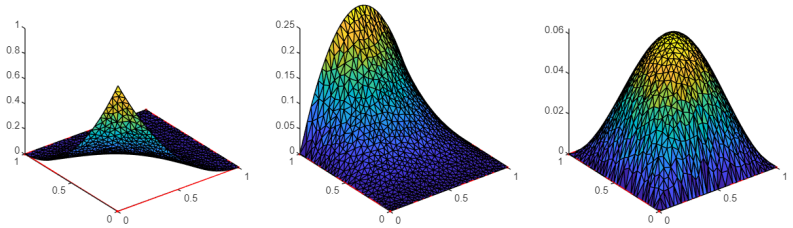
to indicate the unordered list made up by the indices in (a,b).

We recall the result in Corollary 3.1.3 based on the result of Theorem 3.1.2: it holds that

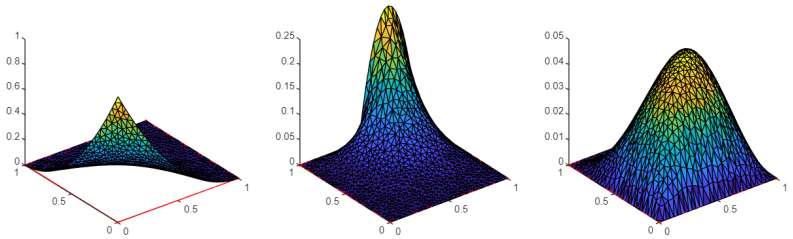
$$\sum_{a=1}^{n}\sum_{b=1}^{n} \hat{M}(a,b)\mu_{ab}(x,y) = M(x,y)\ \ \forall (x,y) \in \overline{P}$$

for any monomial $M(x,y)$ of degree up to 2.

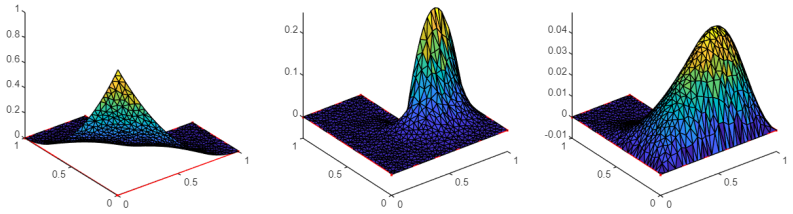
Like we noted earlier, $\mu_{cd} = \mu_{dc}$ for any two indices c,d. That means that, if $c \neq d$, then the function μ_{cd} will be counted twice in the sum (once for $a = c, b = d$ and once for $a = d, b = c$).



(a) P quadrilateral with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1), \boldsymbol{v}_4 = (0,1)$.



(b) P pentagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1), \boldsymbol{v}_4 = (\frac{1}{2},1), \boldsymbol{v}_5 = (0,1)$.



(c) P hexagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,\frac{1}{2}), \boldsymbol{v}_4 = (\frac{1}{2},\frac{1}{2}), \boldsymbol{v}_5 = (\frac{1}{2},1), \boldsymbol{v}_6 = (0,1)$.

Figure 3.1. Surf plot of quadratic coordinates μ_{11}, μ_{34} and μ_{13} (defined with respect to mean value coordinates) over different polygons. The surf plots are obtained by generating a triangulation over the polygon through the `triangle` code provided by Jonathan Shewchuk, imposing a maximum area constraint of 0.001.

On the other hand, the functions of the form μ_{ee} will only be counted once. Thus, the sum above can be rewritten as

$$\sum_{aa} \hat{M}(a,a)\mu_{aa}(x,y) + \sum_{ab;a\neq b} \left(\hat{M}(a,b) + \hat{M}(b,a)\right)\mu_{ab}(x,y) = M(x,y) \quad \forall (x,y) \in \overline{P} \tag{3.4}$$

for any monomial $M(x,y)$ of degree up to 2. We used the notation in (3.3) over these sums and we will continue to do so throughout this chapter.

The idea behind the construction of serendipity coordinates is to reduce the space of quadratic coordinates to a space of $2n$ functions of the form ξ_{ii} or $\xi_{i(i+1)}$, such that a similar constraint holds: we ask that

$$\sum_{ii} \hat{M}(i,i)\xi_{ii}(x,y) + \sum_{i(i+1)} \left(\hat{M}(i,i+1) + \hat{M}(i+1,i)\right)\xi_{i(i+1)}(x,y) = M(x,y) \quad \forall (x,y) \in \overline{P} \tag{3.5}$$

for any monomial $M(x,y)$ of degree up to 2. Note that, throughout this thesis, we treat indices as if they are cyclic, so $n+1 = 1$ and $0 = n$.

Example: If $M(x,y) = x$, then by Theorem 3.1.2 it holds that

$$\sum_a \sum_b \mu_{ab}(x,y)x_a = x \quad \forall (x,y) \in \overline{P}$$

As in (3.4), this result can be rewritten as

$$\sum_{aa} x_a \mu_{aa}(x,y) + \sum_{ab;a\neq b} (x_a + x_b)\mu_{ab}(x,y) = x \quad \forall (x,y) \in \overline{P}$$

because

$$\hat{M}(a,b) = x_a; \quad \hat{M}(b,a) = x_b$$

Thus, the quadratic serendipity functions should satisfy this constraint:

$$\sum_{ii} x_i \xi_{ii}(x,y) + \sum_{i(i+1)} (x_i + x_{i+1})\xi_{i(i+1)}(x,y) = x \quad \forall (x,y) \in \overline{P}$$

Example: If $M(x,y) = y^2$, then by Theorem 3.1.2 it holds that

$$\sum_a \sum_b \mu_{ab}(x,y)y_a y_b = y^2 \quad \forall (x,y) \in \overline{P}$$

As in (3.4), this result can be rewritten as

$$\sum_{aa} (y_a)^2 \mu_{aa}(x,y) + \sum_{ab;a\neq b} 2y_a y_b \mu_{ab}(x,y) = x \quad \forall (x,y) \in \overline{P}$$

because

$$\hat{M}(a,b) = y_a y_b = y_b y_a = \hat{M}(b,a)$$

Thus, the quadratic serendipity functions should satisfy this constraint:

$$\sum_{ii}(y_i)^2\xi_{ii}(x,y)+\sum_{i(i+1)}2y_iy_{i+1}\xi_{i(i+1)}(x,y)=y^2\ \ \forall(x,y)\in\overline{P}$$

Because each of the quadratic serendipity functions is a linear combination of the quadratic coordinates, we can write

$$\xi_{ij}=\sum_{ab}c_{ab}^{ij}\mu_{ab}\ \ \forall\xi_{ij} \tag{3.6}$$

Thus, the problem of constructing serendipity coordinates ξ_{ij} that satisfy (3.5) can be solved by looking for the appropriate values of the coefficients c_{ab}^{ij}.

Rand et al. [2014] show the following:

Theorem 3.2.1. *If ξ_{ij} are defined as in* (3.6) *with c_{ab}^{ij} such that*

$$\begin{aligned}\sum_{ii}\hat{M}(i,i)c_{aa}^{ii}+\sum_{i(i+1)}\big(\hat{M}(i,i+1)+\hat{M}(i+1,i)\big)c_{aa}^{i(i+1)}&=\hat{M}(a,a)\ \ \forall aa\\ \sum_{ii}\hat{M}(i,i)c_{ab}^{ii}+\sum_{i(i+1)}\big(\hat{M}(i,i+1)+\hat{M}(i+1,i)\big)c_{ab}^{i(i+1)}&=\hat{M}(a,b)+\hat{M}(b,a)\ \ \forall ab,\ a\neq b\end{aligned} \tag{3.7}$$

for any monomial $M=M(x,y)$ of degree at most 2, then ξ_{ij} satisfy (3.5).

Proof. Let $M=M(x,y)$ be a monomial of degree at most 2. Then, it holds that

$$\begin{aligned}&\sum_{ii}\hat{M}(i,i)\xi_{ii}(x,y)+\sum_{i(i+1)}\big(\hat{M}(i,i+1)+\hat{M}(i+1,i)\big)\xi_{i(i+1)}(x,y)\\ &=\sum_{ii}\hat{M}(i,i)\left(\sum_{ab}c_{ab}^{ii}\mu_{ab}(x,y)\right)+\sum_{i(i+1)}\big(\hat{M}(i,i+1)+\hat{M}(i+1,i)\big)\left(\sum_{ab}c_{ab}^{i(i+1)}\mu_{ab}(x,y)\right)\\ &=\sum_{ab}\sum_{ii}\hat{M}(i,i)c_{ab}^{ii}\mu_{ab}+\sum_{ab}\sum_{i(i+1)}\big(\hat{M}(i,i+1)+\hat{M}(i+1,i)\big)c_{ab}^{i(i+1)}\mu_{ab}(x,y)\\ &=\sum_{ab}\left(\sum_{ii}\hat{M}(i,i)c_{ab}^{ii}+\sum_{i(i+1)}\big(\hat{M}(i,i+1)+\hat{M}(i+1,i)\big)c_{ab}^{i(i+1)}\right)\mu_{ab}(x,y)\\ &=\sum_{aa}\hat{M}(a,a)\mu_{aa}(x,y)+\sum_{ab;a\neq b}\big(\hat{M}(a,b)+\hat{M}(b,a)\big)\mu_{ab}(x,y)=M(x,y)\qquad(\star\star)\end{aligned}$$

where $(\star\star)$ is due to the application of (3.7) and (3.4). □

Example: For $M(x,y)=x$ (i.e. $\hat{M}(a,b)=x_a$), the constraint that the coefficients have to satisfy is

$$\begin{aligned}\sum_{ii}x_ic_{aa}^{ii}+\sum_{i(i+1)}(x_i+x_{i+1})c_{aa}^{i(i+1)}&=x_a\ \ \forall aa\\ \sum_{ii}x_ic_{ab}^{ii}+\sum_{i(i+1)}(x_i+x_{i+1})c_{ab}^{i(i+1)}&=x_a+x_b\ \ \forall ab,\ a\neq b\end{aligned}$$

Indeed,

$$\begin{aligned}
&\sum_{ii} x_i \xi_{ii}(x,y) + \sum_{i(i+1)} (x_i + x_{i+1})\, \xi_{i(i+1)}(x,y) \\
&= \sum_{ii} x_i \left(\sum_{ab} c_{ab}^{ii} \mu_{ab}(x,y) \right) + \sum_{i(i+1)} (x_i + x_{i+1}) \left(\sum_{ab} c_{ab}^{i(i+1)} \mu_{ab}(x,y) \right) \\
&= \sum_{ab} \sum_{ii} x_i c_{ab}^{ii} \mu_{ab} + \sum_{ab} \sum_{i(i+1)} (x_i + x_{i+1})\, c_{ab}^{i(i+1)} \mu_{ab}(x,y) \\
&= \sum_{ab} \left(\sum_{ii} x_i c_{ab}^{ii} + \sum_{i(i+1)} (x_i + x_{i+1})\, c_{ab}^{i(i+1)} \right) \mu_{ab}(x,y) \\
&= \sum_{aa} x_a \mu_{aa}(x,y) + \sum_{ab;a\neq b} (x_a + x_b)\, \mu_{ab}(x,y) = x
\end{aligned}$$

Example: For $M(x,y) = y^2$ (i.e. $\hat{M}(a,b) = y_a y_b$), the constraint that the coefficients have to satisfy is

$$\begin{aligned}
\sum_{ii} (y_i)^2 c_{aa}^{ii} + \sum_{i(i+1)} 2 y_i y_{i+1} c_{aa}^{i(i+1)} &= (y_a)^2 \quad \forall aa \\
\sum_{ii} (y_i)^2 c_{ab}^{ii} + \sum_{i(i+1)} 2 y_i y_{i+1} c_{ab}^{i(i+1)} &= 2 y_a y_b \quad \forall ab,\ a \neq b
\end{aligned}$$

Indeed,

$$\begin{aligned}
&\sum_{ii} (y_i)^2 \xi_{ii}(x,y) + \sum_{i(i+1)} 2 y_i y_{i+1} \xi_{i(i+1)}(x,y) \\
&= \sum_{ii} (y_i)^2 \left(\sum_{ab} c_{ab}^{ii} \mu_{ab}(x,y) \right) + \sum_{i(i+1)} 2 y_i y_{i+1} \left(\sum_{ab} c_{ab}^{i(i+1)} \mu_{ab}(x,y) \right) \\
&= \sum_{ab} \sum_{ii} (y_i)^2 c_{ab}^{ii} \mu_{ab} + \sum_{ab} \sum_{i(i+1)} 2 y_i y_{i+1} c_{ab}^{i(i+1)} \mu_{ab}(x,y) \\
&= \sum_{ab} \left(\sum_{ii} (y_i)^2 c_{ab}^{ii} + \sum_{i(i+1)} 2 y_i y_{i+1} c_{ab}^{i(i+1)} \right) \mu_{ab}(x,y) \\
&= \sum_{aa} (y_a)^2 \mu_{aa}(x,y) + \sum_{ab;a\neq b} 2 y_a y_b \mu_{ab}(x,y) = y^2
\end{aligned}$$

For $n > 3$, this problem is underconstrained: there are more coefficients c_{ab}^{ij} than the constraint equations determining them. This means there is some freedom when it comes to choosing the coefficients. In order to make a smart choice, we will rewrite this problem as a linear system.

Let us first order the quadratic coordinates based on their indices. We place all functions of the form μ_{aa} first, ordered for a in ascending order. Then we place all functions of the form $\mu_{a(a+1)}$ (ending with μ_{n1}), also ordered for a in ascending order. Finally, we place all other functions in lexicographical order. This gives an array of functions $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{nn} & \mu_{12} & \cdots & \mu_{n1} & \mu_{13} & \cdots & \mu_{(n-2)n} \end{bmatrix}^T$$

Let us now do the same with the quadratic serendipity coordinates. We follow the same ordering, but since we only have to order $2n$ functions, we skip the lexicographical ordering step. This yields an array of functions $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_{11} & \cdots & \xi_{nn} & \xi_{12} & \cdots & \xi_{n1} \end{bmatrix}^T$$

It is now possible to write (3.6) as a linear algebra problem:

$$\boldsymbol{\xi} = \boldsymbol{A}\boldsymbol{\mu}$$

with $\boldsymbol{A}$ being the matrix of the coefficients c_{ab}^{ij}:

$$\boldsymbol{A} = \begin{bmatrix} c_{11}^{11} & \cdots & c_{ab}^{11} & \cdots & c_{(n-2)n}^{11} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{11}^{ij} & \cdots & c_{ab}^{ij} & \cdots & c_{(n-2)n}^{ij} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{11}^{n1} & \cdots & c_{ab}^{n1} & \cdots & c_{(n-2)n}^{n1} \end{bmatrix}$$

The issue of choosing the coefficients c_{ab}^{ij} thus becomes an issue of choosing the structure of $\boldsymbol{A}$. We choose a simple structure:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & | & \boldsymbol{A}' \end{bmatrix} \tag{3.8}$$

with $\boldsymbol{I}$ being the $2n \times 2n$ identity matrix. Note that this structure satisfies the equations in (3.7) for the coefficients of the form c_{aa}^{ij} and $c_{a(a+1)}^{ij}$. Therefore, the problem reduces to finding a good choice of coefficients for the elements in $\boldsymbol{A}'$.

Up until now, we have followed the lead of Rand et al. [2014]. This is where their approach diverges from the one by Hackemack and Ragusa [2018]; and thus we will now follow the latter.

We can isolate a column $\boldsymbol{c}_{ab}$ of $\boldsymbol{A}'$ like so:

$$\boldsymbol{c}_{ab} = \begin{bmatrix} c_{ab}^{11} & \cdots & c_{ab}^{nn} & c_{ab}^{12} & \cdots & c_{ab}^{n1} \end{bmatrix}^T$$

Fixing a monomial $M = M(x, y)$ of degree up to 2, this allows us to rewrite (3.7) for the coefficients in $\boldsymbol{c}_{ab}$ as a product of arrays:

$$\begin{bmatrix} \hat{M}(1,1) & \cdots & \hat{M}(n,n) & \hat{M}(1,2)+\hat{M}(2,1) & \cdots & \hat{M}(n,1)+\hat{M}(1,n) \end{bmatrix} \boldsymbol{c}_{ab} = \hat{M}(a,b)+\hat{M}(b,a)$$

Note that the choice of the term on the right-hand side comes from the fact that, by construction of $\boldsymbol{A}'$ and by the fact that $\boldsymbol{c}_{ab}$ is a column of $\boldsymbol{A}'$, it holds that $a \neq b$.

There are six monomials of degree up to 2, meaning that, if the coefficients in $\boldsymbol{c}_{ab}$ satisfy (3.7), then they must satisfy six constraints equations. Therefore, all of these equations can be condensed into a single linear system:

$$\boldsymbol{B}\boldsymbol{c}_{ab} = \boldsymbol{q}_{ab} \tag{3.9}$$

with $\boldsymbol{B}$ and $\boldsymbol{q}_{ab}$ defined as follows:

$$\boldsymbol{B} = \begin{bmatrix} 1 & \cdots & 1 & 2 & \cdots & 2 \\ x_1 & \cdots & x_n & x_1+x_2 & \cdots & x_n+x_1 \\ y_1 & \cdots & y_n & y_1+y_2 & \cdots & y_n+y_1 \\ (x_1)^2 & \cdots & (x_n)^2 & 2x_1x_2 & \cdots & 2x_nx_1 \\ x_1y_1 & \cdots & x_ny_n & x_1y_2+x_2y_1 & \cdots & x_ny_1+x_1y_n \\ (y_1)^2 & \cdots & (y_n)^2 & 2y_1y_2 & \cdots & 2y_ny_1 \end{bmatrix}; \boldsymbol{q}_{ab} = \begin{bmatrix} 2 \\ x_a+x_b \\ y_a+y_b \\ 2x_ax_b \\ x_ay_b+x_by_a \\ 2y_ay_b \end{bmatrix} \tag{3.10}$$

Each row of $\boldsymbol{B}$ and $\boldsymbol{q}_{ab}$ represents a different constraint equation, corresponding to a different monomial M of degree at most 2. It is easy to see this linear system is indeed equivalent to the equations in (3.7).

Through this reasoning, we can write several linear systems of the form in (3.9), one for each column of $\boldsymbol{A}'$. This actually means we can condense them all into one single matrix equation: indeed, by definition, it holds that

$$\boldsymbol{A}' = \begin{bmatrix} \boldsymbol{c}_{13} & | & \cdots & | & \boldsymbol{c}_{(n-2)n} \end{bmatrix}$$

Therefore, we can write

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{q}_{13} & | & \cdots & | & \boldsymbol{q}_{(n-2)n} \end{bmatrix}$$

and represent all systems in (3.9) at once as

$$\boldsymbol{B}\boldsymbol{A}' = \boldsymbol{Q}$$

Both $\boldsymbol{B}$ and $\boldsymbol{Q}$ are known, by definition. In order to find $\boldsymbol{A}'$ from them, $\boldsymbol{B}$ is inverted by calculating its *Moore-Penrose pseudoinverse*:

$$\boldsymbol{A}' = \boldsymbol{B}^\dagger \boldsymbol{Q}; \;\; \boldsymbol{B}^\dagger = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}$$

For an underconstrained problem, this pseudoinverse yields a solution satisfying a least squares estimation. In practice, this method provides an easy way of solving the problem regardless of the shape of the polygon of definition: it yields valid quadratic interpolants on convex and concave polygons alike.

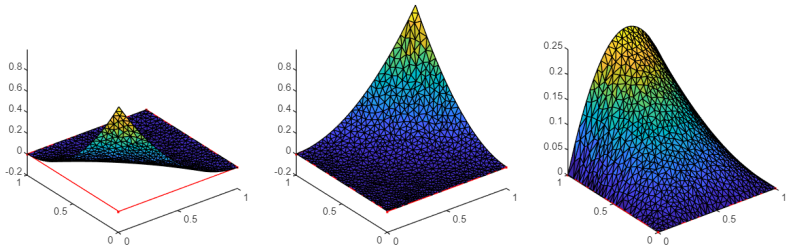
We now know all necessary steps in order to calculate the quadratic serendipity coordinates: we can calculate $\boldsymbol{A}'$ and therefore $\boldsymbol{A}$, then find the serendipity coordinates through (3.6). Note that, because of the structure of $\boldsymbol{A}$ we chose in (3.8), (3.6) can be written as

$$\xi_{ij} = \mu_{ij} + \sum_{ab;b\notin\{a,a+1\}} c_{ab}^{ij}\mu_{ab} \;\; \forall \xi_{ij} \tag{3.11}$$

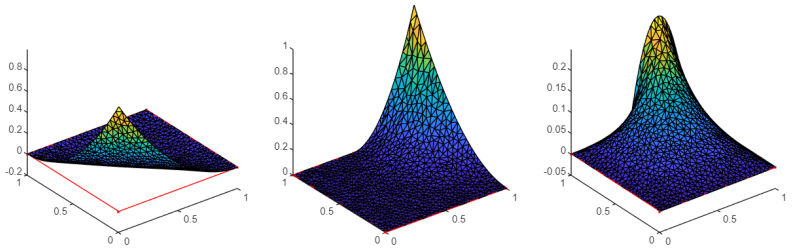
This also implies the following:

Remark 3.2.2. *The serendipity coordinates ξ_{ij} are proportional to the quadratic Bernstein basis polynomials when restricted to the edges of $\overline{P}$.*

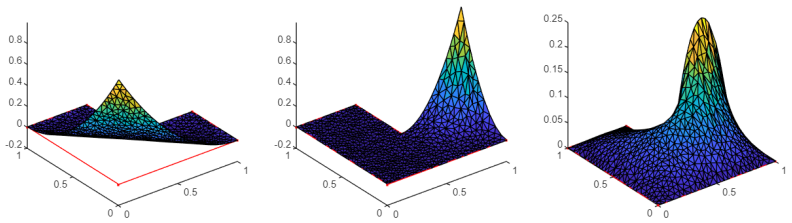
Proof. Because the coordinates λ_a satisfy (2.1a) and (2.1b), it holds that $\{\lambda_i, \lambda_{i+1}\}$ are the linear Bernstein basis polynomials on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$. This implies that $\{\mu_{ii}, \mu_{i(i+1)}, \mu_{(i+1)(i+1)}\}$ are proportional to the quadratic Bernstein basis polynomials on the same edge. Not only that, but because of those same properties of λ_a, it holds that μ_{ab} vanishes on all edges of $\overline{P}$



(a) P quadrilateral with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1), \boldsymbol{v}_4 = (0,1)$.



(b) P pentagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1), \boldsymbol{v}_4 = (\frac{1}{2},1), \boldsymbol{v}_5 = (0,1)$.



(c) P hexagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,\frac{1}{2}), \boldsymbol{v}_4 = (\frac{1}{2},\frac{1}{2}), \boldsymbol{v}_5 = (\frac{1}{2},1), \boldsymbol{v}_6 = (0,1)$.

Figure 3.2. Surf plot of quadratic serendipity coordinates ξ_{11}, ξ_{33} and ξ_{34} (defined with respect to mean value coordinates) over different polygons.

if $b \notin \{a, a+1\}$. Thus, by (3.11), it holds that $\xi_{ij} = \mu_{ij}$ $\forall \xi_{ij}$ on the edges of $\overline{P}$, meaning $\{\xi_{ii}, \xi_{i(i+1)}, \xi_{(i+1)(i+1)}\}$, too, are proportional to the quadratic Bernstein basis polynomials on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$. □

Since the quadratic Bernstein basis polynomials are a basis for the space of quadratic polynomials, this remark implies that the space spanned by serendipity functions is exactly the space of quadratic polynomials when restricted to the edges of $\overline{P}$.

Figure 3.2 shows a few examples of quadratic serendipity coordinates over polygons. As stated in the proof of Remark 3.2.2, each serendipity coordinate ξ_{ij} is equal to the corresponding quadratic coordinate μ_{ij} on the boundary of the polygon.

Before moving on, let us observe the matrix $\boldsymbol{M}$ defined as follows:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{B} & | & \boldsymbol{Q} \end{bmatrix}$$

By definition of $\boldsymbol{B}$ and $\boldsymbol{Q}$, that means the elements of $\boldsymbol{M}$ are the following:

$$\boldsymbol{M} = \begin{bmatrix} 1 & \cdots & 1 & 2 & \cdots & 2 & 2 & \cdots & 2 \\ x_1 & \cdots & x_n & x_1+x_2 & \cdots & x_n+x_1 & x_1+x_3 & \cdots & x_{n-2}+x_n \\ y_1 & \cdots & y_n & y_1+y_2 & \cdots & y_n+y_1 & y_1+y_3 & \cdots & y_{n-2}+y_n \\ (x_1)^2 & \cdots & (x_n)^2 & 2x_1x_2 & \cdots & 2x_nx_1 & 2x_1x_3 & \cdots & 2x_{n-2}x_n \\ x_1y_1 & \cdots & x_ny_n & x_1y_2+x_2y_1 & \cdots & x_ny_1+x_1y_n & x_1y_3+x_3y_1 & \cdots & x_{n-2}y_n+x_ny_{n-2} \\ (y_1)^2 & \cdots & (y_n)^2 & 2y_1y_2 & \cdots & 2y_ny_1 & 2y_1y_3 & \cdots & 2y_{n-2}y_n \end{bmatrix}$$

From this definition, it is clear that the elements of $\boldsymbol{Q}$ follow exactly the same structure as the last n elements of $\boldsymbol{B}$. Not only that, but, because of the ordering of $\boldsymbol{B}$ and $\boldsymbol{Q}$, the columns of $\boldsymbol{M}$ are ordered with the same logic as the quadratic coordinates in the array $\boldsymbol{\mu}$, index-wise. In fact, these two observations are connected: the structure of $\boldsymbol{M}$ is such that (3.4) can be rewritten as

$$\boldsymbol{M}\boldsymbol{\mu}(x,y) = \begin{bmatrix} 1 \\ x \\ y \\ x^2 \\ xy \\ y^2 \end{bmatrix}$$

Thus, the columns of $\boldsymbol{M}$ are actually directly linked to the elements of $\boldsymbol{\mu}$ through (3.4). The reason for this has to do with the properties of c_{ab}^{ij} and with the process through which they were derived in Theorem 3.2.1. We will use $\boldsymbol{M}$ more explicitly for the construction of $\boldsymbol{A}$ in the next chapter.

Chapter 4

Extending the approach

4.1 Reproducibility of polynomials of any order

The previous chapter focused on retreading the work of other papers in order to construct quadratic coordinates and reduce them to quadratic coordinates. This chapter will use that work as a basis and extend that approach to any order of polynomials.

We once again recall the result in Corollary 3.1.3 based on the result of Theorem 3.1.2: it holds that

$$\sum_{a=1}^{n}\sum_{b=1}^{n}\hat{M}(a,b)\mu_{ab}(x,y) = M(x,y) \quad \forall (x,y)\in\overline{P}$$

for any monomial $M(x,y)$ of degree up to 2. The short-hand of $\hat{M}(a,b)$ is used to write this result in a more compact fashion: we would like to prove a similar result through a similar notation for monomials of arbitrary order.

Thus, we introduce the following notation: given a monomial M, we define $\hat{M}$ such that

$$\begin{aligned} \hat{M}(a_1,\cdots,a_q) = x_{a_1}\cdots x_{a_k}y_{a_{k+1}}\cdots y_{a_{k+l}} \quad \text{for } M(x,y)=x^k y^l,\ k+l\neq 0; \\ \hat{M}(a_1,\cdots,a_q) = 1 \quad \text{for } M(x,y)=1 \end{aligned} \tag{4.1}$$

for any ordered list of indices $(a_1,\cdots,a_q)$ such that $k+l\leq q$ and $1\leq a_i\leq n,\ i=1,\cdots,q$.

One way to visualize this definition of $\hat{M}$ is to write M explicitly as a product of x, y and 1 and then "apply" the indices in $(a_1,\cdots,a_q)$ to the product. In practice, M is written as

$$M(x,y) = \underbrace{\underbrace{x\cdot\ldots\cdot x}_{k}\cdot\underbrace{y\cdot\ldots\cdot y}_{l}\cdot 1\cdot\ldots\cdot 1}_{q}$$

and thus $\hat{M}$ can be written as

$$\hat{M}(a_1,\cdots,a_q) = \underbrace{\underbrace{x_{a_1}\cdot\ldots\cdot x_{a_k}}_{k}\cdot\underbrace{y_{a_{k+1}}\cdot\ldots\cdot y_{a_{k+l}}}_{l}\cdot 1\cdot\ldots\cdot 1}_{q}$$

This is also an extension of the notation for $\hat{M}(a,b)$ in (3.2).

Example: If $M(x,y) = x$, then

$$M(x,y) = x \cdot 1$$

and thus

$$\hat{M}(a,b) = x_a \cdot 1$$

Thus $\hat{M}(a,b) = x_a$, which corresponds to the definition of $\hat{M}(a,b)$ in (3.2).

Example: If $M(x,y) = y^2$, then

$$M(x,y) = y \cdot y$$

and thus

$$\hat{M}(a,b) = y_a \cdot y_b$$

Thus $\hat{M}(a,b) = y_a y_b$, which corresponds to the definition of $\hat{M}(a,b)$ in (3.2).

Example: If $M(x,y) = xy$, then

$$M(x,y) = x \cdot y \cdot 1$$

and thus

$$\hat{M}(a,b,c) = x_a \cdot y_b \cdot 1$$

Thus $\hat{M}(a,b,c) = x_a y_b$.

Example: If $M(x,y) = x^2 y$, then

$$M(x,y) = x \cdot x \cdot y$$

and thus

$$\hat{M}(a,b,c) = x_a \cdot x_b \cdot y_c$$

Thus $\hat{M}(a,b,c) = x_a x_b y_c$.

We then prove a more general version of Theorem 3.1.2:

Theorem 4.1.1. *The set of functions of the form $\mu_{a_1 \cdots a_q}$, with $\mu_{a_1 \cdots a_q} = \prod_{i=1}^{q} \lambda_{a_i}$, is such that*

$$\sum_{a_1=1}^{n} \cdots \sum_{a_q=1}^{n} \mu_{a_1 \cdots a_q}(x,y) \hat{M}(a_1, \cdots, a_q) = M(x,y) \;\; \forall (x,y) \in \overline{P}$$

for any monomial $M = M(x,y)$ of degree up to q.

Proof. By induction, let us assume the theorem holds for the set of functions $\mu_{a_1 \cdots a_{q-1}}$ with monomials of degree up to $q-1$. (The base case of the induction argument is given by (3.1)).

Let $M = M(x,y)$ be a monomial of degree at most q. Thus, $M(x,y) = x^k y^l$, $0 \leq k+l \leq q$.

We now write M as

$$M(x,y) = M_1(x,y) M_2(x,y)$$

with $M_1(x,y)$ and $M_2(x,y)$ defined as the following:

- If $k+l<q$, then $M_1(x,y)=x^k y^l$ and $M_2(x,y)=1$;
- If $k+l=q$ and $l\neq 0$, then $M_1(x,y)=x^k y^{l-1}$ and $M_2(x,y)=y$;
- If $k=q$ and $l=0$, i.e. $M(x,y)=x^k$, then $M_1(x,y)=x^{k-1}$ and $M_2(x,y)=x$.

By this definition, M_1 is a monomial of degree at most $q-1$ and M_2 is a monomial of degree at most 1. This definition also implies

$$\hat{M}(a_1,\cdots,a_q)=\hat{M}_1(a_1,\cdots,a_{q-1})\hat{M}_2(a_q)$$

for any ordered sequence $(a_1,\cdots,a_q)$ such that $1\leq a_i\leq n,\ i=1,\cdots,q$. Indeed:

- If $k+l<q$, then

$$\hat{M}(a_1,\cdots,a_q)=x_{a_1}\cdots x_{a_k}y_{a_{k+1}}\cdots y_{a_{k+l}}=\hat{M}_1(a_1,\cdots,a_{q-1})=\hat{M}_1(a_1,\cdots,a_{q-1})\hat{M}_2(a_q)$$

- If $k+l=q$ and $l\neq 0$, then

$$\hat{M}(a_1,\cdots,a_q)=x_{a_1}\cdots x_{a_k}y_{a_{k+1}}\cdots y_{a_{q-1}}y_{a_q}=\hat{M}_1(a_1,\cdots,a_{q-1})\hat{M}_2(a_q)$$

- If $k=q$ and $l=0$, then

$$\hat{M}(a_1,\cdots,a_q)=x_{a_1}\cdots x_{a_{q-1}}x_{a_q}=\hat{M}_1(a_1,\cdots,a_{q-1})\hat{M}_2(a_q)$$

Therefore, it holds that

$$\begin{aligned}
&\sum_{a_1=1}^{n}\cdots\sum_{a_q=1}^{n}\hat{M}(a_1,\cdots,a_q)\mu_{a_1\cdots a_q}(x,y)\\
&=\sum_{a_1=1}^{n}\cdots\sum_{a_q=1}^{n}\hat{M}_1(a_1,\cdots,a_{q-1})\hat{M}_2(a_q)\mu_{a_1\cdots a_{q-1}}(x,y)\lambda_{a_q}(x,y)\\
&=\left(\sum_{a_1=1}^{n}\cdots\sum_{a_{q-1}=1}^{n}\hat{M}_1(a_1,\cdots,a_{q-1})\mu_{a_1\cdots a_{q-1}}(x,y)\right)\left(\sum_{a_q=1}^{n}\hat{M}_2(a_q)\lambda_{a_q}(x,y)\right)\\
&=M_1(x,y)M_2(x,y)=M(x,y) \qquad (\star\star\star)
\end{aligned}$$

where $(\star\star\star)$ is due to the application of the induction hypothesis and (3.1). □

Example: Let us try to prove Theorem 4.1.1 for $q=3$ and $M(x,y)=y^2$.

Then, $M=M_1M_2$ with $M_1(x,y)=y^2$ and $M_2(x,y)=1$. Furthermore,

$$\hat{M}(a,b,c)=y_a y_b;\ \ \hat{M}_1(a,b)=y_a y_b;\ \ \hat{M}_2(c)=1$$

and thus $\hat{M}(a,b,c)=\hat{M}_1(a,b)\hat{M}_2(c)$.

We know from the induction hypothesis that

$$\sum_{a=1}^{n}\sum_{b=1}^{n}y_a y_b\mu_{ab}(x,y)=y^2;\ \ \sum_{c=1}^{n}\lambda_c(x,y)=1$$

Thus, we prove the theorem by showing the following:

$$\begin{aligned}
&\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n} y_a y_b \mu_{abc}(x,y) \\
&= \sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n} y_a y_b \mu_{ab}(x,y)\lambda_c(x,y) \\
&= \left(\sum_{a=1}^{n}\sum_{b=1}^{n} y_a y_b \mu_{ab}(x,y)\right)\left(\sum_{c=1}^{n}\lambda_c(x,y)\right) = y^2 \cdot 1 = y^2
\end{aligned}$$

Example: Let us try to prove Theorem 4.1.1 for $q = 3$ and $M(x,y) = x^2 y$.

Then, $M = M_1 M_2$ with $M_1(x,y) = x^2$ and $M_2(x,y) = y$. Furthermore,

$$\hat{M}(a,b,c) = x_a x_b y_c; \quad \hat{M}_1(a,b) = x_a x_b; \quad \hat{M}_2(c) = y_c$$

and thus $\hat{M}(a,b,c) = \hat{M}_1(a,b)\hat{M}_2(c)$.

We know from the induction hypothesis that

$$\sum_{a=1}^{n}\sum_{b=1}^{n} x_a x_b \mu_{ab}(x,y) = x^2; \quad \sum_{c=1}^{n} y_c \lambda_c(x,y) = y$$

Thus, we prove the theorem by showing the following:

$$\begin{aligned}
&\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n} x_a x_b y_c \mu_{abc}(x,y) \\
&= \sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n} x_a x_b y_c \mu_{ab}(x,y)\lambda_c(x,y) \\
&= \left(\sum_{a=1}^{n}\sum_{b=1}^{n} x_a x_b \mu_{ab}(x,y)\right)\left(\sum_{c=1}^{n} y_c \lambda_c(x,y)\right) = x^2 \cdot y = x^2 y
\end{aligned}$$

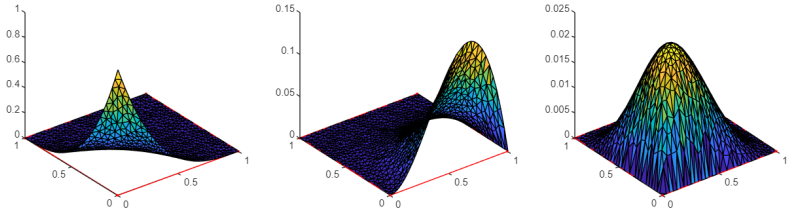
By Remark 3.1.1, it therefore holds that

Corollary 4.1.2. *The set of functions of the form $\mu_{a_1 \cdots a_q}$, with $\mu_{a_1 \cdots a_q} = \prod_{i=1}^{q} \lambda_{a_i}$, can reproduce polynomials of degree up to q.*

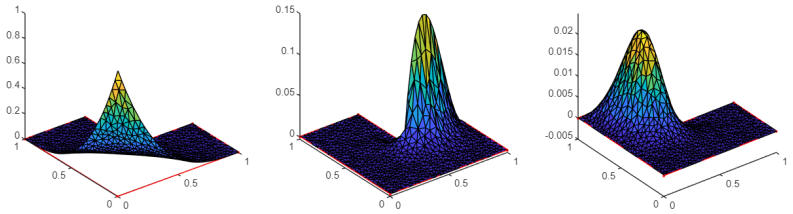
We call these functions the *q-th coordinates* over $\overline{P}$ (defined with respect to mean value coordinates). Figure 4.1 shows a few examples of q-th coordinates for $q = 3, 4$. q-th coordinates are the generalized version of the quadratic coordinates seen in the previous section. We will use them in order to define *q-th serendipity coordinates* through a similar generalized approach.

4.2 Rewriting the polynomial reproducibility property

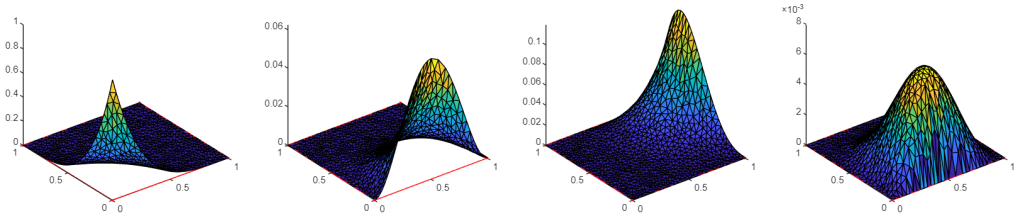
Theorem 4.1.1 and Corollary 4.1.2 have shown us that q-th coordinates can reproduce polynomials of degree up to q. Therefore, our goal is to simply reduce this space to a space of q-th



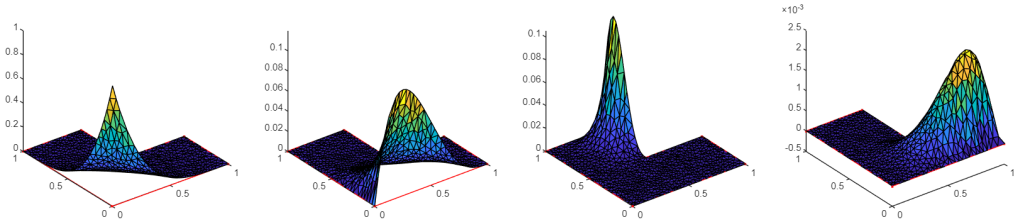
(a) μ_{111}, μ_{122}, μ_{113} for $q = 3$ over P_1.



(b) μ_{111}, μ_{344}, μ_{456} for $q = 3$ over P_2.



(c) μ_{1111}, μ_{1122}, μ_{2333}, μ_{1123} for $q = 4$ over P_1.



(d) μ_{1111}, μ_{1112}, μ_{4555}, μ_{2236} for $q = 4$ over P_2.

Figure 4.1. Surf plot of q-th coordinates for variable value of q over either the quadrilateral P_1 or the hexagon P_2, with P_1 having vertices $\boldsymbol{\nu}_1 = (0,0), \boldsymbol{\nu}_2 = (1,0), \boldsymbol{\nu}_3 = (1,1), \boldsymbol{\nu}_4 = (0,1)$ and P_2 having vertices $\boldsymbol{\nu}_1 = (0,0), \boldsymbol{\nu}_2 = (1,0), \boldsymbol{\nu}_3 = (1,\frac{1}{2}), \boldsymbol{\nu}_4 = (\frac{1}{2},\frac{1}{2}), \boldsymbol{\nu}_5 = (\frac{1}{2},1), \boldsymbol{\nu}_6 = (0,1)$.

serendipity coordinates that share the same property. When working with quadratic coordinates, we rewrote the result of Theorem 3.1.2 as (3.4), in order to avoid stacked sums: we would like to do something similar here as well.

We first generalize the notation in (3.3): we write

$$a_1 \cdots a_q$$

to indicate the unordered list made up by the indices in $(a_1, \cdots, a_q)$. We will use this notation all throughout the rest of the thesis.

Then, given an unordered list of indices $a_1 \cdots a_1$, we define

$$\mathscr{P}(a_1 \cdots a_q)$$

as the set of all ordered sequences made up by the q indices in $a_1 \cdots a_q$. We note a generic element in $\mathscr{P}(a_1 \cdots a_q)$ as $(a_{p_1}, \cdots, a_{p_q})$, to indicate that they are the same indices but arranged in a specific order.

Example: Here are some examples of $\mathscr{P}(a_1 \cdots a_q)$ for different unordered lists $a_1 \cdots a_q$:

- $\mathscr{P}(123) = \{(1,2,3);(1,3,2);(2,1,3);(2,3,1);(3,1,2);(3,2,1)\}$;
- $\mathscr{P}(112) = \{(1,1,2);(1,2,1);(2,1,1)\}$;
- $\mathscr{P}(35) = \{(3,5);(5,3)\}$;
- $\mathscr{P}(2244) = \{(2,2,4,4);(2,4,2,4);(4,2,2,4);(2,4,4,2);(4,2,4,2);(4,4,2,2)\}$;

We use this notation to state the following result:

Lemma 4.2.1. *The set of q-th coordinates $\mu_{a_1 \cdots a_q}$ (defined with respect to the mean value coordinates λ_a) is such that*

$$\sum_{a_1 \cdots a_q} \mu_{a_1 \cdots a_q}(x,y) \left(\sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \hat{M}(a_{p_1}, \cdots, a_{p_q}) \right) = M(x,y) \quad \forall (x,y) \in \overline{P}$$

for any monomial $M = M(x,y)$ of degree up to q.

Proof. The proof simply involves writing the sum in Theorem 4.1.1 and grouping together addends using the same sets of indices:

$$\begin{aligned} M(x,y) &= \sum_{a_1=1}^{n} \cdots \sum_{a_q=1}^{n} \mu_{a_1 \cdots a_q}(x,y) \hat{M}(a_1, \cdots, a_q) \\ &= \sum_{a_1 \cdots a_q} \sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \mu_{a_{p_1} \cdots a_{p_q}}(x,y) \hat{M}(a_{p_1}, \cdots, a_{p_q}) \\ &= \sum_{a_1 \cdots a_q} \mu_{a_1 \cdots a_q}(x,y) \left(\sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \hat{M}(a_{p_1}, \cdots, a_{p_q}) \right) \end{aligned}$$

□

Example: Let us consider the case $q = 3$ explicitly. We know from Theorem 4.1.1 that

$$\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n}\hat{M}(a,b,c)\mu_{abc}(x,y) = M(x,y) \quad \forall (x,y)\in\overline{P}$$

for all monomials $M = M(x,y)$ of degree up to 3.

Let us fix one such monomial M. For the sake of this example, we call $\hat{M}(a,b,c)$ the *associated coefficient* to the function μ_{abc}.

Any function can appear a different number of times depending on how many repeating digits it has:

- If it is a function of the type $\mu_{\hat{a}\hat{a}\hat{a}}$, then it will only be counted once, for $(a,b,c) = (\hat{a},\hat{a},\hat{a})$, with associated coefficient $\hat{M}(\hat{a},\hat{a},\hat{a})$;
- If it is a function of the type $\mu_{\hat{a}\hat{a}\hat{b}}, \hat{a}\neq\hat{b}$ then it will be counted three times:
  - once for $(a,b,c) = (\hat{a},\hat{a},\hat{b})$, with associated coefficient $\hat{M}(\hat{a},\hat{a},\hat{b})$;
  - once for $(a,b,c) = (\hat{a},\hat{b},\hat{a})$, with associated coefficient $\hat{M}(\hat{a},\hat{b},\hat{a})$;
  - once for $(a,b,c) = (\hat{b},\hat{a},\hat{a})$, with associated coefficient $\hat{M}(\hat{b},\hat{a},\hat{a})$;
- If it is a function of the type $\mu_{\hat{a}\hat{b}\hat{c}}, \hat{a}\neq\hat{b}\neq\hat{c}$ then it will be counted six times:
  - once for $(a,b,c) = (\hat{a},\hat{b},\hat{c})$, with associated coefficient $\hat{M}(\hat{a},\hat{b},\hat{c})$;
  - once for $(a,b,c) = (\hat{a},\hat{c},\hat{b})$, with associated coefficient $\hat{M}(\hat{a},\hat{c},\hat{b})$;
  - once for $(a,b,c) = (\hat{b},\hat{a},\hat{c})$, with associated coefficient $\hat{M}(\hat{b},\hat{a},\hat{c})$;
  - once for $(a,b,c) = (\hat{b},\hat{c},\hat{a})$, with associated coefficient $\hat{M}(\hat{b},\hat{c},\hat{a})$;
  - once for $(a,b,c) = (\hat{c},\hat{a},\hat{b})$, with associated coefficient $\hat{M}(\hat{c},\hat{a},\hat{b})$;
  - once for $(a,b,c) = (\hat{c},\hat{b},\hat{a})$, with associated coefficient $\hat{M}(\hat{c},\hat{b},\hat{a})$.

Thus, the equation can be rewritten as

$$\sum_{aaa}\hat{M}(a,a,a)\mu_{aaa} + \sum_{\substack{aab\\ a\neq b}}\big(\hat{M}(a,a,b)+\hat{M}(a,b,a)+\hat{M}(b,a,a)\big)\mu_{aab} +$$
$$+\sum_{\substack{abc\\ a\neq b\neq c}}\big(\hat{M}(a,b,c)+\hat{M}(a,c,b)+\hat{M}(b,a,c)+\hat{M}(b,c,a)+\hat{M}(c,a,b)+\hat{M}(c,b,a)\big)\mu_{abc} =$$
$$= M(x,y) \quad \forall (x,y)\in\overline{P}$$

For example, if $M(x,y) = xy$ (and thus $\hat{M}(a,b,c) = x_a y_b$), then the equation is

$$\sum_{aaa}x_a y_a\mu_{aaa} + \sum_{\substack{aab\\ a\neq b}}(x_a y_a + x_a y_b + x_b y_a)\mu_{aab} +$$
$$+\sum_{\substack{abc\\ a\neq b\neq c}}(x_a y_b + x_a y_c + x_b y_a + x_b y_c + x_c y_a + x_c y_b)\mu_{abc} = xy \quad \forall (x,y)\in\overline{P}$$

On the other hand, if $M(x,y) = x^2 y$ (and thus $\hat{M}(a,b,c) = x_a x_b y_c$), then the equation is

$$\sum_{aaa}(x_a)^2 y_a \mu_{aaa} + \sum_{\substack{aab \\ a\neq b}}\left((x_a)^2 y_b + 2x_a x_b y_a\right)\mu_{aab} + \\ + \sum_{\substack{abc \\ a\neq b\neq c}}(2x_a x_b y_c + 2x_a x_c y_b + 2x_b x_c y_a)\,\mu_{abc} = x^2 y \quad \forall (x,y)\in\overline{P}$$

Either way, the functions are grouped together based on how many times they appear in the stacked sum; which is directly determined by the number of ordered sequences associated to each function's respective sequence of indices. Indeed,

$$\mathscr{P}(aaa) = \{(a,a,a)\} \quad \forall aaa \in \mathscr{A}$$
$$\mathscr{P}(aab) = \{(a,a,b);(a,b,a);(b,a,a)\} \quad \forall aab \in \mathscr{A}, a\neq b$$
$$\mathscr{P}(abc) = \{(a,b,c);(a,c,b);(b,a,c);(b,c,a);(c,a,b);(c,b,a)\} \quad \forall abc \in \mathscr{A}, a\neq b\neq c$$

and thus

$$\sum_{aaa}\hat{M}(a,a,a)\mu_{aaa} + \sum_{\substack{aab \\ a\neq b}}\left(\hat{M}(a,a,b)+\hat{M}(a,b,a)+\hat{M}(b,a,a)\right)\mu_{aab} + \\ + \sum_{\substack{abc \\ a\neq b\neq c}}\left(\hat{M}(a,b,c)+\hat{M}(a,c,b)+\hat{M}(b,a,c)+\hat{M}(b,c,a)+\hat{M}(c,a,b)+\hat{M}(c,b,a)\right)\mu_{abc} =$$
$$= \sum_{aaa}\left(\sum_{(a_1,a_2,a_3)\in\mathscr{P}(aaa)}\hat{M}(a_1,a_2,a_3)\right)\mu_{aaa} + \sum_{\substack{aab \\ a\neq b}}\left(\sum_{(a_1,a_2,a_3)\in\mathscr{P}(aab)}\hat{M}(a_1,a_2,a_3)\right)\mu_{aab} + \\ + \sum_{\substack{abc \\ a\neq b\neq c}}\left(\sum_{(a_1,a_2,a_3)\in\mathscr{P}(abc)}\hat{M}(a_1,a_2,a_3)\right)\mu_{abc}$$
$$= \sum_{abc}\left(\sum_{(a_1,a_2,a_3)\in\mathscr{P}(abc)}\hat{M}(a_1,a_2,a_3)\right)\mu_{abc}$$

We prove one further result involving $\mathscr{P}(a_1\cdots a_q)$:

Lemma 4.2.2. *Let*

$$(\underbrace{a_1,\cdots,a_1}_{s_1},\cdots,\underbrace{a_r,\cdots,a_r}_{s_r})$$

be an ordered sequence of q indices, such that $a_1,\cdots,a_r$ are pairwise distinct and appear in the sequence $s_1,\cdots,s_r$ times, respectively. Furthermore, let $\mathscr{S}_q$ be the set of all permutations over q indices. Then, for every $(a_{p_1},\cdots,a_{p_q}) \in \mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$, the number of permutations $p\in\mathscr{S}_q$ such that $p(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = (a_{p_1},\cdots,a_{p_q})$ is exactly $\frac{q!}{|\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)|}$ (where $|\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)|$ is the cardinality of $\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$).

Proof. By definition, any $(a_{p_1},\cdots,a_{p_q}) \in \mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$ is made up of the same indices as $(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$ and thus can be obtained by permuting it:

$$\forall (a_{p_1},\cdots,a_{p_q}) \in \mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)\,\exists \tilde{p} \in \mathscr{S}_q\, s.t.\, (a_{p_1},\cdots,a_{p_q}) = \tilde{p}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$$

We choose one such $\tilde{p}$ for every $(a_{p_1},\cdots,a_{p_q}) \in \mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$ and construct a set $\mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$ of permutations. This means there is a one-to-one correspondence between elements of $\mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$ and $\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$, which also implies their cardinalities are the same:

$$\forall (a_{p_1},\cdots,a_{p_q}) \in \mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)\ \exists! p \in \mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)\, s.t. \quad (a_{p_1},\cdots,a_{p_q}) = p(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) \tag{4.2a}$$

$$|\mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)| = |\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)| = \frac{q!}{s_1!\cdots s_r!} \tag{4.2b}$$

The last equivalence in (4.2b) is a combinatorics result: $\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$ is the set of unique sequences made up by the indices in $a_1\cdots a_1\cdots a_r\cdots a_r$; and the number of said sequences is a multinomial coefficient.

Every $\tilde{p} \in \mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$ permutes the elements in the r groups of $s_1,\cdots,s_r$ indices with respect to the other groups of elements, but not with respect to the other elements in each group. A permutation $p \in \mathscr{S}_q$ does both. Because of this, any permutation $p \in \mathscr{S}_q$ can be written as

$$p(b_1,\cdots,b_q) = \tilde{p}(p_1(b_1,\cdots,b_{s_1}),\cdots,p_r(b_{q-s_r+1},\cdots,b_q));$$
$$p_1 \in \mathscr{S}_{s_1},\cdots,p_r \in \mathscr{S}_{s_r},\ \tilde{p} \in \mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$$

where, of course, $\mathscr{S}_{s_1},\cdots,\mathscr{S}_{s_r}$ are the sets of all permutations over $s_1,\cdots,s_r$ indices, respectively. We condense this form by writing $p = \tilde{p}(p_1,\cdots,p_r)$.

Now, let $(a_{\hat{p}_1},\cdots,a_{\hat{p}_r})$ be a specific ordered sequence in $\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)$. Let $\hat{p} \in \mathscr{S}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r)$ be the permutation such that $\hat{p}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = (a_{\hat{p}_1},\cdots,a_{\hat{p}_r})$. Then, for $p \in \mathscr{S}_q$, it holds that

$$\begin{aligned}
& p(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = (a_{\hat{p}_1},\cdots,a_{\hat{p}_r}) \\
\iff\ & p(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = \hat{p}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) \\
\iff\ & \tilde{p}(p_1(a_1,\cdots,a_1),\cdots,p_r(a_r,\cdots,a_r)) = \hat{p}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) \\
\iff\ & \tilde{p}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = \hat{p}(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) \\
\iff\ & \tilde{p} = \hat{p} \qquad (\star)
\end{aligned}$$

where ($\star$) is by (4.2a). Thus, the permutations $p \in \mathscr{S}_q$ such that $p(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = (a_{\hat{p}_1},\cdots,a_{\hat{p}_r})$ are exactly the permutations of the form

$$p = \hat{p}(p_1,\cdots,p_r) \tag{4.2c}$$

for some $p_1 \in \mathscr{S}_{s_1},\cdots,p_r \in \mathscr{S}_{s_r}$. By definition, there are $s_1!$ distinct permutations in $\mathscr{S}_{s_1}$, $s_2!$ distinct permutations in $\mathscr{S}_{s_2}$ and so on until $\mathscr{S}_{s_r}$. Therefore, there are exactly

$$s_1!\cdots s_r! = \frac{q!}{\frac{q!}{s_1!\cdots s_r!}} = \frac{q!}{|\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)|}$$

permutations of the form in (4.2c), where the last equivalence comes from (4.2b). Thus, there are exactly $\frac{q!}{|\mathscr{P}(a_1\cdots a_1\cdots a_r\cdots a_r)|}$ permutations $p \in \mathscr{S}_q$ such that $p(a_1,\cdots,a_1,\cdots,a_r,\cdots,a_r) = (a_{\hat{p}_1},\cdots,a_{\hat{p}_r})$. □

Example: Let us consider $q = 3$. $\mathscr{S}_3$ is a set of six permutations $p_1, \cdots, p_6$, such that

$$p_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}; \quad p_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}; \quad p_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix};$$
$$p_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}; \quad p_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}; \quad p_6 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

If we consider the sequence $(1,1,2)$, then

$$\mathscr{P}(112) = \{(1,1,2); (1,2,1); (2,1,1)\}$$

and

$$p_1(1,1,2) = (1,1,2); \quad p_2(1,1,2) = (1,2,1); \quad p_3(1,1,2) = (1,1,2);$$
$$p_4(1,1,2) = (1,2,1); \quad p_5(1,1,2) = (2,1,1); \quad p_6(1,1,2) = (2,1,1)$$

Of course, applying any permutation to $(1,1,2)$ can only yield an element of $\mathscr{P}(112)$ as a result. But, more specifically, we note how each of the three elements of $\mathscr{P}(112)$ is associated to exactly two permutations of $(1,1,2)$, making up the total of the six permutations in $\mathscr{S}_3$. In practice, this is because, since the first two elements of $(1,1,2)$ are the same, the result of applying a permutation to $(1,1,2)$ is only determined by the final position of the third element.

More specifically, we can follow the lead of Lemma 4.2.2 on this restricted case. We construct $\mathscr{S}(1,1,2)$ as explained in the theorem: for the purpose of this example, we pick

$$\mathscr{S}(1,1,2) = \{p_1; p_2; p_5\}$$

As per definition, $p_1(1,1,2) = (1,1,2)$, $p_2(1,1,2) = (1,2,1)$ and $p_5(1,1,2) = (2,1,1)$.

Now we note that, borrowing notation from the proof of Lemma 4.2.2, it holds that

$$p_1 = p_1(p_a, p_I); \quad p_2 = p_2(p_a, p_I); \quad p_3 = p_1(p_b, p_I);$$
$$p_4 = p_2(p_b, p_I); \quad p_5 = p_5(p_a, p_I); \quad p_6 = p_5(p_b, p_I)$$

where

$$\mathscr{S}_2 = \left\{ p_a = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}; p_b = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \right\}; \quad \mathscr{S}_1 = \left\{ p_I = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$

For example, p_1 is such that

$$p_1(1,2,3) = p_1(p_a(1,2), p_I(3)) = p_1(1,2,3) = (1,2,3)$$
$$\implies p_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$

On the other hand, p_4 is such that

$$p_4(1,2,3) = p_2(p_b(1,2), p_I(3)) = p_2(2,1,3) = (2,3,1)$$
$$\implies p_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

In more simple terms, any element of $\mathscr{S}_3$ can be generated by first permuting the first two indices appropriately (and leaving the third index unchanged) and then applying a permutation in $\mathscr{S}(1,1,2)$.

By construction, this means that every element in $\mathscr{S}(1,1,2)$ can generate $2!\cdot 1! = 2$ elements in $\mathscr{S}_3$. Also by construction, the result of applying a permutation in $\mathscr{S}_3$ to $(1,1,2)$ depends uniquely on the permutation in $\mathscr{S}(1,1,2)$ that was used to define it through this process. Overall, this means that, for every element of $\mathscr{P}(112)$, there are 2 permutations in $\mathscr{S}_3$ yielding that element of $\mathscr{P}(112)$ when applied to $(1,1,2)$.

Next, we introduce another notation. Given an unordered list of indices $a_1\cdots a_q$ and a monomial $M = M(x,y)$ of degree up to q, we define $S_M(a_1\cdots a_q)$ as follows:

$$S_M(a_1\cdots a_q) = \frac{|\mathscr{P}(a_1\cdots a_q)|}{q!}\sum_{p\in\mathscr{S}_q}\hat{M}(p(a_1,\cdots,a_q))$$

Here, $\mathscr{S}_q$ and $|\mathscr{P}(a_1\cdots a_q)|$ are defined as in Lemma 4.2.2: $\mathscr{S}_q$ is the set of all permutations over q indices; and $|\mathscr{P}(a_1\cdots a_q)|$ is the cardinality of $\mathscr{P}(a_1\cdots a_q)$, i.e. the number of unique ordered lists made up by the indices in $a_1\cdots a_q$. $(a_1,\cdots,a_q)$ is an arbitrary element of $\mathscr{P}(a_1\cdots a_q)$, i.e. an ordered list made up by the indices in $a_1\cdots a_q$.

Example: Let us define $S_M(112)$ for some $M = M(x,y)$ of degree at most 3.

As per definition,

$$S_M(a_1\cdots a_q) = \frac{|\mathscr{P}(a_1\cdots a_q)|}{q!}\sum_{p\in\mathscr{S}_q}\hat{M}(p(a_1,\cdots,a_q))$$

Because $(a_1,\cdots,a_q)$ is an arbitrary element of $\mathscr{P}(a_1\cdots a_q)$, the definition of $S_M(112)$ involves picking an arbitrary element of $\mathscr{P}(112)$. (We do not yet know that $S_M(112)$ is well-defined with this caveat, but for the moment, we assume it is.) As we have seen in the previous example,

$$\mathscr{P}(112) = \{(1,1,2);(1,2,1);(2,1,1)\}$$

Thus, for the purposes of this example, let us pick $(1,1,2)$ as the arbitrary element. Then,

$$\begin{aligned} S_M(112) &= \frac{|\mathscr{P}(112)|}{3!}\sum_{p\in\mathscr{S}_3}\hat{M}(p(1,1,2)) \\ &= \frac{3}{6}\sum_{p\in\mathscr{S}_3}\hat{M}(p(1,1,2)) = \frac{1}{2}\sum_{p\in\mathscr{S}_3}\hat{M}(p(1,1,2)) \end{aligned}$$

In the previous example, we have also stated that

$$\mathscr{S}_3 = \{p_1;p_2;p_3;p_4;p_5;p_6\}$$

with

$$p_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix};\ p_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix};\ p_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix};$$
$$p_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix};\ p_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix};\ p_6 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

and thus

$$p_1(1,1,2) = (1,1,2);\ \ p_2(1,1,2) = (1,2,1);\ \ p_3(1,1,2) = (1,1,2);$$
$$p_4(1,1,2) = (1,2,1);\ \ p_5(1,1,2) = (2,1,1);\ \ p_6(1,1,2) = (2,1,1)$$

This means

$$\begin{aligned} &\sum_{p\in\mathscr{S}_3} \hat{M}(p(1,1,2)) \\ &= \hat{M}(p_1(1,1,2)) + \hat{M}(p_2(1,1,2)) + \hat{M}(p_3(1,1,2)) + \hat{M}(p_4(1,1,2)) + \\ &\qquad + \hat{M}(p_5(1,1,2)) + \hat{M}(p_6(1,1,2)) \\ &= \hat{M}(1,1,2) + \hat{M}(1,2,1) + \hat{M}(1,1,2) + \hat{M}(1,2,1) + \hat{M}(2,1,1) + \hat{M}(2,1,1) \\ &= 2\hat{M}(1,1,2) + 2\hat{M}(1,2,1) + 2\hat{M}(2,1,1) \end{aligned}$$

and thus

$$\begin{aligned} S_M(112) &= \frac{1}{2}\sum_{p\in\mathscr{S}_3} \hat{M}(p(1,1,2)) \\ &= \frac{1}{2}\left(2\hat{M}(1,1,2) + 2\hat{M}(1,2,1) + 2\hat{M}(2,1,1)\right) \\ &= \hat{M}(1,1,2) + \hat{M}(1,2,1) + \hat{M}(2,1,1) \end{aligned}$$

For example, if $M(x,y) = xy$, then $\hat{M}(a,b,c) = x_a y_b$ and thus

$$S_M(112) = x_1y_1 + x_1y_2 + x_2y_1$$

If instead $M(x,y) = x^2y$, then $\hat{M}(a,b,c) = x_a x_b y_c$ and thus

$$S_M(112) = x_1x_1y_2 + x_1x_2y_1 + x_2x_1y_1 = (x_1)^2y_2 + 2x_1x_2y_1$$

It is worth noting that, in this case, we find that $S_M(112)$ is exactly equal to the sum of $\hat{M}$ applied to all the ordered sequences in $\mathscr{P}(112)$:

$$S_M(112) = \sum_{(a,b,c)\in\mathscr{P}(112)} \hat{M}(a,b,c)$$

We will prove that this is not a coincidence and that this property holds in general for $S_M(a_1\cdots a_q)$.

Because the definition of $S_M(a_1\cdots a_q)$ involves taking an arbitrary ordering of the indices in $a_1\cdots a_q$, we must ensure this definition can be taken without any issue. Thus, we note the following:

Remark 4.2.3. *Let $M = M(x,y)$ be a monomial of degree at most q. Then, $S_M(a_1 \cdots a_q)$ is well-defined; i.e., it holds that*

$$\frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{p \in \mathscr{S}_q} \hat{M}(p(a_1, \cdots, a_q)) = \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{p \in \mathscr{S}_q} \hat{M}(p(a_{p_1}, \cdots, a_{p_q}))$$

for any $(a_1, \cdots, a_q), (a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)$.

Proof. Because $(a_1, \cdots, a_q)$ and $(a_{p_1}, \cdots, a_{p_q})$ are lists made up of the same indices, one can be obtained by permuting the other:

$$\exists \tilde{p} \in \mathscr{S}_q \; s.t. \; (a_{p_1}, \cdots, a_{p_q}) = \tilde{p}(a_1, \cdots, a_q) \tag{4.3a}$$

Let us now consider the set $\tilde{p}(\mathscr{S}_q)$. If $p_1, p_2 \in \mathscr{S}_q$, then $\tilde{p} \circ p_1, \tilde{p} \circ p_2 \in \mathscr{S}_q$: this is because $\tilde{p}$, p_1 and p_2 are permutations; and since permutations are essentially bijective functions, the composition of permutations is a permutation. $\tilde{p}$'s bijectivity also implies that, if $\tilde{p} \circ p_1 = \tilde{p} \circ p_2$, then $p_1 = p_2$.

Thus, because $\mathscr{S}_q$ is a set of $q!$ distinct permutations, it holds that $\tilde{p}(\mathscr{S}_q)$ is also a set of $q!$ distinct permutations. By the pigeonhole principle, this means $\tilde{p}(\mathscr{S}_q) = \mathscr{S}_q$, i.e.

$$\forall p \in \mathscr{S}_q \; \exists! \hat{p} \in \mathscr{S}_q \; s.t. \; \hat{p} = \tilde{p} \circ p \tag{4.3b}$$

Therefore,

$$\begin{aligned} \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{p \in \mathscr{S}_q} \hat{M}(p(a_{p_1}, \cdots, a_{p_q})) &= \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{p \in \mathscr{S}_q} \hat{M}(p(\tilde{p}(a_1, \cdots, a_q))) && (*) \\ &= \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{\hat{p} \in \mathscr{S}_q} \hat{M}(\hat{p}(a_1, \cdots, a_q))) && (**) \end{aligned}$$

where $(*)$ is due to (4.3a) and $(**)$ is due to (4.3b). □

The generalization of the construction of serendipity coordinates then hinges on this result:

Theorem 4.2.4. *The set of q-th coordinates $\mu_{a_1 \cdots a_q}$ (defined with respect to the mean value coordinates λ_a) is such that*

$$\sum_{a_1 \cdots a_q} \mu_{a_1 \cdots a_q}(x,y) S_M(a_1 \cdots a_q) = M(x,y) \;\; \forall (x,y) \in \overline{P}$$

for any monomial $M = M(x,y)$ of degree up to q.

Proof. Let $M = M(x,y)$ be a monomial of degree at most q. We know from Lemma 4.2.1 that

$$\sum_{a_1 \cdots a_q} \mu_{a_1 \cdots a_q}(x,y) \left(\sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \hat{M}(a_{p_1}, \cdots, a_{p_q}) \right) = M(x,y) \;\; \forall (x,y) \in \overline{P}$$

Therefore, proving this theorem can be achieved by proving that

$$S_M(a_1 \cdots a_q) = \sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \hat{M}(a_{p_1}, \cdots, a_{p_q}) \;\; \forall a_1 \cdots a_q$$

Let $a_1 \cdots a_q$ be an unordered sequence of indices. We rewrite the sequence by grouping together equal indices: thus, we write

$$\underbrace{a_1 \cdots a_1}_{s_1} \cdots \underbrace{a_r \cdots a_r}_{s_r}$$

with $a_1, \cdots, a_r$ pairwise distinct and appearing $s_1, \cdots, s_r$ times in the sequence, respectively $(s_1, \cdots, s_r \geq 1)$.

Let us now consider the ordered list $(a_1, \cdots, a_1, \cdots, a_r, \cdots, a_r)$, such that all equal indices are grouped together. From Remark 4.2.3, we know we can define $S(a_1 \cdots a_q)$ like so:

$$S_M(a_1 \cdots a_q) = \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{p \in \mathscr{S}_q} \hat{M}(p(a_1, \cdots, a_1, \cdots, a_r, \cdots, a_r))$$

Then, the proof is a simple matter of grouping together permutations that yield the same sequence of indices:

$$\begin{aligned}
S_M(a_1 \cdots a_q) &= \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{p \in \mathscr{S}_q} \hat{M}(p(a_1, \cdots, a_1, \cdots, a_r, \cdots, a_r)) \\
&= \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \sum_{\substack{p \in \mathscr{S}_q \\ p(a_1, \cdots, a_1, \cdots, a_r, \cdots, a_r) = (a_{p_1}, \cdots, a_{p_q})}} \hat{M}(p(a_1, \cdots, a_1, \cdots, a_r, \cdots, a_r)) \\
&= \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \sum_{\substack{p \in \mathscr{S}_q \\ p(a_1, \cdots, a_1, \cdots, a_r, \cdots, a_r) = (a_{p_1}, \cdots, a_{p_q})}} \hat{M}(a_{p_1}, \cdots, a_{p_q}) \\
&= \frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!} \sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \frac{q!}{|\mathscr{P}(a_1 \cdots a_q)|} \hat{M}(a_{p_1}, \cdots, a_{p_q}) \qquad (\diamond) \\
&= \sum_{(a_{p_1}, \cdots, a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \hat{M}(a_{p_1}, \cdots, a_{p_q})
\end{aligned}$$

where $(\diamond)$ is by the result in Lemma 4.2.2.

□

This result serves as a generalization of the equations in (3.4), expressing the reproducibility of monomials for quadratic coordinates: we now have a similar closed form expression of equations for serendipity coordinates of any order. The specific definition of $S_M(a_1 \cdots a_q)$ also makes this result easy to implement: the permutations in $\mathscr{S}_q$ can be applied to any ordered sequence indiscriminately; and $|\mathscr{P}(a_1 \cdots a_q)|$ can be found either by counting or by the second equivalence in (4.2b).

This newfound generalization can be used to extend the results regarding serendipity coordinates as well.

4.3 Serendipity coordinates of any order

Let $\mathscr{A}$ be the set of all unordered sequences of length q made up of elements between 1 and n; and let $\mathscr{I} \subseteq \mathscr{A}$. We would like to reduce the set of q-th coordinates $\mu_{a_1 \cdots a_q}$, $a_1 \cdots a_q \in \mathscr{A}$ to a

set of q-th serendipity coordinates $\xi_{i_1\cdots i_q}$, $i_1\cdots i_q \in \mathscr{I}$ such that they satisfy an equivalent result to that in Theorem 4.2.4: given a monomial $M = M(x,y)$ of degree at most q, we ask that

$$\sum_{i_1\cdots i_q\in\mathscr{I}} \xi_{i_1\cdots i_q}(x,y)S_M(i_1\cdots i_q) = M(x,y) \quad \forall (x,y)\in\overline{P} \tag{4.5}$$

Of course, we also ask that the serendipity coordinates be written as a linear combination of the q-th coordinates:

$$\xi_{i_1\cdots i_q} = \sum_{a_1\cdots a_q} c^{i_1\cdots i_q}_{a_1\cdots a_q}\mu_{a_1\cdots a_q} \quad \forall i_1\cdots i_q \in \mathscr{I} \tag{4.6}$$

In order to find these coordinates, we prove a generalized version of Theorem 3.2.1:

Theorem 4.3.1. *If $\xi_{i_1\cdots i_q}$ are defined as in* (4.6) *with $c^{i_1\cdots i_q}_{a_1\cdots a_q}$ such that*

$$\sum_{i_1\cdots i_q\in\mathscr{I}} S_M(i_1\cdots i_q)c^{i_1\cdots i_q}_{a_1\cdots a_q} = S_M(a_1\cdots a_q) \quad \forall a_1\cdots a_q \in \mathscr{A} \tag{4.7}$$

for any monomial $M = M(x,y)$ of degree at most q, then $\xi_{i_1\cdots i_q}$ satisfy (4.5).

Proof. Let $M = M(x,y)$ be a monomial of degree at most q. Then, it holds that

$$\begin{aligned}
&\sum_{i_1\cdots i_q\in\mathscr{I}} \xi_{i_1\cdots i_q}(x,y)S_M(i_1\cdots i_q)\\
&= \sum_{i_1\cdots i_q\in\mathscr{I}} \left(\sum_{a_1\cdots a_q} c^{i_1\cdots i_q}_{a_1\cdots a_q}\mu_{a_1\cdots a_q}\right) S_M(i_1\cdots i_q)\\
&= \sum_{a_1\cdots a_q}\sum_{i_1\cdots i_q\in\mathscr{I}} c^{i_1\cdots i_q}_{a_1\cdots a_q}\mu_{a_1\cdots a_q} S_M(i_1\cdots i_q)\\
&= \sum_{a_1\cdots a_q} \mu_{a_1\cdots a_q}\left(\sum_{i_1\cdots i_q\in\mathscr{I}} S_M(i_1\cdots i_q)c^{i_1\cdots i_q}_{a_1\cdots a_q}\right)\\
&= \sum_{a_1\cdots a_q} \mu_{a_1\cdots a_q} S_M(a_1\cdots a_q) = M(x,y) \qquad (\diamond\diamond)
\end{aligned}$$

where $(\diamond\diamond)$ is due to the application of (4.7) and Theorem 4.2.4. □

Just like for quadratic serendipity coordinates, this problem is underconstrained for $n > 3$. Our goal is to follow the same process as in that case, writing the problem as a linear system and using that to find a solution.

We fix an ordering $\hat{i}_1\cdots\hat{i}_q,\cdots,\hat{j}_1\cdots\hat{j}_q$ for elements in $\mathscr{I}$ and an ordering $\hat{a}_1\cdots\hat{a}_q,\cdots,\hat{b}_1\cdots\hat{b}_q$ for elements in $\mathscr{A}\setminus\mathscr{I}$. We also fix an ordering $M_1,\cdots,M_{\frac{(q+1)(q+2)}{2}}$ for monomials of degree up to q. Then, we define the matrix

$$\boldsymbol{M} = \begin{bmatrix}\boldsymbol{B} & | & \boldsymbol{Q}\end{bmatrix}; \tag{4.8}$$

$$\boldsymbol{B} = \begin{bmatrix} S_{M_1}(\hat{i}_1\cdots\hat{i}_q) & \cdots & S_{M_1}(\hat{j}_1\cdots\hat{j}_q)\\ \vdots & \ddots & \vdots \\ S_{M_{\frac{(q+1)(q+2)}{2}}}(\hat{i}_1\cdots\hat{i}_q) & \cdots & S_{M_{\frac{(q+1)(q+2)}{2}}}(\hat{j}_1\cdots\hat{j}_q)\end{bmatrix};\ \boldsymbol{Q} = \begin{bmatrix} S_{M_1}(\hat{a}_1\cdots\hat{a}_q) & \cdots & S_{M_1}(\hat{b}_1\cdots\hat{b}_q)\\ \vdots & \ddots & \vdots \\ S_{M_{\frac{(q+1)(q+2)}{2}}}(\hat{a}_1\cdots\hat{a}_q) & \cdots & S_{M_{\frac{(q+1)(q+2)}{2}}}(\hat{b}_1\cdots\hat{b}_q)\end{bmatrix}$$

and the arrays

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{\hat{i}_1 \cdots \hat{i}_q} & \cdots & \mu_{\hat{j}_1 \cdots \hat{j}_q} & \mu_{\hat{a}_1 \cdots \hat{a}_q} & \cdots & \mu_{\hat{b}_1 \cdots \hat{b}_q} \end{bmatrix}^T;$$

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_{\hat{i}_1 \cdots \hat{i}_q} & \cdots & \xi_{\hat{j}_1 \cdots \hat{j}_q} \end{bmatrix}^T$$

By Theorem 4.2.4 and by (4.5), it then holds that

$$\boldsymbol{M\mu} = \boldsymbol{B\xi} = \begin{bmatrix} M_1 \\ \vdots \\ M_{\frac{(q+1)(q+2)}{2}} \end{bmatrix} \tag{4.9}$$

on $\overline{P}$.

As in the quadratic case, (4.6) prompts us to look for a matrix

$$\boldsymbol{A} = \begin{bmatrix} c^{\hat{i}_1 \cdots \hat{i}_q}_{\hat{i}_1 \cdots \hat{i}_q} & \cdots & c^{\hat{i}_1 \cdots \hat{i}_q}_{\hat{j}_1 \cdots \hat{j}_q} & c^{\hat{i}_1 \cdots \hat{i}_q}_{\hat{a}_1 \cdots \hat{a}_q} & \cdots & c^{\hat{i}_1 \cdots \hat{i}_q}_{\hat{b}_1 \cdots \hat{b}_q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c^{\hat{j}_1 \cdots \hat{j}_q}_{\hat{i}_1 \cdots \hat{i}_q} & \cdots & c^{\hat{j}_1 \cdots \hat{j}_q}_{\hat{j}_1 \cdots \hat{j}_q} & c^{\hat{j}_1 \cdots \hat{j}_q}_{\hat{a}_1 \cdots \hat{a}_q} & \cdots & c^{\hat{j}_1 \cdots \hat{j}_q}_{\hat{b}_1 \cdots \hat{b}_q} \end{bmatrix}$$

of coefficients $c^{i_1 \cdots i_q}_{a_1 \cdots a_q}$, such that

$$\boldsymbol{\xi} = \boldsymbol{A\mu}$$

(4.9) actually points us towards a way of finding a possible form for $\boldsymbol{A}$. Indeed, we observe that

$$\boldsymbol{M\mu} = \boldsymbol{B\xi} = \boldsymbol{BA\mu}$$

This obviously holds if

$$\boldsymbol{BA} = \boldsymbol{M} = \begin{bmatrix} \boldsymbol{B} & | & \boldsymbol{Q} \end{bmatrix} \tag{4.10}$$

where the last equality is from (4.8). The product between $\boldsymbol{B}$ and $\boldsymbol{A}$ yields $\boldsymbol{M}$, whose leftmost elements are exactly the elements of $\boldsymbol{B}$. This suggests writing $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & | & \boldsymbol{A'} \end{bmatrix}$$

where $\boldsymbol{I}$ is the identity matrix of appropriate size. (4.10) can thus be rewritten as

$$\boldsymbol{B}\begin{bmatrix} \boldsymbol{I} & | & \boldsymbol{A'} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B} & | & \boldsymbol{Q} \end{bmatrix}$$

which obviously holds if and only if

$$\boldsymbol{BA'} = \boldsymbol{Q}$$

Thus, just like in the quadratic case, $\boldsymbol{A'}$ (and thus $\boldsymbol{A}$) can be found by inverting $\boldsymbol{B}$ through its Moore-Penrose pseudoinverse:

$$\boldsymbol{A'} = \boldsymbol{B}^{\dagger}\boldsymbol{Q}; \quad \boldsymbol{B}^{\dagger} = \boldsymbol{B}^T(\boldsymbol{BB}^T)^{-1}$$

Assuming the pseudoinverse yields a valid solution, this whole method can be used to find a reduced space of serendipity coordinates of any size, using any set of corresponding unordered sequences (and thus q-th coordinates). Similarly to (3.11), the specific structure of $\boldsymbol{A}$ implies that these serendipity coordinates can be written as

$$\xi_{i_1\cdots i_q} = \mu_{i_1\cdots i_q} + \sum_{a_1\cdots a_q \in \mathscr{A}\setminus\mathscr{I}} c^{i_1\cdots i_q}_{a_1\cdots a_q} \mu_{a_1\cdots a_q} \tag{4.11}$$

In our case, we choose a specific definition of $\mathscr{I}$ based on a specific ordering of $\mathscr{A}$. We define said ordering by first placing all the sequences of the form $a\cdots a, a = 1,\cdots,n$. Then, we place all the sequences of the form $a\cdots a(a+1), a = 1,\cdots,n$. Then come all the sequences of the form $a\cdots a(a+1)(a+1), a = 1,\cdots,n$; and so on until all the ones of the form $a(a+1)\cdots(a+1), a = 1,\cdots,n$. Then come all other sequences ordered in lexicographical-like ordering, i.e. "extended" ascending order: in practice, if $a_1\cdots a_q$ and $b_1\cdots b_q$ are indexed such that $a_1 \leq \cdots \leq a_q$ and $b_1 \leq \cdots \leq b_q$, then $a_1\cdots a_q$ is ordered before $b_1\cdots b_q$ if and only if there exists $l \geq 0$ such that $a_1 = b_1,\cdots,a_l = b_l, a_{l+1} < b_{l+1}$.

We then take $\mathscr{I}$ as the first $r > 0$ elements of $\mathscr{A}$ and we keep the same ordering on both $\mathscr{I}$ and $\mathscr{A}\setminus\mathscr{I}$. Generally, we will assume $r = qn$, because we would like to mirror the methodology in the last chapter (more specifically, generalize Remark 3.2.2). However, we will see later in the thesis that sometimes this value for r is not sufficient, so the following results for the theory of this thesis will be proven for a generic $r \geq qn$.

Example: Let us find an ordering for the set $\mathscr{A}$ of sequences $a_1\cdots a_q, 1 \leq a_1,\cdots,a_q \leq n$, picking $q = 3$ and $n = 4$.

This means we must order all sequences $abc, 1 \leq a,b,c \leq 4$. Following the reasoning outlined above, first we place all the sequences of the form aaa:

$$111, 222, 333, 444$$

Then, we place all sequences of the form $aa(a+1)$:

$$112, 223, 334, 441$$

Then, we place all sequences of the form $a(a+1)(a+1)$:

$$122, 233, 344, 411$$

Finally, we place all other sequences in "extended" ascending order, as explained earlier:

$$113, 123, 124, 133, 134, 224, 234, 244$$

In total, this is the final ordering of the set $\mathscr{A}$ of sequences $abc, 1 \leq a,b,c \leq 4$:

$$111, 222, 333, 444, 112, 223, 334, 441, 122, 233,$$
$$344, 411, 113, 123, 124, 133, 134, 224, 234, 244$$

If we take $\mathscr{I}$ as the first $qn = 3 \cdot 4 = 12$ sequences of $\mathscr{A}$, then $\mathscr{I}$ will be made up of the sequences

$$111, 222, 333, 444, 112, 223, 334, 441, 122, 233, 344, 411$$

Example: Let us find $\boldsymbol{\xi}$ by building $\boldsymbol{M}$ and $\boldsymbol{\mu}$, for $q = 2$ over the unit square P with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1), \boldsymbol{v}_4 = (0,1)$. We look for $qn = 2 \cdot 4 = 8$ serendipity coordinate functions. Note that, for $q = 2$, (3.10) already provides a more explicit construction, but we do not follow it here for the sake of this example.

Similarly to the previous example, we must order the set $\mathscr{A}$ of sequences $a_1 \cdots a_q, 1 \leq a_1, \cdots, a_q \leq n$, but this time taking $q = 2$ and $n = 4$. We start by placing all the sequences of the form aa:

$$11, 22, 33, 44$$

Then, we place all sequences of the form $a(a+1)$:

$$12, 23, 34, 41$$

Finally, we place all other sequences in "extended" ascending order:

$$13, 24$$

In total, this is the final ordering of the set $\mathscr{A}$ of sequences $abc, 1 \leq a, b, c \leq 3$:

$$11, 22, 33, 44, 12, 23, 34, 41, 13, 24$$

We will be using this ordering in the definitions of $\boldsymbol{M}$, $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$. For the latter two, it is as simple as ordering all the functions in the arrays according to their associated unordered list:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{11} & \mu_{22} & \mu_{33} & \mu_{44} & \mu_{12} & \mu_{23} & \mu_{34} & \mu_{41} & \mu_{13} & \mu_{24} \end{bmatrix}^T$$

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_{11} & \xi_{22} & \xi_{33} & \xi_{44} & \xi_{12} & \xi_{23} & \xi_{34} & \xi_{41} \end{bmatrix}^T$$

Of course, $\boldsymbol{\xi}$ only contains $qn = 8$ functions, as stated earlier.

As for $\boldsymbol{M}$, the ordering of its columns relies on the ordering of $\mathscr{A}$ we just found, but the ordering of its rows relies on an ordering of all monomials of degree at most 2. For this purpose, we choose the following ordering:

$$1, x, y, x^2, xy, y^2$$

Thus, we can already write an extended form of $\boldsymbol{M}$:

$$\boldsymbol{M} = \begin{bmatrix} S_1(11) & S_1(22) & S_1(33) & S_1(44) & S_1(12) & S_1(23) & S_1(34) & S_1(41) & S_1(13) & S_1(24) \\ S_x(11) & S_x(22) & S_x(33) & S_x(44) & S_x(12) & S_x(23) & S_x(34) & S_x(41) & S_x(13) & S_x(24) \\ S_y(11) & S_y(22) & S_y(33) & S_y(44) & S_y(12) & S_y(23) & S_y(34) & S_y(41) & S_y(13) & S_y(24) \\ S_{x^2}(11) & S_{x^2}(22) & S_{x^2}(33) & S_{x^2}(44) & S_{x^2}(12) & S_{x^2}(23) & S_{x^2}(34) & S_{x^2}(41) & S_{x^2}(13) & S_{x^2}(24) \\ S_{xy}(11) & S_{xy}(22) & S_{xy}(33) & S_{xy}(44) & S_{xy}(12) & S_{xy}(23) & S_{xy}(34) & S_{xy}(41) & S_{xy}(13) & S_{xy}(24) \\ S_{y^2}(11) & S_{y^2}(22) & S_{y^2}(33) & S_{y^2}(44) & S_{y^2}(12) & S_{y^2}(23) & S_{y^2}(34) & S_{y^2}(41) & S_{y^2}(13) & S_{y^2}(24) \end{bmatrix}$$

Of course, this form is still not explicit, as we have yet to write out the exact value of all the $S_M(ab)$. We already know from the proof of Theorem 4.2.4 that

$$S_M(ab) = \sum_{(c,d)\in\mathscr{P}(ab)} \hat{M}(c,d)$$

for all unordered sequences ab and for all monomials M of degree up to 2. Because $q = 2$, depending on the unordered sequence, there are two possible cases for the definition of $S_M(ab)$:

- If the sequence is of the form aa, then $\mathscr{P}(aa) = \{(a,a)\}$. Thus

$$S_M(aa) = \hat{M}(a,a)$$

- If the sequence is of the form $ab, a \neq b$, then $\mathscr{P}(ab) = \{(a,b);(b,a)\}$. Thus

$$S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a)$$

Let us go over each monomial one by one while keeping in mind this definition for S_M:

- If $M(x,y) = 1$, then $\hat{M}(a,b) = 1$. Thus

$$\begin{aligned} &S_M(aa) = \hat{M}(a,a) = 1 \quad \forall aa \in \mathscr{A} \\ &S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a) = 1 + 1 = 2 \quad \forall ab \in \mathscr{A}, a \neq b \end{aligned}$$

- If $M(x,y) = x$, then $\hat{M}(a,b) = x_a$. Thus

$$\begin{aligned} &S_M(aa) = \hat{M}(a,a) = x_a \quad \forall aa \in \mathscr{A} \\ &S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a) = x_a + x_b \quad \forall ab \in \mathscr{A}, a \neq b \end{aligned}$$

- If $M(x,y) = y$, then $\hat{M}(a,b) = y_a$. Thus

$$\begin{aligned} &S_M(aa) = \hat{M}(a,a) = y_a \quad \forall aa \in \mathscr{A} \\ &S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a) = y_a + y_b \quad \forall ab \in \mathscr{A}, a \neq b \end{aligned}$$

- If $M(x,y) = x^2$, then $\hat{M}(a,b) = x_a x_b$. Thus

$$\begin{aligned} &S_M(aa) = \hat{M}(a,a) = x_a x_a = (x_a)^2 \quad \forall aa \in \mathscr{A} \\ &S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a) = x_a x_b + x_b x_a = 2x_a x_b \quad \forall ab \in \mathscr{A}, a \neq b \end{aligned}$$

- If $M(x,y) = xy$, then $\hat{M}(a,b) = x_a y_b$. Thus

$$\begin{aligned} &S_M(aa) = \hat{M}(a,a) = x_a y_a \quad \forall aa \in \mathscr{A} \\ &S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a) = x_a y_b + x_b y_a \quad \forall ab \in \mathscr{A}, a \neq b \end{aligned}$$

- If $M(x,y) = y^2$, then $\hat{M}(a,b) = y_a y_b$. Thus

$$S_M(aa) = \hat{M}(a,a) = y_a y_a = (y_a)^2 \quad \forall aa \in \mathscr{A}$$
$$S_M(ab) = \hat{M}(a,b) + \hat{M}(b,a) = y_a y_b + y_b y_a = 2y_a y_b \quad \forall ab \in \mathscr{A}, a \neq b$$

Thus, we can substitute all the respective $S_M(ab)$ in $\boldsymbol{M}$ with these more explicit forms:

$$\boldsymbol{M} = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ x_1 & x_2 & x_3 & x_4 & x_1+x_2 & x_2+x_3 & x_3+x_4 & x_1+x_4 & x_1+x_3 & x_2+x_4 \\ y_1 & y_2 & y_3 & y_4 & y_1+y_2 & y_2+y_3 & y_3+y_4 & y_1+y_4 & y_1+y_3 & y_2+y_4 \\ (x_1)^2 & (x_2)^2 & (x_3)^2 & (x_4)^2 & 2x_1x_2 & 2x_2x_3 & 2x_3x_4 & 2x_1x_4 & 2x_1x_3 & 2x_2x_4 \\ x_1y_1 & x_2y_2 & x_3y_3 & x_4y_4 & \begin{matrix} x_1y_2+ \\ +x_2y_1 \end{matrix} & \begin{matrix} x_2y_3+ \\ +x_3y_2 \end{matrix} & \begin{matrix} x_3y_4+ \\ +x_4y_3 \end{matrix} & \begin{matrix} x_4y_1+ \\ +x_1y_4 \end{matrix} & \begin{matrix} x_1y_3+ \\ +x_3y_1 \end{matrix} & \begin{matrix} x_2y_4+ \\ +x_4y_2 \end{matrix} \\ (y_1)^2 & (y_2)^2 & (y_3)^2 & (y_4)^2 & 2y_1y_2 & 2y_2y_3 & 2y_3y_4 & 2y_1y_4 & 2y_1y_3 & 2y_2y_4 \end{bmatrix}$$

More specifically, recalling that $\boldsymbol{v}_1 = (x_1, y_1) = (0,0), \boldsymbol{v}_2 = (x_2, y_2) = (1,0), \boldsymbol{v}_3 = (x_3, y_3) = (1,1), \boldsymbol{v}_4 = (x_4, y_4) = (0,1)$, we can write $\boldsymbol{M}$ numerically:

$$\boldsymbol{M} = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 1 & 0 & 0+1 & 1+1 & 1+0 & 0+0 & 0+1 & 1+0 \\ 0 & 0 & 1 & 1 & 0+0 & 0+1 & 1+1 & 0+1 & 0+1 & 0+1 \\ 0^2 & 1^2 & 1^2 & 0^2 & 2\cdot0\cdot1 & 2\cdot1\cdot1 & 2\cdot1\cdot0 & 2\cdot0\cdot0 & 2\cdot0\cdot1 & 2\cdot1\cdot0 \\ 0\cdot0 & 1\cdot0 & 1\cdot1 & 0\cdot1 & \begin{matrix} 0\cdot0+ \\ +1\cdot0 \end{matrix} & \begin{matrix} 1\cdot1+ \\ +1\cdot0 \end{matrix} & \begin{matrix} 1\cdot1+ \\ +0\cdot1 \end{matrix} & \begin{matrix} 0\cdot0+ \\ +0\cdot1 \end{matrix} & \begin{matrix} 0\cdot1+ \\ +1\cdot0 \end{matrix} & \begin{matrix} 1\cdot1+ \\ +0\cdot0 \end{matrix} \\ 0^2 & 0^2 & 1^2 & 1^2 & 2\cdot0\cdot0 & 2\cdot0\cdot1 & 2\cdot1\cdot1 & 2\cdot0\cdot1 & 2\cdot0\cdot1 & 2\cdot0\cdot1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

In order to find ξ from $\boldsymbol{M}$ and $\boldsymbol{\mu}$, we must find $\boldsymbol{A}$, which involves splitting $\boldsymbol{M}$ between $\boldsymbol{B}$ and $\boldsymbol{Q}$. $\boldsymbol{B}$ is made up of the first $qn = 8$ columns of $\boldsymbol{M}$, while $\boldsymbol{Q}$ is made up of the remaining columns:

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \\ 0 & 1 & 1 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 2 & 0 \end{bmatrix}; \quad \boldsymbol{Q} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Then,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & | & \boldsymbol{A}' \end{bmatrix}; \quad \boldsymbol{A}' = \boldsymbol{B}^\dagger \boldsymbol{Q}; \quad \boldsymbol{B}^\dagger = \boldsymbol{B}^T (\boldsymbol{B}\boldsymbol{B}^T)^{-1}$$

Through numerical testing, we find that

$$\boldsymbol{B}^{\dagger} = \frac{1}{6}\begin{bmatrix} 4 & -7 & -7 & 3 & 6 & 3 \\ 0 & 1 & -1 & 3 & -6 & 3 \\ 2 & -5 & -5 & 3 & 6 & 3 \\ 0 & -1 & 1 & 3 & -6 & 3 \\ 1 & 3 & -2 & -3 & 0 & 0 \\ -1 & 2 & 3 & 0 & 0 & -3 \\ -1 & 3 & 2 & -3 & 0 & 0 \\ 1 & -2 & 3 & 0 & 0 & -3 \end{bmatrix}$$

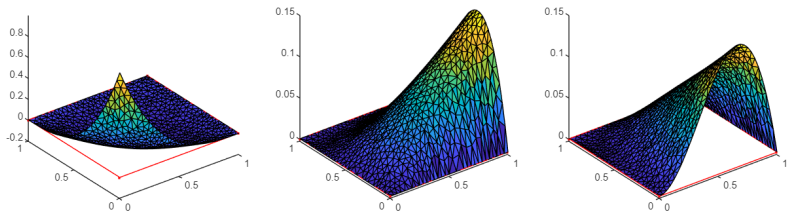
Thus,

$$\boldsymbol{A}' = \boldsymbol{B}^{\dagger}\boldsymbol{Q} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -1 & 0 \\ 0 & -1 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \Longrightarrow \boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & | & \boldsymbol{A}' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

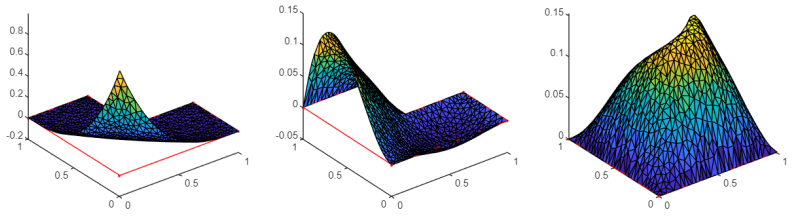
This means that, finally, we can write

$$\begin{bmatrix} \xi_{11} \\ \xi_{22} \\ \xi_{33} \\ \xi_{44} \\ \xi_{12} \\ \xi_{23} \\ \xi_{34} \\ \xi_{41} \end{bmatrix} = \xi = \boldsymbol{A}\boldsymbol{\mu} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{22} \\ \mu_{33} \\ \mu_{44} \\ \mu_{12} \\ \mu_{23} \\ \mu_{34} \\ \mu_{41} \\ \mu_{13} \\ \mu_{24} \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{13} \\ \mu_{22} - \mu_{24} \\ \mu_{33} - \mu_{13} \\ \mu_{44} - \mu_{24} \\ \mu_{12} + \frac{1}{2}\mu_{13} + \frac{1}{2}\mu_{24} \\ \mu_{23} + \frac{1}{2}\mu_{13} + \frac{1}{2}\mu_{24} \\ \mu_{34} + \frac{1}{2}\mu_{13} + \frac{1}{2}\mu_{24} \\ \mu_{41} + \frac{1}{2}\mu_{13} + \frac{1}{2}\mu_{24} \end{bmatrix}$$

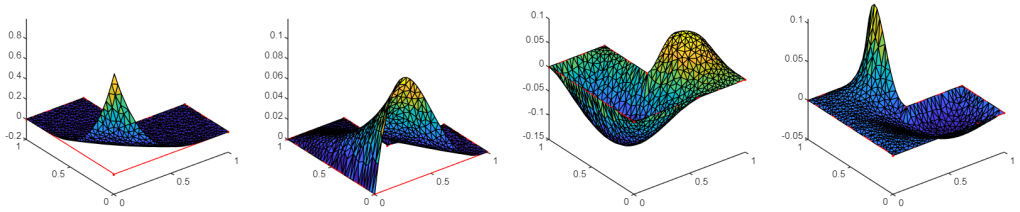
Figure 4.2 shows a few examples of q-th serendipity coordinates over polygons. For higher values of q, each serendipity coordinate $\xi_{i_1 \cdots i_q}$ takes on a noticeably different shape from the corresponding q-th coordinate $\mu_{i_1 \cdots i_q}$ on the interior of the polygon.



(a) ξ_{111}, ξ_{223}, ξ_{122} for $q = 3$ over P_1.



(b) ξ_{111}, ξ_{661}, ξ_{344} for $q = 3$ over P_2.



(c) ξ_{1111}, ξ_{1112}, ξ_{3344}, ξ_{4555} for $q = 4$ over P_2.

Figure 4.2. Surf plot of q-th serendipity coordinates for variable value of q over either the quadrilateral P_1 or the hexagon P_2, with P_1 having vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1), \boldsymbol{v}_4 = (0,1)$ and P_2 having vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,\frac{1}{2}), \boldsymbol{v}_4 = (\frac{1}{2},\frac{1}{2}), \boldsymbol{v}_5 = (\frac{1}{2},1), \boldsymbol{v}_6 = (0,1)$. The coefficients for the serendipity coordinates are found through $\boldsymbol{B}$ and $\boldsymbol{Q}$, constructed as in (4.8). Theorem 4.2.4 was proven to aid with the implementation of $\boldsymbol{B}$ and $\boldsymbol{Q}$, as $S_M(a_1 \cdots a_q)$ is easier to code and compute than $\sum_{(a_{p_1},\cdots,a_{p_q}) \in \mathscr{P}(a_1 \cdots a_q)} \hat{M}(a_{p_1}, \cdots, a_{p_q})$.

Chapter 5

Interpolation functions

5.1 Connection to Bernstein basis polynomials

We started this thesis with the intent of using GBCs and serendipity coordinates for FEM: therefore, our goal is to use our method in order to obtain a proper set of FEM basis functions. FEM basis functions often take the form of *interpolation functions*, i.e. functions $v_{i_1\cdots i_q}$ coupled with interpolation nodes $\boldsymbol{v}_{i_1\cdots i_q}$ such that any function $f = f(x,y)$ possesses an *interpolant* s_f such that

$$s_f(\boldsymbol{v}) = \sum_{i_1\cdots i_q} v_{i_1\cdots i_q}(\boldsymbol{v}) f(\boldsymbol{v}_{i_1\cdots i_q}) \quad \forall \boldsymbol{v} \in \overline{P}; \tag{5.1a}$$

$$s_f(\boldsymbol{v}_{i_1\cdots i_q}) = f(\boldsymbol{v}_{i_1\cdots i_q}) \quad \forall \boldsymbol{v}_{i_1\cdots i_q}; \tag{5.1b}$$

$$s_f = f \quad \forall f \text{ polynomials of degree at most } q \tag{5.1c}$$

(5.1a) and (5.1b) together imply

$$v_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q}) = \delta_{i_1\cdots i_q, j_1\cdots j_q} \tag{5.2}$$

where $\delta_{i_1\cdots i_q, j_1\cdots j_q}$ is the Kronecker delta:

$$\delta_{i_1\cdots i_q, j_1\cdots j_q} = \begin{cases} 1 & i_1\cdots i_q = j_1\cdots j_q \\ 0 & i_1\cdots i_q \neq j_1\cdots j_q \end{cases}$$

The idea is to write every function $v_{i_1\cdots i_q}$ as a linear combination of the serendipity functions $\xi_{i_1\cdots i_q}$:

$$v_{i_1\cdots i_q} = \sum_{j_1\cdots j_q \in \mathscr{I}} \alpha_{j_1\cdots j_q} \xi_{j_1\cdots j_q} \tag{5.3}$$

For s_f defined as in (5.1a), this would satisfy (5.2) and in turn (5.1b). Because serendipity functions can reproduce polynomials of degree up to q, (5.1c) would also be satisfied.

The specific method we intend to employ in this approach makes use of a connection between serendipity functions and Bernstein basis polynomials. By Remark 3.2.2, we have already

proven that quadratic serendipity coordinates are proportional to the quadratic Bernstein basis polynomials on the boundary of the polygon: our first goal is to extend this property to serendipity coordinates of any order.

Let $\mathscr{R}$ be the set of the first qn elements of $\mathscr{A}$, i.e. the set of the first qn elements of $\mathscr{I}$. By definition of $\mathscr{A}$ and $\mathscr{I}$, it holds that $\mathscr{R} \subseteq \mathscr{I} \subseteq \mathscr{A}$ and that all elements in $\mathscr{R}$ are of the form $i\cdots i$ or $i\cdots i(i+1)\cdots(i+1)$. Then, we can state the following:

Remark 5.1.1. *The serendipity coordinates $\xi_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ are proportional to the Bernstein basis polynomials of degree q when restricted to the edges of $\overline{P}$. The serendipity coordinates $\xi_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}\setminus\mathscr{R}$ are null when restricted to the edges of $\overline{P}$.*

Proof. Because the coordinates λ_a satisfy (2.1a) and (2.1b), it holds that $\{\lambda_i, \lambda_{i+1}\}$ are the linear Bernstein basis polynomials on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$ and all other λ_a vanish on that edge. This implies that $\{\mu_{i\cdots i}, \mu_{i\cdots i(i+1)}, \cdots, \mu_{i(i+1)\cdots(i+1)}, \mu_{(i+1)\cdots(i+1)}\}$ are proportional to the Bernstein basis polynomials of degree q on the same edge and are the only coordinates which do not vanish on that edge. Not only that, but because of those same properties of λ_a, it holds that $\mu_{a_1\cdots a_q}$ vanishes on all edges of $\overline{P}$ for all $a_1\cdots a_q \notin \mathscr{R}$. Thus, by (4.11), it holds that $\xi_{i_1\cdots i_q} = \mu_{i_1\cdots i_q}\ \forall \xi_{i_1\cdots i_q}$ on the edges of $\overline{P}$. This means $\xi_{i_1\cdots i_q}$ is null on the edges of $\overline{P}$ for any $i_1\cdots i_q \notin \mathscr{R}$; and it also means $\{\xi_{i\cdots i}, \xi_{i\cdots i(i+1)}, \cdots, \xi_{i(i+1)\cdots(i+1)}, \xi_{(i+1)\cdots(i+1)}\}$, too, are proportional to the Bernstein basis polynomials of degree q on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$ and are the only serendipity coordinates which do not vanish on that edge. □

As with Remark 3.2.2, the Bernstein basis polynomials of degree q are a basis for the space of polynomials of the same degree, so this remark implies that the space spanned by serendipity functions is exactly the space of polynomials of degree q when restricted to the edges of $\overline{P}$.

This remark actually proves one further property of serendipity coordinates:

Corollary 5.1.2. *If $\mathscr{I} = \mathscr{R}$ or $\mathscr{I} = \mathscr{R} \cup \{\hat{i}_1\cdots\hat{i}_q\}$ for some $\hat{i}_1\cdots\hat{i}_q \in \mathscr{A}\setminus\mathscr{R}$, then the serendipity coordinates $\xi_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}$ are linearly independent.*

Proof. Assume

$$\sum_{i_1\cdots i_q \in \mathscr{I}} \alpha_{i_1\cdots i_q}\xi_{i_1\cdots i_q} = 0 \tag{5.4}$$

for some $\alpha_{i_1\cdots i_q} \in \mathbb{R}$. This implies

$$\sum_{i_1\cdots i_q \in \mathscr{I}} \alpha_{i_1\cdots i_q}\xi_{i_1\cdots i_q} = 0 \ \text{ on } [\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]\ \forall i = 1, \cdots, n$$

By Remark 5.1.1, all serendipity coordinates not of the form $\xi_{i\cdots i(i+1)\cdots(i+1)}$ vanish on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$, meaning the above equation can be rewritten as

$$\sum_{i_1\cdots i_q = i\cdots i(i+1)\cdots(i+1)} \alpha_{i_1\cdots i_q}\xi_{i_1\cdots i_q} = 0 \ \text{ on } [\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]\ \forall i = 1, \cdots, n$$

But, by the same remark, serendipity coordinates of the form $\xi_{i\cdots i(i+1)\cdots(i+1)}$ are proportional to the Bernstein basis polynomials on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$; and Bernstein basis polynomials are linearly independent. Thus

$$\begin{aligned} &\alpha_{i\cdots i(i+1)\cdots(i+1)} = 0 \;\; \forall i\cdots i(i+1)\cdots(i+1),\; \forall i = 1,\cdots,n \\ \implies &\alpha_{i_1\cdots i_q} = 0 \;\; \forall i_1\cdots i_q \in \mathscr{R} \end{aligned}$$

If $\mathscr{I} = \mathscr{R}$, this proves the corollary. If $\mathscr{I} = \mathscr{R} \cup \{\hat{i}_1\cdots\hat{i}_q\}$ for some $\hat{i}_1\cdots\hat{i}_q \in \mathscr{A}$, then (5.4) can be rewritten as

$$\alpha_{\hat{i}_1\cdots\hat{i}_q}\xi_{\hat{i}_1\cdots\hat{i}_q} = 0$$

Because none of the serendipity coordinates are null on $\overline{P}$ by construction, this implies

$$\alpha_{\hat{i}_1\cdots\hat{i}_q} = 0$$

Overall,

$$\alpha_{i_1\cdots i_q} = 0 \;\; \forall i_1\cdots i_q \in \mathscr{I}$$

proving the corollary. □

5.2 1D interpolation from Bernstein basis polynomials

Before working with the serendipity functions directly, we first attempt to obtain 1D interpolation functions from the Bernstein basis polynomials themselves; or rather, from polynomials proportional to them. For $k = 0,\cdots,q$, we define the polynomial $b_{k,q} : [0,1] \to \mathbb{R}$ and the node $x_{k,q} \in [0,1]$ as

$$b_{k,q}(x) = x^{q-k}(1-x)^k; \quad x_{k,q} = \frac{k}{q}$$

All polynomials $b_{k,q}$ are exactly proportional to one of the Bernstein basis polynomials for a factor of $\binom{q}{k}$. From the proof of Remark 5.1.1, it is implied that all functions $\xi_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ are equal to some $b_{k,q}$ on some facet of $\overline{P}$.

We also define $l_{k,q} : [0,1] \to \mathbb{R}$ as the Lagrange fundamental polynomial of degree at most q such that

$$l_{k,q}(x_{m,q}) = \delta_{k,m} = \begin{cases} 1 & k = m \\ 0 & k \neq m \end{cases} \quad \forall m = 0,\cdots,q$$

By these definitions, it holds that

$$b_{k,q} = \sum_{m=0}^{q} b_{k,q}(x_{m,q})l_{m,q} = \sum_{m=1}^{q-1} b_{k,q}(x_{m,q})l_{m,q} \;\; \forall k = 1,\cdots,q-1$$

where the second equivalence is because $b_{k,q}(x_{0,q}) = b_{k,q}(x_{q,q}) = 0, k = 1,\cdots,q-1$. (For now, we only focus on the internal polynomials.)

Therefore, if we define the matrix

$$\boldsymbol{C} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,q-1} \\ \vdots & \ddots & \vdots \\ a_{q-1,1} & \cdots & a_{q-1,q-1} \end{bmatrix} = \begin{bmatrix} b_{1,q}(x_{1,q}) & \cdots & b_{1,q}(x_{q-1,q}) \\ \vdots & \ddots & \vdots \\ b_{q-1,q}(x_{1,q}) & \cdots & b_{q-1,q}(x_{q-1,q}) \end{bmatrix}^{-1}$$

it holds that

$$l_{k,q} = \sum_{m=1}^{q-1} a_{k,m} b_{m,q} \quad \forall k = 1, \cdots, q-1 \tag{5.5}$$

which lets us obtain internal interpolation functions from these polynomials, in the form of $l_{k,q}$. Since $b_{0,q}(x_{0,q}) = b_{q,q}(x_{q,q}) = 1$ and $b_{k,q}(x_{0,q}) = b_{k,q}(x_{q,q}) = 0, k = 1, \cdots, q-1$, the interpolation functions on the extremes can be obtained as

$$\begin{aligned} l_{0,q} &= b_{0,q} - \sum_{m=1}^{q-1} b_{0,q}(x_{m,q}) l_{k,q} = b_{0,q} - \sum_{m=1}^{q-1} \left(\frac{q-m}{q}\right)^q l_{k,q} \\ l_{q,q} &= b_{q,q} - \sum_{m=1}^{q-1} b_{q,q}(x_{m,q}) l_{k,q} = b_{q,q} - \sum_{m=1}^{q-1} \left(\frac{m}{q}\right)^q l_{k,q} \end{aligned} \tag{5.6}$$

We state the explicit relation between $l_{k,q}$ and $b_{k,q}$ for $q = 2, 3, 4$:

Remark 5.2.1. *It holds that*

- $$l_{0,2} = b_{0,2} - b_{1,2}; \quad l_{1,2} = 4b_{1,2}; \quad l_{2,2} = b_{2,2} - b_{1,2}$$
- $$l_{0,3} = b_{0,3} - \frac{1}{2}\left(5b_{1,3} - 2b_{2,3}\right); \; l_{1,3} = \frac{9}{2}\left(2b_{1,3} - b_{2,3}\right);$$
  $$l_{2,3} = \frac{9}{2}\left(2b_{2,3} - b_{1,3}\right); \; l_{3,3} = b_{3,3} - \frac{1}{2}\left(-2b_{1,3} + 5b_{2,3}\right)$$
- $$l_{0,4} = b_{0,4} - \frac{13}{3}b_{1,4} + \frac{13}{3}b_{2,4} - b_{3,4}; \; l_{1,4} = 16b_{1,4} - \frac{64}{3}b_{2,4} + \frac{16}{3}b_{3,4};$$
  $$l_{2,4} = -12b_{1,4} + 40b_{2,4} - 12b_{3,4}; \; l_{3,4} = \frac{16}{3}b_{1,4} - \frac{64}{3}b_{2,4} + 16b_{3,4};$$
  $$l_{4,4} = b_{0,4} - b_{1,4} + \frac{13}{3}b_{2,4} - \frac{13}{3}b_{3,4}$$

Proof. It is sufficient to write out every $l_{k,q}$ explicitly and evaluate them over each node. □

We would now like to apply these results in order to properly obtain interpolation functions from the serendipity coordinates. We follow the lead of the method detailed by Floater and Lai [2016], who present a way of obtaining interpolation functions from reduced functions which take the form of Bernstein basis polynomials on the boundary of a polygon: we intend to apply this method to our own serendipity functions as well.

5.3 Interpolation functions from serendipity coordinates

Let us consider the serendipity coordinates $\xi_{i_1 \cdots i_q}$, $i_1 \cdots i_q \in \mathscr{R}$. By definition of $\mathscr{R}$, these coordinates are of the form $\xi_{i \cdots i}$ or $\xi_{i \cdots i(i+1) \cdots (i+1)}$ for some i. Let us attempt to obtain functions satisfying (5.2) and (5.3) from these serendipity coordinates only: we call these functions $\tilde{v}_{i_1 \cdots i_q}$.

By Remark 5.1.1 and its proof, the functions $\{\xi_{i \cdots i}, \xi_{i \cdots i(i+1)}, \cdots, \xi_{i(i+1) \cdots (i+1)}, \xi_{(i+1) \cdots (i+1)}\}$ are proportional to the Bernstein basis polynomials on the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$. In practice, this means that, for any $\boldsymbol{v} = \boldsymbol{v}_i + t(\boldsymbol{v}_{i+1} - \boldsymbol{v}_i)$ $(t \in [0,1])$, it holds that

$$\xi_{i \cdots i \underbrace{(i+1) \cdots (i+1)}_{k}}(\boldsymbol{v}) = b_{k,q}(t) \quad \forall k = 0, \cdots, q \tag{5.7}$$

We define the nodes $\boldsymbol{v}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ such that

$$\boldsymbol{v}_{i\cdots i\underbrace{(i+1)\cdots(i+1)}_{k}} = \frac{k}{q}\boldsymbol{v}_i + \frac{q-k}{q}\boldsymbol{v}_{i+1} \quad \forall k = 0,\cdots,q \tag{5.8}$$

Thus, recalling the definition of $\nu_{i_1\cdots i_q}$, we can utilize (5.7) to rewrite (5.5) as

$$\tilde{\nu}_{i\cdots i\underbrace{(i+1)\cdots(i+1)}_{k}} = \sum_{m=1}^{q-1} \alpha_{k,m}\xi_{i\cdots i\underbrace{(i+1)\cdots(i+1)}_{m}} \quad \forall k = 1,\cdots,q-1 \tag{5.9a}$$

(5.3) trivially holds for $\tilde{\nu}_{i_1\cdots i_q}$ defined this way; and by the definition of $\boldsymbol{v}_{i_1\cdots i_q}$, it is easy to see that (5.2) holds as well.

For functions of the form $\tilde{\nu}_{i\cdots i}$, we must note that the function $\xi_{i\cdots i}$ is proportional to a Bernstein basis polynomial (in the sense of (5.7)) on two edges: it is $b_{q,q}$ on $[\boldsymbol{v}_{i-1}, \boldsymbol{v}_i]$ and $b_{0,q}$ on $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$. Thus, in order to obtain $\tilde{\nu}_{i\cdots i}$ from it, we can utilize (5.7) to rewrite (5.6), but we must subtract the internal interpolation functions from both edges:

$$\begin{aligned}\tilde{\nu}_{i\cdots i} &= \xi_{i\cdots i} - \sum_{k=1}^{q-1}\left(\frac{q-k}{q}\right)^q \xi_{\underbrace{i\cdots i}_{q-k}(i+1)\cdots(i+1)} - \sum_{k=1}^{q-1}\left(\frac{q-k}{q}\right)^q \xi_{\underbrace{(i-1)\cdots(i-1)}_{k}i\cdots i} \\ &= \xi_{i\cdots i} - \sum_{k=1}^{q-1}\left(\frac{q-k}{q}\right)^q \left(\xi_{\underbrace{i\cdots i}_{q-k}(i+1)\cdots(i+1)} + \xi_{\underbrace{(i-1)\cdots(i-1)}_{k}i\cdots i}\right)\end{aligned} \tag{5.9b}$$

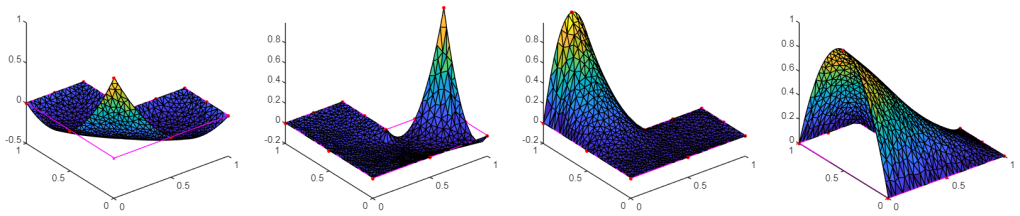
These observations allow us to use Remark 5.2.1 in order to explicitly write out $\tilde{\nu}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ for $q = 2,3,4$:

Remark 5.3.1. *It holds that*

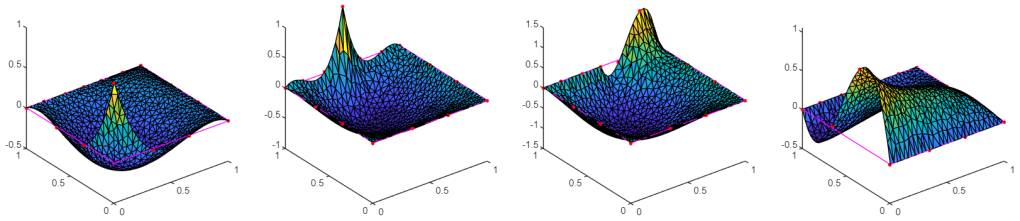
- $$\tilde{\nu}_{ii} = \xi_{ii} - \xi_{(i-1)i} - \xi_{i(i+1)}; \quad \tilde{\nu}_{i(i+1)} = 4\xi_{i(i+1)}$$
- $$\tilde{\nu}_{iii} = \xi_{iii} - \left(\xi_{ii(i+1)} + \xi_{(i-1)ii}\right) + \frac{5}{2}\left(\xi_{i(i+1)(i+1)} + \xi_{(i-1)(i-1)i}\right);$$
$$\tilde{\nu}_{ii(i+1)} = \frac{9}{2}\left(2\xi_{ii(i+1)} - \xi_{i(i+1)(i+1)}\right); \quad \tilde{\nu}_{i(i+1)(i+1)} = \frac{9}{2}\left(2\xi_{i(i+1)(i+1)} - \xi_{ii(i+1)}\right)$$
- $$\begin{aligned}\tilde{\nu}_{iiii} &= \xi_{iiii} - \frac{13}{3}\left(\xi_{iii(i+1)} + \xi_{(i-1)(i-1)(i-1)i}\right) + \\ &+ \frac{13}{3}\left(\xi_{ii(i+1)(i+1)} + \xi_{(i-1)(i-1)ii}\right) - \left(\xi_{i(i+1)(i+1)(i+1)} + \xi_{(i-1)iii}\right);\end{aligned}$$
$$\tilde{\nu}_{iii(i+1)} = 16\xi_{iii(i+1)} - \frac{64}{3}\xi_{ii(i+1)(i+1)} + \frac{16}{3}\xi_{i(i+1)(i+1)(i+1)};$$
$$\tilde{\nu}_{ii(i+1)(i+1)} = -12\xi_{iii(i+1)} + 40\xi_{ii(i+1)(i+1)} - 12\xi_{i(i+1)(i+1)(i+1)};$$
$$\tilde{\nu}_{i(i+1)(i+1)(i+1)} = \frac{16}{3}\xi_{iii(i+1)} - \frac{64}{3}\xi_{ii(i+1)(i+1)} + 16\xi_{i(i+1)(i+1)(i+1)};$$

for $i = 1,\cdots,n$.

In some cases, $\mathscr{I} = \mathscr{R}$ and therefore $\tilde{\nu}_{i_1\cdots i_q}$ are exactly the interpolation functions $\nu_{i_1\cdots i_q}$. Figure 5.1 shows a few examples of such functions. However, in some cases, $\mathscr{R}$ is a proper subset of $\mathscr{I}$ and therefore obtaining $\nu_{i_1\cdots i_q}$ from $\tilde{\nu}_{i_1\cdots i_q}$ requires further work. We note the following:



(a) v_{11}, v_{33}, v_{56} and v_{61} for $q=2$, over the hexagon P with vertices $\boldsymbol{v}_1=(0,0), \boldsymbol{v}_2=(1,0), \boldsymbol{v}_3=(1,\frac{1}{2}), \boldsymbol{v}_4=(\frac{1}{2},\frac{1}{2}), \boldsymbol{v}_5=(\frac{1}{2},1), \boldsymbol{v}_6=(0,1)$.



(b) v_{111}, v_{444}, v_{334} and v_{511} for $q=3$, over the pentagon P with vertices $\boldsymbol{v}_1=(0,0), \boldsymbol{v}_2=(1,0), \boldsymbol{v}_3=(1,1), \boldsymbol{v}_4=(1,\frac{1}{2}), \boldsymbol{v}_5=(0,1)$.

Figure 5.1. Surf plot of interpolation functions obtained from qn q-th serendipity coordinates, for variable value of q over different polygons. The points highlighted in red are the evaluation of the interpolation function over the qn interpolation nodes.

Theorem 5.3.2. *Let $\boldsymbol{v}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ be a set of nodes defined as in* (5.8)*; and let $\boldsymbol{v}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}\setminus\mathscr{R}$ be a miscellaneous set of nodes on the interior of $\overline{P}$. Let $\tilde{v}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ be defined as in* (5.9a) *and* (5.9b)*; and let $v_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}\setminus\mathscr{R}$ be a set of functions satisfying* (5.3) *and* (5.2) *for $\boldsymbol{v}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}$. Then, if $v_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{R}$ are such that*

$$v_{i_1\cdots i_q} = \tilde{v}_{i_1\cdots i_q} - \sum_{j_1\cdots j_q \in \mathscr{I}\setminus\mathscr{R}} \tilde{v}_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q})\, v_{j_1\cdots j_q}$$

it holds that $v_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}$ are interpolation functions with respect to $\boldsymbol{v}_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}$.

Proof. We simply have to prove that $v_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q}) = \delta_{i_1\cdots i_q, j_1\cdots j_q}$ for all $i_1\cdots i_q, j_1\cdots j_q \in \mathscr{I}$. This is already true by hypothesis for $i_1\cdots i_q \in \mathscr{I}\setminus\mathscr{R}$.

Let $i_1\cdots i_q \in \mathscr{R}$, $j_1\cdots j_q \in \mathscr{I}$. Then, if $j_1\cdots j_q \in \mathscr{R}$, it holds that

$$\begin{aligned} v_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q}) &= \tilde{v}_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q}) - \sum_{\hat{j}_1\cdots\hat{j}_q \in \mathscr{I}\setminus\mathscr{R}} \tilde{v}_{i_1\cdots i_q}(\boldsymbol{v}_{\hat{j}_1\cdots\hat{j}_q})\, v_{\hat{j}_1\cdots\hat{j}_q}(\boldsymbol{v}_{j_1\cdots j_q}) \\ &= \tilde{v}_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q}) - \sum_{\hat{j}_1\cdots\hat{j}_q \in \mathscr{I}\setminus\mathscr{R}} \tilde{v}_{i_1\cdots i_q}(\boldsymbol{v}_{\hat{j}_1\cdots\hat{j}_q})\, \delta_{\hat{j}_1\cdots\hat{j}_q, j_1\cdots j_q} \\ &= \tilde{v}_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q}) = \delta_{i_1\cdots i_q, j_1\cdots j_q} \qquad (\Diamond) \end{aligned}$$

where $(\Diamond)$ is by the definition and properties of $\tilde{v}_{i_1\cdots i_q}$.

On the other hand, if $j_1 \cdots j_q \in \mathscr{I} \setminus \mathscr{R}$, it holds that

$$\begin{aligned} \nu_{i_1 \cdots i_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) &= \tilde{\nu}_{i_1 \cdots i_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) - \sum_{\hat{j}_1 \cdots \hat{j}_q \in \mathscr{I} \setminus \mathscr{R}} \tilde{\nu}_{i_1 \cdots i_q}(\boldsymbol{\nu}_{\hat{j}_1 \cdots \hat{j}_q}) \nu_{\hat{j}_1 \cdots \hat{j}_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) \\ &= \tilde{\nu}_{i_1 \cdots i_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) - \sum_{\hat{j}_1 \cdots \hat{j}_q \in \mathscr{I} \setminus \mathscr{R}} \tilde{\nu}_{i_1 \cdots i_q}(\boldsymbol{\nu}_{\hat{j}_1 \cdots \hat{j}_q}) \delta_{\hat{j}_1 \cdots \hat{j}_q, j_1 \cdots j_q} \\ &= \tilde{\nu}_{i_1 \cdots i_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) - \tilde{\nu}_{i_1 \cdots i_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) = 0 = \delta_{i_1 \cdots i_q, j_1 \cdots j_q} \end{aligned} \tag{$\lozenge\lozenge$}$$

where ($\lozenge\lozenge$) is because $i_1 \cdots i_q \neq j_1 \cdots j_q$, since $i_1 \cdots i_q \in \mathscr{R}$ and $j_1 \cdots j_q \in \mathscr{I} \setminus \mathscr{R}$. □

The difficulty now is finding appropriate functions $\nu_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{I} \setminus \mathscr{R}$ of the form in (5.3) such that they satisfy the hypotheses of this theorem. More specifically, such interpolation functions would be of the form

$$\nu_{i_1 \cdots i_q} = \sum_{j_1 \cdots j_q \in \mathscr{I} \setminus \mathscr{R}} \beta_{j_1 \cdots j_q}^{i_1 \cdots i_q} \xi_{j_1 \cdots j_q} \tag{5.10}$$

for appropriate values of $\beta_{j_1 \cdots j_q}^{i_1 \cdots i_q} \in \mathbb{R}$. In other words, only the functions $\xi_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{I} \setminus \mathscr{R}$ would contribute to the linear combination. Indeed, Corollary 5.1.2 states that the functions of the form $\xi_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{R}$ are linearly independent: thus the only serendipity coordinates whose linear combination gives rise to a function vanishing on the boundary must vanish on the boundary themselves; and Remark 5.1.1 states that the functions $\xi_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{I}$ vanish on the boundary of $\overline{P}$ if and only if $i_1 \cdots i_q \notin \mathscr{R}$.

Because $\nu_{i_1 \cdots i_q}$ defined this way already vanish on the boundary, in order for them to satisfy the hypotheses of the theorem, it is sufficient to check they satisfy (5.2) only on the internal nodes. In other words, they must satisfy the equations

$$\nu_{i_1 \cdots i_q}(\boldsymbol{\nu}_{j_1 \cdots j_q}) = \delta_{i_1 \cdots i_q, j_1 \cdots j_q} \quad \forall i_1 \cdots i_q, j_1 \cdots j_q \in \mathscr{I} \setminus \mathscr{R} \tag{5.11}$$

for appropriate interpolation nodes $\boldsymbol{\nu}_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{I} \setminus \mathscr{R}$.

Thus, every function $\nu_{i_1 \cdots i_q}$ is both defined with respect to $|\mathscr{I} \setminus \mathscr{R}|$ coefficients, as in (5.10), and associated with $|\mathscr{I} \setminus \mathscr{R}|$ constraints equations, as in (5.11). The number of coefficients and constraints is the same: therefore, if all the $\xi_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{I} \setminus \mathscr{R}$ are linearly independent, then there exists exactly one valid solution for each $\nu_{i_1 \cdots i_q}, i_1 \cdots i_q \in \mathscr{I} \setminus \mathscr{R}$.

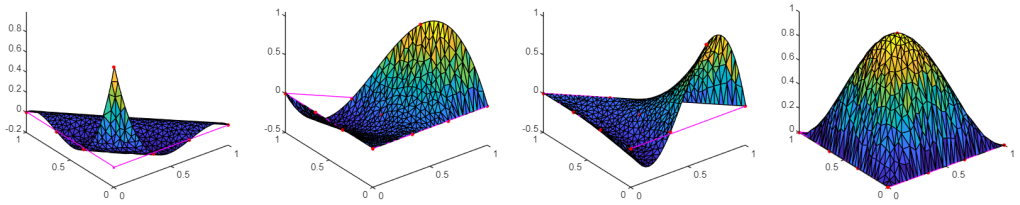
If we write the ordering of $\mathscr{I} \setminus \mathscr{R}$ as $\hat{i}_1 \cdots \hat{i}_q, \cdots, \hat{j}_1 \cdots \hat{j}_q$, then the equations in (5.10) can be collectively written as

$$\boldsymbol{\nu} = \boldsymbol{D} \tilde{\boldsymbol{\xi}} \tag{5.12}$$

where

$$\boldsymbol{\nu} = \begin{bmatrix} \nu_{\hat{i}_1 \cdots \hat{i}_q} & \cdots & \nu_{\hat{j}_1 \cdots \hat{j}_q} \end{bmatrix}^T; \quad \tilde{\boldsymbol{\xi}} = \begin{bmatrix} \xi_{\hat{i}_1 \cdots \hat{i}_q} & \cdots & \xi_{\hat{j}_1 \cdots \hat{j}_q} \end{bmatrix}^T;$$

$$\boldsymbol{D} = \begin{bmatrix} \beta_{\hat{i}_1 \cdots \hat{i}_q}^{\hat{i}_1 \cdots \hat{i}_q} & \cdots & \beta_{\hat{j}_1 \cdots \hat{j}_q}^{\hat{i}_1 \cdots \hat{i}_q} \\ \vdots & \ddots & \vdots \\ \beta_{\hat{i}_1 \cdots \hat{i}_q}^{\hat{j}_1 \cdots \hat{j}_q} & \cdots & \beta_{\hat{j}_1 \cdots \hat{j}_q}^{\hat{j}_1 \cdots \hat{j}_q} \end{bmatrix}$$



(a) ν_{111} (b) ν_{223} (c) ν_{122} (d) ν_{123}

Figure 5.2. Surf plot of interpolation functions obtained from $qn+1$ q-th serendipity coordinates, for $q=3$ over the triangle P with vertices $\boldsymbol{v}_1=(0,0), \boldsymbol{v}_2=(1,0), \boldsymbol{v}_3=(0,1)$. The points highlighted in red are the evaluation of the interpolation function over the $qn+1$ interpolation nodes.

As stated earlier, if all the $\xi_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}\setminus\mathscr{R}$ are linearly independent, then there exists exactly one solution for this system. This also means the matrix $\boldsymbol{D}$ is invertible. Thus, we can write

$$\boldsymbol{D}^{-1}\boldsymbol{\nu}=\tilde{\boldsymbol{\xi}}$$

However, by (5.11), we can easily see that, if there is a linear combination of $\nu_{i_1\cdots i_q}$ that is equal to $\xi_{j_1\cdots j_q}, j_1\cdots j_q \in \mathscr{I}\setminus\mathscr{R}$, then it must be of the form

$$\xi_{j_1\cdots j_q}=\sum_{i_1\cdots i_q\in\mathscr{I}\setminus\mathscr{R}} \xi_{j_1\cdots j_q}(\boldsymbol{v}_{i_1\cdots i_q})\nu_{i_1\cdots i_q}$$

Thus

$$\boldsymbol{D}=\begin{bmatrix} \xi_{\hat{i}_1\cdots\hat{i}_q}(\boldsymbol{v}_{\hat{i}_1\cdots\hat{i}_q}) & \cdots & \xi_{\hat{i}_1\cdots\hat{i}_q}(\boldsymbol{v}_{\hat{j}_1\cdots\hat{j}_q}) \\ \vdots & \ddots & \vdots \\ \xi_{\hat{j}_1\cdots\hat{j}_q}(\boldsymbol{v}_{\hat{i}_1\cdots\hat{i}_q}) & \cdots & \xi_{\hat{j}_1\cdots\hat{j}_q}(\boldsymbol{v}_{\hat{j}_1\cdots\hat{j}_q}) \end{bmatrix}^{-1} \tag{5.13}$$

Unfortunately, proving that the functions $\xi_{i_1\cdots i_q}, i_1\cdots i_q \in \mathscr{I}\setminus\mathscr{R}$ are linearly independent is not straightforward in general. However, Corollary 5.1.2 suggests looking at the case in which $|\mathscr{I}\setminus\mathscr{R}|=1$: indeed, if we write $\mathscr{I}\setminus\mathscr{R}=\{j_1\cdots j_q\}$, then $\{\xi_{j_1\cdots j_q}\}$ is trivially linearly independent. If we then let $\boldsymbol{v}_{j_1\cdots j_q}$ be a point on the interior of $\overline{P}$ such that $\xi_{j_1\cdots j_q}(\boldsymbol{v}_{j_1\cdots j_q})\neq 0$, it is easy to see that

$$\nu_{j_1\cdots j_q}=\frac{\xi_{j_1\cdots j_q}}{\xi_{j_1\cdots j_q}(\boldsymbol{v}_{j_1\cdots j_q})} \tag{5.14a}$$

satisfies both (5.10) and (5.11). Then, Theorem 5.3.2 can be applied in order to find the remaining interpolation functions:

$$\begin{aligned} \nu_{i_1\cdots i_q} &= \tilde{\nu}_{i_1\cdots i_q}-\tilde{\nu}_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q})\nu_{j_1\cdots j_q} \\ &= \tilde{\nu}_{i_1\cdots i_q}-\tilde{\nu}_{i_1\cdots i_q}(\boldsymbol{v}_{j_1\cdots j_q})\frac{\xi_{j_1\cdots j_q}}{\xi_{j_1\cdots j_q}(\boldsymbol{v}_{j_1\cdots j_q})} \quad \forall i_1\cdots i_q\in\mathscr{R} \end{aligned} \tag{5.14b}$$

Figure 5.2 shows an example of such interpolation functions.

Chapter 6

Complications

6.1 Dimension of the space of polynomials

While the explanation for how to apply this method has been quite general so far, there are some difficulties which do not always allow for it to yield correct results. Indeed, since part of the process involves taking a least squares estimation, our method may succeed in its computation even when there is no valid solution, making it necessary to check that there are no issues which would prevent us from properly obtaining the correct result.

The most immediate difficulty concerns the number of serendipity functions. We generally assume we are working with qn serendipity functions, but we cannot do that for any polygon. Indeed, serendipity functions should span the space of polynomials of degree up to q, which has dimension $\frac{(q+1)(q+2)}{2}$: that means we must find at least that many serendipity functions in order for the polynomial reproducibility property to hold. However, in some cases, $qn < \frac{(q+1)(q+2)}{2}$: if we were to try to obtain serendipity coordinate functions in those cases, we would not have enough functions to span the whole space, thus our method would fail. One such case involves cubic serendipity coordinates on a triangle, i.e. $q = 3, n = 3$: then, $qn = 9$ and $\frac{(q+1)(q+2)}{2} = 10$. This means even orders as low as cubic would run into difficulties.

One way to tackle this issue is to simply only work with polygons such that $qn \geq \frac{(q+1)(q+2)}{2}$, essentially imposing that we do not consider polygons with too few sides. In an FEM context, this could be accomplished, for example, by combining adjacent elements to create a bigger polygon when needed. With this approach, Corollary 5.1.2 ensures the linear independence of the serendipity coordinates; and when constructing interpolation functions from them, the interpolation nodes are rigorously defined as in (5.8).

Another way to tackle this issue is to extend the construction of serendipity coordinates so as to find more than qn functions when that would be necessary. This latter method has a few disadvantages: namely, recalling the definitions of $\mathscr{A}$, $\mathscr{I}$ and $\mathscr{R}$ in the previous chapters, this workaround involves working with functions associated with sequences in $\mathscr{I} \setminus \mathscr{R}$, which makes it harder to rigorously prove and define concepts such as their linear independence and their associated interpolation nodes. However, thanks to the work done in previous chapters, it is still possible to obtain worthwhile theoretical results at least for low values of q.

When $q = 3$, this approach would involve taking 10 serendipity coordinates instead of 9 when working with triangles ($n = 3$). That would imply $|\mathscr{I} \setminus \mathscr{R}| = 1$, i.e. $\mathscr{I} = \mathscr{R} \cup \{abc\}$ for some $abc \in \mathscr{A} \setminus \mathscr{R}$. We know from Corollary 5.1.2 that this still allows the resulting serendipity

coordinates to be linearly independent; and building interpolation functions from said serendipity coordinates can be done as in (5.14a) and (5.14b).

When $q = 4$, this approach would involve taking 15 serendipity coordinates instead of 12 when working with triangles ($n = 3$). This is a more tricky case, but it is still possible to work within it. Indeed, these serendipity coordinates are linearly independent. This is because it can be shown that there are exactly 15 elements of $\mathscr{A}$ for $q = 3, n = 4$, meaning there are 15 cubic coordinates: thus, the 15 serendipity coordinates in this case are exactly the cubic coordinates. By Corollary 4.1.2, these coordinates span the space of polynomials of degree up to 3, which has dimension $\frac{(q+1)(q+2)}{2} = 15$: that means they are a basis for that space, thus they are linearly independent. Having shown this property, the interpolation functions can be obtained first through (5.12) and (5.13) and then by following the lead of Theorem 5.3.2.

This latter approach allows for the application of this method over a wider range of polygons, but working with more than qn functions also renders these applications less elegant. Namely, when constructing interpolation functions, each of the extra functions does not have a rigorously defined interpolation node like in (5.8), but rather they are associated to "virtual interpolation nodes" which are randomly picked on the interior of $\overline{P}$.

However, regardless of the approach, there is another issue related to the existence or lack thereof of a valid solution: it concerns the matrix $\boldsymbol{B}$, as defined in (4.8), and the cases in which the construction of its pseudoinverse fails.

6.2 Linear independence of the rows of $\boldsymbol{B}$

We had defined the Moore-Penrose pseudoinverse of $\boldsymbol{B}$ as

$$\boldsymbol{B}^\dagger = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}$$

If defined like this, $\boldsymbol{B}^\dagger$ is a *right inverse* — that is, $\boldsymbol{B}\boldsymbol{B}^\dagger$ is an identity matrix. This is what allows for its use in the solution of the problem: indeed, if $\boldsymbol{A}' = \boldsymbol{B}^\dagger\boldsymbol{Q}$, then $\boldsymbol{B}\boldsymbol{A}' = \boldsymbol{B}\boldsymbol{B}^\dagger\boldsymbol{Q} = \boldsymbol{Q}$, which is our goal. However, this definition is only applicable to matrices with linearly independent rows. If $\boldsymbol{B}$'s rows are not all linearly independent, its Moore-Penrose pseudoinverse can still be defined, but it loses the property of being a right inverse, meaning our method fails. Therefore, we are interested in studying, or at least categorizing, the cases for which $\boldsymbol{B}$'s rows are linearly dependent.

The results described in this section from this point forward have been obtained purely through numerical testing, unless a theoretical explanation is provided.

One of the most notable issues concerns polygons with consecutive collinear vertices, i.e. weakly convex polygons. If a polygon P with n vertices is weakly convex, it can be constructed by taking a strongly convex polygon $\tilde{P}$ with m vertices and adding $n - m$ nodes on its edges: it then holds that, if the construction of our method fails for qm functions on $\tilde{P}$, it will fail for qn functions on P. For example, when $q = 3$, our construction usually succeeds when P is a quadrilateral ($n = 4$), because $qn = 16 > 10 = \frac{(q+1)(q+2)}{2}$. However, if the quadrilateral in question is weakly convex, then it can be constructed by taking a triangle $\tilde{P}$ ($m = 3$) and adding one node to it. Because $qm = 9 < 10 = \frac{(q+1)(q+2)}{2}$, the costruction fails for triangles, therefore it fails for weakly convex quadrilaterals.

To understand the reason for this, it is more helpful to look at the columns of $\boldsymbol{B}$, rather than its rows. Indeed, the dimension of the vector space spanned by $\boldsymbol{B}$'s rows is the same as the dimension of the vector space spanned by $\boldsymbol{B}$'s columns, that being the *rank* of $\boldsymbol{B}$. Therefore, if

$\boldsymbol{B}$'s rank is lower than the number of its rows, then its rows are not linearly independent and our construction fails.

In this situation, let $\tilde{\boldsymbol{B}}$ be the matrix of qm columns constructed from $\tilde{P}$, akin to $\boldsymbol{B}$ with qn columns for P. For the sake of this example, let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$ be the vertices of $\tilde{P}$ and $\{\boldsymbol{v}_{m+1}, \cdots, \boldsymbol{v}_n\}$ be the additional vertices of P. Then, the columns of $\tilde{\boldsymbol{B}}$ can be written as a linear combination of the columns of $\boldsymbol{B}$. More specifically, the columns of $\boldsymbol{B}$ which are constructed only from the coordinates of $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$ are identical to some columns of $\tilde{\boldsymbol{B}}$.

Example: Let us take $q = 2, n = 4, m = n-1 = 3$. Thus, the vertices of $\tilde{P}$ are $\boldsymbol{v}_1 = (x_1, y_1), \boldsymbol{v}_2 = (x_2, y_2), \boldsymbol{v}_3 = (x_3, y_3)$ and the vertices of P are $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$ with $\boldsymbol{v}_4 = t\boldsymbol{v}_3 + (1-t)\boldsymbol{v}_1, 0 < t < 1$.

We recall the definition of $\boldsymbol{B}$ in (3.10) and define $\boldsymbol{B}$, $\tilde{\boldsymbol{B}}$ and $\boldsymbol{q}_{ab}$ like so:

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \\ x_1 & x_2 & x_3 & x_4 & x_1+x_2 & x_2+x_3 & x_3+x_4 & x_1+x_4 \\ y_1 & y_2 & y_3 & y_4 & y_1+y_2 & y_2+y_3 & y_3+y_4 & y_1+y_4 \\ (x_1)^2 & (x_2)^2 & (x_3)^2 & (x_4)^2 & 2x_1x_2 & 2x_2x_3 & 2x_3x_4 & 2x_1x_4 \\ x_1y_1 & x_2y_2 & x_3y_3 & x_4y_4 & \begin{matrix} x_1y_2+ \\ +x_2y_1 \end{matrix} & \begin{matrix} x_2y_3+ \\ +x_3y_2 \end{matrix} & \begin{matrix} x_3y_4+ \\ +x_4y_3 \end{matrix} & \begin{matrix} x_4y_1+ \\ +x_1y_4 \end{matrix} \\ (y_1)^2 & (y_2)^2 & (y_3)^2 & (y_4)^2 & 2y_1y_2 & 2y_2y_3 & 2y_3y_4 & 2y_1y_4 \end{bmatrix}$$

$$= \left[\begin{array}{c|c|c|c|c|c|c|c} \boldsymbol{q}_{11} & \boldsymbol{q}_{22} & \boldsymbol{q}_{33} & \boldsymbol{q}_{44} & \boldsymbol{q}_{12} & \boldsymbol{q}_{23} & \boldsymbol{q}_{34} & \boldsymbol{q}_{41} \end{array} \right];$$

$$\tilde{\boldsymbol{B}} = \begin{bmatrix} 1 & 1 & 1 & 2 & 2 & 2 \\ x_1 & x_2 & x_3 & x_1+x_2 & x_2+x_3 & x_1+x_3 \\ y_1 & y_2 & y_3 & y_1+y_2 & y_2+y_3 & y_1+y_3 \\ (x_1)^2 & (x_2)^2 & (x_3)^2 & 2x_1x_2 & 2x_2x_3 & 2x_1x_3 \\ x_1y_1 & x_2y_2 & x_3y_3 & \begin{matrix} x_1y_2+ \\ +x_2y_1 \end{matrix} & \begin{matrix} x_2y_3+ \\ +x_3y_2 \end{matrix} & \begin{matrix} x_1y_3+ \\ +x_3y_1 \end{matrix} \\ (y_1)^2 & (y_2)^2 & (y_3)^2 & 2y_1y_2 & 2y_2y_3 & 2y_1y_3 \end{bmatrix}$$

$$= \left[\begin{array}{c|c|c|c|c|c} \boldsymbol{q}_{11} & \boldsymbol{q}_{22} & \boldsymbol{q}_{33} & \boldsymbol{q}_{12} & \boldsymbol{q}_{23} & \boldsymbol{q}_{31} \end{array} \right]$$

Note that this definition matches the definition of $\boldsymbol{B}$ in (4.8). Also note that the definition of $\boldsymbol{q}_{ab}$ is an extension of the one in (3.10).

The first, second, third, fifth and sixth columns of $\boldsymbol{B}$ are exactly the first five columns of $\tilde{\boldsymbol{B}}$. Let us analyze the fourth, seventh and eighth columns.

The fourth column, of $\boldsymbol{B}$, i.e. $\boldsymbol{q}_{44}$, can be written as a linear combination of $\boldsymbol{q}_{11}$, $\boldsymbol{q}_{33}$ and $\boldsymbol{q}_{31}$. In order to show it more easily, we introduce a new notation. Recalling that $\boldsymbol{v}_4 = t\boldsymbol{v}_3 + (1-t)\boldsymbol{v}_1$, we can write

$$x_4 = tx_3 + (1-t)x_1$$
$$y_4 = ty_3 + (1-t)y_1$$

We now define $m_{a,i}$ such that

$$m_{a,i} = \begin{cases} 1 & i = 0 \\ x_a & i = 1 \\ y_a & i = 2 \end{cases} \quad \forall a = 1, \cdots, 4$$

Then, the above equations can be rewritten as

$$m_{4,i} = t m_{3,i} + (1-t) m_{1,i} \quad \forall i = 0, 1, 2$$

This means we can use this new definition to state the following:

$$\begin{aligned} m_{4,i} m_{4,j} &= (t m_{3,i} + (1-t) m_{1,i})(t m_{3,j} + (1-t) m_{1,j}) \\ &= t^2 m_{3,i} m_{3,j} + t(1-t) m_{1,i} m_{3,j} + t(1-t) m_{3,i} m_{1,j} + (1-t)^2 m_{1,i} m_{1,j} \\ &= t^2 m_{3,i} m_{3,j} + t(1-t)(m_{1,i} m_{3,j} + m_{3,i} m_{1,j}) + (1-t)^2 m_{1,i} m_{1,j} \quad \forall i, j = 0, 1, 2 \end{aligned}$$

For appropriate choices of i and j, it holds that $m_{4,i} m_{4,j}$, $m_{3,i} m_{3,j}$, $m_{1,i} m_{3,j} + m_{3,i} m_{1,j}$ and $m_{1,i} m_{1,j}$ are elements in a given position of $\boldsymbol{q}_{44}$, $\boldsymbol{q}_{33}$, $\boldsymbol{q}_{31}$ and $\boldsymbol{q}_{11}$, respectively. For example, for $i = 0, j = 2$, it holds that

$$\begin{aligned} m_{4,i} m_{4,j} = 1 \cdot y_4 = y_4 &= (t + (1-t))(t y_3 + (1-t) y_1) \\ &= t^2 y_3 + t(1-t)(y_1 + y_3) + (1-t)^2 y_1 \end{aligned}$$

and $y_4, y_3, y_1 + y_3, y_1$ are the third element of $\boldsymbol{q}_{44}, \boldsymbol{q}_{33}, \boldsymbol{q}_{31}, \boldsymbol{q}_{11}$ respectively. On the other hand, if $i = 1, j = 1$, then

$$\begin{aligned} m_{4,i} m_{4,j} = x_4 \cdot x_4 = (x_4)^2 &= (t x_3 + (1-t) x_1)(t x_3 + (1-t) x_1) \\ &= t^2 (x_3)^2 + t(1-t)(2 x_1 x_3) + (1-t)^2 (x_1)^2 \end{aligned}$$

and $(x_4)^2, (x_3)^2, 2x_1 x_3, (x_1)^2$ are the fourth element of $\boldsymbol{q}_{44}, \boldsymbol{q}_{33}, \boldsymbol{q}_{31}, \boldsymbol{q}_{11}$ respectively.

Thus, because this holds for every element of $\boldsymbol{q}_{44}$, $\boldsymbol{q}_{11}$, $\boldsymbol{q}_{33}$ and $\boldsymbol{q}_{31}$, it can be more succintly stated that

$$\boldsymbol{q}_{44} = (1-t)^2 \boldsymbol{q}_{11} + t^2 \boldsymbol{q}_{33} + t(1-t) \boldsymbol{q}_{31}$$

A similar reasoning can be applied to the other remaining columns of $\boldsymbol{B}$. The seventh column, i.e. $\boldsymbol{q}_{34}$, can similarly be proven to be a linear combination of $\boldsymbol{q}_{33}$ and $\boldsymbol{q}_{31}$. Indeed,

$$\begin{aligned} m_{3,i} m_{4,j} + m_{4,i} m_{3,j} &= m_{3,i}(t m_{3,j} + (1-t) m_{1,j}) + (t m_{3,i} + (1-t) m_{1,i}) m_{3,j} \\ &= t m_{3,i} m_{3,j} + (1-t) m_{3,i} m_{1,j} + t m_{3,i} m_{3,j} + (1-t) m_{1,i} m_{3,j} \\ &= 2t m_{3,i} m_{3,j} + (1-t)(m_{3,i} m_{1,j} + m_{1,i} m_{3,j}) \quad \forall i, j = 0, 1, 2 \end{aligned}$$

Therefore,

$$\boldsymbol{q}_{34} = 2t \boldsymbol{q}_{33} + (1-t) \boldsymbol{q}_{31}$$

Finally, something similar holds for the eighth column, i.e. $\boldsymbol{q}_{41}$, which is a linear combination of $\boldsymbol{q}_{11}$ and $\boldsymbol{q}_{31}$:

$$\begin{aligned} m_{1,i}m_{4,j} + m_{4,i}m_{1,j} &= m_{1,i}(tm_{3,j} + (1-t)m_{1,j}) + (tm_{3,i} + (1-t)m_{1,i})m_{1,j} \\ &= tm_{1,i}m_{3,j} + (1-t)m_{1,i}m_{1,j} + tm_{3,i}m_{1,j} + (1-t)m_{1,i}m_{1,j} \\ &= 2(1-t)m_{1,i}m_{1,j} + t(m_{3,i}m_{1,j} + m_{1,i}m_{3,j}) \quad \forall i,j = 0,1,2 \end{aligned}$$

Therefore,

$$\boldsymbol{q}_{41} = 2(1-t)\boldsymbol{q}_{11} + t\boldsymbol{q}_{31}$$

Thus, all columns of $\boldsymbol{B}$ can be written as a linear combination of some columns of $\tilde{\boldsymbol{B}}$. Indeed, $\boldsymbol{q}_{44}$, $\boldsymbol{q}_{34}$ and $\boldsymbol{q}_{41}$ can be written as a linear combination of $\boldsymbol{q}_{11}$, $\boldsymbol{q}_{33}$ and $\boldsymbol{q}_{31}$; whereas all other columns of $\boldsymbol{B}$, which are such that x_4 and y_4 don't appear in their construction, are exactly equal to some column of $\tilde{\boldsymbol{B}}$.

Note that, in this example, the construction would actually work either way, since it succeeds on $\tilde{P}$ $(q = 2, m = 3)$. Indeed, our construction never poses any issues on $q = 2$. This example simply serves to showcase how, in general, a weakly convex polygon would be associated to a matrix $\boldsymbol{B}$ of lower rank, which is a result that can also be observed for higher values of q and m.

If all columns of $\boldsymbol{B}$ are spanned by $\tilde{\boldsymbol{B}}$ as in this case, that means the rank of $\boldsymbol{B}$ cannot be higher than the rank of $\tilde{\boldsymbol{B}}$. Thus, if $\tilde{\boldsymbol{B}}$'s rank is lower than the dimension of the space of polynomials, e.g. if $\tilde{P}$ is such that $qm < \frac{(q+1)(q+2)}{2}$, then the same will hold for $\boldsymbol{B}$'s rank, meaning $\boldsymbol{B}$'s rows will not be linearly independent and thus the construction will fail.

To avoid this issue, this method must not be applied to weakly convex polygons with too few non-collinear vertices. In general, we can streamline this constraint by deciding to only apply this method to strongly convex and concave polygons in general (for $q \geq 3$).

Unfortunately, there is one more issue regarding $\boldsymbol{B}$'s rank: whenever $n \leq q$, if $\boldsymbol{B}$ is constructed as in (4.8) by taking $\mathscr{I} = \mathscr{R}$, then its rank is less than the number of its rows and thus our method fails. When $q = 3$, this is expected, because $qn \leq q^2 = 9 < 10 = \frac{(q+1)(q+2)}{2}$, as we have already observed. However, the same holds for higher values of q as well: when $q = 4, n = 4$, for instance, it holds that $qn = 16 > 15 = \frac{(q+1)(q+2)}{2}$, yet the construction fails anyway.

If we only work with polygons that allow for exactly qn serendipity coordinates, then getting around this issue translates into exclusively applying this method when not only $qn \geq \frac{(q+1)(q+2)}{2}$, but also $n > q$. We note that the latter implies the former:

Remark 6.2.1. *If $q \geq 2$ and $n > q$, then $qn \geq \frac{(q+1)(q+2)}{2}$.*

Proof. By induction.

When $q = 2$, it holds that $\frac{(q+1)(q+2)}{2} = 6$. Then

$$qn \geq q(q+1) = 2 \cdot 3 = 6$$

Let us now assume $q \geq 2, qn \geq \frac{(q+1)(q+2)}{2}$ $\forall n > q$. We must prove $(q+1)n \geq \frac{(q+2)(q+3)}{2}$ $\forall n > q+1$. Thus let n be such that $n > q+1$. Then, by induction hypothesis and by definition of n, we can write

$$(q+1)n = qn + n \geq \frac{(q+1)(q+2)}{2} + (q+2)$$

and thus

$$(q+1)n \geq \frac{(q+1)(q+2)}{2} + (q+2) = \frac{q^2+3q+2+2(q+2)}{2} = \frac{q^2+5q+6}{2} = \frac{(q+2)(q+3)}{2}$$

□

Therefore, if we only want to apply this method to n-gons which allow for exactly qn serendipity coordinates, we must check that $n > q$ (and that the polygon of application is not weakly convex).

However, in the previous section, we had also provided an alternative approach which extended the application of this method to the cases $q = 3, n = 3$ and $q = 4, n = 3$, by increasing the number of serendipity coordinates to be found. We employ a similar methodology here for the case $q = 4, n = 4$. Indeed, this issue can be avoided by choosing $\mathscr{I}$ such that $\mathscr{I} = \mathscr{R} \cup \{1234\}$: the full ordering of $\mathscr{A}$ is the same as usual, except with sequence 1113 swapped with sequence 1234. This generates 17 serendipity functions instead of 16, which is enough to let $\boldsymbol{B}$ have rank equal to 15, making all its rows linearly independent and letting our construction succeed. Due to this definition of $\mathscr{I}$, Corollary 5.1.2 guarantees the linear independence of the resulting serendipity coordinates, whereas (5.14a) and (5.14b) show how to build the corresponding interpolation functions. For the cases $q = 3, n \geq 4$ and $q = 4, n \geq 5$, it obviously holds that $n > q$, meaning our method succeeds as normal: therefore, this extended approach can be successfully applied to any strongly convex and concave polygons for orders $q = 3$ and $q = 4$.

6.3 Short comparison of the two approaches

Throughout this chapter, we have detailed two different approaches to the application of our method: one which only considers polygons for which our construction works with only qn serendipity coordinates, and another which extends the applicability of the method whenever possible at least for low values of q. Table 6.1 shows a brief comparison of the applicability of the two approaches and their resulting properties.

Note that it is possible to combine the two approaches to obtain a wider range of applicability, for example using approach 2 for $q \leq 4$ and approach 1 for $q > 4$.

Properties \ Approach	Approach 1	Approach 2
Methodology	Only construct serendipity coordinates when the construction succeeds for qn coordinates	Extend the number of serendipity coordinates past qn when needed, to ensure a set of linearly independent serendipity coordinates on as many polygons of a given order as possible
Order	Arbitrary	$q = 2, 3, 4$
Polygons of definition	$\begin{cases} \text{Arbitrary simple polygons} & q = 2 \\ \text{Strongly convex and concave simple } n\text{-gons such that } n > q & \text{otherwise} \end{cases}$	$\begin{cases} \text{Arbitrary simple polygons} & q = 2 \\ \text{Strongly convex and concave simple polygons} & \text{otherwise} \end{cases}$
Number of functions	qn	$\begin{cases} 10 & q = 3, n = 3 \\ 15 & q = 4, n = 3 \\ 17 & q = 4, n = 4 \\ qn & \text{otherwise} \end{cases}$
Interpolation nodes	Rigorously defined	Occasionally "virtual" (randomly picked on the interior of the polygon)

Table 6.1. Comparisons of properties of different approaches to the application of our method.

Chapter 7

Comparisons to other methods

7.1 The method by Floater and Lai

There have been other attempts to construct serendipity coordinates of different orders. One such attempt is by Floater and Lai [2016], who construct functions that can reproduce polynomials of a given order q and are the q-th Bernstein basis polynomials on the boundary of the polygon of definition.

This method does not make use of q-th coordinates, instead constructing serendipity functions directly based on products involving GBCs and standard barycentric coordinates, taking a clue from Bernstein-Bézier functions.

Given a polygon P with vertices $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ $i = 1, \cdots, n$, we define

$$\lambda_{i,-1} \; ; \; \lambda_{i,0} \; ; \; \lambda_{i,1}$$

as the three barycentric coordinates on the triangle with vertices $\boldsymbol{v}_{i-1}$, $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i+1}$. More specifically, each $\lambda_{i,j}$ is defined such that

$$\lambda_{i,j}(\boldsymbol{v}_l) = \delta_{i,l+j}$$

We also take $\phi_i, i = 1, \cdots, n$ to be the Wachspress coordinates with respect to P. We recall that they are defined as

$$\phi_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}; \;\; \omega_i(\boldsymbol{v}) = \frac{\cot \gamma_{i-1} + \cot \beta_i}{||\boldsymbol{v}_i - \boldsymbol{v}||^2}; \;\; i = 1, \cdots, n$$

If $q \geq 3$, we further write

$$M_1, \cdots, M_{\frac{(q-1)(q-2)}{2}}$$

to represent the different monomials of degree up to $q-3$, acting as a basis for the space of polynomials of degree up to $q-3$.

Finally, we also define

$$b = \prod_{i=1}^{n} h_i; \;\; h_i(\boldsymbol{v}) = (\boldsymbol{v}_i - \boldsymbol{v}) \cdot \boldsymbol{n}_i, \; i = 1, \cdots, n$$

where $\boldsymbol{n}_i$ is the outward unit normal vector to the edge $[\boldsymbol{v}_i, \boldsymbol{v}_{i+1}]$.

Then, we can define the functions

$$F_i = \phi_i(\lambda_{i,0})^{q-1};$$
$$F_{i,k} = \binom{q-1}{k}\phi_i(\lambda_{i,1})^k(\lambda_{i,0})^{q-1-k} + \binom{q-1}{k-1}\phi_{i+1}(\lambda_{i+1,0})^{k-1}(\lambda_{i+1,-1})^{q-k}$$

for $i = 1, \cdots, n; k = 1, \cdots, q-1$. Finally, it is possible to prove that the functions

$$\{F_1, \cdots, F_n, F_{1,1}, \cdots, F_{n,q-1}, \frac{b}{\sum_{j=1}^n \omega_j} M_1, \cdots, \frac{b}{\sum_{j=1}^n \omega_j} M_{\frac{(q-1)(q-2)}{2}}\}$$

are all linearly independent and span the space of polynomials of degree up to q. Not only that, but the functions F_i and $F_{i,k}$ are the q-th Bernstein basis polynomials on the boundary of $\overline{P}$, whereas the other functions in the set vanish on the boundary of $\overline{P}$. These functions can be easily defined for any value of q and it is just as easy to obtain interpolation functions from them.

For $q = 2$, the only functions in the set are of the type F_i and $F_{i,1}$, meaning there are qn serendipity functions. For $q \geq 3$, the functions obtained from the monomials also get added, for a total of $qn + \frac{(q-1)(q-2)}{2}$ serendipity functions (qn functions $F_i, F_{i,k}$ and $\frac{(q-1)(q-2)}{2}$ other functions).

The property of polynomial reproducibility can only be proven for ϕ_i being Wachspress coordinates (at least for $q \geq 3$). This binds these results to the limitations of Wachspress coordinates: namely, these serendipity functions are not defined on concave polygons. The usage of barycentric coordinates in the definition also further imposes the polygons of definition not have consecutive collinear vertices: overall, this means these functions can only be defined on strongly convex polygons.

Figure 7.1 shows some examples of functions F_i and $F_{i,k}$, as well as their corresponding interpolation functions.

7.2 The method by Cao et al.

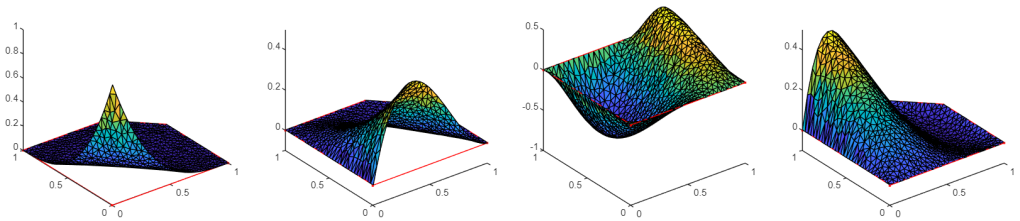
Even when employing q-th coordinates, there are different methods that can be used to reduce them to serendipity coordinates. One such method is the one detailed by Cao et al. [2022]. It focuses on the case $q = 2$ and provides an explicit form of the linear combination making up serendipity coordinates.

Let P be a polygon with vertices $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$. Any quadratic coordinate $\mu_{ij}, j \notin \{i-1, i, i+1\}$ defined with respect to P can be written as

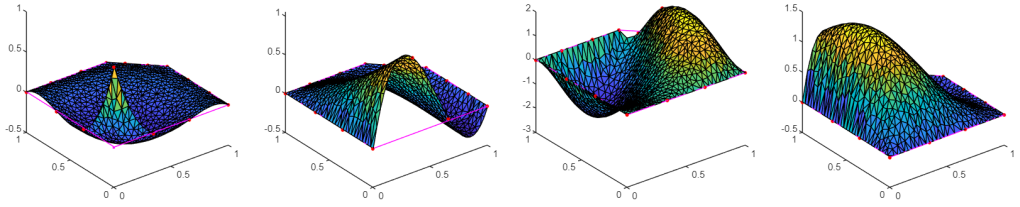
$$2\mu_{ij} = c_{ij}^{i,i}\mu_{ij} + c_{ij}^{j,j}\mu_{ij} + 2c_{ij}^{i,i-1}\mu_{ij} + 2c_{ij}^{i,i+1}\mu_{ij} + 2c_{ij}^{j,j-1}\mu_{ij} + 2c_{ij}^{j,j+1}\mu_{ij}$$

with

$$c_{ij}^{i,i} = \frac{A(\boldsymbol{v}_{i-1}, \boldsymbol{v}_j, \boldsymbol{v}_{i+1})}{A(\boldsymbol{v}_{i-1}, \boldsymbol{v}_i, \boldsymbol{v}_{i+1})};\ c_{ij}^{j,j} = \frac{A(\boldsymbol{v}_{j-1}, \boldsymbol{v}_i, \boldsymbol{v}_{j+1})}{A(\boldsymbol{v}_{j-1}, \boldsymbol{v}_j, \boldsymbol{v}_{j+1})};$$
$$c_{ij}^{i,i-1} = \frac{A(\boldsymbol{v}_j, \boldsymbol{v}_i, \boldsymbol{v}_{i+1})}{2A(\boldsymbol{v}_{i-1}, \boldsymbol{v}_i, \boldsymbol{v}_{i+1})};\ c_{ij}^{j,j-1} = \frac{A(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{v}_{j+1})}{2A(\boldsymbol{v}_{j-1}, \boldsymbol{v}_j, \boldsymbol{v}_{j+1})};$$
$$c_{ij}^{i,i+1} = \frac{A(\boldsymbol{v}_{i-1}, \boldsymbol{v}_i, \boldsymbol{v}_j)}{2A(\boldsymbol{v}_{i-1}, \boldsymbol{v}_i, \boldsymbol{v}_{i+1})};\ c_{ij}^{j,j+1} = \frac{A(\boldsymbol{v}_{j-1}, \boldsymbol{v}_j, \boldsymbol{v}_i)}{2A(\boldsymbol{v}_{j-1}, \boldsymbol{v}_j, \boldsymbol{v}_{j+1})}$$



(a) F_1 (b) $F_{1,1}$ (c) $F_{3,1}$ (d) $F_{4,2}$



(e) ν_1 (f) $\nu_{1,1}$ (g) $\nu_{3,1}$ (h) $\nu_{4,2}$

Figure 7.1. Surf plot of functions of the type F_i and $F_{i,k}$, as well as their corresponding interpolation functions ν_i and $\nu_{i,k}$, over the pentagon P with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,0.7), \boldsymbol{v}_4 = (0.7,1), \boldsymbol{v}_5 = (0,1)$, for $q = 3$. For the sake of this example, ν_i and $\nu_{i,k}$ are defined as the interpolation functions satisfying (5.2) for $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i,k} = \frac{q-k}{q}\boldsymbol{v}_i + \frac{k}{q}\boldsymbol{v}_{i+1}$.

where $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ is the *signed area* of the triangle with vertices $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$:

$$A((x_1,y_1),(x_2,y_2),(x_3,y_3)) = \frac{1}{2}\det\begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}$$

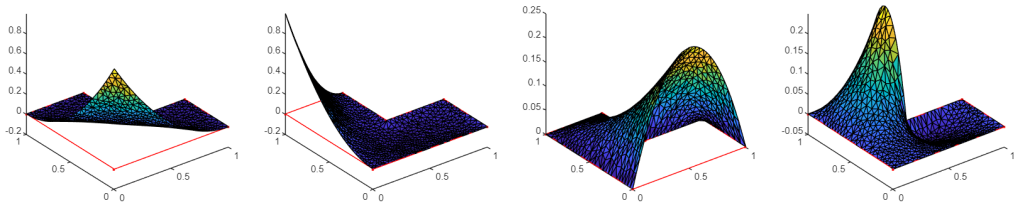
The $2n$ quadratic serendipity coordinates are then written explicitly by making use of this decomposition:

$$\xi_{ii} = \mu_{ii} + \sum_{j \notin \{i-1,i,i+1\}} c_{ij}^{i,i}\mu_{ij}$$

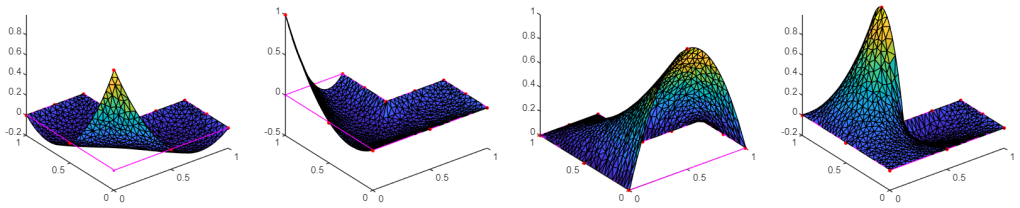
$$\xi_{i(i+1)} = \mu_{i(i+1)} + \sum_{j \notin \{i-1,i,i+1\}} c_{ij}^{i,i+1}\mu_{ij} + \sum_{j \notin \{i,i+1,i+2\}} c_{(i+1)j}^{i+1,i}\mu_{(i+1)j}$$

A result analogous to Remark 3.2.2 can be proven for these coordinates, meaning they, too, are proportional to the quadratic Bernsetin basis polynomials on the boundary of $\overline{P}$. This also implicitly shows they are all linearly independent and gives a method of obtaining interpolation functions from them. It is further possible to prove that these functions do indeed reproduce quadratic polynomials on $\overline{P}$.

This method is limited to the quadratic case, obtaining $2n$ reduced functions from $\frac{n(n+1)}{2}$ quadratic serendipity coordinates: this is the same result achieved by the method by Hackemack and Ragusa [2018], which we have worked on extending. Both methods also share their choice of GBCs of reference: while they can be applied to any set of GBCs on $\overline{P}$, they both choose mean value coordinates for their simple closed form and definability over concave polygons. However,



(a) ξ_{11} (b) ξ_{66} (c) ξ_{12} (d) ξ_{45}



(e) ν_{11} (f) ν_{66} (g) ν_{12} (h) ν_{45}

Figure 7.2. Surf plot of serendipity coordinates ξ_{ij} as obtained through the method by Cao et al. [2022], as well as their corresponding interpolation functions ν_{ij}, over the hexagon P with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1, \frac{1}{2}), \boldsymbol{v}_4 = (\frac{1}{2}, \frac{1}{2}), \boldsymbol{v}_5 = (\frac{1}{2}, 1), \boldsymbol{v}_6 = (0,1)$.

this method specifically has to sacrifice definability over weakly convex polygons, due to the use of signed areas in the definition of its coefficients.

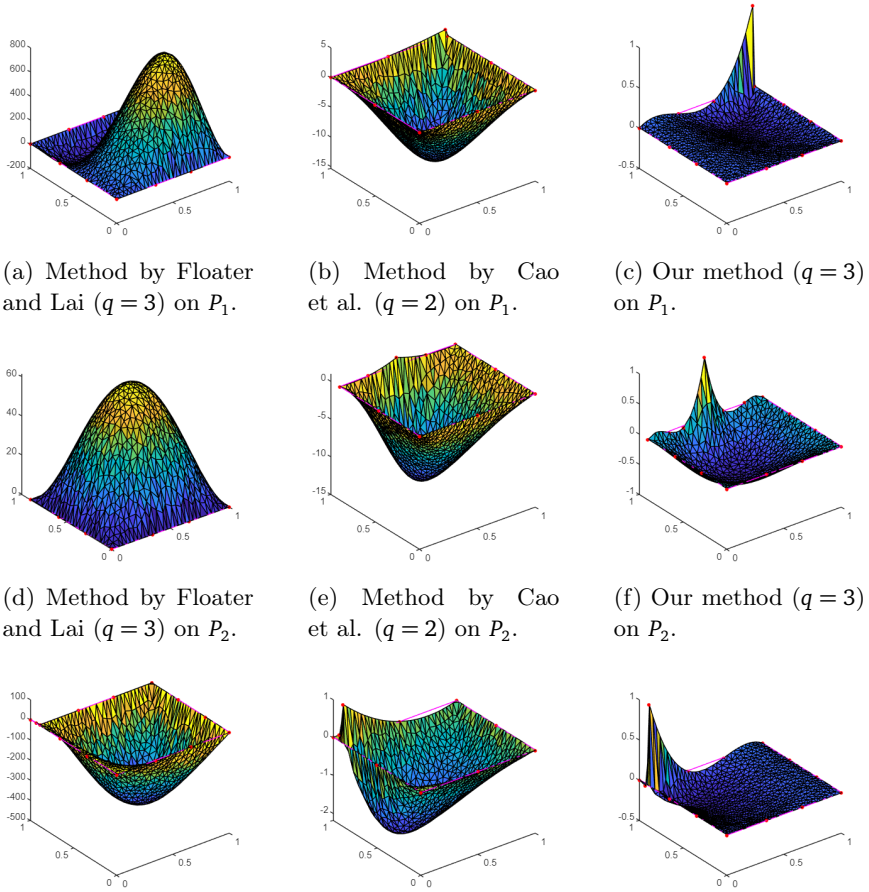
Figure 7.2 shows some examples of serendipity coordinates ξ_{ij} obtained this way, as well as their corresponding interpolation functions.

7.3 Application to edge cases

Some polygons can be very similar in shape to other polygons if they possess characteristics such as very short edges, almost flat angles or very small angles. These represent edge cases which can lead to erratic behavior when constructing functions like the ones studied in this thesis.

Figure 7.3 shows a few comparisons between interpolation functions over such edge cases, obtained through the three different methods seen so far. Indeed, both methods presented in this chapter give rise to interpolation functions whose extremums blow up in magnitude. The usage of barycentric coordinates for one method and division by signed areas for the other means that the resulting functions do not behave ideally when some of the polygon vertices form triangles that are too narrow. The method studied in this thesis, on the other hand, does not employ similar techniques, therefore its interpolation functions have a more controlled shape.

Our method is only vulnerable to edge cases if the polygon in question approaches a polygon for which our construction does not work. Figure 7.4 showcases our method being applied for $q = 3$ to similar polygons as the ones in Figure 7.3, except with one vertex each removed. This way, the shape of those polygons is approaching a triangle rather than a quadrilateral: because our construction fails for $q = 3, n = 3$, the interpolation functions obtained over those polygons



(a) Method by Floater and Lai ($q = 3$) on P_1.

(b) Method by Cao et al. ($q = 2$) on P_1.

(c) Our method ($q = 3$) on P_1.

(d) Method by Floater and Lai ($q = 3$) on P_2.

(e) Method by Cao et al. ($q = 2$) on P_2.

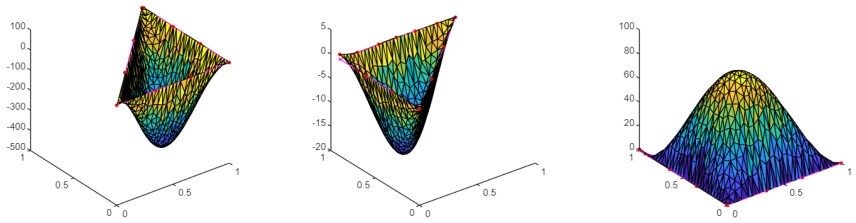
(f) Our method ($q = 3$) on P_2.

(g) Method by Floater and Lai ($q = 3$) on P_3.

(h) Method by Cao et al. ($q = 2$) on P_3.

(i) Our method ($q = 3$) on P_3.

Figure 7.3. Surf plot of interpolation functions obtained through different methods over different polygons. Each interpolation function is relative to the fourth vertex in each polygon, being equal to 1 on it and equal to 0 on all other interpolation nodes. Each polygon serves as a test for behavior in edge cases, codified through a parameter $\epsilon = 0.01$ on the definition of the polygon vertices. The polygon P_1 is a pentagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1-\epsilon), \boldsymbol{v}_4 = (1-\epsilon,1), \boldsymbol{v}_5 = (0,1)$ and serves to test behavior on short edges. The polygon P_2 is a pentagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,0.9), \boldsymbol{v}_4 = (\frac{1}{2}, 0.9+\epsilon), \boldsymbol{v}_5 = (0,0.9)$ and serves to test behavior on almost flat angles. The polygon P_3 is a pentagon with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,0.9), \boldsymbol{v}_4 = (\epsilon,0.9), \boldsymbol{v}_5 = (0,1)$ and serves to test behavior on small angles. Because P_3 is concave and Wachspress coordinates are not defined over concave polygons, (g) uses mean value coordinates in the construction of the interpolation functions, purely for illustrative purposes.



(a) P_4 (b) P_5 (c) P_6

Figure 7.4. Surf plot of interpolation functions obtained through our method over different polygons, for $q = 3$. Each interpolation function is relative to the fourth vertex in each polygon, being equal to 1 on it and equal to 0 on all other interpolation nodes. Each polygon serves as a test for behavior in edge cases, codified through a parameter $\epsilon = 0.01$ on the definition of the polygon vertices. The polygon P_4 is a quadrilateral with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (1,1-\epsilon), \boldsymbol{v}_4 = (1-\epsilon,1)$ and serves to test behavior on short edges. The polygon P_5 is a quadrilateral with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0.9), \boldsymbol{v}_3 = (\frac{1}{2}, 0.9+\epsilon), \boldsymbol{v}_4 = (0,0.9)$ and serves to test behavior on almost flat angles. The polygon P_3 is a quadrilateral with vertices $\boldsymbol{v}_1 = (0,0), \boldsymbol{v}_2 = (1,0), \boldsymbol{v}_3 = (\epsilon,0.9), \boldsymbol{v}_4 = (0,1)$ and serves to test behavior on small angles.

through our method behave erratically. Algebraically, this is the result of trying to invert a matrix whose rows are close to being linearly dependent.

7.4 Comparisons

Table 7.1 shows a comparison between the properties of the different methods. For our method, it is assumed that its application is extended for $q = 3, 4$ and restricted to qn functions for $q \geq 5$.

With $q = 2$, all methods yield exactly $2n$ functions. Of course, in this case, our method is simply the one by Hackemack and Ragusa. It is also the most widely applicable method, as it is defined over all arbitrary simple polygons, unlike the other two.

What our work provides is an extension of that method for orders higher than 2. In fact, our method is the only one which can be defined over concave polygons in such cases: the method by Cao et al. does not have a generalization to higher orders; whereas the method by Floater and Lai can be directly applied for coordinates of any order, but is limited to strongly convex polygons. Our method can be used freely for any order, provided it is applied to polygons with a sufficiently high number of vertices. Unfortunately, the number of vertices is still a pretty substantial constraint, so there is incentive to try to extend our method to smaller polygons even for $q > 2$. As detailed in the previous chapter, there are some caveats which limit this generalization, but it still seems to be possible to find a working extension of our method for $q = 3, 4$ over all strongly convex and concave simple polygons. This still renders our method significantly flexible, especially since there is not much need to investigate orders higher than 4 in ordinary computation (namely, FEM computation): losing the possibility to do so in exchange for the applicability over concave polygons is a favorable trade-off.

Another advantage of our method on higher orders lies in the construction of interpolation functions. All methods are based on the construction of qn functions with boundary behavior

Property \ Method	Our method	Floater and Lai	Cao et al.
GBCs of reference	Mean value	Wachspress	Mean value
Order	Arbitrary	Arbitrary	$q = 2$
Polygons of definition	Arbitrary simple polygons, $q = 2$; Strongly convex and concave polygons, $q = 3, 4$; Strongly convex and concave n-gons such that $n > q$, otherwise	Strongly convex polygons	Strongly convex and concave polygons
Number of functions	$\begin{cases} 10 & q = 3, n = 3 \\ 15 & q = 4, n = 3 \\ 17 & q = 4, n = 4 \\ qn & \text{otherwise} \end{cases}$	$\begin{cases} qn & q = 2 \\ qn + \frac{(q-1)(q-2)}{2} & \text{otherwise} \end{cases}$	$2n = qn$
Interpolation nodes	Occasionally "virtual"	Occasionally "virtual"	Rigorously defined
Vulnerability to edge cases	Only when approaching a polygon for which the construction is undefined	Always	Always
Smoothness	C^∞	C^∞	C^∞

Table 7.1. Comparisons of properties of different methods for obtaining serendipity coordinates.

related to Bernstein basis polynomials: this allows for an explicit definition of their corresponding interpolation nodes, by essentially dividing each edge into q equal parts. However, in order to reproduce polynomials of the needed order, the method by Floater and Lai has to complete this space of qn functions with a few additional functions. These latter functions cannot be associated to an explicit interpolation node in the same fashion as the other ones, so a "virtual interpolation node" is picked for each of them, by essentially choosing a random point on the interior of the polygon of definition. Our method does not run into this issue for polygons with a sufficiently high number of vertices, as it can reproduce polynomials of the needed order through qn functions only. Smaller polygons on higher orders still require more than qn functions, meaning they feature at least one function with a virtual interpolation node, but not having to rely on such a workaround in every case is still an advantage.

Finally, our method provides greater resistance to polygons with short edges, flat angles and small angles. The other two methods always lead to ill-behaved constructions over such polygons, whereas that is only the case for our method over polygons approaching shapes with a low number of vertices. There is still a need to generally avoid such behavior with our method, though, which may manifest into attempting to avoid such edge cases regardless of the number of vertices, rendering this advantage negligible. However, it is an advantage nonetheless and it may come in handy depending on the usage of the method.

It is worth noting that, because both mean value coordinates and Wachspress coordinates are smooth, all three methods yield serendipity coordinates that are C^∞ over their domain of definition.

Chapter 8

Future work

Some of the theoretical results in this thesis have not been properly proven: namely, the ones related to the complications of this method. The result related to linear combinations in the columns of $\boldsymbol{B}$ for weakly convex polygons has only been proven for one specific case: similar results can be shown numerically and could be similarly proven on a case-by-case basis, but a general proof would require more theoretical work. What complicates the proof is the different form taken by the elements of $\boldsymbol{B}$ depending on their associated sequence $a_1 \cdots a_q$, since $S_M(a_1 \cdots a_q)$ depends directly on $|\mathscr{P}(a_1 \cdots a_q)|$. Because of that, one approach to constructing a general proof could involve ignoring the $\frac{|\mathscr{P}(a_1 \cdots a_q)|}{q!}$ factor in $S_M(a_1 \cdots a_q)$, essentially multiplying every column of $\boldsymbol{B}$ and $\tilde{\boldsymbol{B}}$ by $\frac{q!}{|\mathscr{P}(a_1 \cdots a_q)|}$ and looking for linear combinations for those columns instead: if a linear combination exists in that case, then it exists for $\boldsymbol{B}$ and $\tilde{\boldsymbol{B}}$ as well. This is likely the theoretical result which can be proven most easily from the work already done in this thesis.

The result regarding the rank of $\boldsymbol{B}$ for $n \leq q$ has been left without a theoretical explanation as well. In that case, it is not clear why the construction would not work. It does not seem to be related to properties of mean value coordinates themselves, since $\boldsymbol{B}$ can be constructed regardless of the choice of coordinate. For $q = 4, n = 4$, the issue has been avoided by taking $\mathscr{I} = \mathscr{R} \cup \{1234\}$ instead of just $\mathscr{I} = \mathscr{R}$: numerical testing suggests that, even for higher orders, taking $\mathscr{I} = \mathscr{R} \cup \{1 \cdots q\}$ when $n = q$ results in $\boldsymbol{B}$ being full rank. It is possible that this behavior is somehow inherited from the case $q = 3$: as has already been noted, for $q = 3, n = 3$, our construction fails for $\mathscr{I} = \mathscr{R}$, because $qn = 9 < 10 = \frac{(q+1)(q+2)}{2}$, but succeeds for $\mathscr{I} = \mathscr{R} \cup \{123\}$. Even then, however, it remains unclear how this behavior could be passed on to higher values of q. Whatever theoretical explanation exists for this behavior requires more research in order to be found.

Another topic of future work concerns the use of this method for the actual solution of an FEM problem. This thesis lays the groundwork for such an application, by constructing serendipity coordinates, obtaining interpolation functions from them and detailing the cases in which their construction does and does not work. It also provides two different approaches for this method, one restricted to polygons with a large enough number of vertices and the other extended to as many polygons as possible. However, it does not put the method into practice on an actual FEM problem. This also means very little consideration has been given to the aspects of this process that relate to the solution of an FEM problem: for example, one of the two approaches requires only working with large enough polygons, but not much has been said

about how to easily obtain a mesh with only polygons of that kind. Using these functions to solve an FEM problem would allow for the exploration of such aspects, providing a complete outlook on the workings of this method.

Bibliography

J. Cao, Y. Xiao, Y. Xiao, Z. Chen, F. Xue, X. Wei, and Y. J. Zhang. Quadratic serendipity element shape functions on general planar polygons. *Computer Methods in Applied Mechanics and Engineering*, 392:114703, 2022.

M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, and W. Stuetzle. Multiresolution analysis of arbitrary meshes. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1995)*, pages 173–182, 1995.

M. S. Floater. Mean value coordinates. *Computer Aided Geometric Design*, 20(1):19–27, 2003.

M. S Floater and M.-J. Lai. Polygonal spline spaces and the numerical solution of the poisson equation. *SIAM Journal on Numerical Analysis*, 54(2):797–824, 2016.

M. S. Floater, K. Hormann, and G. Kós. A general construction of barycentric coordinates over convex polygons. *Advances in Computational Mathematics*, 24(1–4):311–331, 2006.

M. W. Hackemack and J. C. Ragusa. Quadratic serendipity discontinuous finite element discretization for s_n transport on arbitrary polygonal grids. *Journal of Computational Physics*, 374:188–212, 2018.

K. Hormann and M. S. Floater. Mean value coordinates for arbitrary planar polygons. *ACM Transactions on Graphics*, 25(4):1424–1441, 2006.

M. Meyer, A. Barr, H. Lee, and M. Desbrun. Generalized barycentric coordinates on irregular polygons. *Journal of Graphics Tools*, 7(1):13–22, 2002.

A. F. Möbius. *Der barycentrische Calcul*. J. A. Barth, Leipzig, 1827.

U. Pinkall and K. Polthier. Computing discrete minimal surfaces and their conjugates. *Experimental Mathematics*, 2(1):15–36, 1993.

A. Rand, A. Gillette, and C. Bajaj. Quadratic serendipity finite elements on polygons using generalized barycentric coordinates. *Mathematics of Computation*, 83(290):2691–2716, 2014.

E. L. Wachspress. Mathematics in science and engineering. In *A Rational Finite Element Basis*, volume 114. Academic Press, 1975.

J. Warren. Barycentric coordinates for convex polytopes. *Advances in Computational Mathematics*, 66(1):97–108, 1996.