On Established and New Semiconvexities in the Calculus of Variations

Submitted for the degree of
DOCTOR OF PHILOSOPHY

Sandra Käbisch

July 2016

Supervisor: Dr. Jonathan Bevan

© Sandra Käbisch 2016
Abstract

After introducing the topics that will be covered in this work we review important concepts from the calculus of variations in elasticity theory. Subsequently the following three topics are discussed:

The first originates from the work of Post and Sivaloganathan [Proceedings of the Royal Society of Edinburgh, Section: A Mathematics, 127(03):595–614, 1997] in the form of two scenarios involving the twisting of the outer boundary of an annulus $A$ around the inner. It seeks minimisers of $\int_A \frac{1}{2} |\nabla u|^2 \, dx$ among deformations $u$ with the constraint $\det \nabla u \geq 0$ a.e. as well as of $\int_A \frac{1}{2} |\nabla u|^2 + h(\det \nabla u) \, dx$ in which $h$ penalises volume compression so that $\det \nabla u > 0$ a.e. is imposed on minimisers. In the former case we find infinitely many explicit solutions for which $\det \nabla u = 0$ holds on a region around the inner boundary of $A$. In the latter we expand on known results by showing similar growth properties of the solutions compared to the previous case while contrasting that $\det \nabla u > 0$ holds everywhere.

In the second we introduce a new semiconvexity called $n$-polyconvexity that unifies poly- and rank-one convexity in the sense that for $f: \mathbb{R}^{d \times D} \to \mathbb{R}$ we have that $n$-polyconvexity is equivalent to polyconvexity for $n = \min\{d, D\} =: d \land D$ and equivalent to rank-one convexity for $n = 1$. For $d, D \geq 3$ we gain previously unknown semiconvexities in hierarchical order (2-polyconvexity, …, $(d \land D - 1)$-polyconvexity, weakest to strongest). We further define functions which are ‘$n$-polyaffine at $F$’ and find that they are not necessarily polyaffine for $n < d \land D$ (unlike rank-one affine functions). As one of the main results we obtain that 1-polyconvex (i.e. rank-one convex) functions $f: \mathbb{R}^{d \times D} \to \mathbb{R}$ are the pointwise supremum of 1-polyaffine functions at $F$ for every $F \in \mathbb{R}^{d \times D}$. In addition and among other things, we discuss envelopes, generalised $T_k$ configurations and relations to quasiconvexity.

The third involves a generalisation of the theory of abstract convexity which allows one to include cases like 1-polyconvex functions as the pointwise supremum of 1-polyconvex functions at $F$ for every $F$, while this is not possible within the classical theory. We review the most important results of the classical theory and present results on generalised hull operators, subgradients, conjugations and Legendre-Fenchel transforms for our new theory. In particular we obtain an operator that is reminiscent of the rank-one convexification process via lamination steps for a function. Moreover, we show that directional convexity is a special case of the generalised abstract convexity theory.

Finally, we conclude each topic, pointing out possible directions of further research.
Declaration of Originality

I confirm that the submitted work is my own work and that I have clearly identified and fully acknowledged all material that is entitled to be attributed to others (whether published or unpublished) using the referencing system set out in the programme handbook. I agree that the University may submit my work to means of checking this, such as the plagiarism detection service Turnitin® UK. I confirm that I understand that assessed work that has been shown to have been plagiarised will be penalised.

Signature ___________________________ 9th September 2016
Acknowledgements

First and foremost I would like to thank Jonathan Bevan for his advice and guidance while I undertook this work and for giving me the freedom to pursue my own interests in my research. Furthermore, his expertise and ability to quickly grasp whatever new material I threw at him in our meetings proved to be an invaluable resource for me. Without his talent to put my otherwise confused thoughts into precise and simple words I might have got lost along the way or, the worse for any reader, this thesis would be twice as long!

Furthermore, I am grateful for the financial support of this project through the University Research Scholarship and the Department of Mathematics.

I extend my thanks to past and present members of the maths department, both staff and fellow PhD students, in particular James, Jon, Gary, Dan and Esther, for making it a place that I felt comfortable working in. Additional thanks go to the many people, especially to Fernando, Matt, Vanessa, Sissi and Des, who I have met from other disciplines for the many, often highly philosophical, discussions and the sharing of experience that refreshed my mind.

Similarly, being part of the Team Surrey Volleyball Club provided a different and welcomed layer of (mostly positive) stress to my PhD life and the many friends that I made in and around the club meant that I soon felt very attached to the university.

I also wish to thank Del for being far more than a landlady, and for welcoming me into her home where I could relax, unwind and share a cup of tea.

Furthermore, I am grateful for my friends and family in Germany who, with a mixture of bewilderment and understanding for my love of mathematics, have provided me with a sense of security that made me unafraid to pursue this path.

Finally, my most heartfelt thanks go to my partner Jürgen, who bears at least half the responsibility of our choice of moving to the UK to do a PhD. He had to listen to my mathematical problems on countless occasions even though they were not in the slightest in his field of expertise. Moreover, he would calm me down when I was trying to do too many things at once or gently nudge me on when I needed a push. His continuous balancing effect on me meant that this project came to a successful end.
Contents

1. Introduction 1

2. Variational methods in elasticity theory 7
   2.1. Elasticity and hyperelastic materials 9
   2.2. Existence (of minimisers) 14
       2.2.1. The direct method of the calculus of variations 14
       2.2.2. The Euler-Lagrange equations 21
   2.3. Non-existence (of minimisers) and the formation of microstructure 24
       2.3.1. Gradient Young measures and relaxation 26

3. Annulus twist-maps 33
   3.1. General setting 33
   3.2. Minimisers without volume compression energy 35
   3.3. Minimisers with volume compression energy 46

4. \(n\)-polyconvexity 52
   4.1. Introducing \(n\)-polyconvexity through examples 53
   4.2. Definition and basic properties of \(n\)-polyconvex functions 62
       4.2.1. Proofs of Theorem 4.8 and 4.10 68
       4.2.2. \(n\)-polyaffine functions 74
       4.2.3. 1-polyconvexity as the pointwise supremum of 1-polyaffines 102
       4.2.4. \(n\)-polyconvexity as the pointwise supremum of \(n\)-polyaffines 108
   4.3. Definition and basic properties of \(n\)-polyconvex sets 111
       4.3.1. On \(T_k\) configurations 115
   4.4. Further properties of \(n\)-polyconvexity 119
       4.4.1. \(n\)-polyconvex envelopes 119
       4.4.2. \(n\)-polyconvex hulls and the envelope of the distance function 121
       4.4.3. Relations to quasiconvexity 126
       4.4.4. The quadratic case 129
       4.4.5. On the nonlocality of \(n\)-polyconvexity 134
1. Introduction

This thesis covers topics that arise from the interplay of the areas of the calculus of variations and nonlinear elasticity. Originating from continuum mechanics, elasticity theory seeks configurations of an elastic body with prescribed environmental conditions for which all forces acting on the body are balanced. The existence of such configurations is inherently difficult to prove due to the nonlinear nature of elasticity with large deformations. With the help of hyperelasticity nonlinear elasticity problems can also be formulated as problems of the minimisation of an energy among all configurations that respect the prescribed conditions. The question then is to establish existence of a minimiser which allows the application of the tools of the calculus of variations, in particular the direct method of the calculus of variations. With respect to this task, partly due to the specific requirements of elasticity problems, the convexity notions of quasi-, poly- and rank-one convexity arise, where quasiconvexity is (under certain conditions) both necessary and sufficient for the existence of minimisers. While both nonlinear elasticity and the calculus of variations have thrived under each other’s influence many open questions still remain. After providing a review of the calculus of variations with respect to elasticity theory in Chapter 2 we discuss the following three topics:

- the validity of the Euler-Lagrange equations and the uniqueness of solutions by means of studying maps that twist the outer boundary of an annulus around the inner boundary,

- a new semiconvexity notion called $n$-polyconvexity as a unifying concept of poly- and rank-one convexity which introduces new intermediary semiconvexity notions between the latter two concepts other than quasiconvexity

- a generalisation of abstract convexity (the concept that certain classes of functions can be written as the pointwise supremum of appropriate elementary functions) that allows us to treat rank-one convexity and other convexity notions that cannot yet be described with the current theory of abstract convexity.

Since each of these topics is motivated in its own way within the marked out framework we will now give a more detailed introduction for each in a separate section.
Chapter 3. In Section 2.2.2 of Chapter 2 we indicate the difficulty that a minimiser might not solve the corresponding Euler-Lagrange equations. Another difficulty that can arise from using solutions to the Euler-Lagrange equations is that the solution might not be unique and the one found may not even be a local minimiser of the problem (since the Euler-Lagrange equations only account for stationarity and not necessarily for minimality). Non-uniqueness of solutions to the Euler-Lagrange equations of elasticity problems with mixed boundary conditions is a well known phenomenon, such as in the buckling of a rod with fixed ends. However, for pure displacement boundary conditions things are not so clear. Indeed, it is still an open question whether sufficiently smooth equilibrium solutions to pure displacement boundary-value problems for homogeneous bodies with strictly polyconvex stored energy function $W$ are unique if the domain $\Omega$ is homeomorphic to a ball (see Problem 8, [11]). For sufficiently small strains F. John [25] was able to show uniqueness in the class of twice continuously differentiable functions. On the other hand, in the same paper, he motivated an example for which uniqueness for pure displacement boundary conditions cannot be expected. This example is that of twist maps on an annulus $A = \{x \in \mathbb{R}^2 : a < |x| < b\}$ (notably not homeomorphic to a ball) in two dimensions. Post and Sivaloganathan [48] rigorously prove the existence of multiple equilibrium solutions on the annulus for a functional that energetically penalises the compression of material. More precisely, they consider the minimisation of $I(u) = \int_A \frac{1}{2}|\nabla u|^2 + h(\det \nabla u) \, dx$ (where $h(d) \to +\infty$ as $d \to 0^+$) among all maps that twist the outer boundary of the annulus around the inner boundary $N$ times. This work has led Francfort and Sivaloganathan [24] to suggest the case where the volume compression is not energetically penalised (they simply minimise $I^0(u) = \int_A \frac{1}{2}|\nabla u|^2 \, dx$) but where physicality is at least somewhat retained by only allowing maps $u$ subject to the condition $\det \nabla u \geq 0$ a.e. on $A$. They argue that minimisers cannot solve the Euler-Lagrange equations and instead only satisfy the energy-momentum equations, but a thorough analysis was not carried out. We study this example in greater detail in Section 3.2 and begin with deriving the energy-momentum equations for rotationally symmetric maps, which reduce to an ordinary differential equation. Surprisingly, this ODE has infinitely many explicit solutions that all (except the identity map for zero twists) share the property that a region $H = \{x \in \mathbb{R}^2 : a < |x| < h\}$ is mapped onto the inner boundary. Therefore, the solution is degenerate on this set, i.e. it holds $\det \nabla u = 0$ on $H$, and it does not satisfy the Euler-Lagrange equations when there is a twist involved. This region also contains almost all of the twist, with at most a quarter of a twist being performed outside this set. We then go on to show that a solution to the energy-momentum equations for rotationally symmetric maps is a local minimiser for a large
class of functions with the same amount of twisting. We are able to do so by exploiting
the specific form of the explicit solutions that are available to us. Using techniques similar
to Sivaloganathan and Spector [58] we show that a rotationally symmetric map is always
desirable over a nonsymmetric map
\[ u(r, \theta) = \rho(r, \theta)(\sin(\theta + \psi(r, \theta)), \cos(\theta + \psi(r, \theta)))^T \]
of which either the twist map \( \psi \) or the radial map \( \rho \) does not depend on the angle \( \theta \).
However, whether or not minimisers of \( I^0 \) necessarily are rotationally symmetric remains
open.

In Section 3.3 we then return to the case of minimising
\[ I(u) = \int_A \frac{1}{2} |\nabla u|^2 + h(\det \nabla u) \, dx \]
which was investigated by Post and Sivaloganathan [48]. While they already prove that
rotationally symmetric minimisers solve the Euler-Lagrange equations and are twice dif-
ferentiable it is still possible to have \( \det \nabla u = 0 \) on a set of measure zero. In resemblance
with minimising \( I^0 \) we suspect that if that is the case that this would occur on the inner
boundary of the annulus. However, with slightly stronger assumptions on the function
\( h \) and using techniques of Bauman et al. [13, 14] we are able to prove that rotationally
symmetric minimisers are \( C([a, b]) \cap C^3(a, b) \) and are nondegenerate on the whole
of the annulus (in particular \( \det \nabla u(x) > 0 \) for \( |x| = a \)). We do so by proving that the
auxiliary functions \( d = \det \nabla u(x) \) and \( z = \frac{1}{2} |\nabla u|^2 + f(\det \nabla u) \) with \( f(d) = h'(d)d - h(d) \)
are respectively monotonically increasing and decreasing along the radius of the annulus.
Finally, we are able to prove a maximum principle for the function \( r \mapsto \frac{\rho(r)}{r} \), where
\( \rho(r) = |u(r, \theta)| \) and \( r \in [a, b] \).

Chapter 4. Here we introduce a semiconvexity called \( n \)-polyconvexity which unifies
poly- and rank-one convexity for (extended real-valued) functions from \( \mathbb{R}^{d \times D} \) in the
following way: When \( n = \min\{d, D\} := d \wedge D \) then \( n \)-polyconvexity is equivalent to
polyconvexity and when \( n = 1 \) then \( n \)-polyconvexity is equivalent to rank-one convexity.
Additionally one gains the new convexities for \( n = (d \wedge D - 1)\ldots 2 \) in weakening order. One
of the main motivations to study \( n \)-polyconvexity is that these intermediate concepts sit
between poly- and rank-one convexity and potentially have a relation to quasiconvexity,
which also is implied by polyconvexity and implies rank-one convexity in the finite-valued
case. Therefore, \( n \)-polyconvexity might provide new means to further our understanding
of quasiconvexity. A second reason for studying \( n \)-polyconvexity is that we found a
different way of defining \( n \)-polyaffine functions. Note that in the present theory polyaffine
and rank-one affine functions are equivalent and so one would expect that \( n \)-polyaffine
functions are equivalent to those as well. However, by making the definition dependent
on a point \( F \in \mathbb{R}^{d \times D} \), i.e. considering \( n \)-polyaffine functions at \( F \), we find that \( n \-
polyaffine functions at \( F \) are bigger classes of functions for each \( F \in \mathbb{R}^{d \times D} \) for \( n < d \wedge D \).
than polyaffine (or equivalently quasiaffine or rank-one affine) functions. In the case of \( n = d \wedge D \), we show that the class of \((d \wedge D)\)-polyaffine functions at \( F \) is the same as the class of all polyaffine functions for any \( F \in \mathbb{R}^{d \times D} \).

Motivated by the methods from convex analysis of writing finite convex functions as the pointwise supremum of affine functions we investigate whether the new definition of \( n \)-polyaffine functions at \( F \) (and in particular for \( n = 1 \), i.e. rank-one convexity) allow a similar characterisation for finite \( n \)-polyconvex functions. Note that it is not possible to write rc-but-not-pc functions as the supremum of rank-one affine functions (the abbreviations ‘rc’ and ‘pc’ stand for ‘rank-one convex’ and ‘polyconvex’ respectively). This is because rank-one affine functions are equivalent to polyaffine functions and hence the pointwise supremum of a selection of rank-one affine functions is always polyconvex. However, 1-polyaffine functions at \( F \) are not equivalent to rank-one affine (i.e. polyaffine) functions and with the additional freedom we are indeed able to establish the connection to rank-one convex functions: Any finite function \( f : \mathbb{R}^{d \times D} \rightarrow \mathbb{R} \) is rank-one convex (i.e. 1-polyconvex) if and only if it can be written as the pointwise supremum of 1-polyaffine functions at \( F \) for each point \( F \in \mathbb{R}^{d \times D} \). This closes a gap in the current literature, as for example Dacorogna states that there is no known equivalent to the characterisation of finite polyconvex functions as the pointwise supremum of polyaffine functions for rank-one convex functions [20, p. 174]. For \( 1 < n < d \wedge D \) we are not able to establish an analogous characterisation, unless we modify the notion of \( n \)-polyconvexity once more. We introduce the notion of strong \( n \)-polyconvexity, which still has the property that it unifies poly- and rank-one convexity albeit only for all finite functions. Furthermore, strong \( n \)-polyconvexity implies \( n \)-polyconvexity for any \( 1 \leq n \leq d \wedge D \). With the stronger notion we are able to prove that any finite function \( f : \mathbb{R}^{d \times D} \rightarrow \mathbb{R} \) is strongly \( n \)-polyconvex if and only if it can be written as the pointwise supremum of \( n \)-polyaffine functions at \( F \) for every \( F \in \mathbb{R}^{d \times D} \). Note that, due to our previous results, this has the consequence that in the finite case the notions of strong 1-polyconvexity, 1-polyconvexity and rank-one convexity are all equivalent.

We then consider \( n \)-polyconvex envelopes of non-\( n \)-polyconvex functions. For \( n < d \wedge D \) we can prove that the \( n \)-polyconvex envelope can be obtained in an iterative process that reminds us of the way the rank-one convex envelope of a function can be found.

Furthermore we consider \( n \)-polyconvexity for sets. Here we distinguish the intersectional and functional \( n \)-polyconvex hull of a set, which correspond to finding the smallest \( n \)-polyconvex set that contains it and all those points that cannot be separated by a finite \( n \)-polyconvex function respectively. For rank-one convexity it is known that there are sets for which the intersectional and functional rank-one convex hull differ (by more
than just boundary points). This, for instance, is the case when a set coincides with its intersectional rank-one convex hull, but contains a $T_k$ configuration. We generalise the concept of a $T_k$ configuration to suit the needs of $n$-polyconvexity and we find an example for which the intersectional and functional 2-polyconvex hulls of a set differ in $\mathbb{R}^{3\times 3}$. Furthermore, we review results for obtaining the functional semiconvex hull of a set as the zero set of the semiconvex envelope of the distance function for the semiconvexities of poly-, quasi- and rank-one convexity and conjecture a corresponding version for $n$-polyconvexity.

As one of the motives for studying $n$-polyconvexity we study its relation to quasiconvexity. We are able to conclude that quasiconvexity does not imply 2-polyconvexity for $d, D \geq 2$.

After this the next challenge we attempt is to show that the class of finite 2-polyconvex functions is strictly larger or smaller than the one of finite polyconvex or rank-one convex functions respectively in $\mathbb{R}^{3\times 3}$. (In the extended real-valued case examples of 1-pc-but-not-2-pc and 2-pc-but-not-3-pc functions are provided in Section 4.1.) For this we investigate quadratic functions. We find that Serre’s construction of a quadratic function gives a 1-pc-but-not-2-pc function. However, we argue that pursuing the same method to construct a 2-pc-but-not-3-pc function in $\mathbb{R}^{3\times 3}$ fails. Nevertheless, the approach could provide a 3-pc-but-not-4-pc example in $\mathbb{R}^{5\times 5}$ and we derive a system of inequalities that, if proven to be contradictory, asserts that such a function exists.

Finally, we show that, like polyconvexity and quasiconvexity, $n$-polyconvexity is also nonlocal for $n > 1$.

**Chapter 5** The main result of the previous chapter, namely that finite rank-one functions can be written as the pointwise supremum of 1-polyaffine functions at $F$ for each point $F \in \mathbb{R}^{d\times D}$, gives rise to this chapter in which we take this idea to an abstract setting, removing all specific structures that are not necessary for the basic concepts. In essence, this is the endeavour of the field of abstract convexity which has been set out from standard convex analysis. In abstract convexity a function is abstract convex if it can be written as the pointwise supremum of a special class of functions, the so-called elementary functions. Standard convexity and polyconvexity are special cases of abstract convexity where the elementary functions are affine and polyaffine functions respectively. However, the present form of abstract convexity does not allow us to treat our newly found correspondence between rank-one convex functions and the pointwise supremum of 1-polyaffine functions as a special case of this theory. Nevertheless, we show that it is possible to generalise the concepts of abstract convexity in order to do so. Firstly, we
present some of the existing framework of abstract convexity focusing mainly on the aspects concerning abstract convex functions rather than sets or other abstract convexities. We then go on to generalise this section by allowing the set of elementary functions to depend on the points of the domain over which the functions are defined, contrasting the results to the original work. In particular, we will obtain new notions of subdifferentials, biconjugates and Legendre-Fenchel transforms that take a weaker form than the original definitions due to their more general nature. As a result of this new generalisation of abstract convexity we show that not only rank-one convexity can be considered as a special case, but also other convexities like directional convexity (which includes both separate convexity and rank-one convexity). Returning to \(n\)-polyconvexity we look at the particular class of (extended real-valued) functions that can be written as the pointwise supremum of \(n\)-polyaffines at each point, which we call abstract \(n\)-polyconvex functions. As well as deriving basic properties of abstract \(n\)-polyconvex functions we discuss some subtle differences of the hulls and envelopes generated from abstract \(n\)-polyconvex functions and those of the previous chapter. Finally, we show that the generalisation of abstract convexity also contains directional convexity as a special case. For the elementary functions we define the previously unknown concept of strong directionally affine functions.

In Chapter 6 we conclude each of the last three chapters and point out further directions of research.
2. Variational methods in elasticity theory

Everything around us and in us is made from materials. Therefore, it is not surprising that we began studying the behaviour of materials under external or internal influences which is now known as materials science. Elasticity theory is that part of this endeavour that studies the mechanics of solid bodies that would return to their reference configuration after the force on the body that led to the deformation is removed. As such, it forms a part of continuum mechanics. Most solid materials exhibit elastic behaviour in some range of strains. For some like rubber the range is quite large, for others like steel, it is rather small. Mathematically, under the axiom of the balance of forces, we can derive partial differential equations for the displacement field that describes the motion of the body from its reference configuration to its deformed configuration. For a given set of forces it is then a natural question to ask for the existence of such a displacement field for the derived PDE. In linear elasticity we assume a linear relationship between the stress and the strain tensor, which is commonly known as Hooke’s law. The resulting PDE is then linear. The theory for the existence of solutions to those PDEs is quite vast and well established. However, Hooke’s law is only a first order approximation and so only works for relatively small strains. For example, large twists or bends in a material cannot be successfully described with linear elasticity. Therefore, nonlinear elasticity often has to be employed, which, however, comes at a price. Existence theory for nonlinear PDEs is very cumbersome and in many cases not established. One possibility to prove existence is to use the implicit function theorem. This, however, only works for small forces. If we make further assumptions on the material, in particular that the stress tensor is actually the derivative of an elastic energy of the body (such materials are called hyperelastic materials), we find that the nonlinear PDE is formally nothing but the necessary condition of the displacement field being a minimiser of this elastic energy. Hence, the problem of existence of a solution to the nonlinear PDE is transformed into the problem of existence of a minimiser of a functional. This is the objective of the calculus of variations.

The calculus of variations looks back on a history of over 300 years. The birth of the area is said to be the brachistochrone curve problem posed by Johann Bernoulli in 1696.
The brachistochrone curve is the curve that carries a point-like object from one point to another in the least amount of time assuming the only force acting on the point is gravity and the motion is frictionless. Therefore, the functional assigns a time to each curve that joins the start and end point and the goal is to minimise this functional over all feasible curves. A necessary condition of a minimiser is then that the derivative of the functional with respect to a one-parameter variation of a candidate minimiser is zero. Bernoulli used this technique to find the solution to the brachistochrone curve problem and it was later refined by both Euler and Lagrange. The resulting equations are therefore known as the Euler-Lagrange equations.

We can see that the restriction to hyperelastic materials makes that part of elasticity theory amenable to the theory of the calculus of variations. However, it is not as straightforward as just applying results of the calculus of variations to elasticity problems. Functionals typically considered in the calculus of variations up to that point were convex, a property that is incompatible with functionals encountered in nonlinear elasticity. Therefore, both areas thrived under each other’s influence. A very good review about the advances in the calculus of variations in general and its special application to materials science is [10]. Other sources of comprehensive and introductory material are [47, 43].

In Section 2.1 we introduce the basic ideas of continuum mechanics and quickly develop the way to hyperelastic materials, which form the basis of this work. Since generally the aim of hyperelastic theory is to find minimisers of the corresponding energy for a given set of conditions, we present the theory of the direct method of the calculus of variations in Section 2.2.1 in view of the special requirements of elasticity theory. Included therein is an excerpt of results regarding necessary and sufficient conditions of the existence of minimiser culminating in two existence theorems applicable in elasticity theory (Section 2.2.1) and a discussion about the Euler-Lagrange equations that, in contrast to other areas of the calculus of variations, are not yet known to hold for many elastic energies (Section 2.2.2).

In the last part of this chapter we talk about the nonexistence of minimisers, which is particularly interesting if microstructure can be observed in the material. We present a short summary of Gradient Young measures and relaxation as two techniques that provide the existence of generalised minimisers and discuss their relation to the original problem.
2.1. Elasticity and hyperelastic materials

Elasticity theory is a part of solid continuum mechanics. It describes materials that will return to their initial configuration after the stress that led to a deformation of the body is removed. Classical treatises in the nonlinear theory of elasticity include [18, 38, 2, 54]. Following [18] we denote the body in its reference state, i.e. undeformed configuration, with the region \( \Omega \subseteq \mathbb{R}^n \) when no forces are present. Given the forces acting on the body the question is what region in \( \Omega^\varphi \subseteq \mathbb{R}^n \) the body would occupy in its deformed state. The notation \( \Omega^\varphi \) comes the fact that in continuum mechanics it is assumed that there exists a mapping \( \varphi \) between the reference state and the deformed state, i.e. a material point \( x \in \Omega \) in the reference state corresponds to the material point \( x^\varphi = \varphi(x) \in \Omega^\varphi \) in the deformed state. The deformed state is in equilibrium. Then the axiom of the balance of force and the axiom of the balance of moment state that the forces and moments acting on the deformed body and the stress in the deformed body are in equilibrium. In general we distinguish two kinds of forces, body forces and surface forces. Obviously the forces need to be defined on the deformed body \( \Omega^\varphi \), i.e.

- body force \( f^\varphi : \Omega^\varphi \rightarrow \mathbb{R}^n \)
- surface force \( g^\varphi : \Gamma^\varphi_N \rightarrow \mathbb{R}^n \),

where \( \Gamma^\varphi_N \) is the part of the boundary of the deformed body on which a surface force acts.

The equations of equilibrium in the deformed configuration then read

\[
\begin{align*}
\text{div}^\varphi T^\varphi &= f^\varphi & \text{in } \Omega^\varphi \\
T^\varphi &= T^\varphi^\top & \text{in } \Omega^\varphi, \\
T^\varphi n^\varphi &= g^\varphi & \text{on } \Gamma^\varphi_N,
\end{align*}
\]

(2.1)

where \( T^\varphi \) is the Cauchy stress tensor. This system of equations, including the symmetry of the stress tensor, follows from the above mentioned axioms of force balance and moment balance (the first equation following from the balance of body forces and the third equation from a balance of surface forces). Therefore, the system as such is written in the Eulerian coordinates \( x^\varphi \) which are unknown, and thus, the system is not of much help in this form. If we were to reexpress it in the Lagrangian coordinates \( x \) of our reference configuration we can overcome this obstacle. The specific terms in (2.1)
transform as follows

\[
T(x) = \det \nabla \varphi(x) T^\varphi(x) \nabla \varphi(x)^{-\top}
\]

\[
f(x) = \det \nabla \varphi(x) f^\varphi(x)
\]

\[
g(x) = \det \nabla \varphi(x)|\nabla \varphi(x)|^{-\top} n|g^\varphi(x)|
\]

with the simple relation that

\[
\text{div } T(x) = \det \nabla \varphi(x) \text{div}^\varphi T^\varphi(x). 
\]

The tensor \(T\) is called the \textit{first Piola-Kirchhoff stress tensor}. Then the system (2.1) is in Lagrangian coordinates

\[
\begin{align*}
\text{div } T(x) &= f(x) \quad \text{in } \Omega \\
\nabla \varphi(x) T(x)^\top &= T(x) \nabla \varphi(x)^\top \quad \text{in } \Omega \\
T(x)n &= g(x) \quad \text{on } \Gamma_N.
\end{align*}
\]

These equation hold regardless of the specific material under consideration: gas, liquid or solid. Nevertheless, it is an underdetermined system since in 3D both the function \(\varphi\) (three variables) and the stress tensor \(T\) (six variables due to symmetry) are to be found as part of the solution, but only three equations given through the equation \text{div } T = f. The other six equations are provided by the information about the material, in this case an elastic solid. A material is called elastic if the stress tensor of the deformed configuration \(T^\varphi(x^\varphi)\) can be given as a function of \(x^\varphi\) and \(\nabla \varphi(x)\). We see that then (compare with (2.2)) also the stress tensor of the reference configuration \(T(x)\) is a function of \(x\) and \(\nabla \varphi(x)\) only, i.e.

\[
T(x) = \hat{T}(x, \nabla \varphi(x)).
\]

This equation is also called \textit{constitutive equation} and the function \(\hat{T}\) is called \textit{response function for the first Piola-Kirchhoff stress tensor}. The specific form of \(\hat{T}\) is usually determined by experiment. With \(\hat{T}\) known the system is no longer underdetermined. Assuming similar properties of the acting forces, i.e.

\[
\begin{align*}
f(x) &= \hat{f}(x, \varphi(x)) \\
g(x) &= \hat{g}(x, \nabla \varphi),
\end{align*}
\]
a typical boundary value problem takes the form

\[
\begin{cases}
\text{div} \hat{T}(x, \nabla \varphi(x)) = \hat{f}(x, \varphi(x)) & \text{in } \Omega \\
\hat{T}(x, \nabla \varphi(x))n(x) = \hat{g}(x, \nabla \varphi(x)) & \text{on } \Gamma_N \\
\varphi(x) = \varphi_0(x) & \text{on } \Gamma_D,
\end{cases}
\]

(2.4)

where \( \Gamma_D \) is the Dirichlet boundary and \( \Gamma_N \) the Neumann boundary with \( \Gamma_D \cup \Gamma_N = \partial \Omega \) and \( \Gamma_D \cap \Gamma_N = \emptyset \). A classical example is the deformation of a beam fixed to a wall under the forces of gravity, see Figure 2.1. In that case \( \hat{f}(x, \varphi(x)) = (0, 0, -g)^\top \) is a dead load (i.e. independent of \( \varphi \)) and \( \hat{g} = 0 \) and the Dirichlet boundary corresponds to the part where the beam is fixed to the wall.

Figure 2.1.: Elastic deformation of a beam under the forces of gravity.

Nevertheless, there are some properties that the stress tensor (and thus the response function) fulfills. Above all stands the axiom of material frame independence, meaning that if we choose to perform a rigid body transformation first and then look at the stress tensor of that transformation it would be the same as transforming the stress tensor of the original deformation. In mathematical terms

\[
\hat{T}(x, QF) = Q \hat{T}(x, F) \quad \forall F \in \mathbb{R}^{3 \times 3}_+ \text{ and } \forall Q \in SO(3).
\]

(2.5)

Other properties of the stress tensor arise from the properties of the material, i.e. if the material is homogeneous or isotropic.

However, such a boundary value problem is generally nonlinear since the response function \( \hat{T} \) normally depends nonlinearly on \( \nabla \varphi \). Proving the existence of a solution of a nonlinear partial differential equation is not straightforward. We will talk about existence in the next section in more detail, but it is exactly the following definition of a type of elasticity that allows us to employ a different method in showing existence, namely the direct method of the calculus of variations. This method relies on minimising the energy of the elastic body and in some cases the minimiser can then be established as a solution to the PDE. The type of materials for these such an energy exists are called
hyperelastic materials and their response functions $\hat{T} : \Omega \times \mathbb{R}^3_+ \to \mathbb{R}^{3 \times 3}$ are essentially the derivative of the energy, i.e. there exists a function $\hat{W} : \Omega \times \mathbb{R}^{3 \times 3}_+ \to \mathbb{R}$ which is differentiable in its second argument s.t.

$$\hat{T}(x, F) = \frac{\partial \hat{W}}{\partial F}(x, F) \quad \text{for all } x \in \Omega \text{ and } F \in \mathbb{R}^{3 \times 3}_+. \quad (2.6)$$

The function $\hat{W}$ is called the stored energy function. In the weak form the system (2.4) is

$$\int_{\Omega} \hat{T}(x, \nabla \varphi(x)) : \nabla v(x) \, dx = \int_{\Omega} \hat{f}(x, \varphi(x)) \cdot v(x) \, dx + \int_{\Gamma_N} \hat{g}(x, \nabla \varphi(x)) \cdot v(x) \, dS \quad (2.7)$$

for all sufficiently regular functions $v : \bar{\Omega} \to \mathbb{R}^3$ that vanish on $\Gamma_D$. With (2.6) it is easy to show that the left hand side of the above equation is essentially the Gâteaux-derivative of the functional

$$W(\varphi) := \int_{\Omega} \hat{W}(x, \nabla \varphi(x)) \, dx$$

so that

$$\int_{\Omega} \hat{T}(x, \nabla \varphi(x)) : \nabla v(x) \, dx = W'(\varphi)v.$$ 

Since we want to have the same properties (rewritable as Gâteaux-derivatives) for the right-hand-side of (2.7) as well we retreat to so-called conservative forces which are forces for that functionals $F$ and $G$ exist, s.t.

$$\int_{\Omega} \hat{f}(x, \varphi(x)) \cdot v \, dx = F'(\varphi)v$$

$$\int_{\Gamma_N} \hat{g}(x, \nabla \varphi(x)) \cdot v \, dS = G'(\varphi)v.$$ 

It is then possible to derive the following theorem:

**Theorem 2.1** (Theorem 4.1-1, [18]). Let a hyperelastic material occupy the domain $\Omega \subseteq \mathbb{R}^3$ in its reference configuration and be subject to conservative body forces and conservative surface forces. Then the system (2.4) is formally equivalent to

$$I'(\varphi)v = 0 \quad (2.8)$$

for all smooth enough maps $v : \bar{\Omega} \to \mathbb{R}^3$ s.t. $v|_{\Gamma_D} = 0$ and where $I$ is the functional
defined for smooth enough functions $\psi : \overline{\Omega} \to \mathbb{R}^3$ s.t.

$$I(\psi) = W(\psi) - F(\psi) - G(\psi).$$

**Remark 2.2.** By ‘formally equivalent’ we mean that the (Gâteaux) differentiability of the functionals $W$, $F$ and $G$ is assumed in order to derive equation (2.8), which in many cases is not a given property. The calculation does not work rigorously in those cases.

Note that (2.8) is a necessary condition for a minimiser of the functional $I$.

In many of the cases we will look at – that is mainly examples from rubber elasticity but also from crystallography – there will be no forces present so that $F$ and $G$ can be taken to be 0 and we are minimising the stored energy alone

$$I(\varphi) = \int_{\Omega} \widehat{W}(x, \nabla \varphi(x)) \, dx.$$  

But let us have a closer look at what kind of properties the stored energy should have. For example, the axiom of frame independence of the response function $\widehat{T}$ (see (2.5)) also reflects itself in the stored energy function in the sense that for all $x \in \overline{\Omega}$, $F \in \mathbb{R}^{3 \times 3}_+$ and $Q \in SO(3)$ we have

$$\widehat{W}(x, QF) = \widehat{W}(x, F).$$

Furthermore, if the material is *isotropic* we have also

$$\widehat{W}(x, FQ) = \widehat{W}(x, F)$$

and if the material is *homogeneous* $\widehat{W}$ does not depend on $x$. Moreover, we can assume other properties on $\widehat{W}$ that reflect physical behaviour, i.e. experimentally observed behaviour, of elastic materials, but which are very hard to implement in the nonhyperelastic case. These are that it should cost an infinite amount of energy to compress any volume within the material into a volume of zero measure or vice versa to stretch a volume to a volume of infinite measure. Mathematically, the first is expressed as

$$\widehat{W}(x, F) \to +\infty \text{ as } \det F \to 0^+$$  

(2.9)
and the second as
\[ \overline{W}(x, F) \to +\infty \text{ as } (|F| + |\text{cof } F| + \det F) \to +\infty \quad (2.10) \]

For more detail, the reader is referred to [18, Section 4.6]. However, it should be stressed that precisely the properties (2.9) and (2.10) of an elastic energy cause a lot of difficulty for the theory. Firstly, it renders the energy \( I \) nonconvex and secondly, possibly not (Gâteaux) differentiable for all arguments \( \varphi \). These points are covered in the next section.

### 2.2. Existence (of minimisers)

In Section 2.1 we motivated the typical boundary value problem of elasticity

\[
\begin{align*}
\text{div} \hat{T}(x, \nabla \varphi(x)) &= \hat{f}(x, \varphi(x)) \quad \text{in } \Omega \\
\hat{T}(x, \nabla \varphi(x)) n(x) &= \hat{g}(x, \nabla \varphi(x)) \quad \text{on } \Gamma_N \\
\varphi(x) &= \varphi_0(x) \quad \text{on } \Gamma_D.
\end{align*}
\quad (2.4)
\]

It remains to investigate whether this nonlinear system has a solution. There is the possibility of employing the implicit function theorem as described in Chapter 6 of [18], but the usage of this theorem heavily relies on regularity properties of solutions of the linearised system that are often not fulfilled for a system like the above. However, Theorem 2.1 provides us with other means of showing existence of a solution. In fact, we shift the problem of existence of a solution of a nonlinear PDE to the existence of a minimiser of the respective elastic energy \( I \). Fortunately, the existence of minimisers can be proved for a wide class of such energies \( I \) with the direct method of the calculus of variations.

#### 2.2.1. The direct method of the calculus of variations

With the direct method of the calculus variations we are seeking to find a minimiser of a functional \( I : \mathcal{A} \subseteq X \rightarrow \mathbb{R} \) where \( X \) is a function space and \( \mathcal{A} \) the set of admissible functions with \( \mathcal{A} \neq \emptyset \). Naturally we require \( I \) to be bounded below on the set \( \mathcal{A} \) so that the infimum of \( I \) on \( \mathcal{A} \) exists. The question is whether there also exists \( \overline{u} \in \mathcal{A} \) that reaches this infimum (which is then a minimum), i.e. \( I(\overline{u}) = \inf_{u \in \mathcal{A}} I(u) \). In the most general setting as this, where no structure of \( I \) is known, there is really not much we can do to answer this question. However, by definition of the infimum there exists an infimising sequence \( (u_n)_n \subseteq \mathcal{A} \), i.e. \( \lim_{n \to \infty} I(u_n) = \inf_{u \in \mathcal{A}} I(u) \). In many cases the boundedness of
the sequence \((I(u_n))_n\) will also imply the boundedness of \((u_n)_n\). If then the Banach space \(X\) is reflexive we know that \((u_n)_n\) contains a weakly convergent subsequence, which we will relabel. The weak limit \(\overline{u}\) of this weakly convergent subsequence is our candidate minimiser (assuming that it is admissible, i.e. an element of \(A\)). To guarantee that it actually is a minimiser we need to further impose conditions on \(I\). A sufficient condition that solely uses the weak convergence of the sequence \((u_n)_n\) is sequential weak lower semicontinuity, which is defined as follows:

**Definition 2.3.** Let \(X\) be a Banach space and \(I : X \rightarrow \mathbb{R}\). Then \(I\) is sequentially weakly lower semicontinuous iff

\[
I(\overline{u}) \leq \lim\inf_{u_n \rightharpoonup \overline{u}} I(u_n)
\]

for all weakly convergent sequences \(u_n \rightharpoonup \overline{u}\).

This condition then ensures that \(I(\overline{u})\) is bounded above and below by the infimum of \(I\) and, hence, \(I(\overline{u})\) must be equal to the infimum of \(I\) and so \(\overline{u}\) is a minimiser.

In general in the calculus of variations, the functional \(I\) is of an integral type like we have seen in the previous section, i.e.

\[
I(u) = \int_{\Omega} f(x,[u]) \, dx
\]

(2.11)

where \(\Omega\) is a region in \(\mathbb{R}^n\) and \([u]\) denotes all derivatives of \(u\). In Section 2.1 we had \(f(x,[u]) = \hat{W}(x,\nabla u)\). The two requirements of the methods, namely that some form of \(I\) implies that the infimising sequence is bounded and that \(I\) is sequentially weakly lower semicontinuous are easily assumed, but indeed are two very delicate matters. In fact, the two conditions are properties the specific integrand \(f\) passes on to \(I\), so the question shifts to what conditions we can impose on \(f\) so that \(I\) is sequentially weakly lower semicontinuous and implies the boundedness of its argument. This and more is addressed in detail the book by Dacorogna [20]. Here we want to give a short review of the essential ideas. We start with the discussion of sequential weak lower semicontinuity as it is the vaster topic. Just before being able to state existence theorems (that are applicable in elasticity) we also comment on the boundedness of the minimising sequence.

We now address the question which functionals are (sequentially) weakly lower semicontinuous. For example, any integral functional with a finite and convex integrand is (sequentially) weakly lower semicontinuous. However, in elasticity, the free energy is usually a nonconvex functional, which comes from the fact that the stored energy function \(\hat{W}\) is nonconvex. Therefore, we need to find other convexity conditions of the
The integrand to ensure that the functional is sequentially weakly lower semicontinuous. The first treatise on necessary conditions of sequentially lower semicontinuity of functional $I$ of the form (2.11) up to the first derivative is that of Morrey from 1952 [42]. He derives quasiconvexity as a necessary condition for sequential weak lower semicontinuity, which is weaker than convexity in the sense that every convex function is also quasiconvex. Morrey introduced quasiconvexity only for finitely valued functions. Yet, in elasticity we often assign the value $+\infty$ to functions that do not satisfy constraints like $\det \nabla u > 0$ and so his definition is not directly applicable in that case. Ball and Murat [9] generalised the notion of Morrey’s quasiconvexity and obtained the following definition:

**Definition 2.4** (Quasiconvexity). Let $U \subseteq \mathbb{R}^{n \times m}$ and $g : U \to \mathbb{R} \cup \{+\infty\}$ be Borel-measurable and bounded below and let $1 \leq p \leq \infty$. Then $g$ is $W^{1,p}$-quasiconvex at $F \in U$ iff

$$\int_D g(F + \nabla \varphi(x)) \, dx \geq g(F) \text{ meas}(D)$$

for every bounded open subset $D \subseteq \mathbb{R}^m$ and $\varphi \in W^{1,p}_0(D)$ with $F + \nabla \varphi(x) \in U$ for all $x \in D$. $g$ is called quasiconvex on $U$ if $g$ is quasiconvex at each $F \in U$.

Morrey’s definition is essentially the above one in the case of $p = \infty$ and for $g$ only taking values in $\mathbb{R}$ rather than on the extended real valued line $\mathbb{R} \cup \{+\infty\}$. Note that $p = \infty$ is also the weakest condition, i.e. $W^{1,p}$-quasiconvexity implies $W^{1,q}$-quasiconvexity for all $1 \leq p \leq q \leq \infty$ and that in many cases it is sufficient to look at $W^{1,\infty}$-quasiconvexity. Now, for $g : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{+\infty\}$, Ball and Murat prove the following theorem:

**Theorem 2.5** (Corollary 3.2,[9]). Let $\Omega \subseteq \mathbb{R}^m$ be nonempty and bounded. Define

$$I(u) = \int_{\Omega} g(\nabla u(x)) \, dx.$$ 

If $I$ is sequentially weakly lower semicontinuous on $W^{1,p}(\Omega; \mathbb{R}^n)$, then $g$ is $W^{1,p}$-quasiconvex. (If $p = \infty$ we consider weak*-convergence rather than weak convergence, since $W^{1,\infty}(\Omega; \mathbb{R}^n)$ is not reflexive.)

Yet, this theorem only shows that quasiconvexity is a necessary condition for sequential weak lower semicontinuity. What we really want is a condition on the integrand that is sufficient for sequential weak lower semicontinuity. Unfortunately, it was only possible to prove that quasiconvexity is also sufficient for sequential weak lower semicontinuity if the integrand takes only finite values. Nevertheless, to that end, we still need to impose certain growth conditions on the integrand to make it work.
Definition 2.6 (Definition 8.10, [20]). Let $1 \leq p < \infty$ and $\Omega \subseteq \mathbb{R}^m$ be a bounded open set. Let $f : \Omega \times \mathbb{R}^n \times \mathbb{R}^{n \times m} \to \mathbb{R}$ be a Carathéodory function. The function $f$ satisfies the growth condition $(C_p)$ if for almost every $x \in \Omega$ and for every $(u, F) \in \mathbb{R}^n \times \mathbb{R}^{n \times m}$ the inequalities

$$-\alpha(|u|^q + |F|^r) - \beta(x) \leq f(x, u, F) \leq g(x, u)(1 + |F|^p)$$

hold, where $\alpha, \beta, g \geq 0$, $\beta \in L^1(\Omega)$, $1 \leq q < p$, $1 \leq r < np/(n - p)$ if $p < n$ and $1 \leq r < \infty$ if $p \geq n$ and $g$ is a Carathéodory function. In the case $p = 1$ we assume

$$|f(x, u, F)| \leq \alpha(1 + |F|).$$

Then the following theorem holds:

Theorem 2.7 (Theorem 8.11, [20]). Let $\Omega \subseteq \mathbb{R}^m$ be a bounded open set with Lipschitz boundary. Let $f : \Omega \times \mathbb{R}^n \times \mathbb{R}^{n \times m} \to \mathbb{R}$ be a Carathéodory function such that $F \to f(x, u, F)$ is $W^{1,\infty}$-quasiconvex. Let $1 \leq p < \infty$ and assume that $f$ satisfies the growth condition $(C_p)$. Let

$$I(u) = \int_\Omega f(x, u(x), \nabla u(x)) \, dx,$$

then $I$ is (sequentially) weakly lower semicontinuous in $W^{1,p}(\Omega; \mathbb{R}^m)$.

This theorem was actually also first proved under stronger growth assumptions by Morrey when he introduced quasiconvexity [42]. However, the growth conditions on the integrand $f$ were too restrictive to be applicable in elasticity to allow for conditions like (2.9). The result presented here was refined by Acerbi and Fusco [1]. There is also a version of this theorem for weak* lower semicontinuity in $W^{1,\infty}(\Omega; \mathbb{R}^n)$ with similar but different growth conditions, see Theorem 8.8 [20, p. 378].

Now we are in a position to formulate a first existence theorem. For this task we merely have to add to the previous result, which ensures sequential weak lower semicontinuity of the functional $I$, the existence of a weakly convergent subsequence of a minimising sequence. This can be achieved by strengthening the growth condition of the lower bound on the stored-energy $f$. Furthermore, we assume from now on that the minimisation problem has pure displacement boundary conditions, i.e. the set of admissible functions is a subset of $W^{1,p}_0(\Omega) = \{ u \in W^{1,p}(\Omega) : u - v \in W^{1,p}_0(\Omega) \}$ for some $v \in W^{1,p}(\Omega)$ (this definition implicitly requires that $\Omega$ has a Lipschitz boundary). This assumption does not play a major role, but most of the problems discussed in this work will be of that
Theorem 2.8 (Theorem 8.29,[20]). Let $p > 1$, $\Omega \subseteq \mathbb{R}^m$ be a bounded open set with a Lipschitz boundary. Let $f : \Omega \times \mathbb{R}^n \times \mathbb{R}^{n \times m} \to \mathbb{R}$, $f = f(x,u,F)$, be a Carathéodory function satisfying for almost every $x \in \Omega$, for every $u \in \mathbb{R}^n$

\[
f(x,u,\cdot) \text{ is quasiconvex,}
\]

with

\[
\alpha_1 |F|^p + \beta_1 |u|^q + \gamma_1(x) \leq f(x,u,F) \leq \alpha_2 |F|^p + \beta_2 |u|^r + \gamma_2(x),
\]

where $\alpha_2 \geq \alpha_1 > 0$, $\beta_1 \in \mathbb{R}$, $\beta_2 \geq 0$, $\gamma_1, \gamma_2 \in L^1(\Omega)$, $p > q \geq 1$ and $1 \leq r \leq np/(n - p)$ if $p < n$ and $1 \leq r < \infty$ if $p \geq n$. Let $A \subseteq W_{e}^{1,p}(\Omega;\mathbb{R}^n)$ be weakly closed. Then $I$ defined as in (2.12) admits at least one minimiser $\bar{u} \in A$, i.e.

\[
I(\bar{u}) = \inf_{u \in A} I(u).
\]

We get weak convergence in the theorem since we bound the gradient in $L^p$ using the term $\alpha_1 |F|^p$. Using Poincaré’s inequality we bound the minimising sequence $u_n$ in $W^{1,p}(\Omega)$ which gives the weakly convergent subsequence. The upper bound of $f$ is merely introduced to ensure that there exists $u \in A$ s.t. $I(u) < +\infty$. Then applying sequential weak lower semicontinuity yields that the weak limit is a minimiser.

Note that an important point in the above presentation is that the stored energy function $f$ only assumes finite values as it is only then possible to show that quasiconvexity of $f$ implies sequential weak lower semicontinuity of $I$ with the correct growth conditions; there exists no corresponding result in the extended real valued case. In fact, little is known about quasiconvexity in the extended real valued case, as will be made clear in Appendix A.2. Additionally, the finite-valuedness of $f$ poses quite a restriction to the applicability of this theorem to elasticity theory as many stored-energy functions have the volume compression property that $f(x,u,F) \to \infty$ as $\det F \to 0^+$ (see (2.9)). In those cases it is necessary to define $f(x,u,F) = \infty$ if $\det F \leq 0$ to exclude any unphysical behaviour. Putting the condition $\det \nabla u > 0$ a.e. in $\Omega$ into the set of admissible function $A$ is no option, since it would not be weakly closed. To that end, based on an observation of Morrey [42], Ball [4] introduced another type of convexity called polyconvexity in 1977, which itself lies between convexity and quasiconvexity (please refer to Appendix A.2 for more information). With polyconvexity we are able to find an existence theorem for the
case above. The idea of it comes from another main result about sequential weak lower semicontinuity:

**Theorem 2.9** (Theorem 3.23, [20]). Let $\Omega$ be an open subset of $\mathbb{R}^n$ and $p, q \geq 1$. Let $f : \Omega \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a Carathéodory function satisfying

$$f(x, u, X) \geq \langle a(x), X \rangle + b(x) + c|u|^p$$

for almost every $x \in \Omega$, for every $(u, X) \in \mathbb{R}^m \times \mathbb{R}^k$, for some $a \in L^q'(\Omega; \mathbb{R}^k)$, $1/q + 1/q' = 1$, $b \in L^1(\Omega)$, $c \in \mathbb{R}$ and where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^k$. Let

$$J(u, X) := \int_{\Omega} f(x, u(x), X(x)) \, dx.$$ 

Assume that $X \mapsto f(x, u, X)$ is convex and that

$$u_n \to u \quad \text{in } L^p(\Omega; \mathbb{R}^m) \quad \text{and} \quad X_n \rightharpoonup X \quad \text{in } L^q(\Omega; \mathbb{R}^k).$$

Then

$$\liminf_{n \to \infty} J(u_n, X_n) \geq J(u, X). \tag{2.13}$$

The main feature about this theorem is that $u$ and $X$ are not necessarily related. As long as $u_n$ converges strongly in some $L^p(\Omega; \mathbb{R}^m)$, the convexity in the weakly converging argument $X$ is sufficient to ensure (2.13). This opens the door to an existence theory for functionals with stored-energy functions of the kind with (2.9) and (2.10), since we could use $X = (\nabla u, \text{cof} \nabla u, \text{det} \nabla u)$ as long as we ensure weak convergence of all minors in some $L^q$ and convexity in of $f(x, u, X)$ in $X$. More generally, polyconvexity is defined as follows:

**Definition 2.10.** Let $f : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{+\infty\}$. Then $f$ is called polyconvex iff there exists $g : \mathbb{R}^{\tau(n,m)} \to \mathbb{R} \cup \{+\infty\}$ convex, s.t.

$$f(F) = g(T(F)),$$

where $T : \mathbb{R}^{n \times m} \to \mathbb{R}^{\tau(n,m)}$ maps $F$ to all its minors, i.e.

$$T(F) = (F, \text{adj}_2 F, \ldots, \text{adj}_{n \wedge m} F)$$
with $\tau(n, m) = \sum_{s=1}^{n \wedge m} \binom{n}{s} \binom{m}{s}$.

**Remark 2.11.** In the case of $n = m = 2$ this translates to

$$f(F) = g(F, \det F)$$

and for $n = m = 3$ to

$$f(F) = g(F, \text{cof} F, \det F).$$

By analogy with the finite valued case we now obtain a sufficient condition for sequential weak lower semicontinuity:

**Theorem 2.12** (Theorem 8.16, [20]). Let $\Omega \subseteq \mathbb{R}^n$ be a bounded open set with Lipschitz boundary and $p > \min(n, m)$. Let $f : \Omega \times \mathbb{R}^n \times \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{+\infty\}$ be a Carathéodory function s.t. for almost every $x \in \Omega$ and every $u \in \mathbb{R}^n$, $f(x, u, \cdot)$ is polyconvex with

$$f(x, u, F) = g(x, u, T(F))$$

s.t.

$$g(x, u, X) \geq \langle a(x), X \rangle + b(x) + c|u|^r$$

where $a \in L^p' (\Omega; \mathbb{R}^{n \times m})$, $\frac{1}{p} + \frac{1}{p'} = 1$, $b \in L^1(\Omega)$, $1 \leq r < \frac{np}{n-p}$ if $p < n$ and $1 \leq r < \infty$ if $p \geq n$ and $c \in \mathbb{R}$. Then $I$ as defined in (2.12)

$$I(u) = \int_{\Omega} f(x, u(x), \nabla u(x)) \, dx$$

is (sequentially) weakly lower semicontinuous in $W^{1,p}(\Omega, \mathbb{R}^n)$.

It should be noted that the proof of this theorem heavily relies on the weak convergence of the sequence of $T(\nabla u_n)$ in some $L^q(\Omega)$ for $q \geq 1$ for the weakly convergent sequence $u_n$ in $W^{1,p}(\Omega; \mathbb{R}^n)$ in question. This is the only reason for the assumptions $p > \min(n, m)$, since then the weak convergence of $u_n$ in $W^{1,p}(\Omega; \mathbb{R}^n)$ automatically provides the weak convergence of $T(\nabla u_n)$ in some $L^q(\Omega)$.

With this information at hand an existence theorem for the extended real valued case reads:

**Theorem 2.13** (Theorem 8.31, [20]). Let $\Omega \subseteq \mathbb{R}^n$ be a bounded open set with a Lipschitz boundary, $p > \min(n, m)$. Let $f : \Omega \times \mathbb{R}^n \times \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{+\infty\}$ be a Carathéodory
function s.t. for almost every $x \in \Omega$ and every $u \in \mathbb{R}^n$ \( f(x, u, \cdot) \) is polyconvex with
\[
 f(x, u, F) \geq a(x) + b_1|F|^p, \tag{2.14}
\]
where $a \in L^1(\Omega)$ and $b_1 > 0$. Assume that $\mathcal{A} \subseteq W^{1,p}_v(\Omega)$ is weakly closed and there exists $u_0 \in \mathcal{A}$ s.t. $I(u_0) = \int_\Omega f(x, u_0(x), \nabla u_0(x)) \, dx < \infty$, then
\[
 I(\bar{u}) = \inf_{u \in A} I(u) \tag{2.8}
\]
admits at least one solution $\bar{u} \in \mathcal{A}$.

Remark 2.14. The growth condition \[2.14\] only ensures that the minimising sequence is bounded in $W^{1,p}(\Omega; \mathbb{R}^n)$, and hence contains a weakly convergent subsequence. Then with $p > \min(n, m)$ we find $T(\nabla u_n)$ weakly convergent in some $L^q(\Omega)$. However, we note that other growth conditions can be imposed which lead to the types of weak convergences needed in Theorem \[2.9\], see [20, Remark 8.32(iii)].

2.2.2. The Euler-Lagrange equations

Provided that the question about the existence of a minimiser $\bar{u}$ of a problem
\[
 I(\bar{u}) = \inf_{u \in A} I(u) \tag{2.8}
\]
with a functional of the form
\[
 I(u) = \int_\Omega f(x, u, \nabla u) \, dx \tag{2.12}
\]
could be answered positively, it immediately poses that of finding it. One approach is to look at the first order optimality condition of the functional $I$ at the minimiser $u$ with respect to a one-parameter group of variations. To be more precise, the variations
\[
 u_\varepsilon(x) = u(x) + \varepsilon \varphi(x) \tag{2.15}
\]
lead to the well-known Euler-Lagrange-equations by formally taking the limit
\[
 I'(u)\varphi = \lim_{\varepsilon \to 0} \frac{I(u_\varepsilon) - I(u)}{\varepsilon} = \lim_{\varepsilon \to 0} \int_\Omega (f(x, u_\varepsilon(x), \nabla u_\varepsilon(x)) - f(x, u(x), \nabla u(x))) \, dx = 0 \tag{2.16}
\]
and by setting it to zero as a necessary condition of a minimum. By also taking account of any boundary conditions the minimiser must fulfil, e.g. \( u = u_0 \) on \( \partial \Omega \) in the case of pure displacement boundary conditions, eq. (2.16) has to hold for all \( \varphi \) which preserve the boundary condition, i.e. \( \varphi = 0 \) on \( \partial \Omega \). This yields

\[
\int_{\Omega} f_u(x, u, \nabla u) \cdot \varphi + f_F(x, u, \nabla u) : \nabla \varphi \, dx = 0,
\]

(2.17)

for all \( \varphi \in C^\infty(\mathbb{R}^n) \) with \( \varphi = 0 \) on \( \partial \Omega \) and where the subscripts \( f_F \) and \( f_u \) denote for the partial derivatives \( (f_F)_{ij} = \frac{\partial f}{\partial F_{ij}} \) and \( (f_u)_i = \frac{\partial f}{\partial u_i} \). In the strong form this equation becomes

\[
\text{div}(f_F(x, u, \nabla u)) = f_u(x, u, \nabla u) \quad \text{in } \Omega.
\]

With the appropriate boundary conditions this coincides with equations (2.3) from Section 2.1.

In the history of the calculus of variations the Euler-Lagrange equations have played a very important role in finding the minimisers. Furthermore, in the one-dimensional setting and subject to additional conditions, it can be shown that any solution to the Euler-Lagrange equations is indeed a local minimiser. This is known as Weierstraß’ Sufficiency Theorem, see for example [10]. However, no field theory exists in higher dimensions, so these methods do not apply in particular to elasticity theory. Moreover, in the motivation of the (weak form of the) Euler-Lagrange equations in eq. (2.16) we were so far only formally taking the limit, i.e. we were not paying attention to the convergence of the latter. Potentially that can be done by applying the Dominated Convergence Theorem, but it is necessary that the integrand \( 1/\varepsilon[f(\cdot, u_{\varepsilon}(\cdot), \nabla u_{\varepsilon}(\cdot)) - f(\cdot, u(\cdot), \nabla u(\cdot))] \) is pointwise bounded by an integrable function for all \( \varepsilon \) small enough. Unfortunately, in elasticity, it is exactly this that causes problems. In the compressible case we assume the stored energy function is of the form \( f(x, u, F) = \hat{W}(x, F) \) and usually has the properties (2.9) and (2.10). Eq. (2.9) ensures that the solution satisfies \( \det \nabla u > 0 \) a.e. in the domain. In effect, it would be nice that these two properties actually imply that the minimiser \( u \) is an element of \( W^{1,\infty}(\Omega) \) with

\[
\det \nabla u \geq \delta > 0
\]

(2.18)

for some \( \delta > 0 \). Then for small enough \( \varepsilon \) all \( \det \nabla u_{\varepsilon} \) are bounded away from zero almost everywhere as well. Assuming that the stored energy \( W \) is differentiable we then find that the difference quotient is pointwise bounded independently of \( \varepsilon \). However, it is
still an open problem if suitable growth conditions can imply (2.18) or \( \bar{u} \in W^{1,\infty}(\Omega) \) for global or local minimisers, see for example [11]. Hence, it is also an open problem whether minimisers satisfy (the weak form of) the Euler-Lagrange equations.

Nevertheless, as announced by Ball [5] in 1983, other forms of stationarity can be derived that have to be satisfied by the minimiser. Instead of the variation of the kind (2.15) he uses the variation

\[
\tilde{u}_\varepsilon(x) = u(x + \varepsilon \varphi(x)).
\]  

(2.19)

To distinguish, variations of the form of (2.15) are called outer variations and of (2.19) inner variations. In contrast to outer variations, for inner variations the determinant is better behaved, since

\[
\det \nabla \tilde{u}_\varepsilon(x) = \det ((\nabla u(x + \varepsilon \varphi(x)))(I + \varepsilon \nabla \varphi(x)))
\]

\[
= \det \nabla u(x + \varepsilon \varphi(x)) \det(I + \varepsilon \nabla \varphi(x)).
\]

This is the reason why it is possible that some growth conditions on the stored energy \( W \) ensure the applicability of the dominated convergence theorem. In this case the Euler-Lagrange equations for those variations take a different form, namely

\[
\int_\Omega [\tilde{W}(\nabla u)I - (\nabla u)^T \tilde{W}_F(\nabla u)] : \nabla \varphi \, dx = 0
\]

(2.20)

or in its strong form

\[
\text{div}(\tilde{W}(\nabla u)I - (\nabla u)^T \tilde{W}_F(\nabla u)) = 0
\]

which are commonly referred to as the energy-momentum equations. However, properly checking the differentiability of \( I(\tilde{u}_\varepsilon) \) with respect to \( \varepsilon \) at \( \varepsilon = 0 \) is very involved. To this end Ball [5] provides conditions for \( \tilde{W} \) that, if satisfied, imply the validity of (2.20) for all \( \varphi \in C_c^1(\Omega; \mathbb{R}^n) \).

**Theorem 2.15** (Ball). Let \( 1 \leq p < \infty \), \( \Omega \subseteq \mathbb{R}^n \) and \( u \in A \subseteq W^{1,p}_* (\Omega; \mathbb{R}^n) \) for some \( v \in W^{1,p}(\Omega; \mathbb{R}^n) \) be a \( W^{1,p} \) local minimiser of \( I \) in \( A \) with \( I(u) = \int_\Omega \tilde{W}(x, \nabla u(x)) \, dx \). Let \( \tilde{W} \) satisfy

(i) \( \tilde{W} \geq 0 \), \( \tilde{W}(A) = +\infty \) if \( \det A \leq 0 \) and \( \tilde{W}(A) \to +\infty \) as \( \det A \to 0^+ \),

(ii) \( \tilde{W} \in C^1(\mathbb{R}^{n \times n}_+) \),
(iii) there exist $\theta > 0$ and $C > 0$ such that

$$|A^T \hat{W}(AB)| \leq C(\hat{W}(A) + 1)$$

(2.21)

for all $A, B \in \mathbb{R}^{n \times n}_+$ with $|B - I| \leq \theta$.

where $\mathbb{R}_+^{n \times n} = \{ A \in \mathbb{R}^{n \times n} : \det A > 0 \}$. Then (2.20) holds for all $\varphi \in C^1_c(\Omega; \mathbb{R}^n)$.

Details of the proof for $n = 2$ can be found in [13] or for $n = 3$ in [14]. Note that the assumptions on $\hat{W}$ in (i) are in line with those normally imposed for elasticity problems (see (2.9)). As mentioned before, those assumptions also cause problems in rigorously deriving the validity of the Euler-Lagrange equations (2.17) for $f(x, u, F) = \hat{W}(x, F)$. Nevertheless, Bauman, Owen & Philips [14] show that sufficiently regular solutions to the energy-momentum equations also solve the Euler-Lagrange equations.

### 2.3. Non-existence (of minimisers) and the formation of microstructure

In the previous sections we have discussed the existence of minimisers of a functional $I$ by applying the direct method of the calculus of variations. The method relies on the sequential weak lower semicontinuity of the functional $I$, which led to the conditions involving quasi- and polyconvexity. In conjunction with appropriate growth conditions on the integrand of the functional it was then possible to establish the existence of a minimiser. However, in some cases, the elastic energies do not display even the necessary condition of quasiconvexity or polyconvexity. In those cases, the direct method is not applicable and it is easy to construct an example where the infimum is not attained, i.e. no minimiser exists. For example, in a one-dimensional problem, the integrand of the functional

$$I(u) = \int_0^1 (u_x - a)^2(u_x + b)^2 + u^2 \, dx$$

for $a, b > 0$ is not quasiconvex, since in 1D (as discussed in Appendix A.2), quasiconvexity is equivalent to convexity – and $(u_x - a)^2(u_x + b)^2$ is clearly not convex in $u_x$. Let us consider the problem of infimising $I(u)$ among all maps in $W^{1,4}_0(0,1)$. Since $I(u) \geq 0$ for all $u \in W^{1,4}_0(0,1)$ and since $I(0) < \infty$, the infimum exists. Both terms are minimal if either $u_x = a$ or $u_x = -b$ on the one hand and $u = 0$ on the other. However, it is obvious that the two requirements are contradictory. Nevertheless, the infimum is 0 and
can be approached by a finer and finer oscillation of the gradients \( u_x = a \) and \( u_x = -b \) as depicted in Figure 2.2. At first glance this example may not seem very relevant for the study of materials, but in fact, it does provide a possible explanation of the occurrence of microstructure as it is often observed in alloys, as can be seen in Figure 2.3. Indeed,

![Image of infimising sequence for I.](image1)

Figure 2.2.: Illustration of an infimising sequence for \( I \).

![Image of microstructure in an alloy.](image2)

Figure 2.3.: Microstructure in an alloy of Copper, Aluminium and Nickel. Courtesy of C. Chu and R. D. James, [17].

the methods from the calculus of variation applied to nonconvex energies have proven to be a very successful tools in the prediction of microstructures in materials. Even though microstructures in the ‘real world’ do not have infinitely small oscillations (they are usually of atomic length scales) we can still interpret them as approximations to the
oscillations created by the infimising sequence sketched for the example above.

Note that the weak limit of this sequence, which is the function \( u = 0 \), no longer carries information of the oscillating behaviour of its generating sequence and that \( I(0) > 0 \) (which states that it cannot be the minimiser). However, there is an object corresponding to a limit in a more general sense which would accurately describe the situation that the infimum can be reached by having infinitesimal oscillations of the gradients \( a \) and \( -b \) of with a ratios of \( \frac{b}{a+b} \) and \( \frac{a}{a+b} \). This generalisation was first introduced by L. C. Young in the context of optimal control along with other interesting examples [61]. Since then the theory of those generalised solutions has come a long way. In honour of L. C. Young, these solutions are now called gradient Young measures. A very detailed review about gradient Young measures and their role in the development of microstructure is given in the lecture notes of S. Müller [43].

In the next section we introduce gradient Young measures and present their relation as generalised solutions to the original minimising problem that has no solution. If one is, however, prepared to lose information about the microstructure of the material and is solely interested in its macroscopical behaviour, then another possibility opens up, which is known as relaxation. The relaxed problem will then have the property that minimising sequences of the original problem are also minimising sequences of the relaxed problem, while the functional is (sequentially) weakly lower semicontinuous and hence admits a minimiser. This is also portrayed in the following section.

2.3.1. Gradient Young measures and relaxation

Just as in Section 2.2.1 we are concerned with the problem of minimising a functional of the form

\[
I(u) = \int_{\Omega} f(x, \nabla u(x)) \, dx, \tag{2.22}
\]

in a set of admissible functions \( \mathcal{A} \), but with the crucial difference that we do not expect to have the existence of a minimiser due to a lack of quasiconvexity of the integrand \( f \).

However, in common with the direct method of the calculus of variations, our starting point is again to look at an infimising sequence of \( I \) in \( \mathcal{A} \). We then hope that such a sequence carries some information about how the infimum can be approached even though it is never reached. And indeed, that kind of information is encoded in gradient Young measures, which are special cases of Young measures, themselves directly derived from the so-called Fundamental Theorem of Young measures. To that end we nevertheless need to assume some regularity on the integrand \( f \), even though we do not assume any
convexity conditions. Therefore, let $C_0(\mathbb{R}^d)$ denote the closure of the space of continuous functions on $\mathbb{R}^d$ with compact support. Then the dual of $C_0(\mathbb{R}^d)$ can be identified with the space of signed Radon measures with finite mass $\mathcal{M}(\mathbb{R}^d)$. The corresponding duality pairing is for $\mu \in \mathcal{M}(\mathbb{R}^d)$ and $f \in C_0(\mathbb{R}^d)$

$$\langle \mu, f \rangle = \int_{\mathbb{R}^d} f \, d\mu.$$ 

Further introducing a possible dependence of $\mu$ on $x$ we call a map $\mu : \Omega \to \mathcal{M}(\mathbb{R}^d)$ weak*-measurable if $\langle \mu(\cdot), f \rangle$ is measurable for all $f \in C_0(\mathbb{R}^d)$. It is customary to write $\mu_x$ instead of $\mu(x)$. Then the following theorem can be obtained:

**Theorem 2.16** (Fundamental Theorem of Young measures, [8], Theorem 3.1 [43]). Let $\Omega \subseteq \mathbb{R}^m$ be a set of finite measure and let $F_j : \Omega \to \mathbb{R}^d$ be a sequence of measurable functions. Then there exists a subsequence $F_{jk}$ and a weak*-measurable map $\nu : \Omega \to \mathcal{M}(\mathbb{R}^d)$ such that the following holds.

(i) $\nu_x \geq 0$, $\|\nu_x\|_{\mathcal{M}(\mathbb{R}^d)} = \int_{\mathbb{R}^d} d\nu_x \leq 1$, for a.e. $x \in \Omega$.

(ii) For all $f \in C_0(\mathbb{R}^d)$

$$f(F_{jk}) \rightharpoonup \overline{f} \text{ in } L^\infty(\Omega),$$

where

$$\overline{f}(x) = \langle \nu_x, f \rangle = \int_{\mathbb{R}^d} f \, d\nu_x.$$ 

(iii) Let $K \subseteq \mathbb{R}^d$ be compact. Then

$$\text{supp } \nu_x \subseteq K \quad \text{if} \quad \text{dist}(F_{jk}, K) \to 0 \text{ in measure}.$$ 

(iv) Furthermore one has

$$(i') \quad \|\nu_x\|_{\mathcal{M}(\mathbb{R}^d)} = 1 \quad \text{for a.e. } x \in \Omega$$

if and only if the sequence does not escape to infinity, i.e. if

$$\lim_{M \to \infty} \sup_k \{ |\{F_{jk} \geq M\}| \} = 0.$$
(v) If (i’) holds, and if $A \subseteq E$ is measurable, if $f \in C_0(\mathbb{R}^d)$ and if $f(F_{j_k})$ is relatively compact in $L^1(A)$,

then

$$f(F_{j_k}) \rightharpoonup \tilde{f} \text{ in } L^1(A), \quad \tilde{f}(x) = \langle v_x, f \rangle.$$ 

(vi) If (i’) holds, then in (iii) one can replace ‘if’ with ‘if and only if’.

**Definition 2.17.** The map $\nu : \Omega \to \mathcal{M}(\mathbb{R}^d)$ from the theorem above is called the Young measure generated by (or associated to) the sequence $F_{j_k}$.

However, with the problem of minimising the energy (2.22) in mind, we are not only interested in Young measures, but more specifically in those which are generated by gradients $\nabla u_j$ of some functions $u_j$, rather than just some arbitrary choices of $F_j$. Those Young measures are appropriately called gradient Young measures and defined as follows:

**Definition 2.18.** A (weak* measurable) map $\nu : \Omega \to \mathcal{M}(\mathbb{R}^{n \times m})$ is a $W^{1,p}$-gradient Young measure if there exists a sequence of maps $u_j : \Omega \to \mathbb{R}^n$, s.t.

$$u_j \rightharpoonup u \text{ in } W^{1,p}(\Omega; \mathbb{R}^n) \quad (\star \text{ if } p = \infty)$$

$$\delta_{\nabla u(\cdot)} \rightharpoonup \nu \text{ in } L^\infty_w(\Omega; \mathcal{M}(\mathbb{R}^{n \times m})), $$

where $L^\infty_w(\Omega; \mathcal{M}(\mathbb{R}^{n \times m}))$ is the space of weak* measurable maps $\mu : \Omega \to \mathcal{M}(\mathbb{R}^{n \times m})$ that are (essentially) bounded.

Kinderlehrer and Pedregal establish that gradient Young measures are the ‘natural’ dual objects to quasiconvex functions, [27].

**Theorem 2.19 ([27]).** Let $1 \leq p < \infty$. A (weakly measurable) map $\nu : \Omega \to \mathcal{M}(\mathbb{R}^{n \times m})$ is a $W^{1,p}$-gradient Young measure if and only if $\nu_x \geq 0$ a.e. and there exists $u \in W^{1,p}(\Omega; \mathbb{R}^n)$ such that the following three conditions hold:

(i) $\int_\Omega \int_{\mathbb{R}^{n \times m}} |F|^p \ d\nu_x(F) \ dx < \infty$

(ii) $\langle \nu_x, \text{id} \rangle = \nabla u$, a.e. $x$

(iii) $\langle \nu_x, f \rangle \geq f(\langle \nu_x, \text{id} \rangle)$ for a.e. $x$ and all quasiconvex $f$ with $|f|(F) \leq C(|F|^p + 1)$.
A similar version of this theorem exists for the case $p = \infty$, [20]. An advantage of gradient Young measures is that the above results easily extend to the case where the function $f$ additionally depends on $x$, i.e. $f : \Omega \times \mathbb{R}^{n \times m} \to \mathbb{R}$, $f = f(x, F)$. We are now in a position to present the theorem that shows that gradient Young measures are generalised solutions of minimising

$$I(u) = \int_{\Omega} f(x, \nabla u) \, dx$$

(2.22)

on the set

$$A = \{ u \in W^{1,p}(\Omega; \mathbb{R}^n) : u - v \in W^{1,p}_0(\Omega; \mathbb{R}^n) \} = W^{1,p}_v(\Omega; \mathbb{R}^n)$$

(2.23)

for some given $v \in W^{1,p}(\Omega; \mathbb{R}^n)$. In general, as we motivated in the beginning of this section, this problem does not have a minimiser due to the lack of quasiconvexity of the integrand $f$ in $I$. Nevertheless, if we define the set

$$\mathcal{G} = \{ \nu : \Omega \to \mathcal{M}(\mathbb{R}^{n \times m}) : \nu \text{ is a } W^{1,p}-\text{gradient Young measure with } \langle \nu_x, \text{id} \rangle = \nabla u(x) \text{ a.e. in } \Omega \text{ for some } u \in A \}$$

and look instead at minimising

$$J(\nu) = \int_{\Omega} \langle \nu_x, f \rangle \, dx$$

on $\mathcal{G}$ we find:

**Theorem 2.20** (Thm 4.9, [43]). Suppose that $f$ is continuous and satisfies $c|F|^p \leq f(F) \leq C(|F|^p + 1)$ for some $c, C > 0$ and $p \in (1, \infty)$. Then

$$\inf_{u \in A} I(u) = \min_{\nu \in \mathcal{G}} J(\nu).$$

Moreover, the minimisers of $J$ are Young measures that are generated by the gradients of minimising sequences of $I$.

If however, one is prepared to lose the information about the microstructure, but is still interested in a somewhat generalised solution, then a technique called relaxation is useful. The idea of that technique is similar to the one for gradient Young measures in that we replace the energy $I$ by a related energy that admits a minimiser. From the previous section we know that this energy has to be sequentially weakly lower semicontinuous in order to ensure the existence of a minimiser. In terms of $I$ this means we are looking
for the largest sequentially weakly lower semicontinuous function below \( I \), which is also called the weak lower semicontinuous envelope \( I^{lsc} \) of \( I \). In fact, it can be shown, that the weak lower semicontinuous envelope \( I^{lsc} \) of the form \( I(u) = \int_{\Omega} f(x, \nabla u) \, dx \) is

\[
I^{lsc}(u) = \int_{\Omega} f^{qc}(x, \nabla u) \, dx,
\]

where \( f^{qc} \) denotes the quasiconvex envelope of the integrand \( f \), i.e. the largest quasiconvex function below \( f \). Then, similarly to Theorem 2.20 and for the same set \( A \) as in (2.23) it holds:

**Theorem 2.21** (Relaxation, Thm. 4.5.ii [43]). Let \( p \in (1, \infty) \) and let \( f \) satisfy \( c|F|^p \leq f(F) \leq C(|F|^p + 1) \) for some \( c > 0 \). Then

\[
\inf_{u \in A} I(u) = \min_{u \in A} I^{lsc}(u)
\]

Moreover, a function \( \bar{u} \) is a minimiser of \( I^{lsc} \) in \( A \) if and only if it is a cluster point with respect to weak convergence in \( W^{1,p} \) of a minimising sequence of \( I \).

Roughly speaking, Theorem 2.21 says that the lower semicontinuous envelope \( I^{lsc} \) of \( I \) with the quasiconvexification \( f^{qc} \) of the stored energy function \( f \) is the macroscopically relevant energy. This is to be understood in the way that the minimiser of \( I^{lsc} \) no longer carries information about the microstructure that may otherwise be described by the corresponding infimising sequence of \( I \). However, by Theorem 2.20 we know that the oscillatory behaviour of the infimising sequence can be captured in the corresponding gradient Young measure and that whenever \( f^{qc}(x, \nabla u) \neq f(x, \nabla u) \) we are essentially dealing with a microstructure.

To illustrate Theorems 2.20 and 2.21 we return to the one-dimensional example from the beginning of this section, i.e. of infimising \( I(u) = \int_{0}^{1} f(x, u, u_x) \, dx \) with \( f(x, u, u_x) = (u_x - a)^2(u_x + b)^2 + u^2 \) among all maps \( u \in W^{1,4}(0,1) \). Note that both the theorems as presented are actually defined for functional \( I(u) = \int_{\Omega} f(x, \nabla u) \, dx \), i.e. without the dependency of \( u \) encountered in this example. However, the dependency on \( u \) (and in fact \( x \)) can be overcome easily as shown in [43, Corollary 3.3 and 3.4]. The functional \( J \) defined for gradient Young measures is then defined as \( J(\nu_x) = \int_{\Omega} \langle \nu_x, f(x, u(x), \cdot) \rangle \, dx \) where \( u \) is the weak limit of the sequence that generates the \( \nu_x \) (see Definition 2.18).

Now let \( h \) be a 1-periodic sawtooth function with \( h(0) = 0 \) and \( h'(x) = a \) on \( (0, \frac{b}{a+b}) \) and \( h'(x) = -b \) on \( (\frac{b}{a+b}, 1) \), i.e. \( h \) corresponds to the largest (dashed) function in Figure 2.2. Then the infimising sequence as shown is defined as \( u_j(x) = \frac{1}{j} h(2^j x) \). The gradient of this sequence satisfies \( F_j := \nabla u_j(x) = h(2^j x) \in \{a, -b\} =: K \). By Theorem 2.16(iii)
and (iv) there exists a weak*-measurable map $\nu : \Omega \rightarrow M(\mathbb{R})$ with $|\nu_x|_{M(\mathbb{R})} = 1$ and $\text{supp} \nu_x \subseteq K$, i.e.

$$\nu_x = \mu(x)\delta_a + (1 - \mu(x))\delta_{-b}.$$  

Then taking $f = \text{id}$ in (v) implies that for a subsequence $F_{k_j}$

$$F_{k_j} \rightharpoonup \langle \nu_x, \text{id} \rangle = \mu(x)a - (1 - \mu(x))b.$$  

But since $u_j \rightharpoonup u = 0$ in $W^{1,4}_0(\Omega)$ we have $\mu(x)a - (1 - \mu(x))b = 0$, i.e. $\mu(x) = \frac{b}{a + b}$. Therefore, $\nu_x = \frac{b}{a + b}\delta_a + \frac{a}{a + b}\delta_{-b}$ is a homogeneous gradient Young measure with $\langle \nu, \text{id} \rangle = \nabla u = 0$. Furthermore, since $f(x, u, u_x) \geq 0$ it follows that $J(\nu_x) = \int_\Omega \nu_x f(x, u, \cdot) \, dx \geq 0$ for all $\nu_x \in \mathcal{G}$. In particular, we have $J(\nu) = \int_\Omega \nu f(x, 0, \cdot) \, dx = |\Omega| \int_{\mathbb{R}} (\lambda - a)^2(\lambda + b)^2 \, d\nu(\lambda) = 0$. Therefore, we obtain

$$\inf_{u \in \mathcal{A}} I(u) = \min_{\nu \in \mathcal{G}} J(\nu) = J(\nu) = 0$$

and the gradient Young measure $\nu$ means that, roughly speaking, at each point there is the probability of $\frac{b}{a + b}$ of finding the gradient $a$ and of $\frac{a}{a + b}$ of finding the gradient $-b$, thereby encoding the microstructure of the generalised solution.

In case of the relaxation technique we similarly only need to consider quasiconvexification in the gradient $\nabla u$ independently of $x$ and $u$. We find

$$f^{qc}(x, u, u_x) = \begin{cases} u^2 & u_x \in (-b, a) \\ u^2 + (u_x - a)^2(u_x + b)^2 & \text{else,} \end{cases}$$

which is sketched in Figure 2.4. For this function we see that $u = 0$ is a minimiser of

![Figure 2.4.: Quasiconvexification $f^{qc}$ of $f$](image-url)
\( I^{\text{sc}}(u) = \int_\Omega f^c(x, u, u_x) \, dx \) as now all gradients in the interval \([-b, a]\) (and so \(u_x = 0\)) can be achieved with zero energy. With respect to Theorem 2.21 we have indeed that

\[
\inf_{u \in A} I(u) = \min_{u \in A} I^{\text{sc}} = I^{\text{sc}}(0) = 0,
\]

where \(u = 0\) is the weak limit of the infimising sequence \(u_j\). The minimiser \(u = 0\) of the relaxed problem therefore only accounts for the macroscopically relevant information and does not encode any microstructure.
3. Annulus twist-maps

3.1. General setting

The original example of multiple equilibrium solutions for pure displacement boundary conditions, motivated by John [25] and examined in more detail by Post and Sivaloganathan [48], is that of finding minimisers to an appropriately defined elastic energy of functions that map an annulus to itself while leaving the boundary invariant. Equilibrium solutions are those functions that solve the energy-momentum equations of the energy and which are therefore candidates for minimisers. The following sections will show that there are multiple solutions to those energy-momentum equations. Those solutions correspond to maps that minimise the energy in certain subclasses of maps on the annulus.

To be more precise, we can find a minimiser of the energy in each class of functions that twists the outer boundary around the inner boundary $N$ times while still leaving the boundary invariant. Note that each distinct minimiser solves the same energy-momentum equations.

The following definitions are based on the work of Post & Sivaloganathan [48]. Let $A = \{x \in \mathbb{R}^2 : a < |x| < b\}$ be an annulus, where $0 < a < b$. The energy considered by Post and Sivaloganathan is a standard polyconvex elastic energy

$$ I(u) = \frac{1}{2} \int_A |\nabla u|^2 + h(\det \nabla u) \, dx $$

where $h$ is a convex function with $h(d) \to +\infty$ as $d \to 0^+$, $h(d) = +\infty$ as $d \leq 0$ and $h(d) \to +\infty$ as $d \to +\infty$. The last two properties are defined to mimic the volume compression and expansion penalisation in (2.9) and (2.10). The energy is then defined on the set of admissible functions

$$ \mathcal{A} = \{u \in W^{1,2}(A) : u = \text{id} \text{ on } \partial A\}. $$

In order to define the classes of functions that twist the outer boundary around the inner boundary $N$ times we need the notion of the winding number. The winding number of a closed $C^1$ curve in the plane, i.e. $\gamma : [a, b] \to \mathbb{R}^2$ with $\gamma(a) = \gamma(b)$ and $\gamma(r) = [x(r), y(r)]^T$,
is defined by

\[
\text{wind } \# \gamma = \frac{1}{2\pi} \int_a^b \frac{x(r)y'(r) - x'(r)y(r)}{x^2(r) + y^2(r)} \, dr.
\]

Post and Sivaloganathan then extend this definition to curves which are merely continuous. Then writing \( u \in \mathcal{A} \) in polar coordinates \( u = u(r, \theta) \) we see that for a.e. \( \theta \in [0, 2\pi] \) we have \( u(\cdot, \theta) \in W^{1,1}(a, b) \) and therefore \( u(\cdot, \theta) \) can be redefined on a set of measure zero to be absolutely continuous on \([a, b]\). Therefore the curve \( \gamma_\theta(r) = \frac{u(r, \theta)}{|u(r, \theta)|} \) of this continuous representative is a closed and continuous curve with a well defined winding number. We are now fit to define the classes of \( N \)-twist maps for each \( N \in \mathbb{N} \):

\[
\mathcal{A}_N = \{ u \in \mathcal{A} : \text{wind } \# \gamma_\theta = N \text{ for a.e. } \theta \in [0, 2\pi] \}.
\]

It remains to show that \( I \) obtains a minimum in each \( \mathcal{A}_N \). Post and Sivaloganathan prove the following essential ingredient.

**Lemma 3.1** (Lm. 2.9 [48]). Let \( N \in \mathbb{N} \) and \((u_n)_n \subseteq \mathcal{A}_N \) with \( u_n \rightharpoonup u \) in \( W^{1,2}(A) \). Then \( u \in \mathcal{A}_N \).

This lemma states that \( \mathcal{A}_N \) is closed under weak convergence. The existence of a minimiser then follows easily by applying the direct method of the calculus of variations as introduced in Section 2.2.1.

However, before we present results on the properties of minimisers of \( I \) on \( \mathcal{A}_N \) we have a look at the modified problem which was already hinted at in their original paper and further investigated by Francfort and Sivaloganathan [24] as a case where the Euler-Lagrange equations are not satisfied for the minimisers. The modified elastic energy is

\[
I^0(u) = \int_A \frac{1}{2} |\nabla u|^2 \, dx
\]

which is defined on the set

\[
\mathcal{A}^0 = \{ u \in W^{1,2}(A) : u = \text{id on } \partial A \text{ and } \det \nabla u \geq 0 \text{ a.e. in } A \}
\]

and analogously

\[
\mathcal{A}^0_N = \{ u \in \mathcal{A}^0 : \text{wind } \# \gamma_\theta = N \text{ for a.e. } \theta \in [0, 2\pi] \}.
\]

The condition \( \det \nabla u \geq 0 \) is introduced in order to remain somewhat physically realistic.
Although we would like to have $\det \nabla u > 0$ a.e. in $A$, it is not wise to demand it, since with that stronger assumption $A^0_N$ would not be weakly closed and the existence of a minimiser could not be expected. The Euler-Lagrange equations are formally given by

$$
\begin{cases}
\Delta u = 0 & \text{in } A \\
u = \text{id} & \text{on } \partial A,
\end{cases}
$$

which has the unique solution $u = \text{id}$ that can only be the minimiser for $N = 0$. Therefore we expect that the minimisers $u^N$ of $I^0$ in $A^0_N$ for $N \neq 0$ are degenerate, i.e. $\det \nabla u^N = 0$ on a set of positive measure. This will be addressed in the following section.

On the contrary, for the functional $I$ with volume compression energy, we find that the Euler-Lagrange equations

$$
\begin{cases}
\Delta u + (\text{cof } \nabla u) \nabla \left( h'(\det \nabla u) \right) = 0 & \text{in } A \\
u = \text{id} & \text{on } \partial A,
\end{cases}
$$

(3.1)

admit infinitely many strong solutions. This will be proved in Section 3.3.

### 3.2. Minimisers without volume compression energy

For each $N \in \mathbb{N}$ we consider the problem of finding $\overline{u}$ such that

$$I^0(\overline{u}) = \min_{u \in A^0_N} I^0(u).$$

Since (2.21) is satisfied we know by Theorem 2.15 that the minimisers satisfy the energy-momentum equations, which in this case take the strong form

$$
\begin{cases}
\text{div} \left( \frac{1}{2} |\nabla u|^2 I - \nabla u^T \nabla u \right) = 0 & \text{in } A \\
u = \text{id} & \text{on } \partial A.
\end{cases}
$$

(3.2)

In particular, we want to look for a rotationally symmetric solution, i.e. a solution from the set

$$A^0_{N,\text{sym}} = \{ u \in A^0_N : u(x) = Q^T u(Qx) \text{ for all } Q \in SO(2) \}.$$
That such a solution exists follows from the same arguments as for the nonrotationally symmetric case. Hence we can represent the solution \( u \) in polar coordinates

\[
 u(r, \theta) = \rho(r)e_r(\theta + \psi(r)) \tag{3.3}
\]

where \( e_r(\theta) = [\cos \theta, \sin \theta]^T \). For brevity we shall write from here on \( e_r \) for \( e_r(\theta) \) and \( \tilde{e}_r \) for \( e_r(\theta + \psi(r)) \). Equally we define \( e_\theta(\theta) = [-\sin \theta, \cos \theta]^T \) and use the abbreviations \( e_\theta \) and \( \tilde{e}_\theta \) analogously. We call \( \rho \) the radial map and \( \psi \) the angular map. Then it holds that:

**Lemma 3.2.** Let \( N \in \mathbb{N} \). Then the radial map \( \rho \) of a minimiser of \( I^0 \) in \( A^0_{N,sym} \) is differentiable and satisfies the ODE

\[
\dot{\rho} = \frac{1}{r} \sqrt{\rho^2 - \frac{\omega^2}{\rho^2} - a^2 + \frac{\omega^2}{a^2}}
\]

\[
\rho(a) = a, \quad \rho(b) = b. \tag{3.4}
\]

for some \( \omega \in (0, \infty) \). Furthermore, the angular map \( \psi \) is differentiable with

\[
\dot{\psi} = \frac{\omega}{r\rho^2}
\]

and \( \psi(a) = 0 \) and \( \psi(b) = 2\pi N \).

**Proof.** To prove this we take the weak form of (3.2) and test with an equally rotationally symmetric test function \( \phi \). We can express \( \phi \) as

\[
\phi(r, \theta) = \tilde{\rho}(r)e_r + \tilde{q}(r)e_\theta.
\]

with \( \tilde{\rho}, \tilde{q} \in C^\infty_c((a, b)) \). Furthermore

\[
\nabla u = \dot{\tilde{\rho}} \tilde{e}_r \otimes e_r + \rho \dot{\tilde{\psi}} \tilde{e}_\theta \otimes e_r + \frac{\rho}{r} \tilde{e}_\theta \otimes e_\theta
\]

\[
\nabla \phi = \dot{\tilde{\rho}} \tilde{e}_r \otimes e_r + \tilde{\dot{q}} \tilde{e}_\theta \otimes e_r + \frac{1}{r} [\tilde{\rho} e_\theta \otimes e_\theta - \tilde{q} \tilde{e}_r \otimes e_\theta]
\]

and

\[
|\nabla u|^2 = \rho^2 + \rho^2 \dot{\psi}^2 + \frac{\rho^2}{r^2}.
\]
Therefore,

\[
\frac{1}{2} |\nabla u|^2 I - \nabla u^T \nabla u = \frac{1}{2} \left( \rho^2 + \rho^2 \dot{\psi}^2 + \frac{\rho^2}{r^2} \right) I - \left( (\rho^2 + \rho^2 \dot{\psi}^2) \mathbf{e}_r \otimes \mathbf{e}_r + \frac{\rho^2}{r^2} \mathbf{e}_\theta \otimes \mathbf{e}_\theta \right) + \rho^2 \frac{\dot{\psi}}{r} (\mathbf{e}_r \otimes \mathbf{e}_\theta + \mathbf{e}_\theta \otimes \mathbf{e}_r) + \frac{\rho^2}{r^2} \mathbf{e}_\theta \otimes \mathbf{e}_\theta
\] (3.5)

so that

\[
0 = \int_A \left[ \frac{1}{2} |\nabla u|^2 I - \nabla u^T \nabla u \right] \cdot \nabla \phi \, dx
= 2\pi \int_a^b \frac{r}{2} \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 + \frac{\rho^2}{r^2} \right) \left( \frac{\dot{\rho}}{\rho} + \frac{\dot{\psi}}{r} \right)
- r \left( (\rho^2 + \rho^2 \dot{\psi}^2) \dot{\rho} + \frac{\rho^2 \dot{\psi}}{r} \left( \frac{\dot{\rho}}{\rho} - \frac{\dot{q}}{r} \right) + \frac{\rho^2}{r^2} \dot{\rho} \right) \, dr
= 2\pi \int_a^b \frac{r}{2} \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 - \frac{\rho^2}{r^2} \right) \left( \dot{\rho} - \frac{\dot{\rho}}{\rho} \right) + r \frac{\rho^2 \dot{\psi}}{r} \left( \frac{\dot{q}}{r} - \frac{\dot{q}}{r} \right) \, dr
= -2\pi \int_a^b \frac{r}{2} \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 - \frac{\rho^2}{r^2} \right) \left( \frac{\dot{\rho}}{\rho} \right) + r \rho^2 \dot{\psi} \left( \frac{\dot{q}}{r} \right) \, dr.
\]

Since \( \hat{\rho} \) and \( \hat{q} \) are arbitrary this implies that there exist constants \( c \) and \( \omega \) s.t.

\[
r^2 \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 - \frac{\rho^2}{r^2} \right) = c \quad \text{in} \quad (a, b) \quad (3.6)
\]

and

\[
r \rho^2 \dot{\psi} = \omega \quad \text{in} \quad (a, b). \quad (3.7)
\]

Furthermore, since \( \int_A |\nabla u|^2 \, dx < \infty \) for the minimiser, it follows \( \rho \in W^{1,2}((a, b)) \), which in turn yields \( \rho \in C((a, b)) \). Therefore, we have \( \dot{\psi} = \frac{\omega}{r \rho} \in C((a, b)) \) as well. That \( \omega > 0 \) simply follows from the fact that by (3.7) \( \psi \) is a monotonic function and we want to achieve a positive winding number, i.e. \( \psi(b) = 2\pi N > 0 \) and \( \psi(a) = 0 \). Substituting \( \dot{\psi} \) back into (3.6) we obtain

\[
r^2 \dot{\rho}^2 + \frac{\omega^2}{\rho^2} - \rho^2 = c \quad (3.8)
\]

which also implies that the weak derivative \( \dot{\rho} \) is in fact continuous and is therefore the classical derivative. Since \( \det \nabla u = \frac{\rho^2}{r^2} \geq 0 \), we find that \( \rho^2 \) is monotonically increasing.
Therefore \( \rho \geq a > 0 \) which in turn implies \( \dot{\rho} \geq 0 \). Hence we can solve for \( \dot{\rho} \) in (3.8) to obtain (3.4).

Now we want to prove that \( c = -a^2 + \frac{\omega^2}{\rho^2} \). For this we show that \( \dot{\rho}(a) \) has to be zero, which implies the assertion. Assume there was a point \( r \in (a, b] \) s.t. \( \dot{\rho}(r) = 0 \) and \( \dot{\rho}(r) > 0 \) for \( r \in (r - \delta, r) \) for some \( \delta > 0 \), meaning that we suppose \( \dot{\rho} \) has a zero after a point where it was strictly positive. Then \( r^2 \dot{\rho}^2 \) approaches zero from above as \( r \to r^- \), and hence so does \( \rho^2 - \frac{\omega^2}{\rho^2} + c \) by (3.8). But \( f(x) = x^2 - \frac{\omega^2}{x^2} + c \) is a monotonically increasing function in \( x \), and since \( \rho \) is monotonically increasing itself, the function \( \rho^2 - \frac{\omega^2}{\rho^2} + c \) is strictly positive on \( (r - \delta, r) \) and so monotonically increasing. This is a contradiction as it has to approach zero as \( r \to r^- \). Therefore, if there is a point where \( \dot{\rho} \) is zero, then so it is at \( r = a \), i.e. \( \dot{\rho}(a) = 0 \). Assume now \( \dot{\rho}(a) > 0 \). Then, since \( \dot{\rho} \in C([a, b]) \) and by the reasoning above, it is bounded away from zero on the whole of \([a, b] \), i.e. \( \dot{\rho} \geq \varepsilon > 0 \) for some \( \varepsilon > 0 \). But then it is a standard result that \( u \) solves the Euler-Lagrange equations

\[
\begin{align*}
\Delta u &= 0 \quad \text{in } A \\
u &= \text{id} \quad \text{on } \partial A,
\end{align*}
\]

which admits only the identity as a solution which corresponds to \( N = 0 \). This proves the assertion.

This lemma implies that we can reduce the energy-momentum equations for \( \rho \) and \( \psi \) to just an ODE in \( \rho \) with the initial condition \( \rho(a) = a \). At the moment it might seem strange that there is only the one parameter \( \omega \) left to fit both the boundary condition \( \rho(b) = b \) and to ensure that \( \psi(b) = 2\pi N \). However, what comes to our rescue is the fact that due to the lack of Lipschitz continuity of the right hand side there are infinitely many solutions for each \( \omega \) that differ qualitatively only by the point \( k \in [a, b) \) where \( \dot{\rho} \) departs from zero, which is an additional hidden parameter. A rather unusual albeit anticipated result is now that this ODE, and therefore the energy-momentum equation, has an explicit solution.

**Theorem 3.3.** Let \( N \in \mathbb{N} \). Then there exist \( \omega > 0 \) and \( k \in [a, b) \) s.t.

\[
\rho(r) = \begin{cases} a, & r \in [a, k] \\ \frac{1}{2} \left( \left( a^2 + \frac{\omega^2}{\rho^2} \right) k^2 + \left( a^2 + \frac{\omega^2}{\rho^2} \right) r^2 + 2 \left( a^2 - \frac{\omega^2}{\rho^2} \right) \right)^{\frac{1}{2}}, & r \in (k, b] \end{cases}
\]

is a solution to the ODE in Lemma 3.2. Furthermore, \( \omega \) and \( k \) are uniquely determined.
The corresponding angular map is

\[
\psi(r) = \begin{cases} 
\frac{\omega}{a^2} \ln \left( \frac{r}{a} \right), & r \in [a, k] \\
\frac{\omega}{a^2} \ln \left( \frac{k}{a} \right) + \arctan \left( \frac{1}{2} \left[ \left( \frac{a^2 + \omega^2}{b^2} \right) \frac{b^2}{k^2} + a^2 - \frac{\omega^2}{a^2} \right] \right) - \arctan \left( \frac{a^2}{\omega} \right), & r \in (k, b].
\end{cases}
\]

(3.9)

Proof. It is easy to verify that ρ given as the above indeed solves the ODE. The existence of ω and k is already ensured by the existence of the minimiser, so there is not much to prove. It remains to ensure that the boundary conditions ρ(b) = b and ψ(b) = 2πN are met. Now, the condition ρ(b) = b already fixes ω > 0 as a function of k:

\[
\omega^2 = \frac{4b^2k^2a^2 - a^4(b^2 + k^2)^2}{(b^2 - k^2)^2}.
\]

Inserting this into (3.9) we find ψb := ψ(b) as a continuous function of k. Since

\[
\frac{1}{2\omega} \left[ \left( \frac{a^2 + \omega^2}{b^2} \right) \frac{b^2}{k^2} + a^2 - \frac{\omega^2}{a^2} \right] > \frac{a^2}{\omega} > 0
\]

we find that less than a quarter of a twist is performed after the map leaves the inner radius of the annulus. Therefore \(ψ_b(k = a) < \frac{\pi}{2} < 2\pi N\). Furthermore \(ψ_b\) has a pole at \(k = b\), i.e. \(ψ_b(k) → ∞\) as \(k → b^−\). Additionally it is easy, but cumbersome, to prove that \(ψ_b\) is monotonically increasing. Together this yields the assertion that for each \(N ∈ \mathbb{N}\) there exists a unique \(k ∈ [a, b)\) s.t. \(ψ_b(k) = 2πN\).

At this point it seems useful to provide a visualisation of the solution found for \(N = 1\): see Figure 3.1. We define the set \(H = \{x ∈ \mathbb{R}^2 : a ≤ |x| ≤ k\} ⊆ A\) to be the region that gets crushed to zero volume and denote for it the name hedgehog region. It becomes intuitive why less than a quarter of the twist is performed outside the hedgehog region. (The map \(x → \frac{x}{|x|}\) is commonly referred to as the hedgehog map. In the region \(H\) the solution corresponds to such a map with a twist, which is why we call it the hedgehog region.)

So far we only considered rotationally symmetric maps and for each \(N ∈ \mathbb{N}\) we have found a unique minimiser in \(A^0_{N, sym}\). At the moment it is not clear whether this minimum in \(A^0_{N, sym}\) is also a (local) minimum of all maps that twist the outer boundary around the inner boundary \(N\) times, i.e. in \(A^0_N\). Although we are unable to prove that a minimiser has to be rotationally symmetric (we discuss this later) we can nevertheless show that the minimiser of \(A^0_{N, sym}\) has a lower energy than a large class of twist maps in \(A^0_N\).

Proposition 3.4. Let \(N ∈ \mathbb{N}\) and let \(u\) minimise \(I^0\) in \(A^0_{N, sym}\).
(i) Let $T_u^+ A_N^0 = \{ \varphi \in H_0^1(A) : u + \varepsilon \varphi \in A_N^0 \text{ for sufficiently small } \varepsilon > 0 \}$. Then for each $\varphi \in T_u^+ A_N^0$

$$I^0(u + \varepsilon \varphi) \geq I^0(u)$$

(3.10)

for all sufficiently small $\varepsilon > 0$.

(ii) Let $v \in A_N^0$ be such that $\varphi := v - u$ satisfies

$$\int_H |\nabla \varphi|^2 + 2 \left(1 + \frac{\omega^2}{a^2}\right) \left(\frac{1}{k} - \frac{1}{r}\right) \det \nabla \varphi \, dx \geq 0.$$  

(3.11)

Then

$$I^0(v) \geq I^0(u).$$

Proof. Proof of (i). Let $\varphi \in T_u^+ A_N^0$. Then for all sufficiently small $\varepsilon > 0$ we have that

$$0 \leq \det(\nabla u + \varepsilon \nabla \varphi) = \det \nabla u + \varepsilon \cof \nabla u \cdot \nabla \varphi + \varepsilon^2 \det \nabla \varphi$$

(3.12)

a.e. in $A$. Since $\det \nabla u = 0$ in $H$ it follows (by dividing by $\varepsilon$ and letting $\varepsilon \to 0^+$) that

$$\cof \nabla u \cdot \nabla \varphi \geq 0$$

(3.13)

a.e. in $H$. We will now show that (3.13) is sufficient to show that $\int_A \nabla u \cdot \nabla \varphi \, dx \geq 0$,
which then implies (3.10) since then
\[
I'(u + \varepsilon \varphi) - I'(u) = \varepsilon \int_A \nabla u \cdot \nabla \varphi \, dx + \varepsilon^2 \int_A \frac{1}{2} |\nabla \varphi|^2 \, dx \geq 0
\] (3.14)

for sufficiently small $\varepsilon > 0$.

In the following we denote by $S_R$ the circle $\{ x \in \mathbb{R}^2 : |x| = R \}$ of radius $R$. Note that, since $u$ is smooth on $H$ and $A \setminus H$ and its first derivatives are continuous across the boundary $S_k$, Green’s theorem implies that
\[
\int_A \nabla u \cdot \nabla \varphi \, dx = -\int_H \Delta u \cdot \varphi \, dx.
\]

Note that the domain of integration on the right hand side is simply $H$ as $u$ is harmonic on $A \setminus H$. Next, the specific form of the solution implies that $\Delta u = -\frac{a}{r^2} \left( \frac{\omega^2}{a^2} + 1 \right) \tilde{e}_r$, so that
\[
\int_A \nabla u \cdot \nabla \varphi \, dx = a \left( \frac{\omega^2}{a^2} + 1 \right) \int_H \frac{1}{r^2} \tilde{e}_r \cdot \varphi \, dx.
\] (3.15)

Turning our attention once more to $\text{cof} \nabla u \cdot \nabla \varphi$ we integrate it on the subannulus $B_{R\setminus a} = \{ x \in A : a \leq |x| \leq R \}$ of outer radius $r \leq k$. Using Green’s theorem and the identity $\text{div} \text{cof} \nabla u = 0$ we obtain
\[
\int_{B_{R\setminus a}} \text{cof} \nabla u \cdot \nabla \varphi \, dx = \int_{S_R} (\text{cof} \nabla u)n \cdot \varphi \, dS = a \int_{S_R} \tilde{e}_r \cdot \varphi \, dS.
\] (3.16)

Therefore, by (3.13), it follows that $a \int_{S_R} \tilde{e}_r \cdot \varphi \, dS \geq 0$ for all $R \in [a, k]$. Now note that $\tilde{e}_r \cdot \varphi$ appears in both (3.15) and (3.16). The latter being positive implies that $\int_{H} \nabla u \cdot \nabla \varphi \, dx \geq 0$ must be positive as well.

**Proof of (ii).** Let $v \in A^0$ such that for $\varphi := v - u$ (3.11) holds. Similar to Part (i) we want to show that (3.14) with $\varepsilon = 1$ is not negative by exploiting (3.12) (now with $\varepsilon = 1$). We have that
\[
\text{cof} \nabla u \cdot \nabla \varphi + \nabla \varphi \geq 0
\]
a.e. in $H$. Therefore, by the above and (3.16) we obtain that
\[
a \int_{S_R} \tilde{e}_r \cdot \varphi \, dS \geq -\int_{B_{R\setminus a}} \text{det} \nabla \varphi \, dx
\] (3.17)
for all $R \in [a, k]$. Recall that we have $\int_A \nabla u \cdot \nabla \varphi \, dx = a \left( \frac{\varphi^2}{r^2} + 1 \right) \int_H \frac{1}{r} \tilde{e}_r \cdot \varphi \, dx$. In order to use (3.17) to control the integral the above equality we prove that

$$\int_{B_{R\setminus a}} \det \nabla \varphi \, dx = \frac{1}{2} \int_{S_R} J \varphi \cdot \varphi, r \, dS,$$

(3.18)

where $J$ is the rotation matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\varphi, r = \frac{\partial \varphi}{\partial r} = \frac{1}{r} \frac{\partial \varphi}{\partial \theta}$. This leads to

$$a \int_{S_R} \tilde{e}_r \cdot \varphi \, dS \geq -\frac{1}{2} \int_{S_R} J \varphi \cdot \varphi, r \, dS$$

for all $R \in [a, k]$ so that

$$\int_H \frac{1}{r^2} \tilde{e}_r \cdot \varphi \, dx \geq -\frac{1}{2} \int_H \frac{1}{r^2} J \varphi \cdot \varphi, r \, dx.$$

(3.19)

In general for a $\varphi \in C^2_0(A)$, which we may assume by a standard density argument, it holds that $\det \nabla \varphi = J \varphi, r \cdot \varphi, \theta$ (with $\varphi, r = \frac{\partial \varphi}{\partial r}$) and that

$$\int_{B_{R\setminus a}} \det \nabla \varphi \, dx = \int_a^R \int_0^{2\pi} J \varphi, r \cdot \varphi, \theta \, d\theta \, dr$$

$$= -\int_a^R \int_0^{2\pi} (J \varphi, r)_{, \theta} \cdot \varphi \, d\theta \, dr$$

$$= -\int_0^{2\pi} \int_a^R (J \varphi, \theta)_{, r} \cdot \varphi \, dr \, d\theta$$

$$= -\int_0^{2\pi} J \varphi(R, \theta)_{, \theta} \cdot \varphi(R, \theta) \, d\theta + \int_0^{2\pi} \int_a^R J \varphi, \theta \cdot \varphi, r \, dr \, d\theta$$

$$= -\int_{S_R} J \varphi, r \cdot \varphi \, dS - \int_{B_{R\setminus a}} J \varphi, r \cdot \varphi, r \, dx,$$

where we used that $J$ is antisymmetric in the last equality. Thus, (3.18) follows and we have (3.19). We now prove that

$$-\frac{1}{2} \int_H \frac{1}{r^2} J \varphi \cdot \varphi, r \, dx = \int_H \left( \frac{1}{k} - \frac{1}{r} \right) \det \nabla \varphi \, dx.$$
This can be seen as follows:

\[-\frac{1}{2} \int \frac{1}{r^2} J_\varphi \cdot \varphi, r \, dx = \frac{1}{2} \int_0^{2\pi} \int_0^k \frac{1}{r} J_\varphi \cdot \varphi, \theta \, d\theta \, dr \]

\[= \frac{1}{2k} \int_{H_0} J_\varphi \cdot \varphi, r \, dS - \frac{1}{2} \int_H \frac{1}{r} J_\varphi, r \cdot \varphi, \theta \, d\theta - \frac{1}{2} \int_0^k \frac{1}{r} J_\varphi \cdot (\varphi, r), \theta \, d\theta \, dr \]

\[= \frac{1}{k} \int_H \det \nabla \varphi, d - \frac{1}{2} \int_H \frac{1}{r} \det \nabla \varphi \, dx + \frac{1}{2} \int_H \frac{1}{r} J_\varphi, r \cdot \varphi, r \, dx \]

\[= \int_H \left( \frac{1}{k} - \frac{1}{r} \right) \det \nabla \varphi, dx, \]

where again we used (3.18) and the antisymmetry of \( J \). Thus, we have that

\[I^0(v) - I^0(u) \geq \left( \frac{\omega^2}{a^2} + 1 \right) \int_H \left( \frac{1}{k} - \frac{1}{r} \right) \det \nabla \varphi, dx + \int_{A} \frac{1}{2} |\nabla \varphi|^2 \, dx \]

and the right hand side is nonnegative by assumption (3.11).

This naturally leads to the following result that \( u \) is a minimiser of \( I^0 \) with respect to perturbations with suitably located support.

**Corollary 3.5.** Let \( N \in \mathbb{N} \) and let \( u \) minimise \( I^0 \) in \( A_{N,sym}^0 \). Let \( v \in A_{N,sym}^0 \) such that \( \varphi := v - u \) has support in the annulus \( A(\bar{r}, b) = \{ x \in \mathbb{R}^2 : \bar{r} < |x| < b \} \subseteq A \), where

\[\frac{1}{\bar{r}} = \frac{1}{k} + \frac{a^2}{a^2 + \omega^2}. \quad (3.20)\]

Then \( I^0(v) \geq I^0(u) \).

**Proof.** With Hadamard’s inequality, which in the \( 2 \times 2 \) case is \( 2 |\det F| \leq |F|^2 \), we can estimate \( \int_H |\nabla \varphi|^2 + 2 \left( \frac{\omega^2}{a^2} + 1 \right) \det \nabla \varphi, dx \) below with

\[\int_H |\nabla \varphi|^2 \left[ 1 + \left( \frac{\omega^2}{a^2} + 1 \right) \left( \frac{1}{k} - \frac{1}{r} \right) \right] \, dx.\]

A simple calculation shows that the integrand is nonnegative if and only if \( r \geq \bar{r} \) for \( \bar{r} \) as in (3.20). Thus, with \( \text{supp} \varphi \subseteq A(\bar{r}, b) \) the whole integral is positive and (3.11) is satisfied.

It would be worthwhile to know if the rotationally symmetric minimiser from \( A_{N,sym}^0 \) is also a global minimiser in \( A_N^0 \). A possible approach would be to show by estimating the energy directly that the energy of any map is bigger than the energy of an appropriately
defined rotationally symmetric map. Such an approach was used in [58] for showing that the minimiser of the Dirichlet energy of all maps $u$ from an annulus with inner radius $a$ and outer radius $b$ to an annulus of inner radius $\mu$ and outer radius $\lambda$ with $a\mu < b\lambda$ and $\mu \geq \lambda$ and the boundary conditions $u(x) = \mu x/|x|$ on the inner radius and $u(x) = \lambda x/|x|$ on the outer radius has to be a radial map. This is done by carefully defining a radial symmetrisation $u^{sym}$ for each map $u$. The main feature of this symmetrisation $u^{sym}$ is that for each $r \in (a, b)$ the image of the sphere $S_r$ under $u^{sym}$ should enclose the same volume with the inner radius as the image of $S_r$ under $u$. Since $u^{sym}(R, \theta) = \rho(R) e_r(\theta)$ is taken to be radial, this translates to

$$2\pi(\rho(r)^2 - a^2) = \int_{B_r \setminus B_a} \det \nabla u \, dx,$$

which by differentiation yields

$$\oint_{S_r} \det \nabla u \, dS = \det \nabla u^{sym} = \frac{\rho(r)}{r} \tilde{\rho}(r).$$

They then estimate the Dirichlet energy with

$$|\nabla u|^2 = |u_r|^2 + |u_\tau|^2 \geq \frac{\det \nabla u^2}{|u_r|^2} + |u_\tau|^2,$$

integrate over $S_r$ and apply Jensen’s inequality with a convex function to obtain

$$\oint |\nabla u|^2 \, dS \geq \oint |\nabla u^{sym}|^2 \, dS.$$ However, here we need to preserve the twisting property of a given map $u$. Assuming that in radial coordinates $u(r, \theta) = \rho(r, \theta) e_r(\theta + \psi(r, \theta))$ with $u \in A^0_N \cap C^2(A)$ ($C^2$ is assumed for simplicity) we can define its analogous symmetrisation

$$u^{sym}(r, \theta) = \rho(r) e_r(\theta) \tilde{\psi}(r)$$

by also requiring (3.21), so that (3.22) holds, and additionally that

$$\tilde{\psi}(r) = \oint_{S_r} \psi(r, \theta) \, dS,$$

which preserves the twisting. Unfortunately, the methods used above for the radial symmetrisation fail since the estimate (3.23) already loses most of twisting information (which is stored in $|u_\tau|^2$). For example, if we already take a rotationally symmetric map, then it would agree pointwise with its symmetrisation, so that $|\nabla u|^2 = |u_r|^2 + |u_\tau|^2 = \dot{\rho}^2 + \rho^2 \psi^2 + \frac{\rho^2}{r^2}$. However, (3.23) yields $|\nabla u|^2 \geq \dot{\rho}^2 + \frac{\rho^2}{r^2}$, so that the term involving the twist, i.e. $\rho^2 \psi^2$, is clearly missing. Unfortunately, overall, we were not successful
in showing that the symmetrisation has a lower energy than the map itself. We could, nevertheless, show a part of the desired result, as follows:

**Proposition 3.6.** Let \( u \in A^0_N \cap C^2 (A) \) with \( u(r, \theta) = \rho(r, \theta)e_r(\theta + \psi(r, \theta)) \). Let either \( \rho_\theta = 0 \) or \( \psi_\theta = 0 \). Then its symmetrisation \( u^{\text{sym}}(r, \theta) = \overline{\rho}(r)e_r(\theta + \overline{\psi}(r)) \) defined by (3.21) and (3.24) has lower energy, i.e.

\[
\int_A |\nabla u|^2 \, dx \geq \int_A |\nabla u^{\text{sym}}|^2 \, dx.
\] (3.25)

**Proof.** By using partial integration one obtains from (3.21)

\[
\overline{\rho}^2 (r) = \int_{S_r} \rho^2 (1 + \psi_\theta) \, dS.
\]

(Note that this equation holds also for \( \rho, \theta \neq 0 \) and \( \psi, \theta \neq 0 \).) Since now either \( \rho_\theta = 0 \) or \( \psi_\theta = 0 \) we find in particular that

\[
\overline{\rho}(r)^2 = \int_{S_r} \rho^2 \, dS.
\]

Differentiating this expression and using Hölder’s inequality we find

\[
\dot{\overline{\rho}}^2 (r) \leq \int_{S_r} \rho^2 \, dS.
\]

Furthermore, it holds that

\[
\dot{\psi}^2 = \left( \int_{S_r} \psi_\theta \, dS \right)^2 \leq \int_{S_r} \psi_\theta^2 \, dS.
\]

Together with those two inequalities we obtain (again since either \( \rho_\theta = 0 \) or \( \psi_\theta = 0 \))

\[
\int_{S_r} |u, r|^2 \, dS = \int_{S_r} \rho_r^2 + \rho^2 \psi_r^2 \, dS \geq \overline{\rho}^2 + \int_{S_r} \rho^2 \, dS \int_{S_r} \psi_r^2 \, dS \geq \overline{\rho}^2 + \rho^2 \dot{\psi}^2 = |u^{\text{sym}}, r|^2.
\]
The estimate for \( \int_{S_r} |u,\tau| dS \) follows by the isoperimetric inequality which states that \( \int_{S_r} |u,\tau| dS \geq \int_{S_r} |u^{sym}| dS \). This is true since we have chosen \( u^{sym} \) so that \( u^{sym}(S_r) \) is itself a circle that encloses the same volume with the inner radius as \( u(S_r) \). So obviously

\[
\int_{S_r} |u,\tau|^2 dS \geq \left( \int_{S_r} |u,\tau| dS \right)^2 \geq |u^{sym}|^2
\]

which implies that \( \int_{S_r} \nabla u^2 dS \geq \int_{S_r} |\nabla u^{sym}|^2 dS \), and hence (3.25).

Neither a proof nor counterexample for the case when \( \rho, \theta \neq 0 \) and \( \psi, \theta \neq 0 \) could so far be found.

### 3.3. Minimisers with volume compression energy

Similarly to Section 3.2 we seek a minimiser of

\[
I(u) = \int_A \frac{1}{2} |\nabla u|^2 + h(\det \nabla u) \, dx
\]

among maps \( u \in \mathcal{A}_N \) for each \( N \in \mathbb{N} \), but where this time the local invertibility condition \( \det \nabla u > 0 \) is encoded in the function \( h \) with the properties

1. \( h \) convex with \( h \geq 0 \),
2. \( h \in C^3((0, \infty)) \) and for some positive constants \( s, c_1, c_2 \) and \( d_0 \), \( c_1 d^{-s-k} \leq (-1)^k h^{(k)}(d) \leq c_2 d^{-s-k} \) for \( 0 < d < d_0 \) and \( k = 0, 1, 2 \),
3. \( h(d) = \infty \) for \( d \leq 0 \),
4. For some real number \( \tau \) and positive constants \( c_3, c_4 \) and \( d_1 \), \( c_3 d^\tau \leq h''(d) \leq c_4 d^\tau \) for \( d \geq d_1 \).

Again, instead of looking at the whole of \( \mathcal{A}_N \), we focus on those functions of \( \mathcal{A}_N \) that are rotationally symmetric, i.e.

\[
\mathcal{A}_{N,sym} = \{ u \in \mathcal{A}_N : u(x) = Q^T u(Qx) \text{ for all } Q \in SO(2) \}.
\]

Post and Sivaloganathan [48] assert that minimisers of \( I \) in \( \mathcal{A}_{N,sym} \) solve the Euler-Lagrange equations, however, a detailed proof is omitted as it is considered standard procedure. The procedure involves considering the energy-momentum equations, which
are known to be satisfied by minimisers, and showing that they imply that the Euler-Lagrange equations hold as well. We give the details of this procedure for completeness, which relies on the works [13, 14].

The weak form of the energy-momentum equations (2.20) in this case is (in terms of $d \equiv \text{det} \nabla u$, for brevity)

$$
\int_A \left( \frac{1}{2} |\nabla u|^2 + h(d) \right) I - \nabla u^T \left( \nabla u + h'(d) \text{cof} \nabla u \right) \cdot \nabla \phi \, dx = 0,
$$

which has to hold for all $\phi \in C^\infty_c(A)$. Then (3.26) simplifies for rotationally symmetric maps to:

**Lemma 3.7.** Let $N \in \mathbb{N}$. Then there exists $\omega \geq 0$ s.t. the radial function $\rho$ of the minimiser $u$ as chosen in (3.3) satisfies

$$
\left[ \frac{1}{2} \left( \dot{\rho}^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) + f(d) \right]' = -\frac{1}{r} \left( \dot{\rho}^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) \text{ in } D'(a,b)
$$

with $\rho(a) = a$ and $\rho(b) = b$ and $f(d) = h'(d)d - h(d)$. Furthermore, the angular map satisfies

$$
\dot{\psi} = \frac{\omega \cdot r \rho^2}{r^2}.
$$

**Proof.** Using the same representations for $\nabla u$ and $\nabla \phi$ as in the proof of Lemma 3.2 (where $\phi(r, \theta) = \tilde{\rho}(r) \tilde{e}_r + \tilde{q} \tilde{e}_\theta$) and the fact that $\nabla u^T \text{cof} \nabla u = dI$ we find

$$
\left( \frac{1}{2} |\nabla u|^2 + h(d) \right) I - \nabla u^T \left( \nabla u + h'(d) \text{cof} \nabla u \right) = \left( \frac{1}{2} |\nabla u|^2 - f(d) \right) I - \nabla u^T \nabla u
$$

Therefore, we only differ by the term $-f(d)I$ to the energy-momentum tensor (3.5) without volume compression energy. Proceeding analogously with the equal parts and treating $-f(d)I$ separately we obtain

$$
0 = \int_a^b \frac{r^2}{2} \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 - \frac{\rho^2}{r^2} \right) \left( \frac{\psi}{r} \right)' + f(d)(r \dot{\rho})' + r \rho^2 \dot{\psi} \left( \frac{\dot{\rho}}{r} \right)' \, dr.
$$

47
Upon noting that $r^2 \left( \frac{\dot{\rho}}{r} \right) = (r\dot{\rho}) - \frac{2}{r}(\dot{r})$ we can rewrite

$$0 = \int_a^b \left[ \frac{1}{2} \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 - \frac{\rho^2}{r^2} \right) + f(d) \right] (r\dot{\rho}) - \frac{1}{r} \left( \dot{\rho}^2 + \rho^2 \dot{\psi}^2 - \frac{\rho^2}{r^2} \right) (r\dot{\rho}) + r \rho^2 \dot{\psi} \left( \frac{\dot{\omega}}{r} \right) \cdot dr.$$

Hence, there exists $\omega \in \mathbb{R}$ s.t.

$$r \rho^2 \dot{\psi} = \omega$$

and

$$\left[ \frac{1}{2} \left( \dot{\rho}^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) + f(d) \right]' = -\frac{1}{r} \left( \dot{\rho}^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) \quad \text{in} \ D'(a,b). \quad (3.27)$$

The following results are based on a similar example studied by Bauman et al. in [14]. Their example, however, does not include any twists as the domain studied is a ball rather than an annulus. Using their techniques we are able to show that:

**Proposition 3.8.** The radial map $\rho$ belongs to the class $C([a,b]) \cap C^3(a,b)$ and solves the Euler-Lagrange equations (3.1).

**Proof.** Similarly to the previous section, we already know that $\rho \in W^{1,2}(a,b)$ and that it therefore has an absolutely continuous representative. Also, since $d > 0$ a.e. in $[a,b]$, we find that $\rho^2$ is monotonically increasing, which thereby implies that $\rho \geq a$. Moreover, since the right hand side of (3.27) can be estimated by $|\nabla u|^2$, which is integrable, it is integrable itself. Therefore $\frac{1}{2} \left( \dot{\rho}^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) + f(d)$ has an absolutely continuous representative $\sigma : (a,b) \to \mathbb{R}$. Now define the auxiliary function

$$g(q,\rho,r) = \frac{1}{2} \left( q^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) + f \left( \frac{\rho}{r} q \right).$$

By the properties of the function $h$ we find that $f'(d) = h''(d)d > 0$ and $f(d) \to -\infty$ as $d \to 0^+$. Thus, $g_q = q + f' \left( \frac{\rho}{r} q \right) \frac{\rho}{r} > 0$ and $g(q,\rho,r) \to -\infty$ as $q \to 0^+$ and $g(q,\rho,r) \to \infty$ as $q \to \infty$ for $q > 0$, $\rho > 0$ and $r > 0$. Hence, we can redefine $\dot{\rho}$ on a set of measure zero s.t.

$$\sigma(r) = g(\dot{\rho}(r),\rho(r),r).$$
Now we want to show that \( \dot{\rho} \) is itself continuous on \( (a, b) \). For this, let \( \bar{r} \) be a fixed number in \( (a, b) \) and \( (r_i)_i \subseteq (a, b) \) s.t. \( r_i \to \bar{r} \). Since \( d > 0 \) we know \( \dot{\rho}(r_i) > 0 \). Assume for a contradiction that \( \limsup_{i \to \infty} \dot{\rho}(r_i) = \infty \). Then there is a subsequence s.t.

\[
\sigma(r_{i_j}) = g(\dot{\rho}(r_{i_j}), \rho(r_{i_j}), r_{i_j}) \to \infty
\]
as \( j \to \infty \). However, \( \sigma \) is continuous at \( \bar{r} \), so this is not possible. Therefore \( \dot{\rho}(r_i) \) is a bounded sequence and possesses a convergent subsequence \( \dot{\rho}(r(i_j)) \). Let \( l \) be the limit of that subsequence. Then \( l > 0 \) for the same reasons as above. Furthermore, since \( g \) and \( \sigma \) are continuous

\[
g(\dot{\rho}(\bar{r}), \rho(\bar{r}), \bar{r}) = \sigma(\bar{r}) = \lim_{j \to \infty} \sigma(r_{i_j}) = \lim_{j \to \infty} g(\dot{\rho}(r_{i_j}), \rho(r_{i_j}), r_{i_j})
\]

\[
= g(l, \rho(\bar{r}), \bar{r})
\]

Since \( g(\cdot, \rho(\bar{r}), \bar{r}) \) is a homeomorphism between \( (0, \infty) \) and \( (0, \infty) \), it follows that \( l = \dot{\rho}(\bar{r}) \). The above argument is true for any convergent subsequence \( \dot{\rho}(r(i_j)) \) and the limit has to equal \( \dot{\rho}(\bar{r}) \). Hence, the whole sequence converges with \( \lim_{i \to \infty} \dot{\rho}(r_i) = \dot{\rho}(\bar{r}) \), and so \( \dot{\rho} \) is continuous on \( (a, b) \).

It remains to prove that also \( \ddot{\rho} \) and \( \dddot{\rho} \) exist and are continuous. From Lemma 3.7 we have

\[
g(\dot{\rho}(r), \rho(r), r) = -\int_0^r \frac{1}{s} \left( \dot{\rho}(s) + \frac{\omega^2}{s^2 \rho^2(s)} - \frac{\rho^2(s)}{s^2} \right) ds.
\]

Note, that the right hand side is a continuously differentiable function. Furthermore, \( g(q, \rho, r) \) is differentiable in \( q \), so the function \( \tilde{g}(r, q) := g(q, \rho(r), r) - \int_0^r \frac{1}{s} \left( \dot{\rho}(s) + \frac{\omega^2}{s^2 \rho^2(s)} - \frac{\rho^2(s)}{s^2} \right) ds \) is continuously differentiable with \( \tilde{g}_q > 0 \). Thus, it follows by the implicit function theorem that \( \dot{\rho} \) is continuously differentiable. Additionally, since \( f \in C^2(0, \infty) \) this implies that \( \tilde{g} \) is \( C^2(a, b) \) and so is \( \dot{\rho} \). Therefore, \( \rho \in C^3(a, b) \).

The Euler-Lagrange equations (3.1), i.e. \( \Delta u + (\text{cof } \nabla u) \nabla (h'(d)) = 0 \), for the rotationally symmetric map are equivalent to

\[
[r \dot{\rho} + \rho h'(d)]' = \frac{\rho}{r} + r \rho \dot{\psi}^2 + \dot{\rho} h'(d)
\]

and

\[
r \rho^2 \dot{\psi} = \omega,
\]

(3.28)
where the last equation obviously holds. Inserting this into the first, multiplying it by \( r \dot{\rho} \), which is positive on \((a, b)\), and appropriately grouping terms one can recover (3.27). \(\square\)

Ultimately we want to prove that \( \dot{\rho} \) does not vanish at the inner boundary, as in the previous case. For that the following result will be useful:

**Lemma 3.9.** Let \( N \in \mathbb{N} \). Then the function \( d \) is strictly monotonically increasing on \((a, b)\). The function \( z = \frac{1}{2} |\nabla u|^2 + f(d) \) is strictly monotonically decreasing on \((a, b)\).

**Proof.** From the previous lemma we have that both quantities are differentiable. Assume for a contradiction that \( \dot{d} \leq 0 \). Then

\[
\ddot{\rho} \leq \frac{1}{\rho} (d - \rho^2).
\]

(3.29)

Furthermore, the Euler-Lagrange equations (3.28) are equivalent to

\[
\ddot{\rho} \left( r + \frac{\rho^2}{r} h''(d) \right) = \frac{1}{r} \left( \rho + \frac{\omega^2}{\rho^3} \right) - \dot{\rho} + \frac{\rho}{r} (d - \rho^2) h''(d)
\]

The factor \( r + \frac{\rho^2}{r} h''(d) \) is always positive, so we can use (3.29) on the left hand side to obtain

\[
\dot{\rho} + \frac{r}{\rho} (d - \rho^2) \geq \frac{1}{r} \left( \rho + \frac{\omega^2}{\rho^3} \right).
\]

Multiplying this through by \( \frac{\rho}{r} \) we can then deduce that

\[
- \left( \dot{\rho} - \frac{\rho}{r} \right)^2 \geq \frac{\omega^2}{r^2 \rho^2}
\]

which is impossible since \( \omega \neq 0 \).

For \( z \) we have

\[
z = \frac{1}{2} \left( \rho^2 + \rho^2 \dot{\psi}^2 + \frac{\rho^2}{r^2} \right) + f(d) = \frac{1}{2} \left( \dot{\rho}^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) + f(d) + \frac{\rho^2}{r^2}
\]

Differentiating and using (3.27) we find

\[
\dot{z} = -\frac{1}{r} \left( \rho^2 + \frac{\omega^2}{r^2 \rho^2} - \frac{\rho^2}{r^2} \right) + 2 \frac{\rho \dot{\rho}}{r^2} - \frac{2 \rho^2}{r^3}
\]

\[
= -\frac{1}{r} \left( \rho^2 + \frac{\omega^2}{r^2 \rho^2} + \frac{\rho^2}{r^2} - 2 \frac{\rho \dot{\rho}}{r} \right) = -\frac{1}{r} \left( \left( \dot{\rho} - \frac{\rho}{r} \right)^2 + \frac{\omega^2}{r^2 \rho^2} \right) < 0
\]
Now we are in the position to prove:

**Proposition 3.10.** Let \( N \in \mathbb{N} \). Then \( \dot{\rho} \in C([a, b]) \) with \( \dot{\rho}(a) > 0 \) and \( \dot{\rho}(b) < \infty \).

**Proof.** Since \( d \) is monotonic on \((a, b)\) the limits \( d(r) \) for \( r \to a^+ \) and \( r \to b^- \) exist (possibly \(+\infty\) for \( r \to b^- \)). Therefore, the limits \( \dot{\rho}(r) \) for \( r \to a^+ \) and \( r \to b^- \) equally exist. Assume \( \dot{\rho}(a) = 0 \). Then \( z(a) = -\infty \) and is strictly monotonically decreasing. This is a contradiction. Hence, \( \dot{\rho}(a) > 0 \). Similarly \( \dot{\rho}(b) < \infty \). \( \square \)

As with the degenerate case in Section 3.2, we would like to prove that the rotationally symmetric minimiser is at least a local minimum in \( A_N \). In this case, and in contrast to the convex case from the previous section, we are in a better situation on the one hand since the candidates satisfy the Euler-Lagrange equations, but in a worse on the other, since the functional is now only polyconvex. There exist sufficient conditions derived from field theory of the calculus of variations for a solution to the Euler-Lagrange equation to be a strong local minimiser for problems with polyconvex integrands. For example, Sivaloganathan [56] proves this if a certain generalised Hamilton-Jacobi differential inequality is satisfied. Unfortunately, the work assumes that the Sobolev exponent \( p \) is bigger than the dimension \( n \) of the domain, but here we have \( p = n \).

Another possibility is again the symmetrisation procedure as attempted in Section 3.2. The term \( h(\det \nabla u) \) does not cause any problem, since \( h \) is convex. But the method suffers from the same problem as before in that information on the twisting is lost.

In the proof of the energy reduction of the symmetrised map in the original work 58, the function \( \frac{\xi}{r} \) plays an important part. In our problem we have found that it satisfies a maximum principle:

**Theorem 3.11.** Let \( N \in \mathbb{N} \). Then the function \( \frac{\xi}{r} \) attains no interior local maximum. Hence, \( (\frac{\xi}{r})^{-} \) changes sign only once and \( \frac{\alpha}{r} \leq \frac{\xi}{r} < 1 \) in \((a, b)\) with \( \frac{\xi}{r} = 1 \) at \( a, b \).

**Proof.** Assume there exists an \( r \in (a, b) \) s.t. \( (\frac{\xi}{r})^{-} = 0 \) and \( (\frac{\xi}{r})^{-} \leq 0 \). Then

\[
0 \geq \left( \frac{\rho}{r} \right)^{-} = -\frac{2}{r} \left( \frac{\rho}{r} \right) + \frac{1}{r} \dot{\rho} = \frac{\rho}{r} \ddot{\rho}.
\]  

(3.30)

However, by Theorem 3.9 we have

\[
0 < \dot{d} = \frac{1}{r} \left( \dot{\rho}^2 + \rho \ddot{\rho} - d \right) = \frac{\rho}{r} \ddot{\rho} + \dot{\rho} \left( \frac{\rho}{r} \right) = \frac{\rho}{r} \ddot{\rho}
\]

which contradicts (3.30). \( \square \)
4. \( n \)-polyconvexity

It is a well-known concept of convex analysis that finite-valued convex functions can be equivalently written as the pointwise supremum of affine functions, cf. [49], where we note that affine functions are those functions \( f \) where both \( f \) and \(-f\) are convex. Since polyconvex functions are nothing but convex functions acting on the list of minors the same idea can be applied here. The respective notion for the affine functions is that of the polyaffine functions, i.e. those functions \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) where both \( f \) and \(-f\) are polyconvex. However, the same concept cannot be applied to rank-one convex functions, since the current definition of a rank-one affine function, i.e. all functions \( f \) for which both \( f \) and \(-f\) are rank-one convex, is equivalent to polyaffine functions, cf. [20, Thm. 5.20]. Similarly, quasiaffine functions turn out be the equivalent to polyaffine functions. In [20] the author therefore decides to simply use the term quasiaffine for all three notions of poly-, rank-one- and quasiaffine functions. However, in this work we will use the term polyaffine for the above concepts for two reasons: Firstly, in the aforementioned duality between finite-valued polyconvex functions and their representation as the pointwise supremum of polyaffine functions the term quasiaffine could be confusing, and secondly and more importantly, there does exist a generalisation of this duality concept to rank-one convex functions which will involve a new definition of what it means to be rank-one affine. This definition, in contrast to the original one, is not equivalent to polyaffine functions. This new generalisation has two consequences that could potentially contribute to our understanding of quasiconvexity, which is one of the most important and yet least well understood concepts in the calculus of variations:

One is that this generalisation allows even more concepts to enter the picture. These will include the new convexity notion of \( n \)-polyconvexity, which will be defined rigorously later. At this stage it is worth pointing out that \( n \)-polyconvexity is a unifying concept for both polyconvexity and rank-one convexity and even includes previously unknown generalised convexities that lie in between those two. To be more precise, for a function \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\} \), we have that \((d \land D)\)-polyconvexity (recall that we define \( d \land D := \min\{d, D\} \)) corresponds to our standard definition of polyconvexity and that 1-polyconvexity corresponds to the function being rank-one convex. Additionally however,
there are the new concepts of 2-polyconvexity, 3-polyconvexity up to \((d \wedge D - 1)\)-polyconvexity. Thus, in dimension \(d = D = 3\) we will encounter the first new kind of 2-polyconvexity which is neither polyconvexity nor rank-one convexity. However, the usual implications of ‘\(n\)-polyconvexity implies \((n - 1)\)-polyconvexity’ will remain and thus add to the chain of implications. Since it is known that quasiconvexity lies itself between polyconvexity (i.e. \((d \wedge D)\)-polyconvexity) and rank-one convexity (i.e. 1-polyconvexity) and so the new \(n\)-polyconvexity notions for \(n = 2, \ldots, d \wedge D - 1\) may be studied with respect to quasiconvexity to see what they can teach us about quasiconvexity and vice versa.

The other consequence is that, since now we have found a way to generalise rank-one affine functions so that they allow us to construct a duality theory for rank-one convex functions, a similar idea may be applicable to construct a duality theory for quasiconvexity. However, more work is required to determine if such a generalisation is possible for quasiconvexity and, in particular, what the correct choice for ‘new’ quasiaffine functions would be.

### 4.1. Introducing \(n\)-polyconvexity through examples

The purpose of this section is to provide the reader with examples that will motivate subsequent rigorous definitions to follow in the next sections. Therefore, the explanations in this section are not meant to be complete or thorough, but should illustrate the main ideas as intuitively as possible.

Let us begin by recalling the definition of a rank-one convex function and therein focus solely on a particular point in \(\mathbb{R}^{d \times D}\). Let the function be \(f\) and the point be \(F\). Since \(f\) is rank-one convex it must be true that for any two points \(F_1, F_2\) on any rank-one line through \(F\), i.e. \(\text{rank}(F_1 - F_2) \leq 1\) and there exists \(\lambda \in (0, 1)\) s.t. \(F = \lambda F_1 + (1 - \lambda)F_2\), we have

\[
    f(F) \leq \lambda f(F_1) + (1 - \lambda) f(F_2).
\]

In the future we will call a rank-one line through \(F\) in direction \(u \otimes v\) for \(u \in \mathbb{R}^d, v \in \mathbb{R}^D\) the set \(F_{u \otimes v} := \{F + tu \otimes v : t \in \mathbb{R}\}\). Thus, if we were to restrict the function \(f\) to just one of these lines, we do indeed have the usual notion of standard convexity. In other words, \(f\) restricted to any rank-one line through \(F\) is convex and can be written as the pointwise supremum of affine functions on this line. (This of course holds for all points in \(\mathbb{R}^{d \times D}\), but we completely ignore this for now and keep focussing on the point
Assume for this illustration that for a given rank-one line, i.e. $u$ and $v$ fixed, the supremum is attained for the affine function $g_{u \otimes v}$. Note that $g_{u \otimes v}$ is only defined on the rank-one line through $F$ in direction $u \otimes v$. We can then write

$$g(F + tu \otimes v) = f(F) + tc_{u \otimes v}$$

for some constant $c_{u \otimes v}$. Therefore, if we consider the collection of all rank-one lines through $F$ we end up with a collection of gradients that are independent of each other. This is in contrast to polyaffine functions where this is not the case. Now denote by the rank-one cone at $F$ all matrices $\tilde{F} \in \mathbb{R}^{d \times D}$ with $\text{rank}(\tilde{F} - F) \leq 1$. Thus we are ready to define a function $g$ to be rank-one affine at $F$ if it is affine on all rank-one lines contained in the rank-one cone at $F$. An example of such a rank-one affine function $g$ is illustrated in Figure 4.1. Note that only lines in the rank-one cone at $F$ are depicted, since being rank-one affine at $F$ makes no assumptions about the function values of $g$ at points $\tilde{F}$ with $\text{rank}(\tilde{F} - F) \geq 2$. For the purpose of the drawing we may assume that the particular function $g$ that is depicted has the value $+\infty$ outside the set of matrices with difference up to rank one to $F$. We will return to this remark shortly. For now note furthermore that we have chosen $\mathbb{R}^{\tau(d,D)}$ as the domain of the function, where $\mathbb{R}^{\tau(d,D)}$ is the space containing $T(F)$, the vector of $F$ and all its minors, i.e. $T(F) = (F, \text{adj}_2 F, \ldots, \text{adj}_{d \wedge D} F) \in \mathbb{R}^{\tau(d,D)}$. Note that $\tau(d, D) = \sum_{s=1}^{d \wedge D} \binom{d}{s} \binom{D}{s}$ is the dimension of the space containing the list of minors $T(F)$ as each minor $\text{adj}_s F$ has dimension $\binom{d}{s} \binom{D}{s}$, $s = 1, \ldots, d \wedge D$. The change

Figure 4.1.: Depiction of a rank-one affine function on its rank-one cone (black and dashed lines) and comparison to a polyaffine plane (grey plane)
to $\mathbb{R}^{\tau(d,D)}$ is possible since the map $T$ of $F$ to all minors is linear along rank-one lines and therefore does not change the linearity of $g$. The only purpose of the change in reference space from $\mathbb{R}^{d \times D}$ to $\mathbb{R}^{\tau(d,D)}$ is to illustrate the difference to a polyaffine function which is depicted in grey in Figure 4.1. One can clearly see that the gradients of different rank-one lines through $F$ have a lot more freedom than if they were required to lie in the graph of polyaffine function, in which case they would have to be chosen so that all subgradients lie in the same plane.

Now we define the set $\mathbb{R}A(F)$ of all rank-one affine functions at $F$. For this purpose we call the set $C^1(F) = \{\bar{F} \in \mathbb{R}^{d \times D} : \text{rank}(\bar{F} - F) \leq 1\}$ the rank-one cone at $F$.

**Definition 4.1.** Let $F \in \mathbb{R}^{d \times D}$. Then we call $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ rank-one affine at $F$ if it is affine on all rank-one lines contained in $C^1(F)$. We denote by $\mathbb{R}A(F)$ the set of all rank-one affine functions at $F$.

Again, we would like to point out that our definition of rank-one affine functions is a localised concept, i.e. it depends on a point $F$. Although rank-one affine functions are defined on all of $\mathbb{R}^{d \times D}$ we do not know or care what their values are outside of the rank-one cone at $F$. Thus they may even assume the values $+\infty$ or $-\infty$ there, something which is clearly impossible for standard notions of affine functions, be it affine or polyaffine. In these cases this derives from the requirement that both $f$ and $-f$ are convex (resp. polyconvex) on the whole domain for $f$ to be affine (resp. polyaffine). It then follows that $f$ is finite everywhere. We find, using a similar argument, that any function $f$ in $\mathbb{R}A(F)$ must be finite on the rank-one cone at $F$. However, we stress that $f \in \mathbb{R}A(F)$ can take the value $\pm \infty$ (or indeed any other value) away from the cone at $F$ if desired.

We now want to verify that there is indeed a duality between rank-one convex functions and rank-one affine functions. One part of the duality is that functions that can be written as the pointwise supremum of rank-one affine functions at every point are indeed rank-one convex and we are already able to prove this result.

**Proposition 4.2.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\}$ and suppose

$$f(F) = \sup\{g(F) : g \in \mathbb{R}A(F) \text{ with } g \leq f\}.$$  

Then $f$ is rank-one convex.
Proof. Let $F \in \mathbb{R}^{d \times D}$ be fixed and $F_1, F_2 \in \mathbb{R}^{d \times D}$ such that $\text{rank}(F_1 - F_2) \leq 1$ and $F = \lambda F_1 + (1 - \lambda)F_2$ for some $\lambda \in [0, 1]$. Then, since all $g$ are affine on all rank-one lines through $F$ we have

$$f(F) = \sup_{g \in \mathbb{R}A(F)} g(F) = \sup_{g \leq f} \lambda g(F_1) + (1 - \lambda)g(F_2) \leq \sup_{g \in \mathbb{R}A(F)} \lambda f(F_1) + (1 - \lambda)f(F_2)$$

This part of the duality theory however is the easier one. Indeed this would also have worked with the original notion of rank-one affine functions (which are polyaffine). The more delicate part is that we need to be able to express rank-one convex functions as the pointwise supremum of rank-one affine functions at $F$. Clearly, this does not work with the original notion of rank-one affine functions, since any function as a pointwise supremum of those polyaffine functions would be polyconvex, and it is well known that there exist rank-one convex functions that are not polyconvex whenever $d, D \geq 2$. On the other hand we also know that in the extended real-valued case not all convex functions may be written as the pointwise supremum of affine functions, e.g. the characteristic function of the open set $(0, 1)$ which is defined as 0 on $(0, 1)$ and $+\infty$ everywhere else. Therefore, for the purpose of this result we only consider finite-valued functions. We then have the following result.

**Theorem 4.3.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ be rank-one convex. Then

$$f(F) = \sup\{g(F) : g \in \mathbb{R}A(F) \text{ with } g \leq f\}. \quad (4.2)$$

This is one of the main results of this work and the proof requires a lot of technical details. It will be proved in Section 4.2.3.

Last but not least we will now motivate the notion of $n$-polyconvexity. To this end we return to the observation that rank-one affine functions are now local to a point $F$ and are affine on all rank-one lines that are contained in the rank-one cone at $F$. Thus, this rank-one cone only includes matrices with difference of up to rank one to $F$, however, in $\mathbb{R}^{d \times D}$ we have matrices that differ to $F$ with up to rank $d \wedge D$. Therefore, we extend our definition of rank-one affine functions at $F$ to the following, denoting by
\[ C^n(F) = \{ \tilde{F} \in \mathbb{R}^{d \times D} : \text{rank}(\tilde{F} - F) \leq n \} \] the rank-n cone at \( F \):

**Definition 4.4.** Let \( F \in \mathbb{R}^{d \times D} \) and let \( 1 \leq n \leq d \wedge D \). Then we call \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) \( n \)-th order rank-one affine at \( F \) if it is affine on all rank-one lines contained in \( C^n(F) \). The set \( \mathbb{R}A_n(F) \) contains all \( n \)-th order rank-one affine functions at \( F \).

Clearly, we have that \( \mathbb{R}A_{d \wedge D}(F) \subseteq \mathbb{R}A_{d \wedge D-1}(F) \subseteq \ldots \subseteq \mathbb{R}A^2(F) \subseteq \mathbb{R}A^1(F) \). Furthermore, denote by \( \mathbb{P}A \) the set of all polyaffine functions. Then on \( \mathbb{R}^{d \times D} \) we have \( \mathbb{P}A = \mathbb{R}A_{d \wedge D}(F) \) for all \( F \in \mathbb{R}^{d \times D} \), since \( g \in \mathbb{R}A_{d \wedge D}(F) \) means precisely that \( g(\lambda F_1 + (1-\lambda)F_2) = \lambda g(F_1) + (1-\lambda)g(F_2) \) for all \( F_1, F_2 \in \mathbb{R}^{d \times D} \) with \( \text{rank}(F_1 - F_2) \leq 1 \), which is the original definition of rank-one affine functions and which are equivalent to polyaffine functions. Thus, in light of the known duality between finite-valued polyconvex functions and those that can be written as the pointwise supremum of polyaffine functions and the newly proposed one between finite-valued rank-one convex functions and those that can be written as the pointwise supremum of rank-one affine functions at \( F \), it becomes clear how \( n \)-polyconvexity is meant to act as a unifying concept between polyconvexity and rank-one convexity (namely via the duality of \( n \)-polyconvex functions and those that can be written as the pointwise supremum of \( n \)-th order rank-one affine functions). With the understanding that \( (d \wedge D) \)-th order rank-one affine functions are polyaffine and the known duality for polyconvex functions it should be conceivable that \( (d \wedge D) \)-polyconvex functions are equivalent to polyconvex functions. Similarly, through Proposition 4.2 and Theorem 4.3 we have that 1-polyconvex functions are equivalent to rank-one convex functions.

The first time that we encounter one of the new convexity conditions, i.e. one which is neither the usual polyconvex nor rank-one convex condition, is when we consider a three dimensional example \( (d = D = 3) \) where 2-polyconvexity arises additionally. The following example is therefore set in 3D space and chosen so that the \( n \)-polyconvex envelopes of the given function for \( n = 1, 2, 3 \) will all differ. Again, without introducing the envelopes formally, we expect to have

\[
f^{(d \wedge D)-pc} \leq f^{(d \wedge D-1)-pc} \leq \ldots \leq f^{2-pc} \leq f^{1-pc} \leq f,
\]

where \( f^{n-pc} \) denotes the \( n \)-polyconvex envelope of the function \( f \), i.e. the largest \( n \)polyconvex function below \( f \). As mentioned, the example is chosen so that additionally \( f^{3-pc} \neq f^{2-pc} \neq f^{1-pc} \neq f \), thereby also showing that 2-polyconvexity is in fact a new class of functions that does not coincide with the class of either polyconvex or rank-one convex functions. The function under consideration will have the value \(+\infty\) everywhere.
except for a few carefully selected points that all lie in the diagonal space

$$\text{Diag}(\mathbb{R}^{3 \times 3}) = \left\{ \begin{pmatrix} x & 0 & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{pmatrix} \mid x, y, z \in \mathbb{R} \right\}$$

and on which the function value is always zero. The function is depicted in Figure 4.2, where only the diagonal space is shown and the points with zero energy are black dots with labels $A_1, \ldots, A_8$, $B_1, \ldots, B_4$ and $C_1$ and $C_2$. The grey lines do not have any meaning except to guide the reader’s eye to understand their three dimensional relations to each other. The coordinates of the points $B_i$ and $C_i$ in $\mathbb{R}^3 = \{(x, y, z) : x, y, z \in \mathbb{R}\}$ are

- $B_1 = (3, 3.5, 5.5)$
- $B_2 = (3, 5.5, 4.5)$
- $B_3 = (3, 4.5, 2.5)$
- $B_4 = (3, 2.5, 3.5)$
- $C_1 = (-6, -3, -3)$
- $C_2 = (-4, -3, -3)$

It is not important to know the exact coordinates of the points $A_i$, but it should be noted that they were obtained by rotating the coordinates of the corners of the cube with centre zero and edge length two (e.g. it includes the opposite corners $(1, 1, 1)$ and $(-1, -1, -1)$) so that none of the points coincide with any of their coordinates with another. In the space of diagonal matrices this corresponds to each of the matrices $A_i$ having difference of rank three to $A_j$, $i \neq j$, i.e. $\text{rank}(A_i - A_j) = 3$, for all $i \neq j$. Then we define the zero set of $f$ by

$$K = \{ \text{diag}(A) : A = A_i, i = 1, \ldots, 8 \text{ or } A = B_i, i = 1, \ldots, 4 \text{ or } A = C_1, C_2 \}$$

and $f : \mathbb{R}^{3 \times 3} \to \mathbb{R} \cup \{+\infty\}$ by

$$f(F) = \begin{cases} 0, & F \in K \\ +\infty, & \text{otherwise}. \end{cases}$$

When we talk about the points $A_i$, $B_i$ and $C_i$ now we refer to them in the diagonal space as matrices, rather than vectors in $\mathbb{R}^3$. Then we see that the points $C_1$ and $C_2$ only differ by rank one, the points $B_1, \ldots, B_4$ all differ by rank two and the points $A_1, \ldots, A_8$ all differ by rank three due to the way they were rotated. Additionally, because of the way
Figure 4.2.: Example of a function such that $f^{3\text{-pc}} \neq f^{2\text{-pc}} \neq f^{1\text{-pc}} \neq f$. 
the groups \( \{ A_i \} \), \( \{ B_i \} \), and \( \{ C_i \} \) were placed in space they all have a difference of rank three to each other, e.g. \( \text{rank}(A_i - B_j) = 3 \) for all \( i = 1, \ldots, 8 \) and \( j = 1, \ldots, 4 \). This can also be seen in Figure 4.2 where rank-one connections can only exist on lines parallel to the \( x, y \) or \( z \)-axis. However, from the projections 4.2b, 4.2c and 4.2d it should be clear that always three such lines would be necessary to connect any of the points of the above groups with any other point of a different group.

We now discuss the different \( n \)-polyconvex envelopes of \( f \). Since \( C_1 \) and \( C_2 \) are rank-one connected \( f \) cannot be rank-one convex, thus we immediately obtain \( f^{1-pc} \neq f \). In fact, the zero set \( K^1 \) of \( f^{1-pc} \) now contains the connection between \( C_1 \) and \( C_2 \), i.e.

\[
K^1 = K \cup \{ \text{diag}(\lambda C_1 + (1 - \lambda)C_2) : \lambda \in (0, 1) \}
\]

and \( f^{1-pc} \) is defined analogously to \( f \) in that \( f^{1-pc} \equiv 0 \) on \( K^1 \) and \( f^{1-pc} \equiv +\infty \) on \( \mathbb{R}^{3 \times 3} \setminus K^1 \). It is clear that the rank-one convex envelope of \( f \) must be zero at least on the set \( K^1 \) and since \( f^{1-pc} \) as defined is rank-one convex it must be the desired envelope. Let us now move on to the 2-polyconvex envelope. Notice that the group \( \{ B_i \} \) occupies a \( T_4 \) configuration, simply embedded into 3D space (these will be treated in greater detail in Section 4.3.1). Note that the \( x \)-coordinate of all points \( B_i \) is 3 and that all points \( B_i \) differ by rank two. If we only consider \( B_i \) in the \( yz \)-plane in 2D completely disregarding that the \( x \)-dimension exists, we know how the zero set \( K^2_B \) of the polyconvex envelope of the function that is zero on the set \( \{ B_i \} \) and \( +\infty \) elsewhere on the \( yz \)-plane looks like, see Figure 4.3 or refer to Appendix A.3 for more details. We then define the set \( K^2 \) as

\[
K^2 = K^1 \cup K^2_B
\]

and \( f^{2-pc} \) accordingly. Again, this is not a proof, but it should be intuitive that the zero
set of the 2-polyconvex envelope of \( f \) must at least contain \( K^2 \). Since 2-polyconvexity at a point disregards anything with difference of rank three to it the condition of 2-polyconvexity of the function \( f^{2-pc} \) can be checked on each subset of matrices of difference of up to rank two separately. As all other points, i.e. those in \( \{A_i\}_i \) and \( \{C_i\}_i \), have a difference of rank three to \( K^2_B \) it should be clear that \( f^{2-pc} \) is 2-polyconvex. Thus, we have that \( f^{2-pc} \) is the 2-polyconvex envelope of \( f \) and furthermore that \( f^{2-pc} \neq f^{1-pc} \). Finally, we consider the 3-polyconvex envelope \( f^{3-pc} \) of \( f \) and note that it is the standard polyconvex envelope. We denote its zero set by \( K^3 \). Given that now all points of all point sets \( \{A_i\}_i \), \( \{B_i\}_i \) and \( \{C_i\}_i \) have a difference of up to rank three to each other \( f^{3-pc} \) may have a very complicated zero set. However, all that we would like to show is that \( f^{3-pc} \neq f^{2-pc} \) and this is a simple task. The reason for choosing the points \( A_i \) as the corners of a (rotated) cube centred around zero was so that it holds that

\[
\frac{1}{8} \sum_{i=1}^{8} T(A_i) = 0 = T(0).
\]

(Recall Definition 2.10 for the map \( T \).) Then the polyconvex envelope of \( f \) must satisfy

\[
0 \leq f(0) \leq \frac{1}{8} \sum_{i=1}^{8} f(A_i) = 0 \quad \text{and so } 0 \in K^3. \quad \text{However, } 0 \notin K^2 \quad \text{and hence, } f^{3-pc} \neq f^{2-pc}.
\]

The above record is a roughly chronological in terms of the thought processes that were involved in the discovery of \( n \)-polyconvexity. In Section 4.2 we will take a slightly different start to the story and begin with introducing \( n \)-polyconvexity with inequalities that correspond to the one for polyconvexity, i.e.

\[
f(F) \leq \sum_{i=1}^{\tau(d,D)+1} \lambda_i f(F_i) \quad \text{(4.1)}
\]

where

\[
T(F) = \sum_{i=1}^{\tau(d,D)+1} \lambda_i T(F_i) \quad \text{(4.2)}
\]

for \( F_i \in \mathbb{R}^{d \times D}, i = 1, \ldots, \tau(d,D) + 1 \) and \( \lambda \in \Lambda_{\tau(d,D)+1} \), or that of rank-one convexity

\[
f(F) \leq \sum_{i=1}^{2} \lambda_i f(F_i) \quad \text{(4.3)}
\]
where

\[ F = \sum_{i=1}^{2} \lambda_i F_i \]  \hspace{1cm} (4.4)

for \( \text{rank}(F_1 - F_2) \leq 1 \) and \( \lambda \in \Lambda_2 \). (Refer to Definition A.1 of Appendix A.1 for the definition of \( \Lambda_k \) for \( k \in \mathbb{N} \).) The analogous characterisation of \( n \)-polyconvexity will involve cosets \( F + V \subseteq \mathbb{R}^{d \times D} \) where \( F \in \mathbb{R}^{d \times D} \) and \( V \) is a subspace of \( \mathbb{R}^{d \times D} \) that will only include matrices of up to rank \( n \) and we will then have

\[ f(F) \leq \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i) \]

where

\[ T(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) \]

for \( F_i \in F + V \) and \( \lambda \in \Lambda_{\tau(n)} \). This should be understood as a hybrid approach between poly- and rank-one convexity in the sense that a directional element in the form of the cosets \( F + V \) is involved – similar to rank-one convexity (where we would choose \( V = \text{span}\{u \otimes v\} \) for \( u \in \mathbb{R}^d \) and \( v \in \mathbb{R}^D \) – while the convexity property is transferred to the space of minors in the spirit of polyconvexity. Furthermore, note that if \( n = d \wedge D \) we could take \( V = \mathbb{R}^{d \times D} \) and would then arrive at (4.1) and (4.2), and thus polyconvexity follows. For \( n = 1 \) and \( V = \text{span}\{u \otimes v\} \) for some \( u, v \in \mathbb{R}^d \) we have that \( F + V \) is simply a rank-one line through \( F \) and since on rank-one lines the minors map \( T \) is linear we indeed recover (4.3) and (4.4) of rank-one convexity. However, the conditions on the subspace \( V \) and the function \( \tau(n) \) are a little more intricate and will be dealt with in detail in Section 4.2.

### 4.2. Definition and basic properties of \( n \)-polyconvex functions

In this section we rigorously define the term \( n \)-polyconvexity and investigate the properties of \( n \)-polyconvex functions. Many of the results will be along the lines of similar results for polyconvexity and rank-one convexity that can be found in [20].

\( n \)-polyconvexity is founded on a particular type of subspace whose definition now follows.

**Definition 4.5.** \( V \subseteq \mathbb{R}^{d \times D} \) is called a simple subspace of maximal rank \( n \), or, for short,
simple rank-$n$ subspace, if there exist $u_1, \ldots, u_m \in \mathbb{R}^d$ and $v_1, \ldots, v_m \in \mathbb{R}^D$ s.t.
\[
V = \text{span}\{u_1 \otimes v_1, \ldots, u_m \otimes v_m\}
\]
and
\[
\max_{F \in V} \text{rank}(F) = n.
\]

Note that the terminology of simple is used because the subspace $V$ is spanned by rank-one matrices or tensors, and tensors of rank one are also called simple tensors. The fact that $V$ is a span of simple tensors plays a major part in the following theory. Allowing $V$ to be the span of other matrices, e.g. the span of the identity matrix, which has full rank, would ultimately fail as we will show in an example later. At this point we now want to proceed with the proper definition of $n$-polyconvexity and we are presented with a choice. The choice is between two approaches that define polyconvexity and rank-one convexity. Rank-one convexity is defined via the inequality
\[
f(F) \leq \lambda f(F_1) + (1 - \lambda) f(F_2)
\]
for all $F = \lambda F_1 + (1 - \lambda) F_2$ such that $\text{rank}(F_1 - F_2) \leq 1$, which then can be shown to be equivalent to $f$ being a convex function on each rank-one line $\{F + tu \otimes v\}_{t \in \mathbb{R}}$ for all $F \in \mathbb{R}^{d \times D}, u \in \mathbb{R}^d, v \in \mathbb{R}^D$, whereas polyconvexity is defined via the use of a convex representative $g$ such that $f(F) = g(T(F))$ for all $F \in \mathbb{R}^{d \times D}$, which in turn can then be shown to be equivalent to the inequality
\[
f(F) \leq \sum_{i=1}^I \lambda_i f(F_i)
\]
for all $T(F) = \sum_{i=1}^I \lambda_i T(F_i)$. To define $n$-polyconvexity we could thus choose either the inequality setting or the convex representative setting. We choose the approach via inequalities purely because this is how convexity was defined historically.

**Definition 4.6.** Let $F \in \mathbb{R}^{d \times D}$. Then $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ is called $n$-polyconvex at $F$ if for all simple rank-$n$ subspaces $V \subseteq \mathbb{R}^{d \times D}$ there exists $\beta_V \in \mathbb{R}^{T(d,D)}, c_V \in \mathbb{R}$ such that $f \geq \langle \beta_V, T(\cdot) \rangle + c_V$ on $V$ and it holds that
\[
f(F) \leq \sum_{i=1}^I \lambda_i f(F_i)
\]
(4.5)
for all $I \in \mathbb{N}, \lambda \in \Lambda_I$ and $F_i \in F + V$ such that $T(F) = \sum_{i=1}^I \lambda_i T(F_i)$.

Note that the class of functions that are $n$-polyconvex at a point $F$ is extremely large and in general can be very ill-behaved. For example in $\mathbb{R}$ we have that $f : \mathbb{R} \to \mathbb{R}$ being
1-polyconvex at 0 simply implies that $f(0) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$ for all $x_1, x_2 \in \mathbb{R}$ such that $0 = \lambda x_1 + (1 - \lambda)x_2$. However, this is satisfied by any function $f : \mathbb{R} \to \mathbb{R}$ such that $f(0)$ and $f \geq 0$, so even the characteristic function $\chi_Q$ of the rational numbers ($\chi_Q(x) = 0$ if $x \in Q$ and $\chi_Q(x) = +\infty$ if $x \in \mathbb{R} \setminus Q$) is rank-one convex at 0. The true power of this definition will become apparent in Section 4.2.4, but it should be remarked that Ball [7] also used the notion of rank-one convexity at a point $F$ to define rank-one convexity. Thus, this leads to:

**Definition 4.7.** $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ is called $n$-polyconvex if $f$ is $n$-polyconvex at $F$ for all $F \in \mathbb{R}^{d \times D}$.

Putting Definition 4.6 and 4.7 together we immediately obtain an equivalent statement for $n$-polyconvexity of $f$:

**Theorem 4.8.** $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ is $n$-polyconvex if and only if for all $F \in \mathbb{R}^{d \times D}$ and for all simple rank-$n$ subspaces $V \subseteq \mathbb{R}^{d \times D}$, there exist a convex function $c_{F+V} : \text{co}(T(F+V)) \to \mathbb{R}$ s.t.

$$f|_{F+V} \geq c_{F+V} \circ T|_{F+V} \quad (4.6)$$

and

$$f(F) \leq \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i) \quad (4.7)$$

with $F_1, \ldots, F_{\tau(n)+1} \in F + V$ s.t.

$$T(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i), \quad (4.8)$$

and where

$$\tau(n) = \sum_{s=1}^{n} \binom{d}{s} \binom{D}{s}. \quad (4.8)$$

**Remark 4.9.** Note that $F_1, \ldots, F_{\tau(n)+1} \in F + V$ ensures that rank$(F_i - F_j) \leq n$ for all $i, j \in \{1, \ldots, \tau(n) + 1\}$.

The main difference between this theorem and the definition of $n$-polyconvexity in Definition 4.7 (in particular when comparing to (4.5)) is that in the theorem only convex
combinations with at most $\tau(n) + 1$ elements need to be tested. Before we prove this theorem we also want to present another result which together with the above one corresponds to Part 1 of Theorem 5.6 from [20].

**Theorem 4.10.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$. Then $f$ is $n$-polyconvex if and only if for all simple rank-$n$ subspaces $V \subseteq \mathbb{R}^{d \times D}$ and for all cosets $F + V$ there exists a convex function $g_{F+V} : T(F + V) \to \mathbb{R} \cup \{+\infty\}$ s.t.

$$f|_{F+V} = g_{F+V} \circ T.$$ 

Note that the latter representation closely resembles the definition of polyconvexity as defined by Ball [4]. The proofs of both Theorem 4.8 and 4.10 require quite some technical detail. In order to avoid obscuring the main idea of this section, namely that $n$-polyconvexity unifies the notions of polyconvexity and rank-one convexity, we will postpone them to Section 4.2.1 and instead focus on the following result:

**Corollary 4.11.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$. Then it holds that

(i) $f$ is polyconvex if and only if it is $(d \wedge D)$-polyconvex,

(ii) $f$ is rank-one convex if and only if it is $1$-polyconvex.

**Proof.** In order to prove Corollary 4.11(i) we use Theorem 4.10. It is clear that $(d \wedge D)$-polyconvexity implies polyconvexity since $\mathbb{R}^{d \times D}$ is itself a simple rank-$(d \wedge D)$ subspace (it is spanned by the simple tensors $e_i \otimes e_j$ for $i = 1, \ldots, d, j = 1, \ldots, D$). Then $0 + \mathbb{R}^{d \times D} = \mathbb{R}^{d \times D}$ and we must have a convex function $g$ on $T(\mathbb{R}^{d \times D})$ s.t. $f = g \circ T$.

Now assume we have a polyconvex function $f$. Then there exists a convex $g$ on $\mathbb{R}^{\tau(d,D)}$ s.t. $f = g \circ T$. Therefore, if we now take a simple rank-$(d \wedge D)$ subspace $V$ and $F \in \mathbb{R}^{d \times D}$ (note that this subspace may not span the whole of $\mathbb{R}^{d \times D}$), then we can simply define $g_{F+V} := g|_{\text{co}(T(F+V))}$. Since $g$ is convex, its restriction to a convex set remains convex and thus we have found $g_{F+V}$.

In contrast, to prove Corollary 4.11(ii) we use Theorem 4.8. Assuming that $f$ is 1-polyconvex, we must then show that $f$ is rank-one convex. For this let $F_1, F_2 \in \mathbb{R}^{d \times D}$ such that $\text{rank}(F_1 - F_2) \leq 1$. Then $F_1 - F_2 = u \otimes v$ for some $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^D$. Further let $\lambda \in [0, 1]$ and $F \in \mathbb{R}^{d \times D}$ such that $F = \lambda F_1 + (1 - \lambda)F_2$. By taking $V = \text{span}\{u \otimes v\}$ we have that $F_1, F_2 \in F + V$. Furthermore, since the minors map $T$ is affine on rank-one convex lines we also immediately obtain that $T(F) = \lambda T(F_1) + (1 - \lambda)T(F_2)$.
Thus, we are in the position of applying the properties of 1-polyconvexity to yield
\[ f(F) \leq \lambda f(F_1) + (1 - \lambda) f(F_2), \]
which is the familiar condition of rank-one convexity.

The reverse implication is not entirely obvious since we now allow more than two
elements \( F_i \) to take part in the convex combination \((4.5)\). Let \( V \) be a simple rank-1
subspace and \( F \in \mathbb{R}^{d \times D} \). Further let \( F_1, \ldots, F_k \in F + V \) and \( \lambda \in \Lambda_k \) s.t. \((4.8)\) holds. We
now want to show by induction that then \((4.7)\) must hold for all \( k \in \mathbb{N} \). Clearly, it is true
for \( k = 2 \). Thus, assume it holds for \( k - 1 \) and that \( \lambda_k \neq 0 \) (otherwise there is nothing to
show). We have in particular that
\[
F = \sum_{i=1}^{k} \lambda_i F_i = (1 - \lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} F_i + \lambda_k F_k.
\]
Note that \( \left( \frac{\lambda_1}{1 - \lambda_k}, \ldots, \frac{\lambda_{k-1}}{1 - \lambda_k} \right) \in \Lambda_{k-1} \). If we can show that \( \tilde{F}_{k-1} \) and \( F_k \) are rank-one
connected, then we can use rank-one convexity of \( f \) and the inductive hypothesis on \( \tilde{F}_{k-1} \)
to obtain
\[
f(F) \leq (1 - \lambda_k) f(\tilde{F}_{k-1}) + \lambda_k f(F_k)
\leq (1 - \lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} f(F_i) + \lambda_k f(F_k)
= \sum_{i=1}^{k} \lambda_i f(F_i),
\]
which is \((4.7)\) for arbitrary \( k \). Therefore, it is left to show that \( \tilde{F}_{k-1} \) and \( F_k \) are rank-one
connected. It is the case that
\[
\tilde{F}_{k-1} - F_k = \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} F_i - F_k = \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} (F_i - F) - (F_k - F) \in V,
\]
since the right-hand side is a linear combination of elements of \( V \). Thus this implies
\[
\text{rank}(\tilde{F}_{k-1} - F_k) \leq 1 \text{ since } V \text{ is a simple rank-1 subspace and hence } f \text{ is 1-polyconvex.}
\]
Note that this proof works since the image of any simple rank-1 subspace under the
minors map \( T \) is already a convex set (unlike simple rank-\( n \) subspaces for \( n > 1 \)).

In the polyconvex case we achieve the result that the inequality \((4.5)\) needs to hold
for at most \( \tau(d,D) + 1 \) matrices participating in the convex combination \( T(F) = \)
\[ \sum_{i=1}^{\tau(d,D)+1} \lambda_i T(F_i) \] with \( F_i \in \mathbb{R}^{d \times D} \). When we consider rank-one convexity the choice of \( F_i \) are limited to a rank-one line through \( F \). However, by a notion that was introduced by Dacorogna \[20\] cf. Def. 5.14 and which he refers to as \((H_I)\), we can remove the restriction of the rank-one line at the expense of losing an upper bound for the participating matrices. The condition \((H_I)\) is recursive in the sense that for any given set of matrices that satisfies \((H_I)\) for some \( I \) we can replace any of the matrices by two matrices that have the former as their rank-one average. The collection of matrices then satisfies \((H_{I+1})\). Due to its recursive nature we will refer to \((H_I)\) as a recursive rank-one convex combination. To be slightly more precise, it can be shown that a function \( f \) is rank-one convex if for all matrices \( F_1, \ldots, F_I \in \mathbb{R}^{d \times D} \) and \( \lambda \in \Lambda_I \) such that \((F_i, \lambda_i)_{1 \leq i \leq I}\) satisfy the special condition \((H_I)\) we have \( f(\sum_{i=1}^{I} \lambda_i F_i) \leq \sum_{i=1}^{I} \lambda_i f(F_i) \). In the following definition we will generalise this concept to suit the needs of \( n \)-polyconvexity.

**Definition 4.12.** Let \( I \) be an integer and \( \lambda \in \Lambda_I \). Let \( F_i \in \mathbb{R}^{d \times D} \) for \( 1 \leq i \leq I \). We say that \((\lambda_i, F_i)_{1 \leq i \leq I}\) satisfy the condition \((H^n_I)\) if

(i) either \( I \leq \tau(n) + 1 \) and there exists \( F \in \mathbb{R}^{d \times D} \) and a simple rank-\( n \) subspace \( V \) such that \( F_1, \ldots, F_I \in F + V \) with \( F = \sum_{i=1}^{I} \lambda_i F_i \),

(ii) or there exists \( 2 \leq \hat{\tau} \leq \min\{\tau(n) + 1, I - 1\} \), \( F \in \mathbb{R}^{d \times D} \) and a simple rank-\( n \) subspace \( V \subseteq \mathbb{R}^{d \times D} \) such that, up to a permutation, \( F_1, \ldots, \hat{F}_{\hat{\tau}} \in F + V \) with \( F = \sum_{i=1}^{\hat{\tau}} (\lambda_i F_i)/(\sum_{i=1}^{\hat{\tau}} \lambda_i) \) and if, for \( 2 \leq i \leq I - \hat{\tau} + 1 \), we define

\[
\hat{\lambda}_i = \sum_{i=1}^{\hat{\tau}} \lambda_i, \quad \hat{\lambda}_i = \lambda_i + \hat{\tau} - 1, \quad \hat{F}_i = F_i + \hat{\tau} - 1
\]

then \((\hat{\lambda}_i, \hat{F}_i)_{1 \leq i \leq I - \hat{\tau} + 1}\) satisfy \((H^n_{I-\hat{\tau}+1})\).

In such a case we call \((F_i, \lambda_i)_{1 \leq i \leq I}\) a recursive \( n \)-polyconvex combination.

Figure 4.4 illustrates the above definition of a recursive \( n \)-polyconvex combination. The labelled matrices \( F_i, i = 1, \ldots, 12 \), together with appropriate coefficients \( \lambda \in \Lambda_{12} \) satisfy \((H^n_{12})\). For example, for \( \hat{\tau} = 4 \) the group \((\lambda_i, F_i)_{i=5,\ldots,8}\) defines the new point \( F_{5678} = (\sum_{i=5}^{8} \lambda_i F_i)/\lambda_{5678} \) (indicated by the circle in the shaded plane) with \( \lambda_{5678} = \lambda_5 + \lambda_6 + \lambda_7 + \lambda_8 \). Then \((\lambda_{5678}, F_{5678})\) and the remaining pairs satisfy \((H^n_9)\). With this definition we obtain the following proposition, which compares to Prop. 5.16 in \[20\].

**Proposition 4.13.** Let \( f : \mathbb{R}^{d \times D} \to \overline{\mathbb{R}} \). Then \( f \) is \( n \)-polyconvex if and only if \( f \) is
bounded below by a n-polyconvex function and the expression

\[ f \left( \sum_{i=1}^{I} \lambda_i F_i \right) \leq \sum_{i=1}^{I} \lambda_i f(F_i) \]  

(4.9)

holds whenever \((\lambda_i, F_i)_{1 \leq i \leq I}\) satisfy \((H^n_I)\).

It is trivial to prove that \(f\) is \(n\)-polyconvex if (4.9) is true for all \((\lambda_i, F_i)_{1 \leq i \leq I}\) satisfying \((H^n_I)\). The reverse implication follows the same reasoning as for the case of rank-one convexity \((n = 1)\) and we will omit it as it does not provide any new insights. The proposition will be useful later since it provides an intuition as to how to compute the \(n\)-polyconvex envelope of a given function (see Section 4.4.1).

4.2.1. Proofs of Theorem 4.8 and 4.10

In both proofs we will largely follow the proof given by Dacorogna [20, Th. 5.6] which we generalise to the case of \(n\)-polyconvexity. A part of the proof is to show that \(\text{co}(T(R^{d \times D})) = R^{\tau(d,D)}\). In our case this will take a more general form and we will present it as a separate result.

**Lemma 4.14.** Let \(V\) be a simple rank-\(n\) subspace of \(R^{d \times D}\) and \(F \in R^{d \times D}\). Then there exists a subspace \(T_V \subseteq R^{d \times D}\) s.t.

\[ \text{co}(T(F + V)) = T(F) + T_V, \]

68
i.e. the convex hull of the minors of the coset \( F + V \subseteq \mathbb{R}^{d \times D} \) is itself a coset in \( \mathbb{R}^{\tau(d,D)} \).

Furthermore, \( \dim(T_V) \leq \tau(n) := n \sum_{i=1}^{d} (\binom{D}{i}) \) (with \( n \leq d \wedge D \)).

**Proof.** The proof consists of two parts. The first part is to show that \( \text{co}(T(F+V)) \) is an affine space and the second is that the subspace \( T_V \) defining this affine space has dimension less than \( \tau(n) = \sum_{i=1}^{n} (\binom{d}{i}) \).

Define

\[
T_V := \text{span}\{T(F_V) - T(F) : F_V \in F + V\}.
\]

We then show that \( C := \text{co}(T(F + V)) - T(F) = T_V \). It is easy to see that \( C \subseteq T_V \) since \( C = \text{co}(T(F + V)) - T(F) = \text{co}(T(F + V) - T(F)) \) and the latter representation naturally lies in the span of \( T(F + V) - T(F) \). Therefore, assume for a contradiction that \( C \neq T_V \). Then by the separation theorem [A.2] there exists \( \alpha \in T_V, \alpha \neq 0 \) and \( \beta \in \mathbb{R} \) s.t.

\[
\langle \alpha, X \rangle \leq \beta \quad \text{for all } X \in C. \quad (4.10)
\]

Due to the way \( T_V \) was defined there also exists \( F_V \in F + V \) s.t.

\[
\langle \alpha, T(F_V) - T(F) \rangle \neq 0.
\]

(Otherwise \( \langle \alpha, X \rangle = 0 \) for all \( X \in T_V \), implying that \( \alpha = 0 \) since \( \alpha \) is an element of \( T_V \) itself.) Since \( F_V \in F + V \) and \( V \) is simple subspace there exist \( u_i, v_i \in \mathbb{R}^d, \lambda_i \in \mathbb{R}, i = 1, \ldots, m \) s.t.

\[
F_V = F + \sum_{i=1}^{m} \lambda_i u_i \otimes v_i.
\]

Now define the matrices

\[
F_l := F + \sum_{i=1}^{l} \lambda_i u_i \otimes v_i
\]

for \( l = 0, \ldots, m \). Then \( F_l \) and \( F_{l-1} \) are rank-one connected for \( l = 1, \ldots, m \). Furthermore \( F_m = F_V \) and \( F_0 = F \) and thus

\[
\langle \alpha, T(F_m) - T(F) \rangle \neq 0
\]
and
\[ \langle \alpha, T(F_0) - T(F) \rangle = 0. \]

Therefore there exists \( \hat{l} \in \{1, \ldots, m\} \) s.t. \( \langle \alpha, T(F_{\hat{l}}) \rangle \neq \langle \alpha, T(F_{\hat{l}-1}) \rangle \). We then define
\[ \hat{F}(\lambda) = F_{\hat{l}-1} + \lambda u_{\hat{l}} \otimes v_{\hat{l}} \]
for all \( \lambda \in \mathbb{R} \). Thus, \( \hat{F}(\lambda) \in F + V \) for all \( \lambda \in \mathbb{R} \) and \( \hat{F}(0) = F_{\hat{l}-1} \) and \( \hat{F}(1) = F_{\hat{l}} \). Since the map to minors is affine on rank-one lines we then obtain
\[ T(\hat{F}(\lambda)) = T(F_{\hat{l}-1}) + \lambda \left( T(F_{\hat{l}}) - T(F_{\hat{l}-1}) \right). \]

Finally, since \( T(\hat{F}(\lambda)) - T(F) \in C \) for all \( \lambda \in \mathbb{R} \) and (4.10) we must have
\[ \beta \geq \langle \alpha, T(\hat{F}(\lambda)) - T(F) \rangle \\
= \langle \alpha, T(F_{\hat{l}-1}) - T(F) \rangle + \lambda \left( \langle \alpha, T(F_{\hat{l}}) \rangle - \langle \alpha, T(F_{\hat{l}-1}) \rangle \right) \neq 0 \]

for all \( \lambda \in \mathbb{R} \). This is clearly a contradiction since this cannot be true for all \( \lambda \in \mathbb{R} \).

It remains to show that \( \dim(T_V) \leq \tau(n) \). Again, we will defer the proof of this to a later stage, namely Section 4.2.2 when we learn more about the structure of the space \( T_V \).

Note that it is an integral part of the proof that \( V \) is a simple subspace. The following example shows that the result is false if that condition is violated.

**Example 4.15.** Consider \( \mathbb{R}^{2 \times 2} \) and let \( V = \text{span}\{e_1 \otimes e_1 + e_2 \otimes e_2\} = \left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \mid a \in \mathbb{R} \right\} \).

Then
\[ T(V) = \left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, a^2 \right\} \subseteq \mathbb{R}^5 \]

Since \( a^2 \), the last entry of \( T(V) \), is always positive we also have
\[ \text{co}(T(V)) \subseteq \mathbb{R}^{2 \times 2} \times \mathbb{R}_+ \]

and so it cannot be a subspace.
Proof of Theorem 4.8. The implication that (4.7) holds if \( f \) is \( n \)-polyconvex is trivial, since it would then hold for the particular choice \( I = \tau(n) + 1 \). We now want to show that \( I \) can always be taken to be equal to \( \tau(n) + 1 \). Thus, let \( X \in \text{co}(T(F + V)) \), \( F_i \in F + V \) and \( \lambda_i \in \Lambda_I \) s.t.

\[
X = \sum_{i=1}^{I} \lambda_i T(F_i).
\]

Similarly to Dacorogna’s proof we first show that \( I = \tau(n) + 2 \) is sufficient. We define

\[
T(\text{epi} f|_{F + V}) := \{(T(F), \mu) \in T(F + V) \times \mathbb{R} : f(F) \leq \mu \} \subseteq \text{co}(T(F + V)) \times \mathbb{R},
\]

where now \( \text{co}(T(F + V)) \times \mathbb{R} = (T(F) + T_V) \times \mathbb{R} = (T(F), 0) + T_V \times \mathbb{R} \) with \( T_V \) defined as in Lemma 4.14. Recall that \( T_V \) is a subspace of \( \mathbb{R}^{d \times D} \) with dimension no more than \( \tau(n) \). Thus, \( \text{co}(T(F + V)) \times \mathbb{R} \) is an affine space with dimension \( \dim(T_V) + 1 \leq \tau(n) + 1 \). Denote \( \hat{\tau} = \dim(T_V) \). Then by applying the usual Carathéodory Theorem, see Theorem A.3, for the coset we obtain \( I = \hat{\tau} + 2 \). A further step in the proof is to show that this number can be further reduced to \( I = \hat{\tau} + 1 \). The reasoning is completely analogous to the proof given in Theorem 5.6 in [20] and hence we will omit it here. In Lemma 4.14 we claim that \( \hat{\tau} \leq \tau(n) = n^n \binom{D}{2} \) (proof to follow) so for the purposes of the theorem we may choose \( I = \tau(n) \).

Proof of Theorem 4.10. The implication ‘\( \Leftarrow \)’ is relatively straightforward. Let \( F \in \mathbb{R}^{d \times D} \) and \( V \) be a simple rank-\( n \) subspace of \( \mathbb{R}^{d \times D} \). Then there exists a function \( g_{F + V} \), s.t. \( g_{F + V} \) is convex on \( \text{co}(T(F + V)) \subseteq \mathbb{R}^{\tau(d, D)} \) with \( f = g_{F + V} \circ T \) on \( F + V \). Thus, for \( c = g_{F + V} \) the inequality (4.6) is satisfied. Then for \( F_1, \ldots, F_{\tau(n) + 1} \in F + V \) satisfying (4.8) we use the convexity of \( g_{F + V} \) on its coset \( F + V \) and obtain (4.7).

The implication ‘\( \Rightarrow \)’ requires more work and we will use the results of Theorem 4.8 to prove the assertion. Assume \( f \) is \( n \)-polyconvex. Then (4.7) holds for all simple rank-\( n \) subspaces \( V \subseteq \mathbb{R}^{d \times D} \) and \( F + V \in \mathbb{R}^{d \times D}/V \) and \( F_1, \ldots, F_{\tau(n) + 1} \in F + V \) satisfying (4.8) and let \( V \) and \( F + V \) be fixed. We then need to show that there exists a convex function \( g_{F + V} : \text{co}(T(F + V)) \rightarrow \mathbb{R} \cup \{+\infty\} \) with \( f|_{F + V} = g_{F + V} \circ T \). Let \( I \geq \tau(n) + 1 \) be an integer and define the function \( g_I : \text{co}(T(F + V)) \rightarrow \mathbb{R} \cup \{+\infty\} \) such that

\[
g_I(X) = \inf \left\{ \lambda_i f(F_i) : \sum_{i=1}^{I} \lambda_i = 1, \lambda_i \geq 0, \sum_{i=1}^{I} \lambda_i T(F_i) = X \text{ and } F_1, \ldots, F_I \in F + V \right\}.
\]
Along the lines of the proof in [20] we will show that, without loss of generality, $I$ can be taken to be equal to $\tau(n)+1$. We then take $g_{F+V} = g_{\tau(n)+1}$ and show that $g_{F+V}$ is convex and satisfies $f(F) = g_{F+V}(T(F))$. Note that in the original version of this proof for the polyconvex case $g_I$ was defined on $\mathbb{R}^{\tau(d,D)}$ instead of its corresponding version for $n = d \wedge D$ here, where it is defined on $\text{co}(T(\mathbb{R}^{d \times D}))$. Using $g_I$ defined on $\mathbb{R}^{\tau(d,D)}$ requires to check whether $g_I$ is actually well defined, i.e. whether for each $X \in \mathbb{R}^{\tau(d,D)}$ there exist $I \in \mathbb{N}$ and $F_1, \ldots, F_I \in \mathbb{R}^{d \times D}$, $\lambda \in \Lambda_I$ s.t. $X = \sum_{i=1}^{I} \lambda_i T(F_i)$, or in other words, whether $\mathbb{R}^{\tau(d,D)} = \text{co}(T(\mathbb{R}^{d \times D}))$. We avoid this step since we define $g_I$ on $\text{co}(T(F+V))$ straight away.

We now show that $g_{F+V}$ is convex. Let $X, Y \in \text{co}T(F+V)$ and $\mu \in [0,1]$. We want to prove that

$$\mu g_{F+V}(X) + (1-\mu)g_{F+V}(Y) \geq g_{F+V}(\mu X + (1-\mu)Y).$$

(4.11)

Fix $\varepsilon > 0$. Then from the considerations above there exist $\lambda, \tilde{\lambda} \in \Lambda_{\tau(n)+1}$ and $F_i, \tilde{F}_i \in F+V$ s.t.

$$\mu g_{F+V}(X) + (1-\mu)g_{F+V}(Y) + \varepsilon \geq \mu \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i) + (1-\mu) \sum_{i=1}^{\tau(n)+1} \tilde{\lambda}_i f(\tilde{F}_i)$$

(4.12)

with

$$\sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) = X, \quad \sum_{i=1}^{\tau(n)+1} \tilde{\lambda}_i T(\tilde{F}_i) = Y.$$  
(4.13)

Upon redefining for $1 \leq i \leq \tau(n) + 1$

$$\tilde{\lambda}_i = \mu \lambda_i \quad \tilde{\lambda}_i = (1-\mu) \tilde{\lambda}_i$$

$$\tilde{F}_i = F_i \quad \tilde{F}_i = \hat{F}_i$$

both (4.12) and (4.13) can be written as

$$\mu g_{F+V}(X) + (1-\mu)g_{F+V}(Y) + \varepsilon \geq \sum_{i=1}^{2\tau(n)+2} \tilde{\lambda}_i f(\tilde{F}_i)$$

(4.14)
with $\tilde{\lambda} \in \Lambda_{2r(n)+2}$ and

$$
\sum_{i=1}^{2r(n)+2} \tilde{\lambda}_i T(\tilde{F}_i) = \mu X + (1 - \mu)Y.
$$

Then, taking the infimum over $(\tilde{\lambda}_i, \tilde{F}_i)$ in (4.14), and noticing that $\varepsilon$ was arbitrary we do indeed have (4.11), i.e. $g_{F+V}$ is convex.

The final step is to prove that $f \leq g_{F+V}$ on $T(F + V)$. Take $\tilde{F} \in F + V$. Because we assume (4.7) holds for all $F_1, \ldots, F_{\tau(n)+1} \in F + V$ such that (4.8) holds for $\tilde{F}$, taking the infimum on both sides of the inequality we immediately obtain that $f(\tilde{F}) \leq g(T(\tilde{F}))$.

Since also for $X = T(\tilde{F})$ in the evaluation of $g_{F+V}(X = T(\tilde{F}))$ a trivial candidate convex combination is $\tilde{F}$ itself, we also obtain $g_{F+V}(T(\tilde{F})) \leq f(\tilde{F})$, and hence, $f = g_{F+V} \circ T$.

The proof of Theorem 4.8 includes a reference to a particular choice of the convex representative $g_{F+V}$. This is the purpose of the following theorem.

**Theorem 4.16.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ be $n$-polyconvex. Then for any $F \in \mathbb{R}^{d \times D}$ and $V \subseteq \mathbb{R}^{d \times D}$ simple rank-$n$ we define $g_{F+V} : \text{co} T(F + V) \to \mathbb{R} \cup \{+\infty\}$ by

$$
\begin{align*}
g_{F+V}(X) := \inf \left\{ \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i) : \lambda \in \Lambda_{\tau(n)+1}, \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) = X, F_i \in F + V \right\}.
\end{align*}
$$

Then $g_{F+V}$ is convex on $T(F + V)$ and

$$
f(\tilde{F}) = g_{F+V}(T(\tilde{F}))
$$

for all $\tilde{F} \in F + V$. Moreover, for every $X \in \text{co} T(F + V)$

$$
g_{F+V}(X) = \sup \{ G(X) : G : \text{co} T(F + V) \to \mathbb{R} \cup \{+\infty\} \text{ convex and } f = G \circ T \text{ on } F + V \}.
$$

Note that $g_{F+V}$ defined by (4.15) is also called the Busemann representative of the function $\tilde{g}_{F+V} : T(F + V) \to \mathbb{R}$ with $\tilde{g}_{F+V}(T(\tilde{F})) = f(\tilde{F})$ for all $\tilde{F} \in F + V$. The function $\tilde{g}_{F+V}$ is defined on the nonconvex set $T(F + V)$ (if $n > 1$) and according to Busemann et al. [10] the convex representative may not be unique. However, this particular
choice of representative is the largest of all possible choices, which simply follows from
the definition of $g_{F+V}$ directly. To see this let $G : \text{co} T(F + V) \to \mathbb{R} \cup \{+\infty\}$ be another
representative, i.e. $f = G \circ T$ on $T(F + V)$ and $G$ convex. Then for $X \in \text{co} T(F + V)$
we have in particular that $G(X) \leq \sum_{i=1}^{(n)+1} \lambda_i f(F_i)$ for all $F_i \in F + V$ and $\lambda \in \Lambda_{r(n)+1}$
such $X = \sum_{i=1}^{(n)+1} \lambda_i F_i$. Hence, by taking the infimum we obtain $G(X) \leq g_{F+V}(X)$.
Since the first part of this proof is basically included in the proof of Theorem 4.8 there is
nothing more to show.

4.2.2. $n$-polyaffine functions

We now formalise the aforementioned concept of $n$-polyaffine functions. In the intro-
ductive section these were referred to as $n$-th order rank-one affine functions. The
reason for using the two different names will become apparent in Corollary 4.21. This
is a consequence of Theorem 4.20, which is also the main result of this section. The
underlying mechanism of the presented proof is not particular efficient and it could be
that the proofs could be shortened.

We now define the notion of $n$-polyaffines.

Definition 4.17. Let $F \in \mathbb{R}^{d \times D}$. Then $h : \mathbb{R}^{d \times D} \to \mathbb{R}$ is called $n$-polyaffine at $F$
if $h$ and $-h$ are $n$-polyconvex at $F$.

Furthermore, we define the notion of the rank-$n$ cone at $F$, which will be of frequent
use in this section.

Definition 4.18. Let $F \in \mathbb{R}^{d \times D}$. Then we call

$$C^n(F) = \{ \tilde{F} : \text{rank}(\tilde{F} - F) \leq n \}$$

the rank-$n$ cone at $F$.

The following lemma shows that $(d \wedge D)$-polyaffine functions at $F$ are nothing but the
well-known quasiaffine functions independently of $F$. However, we will see in the main
result of this section that $n$-polyaffine functions at $F$ are more general than quasiaffine
functions for $n < d \wedge D$ and in that they depend on $F$.

Lemma 4.19. Let $F \in \mathbb{R}^{d \times D}$ and $h : \mathbb{R}^{d \times D} \to \mathbb{R}$. Then $h$ is $n$-polyaffine at $F$ if and
only if $h$ is finite on the rank-$n$ cone $C^n(F)$ and it holds that

$$h(F) = \sum_{i=1}^{l} \lambda_i h(F_i) \quad (4.16)$$
for all \( F_1, \ldots, F_I \in F + V \) for some \( V \) simple rank-\( n \) with \( T(F) = \sum_{i=1}^I \lambda_i T(F_i) \), \( \lambda \in \Lambda_I \).

The proof of this lemma is straightforward and will be omitted. Recall that a function that is \( n \)-polyconvex at \( F \) may be rather badly behaved and the class of functions that are \( n \)-polyconvex at a given \( F \) is very large. On the contrary, the above lemma shows that this is not the case for \( n \)-polyaffine functions at \( F \) and that \( n \)-polyaffine functions at \( F \) are finite on the whole rank-\( n \) cone at \( F \). That the class of \( n \)-polyaffine functions at \( F \) is indeed much smaller and takes a familiar appearance is the topic of the main theorem of this section, which we present now.

**Theorem 4.20.** Let \( F \in \mathbb{R}^{d \times D} \) and \( h : \mathbb{R}^{d \times D} \to \mathbb{R} \). Then the following three statements are equivalent:

(i) \( h \) is \( n \)-polyaffine at \( F \),

(ii) \( h \) is affine on any rank-one line contained in \( \mathcal{C}^n(F) \),

(iii) there exists \( \beta \in \mathbb{R}^{\tau(d, D)} \) such that

\[
h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F)
\]

for all \( \tilde{F} \in \mathcal{C}^n(F) \).

Here we use the wording ‘affine on any rank-one line contained in \( \mathcal{C}^n(F) \)’. This is simply supposed to mean that whenever there is a rank-one line \( \{ A + tu \otimes v : t \in \mathbb{R} \} \) for some \( A \in \mathbb{R}^{d \times D} \) that this set is completely contained in the rank-\( n \) cone \( \mathcal{C}^n(F) \) centred at \( F \) with \( \mathcal{C}^n(F) = \{ F \in \mathbb{R}^{d \times D} : \text{rank}(F - F) \leq n \} \).

**Corollary 4.21.** Let \( h : \mathbb{R}^{d \times D} \to \mathbb{R} \) and \( F \in \mathbb{R}^{d \times D} \). Then \( h \) is \( n \)-polyaffine at \( F \) if and only if it is \( n \)-th order rank-one affine at \( F \).

Note once more that for \( n = d \wedge D \) the theorem is nothing but the well known result that all quasiaffine functions are those that are linear combinations of minors. This was first established for \( d = D = 1, 2, 3 \) in Ball [4] and proved in general by Dacorogna [20, Th. 5.20]. Further references on preceding work on this result for \( n = d = D \) can also be found in [20, Sec. 5.3.1]. The general proof presented by Dacorogna is very technical and uses induction over the number of rows in the matrices, i.e. over \( d \) in \( \mathbb{R}^{d \times D} \). This approach is not possible in our even more general case where \( n \) can be smaller than \( d \wedge D \) and our approach is necessarily more technical. Therefore we postpone the proof of the
theorem, following two technical lemmas and some instructive examples. However, the main steps of the proof regarding the implication \((ii) \Rightarrow (iii)\) can be described as follows:

**Step 1.** Looking at equation \((4.17)\) it is obvious that it is equivalent to find \(\beta \in \mathbb{R}^T(d,D)\) such that

\[
\langle \beta, T(F + \sum_{i=1}^{n} u^i \otimes v^i) - T(F) \rangle = h(F + \sum_{i=1}^{n} u^i \otimes v^i) - h(F) \tag{4.18}
\]

for all \(u^i \in \mathbb{R}^d, v^i \in \mathbb{R}^D, i = 1, \ldots, n\). Similar to the matrix determinant lemma which asserts that \(\det(\mathbf{F} + u \otimes v) = \det(\mathbf{F}) + u^T \text{cof}(\mathbf{F})v\) we want to use the generalised formulas in Lemmas 4.22 and 4.25 of this identity to pull the coefficients \(u_i\) and \(v_i\) out and using them as independent variables. For example, in the case of \(n = 1\) and \(d = D = 2\) we will have to satisfy the equation

\[
2 \sum_{i,j=1}^{2} \langle \beta, T(\mathbf{F} + e_i \otimes e_j) - T(F) \rangle u_i v_j = 2 \sum_{i,j=1}^{2} [h(\mathbf{F} + e_i \otimes e_j) - h(F)] u_i v_j.
\]

**Step 2.** Noting that \(u_i\) and \(v_j\) are independent variables the above equation can only be true if the coefficients \(\langle \beta, T(\mathbf{F} + e_i \otimes e_j) - T(F) \rangle\) and \(h(\mathbf{F} + e_i \otimes e_j) - h(F)\) coincide, thus giving a system of linear equations. In Lemma 4.27 we derive the system of equations that needs to be satisfied for the more general equation \((4.18)\).

**Step 3.** Finally, in Lemma 4.28 we show that the system derived in Step 2 has a solution by carefully reducing the number of equations in the system and showing that it can be written in reduced row echelon form.

**Step 1.**

We begin by the presenting the first lemma to assist the proof.

**Lemma 4.22.** Let \(F \in \mathbb{R}^{d \times D}\) and \(\gamma : \mathbb{R}^{d \times D} \to \mathbb{R}^m\) be affine along rank-one lines contained in \(C^1(F)\). Then

\[
\gamma(F + u \otimes v) = \gamma(F) + \sum_{i=1}^{d} \sum_{j=1}^{D} [\gamma(F + e_i \otimes e_j) - \gamma(F)] u_i v_j \tag{4.19}
\]

holds for all \(u \in \mathbb{R}^d\) and \(v \in \mathbb{R}^D\).

**Proof.** Let the assumptions of the lemma hold. We now prove this lemma by induction over \(D\). For the base case assume \(D = 1\), i.e. \(\gamma : \mathbb{R}^d \to \mathbb{R}^m\). Then \(C^1(F) = \mathbb{R}^d\) since
rank-one convexity is not a restriction in the vectorial case and $\gamma$ is simply affine. It is then an easy task to verify that

$$
\gamma(F + u) = \gamma(F) + \sum_{i=1}^{d} \left[ \gamma(F + e_i) - \gamma(F) \right] u_i
$$

for all $u \in \mathbb{R}^d$.

Now assume (4.19) holds for $D$ and consider the case $D + 1$, so $\gamma : \mathbb{R}^{d \times (D+1)} \to \mathbb{R}^m$. For any $v \in \mathbb{R}^{D+1}$ we denote by $v^D$ the first $D$ entries of $v$ and by $v_{D+1}$ the last entry of $v$ (we use the superscript instead of subscript for $v^D$ solely so that it is notationally easier to access the components $v^D_i$). Then we can write

$$
v = \begin{pmatrix} v^D \\ v_{D+1} \end{pmatrix} = \begin{pmatrix} v^D \\ 0 \end{pmatrix} + v_{D+1}e_{D+1}.
$$

and using rank-one linearity we find

$$
\gamma(F + u \otimes v) = \gamma\left(F + u \otimes \begin{pmatrix} v^D \\ 0 \end{pmatrix} + v_{D+1} \otimes e_{D+1}\right)
= \gamma\left(F + u \otimes \begin{pmatrix} v^D \\ 0 \end{pmatrix}\right) + v_{D+1} \left[ \gamma(F + u \otimes e_{D+1}) - \gamma(F) \right]. \tag{4.20}
$$

We then define the auxiliary functions $\gamma_1 : \mathbb{R}^{d \times D} \to \mathbb{R}^m \cup \{+\infty\}$ s.t.

$$
\gamma_1(u \otimes v_1) = \gamma\left(F + u \otimes \begin{pmatrix} v_1 \\ 0 \end{pmatrix}\right)
$$

for all $u \in \mathbb{R}^d$ and $v_1 \in \mathbb{R}^D$ and $\gamma_2 : \mathbb{R}^d \to \mathbb{R}^m \cup \{+\infty\}$ s.t.

$$
\gamma_2(u) = \gamma(F + u \otimes e_{D+1}).
$$

Since both $\gamma_1$ and $\gamma_2$ and $n$-polyaffine at $0$ we can now use the inductive hypothesis on both $\gamma_1$ and $\gamma_2$ and obtain together with (4.20) that

$$
\gamma(F + u \otimes v) = \gamma_1(u \otimes \begin{pmatrix} v^D \\ 0 \end{pmatrix}) + v_{D+1} \left[ \gamma_2(u) - \gamma(F) \right]
$$
Example 4.23. Let \( \gamma = \det \) on \( \mathbb{R}^{d \times d} \) and \( F \in \mathbb{R}^{d \times d} \). Since \( \det \) is quasiaffine it is also 1-polyaffine at every \( F \in \mathbb{R}^{d \times d} \). Thus, Lemma 4.22 applies and we obtain

\[
(4.21)
\]

For any \( A \in \mathbb{R}^{d \times d} \) denote by \( M_{ij}(A) \) the \( ij \)-th minor of \( A \), i.e. the matrix that is obtained when deleting the \( i \)-th row and the \( j \)-th column. Then, by expanding the determinant of \( F + e_i \otimes e_j \) along the \( i \)-th row, we find

\[
\det(F + e_i \otimes e_j) = \sum_{k=1}^{d} (-1)^{i+k}[F + e_i \otimes e_j]_{ik} \det(M_{ik}(F + e_i \otimes e_j)).
\]

Furthermore, \( M_{ik}(F + e_i \otimes e_j) = M_{ik}(F) \) and \( [F + e_i \otimes e_j]_{ik} = F_{ik} + \delta_{jk} \). Hence,

\[
\det(F + e_i \otimes e_j) = \sum_{k=1}^{d} (-1)^{i+k}F_{ik} \det(M_{ik}(F)) + \sum_{k=1}^{d} (-1)^{i+k} \delta_{jk} \det(M_{ik}(F)) = \det(F) + (-1)^{i+j} \det M_{ij}(F) = \det(F) + [\text{cof}(F)]_{ij}.
\]

Inserting this into (4.21) we obtain

\[
\det(F + u \otimes v) = \det(F) + \sum_{i,j=1}^{d} [\text{cof}(F)]_{ij}u_iv_j = \det(F) + u^T \text{cof}(F)v,
\]

78
which is exactly the well known matrix determinant lemma, thus proving that Lemma 4.22 provides a generalisation for all 1-polyaffine functions at a point \( F \in \mathbb{R}^{d \times D} \).

The next step is to generalise the above formula even further to the case of \( n \)-polyaffine functions at \( F \in \mathbb{R}^{d \times D} \). In order to make the following as accessible as possible we introduce some shorthand notation.

**Definition 4.24.** Let \( S \) be a finite set with \( |S| = n \) and let \( k \) be an integer such that \( 0 \leq k \leq n \). We then denote by \( \binom{S}{k} \) the set of all \( k \)-combinations, i.e. the set of all subsets of \( S \) with \( k \) distinct elements. Furthermore it is the case that

\[
\left| \binom{S}{k} \right| = \binom{n}{k}.
\]

For \( k = 0 \) we mean \( \binom{S}{k} \) to be the empty set.

For brevity we also set

\[
S^n := \{1, \ldots, n\}.
\]

Then the aforementioned generalisation of Lemma 4.22 from 1- to \( n \)-polyaffine functions at \( F \) is the following lemma.

**Lemma 4.25.** Let \( F \in \mathbb{R}^{d \times D} \) and \( \gamma : \mathbb{R}^{d \times D} \to \mathbb{R}^m \) be affine on rank-one lines contained in \( C^n(F) \). Then

\[
\gamma \left( F + \sum_{x=1}^{n} u^x \otimes v^x \right) = \sum_{k=0}^{n} \sum_{I^{n,k} \in \binom{S^n}{k}} \sum_{i^x \in \{1, \ldots, d\}} \sum_{j^x \in \{1, \ldots, D\}} \sum_{x \in I^{n,k}} \left( \sum_{l=0}^{k} (-1)^{k-l} \sum_{J^{n,k}_l \in \binom{I^{n,k}}{l}} \gamma \left( F + \sum_{y \in J^{n,k}_l} e_{iy} \otimes e_{j^y} \right) \right) \prod_{x \in I^{n,k}_l} u^x_{i^x} v^x_{j^x} \right) \tag{4.22}
\]

for all \( u^1, \ldots, u^n \in \mathbb{R}^d \) and \( v^1, \ldots, v^n \in \mathbb{R}^D \).

To get a feel for this cumbersome notation we explicitly write down the form (4.22) takes for \( n = 2 \) and \( n = 3 \) for \( d = D = 3 \).
Example 4.26.  \( i \)  \( n = 2, d = D = 3: \)

\[
\gamma(F + u^1 \otimes v^1 + u^2 \otimes v^2) = \gamma(F) + \sum_{x=1}^{2} \sum_{i=1}^{d} \left[ \gamma(F + e_i \otimes e_j) - \gamma(F) \right] u^x_i v^x_j
\]

\[
+ \sum_{i=1, j=1}^{d} \left[ \gamma(F + e_i \otimes e_j + e_k \otimes e_l) - \gamma(F + e_i \otimes e_j) \right] v^3_k v^3_l
\]

\(- \gamma(F + e_k \otimes e_l) + \gamma(F) \] u^1_i v^1_j u^2_k v^2_l\]

\( ii \)  \( n = 3, d = D = 3: \)

\[
\gamma(F + u^1 \otimes v^1 + u^2 \otimes v^2 + u^3 \otimes v^3) = \gamma(F)
\]

\[
+ \sum_{x=1}^{3} \sum_{i=1, j=1}^{d} \left[ \gamma(F + e_i \otimes e_j) - \gamma(F) \right] u^x_i v^x_j
\]

\[
+ \sum_{x, y \in \{1, 2, 3\}, i=1, j=1}^{d} \sum_{x < y}^{d} \left[ \gamma(F + e_i \otimes e_j + e_k \otimes e_l) - \gamma(F + e_i \otimes e_j) \right]
\]

\[- \gamma(F + e_k \otimes e_l) + \gamma(F) \] u^3_k v^3_l u^3_m v^3_n\]

\[
+ \sum_{i=1, j=1}^{d} \left[ \gamma(F + e_i \otimes e_j + e_k \otimes e_l + e_o \otimes e_p) - \gamma(F + e_i \otimes e_j + e_k \otimes e_l) \right]
\]

\[- \gamma(F + e_i \otimes e_j + e_o \otimes e_p) - \gamma(F + e_k \otimes e_l + e_o \otimes e_p) + \gamma(F + e_i \otimes e_j)\]

\[
+ \gamma(F + e_k \otimes e_l) + \gamma(F + e_o \otimes e_p) - \gamma(F) \] u^1_i v^1_j u^2_k v^2_l u^3_o v^3_p\]

We see that the terms \( \gamma(F + e_i \otimes e_j) - \gamma(F) \) appear twice in the first case and three times in the second. Similarly \( \gamma(F + e_i \otimes e_j + e_k \otimes e_l) - \gamma(F + e_i \otimes e_j) \) - \( \gamma(F + e_k \otimes e_l) + \gamma(F) \) appears once for \( n = 2 \), but three times for \( n = 3 \). This is due to the fact that the sum \( \sum_{i=1}^{n} u^i \otimes v^i \) is symmetric with respect to the ordering of \( u^i \otimes v^i \), but this repetition carries no new information about the function \( \gamma \) other than this symmetry.

In order to further ease the proof of Lemma 4.25 we introduce more shorthand notations. Let \( \mathcal{A} = (A_1, \ldots, A_m) \in (\mathbb{R}^{d \times D})^m \) be an \( m \)-tuple and \( J \subseteq \{1, \ldots, m\} \) a set of indices.
Then we define
\[
\Lambda^F_J = \gamma \left( F + \sum_{x \in J} A_x \right)
\] (4.23)
and by convention \( \Lambda^F_\emptyset = \gamma(F) \). Further we define the set \( E = \{ e_i \otimes e_j : i = 1, \ldots, d, j = 1, \ldots, D \} \) as the set of canonical matrices in \( \mathbb{R}^{d \times D} \). For the \( k \)-tuple \( \mathcal{E} \in E^k \) with \( \mathcal{E}_y = e_{i_y} \otimes e_{j_y}, y = 1, \ldots, k \) and an index set \( I = \{ x_1, \ldots, x_k \} \subseteq S^n \) where \( x_i < x_j \) for \( i < j \) and matrices \( u^1 \otimes v^1, \ldots, u^n \otimes v^n \) we then define
\[
U_{\mathcal{E}}(I) = \prod_{m=1}^{k} u_{i_m}^{x_m} v_{j_m}^{x_m}.
\]
and by a similar convention \( U_{\emptyset}(\emptyset) = 1 \). For instance, with this new notation we can replace the sum \( \sum_{i \in \{1, \ldots, d\}} \sum_{j \in \{1, \ldots, D\}} \sum_{x \in I_{n,k}} \gamma\left( \sum_{y \in J_{n,k}^x} e_{i_y} \otimes e_{j_y} \right) \) by \( \sum_{J_{n,k}^x} \Lambda^F_{\mathcal{E}}(J_{n,k}^x) \). Therefore, in the new notation eq. (4.22) reads as
\[
\Lambda^F_\emptyset + \sum_{x=1}^{n} u^x \otimes v^x
\]
and by a similar convention \( U_{\emptyset}(\emptyset) = 1 \). For instance, with this new notation we can replace the sum \( \sum_{i \in \{1, \ldots, d\}} \sum_{j \in \{1, \ldots, D\}} \sum_{x \in I_{n,k}} \gamma\left( \sum_{y \in J_{n,k}^x} e_{i_y} \otimes e_{j_y} \right) \) by \( \sum_{\mathcal{E} \in E^k} \Lambda^F_{\mathcal{E}}(J_{n,k}^x) \). Therefore, in the new notation eq. (4.22) reads as
\[
\Lambda^F_\emptyset + \sum_{x=1}^{n} u^x \otimes v^x (\emptyset) = \sum_{k=0}^{n} \sum_{I_{n,k} \subseteq \{1, \ldots, d\}} \Lambda_{\mathcal{E}}(J_{n,k}^x) \sum_{\mathcal{E} \in E^k} \Lambda^F_{\mathcal{E}}(J_{n,k}^x) U_{\mathcal{E}}(I_{n,k}^x).
\] (4.24)

We will prove the validity of this equation by induction, the base case of which is already dealt with by Lemma 4.22.

**Proof of Lemma 4.25** Let \( F \in \mathbb{R}^{d \times D} \) and \( \gamma : \mathbb{R}^{d \times D} \to \mathbb{R}^m \) be \( n \)-polyaffine at \( F \). As indicated, we prove the validity of the shorthand equality (4.24) of the lemma by induction over \( n \). The base case \( n = 1 \) is already provided by Lemma 4.19. Thus, assume that (4.24) holds for \( n \) and consider the case \( n + 1 \), i.e. \( \gamma \) is affine on all rank-one lines that are
contained in the rank-\((n+1)\) cone \(C^{n+1}(F)\). We want to show that

\[
\Lambda_{\emptyset} \left( F + \sum_{x=1}^{n+1} u^x \otimes v^x \right) (\emptyset) = \sum_{k=0}^{n+1} \sum_{I_n,k \in \binom{S^{n+1}}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n+1,k} \in \binom{S^k}{l}} \Lambda_{E_{k}}^{F}(I_{l}^{n+1,k}) \right] U_{E_{n+1}}(I_{n+1,k}). \tag{4.25}
\]

Note that \(\gamma(F + \sum_{x=1}^{n+1} u^x \otimes v^x) = \gamma(F + \sum_{x=1}^{n} u^x \otimes v^x + u^{n+1} \otimes v^{n+1})\) and that therefore \(\gamma\) is affine on rank-one lines contained in \(C^n(F + u^{n+1} \otimes v^{n+1})\) for all \(u^{n+1} \otimes v^{n+1}\). Hence, we obtain with the inductive hypothesis that

\[
\Lambda_{\emptyset} \left( F + \sum_{x=1}^{n+1} u^x \otimes v^x \right) (\emptyset) = \sum_{k=0}^{n} \sum_{I_n,k \in \binom{S^n}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in \binom{S^k}{l}} \Lambda_{E_{k}}^{F+u^{n+1} \otimes v^{n+1}}(J_{l}^{n,k}) \right] U_{E_{n+1}}(I_{n,k}).
\]

We may now also use that \(\gamma\) is affine on rank-one lines contained in \(C^1(F + \sum_{i=1}^{n} u^i \otimes v^i)\) for all \(u^i \in \mathbb{R}^d, v^i \in \mathbb{R}^D\), so that for \(E_{k} \in E^k\) and \(J \subseteq \{1, \ldots, k\}\) we can rewrite

\[
\Lambda_{E_{k}}^{F+u^{n+1} \otimes v^{n+1}}(J) \text{ by using eq. (4.19) as}
\]

\[
\Lambda_{E_{k}}^{F+u^{n+1} \otimes v^{n+1}}(J) = \Lambda_{E_{k}}^{F}(J) + \sum_{E_{n+1} \in E} \left[ \Lambda_{E_{k},E_{n+1}}^{F}(J \cup \{n+1\}) - \Lambda_{E_{k},E_{n+1}}^{F}(J) \right] U_{E_{n+1}}(\{n+1\}).
\]
Inserting this into the above equation we find

\[
\Lambda_{\mathcal{F}} F^{n+1} \otimes v^x \mathbf{(\varnothing)} \\
= \sum_{k=0}^{n} \sum_{I_{n,k} \in \binom{S}{k}} \sum_{E_k \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in \binom{S}{k}} \Lambda_{\mathcal{E}_k} (I_{n,k}) \right] U_{E_{n+1}} \left( \{n+1\} \right) U_{E_{n+k}} (I_{n,k}) \\
+ \sum_{E_{n+1} \in E} \left[ \Lambda_{\mathcal{E}_{n+1}} (J_{l}^{n,k} \cup \{n+1\}) - \Lambda_{\mathcal{E}_k} (I_{n,k}) \right] \right] \\
\sum_{k=0}^{n} \sum_{I_{n,k} \in \binom{S}{k}} \sum_{E_k \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in \binom{S}{k}} \Lambda_{\mathcal{E}_k} (I_{n,k}) \right] U_{E_{n+1}} \left( \{n+1\} \right) U_{E_{n+k}} (I_{n,k}) \\
+ \sum_{k=0}^{n} \sum_{I_{n,k} \in \binom{S}{k}} \sum_{E_k \in E^k \times E} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in \binom{S}{k}} \Lambda_{\mathcal{E}_k} (I_{n,k}) \right] U_{E_{n+1}} \left( \{n+1\} \right).
\]

We want this to be equal to the expression (4.25). It might not have been clear at first why we have been using the superscripts \(I_{n+1,k}\) and \(J_{l}^{n+1,k}\), but the reader might sense now that this is merely to keep track of the terms that are possibly involved, whereby we are better able to distinguish it to the index sets \(I_{n,k}\) and \(J_{l}^{n,k}\). Note that \(S^{n+1} = \{1, \ldots, n+1\}\) and so the superscripts \(I_{n+1,k}\) and \(J_{l}^{n+1,k}\) indicate that both sets may now contain the element \(n+1\), in contrast to \(I_{n,k}\) and \(J_{l}^{n,k}\). Furthermore, the set of \(k\)-combinations \(\binom{S^{n+1}}{k}\) of \(S^{n+1}\) occurs. In order to connect the above expression to (4.26) we need to relate this set to \(\binom{S}{k}\). For this we introduce another notation. Let \(\mathcal{P}\) be a set of sets and \(A\) be another set. Then we denote by \(\mathcal{P}^{\cup A}\) the set of sets defined by

\[
\mathcal{P}^{\cup A} = \{ P \cup A : P \in \mathcal{P} \},
\]

i.e. the set of the union of each set element of \(\mathcal{P}\) with the set \(A\). With this notation it
can be seen that for $0 < k < n + 1$ we have
\[
\binom{S^{n+1}}{k} = \binom{S^n}{k} \cup \binom{S^n}{k-1}^{\cup \{n+1\}}.
\] (4.27)

Using the superscript notation also helps to clarify in which order the operations are undertaken just like the power operator $x^y$ is applied before the multiplication operator $x \cdot y$, i.e. there is the implicit parenthesis of $P \cup Q^{\cup A} := P \cup (Q^{\cup A})$ which is not the same as $(P \cup Q)^{\cup A}$. Leaving (4.26) as it is for now and instead returning our attention to (4.25) we can separate out the case for $k = 0$, which is simply $\Lambda F^0(\emptyset, \emptyset)$, and the case $k = n + 1$ in order to apply (4.27). Using (4.27) we can now rewrite (4.25) as
\[
\Lambda F^+ \sum_{x=1}^{n+1} u^x \otimes v^x (\emptyset) = \Lambda F^0(\emptyset)
\]
\[+ \sum_{k=1}^{n} \sum_{I_n,k \in \binom{S^n}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{n+1,k} \in \binom{S^k}{l}} \Lambda F^E(J_{n+1,k}) \right] U_E(I_{n+1,k})
\]
\[+ \sum_{E \in E^{n+1}} \left[ \sum_{l=0}^{n+1} (-1)^{n+1-l} \sum_{J_{n+1,n+1} \in \binom{S^{n+1}}{l}} \Lambda F^E(J_{n+1,n+1}) \right] U_E(S^{n+1})
\]
\[= \Lambda F^0(\emptyset)
\]
\[+ \sum_{k=1}^{n} \sum_{I_n,k \in \binom{S^n}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{n,k} \in \binom{S^k}{l}} \Lambda F^E(J_{n,k}) \right] U_E(I_{n,k})
\]
\[+ \sum_{k=1}^{n} \sum_{I_n,k \in \binom{S^n}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{n+1,k} \in \binom{S^k}{l}} \Lambda F^E(J_{n+1,k}) \right] U_E(I_{n,k+1} \cup \{n+1\})
\]
\[+ \sum_{E \in E^{n+1}} \left[ \sum_{l=0}^{n+1} (-1)^{n+1-l} \sum_{J_{n+1,n+1} \in \binom{S^{n+1}}{l}} \Lambda F^E(J_{n+1,n+1}) \right] U_E(S^{n+1})
\]
Noticing that in the last equality we can put the first line into the second by changing the
lower limit of the sum from 1 to 0 and the last line into the second last line by changing
the upper limit of the sum of the second last line from \(n\) to \(n+1\) we obtain

\[
\begin{align*}
F_{\emptyset} + \sum_{i=1}^{n+1} u^i \otimes v^i \\
\Lambda_{\emptyset} = \\
&+ \sum_{k=0}^{n} \sum_{I^{n,k} \in \binom{S^n}{k}} \sum_{\mathcal{E} \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^{n,k} \in \binom{S^k}{l}} \Lambda^F_{\mathcal{E}}(J_l^{n,k}) \right] U_{\mathcal{E}}(I^{n,k}) \\
&+ \sum_{k=1}^{n+1} \sum_{I^{n,k-1} \cup \{n+1\} \in \binom{S^{n+1}}{k}} \sum_{\mathcal{E} \in E^{k+1}} \left[ \sum_{l=0}^{k+1} (-1)^{k+1-l} \sum_{J_l^{n+1,k+1} \in \binom{S^{k+1}}{l}} \Lambda^F_{\mathcal{E}}(J_l^{n+1,k+1}) \right] U_{\mathcal{E}}(I^{n,k-1} \cup \{n+1\})
\end{align*}
\]

Now we can perform a shift of the index \(k\) in the second summation to yield

\[
\begin{align*}
F_{\emptyset} + \sum_{i=1}^{n+1} u^i \otimes v^i \\
\Lambda_{\emptyset} = \\
&+ \sum_{k=0}^{n} \sum_{I^{n,k} \in \binom{S^n}{k}} \sum_{\mathcal{E} \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^{n,k} \in \binom{S^k}{l}} \Lambda^F_{\mathcal{E}}(J_l^{n,k}) \right] U_{\mathcal{E}}(I^{n,k}) \\
&+ \sum_{k=0}^{n+1} \sum_{I^{n,k-1} \cup \{n+1\} \in \binom{S^{n+1}}{k}} \sum_{\mathcal{E} \in E^{k+1}} \left[ \sum_{l=0}^{k+1} (-1)^{k+1-l} \sum_{J_l^{n+1,k+1} \in \binom{S^{k+1}}{l}} \Lambda^F_{\mathcal{E}}(J_l^{n+1,k+1}) \right] U_{\mathcal{E}}(I^{n,k-1} \cup \{n+1\})
\end{align*}
\]

(4.28)

If we compare this to (4.26) we see that we now merely need to equate the last two lines
of (4.26) and the last two lines of the above equation. Let \(lhs\) be equal to the last two
lines of (4.26) and \(rhs\) equal to the last two lines of (4.28), i.e.

\[
\begin{align*}
lhs = \sum_{k=0}^{n} \sum_{\mathcal{E} \in \binom{S^n}{k}} \sum_{I^{n,k} \in \binom{S^n}{k}} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^{n,k} \in \binom{S^k}{l}} \Lambda^F_{(\mathcal{E}_{k,E_{n+1}})}(J_l^{n,k} \cup \{k+1\}) \right] U(\mathcal{E}_{k,E_{n+1}})(I^{n,k} \cup \{n+1\})
\end{align*}
\]

85
and

\[ \text{rhs} = \sum_{k=0}^{n} \sum_{I^k \cup \{n+1\} \in (S_{n}^k \cup \{n+1\})} \left[ \sum_{l=0}^{k+1} (-1)^{k+1-l} \sum_{J^{n+1,k+1}_l \in (S_{l}^{k+1})} \Lambda^F_E(J^{n+1,k+1}_l) \right] U_E(I^k \cup \{n+1\}). \]

By dropping \( \cup \{n+1\} \) in rhs (which changes nothing about the elements that are summed over) and changing \( (E, E_{n+1}) \in E^k \times E \) to \( E \in E^{k+1} \) in lhs we see that we now only need to equate both square brackets of lhs and rhs, which we denote by lhs[\( \cdot \)] and rhs[\( \cdot \)] respectively. In rhs[\( \cdot \)] we sum over the elements in \( (S_{l}^{k+1}) \), for which we have for \( 0 < l < k + 1 \)

\[ \left( S_{l}^{k+1} \right) = \left( S_{l}^{k} \right) \cup \left( S_{l-1}^{k} \right). \]

Thus separating the case \( l = 0 \) and \( l = k + 1 \) we find

\[ \text{rhs}[\cdot] = (-1)^{k+1} \Lambda^F(\emptyset) + \sum_{l=1}^{k} (-1)^{k+1-l} \sum_{J^{n+1,k+1}_l \in (S_{l}^{k+1})} \Lambda^F_E(J^{n+1,k+1}_l) + \Lambda^F_E(S^{k+1}) \]

\[ = (-1)^{k+1} \Lambda^F(\emptyset) + \sum_{l=1}^{k} (-1)^{k+1-l} \sum_{J^{n,k}_l \in (S_{l}^{k})} \Lambda^F_E(J^{n,k}_l) \]

\[ + \sum_{l=1}^{k} (-1)^{k+1-l} \sum_{J^{n,k}_{l-1} \cup \{k+1\} \in (S_{l-1})} \Lambda^F_E(J^{n,k}_{l-1} \cup \{k+1\}) + \Lambda^F_E(S^{k+1}) \]

and again by putting the first two and the last two terms together, dropping \( \{k+1\} \) in the last sum and performing an index shift it is

\[ \text{rhs}[\cdot] = -\sum_{l=0}^{k} (-1)^{k-l} \sum_{J^{n,k}_l \in (S_{l}^{k})} \Lambda^F_E(J^{n,k}_l) + \sum_{l=1}^{k+1} (-1)^{k+1-l} \sum_{J^{n,k}_{l-1} \cup \{k+1\} \in (S_{l-1})} \Lambda^F_E(J^{n,k}_{l-1} \cup \{k+1\}) \]

\[ = -\sum_{l=0}^{k} (-1)^{k-l} \sum_{J^{n,k}_l \in (S_{l}^{k})} \Lambda^F_E(J^{n,k}_l) + \sum_{l=0}^{k} (-1)^{k-l} \sum_{J^{n,k}_l \in (S_{l}^{k})} \Lambda^F_E(J^{n,k}_l \cup \{k+1\}) \]

86
\[
= \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in (S^{k})} [\Lambda_{E}^{F}(J_{l}^{n,k} \cup \{k + 1\}) - \Lambda_{E}^{F}(J_{l}^{n,k})],
\]

which is equal to \( \text{lhs}[\cdot] \).

Notice that equation (4.24) for \( \Lambda_{E}^{F} \) holds for all maps \( \gamma : \mathbb{R}^{d \times D} \to \mathbb{R}^{m} \) that are \( n \)-polyaffine at \( F \in \mathbb{R}^{d \times D} \). Thus, it holds in particular for the minors map \( T \), which is polyaffine. In view of proving the existence of \( \beta \in \mathbb{R}^{\tau(d,D)} \) for any \( h \) polyaffine at \( F \) such that

\[
h(\tilde{F}) - h(F) = \langle \beta, T(\tilde{F}) - T(F) \rangle
\]

for all \( \tilde{F} \in C^{n}(F) \), we see that it is sufficient to apply formula (4.24) to both \( h \) and \( T \) with \( \tilde{F} = F + \sum_{x=1}^{n} u_{x}^{r} \otimes v_{x}^{r} \), i.e. to

\[
h_{\otimes}^{F}(\mathcal{O}) - h_{\otimes}^{F}(\mathcal{O}) = \langle \beta, T_{\otimes}^{F}(\mathcal{O}) - T_{\otimes}^{F}(\mathcal{O}) \rangle
\]

where \( h_{\otimes}^{F} \) and \( T_{\otimes}^{F} \) are defined like \( \Lambda_{E}^{F} \) for \( \gamma = h \) and \( \gamma = T \) respectively.

Step 2. We can then infer the following:

**Lemma 4.27.** Let \( h \) be \( n \)-polyaffine at \( F \). Then there exists \( \beta \in \mathbb{R}^{\tau(d,D)} \) such that

\[
h(\tilde{F}) - h(F) = \langle \beta, T(\tilde{F}) - T(F) \rangle
\]

for all \( \tilde{F} \in C^{n}(F) \) if and only if there exists \( \beta \in \mathbb{R}^{\tau(d,D)} \) that solves the system of linear equations

\[
\langle \beta, \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in (S^{k})} T_{E}^{F}(J_{l}^{n,k}) \rangle = \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_{l}^{n,k} \in (S^{k})} h_{E}^{F}(J_{l}^{n,k}),
\]

for all \( k = 1, \ldots, n \) and \( E \in E^{k} \).

The proof of this lemma consists of simply expanding equation (4.29) and recognising that all coefficients of the factors \( U(T^{n,k}) \) must coincide, which is precisely the given system of linear equations. For example, consider the case \( h : \mathbb{R}^{2 \times 2} \to \mathbb{R} \cup \{+\infty\} \).
1-polyaffine at \( F \in \mathbb{R}^{2 \times 2} \) and \( T(\tilde{F}) = [\tilde{F}, \det(\tilde{F})] \). Then the expansion of the equation

\[
\langle \beta, T(F + u \otimes v) - T(F) \rangle = h(F + u \otimes v) - h(F),
\]

which has to hold for all \( u, v \in \mathbb{R}^2 \), becomes

\[
\langle \beta, T(F) + \sum_{i,j=1}^{2} [T(F + e_i \otimes e_j) - T(F)] u_i v_j - T(F) \rangle = h(F) + \sum_{i,j=1}^{2} [h(F + e_i \otimes e_j) - h(F)] u_i v_j - h(F)
\]

which simplifies to

\[
\langle \beta, \sum_{i,j=1}^{2} [T(F + e_i \otimes e_j) - T(F)] u_i v_j \rangle = \sum_{i,j=1}^{2} [h(F + e_i \otimes e_j) - h(F)] u_i v_j.
\]

Therefore we obtain the system

\[
\langle \beta, T(F + e_i \otimes e_j) - T(F) \rangle = h(F + e_i \otimes e_j) - h(F),
\]

\( i, j \in \{1, 2\} \). This concludes Step 2 in proving ‘(ii) \( \Rightarrow \) (iii)’ of Theorem 4.20.

For Step 3 we first continue to discuss the given example. When we identify \( \beta = [\beta^1 | \beta^2] \) with \( \beta^1 \in \mathbb{R}^{2 \times 2}, \beta^2 \in \mathbb{R} \) with the row vector \( \begin{bmatrix} \beta_{11}^1 & \beta_{12}^1 & \beta_{21}^1 & \beta_{22}^1 & | & \beta^2 \end{bmatrix} \) we find the above equivalent to

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & | & F_{22} \\
0 & 1 & 0 & 0 & | & -F_{21} \\
0 & 0 & 1 & 0 & | & -F_{12} \\
0 & 0 & 0 & 1 & | & F_{11}
\end{bmatrix} \begin{bmatrix} \beta_{11}^1 & \beta_{12}^1 & \beta_{21}^1 & \beta_{22}^1 & | & \beta^2 \end{bmatrix} \begin{bmatrix} h(F + e_1 \otimes e_1) - h(F) \\
h(F + e_1 \otimes e_2) - h(F) \\
h(F + e_2 \otimes e_1) - h(F) \\
h(F + e_2 \otimes e_2) - h(F)
\end{bmatrix},
\]

which, since the matrix on the left is given in row echelon form and has full rank, obviously has a solution (in fact, infinitely many). Note that had we considered \( h \) polyaffine instead of just 1-polyaffine we would have to satisfy more equations, namely

\[
\langle \beta, T(F + e_i \otimes e_j + e_k \otimes e_l) - T(F + e_i \otimes e_j) - T(F + e_k \otimes e_l) + T(F) \rangle
\]

\[
= h(F + e_i \otimes e_j + e_k \otimes e_l) - h(F + e_i \otimes e_j) - h(F + e_k \otimes e_l) + h(F)
\]

88
for \( i, j, k, l \in \{1, 2\} \). Although this seemingly adds eight equations for only one remaining free parameter \( \beta^2 \) we can show that they are all equivalent to just the one equation when \( i = j = 1 \) and \( k = l = 2 \) (we will only do this in the general case). Thus it must additionally hold that

\[
\begin{bmatrix}
0 & 0 & 0 & 0 \\
\end{bmatrix}
\beta^T = h(F + e_i \otimes e_j + e_k \otimes e_l) - h(F + e_i \otimes e_j)
- h(F + e_k \otimes e_l) + h(F),
\]

which still gives a system of full rank that therefore admits a solution. In this latter case the solution is also unique. The existence of a solution to this system in general is the content of the next lemma.

**Lemma 4.28.** The system of linear equations given in Lemma 4.27 has at least one solution. The solution space of the homogeneous system has dimension \( \frac{d^i D^i}{i = n + 1} \).

**Proof.** Let us consider \( \gamma \)-polyaffine at \( F \) and let \( A = \{A_1, \ldots, A_n\} \subseteq C^1(0) \), i.e. \( \text{rank}(A_i) \leq 1, i = 1, \ldots, n \). Recall that for \( J \subseteq S^n \) we define

\[
\Lambda^F_A(J) = \gamma \left( F + \sum_{j \in J} A_j \right).
\]

Note that in the previous working we only considered \( A_i = E_i \) with \( E_i \in E \subseteq C^1(0) \), however, the following does not necessarily need the restriction to the set \( E = \{e_i \otimes e_j : i \in \{1, \ldots, d\}, j \in \{1, \ldots, D\}\} \). Instead the next results will hold for the more general case of \( A_i \in C^1(0) \). From \( \Lambda^F_A \) we will now define a new function \( \Lambda^{F,k}_A \) that has a recursive property. From this and the recognition that \( \gamma \) is affine on rank-one lines contained in \( C^n(F) \) we can deduce a couple of helpful properties for \( \Lambda^{F,k}_A \). For \( J \subseteq S^n \) we define \( \Lambda^{F,k}_A \) as follows:

\[
\Lambda^{F,k}_A(J) = \Lambda^{F,k-1}_A(J) - \sum_{\substack{J^k \in (S^n)^k \\ \{J^k\} \subseteq J}} \Lambda^{F,k-1}_A(J^k), \quad 1 \leq k \leq |J|
\]

with

\[
\Lambda^{F,0}_A(J) = \Lambda^F_A(J) - \Lambda^F_A(\emptyset).
\]

The properties that we want to show are:
(i) \( \Lambda_{A}^{F|J|}(J) = 0 \) \hspace{1cm} (4.32)

(ii) If \( A_1 \) and \( A_2 \) are rank-one connected, then for any \( J \) with \( 1, 2 \notin J \) and \( k \leq |J| + 2 \) we have
\[
\Lambda_{A}^{F,k}({\{1, 2\} \cup J}) = \Lambda_{A}^{F,k}({\{1\} \cup J}) + \Lambda_{A}^{F,k}({\{2\} \cup J}) - \Lambda_{A}^{F,k}(J), \tag{4.33}
\]

(iii) If \( A_1 \) and \( A_2 \), \( A_1 \) and \( A_3 \), \( A_2 \) and \( A_3 \) and \( A_2 \) and \( A_4 \) are rank-one connected and \( 1, 2, 3, 4 \notin J \) then
\[
\Lambda_{A}^{F,|J|+1}({\{1, 4\} \cup J}) + \Lambda_{A}^{F,|J|+1}({\{2, 3\} \cup J}) = 0. \tag{4.34}
\]

With these results we will be able to show that many of the equations in the linear system (4.30) are either zero or linearly dependent. For instance, property (i) and (ii) together imply that also \( \Lambda_{A}^{F,|J|+1}({\{1, 2\} \cup J}) = 0 \), meaning that any equations will vanish if any two of the participating rank-one tensors have a rank-one connection. This is the case when two of the basis tensors \( e_i \otimes e_j \) and \( e_k \otimes e_l \) have either \( i = k \) or \( j = l \) and thus, these cases can be disregarded. The remaining cases therefore satisfy \( e_i \otimes e_{j_1}, \ldots, e_i \otimes e_{j_k} \) with \( i^x \neq i^y \) and \( j^x \neq j^y \) for all \( x \neq y \).

**Part (ii).** For now we will focus on proving (iii). (Note that (i) is trivial.) The proof follows in two steps. The first is to prove the similar formula
\[
\Lambda_{A}^{F,k}({\{1, 2\} \cup J}) = \Lambda_{A}^{F,k}({\{1\} \cup J}) + \Lambda_{A}^{F,k}({\{2\} \cup J}) - \Lambda_{A}^{F,k}(J) \\
- \sum_{l=1}^{k-1} \sum_{J^{l-1} \in \binom{J}{l-1}} \Lambda_{A}^{F,l}({\{1, 2\} \cup J^{l-1}}) \hspace{1cm} (4.35)
\]
and then the final result (4.33). We prove both by induction. To prove the base case \( k = 0 \) of (4.35) we have to show that \( \Lambda_{A}^{F,0}({\{1, 2\} \cup J}) = \Lambda_{A}^{F,0}({\{1\} \cup J}) + \Lambda_{A}^{F,0}({\{2\} \cup J}) - \Lambda_{A}^{F,0}(J) \). By denoting \( \tilde{F} = F + \sum_{j \in J} A_j \) it is then equivalent to show
\[
\gamma(\tilde{F} + A_1 + A_2) = \gamma(\tilde{F} + A_1) + \gamma(\tilde{F} + A_2) - \gamma(\tilde{F}).
\]
Knowing that \( \gamma \) is affine on the line through \( \tilde{F} + A_1 \) and \( \tilde{F} + A_2 \) (since \( A_1 \) and \( A_2 \) are
rank-one connected by assumption) allows us to observe that
\[
\gamma(\tilde{F} + A_1 + A_2) = \frac{1}{2} \gamma(\tilde{F} + 2A_1) + \frac{1}{2} \gamma(\tilde{F} + 2A_2).
\]
Then using linearity of \(\gamma\) along the rank-one lines through \(\tilde{F}\) and \(F + A_i, i = 1, 2\) we get
\[
\gamma(\tilde{F} + A_1 + A_2) = \frac{1}{2} \left[ \gamma(\tilde{F}) + 2\left(\gamma(F + A_1) - \gamma(\tilde{F})\right) + \gamma(\tilde{F}) + 2\left(\gamma(F + A_2) - \gamma(\tilde{F})\right) \right]
\]
\[
= \gamma(\tilde{F} + A_1) + \gamma(\tilde{F} + A_2) - \gamma(\tilde{F}).
\]
Thus the base case \(k = 0\) holds. Now assume (4.35) holds for \(k\). Note that since \(1, 2 \notin J\) it is
\[
\left(\{1, 2\} \cup J_{k+1}\right) = \left(\{1, 2\} \cup J_{k-1}\right) \cup \left(\{1\} \cup J_{k-1}\right) \cup \left(\{2\} \cup J_{k-1}\right) \cup \left(\{1\} \cup J_k\right) \cup \left(\{2\} \cup J_k\right) \cup \left(\{1\} \cup J_{k+1}\right) \cup \left(\{2\} \cup J_{k+1}\right).
\]
(4.36)
Then by the definition of \(\Lambda_{F,k}\) and (4.36) we obtain
\[
\Lambda_{F,k+1}(\{1, 2\} \cup J) = \Lambda_{F,k}(\{1, 2\} \cup J) - \sum_{S \in \binom{\{1, 2\} \cup J}{k+1}} \Lambda^F_{A}(S)
\]
\[
= \Lambda_{F,k}(\{1, 2\} \cup J) - \sum_{J^{k-1} \in \binom{J}{k-1}} \Lambda^F_{A}(\{1, 2\} \cup J^{k-1}) - \sum_{J^k \in \binom{\{1\}}{k}} \Lambda^F_{A}(\{1\} \cup J^k)
\]
\[
- \sum_{J^k \in \binom{\{2\}}{k}} \Lambda^F_{A}(\{2\} \cup J^k) - \sum_{J^{k+1} \in \binom{\{1\}}{k+1}} \Lambda^F_{A}(\{1\} \cup J^{k+1})
\]
By using the inductive hypothesis we find
\[
\Lambda_{F,k+1}(\{1, 2\} \cup J) = \Lambda_{F,k}(\{1\} \cup J) + \Lambda_{F,k}(\{2\} \cup J) - \Lambda_{F,k}(J)
\]
\[
- \sum_{l=1}^{k-1} \sum_{J^{l-1} \in \binom{\{1\}}{l-1}} \Lambda^F_{A}(\{1, 2\} \cup J^{l-1}) - \sum_{J^{k-1} \in \binom{\{1\}}{k-1}} \Lambda^F_{A}(\{1, 2\} \cup J^{k-1})
\]
\[
- \sum_{J^k \in \binom{\{1\}}{k}} \Lambda^F_{A}(\{1\} \cup J^k) - \sum_{J^k \in \binom{\{2\}}{k}} \Lambda^F_{A}(\{2\} \cup J^k) - \sum_{J^{k+1} \in \binom{\{1\}}{k+1}} \Lambda^F_{A}(J^{k+1}).
\]
By rearranging the order of those terms and applying the definition for $\Lambda_{A,F,k+1}^{J}$ we deduce

$$\Lambda_{A}^{F,k+1}(\{1,2\} \cup J) = \Lambda_{A}^{F,k}(\{1\} \cup J) - \sum_{J^{k} \in (J)} \Lambda_{A}^{F,k}(\{1\} \cup J^{k}) - \sum_{J^{k+1} \in (J)} \Lambda_{A}^{F,k}(J^{k+1})$$

$$+ \Lambda_{A}^{F,k}(\{2\} \cup J) - \sum_{J^{k} \in (J)} \Lambda_{A}^{F,k}(\{2\} \cup J^{k}) - \sum_{J^{k+1} \in (J)} \Lambda_{A}^{F,k}(J^{k+1})$$

$$- \left( \Lambda_{A}^{F,k}(J) - \sum_{J^{k+1} \in (J)} \Lambda_{A}^{F,k}(J^{k+1}) \right)$$

$$- \sum_{l=1}^{k-1} \sum_{J^{l-1} \in (J)} \Lambda_{A}^{F,l}(\{1,2\} \cup J^{l-1}) - \sum_{J^{k-1} \in (J)} \Lambda_{A}^{F,k}(\{1,2\} \cup J^{k-1})$$

where each term of the last equality corresponds to each row of the previous equality.

Now we want to show that it is possible to completely remove the last term of eq. (4.35) to obtain (4.33). We do so by showing by induction over $|J|$ that $\Lambda_{A}^{F,J+1}(\{1,2\} \cup J) = 0$. In the base case we have $|J| = 0$, i.e. $J = \emptyset$. Then by definition of $\Lambda_{A}^{F,k}$ we have

$$\Lambda_{A}^{F,1}(\{1,2\}) = \Lambda_{A}^{F,0}(\{1\}) - \Lambda_{A}^{F,0}(\{2\}) = \gamma(F + A_{1} + A_{2}) - \gamma(F + A_{1}) - \gamma(F + A_{2}) + \gamma(F)$$

which, as we have seen in the base case of the previous induction, must be zero. Now consider the case $|J| + 1$, i.e. we have a set $\tilde{J}$ with $|\tilde{J}| = |J| + 1$ and assume the hypothesis holds for all sets $\tilde{L}$ with $|\tilde{L}| \leq |J|$. Then it is by eq. (4.35)

$$\Lambda_{A}^{F,J+2}(\{1,2\} \cup \tilde{J}) = \Lambda_{A}^{F,J+2}(\{1\} \cup \tilde{J}) + \Lambda_{A}^{F,J+2}(\{2\} \cup \tilde{J}) - \Lambda_{A}^{F,J+2}(\{1,2\} \cup \tilde{J})$$

By the inductive hypothesis the last terms vanish and the first three terms are all equal to zero by property (4.34) and thus, $\Lambda_{A}^{F,J+2}(\{1,2\} \cup \tilde{J}) = 0$. Together with (4.35) this proves (4.33).

We now continue to prove property (4.34). For this let $A_{1}$ and $A_{2}$, $A_{2}$ and $A_{3}$, $A_{1}$ and $A_{3}$ and $A_{4}$ be rank-one connected. Further let $J$ be an index set with $|J| + 4 \leq n$ and $1, 2, 3, 4 \notin J$. We want to show $\Lambda_{A}^{F,J+1}(\{1,4\} \cup J) + \Lambda_{A}^{F,J+1}(\{2,3\} \cup J) = 0$. Observe
that by using that $\gamma$ is affine on rank-one connections between $A_1$ and $A_2$, $A_1$ and $A_3$, and $A_3$ and $A_4$ and \[4.33\]

$$\Lambda^F_{A}^{\gamma}(\{1, 3, 4\} \cup J) = \Lambda^F_{A}^{\gamma}(\{1, 3, 4\} \cup J) + \Lambda^F_{A}^{\gamma}(\{2, 3, 4\} \cup J) - \Lambda^F_{A}^{\gamma}(\{3, 4\} \cup J)$$

$$= \Lambda^F_{A}^{\gamma}(\{3, 4\} \cup J) + \Lambda^F_{A}^{\gamma}(\{1, 4\} \cup J) - \Lambda^F_{A}^{\gamma}(\{4\} \cup J)$$

$$+ \Lambda^F_{A}^{\gamma}(\{2, 3\} \cup J) + \Lambda^F_{A}^{\gamma}(\{3, 4\} \cup J) - \Lambda^F_{A}^{\gamma}((3) \cup J) = 0,$$

since all terms in the last equality must be zero by (i). Additionally we have that

$$\left(\begin{array}{c} J \\ |J| + 2 \end{array}\right) = \bigcup_{I^3 \in (S^4_3)} \left\{ J \cup I^2 \cup \bigcup_{I^3 \in (S^4_3)} \left(\begin{array}{c} J \\ |J| - 1 \end{array}\right) \cup \left(\begin{array}{c} J \\ |J| - 2 \end{array}\right) \right\} \cup \{1, 2, 3, 4\}.$$

Then by using the definition of $\Lambda^F_{A}^{\gamma}$ and the above equality we obtain

$$0 = \Lambda^F_{A}^{\gamma}(S^4 \cup J) = \Lambda^F_{A}^{\gamma}(S^4 \cup J) - \sum_{J^{|J|+2} \in (S^4_2)} \Lambda^F_{A}^{\gamma}(J^{|J|+2})$$

$$= \Lambda^F_{A}^{\gamma}(S^4 \cup J) - \sum_{I^2 \in (S^4_2)} \Lambda^F_{A}^{\gamma}(I^2 \cup J)$$

$$- \sum_{J^{|J|-1} \in (|J|-1)} \sum_{I^3 \in (S^4_3)} \Lambda^F_{A}^{\gamma}(I^3 \cup J^{|J|-1}) - \sum_{J^{|J|-2} \in (|J|-2)} \Lambda^F_{A}^{\gamma}(S^4 \cup J^{|J|-2}).$$

We now notice that all negative terms have an index set of $|J| + 2$ elements. Thus, by property \[4.33\] all those negative terms are zero that contain at least one pair of indices that corresponds to a pair of rank-one connected matrices. This is the case for all sets $I^3 \in (S^4_3)$ and $S^4$ itself. For the sets $I^2 \in (S^4_2) = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\}$ all but the sets $\{1, 4\}$ and $\{2, 3\}$ have a rank-one connection by assumption. Therefore we conclude

$$0 = \Lambda^F_{A}^{\gamma}(S^4 \cup J) - \Lambda^F_{A}^{\gamma}(\{1, 4\} \cup J) - \Lambda^F_{A}^{\gamma}(\{2, 3\} \cup J).$$

To the term $\Lambda^F_{A}^{\gamma}(S^4 \cup J)$ the same reasoning as above for $\Lambda^F_{A}^{\gamma}(\{1, 4\} \cup J)$ applies and it vanishes, leaving us with

$$\Lambda^F_{A}^{\gamma}(\{1, 4\} \cup J) + \Lambda^F_{A}^{\gamma}(\{2, 3\} \cup J) = 0.$$
which is exactly property [iii].

We notice that all of the above properties are true for $\Lambda^{F,k}_{A}$. Furthermore, we claim that the system of equations given in (4.30) can be rewritten as

$$\langle \beta, T^{F,k-1}_E(\mathcal{I}^{n,k}) \rangle = h^{F,k-1}_E(\mathcal{I}^{n,k}),$$

for all $\mathcal{I}^{n,k} \in \binom{S_n}{k}$ and $E \in E^k$, $k = 1, \ldots, n$. To see this consider again in general $\Lambda^{F}_{A}$ and w.l.o.g. $\mathcal{I}^{n,k} = S^k$. We want to prove that

$$\sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F}_{A}(J_l) = \sum_{l=1}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F,0}_{A}(J_l)$$

and for each $i \leq k-1$ that

$$\sum_{l=i+1}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F,i}_{A}(J_l) = \sum_{l=i}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F,i-1}_{A}(J_l).$$

Then by the above chain of equalities we can conclude

$$\sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F}_{A}(J_l) = \sum_{J_k \in \binom{S^k}{k}} \Lambda^{F,k-1}_{A}(J_k) = \Lambda^{F,k-1}_{A}(S^k).$$

By definition of $\Lambda^{F,i}_{A}$ we find

$$\sum_{l=i+1}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F,i}_{A}(J_l) = \sum_{l=i}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \left[ \Lambda^{F,i-1}_{A}(J_l) - \sum_{j_l \in \binom{S^k}{l}} \Lambda^{F,i-1}_{A}(J_l^i) \right].$$

Adding the zero $(-1)^{k-i} \sum_{J_i \in \binom{S^k}{i}} \Lambda^{F,i-1}_{A}(J_i) - (-1)^{k-i} \sum_{J_i \in \binom{S^k}{i}} \Lambda^{F,i-1}_{A}(J_i^i)$ and working both terms into each of the above we obtain

$$C^k_i(S^k) = \sum_{l=i}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \Lambda^{F,i-1}_{A}(J_l) - \sum_{l=i}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \sum_{J_l^i \in \binom{S^k}{l}} \Lambda^{F,i-1}_{A}(J_l^i)$$

$$= C^k_{i-1}(S^k) - \sum_{l=i}^{k} (-1)^{k-l} \sum_{J_l \in \binom{S^k}{l}} \sum_{J_l^i \in \binom{S^k}{l}} \Lambda^{F,i-1}_{A}(J_l^i).$$

(4.39)
Therefore, it remains to show that the subtracted terms equate to zero. For this note that in the double summation \( \sum_{J_l \in (S^k)} \sum_{J_l i \in (J_l)} \) we always choose a set \( J_l i \) of \( i \) elements. Thus it is appropriate to ask how many times an arbitrary set \( J_i \in (S^k) \) would occur in this double summation. Hence for fixed \( J_i \in (S^k) \) it is necessary that \( J_l \subseteq J_i \in (S^k) \) and there are \( l - i \) more elements in \( J_l \) from the set \( S^k \setminus J_i \). Altogether the set \( J_i \) is included in \( \binom{k}{l-i} \) sets from \( (S^k) \) and therefore

\[
\sum_{J_l \in (S^k)} \sum_{J_l i \in (J_l)} \Lambda^{F,i-1}_A(J_l) = \binom{k-i}{l-i} \sum_{J_l i \in (S^k)} \Lambda^{F,i-1}_A(J_i).
\]

Plugging this into (4.39) we obtain

\[
C^k_i(S^k) = C^k_{i-1}(S^k) - \sum_{l=i}^k (-1)^{k-l} \binom{k-i}{l-i} \sum_{J_l i \in (S^k)} \Lambda^{F,i-1}_A(J_l).
\]

It is a simple task to check that

\[
\sum_{l=i}^k (-1)^{k-l} \binom{k-i}{l-i} = 0
\]

and thus,

\[
C^k_i(S^k) = C^k_{i-1}(S^k).
\]

Let us now once more consider the system of equations (4.30) that is to be solved. This system includes a vast number of equations. More precisely, for each \( k = 1, \ldots, n \) we can choose the matrices \( E_1, \ldots, E_k \in \{ e_i \otimes e_j : i = 1, \ldots, d, j = 1, \ldots, D \} = E \) for which the maps \( T^{F,k}_E \) and \( h^{F,k}_E \) will be defined for \( E = (E_1, \ldots, E_k) \). Note that the order of the matrices does not matter, however we may choose a matrix multiple times. Therefore we initially have \( (dD)^k \) choices. Surprisingly most of these choices amount to the respective equation being the trivial equality \( 0 = 0 \). This is the case whenever there are \textit{any two} matrices \( E_l \) and \( E_m \) that are rank-one connected, due to property [iii] of \( \Lambda^{F,k}_A \) for both \( T^{F,k}_E \) and \( h^{F,k}_E \). Therefore we only need to consider the choices \( E_l = e_i \otimes e_j \) with \( i \neq m \) and \( j \neq j^m \) for all \( l \neq m \), i.e. when we consider the matrix \( A = \sum_{l=1}^k E_l \) the matrix \( A \) consists solely of zeros and exactly \( k \) ones, where each of the ones is contained in a different row and column. Furthermore, selecting any \( k \) rows and \( k \) columns and considering the options of placing exactly \( k \) ones into it in the above manner will lead
to \(k\) choices. However, all of those \(k\) cases can be reduced to just one since every such choice can be referred back to the case where the first one is placed on the first selected row and column element, the second one on the second row and column element and so forth. So, for instance, in the case of the first \(k\) selected rows and columns the choices for the matrix \(A\) can be

\[
\begin{pmatrix}
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
1 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 \\
\end{pmatrix}
\]

and

\[
\begin{pmatrix}
0 & 1 & 0 & \ldots & 0 \\
1 & 0 & 0 & \ldots & 0 \\
1 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 \\
\end{pmatrix}
\].

Both cases are related by property (iii) and therefore simply changes the sign on both sides of the equation (4.37) or (4.30) respectively. Hence, we only need to consider one case for each possible choice of \(k\) selected rows and columns, of which there are \(\binom{D}{k} \binom{D}{k}\) many. This already reminds us of the dimension of the \(k\)-th adjugate \(\text{adj}_k\). We now want to argue that all remaining equations from \(T_{\mathcal{E}}^{F,k-1}(I^{n,k})\) are automatically in reduced row echelon form. Since \(T(\cdot) = [\text{adj}_1(\cdot), \ldots, \text{adj}_{d\wedge D}(\cdot)]\) we can also write \(T_{\mathcal{E}}^{F,k}(\cdot) = [\text{adj}_{1\mathcal{E}}^{F,k}(\cdot), \ldots, \text{adj}_{d\wedge D\mathcal{E}}^{F,k}(\cdot)]\), since the definition of \(T_{\mathcal{E}}^{F,k}\) is acting componentwise on the entries of \(T\). Then consider \(\text{adj}_{s\mathcal{E}}^{F,k-1}(I^{n,k})\) for any choice of matrices \(E_i\) as identified as relevant in the above, i.e. consisting of \(k\) ones each in a different row and column. Because each element of \(\text{adj}_{s\mathcal{E}}^{F,k-1}(I^{n,k})\), i.e. \(\text{adj}_{s\mathcal{E}}^{F,k-1}(\cdot)\) \(\alpha\beta\) for \(\alpha \in \{1, \ldots, \binom{D}{s}\}\) and \(\beta \in \{1, \ldots, \binom{D}{s}\}\), is only working on a \(s \times s\) submatrix we notice that it can effectively only contain at most \(s\) rank-one matrices \(E_i\) that are nonzero on the particular \(s \times s\) submatrix. To be more precise, let \(r(\alpha)\) denote the \(s\) rows selected and \(c(\beta)\) the \(s\) columns selected for \(\alpha, \beta\). Furthermore let \(\tilde{T}^{n,k} = \{l \in I^{n,k} : i^l \in r(\alpha), j^l \in c(\beta)\}\), which is the index set of matrices \(e_i^l \otimes e_j^l\) that would still be selected when considering the \(s \times s\) submatrix for \(\text{adj}_{s}(\cdot)\) \(\alpha\beta\). Then clearly we have \(\text{adj}_{s\mathcal{E}}^{F,k-1}(I^{n,k}) = \text{adj}_{s\mathcal{E}}^{F,k-1}(\tilde{T}^{n,k})\), since the other matrices are ‘not seen’ by the adjugate at the \(\alpha, \beta\) entry. From this consideration we can conclude two things. The first is that for \(s < k\) we obtain

\[
\text{adj}_{s}^{F,k-1}(I^{n,k}) = 0,
\]
due to $|\hat{I}^{n,k}| \leq s < k$ and property [1]. Similarly, if $s = k$ we obtain

$$\text{adj}_{k}^{F,k-1}(I^{n,k}) = 0,$$

if $I^{n,k} \neq \{ l \in I^{n,k} : i^l \in r(\alpha), j^l \in c(\beta) \}$. Thus, we now only need to show that if $I^{n,k} = \{ l \in I^{n,k} : i^l \in r(\alpha), j^l \in c(\beta) \}$ then

$$[\text{adj}_{k}^{F,k-1}(I^{n,k})]_{\alpha \beta} = \pm 1.$$

We do so by showing the equation

$$\det_{\mathcal{E}}^{F,k-1}(S^k) = \det_{\mathcal{E}}^{0}(S^k) = \det \left( \sum_{l=1}^{k} E_{l} \right), \tag{4.40}$$

where $\mathcal{E} = (E_1, \ldots, E_k)$ with $E_l = e_{i^l} \otimes e_{j^l}, l = 1, \ldots, k$ is such that $i^l \neq i^m$ and $j^l \neq j^m$ for all $l \neq m$ (every one is in different row and column). As with many proofs in this section we also use induction here to verify the claim. The base case for $k = 1$ is easily dealt with since in dimension $1 \times 1 \det$ is simply the identity function. Now assume (4.40) holds for $k - 1$. Furthermore, denote by $M_{ij}(F)$ the matrix obtained by deleting the $i$-th row and $j$-th column of $F$ and analogously to our familiar notation $\Lambda^F$ we denote by $M^F_{i,j,\mathcal{E}}(J)$ the matrix $M_{ij}(F + \sum_{j \in J} E_j)$. In the following we want to expand the determinant along the $i^k$-th row. In order to do so we need to identify in which cases the $i^k$-th row contains an entry generated by a matrix $E_j$. We have for $0 < l < k$ that

$$\left( S^k \atop l \right) = \left( S^{k-1} \atop l \right) \cup \left( S^{k-1} \cup \{k\} \atop l - 1 \right), \tag{4.41}$$

and in the first family of sets in the union the index excludes $k$ and so the $i^k$-th row will not include a one, in contrast to the second family. Then we obtain that

$$\det_{\mathcal{E}}^{F,k-1}(S^k) = \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l \in (S^k \atop l)} \det_{\mathcal{E}}^{F}(J_l)$$

$$= \det_{\mathcal{E}}^{F}(S^k) + (-1)^k \det_{\mathcal{E}}^{F}(\emptyset) + \sum_{l=1}^{k-1} \sum_{J_l \in (S^{k-1} \atop l)} (-1)^{k-l} \det_{\mathcal{E}}^{F}(J_l)$$

$$+ \sum_{l=1}^{k-1} \sum_{J_{l-1} \in (S^{k-1} \atop l-1)} \det_{\mathcal{E}}^{F}(J_{l-1} \cup \{k\})$$

97
\[
\begin{align*}
&= \sum_{l=0}^{k-1} (-1)^{k-l} \sum_{J_l \in \binom{S^{k-1}}{l}} \det_{E}^F (J_l) \\
&\quad + \sum_{l=1}^{k} (-1)^{k-l} \sum_{J_{l-1} \in \binom{S^{k-1}}{l-1}} \det_{E}^F (J_{l-1} \cup \{k\}),
\end{align*}
\]
where the first equality follows by the property (4.38) of $\Lambda_{F,k}^E$ for $\Lambda = \det$, the second by separating out $l = 0$ and $l = k$ and using (4.41) and the third equality by the first and second terms with the respective sums by changing the limits appropriately. Now by expanding the determinant along the $i^k$-th row we find
\[
\begin{align*}
\det_{E}^{F,k-1} (S^k) &= \sum_{l=0}^{k-1} (-1)^{k-l} \sum_{J_l \in \binom{S^{k-1}}{l}} \sum_{m=1}^{D} (-1)^{k+m} F^{i^k m} \det (M^{F,\{k\}}^{i^k m, E}(J_l)) \\
&\quad + \sum_{l=1}^{k} (-1)^{k-l} \sum_{J_{l-1} \in \binom{S^{k-1}}{l-1}} \sum_{m=1}^{D} (-1)^{k+m} [F + e_{i^k} \otimes e_{j^k}]^{i^k m} \det (M^{F,\{k\}}^{i^k m, E}(J_{l-1})).
\end{align*}
\]
Performing an index shift in the second sum and noting that $[F+e_{i^k} \otimes e_{j^k}]^{i^k m} = F^{i^k m} + \delta_{j^k m}$ the above reduces to
\[
\begin{align*}
\det_{E}^{F,k-1} (S^k) &= -\sum_{l=0}^{k-1} (-1)^{k-l} \sum_{J_l \in \binom{S^{k-1}}{l}} \sum_{m=1}^{D} (-1)^{k+m} \delta_{j^k m} \det (M^{F,\{k\}}^{i^k m, E}(J_l)) \\
&\quad + \sum_{l=0}^{k-1} (-1)^{k-1-l} \sum_{J_l \in \binom{S^{k-1}}{l}} (-1)^{j^k} \det (M^{F,\{k\}}^{i^k j^k, E}(J_l)) \\
&\quad = (-1)^{j^k} \det_{E}^{M^{i^k j^k, \{k\}} (F), \{k\}-2} (S^{k-1}),
\end{align*}
\]
where in a slight abuse of notation $\det_{E}^{M^{i^k j^k, \{k\}} (F), \{k\}-2} (J)$ is supposed to refer to $\det(M^{i^k j^k, \{k\}} (F + \sum_{j \in J} E_j))$ as compared to $\det(M^{i^k j^k, \{k\}} (F) + \sum_{j \in J} E_j)$. Then by the inductive hypothesis
\[
\det_{E}^{F,k-1} (S^k) = (-1)^{j^k} \det \left( M^{i^k j^k} \left( \sum_{l=1}^{k-1} E_l \right) \right) = \det \left( \sum_{l=1}^{k} E_l \right),
\]
thus proving (4.40). Note that this is completely independent of the chosen point $F \in \mathbb{R}^{d \times D}$. With this it is easy to see that in fact the system (4.30) has a solution. \(\square\)
Finally we are able to prove Theorem 4.20.

Proof of Theorem 4.20. We proof this theorem by showing the implications ‘(i) ⇒ (ii)’, ‘(ii) ⇒ (iii)’ and ‘(iii) ⇒ (i)’. With all of the above work we have now already proved ‘(ii) ⇒ (iii)’, ‘(iii) ⇒ (i)’ is trivial and thus we proceed with showing ‘(i) ⇒ (ii)’. Fortunately, the proof is straightforward.

‘(i) ⇒ (ii)’. Let $h$ be $n$-polyaffine at $F$. Further let $A \in \mathbb{R}^{d \times D}$, $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^D$ such that the rank-one line $\{A + tu \otimes v : t \in \mathbb{R}\}$ is completely contained in $C^n(F)$. This implies that $m := \text{rank}(A - F) \leq n$ and so there exists $u_i \in \mathbb{R}^d, v_i \in \mathbb{R}^D, i = 1, \ldots, m$ such that $A = F + \sum_{i=1}^{m} u_i \otimes v_i$. Now define $V := \text{span}(\{u_i \otimes v_i : i = 1, \ldots, m\} \cup \{u \otimes v\})$. If $m < n$ then $V$ is clearly a simple rank-$n$ subspace. In the case of $m = n$ we still need to show that the tensor $u \otimes v$ does not add to the rank of $V$. It is easy to see that this is the case when either $u$ or $v$ is a linear combination of $u_i, i = 1, \ldots, n$ or $v_i, i = 1, \ldots, m$ respectively. Thus, the only case that remains to investigate is when $\{u_i : i = 1, \ldots, n\} \cup \{u\}$ and $\{v_i : i = 1, \ldots, n\} \cup \{v\}$ are both linearly independent sets. In this case, consider the matrix $B = \sum_{i=1}^{n} u_i \otimes v_i + u \otimes v$. Clearly, $\text{col}(B) \subseteq \text{span}(\{u_i \cup \{u\})$ and hence $\text{rank}(B) \leq n + 1$. Now let $\tilde{u} = \sum_{i=1}^{n} \lambda_i u_i + \lambda u \in \text{span}(\{u_i \cup \{u\})$. We then want to show that $\tilde{u} \in \text{col}(B)$, i.e. that there exists $x \in \mathbb{R}^D$ s.t. $Bx = \tilde{u}$. It is

$$\tilde{u} = Bx \iff \sum_{i=1}^{n} \lambda_i u_i + \lambda u = \sum_{i=1}^{n} (v_i, x) u_i + (v, x)u,$$

which, due to linear independence of $\{u_i \cup \{u\}$ is equivalent to $\lambda_i = (v_i, x), i = 1, \ldots, n$ and $\lambda = (v, x)$. These equations in turn are equivalent to the system

$$ \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \\ v \end{pmatrix} x = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \lambda \end{pmatrix}$$

This system always has a solution since the matrix on the left has got rank $n + 1$, since its rows span a $(n + 1)$-dimensional vector space. Thus, we conclude that $\text{col}(B) = \text{span}(\{u_i \cup \{u\})$ and hence $\text{rank}(B) = \text{dim} \text{col}(B) = n + 1$, which is a contradiction to $\text{rank}(B) \leq n$. Therefore $V$, as defined above, is a simple rank-$n$ subspace. Now define $\varphi(t) = h(A + tu \otimes v) = h(F + \sum_{i=1}^{m} u_i \otimes v_i + tu \otimes v)$. We have to show that $\varphi$ is a linear function. We do so by constructing $\{F_i\}_{i=1,...,m+1} \subseteq F + V$ such that $T(F) = \sum_{i=1}^{m+1} \lambda_i T(F_i)$ for $\lambda \in \Lambda_{m+1}$ and using that $h$ satisfies (4.16). We claim that it
holds that
\[
T(F) = \frac{1}{2k} T\left(F + \sum_{i=1}^{k} u_i \otimes v_i\right) + \sum_{i=1}^{k} \frac{1}{2i} T\left(F + \sum_{l=1}^{i-1} u_l \otimes v_l - u_i \otimes v_i\right),
\] (4.42)
which we prove by induction. For \(k = 1\) the above equation simplifies to
\[
T(F) = \frac{1}{2} T(F + u_1 \otimes v_1) + \frac{1}{2} T(F - u_1 \otimes v_1),
\]
which is true simply by linearity of \(T\) along rank-one lines. Now assume (4.42) holds for \(k\) and consider the case \(k + 1\). Then
\[
T(F) = \frac{1}{2k} T\left(F + \sum_{i=1}^{k+1} u_i \otimes v_i\right) + \sum_{i=1}^{k} \frac{1}{2i} T\left(F + \sum_{l=1}^{i-1} u_l \otimes v_l - u_i \otimes v_i\right)
\]
\[= \frac{1}{2k+1} T\left(F + \sum_{i=1}^{k+1} u_i \otimes v_i\right) + \frac{1}{2(k+1)} T\left(F + \sum_{i=1}^{k} u_i \otimes v_i - u_{k+1} \otimes v_{k+1}\right)
\]
\[+ \sum_{i=1}^{k} \frac{1}{2i} T\left(F + \sum_{l=1}^{i-1} u_l \otimes v_l - u_i \otimes v_i\right)
\]
\[= \frac{1}{2k+1} T\left(F + \sum_{i=1}^{k+1} u_i \otimes v_i\right) + \sum_{i=1}^{k} \frac{1}{2i} T\left(F + \sum_{l=1}^{i-1} u_l \otimes v_l - u_i \otimes v_i\right),
\]
which proves (4.42), and which we use with \(m = k\). We then obtain that
\[
T(F) = \frac{1}{2^m} T(A) + \sum_{i=1}^{m} \frac{1}{2i} T\left(F + \sum_{l=1}^{i-1} u_l \otimes v_l - u_i \otimes v_i\right).
\]
Note that further \(T(A) = \frac{1}{1+t} T(A - u \otimes v) + \frac{1}{1+t} T(A + tu \otimes v)\). Hence, by splitting the first term of the last equation once more we obtain
\[
T(F) = \frac{1}{2^m} \left[ \frac{t}{1+t} T(A - u \otimes v) + \frac{1}{1+t} T(A + tu \otimes v) \right]
\[+ \sum_{i=1}^{m} \frac{1}{2i} T\left(F + \sum_{l=1}^{i-1} u_l \otimes v_l - u_i \otimes v_i\right),
\]
where all the matrices participating in the convex combination are elements of \(F + V\).
Then we find that
\[
\begin{aligned}
  h(F) &= 1 + \frac{1}{2m} \left[ \frac{t}{1 + t} h(A - u \otimes v) + \frac{1}{1 + t} h(A + tu \otimes v) \right] \\
  &\quad + \sum_{i=1}^{m} \frac{1}{2^i} h \left( F + \sum_{i=1}^{i-1} u_i \otimes v_i - u_i \otimes v_i \right), \\
\end{aligned}
\]
which, after solving for \(h(A + tu \otimes v)\) yields
\[
  h(A + tu \otimes v) = 2^m (1 + t) h(F) - th(A - u \otimes v) - 2^n (1 + t)c,
\]
which is a linear function in \(t\).

**Corollary 4.29.** Let \(h : \mathbb{R}^{d \times D} \to \mathbb{R}\). Then the following statements are equivalent:

1. \(h\) is polyaffine,
2. \(h\) is \((d \land D)\)-polyaffine at \(F\) for arbitrary \(F \in \mathbb{R}^{d \times D}\).

For completeness we are now also in the position to prove the upper bound on the dimension of the subspace \(T_V\) from Theorem 4.8.

**Proof of** \(\tau(n) \leq \sum_{s=1}^{n} \binom{d}{s} \binom{D}{s}\) of Theorem 4.8. Recall that for \(F \in \mathbb{R}^{d \times D}\) and \(V \subseteq \mathbb{R}^{d \times D}\) simple rank-\(n\) we define \(T_V = \text{co}(T(F + V) - T(F))\). We have shown in the proof of Theorem 4.8 that \(T_V\) is indeed a subspace. Note that \(F + V \subseteq C^n(F)\) for all \(V \subseteq \mathbb{R}^{d \times D}\) simple rank-\(n\) and thus \(T_V \subseteq \text{co}(T(C^n(F)) - T(F)) \subseteq \text{span}(T(C^n(F)) - T(F))\). We now show that \(C := \text{span}(T(C^n(F)) - T(F))\) is a subspace with \(\dim C \leq \sum_{s=1}^{n} \binom{d}{s} \binom{D}{s} = \tau(n)\) thus proving the assertion.

For \(\tilde{F} \in C^n(F)\) we have \(\tilde{F} = F + \sum_{x=1}^{n} u_x \otimes v_x\) for some \(u_x \in \mathbb{R}^d, v_x \in \mathbb{R}^D, x = 1, \ldots, n\) and with (4.24) and the notation \(\Lambda_{\varepsilon}^{F,k}\) for \(\Lambda = T\) we obtain that
\[
  T(\tilde{F}) - T(F) = T_{\varepsilon}^{1} + \sum_{x=1}^{n} u_x \otimes v_x (\emptyset) - T_{\varepsilon}^{1}(\emptyset) \\
  = \sum_{k=1}^{n} \sum_{\varepsilon \in \left\{ E_{1}, \ldots, E_{k} \right\}} T_{\varepsilon}^{F,k-1}(S^{k}) U_{\varepsilon}(I^{n,k}).
\]
Hence, as \(U_{\varepsilon}(I^{n,k})\) are just coefficients, \(C \subseteq \text{span}\{T_{\varepsilon}^{F,k-1}(S^{k}) : \varepsilon \in E^{k}, k = 1, \ldots, n\}\). In the foregoing sections we argued that most of the vectors \(T_{\varepsilon}^{F,k-1}\) are zero, namely those where \(\varepsilon = (E_{1}, \ldots, E_{k}) \in E^{k}\) contains any two matrices \(E_{i}\) and \(E_{j}\) with a one in the same row or column. Thus, only choices of \(\varepsilon \in E^{k}\) are nonzero where each one of \(E_{i}\) is in
a different row and column to any of the other ones in $E_j$. Selecting $k$ rows and columns of a matrix $F \in \mathbb{R}^{d \times D}$ there are then exactly $k!$ possible ways of distributing ones in such a way. However, by property (iii) of $T_{E,k}^k$, all of these are either the same or the negative of one of the choices. Therefore, for each $k$ and a fixed choice of $k$ rows and columns, there is only one linearly independent element $T_{E,k}^k(S^k)$ for $E$ containing only ones in those chosen rows and columns. As there are $\binom{d}{k}$ and $\binom{D}{k}$ possible ways of fixing rows and columns respectively for each $k$ we have a maximum of $\sum_{k=1}^{n} \binom{d}{k} \binom{D}{k}$ linearly independent vectors in $\{T_{E,k}^k(S^k) : E \in E^k\}$ and therefore also $\dim C \leq \sum_{s=1}^{n} \binom{d}{s} \binom{D}{s} = \tau(n)$. 

4.2.3. 1-polyconvexity as the pointwise supremum of 1-polyaffines

We now want to turn our attention to equivalent formulations of $n$-polyconvexity in terms of subdifferentials or as we would call them at this stage: $n$-polyaffines. The following result for the 1-polyconvex case is the most striking original result for this theory as it establishes a new, yet equivalent, characterisation of finite rank-one convex functions. This new characterisation approaches 1-polyconvexity from the viewpoint of convex analysis or the more general field of abstract convexity. Given the previous definitions of rank-one, poly- and quasiaffine functions and the result on their equivalence it was probably deemed impossible to achieve such a characterisation for rank-one convexity. However, with the idea of attributing the property of being rank-one affine to the function for a given point $F \in \mathbb{R}^{d \times D}$ we were able to show that $n$-polyaffine functions at $F \in \mathbb{R}^{d \times D}$ are more general than rank-one, poly- and quasiaffine functions and the new characterisation becomes possible. We are going to treat the case $n = 1$ separately from $n > 1$ as in the latter case we will need to make further assumptions. Note that for $(d \wedge D)$-polyconvexity it is already known that finite polyconvex functions can be written as the pointwise supremum of polyaffine function and $(d \wedge D)$-polyaffine functions at $F \in \mathbb{R}^{d \times D}$ are simply polyaffine functions for any $F \in \mathbb{R}^{d \times D}$.

The main result of this section is the following.

**Theorem 4.30.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$, i.e. $f$ takes only finite values. Then the following conditions are equivalent:

(i) $f$ is rank-one convex

(ii) for every $F \in \mathbb{R}^{d \times D}$ there exists $\beta(F) \in \mathbb{R}^{\tau(d,D)}$ such that

$$f(\tilde{F}) - f(F) \geq \langle \beta(F), T(\tilde{F}) - T(F) \rangle$$  \hspace{1cm} (4.43)
for all $\tilde{F} \in C^1(F)$ and where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^{\tau(d,D)}$.

(iii) at each $F \in \mathbb{R}^{d \times D}$ $f$ can be written as the pointwise supremum of rank-one affine functions at $F$ below $f$, i.e.

$$f(F) = \sup \{ h(F) : h(\cdot) = \langle \beta(F), T(\cdot) - T(F) \rangle + h(F), h \leq f \text{ on } C^1(F), \beta(F) \in \mathbb{R}^{\tau(d,D)} \}. \quad (4.44)$$

Note this theorem (together with Theorem 4.8) compares to Theorem 5.6 in [20] for the polyconvex case. The author of [20] remarks that ‘[there] is no known equivalent to Theorem 5.6 for rank-one convex functions’. The above theorem fills this gap.

Proof of Theorem 4.30. (i) ⇒ (ii) Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ be rank-one convex. As is the case for proving the existence of a subdifferential for the convex functions at a point $F \in \mathbb{R}^{d \times D}$ we also define the one-sided directional derivative $f'(F; \cdot)$. However, here we only define it at those points where we can guarantee it exists, i.e. $f'(F; \cdot) : C^1(0) \to \mathbb{R}$ s.t.

$$f'(F; Y) := \lim_{t \to 0} \frac{f(F + tY) - f(F)}{t}. \label{eqn:directional_derivative}$$

Each of those limits exists and is finite since $f$ is rank-one convex on $\mathbb{R}^{d \times D}$ and thus convex on the infinite line $\{F + tY : t \in \mathbb{R}\}$ with $Y \in C^1(0)$. The difference quotient on the right is a monotonically increasing function of $t$. Note that the directional derivative is usually defined on the whole space instead of just the cone of rank-one directions. Furthermore $f'(F; \cdot)$ is a positively one-homogeneous function, i.e. $f'(F; tY) = tf'(F; Y)$ for all $t > 0$ and $Y \in C^1(0)$. It is also rank-one convex on $C^1(0)$, i.e. whenever there are $Y_1, Y_2 \in C^1(0)$ s.t. $\text{rank}(Y_1 - Y_2) \leq 1$ then $f'(F; tY_1 + (1 - t)Y_2) \leq tf'(F; Y_1) + (1 - t)f'(F; Y_2)$. This follows from the same methods used for $f'(F; \cdot)$ in standard convex analysis. Moreover, it holds that

$$f(F + tY) - f(F) \geq tf'(F; Y) \quad (4.45)$$

for all $t > 0$ and $Y \in C^1(0)$. The aim is to show that there exists a $\beta \in \mathbb{R}^{d \times D}$ s.t.

$$f'(F; Y) \geq \langle \beta, Y \rangle$$
for all $Y \in \mathcal{C}^1(0)$. If this is the case we are finished, since then

$$f(\tilde{F}) - f(F) \geq f'(F; \tilde{F} - F) \geq \langle \beta, \tilde{F} - F \rangle = \langle \hat{\beta}, T(\tilde{F}) - T(F) \rangle$$

where $\hat{\beta} = [\beta, 0, \ldots, 0] \in \mathbb{R}^{\tau(d,D)}$ is only filling up the places corresponding to the adjugates $\text{adj}_s(F)$ in $T(F)$ for $s > 1$ with zeros. In order to prove this claim we proceed in two steps. First we define a function $g$ such that $g \leq f'(F; \cdot)$ on $\mathcal{C}^1(F)$ and which is convex. Then we show that $g$ is proper by showing that $g(0) = 0$ and therefore that $g$ has a subdifferential at 0. This subdifferential is exactly what we need.

We define $g : \text{co}(\mathcal{C}^1(0)) \to \mathbb{R} \cup \{-\infty\}$ as follows:

$$g(X) = \inf \{ \lambda_i f'(F; Y_i) : X = \sum_{i=1}^{k} \lambda_i Y_i, k \in \mathbb{N}, \lambda \in \Lambda_k, Y_i \in \mathcal{C}^1(0) \}.$$ 

To prove that $g$ is convex let $X_1, X_2 \in \text{co}(\mathcal{C}^1(0))$ and $\varepsilon > 0$. Then there exist $Y_1^1, Y_2^1 \in \mathcal{C}^1(0)$, $i = 1, \ldots, k$ and $\lambda^1, \lambda^2 \in \Lambda_k$ such that $X_j = \sum_{i=1}^{k} \lambda^j_i Y_i^j$ and $g(X_j) + \varepsilon \geq \sum_{i=1}^{k} \lambda^j_i f(Y_i^j)$. Furthermore, for $\mu \in [0, 1]$ we then have $\mu X_1 + (1 - \mu) X_2 = \sum_{i=1}^{k} \mu \lambda^1_i Y_i^1 + (1 - \mu) \lambda^2_i Y_i^2$ and thus by the definition of $g$ and the above

$$g(\mu X_1 + (1 - \mu) X_2) \leq \sum_{i=1}^{k} \mu \lambda^1_i f'(F; Y_i^1) + (1 - \mu) \lambda^2_i f'(F; Y_i^2) \leq \mu g(X_1) + (1 - \mu) g(X_2) + 2\varepsilon.$$ 

Since $\varepsilon$ is arbitrary we find that $g$ is convex on $\text{co}(\mathcal{C}^1(0))$. However, in the definition of $g$ it is not excluded that $g \equiv -\infty$, so we still need to show that $g$ is proper. For that we show that $g(0) = 0$. Obviously $g(0) \leq f'(F; 0) = 0$. It remains to show that $g(0) \geq 0$. Therefore let $Y_i \in \mathcal{C}^1(0)$ such that $0 = \sum_{i=1}^{k} \lambda_i Y_i$ for some $\lambda \in \Lambda_k$. We show that $Y_1, \ldots, Y_k$ form a degenerate $T_k$ configuration with 0 as the common point of rank-one connections. ($T_k$ configurations will be introduced in Section 4.3.1 A degenerate $T_k$ configurations $Y_1, \ldots, Y_k$ is one which is not a $T_k$ configuration but for which a sequence of $T_k$ configuration $Y_1^\varepsilon, \ldots, Y_k^\varepsilon$ exists such that $Y_i^\varepsilon \to Y_i$ as $\varepsilon \to 0$.) We claim that then it must hold that

$$0 = f'(F; 0) \leq \sum_{i=1}^{k} \lambda_i f'(F; Y_i).$$

If that is the case then $g(0) \geq 0$ follows easily since $g(0)$ is simply the infimum of all such configurations $Y_1, \ldots, Y_k \in \mathcal{C}^1(0)$ by definition. If $f'(F; \cdot)$ was finite and rank-one convex
on all of $\mathbb{R}^{d \times D}$ then this would be a known result. However, $f'(F; \cdot)$ is only defined on $C^1(0)$, so we cannot apply this result. Nevertheless, we can make use of the fact that $f'(F; \cdot)$ is the directional derivative of a finite and rank-one convex function defined on all of $\mathbb{R}^{d \times D}$. If for $f$ itself it holds that

$$f(F) = f\left(F + \sum_{i=1}^{k} t \lambda_i Y_i\right) \leq \sum_{i=1}^{k} \lambda_i f(F + t Y_i) \quad (4.46)$$

for all $t > 0$ then

$$0 \leq \frac{\sum_{i=1}^{k} \lambda_i f(F + t Y_i) - f(F)}{t} = \sum_{i=1}^{k} \lambda_i \frac{f(F + t Y_i) - f(F)}{t}.$$

The limit of the above for $t \to 0^+$ exists since the limit of each summand on the right exists (and is finite). From the left we obtain that the limit must be positive, so overall that

$$0 \leq \sum_{i=1}^{k} f'(F; Y_i),$$

which is the desired result.

Therefore, last but not least, we must prove that $Y_1, \ldots, Y_k$ form a degenerated $T_k$ configuration. For this we construct a sequence $Y_1^\varepsilon, \ldots, Y_k^\varepsilon$ of approximating nondegenerate $T_k$ configuration such that $Y_i^\varepsilon \to Y_i$ as $\varepsilon \to 0$. The idea for the construction was taken from the example of a particular degenerated $T_4$ configuration in [28], see Example 4.18. Define

$$X_i^\varepsilon = \varepsilon \sum_{j=1}^{i} \lambda_j Y_j, \quad (4.47)$$

$$Y_i^\varepsilon = Y_i + X_i^\varepsilon. \quad (4.48)$$

The situation is depicted in Figure 4.5. Then we have that $\operatorname{rank}(Y_{i+1}^\varepsilon - X_i^\varepsilon) = \operatorname{rank}(Y_{i+1}) \leq 1$ and $X_{i+1}^\varepsilon \in [X_i^\varepsilon, Y_{i+1}^\varepsilon]$ with

$$X_{i+1}^\varepsilon = \frac{1}{1 + \varepsilon \lambda_{i+1}} X_i^\varepsilon + \frac{\varepsilon \lambda_{i+1}}{1 + \varepsilon \lambda_{i+1}} Y_{i+1}^\varepsilon \quad (4.49)$$

for $i = 1, \ldots, k$ with the indices taken modulo $k$ in the sense that $k + 1$ corresponds
Figure 4.5.: Degenerated $T_4$ configuration $\{Y_1, \ldots, Y_4\}$ with approximating nondegenerate $T_4$ configuration $\{Y_1^\varepsilon, \ldots, Y_4^\varepsilon\}$ and auxiliary points $\{X_1^\varepsilon, \ldots, X_4^\varepsilon\}$.

It follows from rank-one convexity that we must have

$$f(F) \leq \sum_{i=1}^k \lambda_i(\mu^\varepsilon) f(F + Y_i^\varepsilon).$$

(4.50)

Furthermore, in a simple but rather long calculation we can show that as $\varepsilon \to 0$ we have $\lambda_i(\mu^\varepsilon) \to \lambda_i$. We omit the proof of this here and instead refer to appendix A.5. Thus, in the limit of (4.50) we recover (4.46), which proves that $g$ is proper.

Now, since $g$ is proper and convex on $C^1(0)$ it has a subdifferential at 0, so there exists a $\beta \in \mathbb{R}^{d \times D}$

$$g(X) \geq \langle \beta, X \rangle$$

for all $X \in \mathbb{R}^{d \times D}$. So in particular we obtain

$$f'(F; Y) \geq g(Y) \geq \langle \beta, Y \rangle$$

for all $Y \in C^1(0)$, which finishes the first implication.
‘(ii) ⇒ (iii)’ Let \( f : \mathbb{R}^{d \times D} \rightarrow \mathbb{R} \) be such that for all \( F \in \mathbb{R}^{d \times D} \) there exists \( \beta(F) \in \mathbb{R}^{r(d,D)} \) such that (4.43) holds. Then at each \( F \) define \( h_F : \mathcal{C}^1(F) \rightarrow \mathbb{R} \) with \( h_F(\tilde{F}) = \langle \beta(F), T(\tilde{F}) - T(F) \rangle + f(F) \). Clearly \( h_F \leq f \) on \( \mathcal{C}^1(F) \) and \( h_F(F) = f(F) \). Thus (4.44) is true.

‘(iii) ⇒ (i)’ Assume (iii) holds. Let \( F_1, F_2 \in \mathbb{R}^{d \times D} \) be such that \( \text{rank}(F_1 - F_2) \leq 1 \), let \( \lambda \in (0,1) \) and \( F = \lambda F_1 + (1 - \lambda)F_2 \). For \( \varepsilon > 0 \) there exists \( \beta(F) \) such that for \( h(\cdot) = \langle \beta(F), T(\cdot) - T(F) \rangle + h(F) \) we have \( h \leq f \) on \( \mathcal{C}^1(F) \) and \( f(F) - \varepsilon \leq h(F) \). Thus, due to the linearity of \( T \) on rank-one connections and the linear character of \( h \), we have
\[
f(F) - \varepsilon \leq h(F) = h(\lambda F_1 + (1 - \lambda)F_2) = \lambda h(F_1) + (1 - \lambda)h(F_2) \leq \lambda f(F_1) + (1 - \lambda)f(F_2).
\]
Hence, since \( \varepsilon > 0 \) is arbitrary, \( f \) is rank-one convex.

Note that this version of 1-polyconvexity is only generally true for finite functions as the following example shows.

**Example 4.31.** Let

\[
A_1 = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}, \quad A_4 = \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}
\]

and \( f : \mathbb{R}^{2 \times 2} \rightarrow \mathbb{R} \) such that for some \( f_0 \in \mathbb{R} \)

\[
f(F) = \begin{cases} (1 - s)f_0, & F = sA_i, s \in [0,1] \\ +\infty, & \text{otherwise} \end{cases}
\]

Because none of the \( A_i, i = 1, \ldots, 4 \) are rank-one connected \( f \) is rank-one convex for all \( f_0 \in \mathbb{R} \). Furthermore we have \( A_i \in \mathcal{C}^1(0) \) and it is \( T(0) = \sum_{i=1}^4 \frac{1}{4}T(A_i) \). However, for \( f_0 > 0 \) there exists no \( \beta \in \mathbb{R}^{r(d,D)} \) such that

\[
f(F) - f(0) \geq \langle \beta, T(F) \rangle
\]

for all \( F \in \mathcal{C}^1(F) \). If so, we would have that \( f(A_i) - f(0) \geq \langle \beta, T(A_i) \rangle \) or equivalently \(-f_0 \geq \langle \beta, T(A_i) \rangle \). Since \( 0 = \sum_{i=1}^4 \frac{1}{4}T(A_i) \) this would imply \(-f_0 \geq 0 \), but we have \( f_0 > 0 \).

Hence, finiteness of the function \( f \) is necessary for Theorem 4.30. In fact, the \( A_i \) form a degenerated \( T_4 \) configuration and the rank-one convex hull of the set \( \{ A_i \}_{i=1,...,4} \) is the set \( S = \{ sA_i : s \in [0,1], i = 1, \ldots, 4 \} \), cf. [59] Th. 2. Thus for any finite function \( \tilde{f} \)
with \( \tilde{f}|_S = f \) and \( \tilde{f} \) being rank-one convex we must necessarily have that \( f_0 \leq 0 \).

**Remark 4.32.** (i) The crucial step in this proof was to show that \( f'(F; Y) \geq \langle \beta, Y \rangle \) for all \( Y \in \mathcal{C}^1(0) \). In the extended real-valued case from example 4.31 this fails if \( f_0 > 0 \). With \( f'(F; \cdot) \) still defined as in (4.45) on \( \mathcal{C}^1(0) \) all limits still exists, however they may now take the value \( +\infty \). Indeed we have \( f'(0; A_i) = \lim_{t \searrow 0} 1/t (f(tA_i) - f(0)) = \lim_{t \searrow 0} 1/t ((1 - t)f_0 - f_0) = -f_0 \) for \( i = 1, \ldots, 4 \). However, for all other directions (including \( -A_i \)) we have \( f'(0; Y) = +\infty \), \( Y /\in S = \{ sA_i : s \geq 0, i = 1, \ldots, 4 \} \). If there existed \( \beta \in \mathbb{R}^{2\times 2} \) such that \( f'(0; Y) \geq \langle \beta, Y \rangle \) for all \( Y \in \mathcal{C}^1(0) \) then it would have to hold that

\[
-4f_0 = \sum_{i=1}^{4} f'(0; A_i) \geq \sum_{i=1}^{4} \langle \beta, A_i \rangle = 0,
\]

which is not true if \( f_0 > 0 \). Again it seems that the case \( f_0 > 0 \) can be excluded if \( f'(0; \cdot) \) is finite on all of \( \mathcal{C}^1(0) \) since the \( A_i \) form a degenerated \( T_4 \) configuration. However, we were not able to conclude that from just \( f'(0; \cdot) \) itself since the approximating (non-)degenerate \( T_k \) configurations used in the proof need \( f \) to be finite around a neighbourhood of the rank-1 cone at 0.

(ii) Kirchheim and Kristensen [29] consider \( D \)-convexity and are able to prove that any function \( f : C \to \mathbb{R} \), where \( C \) is an open subset of a vector space \( V \), that has a \( D \)-subcone at a point \( x \in C \), will also possess a subdifferential at \( x \in C \). Proving that if \( C \) is furthermore an open convex cone and \( f \) additionally \( D \)-convex and positively 1-homogeneous, that it possesses a \( D \)-subcone at any \( x \in C \cap D \) they conclude that such \( f \) has indeed a subdifferential at those \( x \). Our result is not quite comparable to their situation as we make slightly different assumptions, but in other ways they are similar. In terms of \( D \)-convexity our result would read that for any \( x \in V \) and any \( D \)-convex function that is finite in a neighbourhood of \( x + D \) we have that \( f \) has a subdifferential at \( x \) on \( x + D \). Note that we do not need the assumption of homogeneity as on the set \( x + D \) itself the existence of a \( D \)-subcone on \( x + D \) is trivial by \( D \)-convexity. Hence, Kirchheim’s and Kristensen’s assumptions are stronger, but so are their conclusions as in their case the subdifferential inequality does not only hold on the set \( x + D \) but on the whole open convex cone \( C \).

#### 4.2.4. \( n \)-polyconvexity as the pointwise supremum of \( n \)-polyaffines

In the beginning of the previous section we already alluded to the fact that we need to make further assumptions in order to prove a similar result of equivalence between
finite \( n \)-polyconvex functions and their representation as the pointwise supremum of \( n \)-polyaffine functions for \( 1 < n < d \wedge D \). Those further assumptions will take the form of a new definition that we call strong \( n \)-polyconvexity. In this case we can show that finite strong \( n \)-polyconvex functions can be written as the pointwise supremum of \( n \)-polyaffine functions and we discuss the relationships between strong \( 1 \)-polyconvexity and rank-one convexity, and \( (d \wedge D) \)-polyconvexity and polyconvexity. It is easy to show that any function that can be written as the pointwise supremum of \( n \)-polyaffine functions is \( n \)-polyaffine. However, the reverse is not trivial. To begin with we present the statement without the use of the further assumptions needed to prove the result as a conjecture.

**Conjecture 4.33.** Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \), i.e. \( f \) takes only finite values. Then the following conditions are equivalent:

(i) \( f \) is \( n \)-polyconvex

(ii) for every \( F \in \mathbb{R}^{d \times D} \) there exists \( \beta(F) \in \mathbb{R}^{(d,D)} \) such that

\[
f(\tilde{F}) - f(F) \geq \langle \beta(F), T(\tilde{F}) - T(F) \rangle
\]

for all \( \tilde{F} \in C^n(F) \) and where \( \langle \cdot, \cdot \rangle \) denotes the scalar product in \( \mathbb{R}^{(d,D)} \).

(iii) at each \( F \in \mathbb{R}^{d \times D} \) \( f \) can be written as the pointwise supremum of \( n \)-polyaffine functions at \( F \) below \( f \), i.e.

\[
f(F) = \sup\{ h(F) : h(\cdot) = \langle \beta(F), T(\cdot) - T(F) \rangle + h(F), h \leq f \text{ on } C^n(F), \beta(F) \in \mathbb{R}^{(d,D)} \}.\]

It is easy to see that (ii) \( \Rightarrow \) (iii) \( \Rightarrow \) (i), so the only missing implication is to prove (i) \( \Rightarrow \) (ii). An analogue version of the proof of \( n = 1 \) for arbitrary \( n \) would require to prove that \( f(F) \leq \sum_{i=1}^{k} \lambda_i f(F_i) \) for any configuration \( F_i \in C^n(F) \) such that \( T(F) = \sum_{i=1}^{k} \lambda_i T(F_i) \) only using that \( f \) is polyconvex on affine spaces \( F + V \) for \( V \) simple rank-\( n \). For \( n = 1 \) the matrices \( F_i \) are a degenerate \( T_k \) configuration with centre \( F \) and a construction of approximating nondegenerate \( T_k \) configurations allowed us to prove the desired inequality. In the case of \( n > 1 \) we were unable to construct an approximating \( T_k^{n-pc} \) configuration of \( \{F_i\}_{i=1,...,k} \) that would allow for a similar reasoning.

However, the above motivates the following definition:

**Definition 4.34.** Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) and \( F \in \mathbb{R}^{d \times D} \). Then \( f \) is called strongly \( n \)-polyconvex at \( F \) if there exists a convex function \( c_F : \text{co} T(C^n(F)) \to \mathbb{R} \) such that
f|_{\mathcal{C}^n(F)} \geq c_F \circ T \text{ on } \mathcal{C}^n(F) \text{ and for all } F_i \in \mathcal{C}^n(F), \ i = 1, \ldots, k \text{ such that}

T(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i)

for some } \lambda \in \Lambda_{\tau(n)+1} \text{ we have that}

f(F) \leq \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i).

(4.51)

Furthermore, } f \text{ is called strongly } n \text{-polyconvex if it is strongly } n \text{-polyconvex at } F \text{ for every } F \in \mathbb{R}^{d \times D}.

Note that strong } n \text{-polyconvexity does not require simple rank-} n \text{ subspaces. For strongly } n \text{-polyconvex functions we are able to show the following.}

**Theorem 4.35.** Let } f : \mathbb{R}^{d \times D} \to \mathbb{R}, \text{ i.e. } f \text{ only takes finite values. Then the following conditions are equivalent:}

(i) } f \text{ is strongly } n \text{-polyconvex}

(ii) for every } F \in \mathbb{R}^{d \times D} \text{ there exists } \beta(F) \in \mathbb{R}^{\tau(d,D)} \text{ such that}

f(\tilde{F}) - f(F) \geq \langle \beta(F), T(\tilde{F}) - T(F) \rangle

for all } \tilde{F} \in \mathcal{C}^n(F) \text{ and where } \langle \cdot, \cdot \rangle \text{ denotes the scalar product in } \mathbb{R}^{\tau(d,D)}.

(iii) at each } F \in \mathbb{R}^{d \times D} f \text{ can be written as the pointwise supremum of } n \text{-polyaffine functions at } F \text{ below } f, \text{ i.e.}

f(F) = \sup \{ h(F) : h(\cdot) = \langle \beta(F), T(\cdot) - T(F) \rangle + h(F), h \leq f \text{ on } \mathcal{C}^n(F), \beta(F) \in \mathbb{R}^{\tau(d,D)} \}.

**Proof.** The implications ‘(ii) } \Rightarrow (iii) \Rightarrow (i)’ \text{ are trivial. We now prove ‘(i) } \Rightarrow (ii)’. Let } f : \mathbb{R}^{d \times D} \to \mathbb{R} \text{ be strongly } n \text{-polyconvex and } F \in \mathbb{R}^{d \times D}. \text{ We then define the function}

g_F : \text{co } T(\mathcal{C}^n(F)) \to \mathbb{R} \text{ such that}

\begin{align*}
g_F(X) &= \inf \{ \sum_{i=1}^{\tau(n)+1} f(F_i) : X = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i), F_i \in \mathcal{C}^n(F) \}.
\end{align*}
Note that we skip the step where we show that taking at most \( \tau(n) + 1 \) elements participating in the convex combination \( \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) \) as opposed to an arbitrary number \( I \) is sufficient. Furthermore, it is easy to show that \( g_F \) is convex and \( g_F(T(\tilde{F})) \leq f(\tilde{F}) \) for all \( \tilde{F} \in \mathcal{C}^n(F) \). Additionally, it is true that \( g_F(T(F)) = f(F) \) since \( f \) is strongly \( n \)-polyconvex at \( F \) and thus (4.51) holds. Thus, \( g_F \) is finite at \( T(F) \) and therefore subdifferentiable, meaning that there exists \( \beta_F \in \mathbb{R}^{\tau(d,D)} \) such that

\[
g_F(X) - g_F(T(F)) \geq \langle \beta_F, X - T(F) \rangle.
\]

for all \( X \in \text{co} T(\mathcal{C}^n(F)) \). Hence, by using \( f \leq g_F \circ T \) on \( \mathcal{C}^n(F) \) and \( f(F) = g_F(T(F)) \), we obtain that

\[
f(\tilde{F}) - f(F) \geq \langle \beta_F, T(\tilde{F}) - T(F) \rangle
\]

for all \( \tilde{F} \in \mathcal{C}^n(F) \).

**Remark 4.36.** It is obvious that polyconvexity, \((d \wedge D)\)-polyconvexity and strong \((d \wedge D)\)-polyconvexity are all equivalent.

### 4.3. Definition and basic properties of \( n \)-polyconvex sets

Throughout the previous sections we have closely aligned with the structure and content of [20]. This section will be no different and the corresponding part of Dacorogna’s monograph is largely a copy of [21], which we follow. We will also define \( n \)-polyconvexity for sets in the way that carries as many properties from convex analysis to this generalised case as possible and so that we will be able to establish the link to generalisation of abstract convexity as presented in Section 5. Due to the generalised approach taken in Section 4.2, we will be able to recover many more connections to standard convex analysis than given in [21].

**Definition 4.37.** Let \( E \subseteq \mathbb{R}^{d \times D} \).

(i) \( E \) is polyconvex if there exists a convex set \( K \subseteq \mathbb{R}^{\tau(d,D)} \) such that

\[
\pi(K \cap T(\mathbb{R}^{d \times D})) = E,
\]

where \( \pi \) denotes the orthogonal projection on the first \( d \times D \) components of \( \mathbb{R}^{\tau(d,D)} \) in \( \mathbb{R}^{d \times D} \). Equivalently, \( E \) is polyconvex if there exists a convex set \( K \subseteq \mathbb{R}^{\tau(d,D)} \)
such that

\[ \{ F \in \mathbb{R}^{d \times D} : T(F) \in K \} = E. \]

(ii) \( E \) is rank-one convex if for every \( \lambda \in [0, 1] \) and \( F_1, F_2 \in E \) such that \( \text{rank}(F_1 - F_2) \leq 1 \) we have

\[ \lambda F_1 + (1 - \lambda)F_2 \in E. \]

(iii) \( E \) is \( n \)-polyconvex if for all simple rank-\( n \) subspaces \( V \subseteq \mathbb{R}^{d \times D} \) and for all \( F \in E \) there exists a convex set \( K_{F,V} \subseteq \mathbb{R}^{\tau(d,D)} \) such that

\[ \pi(K_{F,V} \cap T(F + V)) = E \cap (F + V), \]

where \( \pi \) is defined as above. Equivalently, \( E \) is \( n \)-polyconvex if for all \( F \in E \) and simple rank-\( n \) subspaces \( V \) there exists a convex set \( K_{F,V} \subseteq \mathbb{R}^{\tau(d,D)} \) such that

\[ \{ \tilde{F} \in F + V : T(\tilde{F}) \in K_{F,V} \} = E \cap (F + V). \quad (4.52) \]

(iv) \( E \) is quasiconvex if

\[ F + \nabla \varphi(x)R \in E, \text{ a.e. } x \in D, \]

for some \( R \in O(n) \) and some \( \varphi \in W_{\text{per}} \) where \( W_{\text{per}} \) is the subspace of periodic functions in \( W^{1,\infty}((0,1)^d; \mathbb{R}^D) \) whose gradients take only a finite number of values.

Remark 4.38. Other definitions of quasiconvex sets exist, cf. [43, 63].

We then also obtain the more intuitive version of \( n \)-polyconvexity for sets:

Theorem 4.39. Let \( E \subseteq \mathbb{R}^{d \times D} \). Then the following conditions are equivalent

(i) \( E \) is \( n \)-polyconvex.

(ii) For all \( F \in \mathbb{R}^{d \times D} \) and simple rank-\( n \) subspaces \( V \subseteq \mathbb{R}^{d \times D} \) we have

\[ \sum_{i=1}^I \lambda_i T(F_i) = T \left( \sum_{i=1}^I \lambda_i F_i \right) \Rightarrow \sum_{i=1}^I \lambda_i F_i \in E \cap (F + V), \]

\[ F_i \in E \cap (F + V), (\lambda_1, \ldots, \lambda_I) \in \Lambda_I. \]
Moreover one can take $I = \tau(n) + 1$.

(iii) For all $F \in \mathbb{R}^{d \times D}$ and simple rank-$n$ subspaces $V \subseteq \mathbb{R}^{d \times D}$ we have

$$E \cap (F + V) = \pi(\text{co}T(E \cap (F + V)) \cap T(F + V))$$

(4.53)

or equivalently

$$E \cap (F + V) = \{\tilde{F} \in \mathbb{R}^{d \times D} : T(\tilde{F}) \in \text{co}T(E \cap (F + V))\}.$$

Proof. ‘(i) $\Rightarrow$ (ii)’: Let $E \subseteq \mathbb{R}^{d \times D}$ be $n$-polyconvex and let $F \in E$ and $V \subseteq \mathbb{R}^{d \times D}$ a simple rank-$n$ subspace be fixed. Then by definition there exists a convex set $K_{F,V} \subseteq \mathbb{R}^{\tau(d,D)}$ such that

$$\{\tilde{F} \in F + V : T(\tilde{F}) \in K_{F,V}\} = E \cap (F + V).$$

(4.52)

Further let $F_1, \ldots, F_I \in E \cap (F + V)$ and $\lambda \in \Lambda_I$ such that

$$T(\sum_{i=1}^I \lambda_i F_i) = \sum_{i=1}^I \lambda_i T(F_i).$$

Since $F_i \in E \cap F + V$ we have by (4.52) that $T(F_i) \in K_{F,V}$ for all $i = 1, \ldots, I$. Because $K_{F,V}$ is convex this implies that $T(\sum_{i=1}^I \lambda_i F_i) = \sum_{i=1}^I \lambda_i T(F_i) \in K_{F,V}$ and thus, $\sum_{i=1}^I \lambda_i F_i \in E \cap (F + V)$.

‘(ii) $\Rightarrow$ (iii)’: We have to show that (4.53) is true. It is evident that $E \cap (F + V)$ is included in the right hand side and it remains to show the opposite inclusion. Therefore, let $\tilde{F} \in \pi(\text{co}T(E \cap (F + V)) \cap T(F + V))$. Then we have that $T(\tilde{F}) \in \text{co}T(E \cap (F + V))$ and thus there exist $F_1, \ldots, F_I \in E \cap (F + V)$ and $\lambda \in \Lambda_I$ such that

$$T(\tilde{F}) = \sum_{i=1}^I \lambda_i T(F_i).$$

By using (ii) we conclude $\tilde{F} \in E \cap (F + V)$.

The case ‘(iii) $\Rightarrow$ (i)’ is trivial.

Similarly to convex analysis we also obtain a relationship between $n$-polyconvex sets and the corresponding version for functions via the use of characteristic functions.
Proposition 4.40. Let \( E \subseteq \mathbb{R}^{d \times D} \) and \( \chi_E \) denote the characteristic function of \( E \):

\[
\chi_E(F) = \begin{cases} 
0, & F \in E \\
+\infty, & \text{otherwise}. 
\end{cases}
\]

Then \( E \) is \( n \)-poly convex if and only if \( \chi_E \) is \( n \)-poly convex.

Remark 4.41. Note that the same is true for convexity. In the case of quasiconvexity this is not clear since very little is known about the quasiconvexity of functions allowed to take the value \(+\infty\).

The proof of the proposition is straightforward. We also obtain the expected implications for \( n \)-poly convexity:

Theorem 4.42. Let \( E \subseteq \mathbb{R}^{d \times D} \). Then

(i) The following implications hold

\[ E \ (d \wedge D)\text{-pc} \Rightarrow E \ (d \wedge D - 1)\text{-pc} \Rightarrow \ldots \Rightarrow E \ 2\text{-pc} \Rightarrow E \ 1\text{-pc}. \]

(ii) \( E \) is poly convex if and only if it is \( (d \wedge D)\)-poly convex.

(iii) \( E \) is rank-one convex if and only if it is \( 1\)-poly convex.

The proof of this theorem is elementary. The definitions in this section followed Dacorogna [20]. They are of an intersectional nature in the sense of standard convexity in which a set is convex if and only if it is the intersection of all halfspaces that contain the set. Similarly, for every \( F \in \mathbb{R}^{d \times D} \) and simple rank-\( n \) subspace the intersection \( E \cap (F + V) \) has to be the intersection of appropriate ‘halfspaces’. The appropriate ‘halfspaces’ here are sets \( \{F \in \mathbb{R}^{d \times D} : \langle \beta, T(F) \rangle \leq \alpha \} \) for some \( \beta \in \mathbb{R}^{r(d,D)} \) and \( \alpha \in \mathbb{R} \) as since \( n \)-poly convexity is based on convexity in the space of minors. For a non-\( n \)-poly convex set we thus call the smallest set that contains it its intersectional \( n \)-poly convex hull. This set-theoretic approach is known for rank-one convexity as the lamination convex hull. However, in Section 4.4.2 we will also consider a different hull which is based on separating points from the hull with finite \( n \)-poly convex functions. Due to the use of functions, and along with Matoušek and Plecháč for directional convexity [40], we use the terminology functional \( n \)-poly convex hull. While a \( n \)-poly convex set with respect to Definition 4.37 coincides with its intersectional \( n \)-poly convex envelope, its functional \( n \)-poly convex envelope may differ. This effect is known for rank-one convexity and it
occurs in particular when the set contains a so-called $T_k$ configuration. To this end the
next section considers $T_k$ configurations as well as a generalisation for $n$-polyconvexity.

### 4.3.1. On $T_k$ configurations

The definition of $n$-polyconvex sets as in Definition 4.37 is of an intersectional nature,
i.e. sets that satisfy this definition coincide with their intersectional $n$-polyconvex hull.
However, it is known that whenever a rank-one convex set in the above sense includes
a $T_k$ configuration, then the functional rank-one convex hull of the set will be strictly
larger. Their relevance have been noted by many authors including Aumann and Hart [3],
Tartar [60] and Scheffer [52]. The effects of their existence on the computation of
semiconvex envelopes or hulls were studied by many, e.g. [40, 39, 31, 33, 32, 22] to name a
few and $T_k$ configurations are an essential part of the analysis. In this section we present
a definition for $T_k$ configurations in rank-one convexity and expand this definition to
so-called $T_k^{pc}$ configuration that allow similar effects to occur in $n$-polyconvexity for
$n > 1$. We provide an example of a $T_k^{pc}$ configuration in $\mathbb{R}^{3\times 3}$ that does not include
$T_k$ configuration in the usual sense. Let us begin with a definition of the usual $T_k$
configuration.

**Definition 4.43.** We say $(A_1, \ldots, A_k) \in (\mathbb{R}^{d \times D})^k$ is a $T_k$ configuration if there exist
$\lambda_1, \ldots, \lambda_k \in (0, 1)$ and $(J_1, \ldots, J_k) \in (\mathbb{R}^{d \times D})^k$ s.t. $\text{rank}(J_i - A_i) = 1$, $i = 1, \ldots, k$
(where $J_0$ means $J_k$) and

$$J_i = \lambda_i J_{i-1} + (1 - \lambda_i) A_i, \quad i = 1, \ldots, k.$$  

Kirchheim et al. [30, Def. 3.9] use a slightly different definition to this, but the only
difference between the two is that the given one focusses on the additional points $J_i$ that
are needed, whereas their one focusses on the rank-one directions it takes to get to them.
The simplest and most famous case is a $T_4$ configuration in the plane.

**Example 4.44.** Let $A_1 = \text{diag}(-1, 2)$, $A_2 = \text{diag}(2, 1)$, $A_3 = \text{diag}(1, -2)$ and $A_4 =
(-2, -1)$. Then $(A_1, \ldots, A_4)$ forms a $T_4$ configuration. The additional points are $J_1 =
(-1, 1), J_2 = (1, 1), J_3 = (1, -1)$ and $J_4 = (-1, -1)$.

A general $T_k$ configuration can be depicted as in Figure 4.7. Note that the configuration
need not be planar. We will show in the following that similar configurations exist that
render the computation of the $n$-polyconvex envelopes difficult. The example which we
Figure 4.6.: The $T_4$ configuration given by $A_1,\ldots,A_4$ with auxiliary points $J_1,\ldots,J_4$ with coordinates being the diagonal entries of the matrices. The black lines indicate the rank-one connections.

Figure 4.7.: A $T_k$ configuration.
present is for the case of 2-polyconvexity in dimension $d, D \geq 3$. In this case we will call the given configuration a $T_k^{2-pc}$ configuration. Again, we will present the example first to stimulate the developing intuition and then give the rigorous definition of $T_k^{np-pc}$ thereafter.

**Example 4.45.** Consider $\mathbb{R}^{3 \times 3}$ and let

$$
\begin{align*}
A_{11} &= \text{diag}(-1, -2, 1/2) & A_{31} &= \text{diag}(1, -2, -1/2) \\
A_{12} &= \text{diag}(-1, -1/2, 2) & A_{32} &= \text{diag}(1, -1/2, -2) \\
A_{21} &= \text{diag}(2, 1/2, 1) & A_{41} &= \text{diag}(-1/2, 2, -1) \\
A_{22} &= \text{diag}(1/2, 2, 1) & A_{42} &= \text{diag}(-2, 1/2, -1)
\end{align*}
$$

Further let $A_i = \{A_{i1}, A_{i2}\}$. Then $(A_1, \ldots, A_4)$ forms a $T_4^{2-pc}$ configuration with the help of the auxiliary points

$$
J_1 = \text{diag}(-1, -1, 1) \quad J_2 = \text{diag}(1, 1, 1) \quad J_3 = \text{diag}(1, -1, -1) \quad J_4 = \text{diag}(-1, 1, -1)
$$

The intuition to the example is the same as for the standard $T_k$ configuration. One can clearly see in Figure 4.8 that the projection onto the $xz$-plane is reminiscent of the usual $T_4$ configuration. Note however, that the above set of points does not contain a $T_4$ configuration. This can easily be checked by closely investigating the involved points. No two points have a rank-one connection and only a few pairs have a rank-two connection. Those can be drawn in the following graph, Figure 4.9. Thus it is sufficient to check whether either $\{A_{11}, A_{12}, A_{31}, A_{32}\}$ or $\{A_{21}, A_{22}, A_{41}, A_{42}\}$ contains a $T_4$ configuration. This is a straightforward task and will be omitted. Therefore, the given set does not contain a standard $T_k$ configuration. Instead, we now need to use 2-polyconvexity rather than rank-one convexity to construct a similar configuration. For this note that $A_{i1}$ and $A_{i2}$ always lie on a connected set of the manifold of all matrices with determinant equal to one. The relevant parts of this manifold are indicated with dashed lines in Figure 4.8. However, none of the matrices on the dashed lines are part of the set. As soon as we start to add the mid-point of one of those dashed lines, take $J_1$ for example, we then inductively gain $J_2$, then $J_3$, $J_4$ and arrive back at $J_1$ through the relation

$$
J_i = \frac{4}{9} A_{i1} + \frac{4}{9} A_{i2} + \frac{1}{9} J_{i-1}
$$

where $J_0 = J_4$. With this we present the formal definition of a $T_k^{np-pc}$ configuration.
Figure 4.8.: A $T_4^{2\text{-pc}}$ configuration. The grey triangles indicate the 2-pc planes. The dashed lines indicate matrices of determinant 1.

Figure 4.9.: Connections of rank two in the set \( \{A_{ij}\}_{i=1,\ldots,4, j=1,2} \).
Definition 4.46. Let $A_i = \{A_{i1}, \ldots, A_{il_i}\} \subseteq \mathbb{R}^{d \times D}$ for $l_i \in \mathbb{N}, i = 1, \ldots, k$ such that there exists a simple rank-$n$ subspace $V \subseteq \mathbb{R}^{d \times D}$ with $A_{ij} \in A_{i1} + V$ for all $j \in \{1, \ldots, l_i\}$. Then $(A_1, \ldots, A_k)$ is called a $T_{k}^{n\text{-pc}}$ configuration if there exist $J_1, \ldots, J_k \subseteq \mathbb{R}^{d \times D}$ such that for all $i \in \{1, \ldots, k\}$ there exist $\lambda^i \in \Lambda^l_{i+1}$ with

$$T(J_i) = \sum_{j=1}^{l_i} \lambda^i_j T(A_{ij}) + \lambda^i_{l_i+1} T(J_{i-1}).$$

Remark 4.47. This definition allows for the addition of just one auxiliary point $J_i$ in each group $A_i$. It is not clear to the author whether this definition should be written to allow for a family of points $\{J_{ij}\}_j$ for each $i$ so that the whole family can be used to obtain a new family $\{J_{(i+1)j}\}_j$.

We will discuss the implications that $T_{k}^{n\text{-pc}}$ configurations have on the functional $n$-polyconvex hull of a set in the next section.

4.4. Further properties of $n$-polyconvexity

4.4.1. $n$-polyconvex envelopes

No convexity notion can ever be fully understood without investigating the convex envelopes it generates. In our case we consider the $n$-polyconvex envelope of a function $f$, i.e. the largest $n$-polyconvex function below $f$.

Definition 4.48. Let $f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\}$. Then we denote the $n$-polyconvex envelope of $f$, i.e. the largest $n$-polyconvex function below $f$, by $P_n f$, i.e. it holds that

$$P_n f(F) = \sup \{g(F) : g \leq f \text{ and } g \text{ n-polyconvex}\}.$$

Denoting by $Cf, Pf, Qf$ and $Rf$ the usual convex, polyconvex, quasiconvex and rank-one convex envelopes respectively, we have that

$$Cf \leq Pf = P_{d\times D} f \leq P_{d\times D-1} f \leq \ldots \leq P_n f \leq \ldots \leq P_2 f \leq P_1 f = Rf \leq f$$

and a natural question is whether we can fit the quasiconvex envelope $Qf$ into this picture. However, we will be mostly unable to answer this question and instead focus on proving...
the more obvious relations. It is well known that the convex and polyconvex envelope can be represented by a finite number of convex combinations. In the polyconvex case this result reads as follows (cf. [20, Thm. 6.8]). For \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\} \) and bounded below by a polyconvex function we have

\[
Pf(F) = \inf \left\{ \sum_{i=1}^{\tau(d,D)+1} \lambda_i f(F_i) : \lambda \in \Lambda_{\tau(d,D)+1}, F_i \in \mathbb{R}^{d \times D}, T(F) = \sum_{i=1}^{\tau(d,D)+1} \lambda_i T(F_i) \right\}.
\]

Similarly, but with a subtle difference Dacorogna proves an analogous result for the rank-one convex envelope [20, Thm. 6.10]. Given \( f \) as above and bounded below by a rank-one convex function it holds that

\[
Rf(F) = \inf \left\{ \sum_{i=1}^{I} \lambda_i f(F_i) : \lambda \in \Lambda_I, F_i \in \mathbb{R}^{d \times D}, F = \sum_{i=1}^{I} \lambda_i F_i, (\lambda_i, F_i)_{1 \leq i \leq I} \text{ satisfy } (H_I) \right\}.
\]

Note that \( I \) can be arbitrarily large whereas in the polyconvex case just \( \tau(d,D) + 1 \) elements in the convex combination were sufficient. Alternatively, we can take \( R_0 f = f \) and then inductively define \( R_k \) by

\[
R_{k+1} f(F) = \inf \{ \lambda R_k f(F_1) + (1 - \lambda) R_k (F_2) : \lambda \in [0,1], \text{ rank}(F_1 - F_2) \leq 1, F = \lambda F_1 + (1 - \lambda) F_2 \}.
\]

Then it also holds

\[
Rf = \lim_{k \to \infty} R_k f = \inf_{n \in \mathbb{N}} R_k f.
\]

The next theorem will show that the \( n \)-polyconvex envelope displays very much the same behaviour as the rank-one convex envelope.

**Theorem 4.49.** Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) and be bounded below by a \( n \)-polyconvex function.
(i) The following holds

\[ P_n f(F) = \inf \left\{ \sum_{i=1}^I \lambda_i f(F_i) : \lambda \in \Lambda_I, F_i \in \mathbb{R}^{d \times D}, \right\} \]

\[ F = \sum_{i=1}^I \lambda_i F_i, (\lambda_i, F_i)_{1 \leq i \leq I} \text{ satisfy } (H^n_I). \]

(ii) Let \( P_n^0 f = f \) and for \( k \in \mathbb{N} \)

\[ P^{k+1}_n f(F) = \inf \left\{ \sum_{i=1}^{\tau(n)+1} \lambda_i P^n_k f(F_i) : \lambda \in \Lambda_{\tau(n)+1}, \right\} \]

\[ V \text{ simple rank-} n, F_i \in F + V, T(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) \right\}. \] (4.54)

Then \( P_n f = \lim_{k \to \infty} P^n_k f = \inf_{k \in \mathbb{N}} P^n_k f \).

Similar to Theorem 4.13 the proof is absolutely analogous and will be omitted.

4.4.2. \( n \)-polyconvex hulls and the envelope of the distance function

The first definitions and results in this section will consider hulls from an intersectional approach, i.e. by considering points from a given set and adding all those points that can be obtained by relevant convex combinations. In the latter part of this section we will investigate different kind of hulls arising from a separational approach, i.e. all points that cannot be separated from the set by finite \( n \)-polyconvex functions. Following \[40\] for \( D \)-convexity we will call these hulls functional \( n \)-polyconvex hulls.

Let us begin with the hulls from an intersectional point of view.

**Definition 4.50.** Let \( K \subseteq \mathbb{R}^{d \times D} \). Then we denote by \( P^n_n K \) the smallest \( n \)-polyconvex set that contains \( K \) and refer it as the intersectional \( n \)-polyconvex hull of \( K \).

Then the following theorem corresponds to its counterpart for envelopes, Theorem 4.49.

**Theorem 4.51.** Let \( K \subseteq \mathbb{R}^{d \times D} \). Let \( P^n_0 K := K \) and define inductively

\[ P^{k+1}_n K = \{ F \in \mathbb{R}^{d \times D} : \exists V \subseteq \mathbb{R}^{d \times D} \text{ simple rank-} n, F_i \in P^n_k K \cap (F + V), \]

\[ i = 1, \ldots, \tau(n) + 1, \lambda \in \Lambda_{\tau(n)+1}, T(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) \} \].
Then $P_n^\cap K = \bigcup_{k \in \mathbb{N}} P^k_n K$. Moreover, if $K$ is open then $P_n^\cap K$ is open.

**Remark 4.52.**  
(i) For the intersectional rank-one convex hull (which corresponds to $P^1_\cap K$) this kind of hull is also often referred to as the lamination convex hull.

(ii) For $n = d \wedge D$, i.e the polyconvex case, we find that $P^1_{d \wedge D} K = P^\cap_n K$ and so no lamination is necessary.

The proof of the above theorem is analogous to the envelope version and again, we omit most of it here. However, since the proof of the openness of the rank-one convex hull of an open set is only hinted at, we will discuss this part in more detail.

**Proof of openness of $P_\cap_n K$ for open $K$ in Theorem 4.51.**  
Let $K$ be open. We simply prove that then $P^1_\cap_n K$ must be open and the whole statement follows by induction. Let $F \in P^1_\cap_n K$. Then there exist a simple rank-$n$ subspace $V$ and matrices $F_i, i = 1, \ldots, k$, $k \leq \tau(n) + 1$, belonging to $K \cap (F + V)$, and $\lambda \in \Lambda_k$ such that $T(F) = \sum_{i=1}^{k} \lambda_i T(F_i)$.

Since $K$ is open and $k$ finite there exists $\varepsilon > 0$ such that $B_\varepsilon(F_i) \subseteq K$ for all $i = 1, \ldots, k$.

We claim that then $B_\varepsilon(F) \subseteq P^1_\cap_n K$, proving that $P^1_\cap_n K$ is indeed open. Let $\widetilde{F} \in B_\varepsilon(F)$. Then define $\widetilde{F}_i = F_i + \widetilde{F} - F \in K$. From Lemma A.8 of Appendix A.4 we have that also $T(\widetilde{F}) = \sum_{i=1}^{k} \lambda_i T(\widetilde{F}_i)$ and so $\widetilde{F} \in P^1_n K$.

The proof for the rank-one convex case can be found after Theorem 7.17 in [20]. The same applies to the following, cf. [20, Prop. 7.22].

**Proposition 4.53.** Let $K \subseteq \mathbb{R}^{d \times D}$ and $\chi_K$ its characteristic function. Then

$$P_n \chi_K = \chi_{P^\cap_n K}$$

where $P_n \chi_K$ is the $n$-polyconvex envelope of the characteristic function $\chi_K$ as defined in Definition 4.48.

The proof of this proposition is simple and in fact it holds that $P_n^k \chi_K = \chi_{P^k_n K}$ for each $k \in \mathbb{N}$, which can be proved by induction. Next we observe that the preservation of the convexity properties of a set is maintained for the interior of the set also in the case of $n$-polyconvexity. However, as to be expected, this is not true for the closure.

**Proposition 4.54.**  
(i) Let $K \subseteq \mathbb{R}^{d \times D}$ be $n$-polyconvex. Then int $K$ is $n$-polyconvex.

(ii) There exists $K \subseteq \mathbb{R}^{2 \times 2}$ such that $K$ is polyconvex (and hence $n$-polyconvex for any $n$), but $\overline{K}$ is not separately convex (and hence not $n$-polyconvex for any $n$).
The proof of ‘(i)’ is analogous to the proof of [20, Prop. 7.24] and ‘(ii)’ was only given for completeness. The particular example can be found in the same reference.

We now consider the functional $n$-polyconvex hulls of a set. They are defined as follows.

**Definition 4.55.** Let $K \subseteq \mathbb{R}^{d \times D}$. Then we denote by $P_n^f K$ all points that cannot be separated from $K$ by a finite $n$-polyconvex function, i.e.

$$P_n^f K = \{ F \in \mathbb{R}^{d \times D} : f(F) \leq 0 \text{ for all } f : \mathbb{R}^{d \times D} \to \mathbb{R} \text{ $n$-polyconvex with } f \leq 0 \text{ on } K \}.$$  

We call $P_n^f K$ the functional $n$-polyconvex hull of $K$.

It is clear that we have the usual cascade of hulls, i.e. $R^f K = P_1^f K \subseteq P_2^f K \subseteq \ldots \subseteq P_{d \wedge D - 1}^f K \subseteq P_{d \wedge D}^f K = P^f K$, where $P^f K$ and $R^f K$ denote the usual functional polyconvex and rank-one convex hull of $K$. The following theorem asserts that the functional $n$-polyconvex hull is in general a stronger notion than the intersectional $n$-polyconvex hull (particularly for open sets).

**Theorem 4.56.** Let $K \subseteq \mathbb{R}^{d \times D}$. Then

$$P_n^f K \subseteq P_n^f K.$$  

The proof of this theorem is trivial since the functional $n$-polyconvex hulls are all closed, $n$-polyconvex sets. Thus, by definition $P_n^f K \subseteq P_n^f K$. Furthermore, we suspect that the two hulls are inherently different in many cases. For example, we believe that there exist compact $n$-polyconvex sets (i.e. their intersectional $n$-polyconvex hull coincides with the set) which have a much larger functional $n$-polyconvex hull. These cases occur whenever there is a $n$-polyconvex set of finitely many points that form a $T_k^{n\text{-pc}}$ configuration. Then at least the auxiliary points of the configuration must be included in the functional $n$-polyconvex hull. We conjecture the following.

**Conjecture 4.57.** For all $1 \leq n < d \wedge D$ there exists a set $K$ of finitely many points such that $K$ is intersectionally $n$-polyconvex, but not functionally $n$-polyconvex.

The case $n = 1$ is dealt with by the well known $T_4$ configuration in the plane. Considering Example 4.45 we have proved that this is true for $n = 2$. We think that $T_k^{n\text{-pc}}$ configurations also provide examples for $n > 2$.

Whenever a $T_k^{n\text{-pc}}$ configuration exists in a set its functional $n$-polyconvex hull will be nontrivial. Thus, knowing whether such a configuration exists is helpful for computing...
the $n$-polyconvex hulls of sets. For rank-one convexity and finite sets Kreiner et al. [33] provide a way of testing for $T_k$ configurations. Conversely, as proved in [59, Thm. 1], for a compact set in $\mathbb{R}^{2 \times 2}$ without any rank-one connections a nontrivial rank-one convex hull implies the existence of a $T_4$ configuration.

In general, for poly-, quasi- and rank-one convexity there exists a relationship between the functional semiconvex hull and the zero set of the respective semiconvex envelope of the distance function. The distance function $\text{dist}_K$ to an arbitrary set $K \subseteq \mathbb{R}^{d \times D}$ is defined as

$$\text{dist}_K(F) = \inf_{\tilde{F} \in K} ||F - \tilde{F}||.$$ 

For quasi- and rank-one convexity the following theorem applies.

**Theorem 4.58.** Let $K \subseteq \mathbb{R}^{d \times D}$ be a compact set. Then the functional $\ast$-convex hull of $K$ is given by the zero set of the $\ast$-convex envelope of the distance function, i.e.

$$K^\ast = \{ F \in \mathbb{R}^{d \times D} : \text{dist}_K^\ast(F) = 0 \},$$

where $\ast \in \{qc, rc\}$.

The case for quasiconvexity was proved by Zhang [62] and for rank-one convexity (in fact in greater generality for directional convexity) by Matoušek [39]. The case for polyconvexity is not as straightforward. Here the zero set depends on the power $p$ of the $p$-distance function $\text{dist}^p_K$, whereas for quasi-, cf. [61] [63], and rank-one convexity this is not the case. Šilhavý [53] shows that only for $p \geq d \wedge D$ does the same apply for the polyconvex hull of a compact set or when $p = 1$ one obtains the convex hull of the set instead. In fact, for any integer $1 \leq p \leq d \wedge D$ Šilhavý defines the so-called $s$-polyconvex hull for the notions of $s$-polyconvexity, which was first defined in [12]. Note that $s$-polyconvexity is a concept that unifies standard convexity and polyconvexity in the sense that for $s = 1$ it is equivalent to convexity and for $s = d \wedge D$ it is equivalent to polyconvexity and should not be confused with $n$-polyconvexity as defined here. A function $f : \mathbb{R}^{d \times D} \rightarrow \mathbb{R}$ is $s$-polyconvex if there exists a convex function $g : \mathbb{R}^{\tau(s)} \rightarrow \mathbb{R}$ such that $f(F) = g(\text{adj}_1 F, \ldots, \text{adj}_s F)$, where $\tau(s) = \sum_{i=1}^s \binom{D}{i}$. The functional $s$-polyconvex hull $K^{(s)\text{-pc}}$ of a compact set $K$ and the $s$-polyconvex envelope $f^{(s)\text{-pc}}$ of a function $f$ is defined as usual (the parenthesis $(s)$-pc is used to distinguish from the $n$-polyconvex case $n$-pc).

**Theorem 4.59** (Thm. 1.1, [53]). Let $K \subseteq \mathbb{R}^{d \times D}$ be compact. Further let $s$ be an integer with $1 \leq s \leq d \wedge D$. Then the $s$-polyconvex hull $K^{(s)\text{-pc}}$ of $K$ is given by the zero set of
the $s$-polyconvex envelope $(\text{dist}_K^p)^{(s)\text{-pc}}$ of the $p$-distance function of $K$, i.e.

$$K^{(s)\text{-pc}} = \{ F \in \mathbb{R}^{d \times D} : (\text{dist}_K^p)^{(s)\text{-pc}}(F) = 0 \},$$

for $p \in [s, s+1)$ or $d \wedge D \leq p < \infty$ when $s = d \wedge D$.

Note that the case $s = d \wedge D$ produces the polyconvex hull. Rephrasing the above results for poly- and rank-one convexity into the language of $n$-polyconvexity we have that the functional $n$-polyconvex hull $K^{n\text{-pc}}$ is given by the zero set of the $n$-polyconvex envelope of $\text{dist}^n_K$ for $n = d \wedge D$ or $n = 1$ respectively. Thus it seems reasonable to conjecture the following:

**Conjecture 4.60.** Let $K \subseteq \mathbb{R}^{d \times D}$ be compact. Then the functional $n$-polyconvex hull $K^{n\text{-pc}}$ of the set $K$ is given by the zero set of $n$-polyconvex envelope of the $n$-th power of the distance function, i.e.

$$P^n_f K = \{ F \in \mathbb{R}^{d \times D} : P^n_f(\text{dist}^n_K)(F) = 0 \}.$$

We were not successful in proving the claim as we encountered the similar difficulties as in Section 4.2.4 for Conjecture 4.33. One such difficulty is that we do not have a global function $g$ that is convex on $T(F + V)$ for all $F \in \mathbb{R}^{d \times D}$ and $V \subseteq \mathbb{R}^{d \times D}$ simple rank-$n$ that represents $f$ in the usual way $f = g \circ T$ on the whole of $\mathbb{R}^{d \times D}$. Having access to such a global representative is by no means a guarantee to success, but it is for instance among the essential ingredients of the proof by Šilhavý for $s$-polyconvexity. On the other hand Matoušek’s proof for directional convexity (applicable for $n = 1$) does not carry over for $n > 1$. The main obstacle here is the nonlocality of $n$-polyconvexity for $n > 1$.

In almost all cases the respective relaxation of the distance function (or any other function) cannot be found by analytical means. Instead numerical approaches are necessary. In the case of rank-one convexity Dolzmann [22] proved the convergence of an iterative convexification along rank-one lines to the rank-one convex envelope of a function when the mesh size approaches zero. A necessary requirement to allow to work on a finite set of points is that the respective function has to agree with its rank-one convex envelope outside a fixed ball in the domain. We will return to a slight modification of the proposed algorithm in Section 5.3.3 after the considerations in Chapter 5.
4.4.3. Relations to quasiconvexity

Quasiconvexity is extremely important for the calculus of variations as it has been shown by Morrey [41] to be necessary and sufficient (with some additional assumptions) for the weak lower semicontinuity of multiple integrals. However it is still not well understood and apart from polyconvexity and rank-one convexity as stronger and weaker notions respectively there are not many more tools to establish whether a function is quasiconvex or not. Here we want to investigate whether the new concepts of $n$-polyconvexity could provide such tools.

It is clear that we have the following implications:

$$
(d \wedge D - 1)\text{-pc} \implies \cdots \implies 2\text{-pc}
$$

$$
\text{pc} = (d \wedge D)\text{-pc} \implies \text{qc} \implies \text{rc} = 1\text{-pc}
$$

where the one ‘quasiconvexity implies rank-one convexity’ only remains true in the finite-valued case. Another figure to illustrate how quasiconvexity could fit into the $n$-polyconvexity framework is Figure 4.10. It is clear that the set of polyconvex functions

![Figure 4.10.](image)

Figure 4.10.: Possible arrangements of the set of quasiconvex functions in the various sets $n$-polyconvex functions

is contained in the set of quasiconvex functions, and that this in turn is contained in the set of rank-one convex functions. It is not clear, however, whether or not it includes or is included in any set of $n$-polyconvex functions for $n = 2, \ldots, d \wedge D - 1$. There are more
configurations possible than depicted in the figure, but the given ones seem to be the most plausible.

The following theorem will show that quasiconvexity does not imply 2-polyconvexity in any dimension, which already excludes many of the possible cases for a relationship between quasiconvexity and \( n \)-polyconvexity.

**Theorem 4.61.** Let \( d, D \geq 2 \). Then there exists \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) such that \( f \) is quasiconvex but not 2-polyconvex.

**Proof.** First we consider the case for square matrices, i.e. \( d = D \geq 2 \). The proof uses Šverák’s examples of quasiconvex functions that are also used to show that the rank-one convex hull of the standard \( T_4 \) configuration of Example 4.44 equals its quasiconvex hull [65]. For \( l \in \mathbb{N} \) we define

\[
f_l(F) = f_l^{\text{sym}} \left( \frac{F + F^T}{2} \right)
\]

where

\[
f_l^{\text{sym}}(F) = \begin{cases} 
|\det F|, & \# \text{neg. eigenvalues of } F = l \\
0, & \text{otherwise.}
\end{cases}
\]

According to Šverák the \( f_l \) defined in such a way are quasiconvex for all \( l \in \mathbb{N} \). Consider the function \( f_0 \) and take the points

\[
F_1 = \text{diag}(-1, 1, 1, \ldots, 1) \\
F_2 = \text{diag}(1, 0, 1, \ldots, 1) \\
F_3 = \text{diag}(0, -2, 1, \ldots, 1) \\
F_4 = \text{diag}(-2, -1, 1, \ldots, 1)
\]

Note that \( F_1, \ldots, F_4 \) form the usual 2D Tartar-Scheffer configuration embedded in \( d \)-dimensional space. Due to the diagonal form of the \( F_i \) with all but the first two values being equal to one we find that \( T(F_i) \) is very simple. It consists solely of the values of \( F_i \),
its determinant \( \det F \) or zero. We have,

\[
\begin{align*}
\det F_1 &= -1 \\
\det F_2 &= 0 \\
\det F_3 &= 0 \\
\det F_4 &= 2.
\end{align*}
\]

Now let \( F = \text{diag}(\frac{1}{10}, \frac{1}{10}, 1, \ldots, 1) \). Then it is easy to verify that

\[
T(F) = \frac{1}{4}T(F_1) + \frac{61}{100}T(F_2) + \frac{1}{100}T(F_3) + \frac{13}{100}T(F_4).
\]

(4.56)

Since only \( F \) and \( F_2 \) have no negative eigenvalues we obtain

\[
f_0(F) = \frac{1}{100} > 0 = \frac{61}{100}f_0(F_2) = \frac{1}{4}f_0(F_1) + \frac{61}{100}f_0(F_2) + \frac{1}{100}f_0(F_3) + \frac{13}{100}f_0(F_4).
\]

However, this implies that \( f_0 \) cannot be 2-polyconvex, since \( F_1, \ldots, F_4 \) are elements of \( I + V \) for \( V = \text{span}\{e_1 \otimes e_1, e_2 \otimes e_2\} \) and satisfy (4.56). If \( f \) was 2-polyconvex, then the opposite inequality must hold by (4.7) of Theorem 4.8.

For the general case of non-square matrices we can simply use the projection onto the first \( d \wedge D \) rows and columns and proceed as above, proving the claim in general.

Remark 4.62. Using the same function it also possible to show that the functional 2-polyconvex hull of the \( T_4^{2pc} \) configuration presented in Example 4.45 differs from its functional 1-polyconvex hull. It is clear that the functional 2-polyconvex hull must at least contain the points \( J_i, i = 1, \ldots, 4 \). However, the point \( J_3 \) can be separated from the set \( \{A_{ij} | i = 1, \ldots, 4, j = 1, 2\} \) by using \( f_0 : \mathbb{R}^{2 \times 2} \to \mathbb{R} \) defined via (4.55). Since \( f_0 \) is quasiconvex in \( \mathbb{R}^{2 \times 2} \) it follows that \( g : \mathbb{R}^{3 \times 3} \to \mathbb{R}, g(F) = f_0(M_{33}(F)), \) where \( M_{33}(F) \) denotes the matrix obtained from \( F \) by deleting the third row and column, is also quasiconvex. Finally, the function \( h : \mathbb{R}^{3 \times 3} \to \mathbb{R} \) with \( h(F) = g(F - \text{diag}(1/2, 1/2, 1)) \) is also quasiconvex and has the property that \( h(A_{ij}) = 0 \) for all \( i = 1, \ldots, 4, j = 1, 2, \) but \( h(J_2) = 1/4 > 0 \).

We were not successful in excluding or asserting any other possible relation between quasiconvexity and \( n \)-polyconvexity. For example, it would be interesting to find a finite 2-polyconvex function that is not quasiconvex. However, even finding a finite 2-polyconvex function that is not polyconvex is a nontrivial task and the motivation of the next section.
4.4.4. The quadratic case

We now treat the quadratic case as it is considered in [20] with respect to $n$-polyconvexity. In particular, in the 3-dimensional setting we hope to find a finite-valued example of a function that is 1-polyconvex, but not 2-polyconvex, and a function that is 2-polyconvex, but not 3-polyconvex. In order to do so, we investigate what the equivalent conditions are for $n$-polyconvexity in the case of a quadratic function. For completeness, we also include the cases of convexity and quasiconvexity, cf. [20, Lemma 5.27].

**Theorem 4.63** (cf. Lm. 5.27, [20]). Let $M$ be a symmetric matrix in $\mathbb{R}^{(d\times D)\times(d\times D)}$ and let

$$f(F) = \langle MF, F \rangle.$$ 

Then the following holds

(i) $f$ is convex if and only if

$$f(F) \geq 0$$

for every $F \in \mathbb{R}^{d\times D}$.

(ii) $f$ is quasiconvex if and only if

$$\int_{\Omega} f(\nabla\varphi(x)) \, dx \geq 0$$

for every bounded open set $\Omega \subseteq \mathbb{R}^d$ and for every $\varphi \in W_{0}^{1,\infty}(D; \mathbb{R}^D)$.

(iii) $f$ is $n$-polyconvex if and only if there exists $\alpha \in \mathbb{R}^{\sigma(2)}$ such that

$$f(F) \geq \langle \alpha, \text{adj}_2 F \rangle$$  \hspace{1cm} (4.57)

for all $F \in \mathbb{R}^{d\times D}$ such that $\text{rank}(F) \leq n$.

**Remark 4.64.** Note that in the original lemma the cases of polyconvexity and rank-one convexity are treated separately. Here they both fall under the case [iii], where polyconvexity directly takes the expected form for $n = d \wedge D$. For rank-one convexity, i.e. $n = 1$, eq. (4.57) then simplifies to the condition found in the original lemma, namely $f(a \otimes b) \geq 0$ for all $a, b \in \mathbb{R}^d$, since for rank-one matrices $F$ we have $\text{adj}_2 F = 0$.  

129
We now only prove part (iii) of Theorem 4.63 since the other proofs can be found in [20].

**Proof of Th. 4.63 (iii).** First we show that similarly to all other convexities also $n$-pc convexity can simply be considered around the point $F = 0$. Recall that by Theorem 4.33 a finite valued function $f$ is $n$-polyconvex if and only if for all $F \in \mathbb{R}^{d \times D}$ there exists $\beta(F) \in \mathbb{R}^{\tau_{(d,D)}}$ such that

$$f(\tilde{F}) - f(F) \geq \langle \beta(F), T(\tilde{F}) - T(F) \rangle$$  \hspace{1cm} (4.58)

for all $\tilde{F} \in C^n(F)$, i.e. rank($\tilde{F} - F$) $\leq n$. We want to show that for quadratic $f$ this condition simplifies to

$$f(\tilde{F}) \geq \langle \beta, T(\tilde{F}) \rangle$$  \hspace{1cm} (4.59)

for some $\beta \in \mathbb{R}^{\tau_{(d,D)}}$ and all $\tilde{F} \in C^n(0)$, i.e. rank($\tilde{F}$) $\leq n$. Then we show that this in fact is equivalent to

$$f(\tilde{F}) \geq \langle \beta_2, \text{adj}_2 \tilde{F} \rangle$$  \hspace{1cm} (4.60)

for some $\beta_2 \in \mathbb{R}^{\sigma(2)}$ and all $\tilde{F} \in C^n(0)$, i.e. rank($\tilde{F}$) $\leq n$.

For the first step it is clear that (4.58) implies (4.59) by taking $F = 0$. Therefore, assume now that (4.59) holds for some $\beta \in \mathbb{R}^{\tau_{(d,D)}}$. Then fix $F \in \mathbb{R}^{d \times D}$ and consider $f(\tilde{F}) - f(F)$ for all $F \in C^n(F)$. We have, using the quadratic form of $f$ and symmetry of $M$, that

$$f(\tilde{F}) - f(F) = f(F + (\tilde{F} - F)) - f(F) = 2 \langle MF, \tilde{F} - F \rangle + f(\tilde{F} - F).$$

Applying (4.59) we then obtain

$$f(\tilde{F}) - f(F) \geq 2 \langle MF, \tilde{F} - F \rangle + \langle \beta, T(\tilde{F} - F) \rangle.$$  

In Lemma A.9 we show that for fixed $F \in \mathbb{R}^{d \times D}$ and $\beta \in \mathbb{R}^{\tau_{(d,D)}}$ there exists $\tilde{\beta} \in \mathbb{R}^{\tau_{(d,D)}}$ such that

$$\langle \beta, T(\tilde{F} - F) \rangle = \langle \tilde{\beta}, T(\tilde{F}) - T(F) \rangle$$  \hspace{1cm} (4.61)

for all $\tilde{F} \in \mathbb{R}^{d \times D}$. Inserting this into the last inequality we have with $\hat{\beta} = \tilde{\beta} +$
(2MF, 0, . . . , 0) that
\[ f(\tilde{F}) - f(F) \geq \langle \tilde{\beta}, T(\tilde{F}) - T(F) \rangle \]
and so in fact (4.59) implies (4.58).

It remains to show the equivalence of (4.59) and (4.60) when \( f \) is quadratic. The
direction (4.60) implies (4.59) is trivial. Thus, assume (4.59) holds. Then for \( \beta = (\beta_1, \ldots, \beta_n) \), \( \beta_i \in \mathbb{R}^{\sigma(i)} \) and taking \( \tilde{F} = \varepsilon F \) with \( F \in C^n(0) \) and \( \varepsilon > 0 \) we have
\[
\varepsilon^2 f(F) \geq \varepsilon \langle \beta_1, F \rangle + \varepsilon^2 \langle \beta_2, \text{adj}_2 F \rangle + O(\varepsilon^3).
\] (4.62)
Dividing this inequality by \( \varepsilon \) and letting \( \varepsilon \to 0 \) we find \( 0 \geq \langle \beta_1, F \rangle \) for all \( F \in C^n(F) \).
Thus, taking \( F = \pm e_i \otimes e_j \) this implies \( \beta_1 = 0 \). Returning to (4.62) with \( \beta_1 = 0 \) and dividing by \( \varepsilon^2 \) we find for \( \varepsilon \to 0 \)
\[
f(F) \geq \langle \beta_2, \text{adj}_2 F \rangle,
\]
which is the desired inequality. \( \Box \)

It is known, cf. [20, Thm. 5.25] that in the quadratic case rank-one convexity and
quasiconvexity are equivalent. Similarly, for \( d = D = 2 \) rank-one convexity and poly-
convexity are equivalent, but for \( d, D \geq 3 \) it is known that rank-one convexity does not
imply polyconvexity. One such counterexample for a function that is rank-one convex,
but not polyconvex is presented by Serre, cf. [6] and references therein. In fact, his
example can be taken to show that in general rank-one convexity does not imply (strong)
2-polyconvexity for \( d, D \geq 3 \), which is summarised in the following theorem.

**Theorem 4.65.** Let \( d, D \geq 3 \). Then there exists a quadratic function \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \)
such that \( f \) is rank-one convex but not 2-polyconvex.

**Proof.** We consider only the \( 3 \times 3 \) case, knowing that it is possible to embed it into the
higher dimensional cases. As mentioned before, the particular counterexample is that of
Serre, see Ball [6]. His function \( f \) is defined as
\[
f(F) = (F_{11} - F_{32} - F_{23})^2 + (F_{12} - F_{31} - F_{13})^2 + (F_{21} - F_{31} - F_{13})^2 + F_{22}^2 + F_{33}^2.
\] (4.63)
Note, that \( f \) is nonnegative (and therefore trivially also polyconvex) and that the zero
set of $f$ forms the linear space
\[
V = \left\{ \begin{pmatrix} b+d & c-a & a \\ c+a & 0 & b \\ c & d & 0 \end{pmatrix} : a, b, c, d \in \mathbb{R} \right\}.
\]

This space has the crucial property that it contains no matrices of rank one, or in other words, all nonzero matrices of $V$ have either rank two or three. This can be seen easily by computing the image of $V$ under the cofactor map. We have
\[
\text{cof}(V) = \left\{ \begin{pmatrix} -bd & bc & d(c+a) \\ ad & -ac & c(c-a)-d(b+d) \\ b(c-a) & a(a+c)-b(b+d) & a^2-c^2 \end{pmatrix} : a, b, c, d \in \mathbb{R} \right\}.
\]

If $V$ contained a matrix of rank one, then cofactor image of $V$ must contain the zero matrix for nontrivial choices of $a, b, c, d \in \mathbb{R}$. So assume for a start that $b \neq 0$. Then this implies that $c = d = 0$ from the first two entries. This in turn implies by the \{3, 3\}-entry that $a = 0$ from which $b = 0$ follows at the \{3, 2\}-entry. Therefore we must have $b = 0$. Now assume $a \neq 0$. Then we must have $c = d = 0$ by the second row and then once more $a = 0$ from the \{3, 3\}-entry. Thus $a = b = 0$. Yet again, from \{3, 3\}, we obtain $c = 0$ and then finally from \{2, 3\} that also $d = 0$. Thus, the preimage of the zero matrix of the cofactor map is only the zero matrix itself.

We now want to modify $f$ in such a way that destroys its polyconvexity (and in doing so we will incidentally also remove its 2-polyconvexity). Knowing that $V$ does not contain any matrices of rank one it is then easy to conclude that
\[
\varepsilon := \inf \{ f(a \otimes b) : a \in \mathbb{R}^d, b \in \mathbb{R}^D, |a \otimes b| = 1 \} > 0.
\]

Then, the function $g$ defined as
\[
g(F) = f(F) - \varepsilon |F|^2
\]
is still rank-one convex, since $g(a \otimes b) = f(a \otimes b) - \varepsilon |a \otimes b|^2 \geq 0$. However, it is now no longer 2-polyconvex in the sense that there exists no $\alpha \in \mathbb{R}^{\sigma(2)}$ such that $g(F) - \langle \alpha, \text{adj}_2(F) \rangle \geq 0$.  

132
If there was such an $\varepsilon$, then by taking $F \in V$ with $b = 1$ and $a = c = d = 0$ we must have

$$-2\varepsilon - \left\langle \alpha, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \right\rangle \geq 0,$$

i.e.

$$\alpha_{32} \geq 2\varepsilon > 0.$$

Similarly, by setting each of the variables $a, c, d$ to one and the others to zero we obtain the three additional inequalities

$$-\alpha_{32} - \alpha_{33} \geq 3\varepsilon > 0$$
$$-\alpha_{23} + \alpha_{33} \geq 3\varepsilon > 0$$
$$\alpha_{23} \geq 2\varepsilon > 0.$$

Adding the second and third equation we find that it must also hold that $-\alpha_{32} - \alpha_{23} > 0$, however, this is clearly impossible due to the first and last inequality.

Having established that rank-one convexity does not imply 2-polyconvexity when $d, D \geq 3$ in the quadratic case it is a natural question to ask whether $n$-polyconvexity implies $(n + 1)$-polyconvexity for $n > 1$. For example, we wonder whether in $\mathbb{R}^{3 \times 3}$ there exists a strongly 2-polyconvex and quadratic function that is not polyconvex. It is tempting to try to construct such a function in a similar way as Serre’s function (4.63). If we could find a linear space $V \subseteq \mathbb{R}^{3 \times 3}$ of matrices that contains no nonzero matrix with rank less than three, we could then define a nonnegative and quadratic function $f$ with $V$ as its zero set. Upon defining $\varepsilon = \inf f(F) : \text{rank}(F) \leq 2, |F| = 1$ we then again have $\varepsilon > 0$ (since $f$ strictly positive on $C^2(0)$ as the zero set of $f$ would only contain matrices of rank three). It would then remain to investigate whether $g(F) = f(F) - \varepsilon |F|^2$ and for all $\alpha \in \mathbb{R}^{\sigma(2)}$ there exists $F_\alpha$ such that $g(F) - \langle \alpha, \text{adj}_2 F \rangle < 0$ and good candidates would be $F_\alpha \in V$. However, this endeavour is doomed to fail from the very start since there exists no such $V$. If there was such a $V$ we must have $\det(F) \neq 0$ for all $F \in V$, $F \neq 0$. Representing $F = \lambda_1 F_1 + \ldots + \lambda_k F_k$ where $F_i$ form a basis of $V$ we have that $\det(F)$ is a polynomial of degree three for each $\lambda_i$. Therefore choosing $\lambda_2 \neq 0$ and $\lambda_3 = \ldots = \lambda_k = 0$ we can still find $\lambda_1 \in \mathbb{R}$ such that $\det(F) = 0$.

Although it is not possible to use this construction of a 2-polyconvex but not 3-polyconvex function in $\mathbb{R}^{3 \times 3}$ there may still exist a different way of finding such a
function. Keeping the method of construction it may however be possible to construct examples of strongly $n$-polyconvex but not strongly $(n+1)$-polyconvex functions for some $n > 1$ in $\mathbb{R}^{d \times D}$ with $d, D > 3$. The essence is to find a subspace of $\mathbb{R}^{d \times D}$ of which each element has at least rank $n+1$. This and other questions have been addressed by several authors, cf. [23] and [15]. In the latter reference we can find a subspace of matrices of $\mathbb{R}^{4 \times 4}$ that has constant rank 3. The space $V$ is the span of the following matrices

\[
\begin{pmatrix}
0 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0
\end{pmatrix},
\begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0
\end{pmatrix},
\begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 \\
0 & 1 & 1 & 1
\end{pmatrix},
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0
\end{pmatrix},
\begin{pmatrix}
0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 \\
1 & 1 & 0 & 0
\end{pmatrix}.
\]

It is easy to define a nonnegative quadratic function $f$ such that $f(F) = 0$ if and only if $F \in V$. Again this $f$ is polyconvex, but we would hope to modify it such that it remains 2-polyconvex, but not 3-polyconvex. To this end we proceed as above and define $\varepsilon = \inf\{f(F) : \text{rank}(F) \leq 2 \text{ and } |F| = 1\}$. Then $g$ defined through $g(F) = f(F) - \varepsilon |F|^2$ is quadratic and 2-polyconvex. However, at the present time we were unable to conclude whether $g$ is 3-polyconvex or even polyconvex. The easiest path to successfully show that no $\alpha \in \mathbb{R}^{6 \times 6}$ exists such that

\[
g(F) - \langle \alpha, \text{adj}_2 F \rangle \geq 0 \quad (4.64)
\]

for all $F \in \mathbb{R}^{4 \times 4}$ would be to hope that for any $\alpha$ we can find $F \in V$ such that (4.64) does not hold. Unfortunately the first five inequalities that $\alpha$ needs to satisfy by choosing each of the basis vectors above do not yield a contradiction like in $\mathbb{R}^{3 \times 3}$. It may however still be possible to show that a system of inequalities obtained by including more than just the five inequalities, but also other combinations of the five matrices does not allow a feasible solution. We did not pursue the matter further and it may well be that either it is not sufficient to search for counterexamples in $V$ but instead in the whole of $\mathbb{R}^{4 \times 4}$ or that indeed $g$ is actually 3-polyconvex or even polyconvex.

### 4.4.5. On the nonlocality of $n$-polyconvexity

Since the works of Kristensen [35, 34] it is known that both polyconvexity and quasiconvexity are nonlocal concepts whereas rank-one convexity is equivalent to the positive definiteness of the second derivative of the function $f$ at every point $F \in \mathbb{R}^{d \times D}$ for all rank-one directions $u \otimes v$ with $u \in \mathbb{R}^d, v \in \mathbb{R}^D$, which is a local condition. To be more precise, in the case of either poly- and quasiconvexity, there exists a function
$f : \mathbb{R}^{d \times D} \to \mathbb{R}$ that is not poly- or quasiconvex, but agrees with a poly- or quasiconvex function on any ball of radius one around any point $F \in \mathbb{R}^{d \times D}$. This is true for $d, D \geq 2$ in the case of polyconvexity and $d \geq 3, D \geq 2$ in the case of quasiconvexity. Here we extend these results to $n$-polyconvexity and we will prove that $n$-polyconvexity is nonlocal for $d, D \geq 2$ and all $n \geq 2$.

**Theorem 4.66.** Assume that $d, D \geq 2$ and $2 \leq n \leq d \wedge D$. Then there exists a smooth function $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ such that $f$ is not $n$-polyconvex, but that its restriction to any ball $B \subseteq \mathbb{R}^{d \times D}$ of radius one can be extended to a $n$-polyconvex function $f_B : \mathbb{R}^{d \times D} \to \mathbb{R}$.

**Remar 4.67.** This is not a surprising result for two reasons. Firstly, $n$-polyconvexity requires the function $f$ to have a convex representative for any simple rank-$n$ subspace of $\mathbb{R}^{d \times D}$. For example, we can choose the space of the first $n$ rows and columns being arbitrary and all others equal to zero, i.e. $V = \text{span}\{e_i \otimes e_j : 1 \leq i \leq n, 1 \leq j \leq n\}$ and there must exist $g_V : \mathbb{R}^{r(d,D)} \to \mathbb{R}$ such that $f|_V = g_V \circ T$ on $V$. Defining a projection operator $P : \mathbb{R}^{n \times n} \to V$ with $[P(F)]_{ij} = F_{ij}$ if $i, j \in \{1, \ldots, n\}$ and $[P(F)]_{ij} = 0$ otherwise, the condition $f|_V = g_V \circ T$ on $V$ is equivalent to the function $\tilde{f} = f \circ P$ being polyconvex. Thus, by carefully choosing the function $f$ we can refer the question of nonlocality for $n$-polyconvexity back to simply polyconvexity. We will do this in detail in the following proof.

The second reason is that each representative $g_V$ for each simple rank-$n$ subspace $V$ is defined on the set $T(V)$ and for all $n > 1$ the set $T(V)$ is a nonconvex. As Busemann et al. [16] remark, convexity on nonconvex sets is in general nonlocal and there is no reason to expect anything different in this particular case. On the other hand, when $n = 1$, $T(V)$ is a convex set already and thus convexity of $g_V$ is local. Therefore, rank-one convexity is local.

**Proof.** The proof is fairly straightforward and is based on the observation that any counterexample $f_{2 \times 2}$ to the locality of polyconvexity in the $\mathbb{R}^{2 \times 2}$ case also provides a counterexample $f$ in $\mathbb{R}^{d \times D}$ for $d, D \geq 2$ by simply defining

$$f(F) := f_{2 \times 2} \left( \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \right).$$

With this definition it is known that $f$ is polyconvex if and only if $f_{2 \times 2}$ is polyconvex. Here we show that $f$ is $n$-polyconvex for any $2 \leq n \leq d \wedge D$ if and only if $f_{2 \times 2}$ is polyconvex. It is easy to see that if $f_{2 \times 2}$ is polyconvex that then $f$ is $n$-polyconvex for $n \leq d \wedge D$ (since $f$ must be polyconvex). Thus it only remains to show that $f$ being
2-polyconvex implies that $f_{2\times2}$ must be polyconvex, which is sufficient for the statement to hold for $n \geq 2$. Thus assume that $f$ is 2-polyconvex. Then consider the simple rank-2 subspace $V = \text{span}\{e_i \otimes e_j : i, j \in \{1, 2\}\}$, i.e. the space of all matrices

$$F = \begin{bmatrix} F_{11} & F_{12} & 0 & \ldots & 0 \\ F_{21} & F_{22} & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \in V \subseteq \mathbb{R}^{d \times D}.$$  

Since $f$ is 2-polyconvex there exists $g_V : \text{co}(T(V)) \to \mathbb{R}$ such that $f = g_V \circ T$. Now define $g : \mathbb{R}^5 \to \mathbb{R}$ such that $g(X) = g_V(X_V)$ where $X_V \in \mathbb{R}^{\tau(d,D)}$ is the element that contains mostly zeros and only the possible five nonzero entries from $X$ at their corresponding places in the bigger space. Then $g$ inherits its convexity from $g_V$ and we also obtain for $F_{2\times2} \in \mathbb{R}^{2\times2}$ and $F \in V$ with $F_{ij} = F_{ij}^{2\times2}$ for $i, j = 1, 2$ that $g(T(F_{2\times2})) = g_V(T(F)) = f(F) = f_{2\times2}(F_{2\times2})$ and so $f_{2\times2}$ is polyconvex. Therefore, by Kristensen’s result [35], $n$-polyconvexity is nonlocal since there exists a function $f$ that is not $n$-polyconvex but agrees with a polyconvex function (and thus $n$-polyconvex function) on any ball of radius one. $\square$
5. A generalisation of abstract convexity

The previous chapter introduced the concept of $n$-polyconvexity and one of the main findings in the theory of $n$-polyconvexity was that any finite strong $n$-polyconvex function can be written as the pointwise supremum of $n$-polyaffine functions at the respective points. This is reminiscent of the concepts used in convex analysis for convex functions and has been made abstract in the theory of abstract convexity to include cases like polyconvexity as the supremum of polyaffines (known as quasiaffines in the current literature) as well as the relevant notions with respect to sets. The field of abstract convexity has many contributors and a comprehensive list of references can be found in the monographs [51, 55, 46]. The main point to be made here is that none of the contributions are capable of including the case of rank-one convexity (and $n$-polyconvexity for $n < d \land D$) into the abstract convexity framework. However, with the work of the previous chapter we can see that it is possible to generalise the concepts of abstract convexity to allow for such an inclusion. Abstract convexity rests on the observation that convex objects like functions or sets can be characterised by a simpler class of convex objects, e.g. affine functions or halfspaces, and that this characterisation does not in particular depend on the structure of the underlying involved spaces. Therefore, the methods of convex analysis can be generalised to include cases like polyconvexity, the essence of which is captured in abstract convexity theory. The way forward with respect to $n$-polyconvex functions is to allow the above elementary functions, in our case $n$-polyaffine functions, to depend on the point at which they form the supremum. Hence, this chapter will be devoted to this generalisation of abstract convexity.

5.1. Classical abstract convexity

The definitions and results presented in this subsection are taken from Singer’s monograph [55]. In sections 5.1.2 and 5.1.3 we base most of our definitions on the book by Rubinov [51], since it does not need the notion of dualities to define a subdifferential. As far as dualities or conjugations are concerned we will again use Singer’s work, however,
they will not be the focus of our attention.

5.1.1. Basic definitions and properties

The only building block necessary for the theory of abstract convexity is that the building blocks of the abstract convex object, e.g. affine planes for convex sets or affine functions for convex functions, are part of a partially ordered set in which all subsets have both a supremum and an infimum, i.e. a complete lattice \((E, \leq)\). With that we are able to define the abstract convexity notion of \(\mathcal{M}\)-convexity as follows.

**Definition 5.1.** Let \(E\) be a complete lattice and \(\mathcal{M} \subseteq E\). An element \(x \in E\) is said to be convex with respect to the set \(\mathcal{M}\) or \(\mathcal{M}\)-convex, if

\[
x = \sup\{m \in \mathcal{M} : m \leq x\}.
\]

Furthermore, we denote by \(\mathcal{C}(\mathcal{M})\) the set of all \(\mathcal{M}\)-convex elements of \(E\), i.e.

\[
\mathcal{C}(\mathcal{M}) = \{x \in E : x \text{ is } \mathcal{M}\text{-convex}\}.
\]

The set \(\mathcal{C}(\mathcal{M})\) is also called the convexity system generated by \(\mathcal{M}\).

For example, for \(X \subseteq \mathbb{R}^n\), consider the lattice \(E = \mathbb{R}^X\) as the set of all functions from \(X\) to \(\mathbb{R}\) with the partial order \(\leq\) being the pointwise comparison, i.e. \(f_1, f_2 \in \mathbb{R}^X\) with \(f_1 \leq f_2\) if and only if \(f_1(x) \leq f_2(x)\) for all \(x \in X\). The subset \(\mathcal{M} \subseteq E\) could then be chosen as the set of all affine functions from \(X\) to \(\mathbb{R}\) and \(\mathcal{M}\)-convexity is then convexity in the ‘usual sense’.

The following is to be expected from the above definition.

**Proposition 5.2.** Let \(E\) be a complete lattice and \(\mathcal{M}_1, \mathcal{M}_2 \subseteq E\). We have

\[
\mathcal{M}_1 \subseteq \mathcal{C}(\mathcal{M}_1) \tag{5.1}
\]

\[
\mathcal{M}_1 \subseteq \mathcal{M}_2 \Rightarrow \mathcal{C}(\mathcal{M}_1) \subseteq \mathcal{C}(\mathcal{M}_2).
\]

Another important concept is the following:

**Definition 5.3.** Let \(E\) be a complete lattice and let \(\mathcal{M} \subseteq E\). For any \(x \in E\) the \(\mathcal{M}\)-convex hull of \(x\) is the element \(\text{co}_\mathcal{M} x \in E\) defined by

\[
\text{co}_\mathcal{M} x = \sup\{h \in \mathcal{C}(\mathcal{M}) : h \leq x\}
\]
The $\mathcal{M}$-convex hull satisfies the following:

**Proposition 5.4.** (i) For any $x \in E$ we have

$$\text{co}_\mathcal{M} x \in \mathcal{C}(\mathcal{M})$$

and $\text{co}_\mathcal{M} x$ is the greatest $\mathcal{M}$-convex minorant of $x$:

$$\text{co}_\mathcal{M} x = \max \{ h \in \mathcal{C}(\mathcal{M}) : h \leq x \}.$$

(ii) $\mathcal{C}(\mathcal{M})$ is the set of all fixed points of the mapping $\text{co}_\mathcal{M} : E \to E$

$$\mathcal{C}(\mathcal{M}) = \{ x \in E : \text{co}_\mathcal{M} x = x \}.$$

The following theorem is quite concise to read, but nonetheless important:

**Theorem 5.5.** For any $x \in E$ we have

$$\text{co}_\mathcal{M} x = \sup \{ m \in \mathcal{M} : m \leq x \}.$$

The validity of this theorem is based on the observation that $\mathcal{M} \subseteq \mathcal{C}(\mathcal{M})$, see Prop. 5.2.

This very general definition of $\mathcal{M}$-convexity can yet be applied to various settings, including sets or functions. In the latter we will turn our attention of the applications of $\mathcal{M}$-convexity to functions only, in which case it is customary to switch the notation to $W$-convexity where the lattice $E$ is the set $\mathbb{R}^X$ of functions from a set $X$ to $\mathbb{R}$ and $W \subseteq \mathbb{R}^X$. Note that Rubinov uses the terminology of $H$-convexity [37, 51].

### 5.1.2. Subdifferentials

In Singer’s work subdifferentials get introduced much later, only after the study of dualities and conjugations. This is reasonable as with dualities or conjugations at your disposal there are many more properties that can be derived for subdifferentials. However, at this stage we are content with simply providing the definition for them without the necessity of using dualities, which follows more Rubinov’s approach [51].
Definition 5.6. Let $X$ be a set, $E = \mathbb{R}^X$ and $W \subseteq \mathbb{R}^X$. Further let $f \in \mathbb{R}^X$ and $x_0 \in X$ with $f(x_0) \in \mathbb{R}$, i.e. $f$ is finite at $x_0$. Then $w \in W$ is called an abstract subgradient (or $W$-subgradient) of $f$ at $x_0$ if

$$f(x) - f(x_0) \geq w(x) - w(x_0)$$

for all $x \in X$. Further we denote the $W$-subdifferential of $f$ at $x_0$ by the set $\partial_W f(x_0)$ as the set of all subgradients $w$ of $f$ at $x_0$, i.e.

$$\partial_W f(x_0) = \{w_0 \in W : f(x) - f(x_0) \geq w(x) - w(x_0) \text{ for all } x \in X\}.$$ 

The following proposition shows that in some way subgradients are simply support functions that are possibly shifted vertically. Denote by $V_W$ the closure of $W$ under vertical shifts, i.e. $V_W = \{h : h(\cdot) = w(\cdot) - c, w \in W, c \in \mathbb{R}\}$.

**Proposition 5.7** (cf. Prop. 1.2, [51]). Let $f : X \to \mathbb{R}$ and $x_0 \in X$ such that $f(x_0) \in \mathbb{R}$. Then the $W$-subdifferential $\partial_W f(x_0)$ is nonempty if and only if

$$f(x_0) = \max\{h(x_0) : h \leq f \text{ and } h \in V_W\}.$$ 

5.1.3. Dualities and Legendre-Fenchel transforms

As with the previous section of subdifferentials Singer’s monograph [55] contains a much larger collection of results than Rubinov’s introduction to dualities and Legendre-Fenchel transform in [51]. We will take the less detailed approach of Rubinov and directly consider conjugations (which are special dualities) and only cover those properties that will be relevant for us later. For two sets $X$ and $W$ a conjugation is a mapping $c : \mathbb{R}^X \to \mathbb{R}^W$ such that for each index set $I$ we have

$$c \left( \inf_{i \in I} f_i \right) = \sup_{i \in I} c(f_i)$$

(5.2)

and if

$$c(f + d) = c(f) + d$$

(5.3)

for all $f \in \mathbb{R}^X$ and $d \in \mathbb{R}$, and where $+$ and $\pm$ are the upper and lower addition that extend the usual addition $a + b$ when not both $a$ and $b$ are equal to $+\infty$ or $-\infty$ to the
case where both a and b may be $+\infty$ or $-\infty$. In this case it is defined

$$+\infty + (-\infty) = -\infty + (+\infty) = +\infty$$

and

$$+\infty + (-\infty) = -\infty + (+\infty) = -\infty.$$

It is customary to write $f^c$ instead of $c(f)$, which we will do from now on. With the help of coupling functions it can be shown that all conjugations appear via a construction that resembles the well-known Fenchel conjugate.

Theorem 5.8 (Thm. 8.2, [55]). Let $X$ and $W$ be two sets. Then $c: \overline{\mathbb{R}}^X \to \overline{\mathbb{R}}^W$ is a conjugation if and only if there exists a coupling function $\varphi: X \times W \to \mathbb{R}$ such that

$$f^c(w) = \sup_{x \in X} \{ \varphi(x, w) + -f(x) \} \quad (5.4)$$

for all $f \in \overline{\mathbb{R}}^X$ and $w \in \overline{\mathbb{R}}^W$. Furthermore, $\varphi$ in the above equation is uniquely determined by $c$ via the formula

$$\varphi(x, w) = (\chi_x)^c(w).$$

Definition 5.9. For any coupling function $\varphi: X \times W \to \overline{\mathbb{R}}$ we call the conjugation $c(\varphi)$ defined through (5.4) the (Fenchel-Moreau) conjugation associated to $\varphi$. Vice versa, given a conjugation $c$ we call the uniquely determined function $\varphi_c$ the associated coupling function to the conjugation $c$.

In the case where $W \subseteq \overline{\mathbb{R}}^X$ and the coupling function $\varphi$ is the so-called natural coupling function $\varphi_{nat}$ with $\varphi_{nat}(x, w) = w(x)$ the above reduces to $f^c(w) = \sup_{x \in X} \{ w(x) + -f(x) \}$ which is commonly denoted by $f^*$ and referred to as the (generalised) Fenchel conjugate of $f$. This is also the case that we are concerned about mostly.

By duality theory every conjugation $c$ has a dual $c'$, which is itself a duality and in this case also a conjugation (and thus called the dual conjugation). In general the dual conjugation $c': \overline{\mathbb{R}}^W \to \overline{\mathbb{R}}^X$ is defined by

$$c'(w) = \inf \{ f \in \overline{\mathbb{R}}^X : f^c \leq w \}.$$
As mentioned just before, \( c' \) is a conjugation (cf. [55, Thm. 8.1]) and so it has an associated coupling function \( \varphi_{c'} : W \times X \to \mathbb{R} \). It turns out that the two coupling functions \( \varphi_c \) and \( \varphi_{c'} \) satisfy that \( \varphi_c(x, w) = \varphi_{c'}(w, x) \) for all \( x \in X \) and \( w \in W \). Thus, we may write

\[
h_{c'}(x) = \sup_{w \in W} \{ \varphi_{c'}(w, x) + -h(w) \} = \sup_{w \in W} \{ \varphi_c(x, w) + -h(w) \}.
\]

For \( f \in \mathbb{R}^X \) the function \( f^{cc'} \) is called the biconjugate of \( f \). Considering again the case when \( W \subseteq \mathbb{R}^X \) and \( \varphi = \varphi_{nat} \) it is customary to write \( h^* \) instead of \( h^{cc'} \) although this is an abuse of notation since we normally require \( X \subseteq \mathbb{R}^W \), which may not be the case. Nevertheless, it is still common to write \( f^{**} \) for the biconjugate of \( f \) in this setting.

5.1.4. Biconjugates and abstract convex hulls

The following theorem is the main motivation for the use of dualities or conjugations. It states that for \( W \subseteq \mathbb{R}^X \) the biconjugate \( f^{**} \) of a function \( f \) corresponds to its \((W + \mathbb{R})\)-convex hull \( f_{co(W+\mathbb{R})} \). Note that we are taking the closure of \( W \) under vertical shifts. It also turns out that \( f_{co(W+\mathbb{R})} = f_{co(W+\mathbb{R}^+)} \).

**Theorem 5.10.** Let \( X \) be a set and \( W \subseteq \mathbb{R}^X \) and \( \varphi = \varphi_{nat} : X \times W \to \mathbb{R} \) the natural coupling function. Then denoting \( f^{cc(\varphi)\varphi'} \) by \( f^{**} \) we have

\[
f^{**} = f_{co(W+\mathbb{R})} = f_{co(W+\mathbb{R}^+)}.
\]

**Corollary 5.11.** (Cor. 8.2, [55]). Let \( X \) and \( W \) as above. Then the following statements are equivalent

(i) \( f \in C(W + \mathbb{R}) \),

(ii) \( f \in C(W + \mathbb{R}^+) \),

(iii) \( f = f^{**} \),

(iv) \( f = \sup_{w \in W} \{ w + -f^*(w) \} \).

5.2. Generalised abstract convexity in the case of \( E = \mathbb{R}^X \)

We will now specialise on the case of extended real-valued functions, i.e. \( E = \mathbb{R}^X \), where \( X \subseteq \mathbb{R}^n \). Then for a subset \( W \subseteq \mathbb{R}^X \) we have the usual notion of \( W \)-convexity as defined
in Definition 5.1 (where $W$ is used instead of $M$) and the set of closed convex functions can be characterised by the set of affine functions. In particular, polyconvex functions fall into this category. The subset $W \subseteq \mathbb{R}^X$ where $X = \mathbb{R}^{d \times D}$ is then the set of polyaffine functions, i.e.

$$W = \{ h \in \mathbb{R}^X : h(\cdot) = \langle \beta, T(\cdot) - T(F) \rangle + h(F) \text{ on } \mathbb{R}^{d \times D}, \beta \in \mathbb{R}^{\tau(d,D)} \}$$

and we have that indeed $C(W)$ only contains polyconvex functions. However, it is not possible to find such a simple set (by that we mean a set of functions that are affine in some sense) for the cases of rank-one convexity (or even quasiconvexity). The reason for that is that the classical definitions of rank-one affine or quasiaffine functions coincide with that of polyaffine functions as discussed in the introduction of this chapter. Equally, it was hinted what generalisation is necessary (at least for the case of rank-one convexity) in order develop an analogous duality between rank-one convex functions and supremum of rank-one affine functions (in the new sense). This involved localising what it means to be rank-one affine to points $F \in \mathbb{R}^{d \times D}$ rather than on the whole of $\mathbb{R}^{d \times D}$. In other words, the set of functions to account for the particular convexity notion may now additionally depend on the points of the domain of the functions. In a way, Rubinov et al. [50] also give a generalisation of abstract convexity which is a localisation of abstract convexity. They define abstract convexity with respect to a fixed subset (the notion is called $H$-convexity with respect to a subset). Yet, this generalisation is still sufficient to include cases like rank-one convexity and we take it a step further and allow the subset to vary. The following will make this more precise.

5.2.1. Basic definitions and properties

**Definition 5.12.** Let $E = \mathbb{R}^X$, i.e. the complete lattice of functions from a set $X$ to $\mathbb{R}$. For each $x \in X$ let $W_x \subseteq \mathbb{R}^X$. Then we call $f \in \mathbb{R}^X$ convex with respect to the family $W_X = \{ W_x \}_{x \in X}$, or $W_X$-convex if

$$f(x) = \sup \{ w(x) : w \in W_x \text{ and } w \leq f \}$$

for all $x \in X$. Furthermore, we denote by $C(W_X)$ the set of all $W_X$-convex functions of $\mathbb{R}^X$, i.e.

$$C(W_X) = \{ f \in \mathbb{R}^X : f \text{ is } W_X \text{-convex} \}.$$
$W_X$-convex hull of $f$ is the element $W_X$-co $f \in E$ defined by

$$W_X$-co $f = \sup\{ h \in \mathcal{C}(W_X) : h \leq f \}$$

Note that due to the possible dependence on $x$ of the sets $W_x$ we now lose a few properties that would have held otherwise. For example, Prop. 5.2 now only becomes:

**Proposition 5.14.** Let $E = \mathbb{R}^X$ and $\{ W^1_x \}_{x \in X}, \{ W^2_x \}_{x \in X} \subseteq \mathcal{P}(E)$. Then

$$W^1_x \subseteq W^2_x \text{ for all } x \in X \Rightarrow \mathcal{C}(W^1_X) \subseteq \mathcal{C}(W^2_X).$$

Note that there is no equivalent of (5.1) in this case. Indeed, in general we have for $x \in X$ that

$$W_x \not\subseteq \mathcal{C}(W_X).$$

For that reason an equivalent of Th. 5.5 in that form does not exist. However, we do have the following:

**Theorem 5.15.** For any $f \in E = \mathbb{R}^X$ and convexity system $W_X$ we have

$$(W_X$-co $f)(x) \leq \sup\{ w(x) : w \in W_x \text{ and } w \leq f \}.$$
It is intuitive that all \( W_X \)-laminates of a function \( f \) share the same \( W_X \)-convex hull, namely \( W_X \)-co \( f \), which the next theorem is expressing.

**Theorem 5.17.** Let \( E = \mathbb{R}^X \) and \( \mathcal{C}(W_X) \) a convexity system on \( E \). Let \( f \in E \) such that there exists \( g \in \mathcal{C}(W_X) \) with \( f \geq g \). Then

\[
W_X \text{-co } f = W_X \text{-lam}_\infty f \leq \ldots \leq W_X \text{-lam}_{i+1} f \\
\leq W_X \text{-lam}_i f \leq W_X \text{-lam}_{i-1} f \leq \ldots \leq W_X \text{-lam}_1 f \leq f,
\]

where \( W_X \text{-lam}_\infty f := \lim_{i \to \infty} W_X \text{-lam}_i f \).

**Proof.** First we prove that for \( h \in \mathcal{C}(W_X) \) with \( h \leq f \) it follows that \( h \leq W_X \text{-lam}_i f \) for all \( i \in \mathbb{N} \). Thus assume \( h \in \mathcal{C}(W_X) \) and \( h \leq f \). For \( i = 1 \) we have \( h(x) = \sup \{ w(x) : w \in W_x \text{ and } w \leq h \} \leq \sup \{ w(x) : w \in W_x \text{ and } w \leq f \} = W_X \text{-lam}_1 f(x) \). Now assume that \( h \leq W_X \text{-lam}_i f \) is true for some \( i > 1 \). Then,

\[
h(x) = \sup_{w \in W_x, w \leq h} w(x) \leq \sup_{w \in W_x, w \leq W_X \text{-lam}_i f} w(x) = W_X \text{-lam}_{i+1} f(x)
\]

and so \( h \leq W_X \text{-lam}_{i+1} f \) and hence, \( h \leq W_X \text{-lam}_i f \) for all \( i \in \mathbb{N} \) by induction. Thus this holds in particular for \( h = W_X \text{-co } f \). Furthermore define \( f^\infty = \lim_{i \to \infty} W_X \text{-lam}_i f \).

The limit exists since it is a monotonically decreasing sequence that is bounded below by \( g \). Then we must have that \( f^\infty = W_X \text{-lam } f^\infty = \sup \{ w(x) : w \in W_x \text{ and } w \leq f^\infty \} \). Therefore \( f^\infty \in \mathcal{C}(W_X) \) and hence \( W_X \text{-co } f = f^\infty \).

In this section we assumed that we have the lattice \( E = \mathbb{R}^X \). In fact, the above is generalisable even more to the case where we have the lattice \( E = F^X \), where \( F \) itself is a lattice. Thus, we could consider the convexity of set-valued functions from \( X \) to \( F \) where \( F \) is a lattice of sets. However, in the present case we will restrict our attention to the case of \( E = \mathbb{R}^X \) as we are motivated mainly by the semiconvex functions from the calculus of variations.

### 5.2.2. Generalised subdifferentials

Analogously to the subdifferentials defined for standard abstract convexity notions we now want to define subdifferentials for the generalised abstract convexity of the previous section. It is not difficult to imagine how this would look like given that subdifferentials are defined for a point \( x_0 \in X \) of a function \( f \in \mathbb{R}^X \) and that in the above generalisation we have \( f \) as the pointwise supremum of functions from the sets \( W_x \subseteq \mathbb{R}^X \).
Definition 5.18. Let $X$ be a set, $E = \mathbb{R}^X$, $W_x \subseteq \mathbb{R}^X$ for all $x \in X$ and $W_X = \{W_x\}_{x \in X}$. Further let $f \in \mathbb{R}^X$ and $x_0 \in X$ with $f(x_0) \in \mathbb{R}$. Then $w_{x_0} \in W_{x_0}$ is called the $W_X$-subgradient of $f$ at $x_0$ if

$$f(x) - f(x_0) \geq w_{x_0}(x) - w_{x_0}(x_0)$$

for all $x \in X$. Further we denote the subdifferential of $f$ by the set $\partial W_X f(x_0)$ as the sets of all $W_X$-subgradients $w_{x_0}$ of $f$ at $x_0$, i.e.

$$\partial W_X f(x_0) = \{w_{x_0} \in W_{x_0} : f(x) - f(x_0) \geq w_{x_0}(x) - w_{x_0}(x_0) \text{ for all } x \in X\}.$$

Analogously we define the sets $V_{W_x}$ as the closure of each $W_x$ under vertical shifts, so that for each $x \in X$ we have $V_{W_x} = \{h : h(\cdot) = w_x(\cdot) - c, w_x \in W_x, c \in \mathbb{R}\}$. Then yet again we obtain:

Proposition 5.19. Let $f \in \mathbb{R}^X$ and $x_0 \in X$ such that $f(x_0) \in \mathbb{R}$. Then the $W_X$-subdifferential $\partial W_X f(x_0)$ of $f$ at $x_0$ is nonempty if and only if

$$f(x_0) = \max\{h(x_0) : h \leq f \text{ and } h \in V_{W_{x_0}}\}.$$

The proof can be transferred one to one from the original setting of Proposition 5.7 and will be omitted.

5.2.3. Dualities and Legendre-Fenchel transforms

In the corresponding section for original abstract convexity we discussed conjugations $c$ between the sets $\mathbb{R}^X$ and $\mathbb{R}^W$ where $X$ and $W$ are two sets and in particular when $W \subseteq \mathbb{R}^X$. It is obvious that we now immediately have a family of conjugations at our disposal, namely $c_x = \{c_x\}_{x \in X}$ where $c_x$ is a conjugation between $\mathbb{R}^X$ and $\mathbb{R}^{W_x}$ for each $x \in X$. By Theorem 5.8 each $c_x$ defines or is defined by a coupling function $\varphi_x : X \times W_x \to \mathbb{R}$ such that $f^{\varphi_x}(w_x) = \sup_{y \in X} \{\varphi_x(y, w_x) + f(y)\}$ for all $f \in \mathbb{R}^X$ and $w_x \in \mathbb{R}^{W_x}$. Furthermore, each of the conjugations $c_x$ has a dual conjugation $c'_x$ and we define $c'_X = \{c'_x\}_{x \in X}$ as the family of all dual conjugations. Accordingly, for $h \in \mathbb{R}^{W_x}$ we have that

$$h^{c'_x}(y) = \sup_{w_x \in W_x} \{\varphi_x(w_x, x) + h(w_x)\} = \sup_{w_x \in W_x} \{\varphi_x(x, w_x) + -h(w_x)\}$$
Then for \( f \in \mathbb{R}^X \) we call \( f^{c_x \epsilon'_x} \) the \textit{biconjugate of } \( f \text{ at } x \). In the case of \( W_x \subseteq \mathbb{R}^X \) for all \( x \in X \) it is now no longer appropriate to use the notation \( f^{**} \) for the biconjugate since it depends on \( x \) and would otherwise be ambiguous. More implications of this dependence will become apparent in the next section.

### 5.2.4. Biconjugates and abstract convex hulls

In this section we only consider the case when \( W_x \subseteq \mathbb{R}^X \) for all \( x \in X \). From the previous section we have the biconjugates \( f^{c_x \epsilon'_x} \) for each \( x \in X \). Taking these we now define a new function \( f^{c_X \epsilon'_X} \) that we will be able to relate back to the \( W_X \)-laminate operator from Section 5.2.1.

**Definition 5.20.** Let \( X \) be a set and \( W_X = \{ W_x \}_{x \in X} \) with \( W_x \subseteq \mathbb{R}^X \). Further let \( c_X = \{ c_x \}_{x \in X} \) be the family of conjugations associated to the natural coupling function \( \varphi_{\text{nat}} : X \times W_x \rightarrow \mathbb{R}, \varphi_{\text{nat}}(x, w) = w(x) \) for each \( x \in X \) and \( c'_X = \{ c'_x \}_{x \in X} \) the family of their respective dual conjugations. Then we call the function \( f^{c_X \epsilon'_X} \) defined by

\[
f^{c_X \epsilon'_X}(x) = f^{c_x \epsilon'_x}(x)
\]

the \( X \)-biconjugate of \( f \).

Now let \( W_X = \{ W_x \}_{x \in X} \) with \( W_x \subseteq \mathbb{R}^X \) for all \( x \in X \). Then by an abuse of notation we denote by \( W_X + \mathbb{R} \) the family \( \{ W_x + \mathbb{R} \}_{x \in X} \), i.e. the closure of \( W_x \) under vertical shifts for each \( x \in X \). Then we obtain the following relation between the \( X \)-biconjugate \( f^{c_X \epsilon'_X} \) of \( f \) and the \( (W_X + \mathbb{R}) \)-laminate (\( W_X + \mathbb{R} \))-lam\((f) \) of \( f \).

**Theorem 5.21.** With the same assumption as in Definition 5.20 we have that

\[
f^{c_X \epsilon'_X} = (W_X + \mathbb{R}) \text{-lam}(f).
\]

**Proof.** By [55, Thm. 8.5] it holds in general that \( f^{c_x \epsilon'_x} = \sup \{ \varphi_{c_x}(\cdot, w_x) + r : w_x \in W_x, r \in \mathbb{R}, \varphi_{c_x}(\cdot, w_x) + r \leq f \} = \sup \{ \varphi_{c_x}(\cdot, w_x) + r : w_x \in W_x, r \in \mathbb{R}, \varphi_{c_x}(\cdot, w_x) + r \leq f \} \). Specifically for \( \varphi_{c_x} = \varphi_{\text{nat}} \) for all \( x \in X \) we obtain at \( x \)

\[
f^{c_x \epsilon'_x}(x) = \sup \{ w_x(x) + r : w_x \in W_x, r \in \mathbb{R} \} = \sup \{ w_x(x) + r : w_x \in W_x, r \in \mathbb{R} \} = ((W_X + \mathbb{R}) \text{-lam } f)(x)
\]

and thus, since the above holds for all \( x \in X \), we deduce \( f^{c_X \epsilon'_X} = (W_X + \mathbb{R}) \text{-lam } f \). \( \Box \)
Therefore we can deduce analogously:

**Corollary 5.22.** Let $X$ be a set and $W_X$ be as above. Then the following statements are equivalent

(i) $f \in C(W_X + \overline{\mathbb{R}}),$

(ii) $f \in C(W_X + \mathbb{R}),$

(iii) $f = f^{cx}c_X,$

(iv) $f(x) = \sup_{w_x \in W_x} \{w_x(x) + -f^{cx}(w_x)\}$ for all $x \in X.$

Note that in all the above we have used the notion $c^*_X$ instead of $c'_X.$ This is due to the observation that $c^*_X$ is not the dual of $c_X$ and therefore $c'_X$ could lead to confusion.

The fact that $c^*_X$ is in general not the dual of $c_X$ can be easily acknowledged with the equivalence to the $W_X$-laminate operator, which does not in general equal the $W_X$-convex hull operator and only converges to it by potentially infinite repeated application. However, this raises the question whether there exists a suitable target space $F$ and a conjugation $c:F \to \mathbb{R}^X$ such that the biconjugate $f^{c'_c}$ does equal the $W_X$-convex hull of $f,$ where $c':F \to \mathbb{R}^X$ is the dual conjugate of $c.$ In the following we will show that there is always a trivial choice of such a conjugation.

**Proposition 5.23.** Let $X$ be a set $C(W_X)$ be the convexity system defined by $W_X = \{W_x\}_{x \in X}$ for $W_x \subseteq \overline{\mathbb{R}}^X$ for all $x \in X$ and where $W_x$ is closed under vertical shifts for all $x \in X.$ Then the map $c_0: \mathbb{R}^X \to (C(W_X) \to \mathbb{R})$ defined such that

\[(c_0(f))(g) = \sup_{x \in X} \{g(x) + -f(x)\}\] (5.5)

is a conjugation and its biconjugate satisfies

\[c'_0c_0f = W_X\text{-co } f,\]

i.e. the biconjugate $c'_0c_0$ is corresponds to the $W_X$-convex hull operator.

**Proof.** For brevity we denote the space $(C(W_X) \to \mathbb{R})$ by $F.$ Note that $F$ is not the dual space of $C(W_X)$ as we do not require the maps from $C(W_X)$ into $\mathbb{R}$ to be linear or continuous. In (5.5) we could replace the term $g(x)$ by $\varphi(x, g)$ if $\varphi : X \times C(W_X) \to \overline{\mathbb{R}}$ is the coupling function defined by $\varphi(x, g) = g(x).$ Thus, by Theorem 8.2 of [55] it follows that $c_0$ is a conjugation.
We now show that its dual $c'_0 : F \to \mathbb{R}^X$ is such that $c'_X c_X : \mathbb{R}^X \to \mathbb{R}^X$ is the $W_X$-convex hull operator, i.e. $c'_0 c_0 = W_X$-co. As usual $c'_0$ is defined as

$$c'_0 G = \inf\{h \in \mathbb{R}^X : c_0 h \leq G\}$$

$$= \inf\{h \in \mathbb{R}^X : \sup_{x \in X} \{g(x) + -h(x)\} \leq G(g) \text{ for all } g \in \mathcal{C}(W_X)\} \quad (5.6)$$

for $G \in F = (\mathcal{C}(W_X) \to \mathbb{R})$. Let $f \in \mathbb{R}^X$. First we show that $c'_0 c_0 f \leq W_X$-co $f$ by proving that $c_0(W_X$-co $f) \leq c_0 f$, making $W_X$-co $f$ a candidate for the attaining the infimum in \[5.6\] with $G = c_0 f$. Thus we must show that $c_0(W_X$-co $f)(g) \leq (c_0 f)(g)$ for all $g \in \mathcal{C}(W_X)$, which is equivalent to

$$\sup_{x \in X} \{g(x) + -W_X$-co $f(x)\} \leq \sup_{x \in X} \{g(x) + -f(x)\}.$$ 

We may assume that for $g \in \mathcal{C}(W_X)$ the right hand side is finite and that the supremum is attained for $\overline{x} \in X$, i.e. we have that

$$g(x) - f(x) \leq g(\overline{x}) - f(\overline{x})$$

for all $x \in X$ (we can replace $+$ by $+$ since the right hand is finite). Rearranging the inequality we find that $g(x) - g(\overline{x}) + f(\overline{x}) \leq f(x)$ for all $x \in X$. Thus, since $\mathcal{C}(W_X)$ is closed under vertical shift the function $g - g(\overline{x}) + f(\overline{x})$ is $W_X$-convex and minorises $f$. Because $W_X$-co $f$ is the largest function in $\mathcal{C}(W_X)$ that minorises $f$ it also follows that $g - g(\overline{x}) + f(\overline{x})$ minorises $W_X$-co $f$, i.e.

$$g(x) - g(\overline{x}) + f(\overline{x}) \leq W_X$-co $f(x)$$

for all $x \in X$. Again, by rearranging we find that $\sup_{x \in X}\{g(x) + -W_X$-co $f(x)\} \leq g(\overline{x}) - f(\overline{x}) = \sup_{x \in X}\{g(x) + -f(x)\}$. Therefore, we have $c_X W_X$-co $f \leq c_X f$ and hence $c'_X c_X f \leq W_X$-co $f$.

To show the reverse inequality, let $h = c'_X c_X f$. We now simply test $c_X h \leq c_X f$ with $W_X$-co $f \in \mathcal{C}(W_X)$. Clearly, $(c_X f)(W_X$-co $f) = \sup_{x \in X}\{W_X$-co $f(x) + -f(x) \leq 0\}$ and hence we must have that $\sup_{x \in X}\{W_X$-co $f(x) + -h(x)\} \leq 0$. This is only possible if $h \geq W_X$-co $f$. \[\square\]

Although this proposition proves that $c'_0 c_0 = W_X$-co it does not provide any advantages. This is because the space $F = (\mathcal{C}(W_X) \to \mathbb{R})$ is extremely large. On the other hand, in classical abstract convexity it is known that a much smaller space can be used,
in particular \( c_W : \mathbb{R}^X \to F \) with \( F = (W \to \mathbb{R}) \) and \( c_W \) defined analogously to \( c_W \) also satisfies \( c_W'c_W \) = \( W_X\)-co and \( W \) can be much smaller than \( C(W) \) (e.g. affine functions as opposed to convex functions). Therefore, a natural question is whether another conjugation \( c_X : \mathbb{R}^X \to F \) with a smaller target space \( F \) can be defined such that \( c_X'c_X = W_X\)-co for a convexity system \( C(W_X) \) in our generalised abstract convexity theory. A natural candidate would be to consider \( c_X \) with the target space \( F = \left( X \to \mathbb{R}^W \right) \) whereby we mean that an element \( g \in F \) is a function on \( X \) and each \( g(x) \) is in turn a function from \( W_x \) into \( \mathbb{R} \). In particular, we could define \( (c_X f)(x) = c_x f \in (W_x \to \mathbb{R}) \).

Note that this is simply a different notation to writing \( c_X = \{ c_x \}_{x \in X} \) as in Definition 5.20. In fact, \( c_X \) is a conjugation since it satisfies (5.2) and (5.3). We formalise this in the following proposition.

**Proposition 5.24.** Let \( X \) and \( W_X \) be as in Theorem [5.21]. Then \( c_X \) is a conjugation.

**Proof.** We need to verify (5.2) and (5.3). Let \( I \) be an index set and \( \{ f_i \}_{i \in I} \subseteq \mathbb{R}^X \) a family of functions. Then, since each \( c_x \) is a conjugation we have

\[
c_X \left( \inf_{i \in I} f_i \right)(x) = c_x \left( \inf_{i \in I} f_i \right) = \sup_{i \in I} c_x (f_i) = \sup_{i \in I} c_X (f_i)(x) = \left( \sup_{i \in I} c_X (f_i) \right)(x).
\]

Thus, \( c_X (\inf_{i \in I} f_i) = \sup_{i \in I} c_X (f_i) \). Now let \( d \in \mathbb{R} \). In the following we will identify \( d \) with the constant function in various spaces. We then have

\[
c_X (f + d)(x) = c_x (f + d) = c_x (f) + -d = c_X (f)(x) + -d(x) = (c_X (f)) + -d(x),
\]

and therefore \( c_X (f + d) = c_X (f) + -d \).

We now consider the dual conjugation \( c_X' \) of \( c_X \). By the theory of duality it holds that \( c_X c_X' c_X = c_X \) and \( c_X' c_X c_X' = c_X' \). Thus, \( c_X c_X' \) is a hull operator. We therefore suspect that \( c_X' \neq c_X' \) in general as discussed before. We did not define the operator \( c_X \) in our new notation, but in the context of \( F = \left( X \to \mathbb{R}^W \right) \) the following is consistent. For \( g \in \left( X \to \mathbb{R}^W \right) \) we define \( c_X (g) = \tilde{g} \in \mathbb{R}^X \) where

\[
\tilde{g}(x) = g(x)c_X'(x),
\]

and hence, when \( g = c_X (f) \) we obtain \( f^c \in X' \rightarrow (c_X' \circ c_X (f))(x) = c_X (f)(x)c_X'(x) = c_X' (f)(x)c_X'(x) =

150
Then it holds that $W \cap \mathbb{R}^X$ where the last equality follows from $f^{c_x} \leq g(x)$ for all $x \in X$.

where the last equality follows from $f^{c_x}, g(x) \in \mathbb{R}^{W_x}$ where $\mathbb{R}^{W_x}$ is a lattice with partial ordering '<' as a pointwise comparison, i.e. for $f_1, f_2 \in \mathbb{R}^{W_x}$ we have $f_1 \leq f_2$ if and only if $f_1(w_x) \leq f_2(w_x)$ for all $w_x \in W_x$. The following example will show that $c'_X c_X f$ does not always yield the desired result of being equal to $W_X \cdot f$. Note, however, that duality theory requires that $c'_X c_X$ is a hull operator, i.e. it must hold that $c'_X c_X (c'_X c_X) = c'_X c_X$.

**Example 5.25.** Let $X = \mathbb{R}^{2 \times 2}$ and let $W_X = \{W_F \mid F \in X\}$ with $W_F = \{w_F : X \to \mathbb{R} : w_F$ is 1-polyaffine at $F\}$. Define the function $f : \mathbb{R}^{2 \times 2} \to \mathbb{R}$ such that

$$f(F) = \begin{cases} 0, & F \neq 0 \\ 1, & F = 0. \end{cases}$$

Then it holds that $W_X \cdot f \equiv 0$ while $c'_X c_X f = f$.

**Proof.** Let $h = c'_X c_X f$. Since $f$ trivially satisfies $c_X f \leq c_X f$ it follows that $h \leq f$. This in turn implies that $c_X h \geq c_X f$, but $h$ also satisfies $c_X h \leq c_X f$ by definition and hence $c_X h = c_X f$. In particular this implies that $(c_X h)(I) = (c_X f)(I)$, or equivalently $(c_X h)(w_1) = (c_X f)(w_1)$ for all $w_1 \in W_I$, where $I \in \mathbb{R}^{2 \times 2}$ is the identity matrix. Note that $f$ is 1-polyaffine at $F = I$, i.e. $f \in W_I$. Thus we must have that $(c_I h)(f) = (c_I f)(f) = \sup_{F \in \mathbb{R}^{2 \times 2}} \{f(F) - f(I)\} = 0$. It holds that $0 = (c_I h)(f) = \sup_{F \in \mathbb{R}^{2 \times 2}} \{f(F) - f(I)\} = \max\{\sup_{F \neq 0} (-h(F)), 1 - h(0)\}$, which implies that both $\sup_{F \neq 0} (-h(F)) \leq 0$ and $1 - h(0) \leq 0$ must be true. These two conditions imply that $h \geq f$, which together with the inequality $h \leq f$ as observed above, implies $h = f$, as claimed.

Although $W_X \cdot f \neq c'_X c_X f$ for this example it is still true that $c'_X c_X$ is a hull operator, albeit not a very useful one, as in the particular case we have that $c'_X c_X (c'_X c_X f) = c'_X c_X f = f$. Furthermore, if instead we choose $W_F$ as the set of polyaffine functions from $\mathbb{R}^{2 \times 2} \to \mathbb{R}$ for all $F \in \mathbb{R}^{2 \times 2}$ we do obtain that $W_{\mathbb{R}^{2 \times 2}} \cdot f = c'_X c_X f = 0$ as desired.

The definition of the conjugation $c_X : \mathbb{R}^{X} \to (X \to \mathbb{R}^{W})$ was an attempt to choose a smaller target space than $\mathbb{R}(W_X) \to \mathbb{R}$ for $c_0$, but it does not seem to be suitable for $W_X$-convexity. It is currently not clear whether there exists a better definition of a conjugation $c_X$ and its target space such that $c'_X c_X = W_X \cdot c_0$ or whether the operator...
5.3. *n*-polyconvexity in the abstract convexity framework

In this section we want to highlight some subtle differences between strongly *n*-polyconvex functions and those that can be written as the pointwise supremum of *n*-polyaffine functions. Due to Theorem 4.35 the differences only exist for in the extended real-valued case. Given the strict inclusion (in the extended real-valued case) of the set of strong *n*-polyconvex functions in that which contains functions that can be written as the pointwise supremum of *n*-polyaffines we define the latter class as abstract *n*-polyconvex functions. This is contained in Section 5.3.1.

In Section 5.3.2 following the characterisation of the intersectional *n*-polyconvex hull of a set as the zero set of the *n*-polyconvex envelope of its characteristic function, we define the abstract *n*-polyconvex hull of a set as the smallest abstract *n*-polyconvex set that contains the set. A set is abstract *n*-polyconvex if its characteristic function is abstract *n*-polyconvex. We will show that the abstract *n*-polyconvex hull contains the *n*-polyconvex hull as is itself contained in the strong functional *n*-polyconvex hull, while all concepts are generally different.

In Section 5.3.3 we consider the *W*-*X*-laminate operator of *n*-polyconvexity as defined in general in Section 5.2.1. Like the iterative process defined by (4.54) to compute the *n*-polyconvex envelope of a function, the laminate operator can be used to compute the abstract *n*-polyconvex envelope of a function. We compare the two methods.

5.3.1. Definition and basic properties for functions and sets

In Sec. 4.2.4 we have already demonstrated how finite strongly *n*-polyconvex functions can be written as the pointwise supremum of *n*-polyaffine functions. In the abstract convexity setting we would define for \( F \in \mathbb{R}^{d \times D} \) the family of functions \( W^n_F = \{ h : \mathbb{R}^{d \times D} \to \mathbb{R} : h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F), \forall \tilde{F} \in C^n(F), \beta \in \mathbb{R}^{(d,D)} \} \) and then \( W^n_{\mathbb{R}^{d \times D}} = \{ W^n_F \}_{F \in \mathbb{R}^{d \times D}} \). A finite function \( f \) is then strongly *n*-polyconvex if and only if it is \( W^n_{\mathbb{R}^{d \times D}} \)-convex. However, if the function is allowed to assume the value \(+\infty\) then strong *n*-polyconvexity does not generally imply that the function can be written as the pointwise supremum of *n*-polyaffine function, i.e. that the function is \( W^n_{\mathbb{R}^{d \times D}} \)-convex. In order to distinguish them (from the possibly larger) class of (strongly) *n*-polyconvex functions defined in Section 4.2 we will hence refer to them as abstract *n*-polyconvex functions. With respect to the notion of *n*-polyaffine functions at \( F \) for any \( F \in \mathbb{R}^{d \times D} \) as...
defined in Definition 4.17 and to the generalised theory of abstract convexity in Sec. 5.2 it is clear how we define abstract $n$-polyconvex functions.

**Definition 5.26.** Let $f : \mathbb{R}^{d \times D} \to \mathbb{R}$. Then we call $f$ abstract $n$-polyconvex if

$$f(F) = \sup\{h(F) : h \in W^n_F, h \leq f\},$$

for all $F \in \mathbb{R}^{d \times D}$ and where

$$W^n_F = \{h : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\} | h \text{ is } n\text{-polyaffine at } F\}.$$ 

We denote by $\mathcal{F}^n$ the set of all abstract $n$-polyconvex functions.

Note that by defining $W^n_{\mathbb{R}^{d \times D}} = \{W^n_F\}_{F \in \mathbb{R}^{d \times D}}$ we have $\mathcal{F}^n = \mathcal{C}(W^n_{\mathbb{R}^{d \times D}})$. It is not difficult to show that abstract $n$-polyconvex functions are indeed strongly $n$-polyconvex.

Proposition 5.27. Let $f \in \mathcal{F}^n$. Then $f$ is (strongly) $n$-polyconvex. If $f$ is finite and strongly $n$-polyconvex, then $f \in \mathcal{F}^n$.

**Proof.** Let $F \in \mathbb{R}^{d \times D}$ and $F_1, \ldots, F_{r(n)+1} \in \mathcal{C}^n(F)$ such that $F = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i)$. For $f$ to be strongly $n$-polyconvex we have to show that there exists a convex $c_F : \text{co}(T(\mathcal{C}^n(F))) \to \mathbb{R} \cup \{+\infty\}$ such that $f \geq c_F \circ T$ on $\mathcal{C}^n(F)$ and that $f(F) \leq \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i)$. This can be done as follows. First note that for each $h$ that is $n$-polyaffine at $F$ we have $h(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i h(F_i)$. For $\varepsilon > 0$ there exists $h_\varepsilon \in W^n_F$ such that $h_\varepsilon \leq f$ and $f(F) - \varepsilon \leq h_\varepsilon(F)$. Thus,

$$f(F) - \varepsilon \leq h_\varepsilon(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i h_\varepsilon(F_i) \leq \sum_{i=1}^{\tau(n)+1} \lambda_i f(F_i).$$

Since $\varepsilon$ is arbitrary the claim follows. For the lower bound $c$ it is sufficient to choose any of the supporting $n$-polyaffine functions, i.e. $h \in W^n_F$ with $h \leq f$. By Theorem 4.20 for each of these there exists a $\beta \in \mathbb{R}^{\tau(d,D)}$ such that $h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F)$ for all $\tilde{F} \in \mathcal{C}^n(F)$. Thus, choosing $c(X) = \langle \beta, X - T(F) \rangle + h(F)$ for all $X \in \text{co}(T(F + V))$ gives the required lower bound.

For finite $f$ the reverse implication follows by Theorem 4.35. \hfill \square

Remark 5.28. The finiteness of $f$ for the reverse implication of the above proposition is a sufficient condition. In convex analysis a convex function (not necessarily finite) can be written as the pointwise supremum of affine functions if and only if it closed, i.e. when its epigraph is a closed set. Finite convex functions are always closed.
At the present time it is not clear to the author how to set up an analogous abstract convexity theory for sets. However, in light of Proposition 4.53 we believe it is consistent to define abstract \( n \)-polyconvex sets as follows.

**Definition 5.29.** Let \( K \subseteq \mathbb{R}^{d \times D} \). We say that \( K \) is abstract \( n \)-polyconvex if the characteristic function \( \chi_K \) is abstract \( n \)-polyconvex. We denote by \( K^n \) the set of all abstract \( n \)-polyconvex sets in \( \mathbb{R}^{d \times D} \).

Not unexpectedly we find that abstract \( n \)-polyconvex sets are \( n \)-polyconvex.

**Proposition 5.30.** Let \( K \subseteq \mathbb{R}^{d \times D} \) be abstract \( n \)-polyconvex. Then \( K \) is \( n \)-polyconvex.

**Proof.** Let \( K \subseteq \mathbb{R}^{d \times D} \) be abstract \( n \)-polyconvex. Then take any \( F \in \mathbb{R}^{d \times D} \) and a simple rank-\( n \) subspace \( V \subseteq \mathbb{R}^{d \times D} \). We then have to show that whenever \( T(F) = \sum_{i=1}^{\tau(n)+1} \lambda_i T(F_i) \), \( \lambda \in \Lambda_{\tau(n)+1} \) and \( F_i \in F + V \) that then \( F \in K \). Note that \( F \in K \) if and only if \( \chi_K(F) = 0 \). Given that \( \chi_K \) is abstract \( n \)-polyconvex the latter follows immediately and so \( K \) is \( n \)-polyconvex.

### 5.3.2. Abstract \( n \)-polyconvex envelopes and hulls

By analogy with \( n \)-polyconvexity, we define abstract \( n \)-polyconvex envelopes and hulls as follows:

**Definition 5.31.** (i) Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \). Then we denote by \( P_n f \in F^n \) the largest abstract \( n \)-polyconvex function below \( f \).

(ii) Let \( K \subseteq \mathbb{R}^{d \times D} \). Then we denote by \( P_n \cap K \in K^n \) the smallest abstract \( n \)-polyconvex set that contains \( K \).

The notation \( P_n \cap K \) for a set \( K \) was chosen due to its similarity with the intersectional \( n \)-polyconvex hull \( P_n \cap K \).

**Proposition 5.32.** Let \( K \subseteq \mathbb{R}^{d \times D} \). Then

\[
P_n \chi_K = \chi_{P_n \cap K}.
\]

Thus \( P_n \cap K \) is the zero set of the abstract \( n \)-polyconvex envelope \( P_n \chi_K \) of the characteristic function \( \chi_K \).

**Proof.** By definition the set \( P_n \cap K \) is abstract \( n \)-polyconvex and hence, so is the characteristic function \( \chi_{P_n \cap K} \). Since further \( K \subseteq P_n \cap K \) we have \( \chi_{P_n \cap K} \leq \chi_K \) and therefore we must have that \( P_n \chi_K \geq \chi_{P_n \cap K} \).
To show the reverse inequality we define the zero set \( \tilde{K} \) of the function \( \mathcal{P}_n \chi_K \), i.e. \( \tilde{K} = \{ F \in \mathbb{R}^{d \times D} : \mathcal{P}_n \chi_K(F) = 0 \} \). We then prove that \( \chi_{\tilde{K}} = \mathcal{P}_n \chi_K \). If this is true, then \( \tilde{K} \) is abstract \( n \)-polyconvex and since \( K \subseteq \tilde{K} \) we have that \( \mathcal{P}_n K \subseteq \tilde{K} \). Thus \( \chi_{\tilde{K}} \leq \chi_{\mathcal{P}_n K} \) or equivalently \( \mathcal{P}_n \chi_K \leq \chi_{\mathcal{P}_n K} \), which concludes the proof.

It remains to show that \( \chi_{\tilde{K}} = \mathcal{P}_n \chi_K \). It is easy to see that \( \chi_{\tilde{K}} = 0 = \mathcal{P}_n \chi_K \) on \( \tilde{K} \). If \( F \notin \tilde{K} \) we have \( \chi_{\tilde{K}}(F) = +\infty \) and we must show that also \( \mathcal{P}_n \chi_K(F) = +\infty \). This follows if we can show that \( \mathcal{P}_n \chi_K \) only assumes the values 0 or \( +\infty \). Assume otherwise, i.e. there exists \( F \in \mathbb{R}^{d \times D} \) such that \( 0 < \mathcal{P}_n \chi_K(F) < +\infty \). Then consider \( f = 2\mathcal{P}_n \chi_K \). Since abstract \( n \)-polyconvexity is invariant under the multiplication of a positive scalar \( f \) is also abstract \( n \)-polyconvex. Furthermore, it holds that \( f \leq \chi_K \) and \( f(F) > \mathcal{P}_n \chi_K(F) \). Therefore, \( \mathcal{P}_n \chi_K \) cannot be the largest abstract \( n \)-polyconvex function below \( f \), and so we must have \( \mathcal{P}_n \chi_K(F) = +\infty \).

The above proposition is reminiscent of Proposition 4.53. Given the similarity of the two envelope or hull definition the question is if and how they differ and to put it into relation with the functional \( n \)-polyconvex hulls from Definition 4.55. To this end we also define the strong \( n \)-polyconvex envelope \( \mathcal{P}_n f \) of a function \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) as well as the strong intersectional (strong functional) \( n \)-polyconvex hull \( \mathcal{P}_n K (\mathcal{P}_n f) \) of set \( K \subseteq \mathbb{R}^{d \times D} \) in analogy to the usual functional \( n \)-polyconvex envelope and hull.

**Proposition 5.33.** (i) Let \( K \subseteq \mathbb{R}^{d \times D} \). Then

\[
\mathcal{P}_n K \subseteq \mathcal{P}_n^\uparrow K \subseteq \mathcal{P}_n f
\]

and in general \( \mathcal{P}_n K \subseteq \mathcal{P}_n^\uparrow K \subseteq \mathcal{P}_n f \), whereas \( \mathcal{P}_n K \not\subseteq \mathcal{P}_n f \).

(ii) Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \). Then

\[
\mathcal{P}_n f \leq \mathcal{P}_n f
\]

and in general \( \mathcal{P}_n f \leq \mathcal{P}_n f \).

**Proof.** For (i) let \( K \subseteq \mathbb{R}^{d \times D} \). Then by definition \( K \subseteq \mathcal{P}_n^\uparrow K \). Since \( \mathcal{P}_n^\uparrow K \) is \( n \)-polyconvex we obtain \( \mathcal{P}_n K \subseteq \mathcal{P}_n (\mathcal{P}_n^\uparrow K) = \mathcal{P}_n^\uparrow K \). For the second inclusion we will show that whenever \( F \notin \mathcal{P}_n^\uparrow K \) we have that \( F \notin \mathcal{P}_n K \). Note that by Proposition 5.32 it holds that

\[
F \notin \mathcal{P}_n^\uparrow K \iff \chi_{\mathcal{P}_n^\uparrow K}(F) = +\infty \iff \mathcal{P}_n \chi_K(F) = +\infty.
\]
Let $F \not\in \overline{P}_n K$. Then there exists a strongly $n$-polyconvex function $f$ such that $f \leq 0$ on $K$ and $f(F) > 0$. Now denote by $f_k = kf$ where $k \in \mathbb{N}$. Then $f_k$ is finite and strongly $n$-polyconvex and thus abstract $n$-polyconvex by Proposition 5.27 for all $k \in \mathbb{N}$. Furthermore $f_k \leq 0$ on $K$ and therefore $f_k \leq \chi_K$. Hence we also have that $f_k \leq \overline{P}_n \chi_K$ for all $k \in \mathbb{N}$. Since $f_k(F) \to +\infty$ as $k \to \infty$ it follows that $\overline{P}_n \chi_K(F) = +\infty$.

We treat the counterexamples for the specific cases in Example 5.34.

Note that ‘(ii)’ is trivial since any abstract $n$-polyconvex function is also (strongly) $n$-polyconvex. Thus the largest abstract $n$-polyconvex function below $f$ is necessarily smaller or equal to the largest $n$-polyconvex function below $f$.

Example 5.34. (i) We define the set $K_1 = \pm\{ \text{diag}(1, y) : y \in (0, 1] \}$. Then $K_1$ is a rank-one convex set. However, $\chi_{K_1}$ is not abstract rank-one convex, so $K_1$ is not abstract rank-one convex. Instead we have that $\overline{P}_1 \chi_{K_1}(F) = 0$ if and only if $F \in \overline{P}_n K_1 = K_1 \cup \{ \text{diag}(x, 0) : |x| \leq 1 \}$, i.e. it contains the whole line segment joining the open ends of both lines. The set $K_1$ is depicted in Figure 5.1a. In this situation the abstract rank-one convex hull of $K_1$ coincides with the functionally rank-one convex hull $K_1$ (which can be shown by using Šverák’s quasiconvex functions from Section 4.4.3).

(ii) We define the set $K_2 = \pm\{ \text{diag}(x, y) : y = (x - 1), \ x \in (0, 1] \}$. Then $K_2$ is abstract rank-one convex since $\chi_{K_2}$ is abstract rank-one convex. However, $K_2$ is not closed and does not coincide with its functionally rank-one convex hull, which also includes the line segment between the two open ends of the diagonal lines. The set $K_2$ is depicted in Figure 5.1b.

(a) Non-closed intersectionally rank-one set $K_1$ with closed abstract rank-one convex hull. (b) Non-closed abstract rank-one convex set $K_2$.

Figure 5.1.: Examples of sets with closed and nonclosed abstract rank-one convex hull.

The examples give the intuition that the abstract $n$-polyconvex hulls are closed in the directions of the simple rank-$n$ subspaces $V$, however, that outside of those direction
Proposition 5.35. Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) be abstract \( n \)-polyconvex. Then \( f|_{F+V} \) is lower semicontinuous for all \( F \in \mathbb{R}^{d \times D} \) with \( f(F) < \infty \) and \( V \subseteq \mathbb{R}^{d \times D} \) simple rank-\( n \).

Proof. Let \( F \in \mathbb{R}^{d \times D} \) and \( V \) a simple rank-\( n \)-subspace. Further let \( \{F_i\}_i \subseteq F + V \) with \( F_i \to F \) as \( i \to \infty \). Then, since \( f \) is abstract \( n \)-polyconvex it is the pointwise supremum of \( n \)-polyaffines at each point. Thus, at \( F \) there exists \( h_F \) \( n \)-polyaffine at \( F \) such that \( f(F) = h(F) \) and \( h \leq f \). Using the existence of \( \beta \in \mathbb{R}^{\tau(d,D)} \) with \( h(\cdot) = \langle \beta, T(\cdot) - T(F) \rangle + f(F) \) we obtain

\[
f(F) \leq f(F_i) - \langle \beta, T(F_i) - T(F) \rangle.
\]

for all \( F_i \). Taking the lim inf on the right hand side and noting that \( T(F_i) \to T(F) \) we find \( f(F) \leq \liminf_{i \to \infty} f(F_i) \), i.e. \( f \) is lower semicontinuous on \( F + V \) \( \square \).

Remark 5.36. In contrast to this result it is in general not true that \( n \)-polyaffine functions are lower semicontinuous in the directions \( F + V \). Taking the set \( K_1 \) from Example 5.34 and considering \( \chi_{K_1} \) we find that this function is not lower semicontinuous in the vertical direction at the point \( F = \text{diag}(1,0) \).

5.3.3. Legendre-Fenchel transforms for \( n \)-polyconvexity

Having defined abstract \( n \)-polyconvex envelopes and hulls an immediate problem is how these can be computed. This was the motivation of Section 5.2.4 and here we will write down what that means for \( n \)-polyconvexity.

Definition 5.37. Let \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \) and \( F \in \mathbb{R}^{d \times D} \).

(i) Then the function \( f^F : W^n_F \to \mathbb{R} \) defined by

\[
f^F(h_F) = \sup_{\tilde{F} \in \mathbb{R}^{d \times D}} \{ h_F(\tilde{F}) + f(\tilde{F}) \}
\]

is called the \( W^n_F \)-conjugate of \( f \).

(ii) The function \( f^{FF} : \mathbb{R}^{d \times D} \to \mathbb{R} \) defined by

\[
f^{FF}(\tilde{F}) = \sup_{h_F \in W^n_F} \{ h_F(\tilde{F}) + f^F(h_F) \}
\]

is called the \( W^n_F \)-biconjugate of \( f \).
(iii) The function \( f^{F_n,1} : \mathbb{R}^{d \times D} \rightarrow \mathbb{R} \) defined by

\[
f^{F_n,1}(F) = f^{FF}(F)
\]

is called the first \( F_n \)-biconjugate.

(iv) For \( i > 1 \) the function \( f^{F_n,i} : \mathbb{R}^{d \times D} \rightarrow \mathbb{R} \) defined by

\[
f^{F_n,i} = (f^{F_n,i-1})^{F_n,1}
\]

is called the \( i \)-th \( F_n \)-biconjugate.

These definitions correspond to the dualities presented in Section 5.2.4 albeit with a more tailored choice of name. Note that for any of the \( F_n \)-biconjugates we only need to evaluate \( f^{FF} \) at \( F \) although it is theoretically defined on all of \( \mathbb{R}^{d \times D} \). Furthermore it is possible to restrict the space of matrices the supremum is searched in in the \( W_n^{F} \)-conjugate.

**Proposition 5.38.** It is sufficient to consider the \( W_n^{F} \)-conjugate of \( f \) as the supremum over all \( \tilde{F} \in C^n(F) \) instead of all of \( \mathbb{R}^{d \times D} \), i.e.

\[
f^{F}(h_F) = \sup_{\tilde{F} \in C^n(F)} \{ h_F(\tilde{F}) + -f(\tilde{F}) \}.
\]

**Proof.** Let \( h_F \in W_n^{F} \) be any \( n \)-polyaffine function at \( F \). Then define \( \tilde{h}_F \) as \( \tilde{h}_F = h_F \) on \( C^n(F) \) and \( \tilde{h}_F = -\infty \) on \( \mathbb{R}^{d \times D} \setminus C^n(F) \). Clearly \( \tilde{h}_F \leq h_F \). Thus, \( f^{F}(\tilde{h}_F) \leq f^{F}(h_F) \). Therefore, \( \tilde{h}_F(F) + -f^{F}(\tilde{h}_F) \geq h_F(F) + -f^{F}(h_F) \) and so \( \tilde{h}_F \) is a preferable over \( h_F \). Hence, for \( f^{F}(h_F) \) we may always redefine \( h_F \equiv -\infty \) on \( \mathbb{R}^{d \times D} \setminus C^n(F) \) and the supremum will not be affected. \( \Box \)

By Theorem 5.21 and the fact that \( W_n^{F} \) is closed under vertical shifts we have that

\[
f^{F_n,1} = W_n^{\mathbb{R}^{d \times D}-\text{lam}}(f)
\]

and hence that (cf. Theorem 5.17)

\[
\overline{P}_n f = W_n^{\mathbb{R}^{d \times D}-\text{co}} f = \lim_{i \rightarrow \infty} W_n^{\mathbb{R}^{d \times D}-\text{lam}}(f) = \lim_{i \rightarrow \infty} f^{F_n,i}. \tag{5.7}
\]

Comparing the two different envelopes \( P_n f \) and \( \overline{P}_n f \) and their stages of computation (Theorems 4.51 and equation (5.7)) one can clearly see the similarity between the two.
It is tempting to think that their intermediate stages $P_n^k f$ and $f^{F_n,k}$ will be similar, if not the same. Certainly, one can expect that in the extended real-valued case the intermediate steps may differ marginally since we know from Example 5.34 that $P_1 f$ may not be lower semicontinuous on $F + V$ for any $F \in \mathbb{R}^{d \times D}$ and $V \subseteq \mathbb{R}^{d \times D}$ simple rank-1 whereas $f^{F_1,k}$ must be. In fact, even when the abstract $n$-polyconvex envelope does not lower the energy of some points from $+\infty$ to something finite there can still be a substantial difference. The example here is again Example 4.31. Already then we established that the given function is rank-one convex, but cannot be written as the pointwise supremum of rank-one affine functions at 0. We concluded that for any such function we have $f^{F_1,1}(0) = 0 < f(0) = P_1^1(f)(0)$.

5.4. Directional convexity in the abstract convexity framework

In this section we present how a few known convexity notions now fit into the theory of abstract convexity, which was previously impossible without the above generalisation. Such examples are strong $n$-polyconvexity, separate convexity, directional convexity (or $D$-convexity). Separate convexity is a special case of directional convexity and will not be covered on its own, but we want to point out that Tartar remarked \[60\], ‘There is no possible duality theory for separately convex functions, at least in the sense that we understand it for convex functions, as they are not bounded below by simple functions.’ This is obsolete now.

Originally directional convexity for a function is defined as the function being convex along lines parallel to those contained in a set of directions $D$, [40]. This definition does not contain a reference to a function being directionally convex at a point. We include this notion as well as a new definition that we refer to as strong directional convexity.

**Definition 5.39.** Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $x \in \mathbb{R}^n$ and let $D \in \mathbb{R}^n$ be a set of directions.

(i) $f$ is directionally convex at $x$ (or $D$-convex at $x$) if for all $x_1, x_2 \in \mathbb{R}^n$ such that $x_2 - x_1 = s \cdot d$ for some $d \in D$ and $s \in \mathbb{R}$ and $x = \lambda x_1 + (1 - \lambda) x_2$ we have that

$$f(x) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

(ii) $f$ is directionally affine at $x$ (or $D$-affine at $x$) if $f$ and $-f$ are $D$-convex at $x$.

(iii) $f$ is directionally convex (or $D$-convex) if it is $D$-convex at $x$ for every $x \in \mathbb{R}^n$. 

159
(iv) \( f \) is strongly directionally convex at \( x \) (or strongly \( D \)-convex at \( x \)) if for all \( x_1, \ldots, x_n \in x + \mathbb{R}D \) such that \( x = \sum_{i=1}^{n} \lambda_i x_i \) for \( \lambda \in \Lambda_n \) we have
\[
f(x) \leq \sum_{i=1}^{n} \lambda_i f(x_i).
\]

(v) \( f \) is strongly directionally affine at \( x \) (or strongly \( D \)-affine at \( x \)) if \( f \) and \( -f \) are strongly \( D \)-convex at \( x \).

(vi) \( f \) is strongly directionally convex (or \( D \)-convex) if it is \( D \)-convex at \( x \) for every \( x \in \mathbb{R}^n \).

It is clear that our way of defining directionally convex functions as being directionally convex at each point is equivalent to the original definition presented by Matoušek and Plecháč [40, Def. 1.2]. What is not immediately clear is whether strong \( D \)-convexity and \( D \)-convexity, or strong \( D \)-affineness and \( D \)-affineness at a point are different concepts. Example 4.31 is of course a case where a \( D \)-convex function is not strongly \( D \)-convex (choosing \( D = C^1(0) \in \mathbb{R}^{2 \times 2} \)). Now consider \( D = \mathbb{R}^2 \), which generates usual convexity on \( \mathbb{R}^2 \). Then the function \( f : \mathbb{R}^2 \to \mathbb{R} \) such that \( f(x) = 0 \) if \( x \not\in \text{span}\{e_1\} \) and \( f(se_1) = s \) is \( D \)-affine at zero, but it is not strongly \( D \)-affine at zero since \( 0 = f(0) \neq \frac{1}{\sqrt{3}}f([1,0]^T) + \frac{1}{\sqrt{3}}f([-1/2,1]^T) + \frac{1}{\sqrt{3}}f([-1/2,-1]^T) \), but \( 0 = \frac{1}{3}[1,0]^T + \frac{1}{3}[-1/2,1]^T + \frac{1}{3}[-1/2,-1]^T \). A function that is strongly \( D \)-affine at \( x \) for some point \( x \) has to satisfy much stronger conditions. In fact, the next proposition shows that strong \( D \)-affine functions are an affine map on the whole of \( x + D \), meaning that there exists \( \beta \in \mathbb{R}^n \) such that \( f(y) = \langle \beta, y - x \rangle + f(x) \) for all \( y \in x + \mathbb{R}D = \{x + sd : s \in \mathbb{R}, d \in D\} \). In contrast, for just \( D \)-affine functions at \( x \) the gradients of \( f \) in directions of \( D \) through \( x \) are independent of each other as the above function shows.

**Proposition 5.40.** Let \( D \subseteq \mathbb{R}^n \). Further let \( x \in \mathbb{R}^n \) and \( f : \mathbb{R}^n \to \mathbb{R} \) be strongly \( D \)-affine at \( x \) with \( |f(x)| < \infty \). Then there exists \( \beta \in \mathbb{R}^n \) such that
\[
f(y) = \langle \beta, y - x \rangle + f(x).
\]

for all \( y \in x + \text{span} \, D \).

**Proof.** Let \( D \) be as in the proposition and \( f \) be \( D \)-affine at 0 with \( f(0) = 0 \) (the result for arbitrary \( x \) follows by translating \( f \) appropriately). We choose a basis \( x_1, \ldots, x_m \in D \) of \( \text{span} \, D \). It is easy to see that there exists \( \beta \in \mathbb{R}^n \) such that \( f(sx_i) = \langle \beta, sx_i \rangle \)
for all \( s \in \mathbb{R} \) and \( i = 1, \ldots, m \). We now want to show that \( f(y) = \langle \beta, y \rangle \) for all \( y \in \mathbb{R}^n \). Since \( x_1, \ldots, x_m \) are a basis there exist \( s_i \in \mathbb{R} \) such that \( y = \sum_{i=1}^m s_i x_i \). Hence, \( \frac{1}{m+1} y - \sum_{i=1}^m \frac{s_i}{m+1} x_i = 0 \). Therefore, as \( f \) is strongly \( D \)-affine at zero, it must hold that \( 0 = f(0) = \frac{1}{m+1} f(y) + \sum_{i=1}^m \frac{1}{m+1} f(-s_i x_i) \), which can be rearranged to \( f(y) = \sum_{i=1}^m \langle \beta, s_i x_i \rangle = \langle \beta, y \rangle \).

\[ \square \]

**Remark 5.41.** We did not define a notion of strong \( n \)-polyaffine functions at \( F \in \mathbb{R}^{d \times D} \) for \( n \)-polyconvexity. In fact, a function that is \( n \)-polyaffine at a point \( F \) is automatically strongly \( n \)-polyaffine at \( F \) (a function is strongly \( n \)-polyaffine at \( F \) if itself and its negative are strongly \( n \)-polyconvex at \( F \) as of Definition 4.34). This is a consequence of the fact that \( n \)-polyaffine functions are affine on any rank-one line contained in \( C^n(F) \) and not only those which pass through \( F \) itself.

The next theorem shows that \( D \)-convexity is a special case of abstract convexity, by showing that finite \( D \)-convex functions can be written as the supremum of \( D \)-affine functions.

**Theorem 5.42.** Let \( D \subseteq \mathbb{R}^n \) be a set of directions and \( f : \mathbb{R}^n \to \mathbb{R} \), i.e. \( f \) assumes only finite values. Then the following statements are equivalent:

1. \( f \) is \( D \)-convex,
2. for each \( x \in \mathbb{R}^n \) there exists \( \beta \in \mathbb{R}^n \) such that
   \[ f(y) - f(x) \geq \langle \beta, y - x \rangle \]
   for all \( y \in x + D \),
3. \( f \) can be written as the pointwise supremum of strongly \( D \)-affine functions, i.e.
   \[ f(x) = \sup \{ h(x) : h \leq f \text{ on } x + D, \ h \text{ strongly } D \text{-affine} \} , \]
4. \( f \) is strongly \( D \)-convex.

**Proof.** It is easy to see that ‘(ii) \( \Rightarrow \) (iii) \( \Rightarrow \) (iv) \( \Rightarrow \) (i)’. For ‘(i) \( \Rightarrow \) (ii)’ we proceed analogously to the proof of Theorem 4.30 and consider the set \( \text{co}(x+D) = x + \text{co}(D) = x + \text{span}(D) \). Then we define \( g : x + \text{span } D \to \mathbb{R} \) such that

\[ g(X) = \inf \{ \sum_{i=1}^{n+1} \lambda_i f(x_i) : X = \sum_{i=1}^{n+1} x_i, \ x_i \in x + D \} . \]
Note that it would be sufficient to choose \( \dim(\text{span } D) \) elements instead of \( n+1 \) elements in the convex combination of \( X \). It follows that \( g \) is convex and \( g \leq f \) on \( x + \mathbb{R}D \) as for each \( y \in x + \mathbb{R}D \) we must have \( g(y) \leq f(y) \) as \( y \) is a candidate itself. We still have to show that \( g(x) = f(x) \), which follows if we can show that \( f(x) \leq \sum_{i=1}^{n+1} \lambda_i x_i \) whenever \( x_i \in x + \mathbb{R}D \) such that \( x = \sum_{i=1}^{n+1} \lambda_i x_i \) for \( \lambda \in \Lambda \). To do that we simply follow the same method used in Theorem 4.30 by constructing a sequence of approximating nondegenerate \( T_{n+1} \) configuration to the degenerate configuration \( x_1, \ldots, x_{n+1} \). We will omit the details here. Then, since \( g \) is convex, it is subdifferentiable, so there exists \( \beta \in \mathbb{R}^n \) such that \( g(X) - g(x) \geq \langle \beta, X - x \rangle \) for all \( X \in x + \text{span } D \). Using that \( g \leq f \) on \( x + \mathbb{R}D \) and \( g(x) = f(x) \) the claim follows.

Thus, for \( X = \mathbb{R}^n \) we define \( W_x = \{ h : \mathbb{R}^n \to \mathbb{R} : h(y) = \langle \beta, y - x \rangle + h(x) \text{ for all } y \in x + \mathbb{R}D \} \subseteq \mathbb{R}^X \) and then \( W_{\mathbb{R}^n} = \{ W_x \}_{x \in \mathbb{R}^n} \). A finite function is then \( D \)-convex if and only if it is strongly \( D \)-convex or equivalently if and only if it is \( W_{\mathbb{R}^n} \)-convex, making it a special case of our generalised abstract convexity theory.
Chapter 3. In this chapter we investigated the deformations of an annulus $A$ by twisting the outer boundary around the inner boundary a number of times. This was motivated by John [25] as an example of the occurrence of multiple equilibrium solutions in pure displacement boundary problems. Post and Sivaloganathan [48] formalised this suggestion and studied the problem of minimising the functional $I(u) = \int_A \frac{1}{2} |\nabla u|^2 + h(\det \nabla u) \, dx,$ which is a typical energy used in elasticity, among all maps for which the angular map $\psi_\theta(r) = \psi(r, \theta)$ has a winding number of $N$ for $N \in \mathbb{N}$ for every $\theta \in [0, 2\pi)$. Considering the class of rotationally symmetric maps their research has shown that for each $N \in \mathbb{N}$ a rotationally symmetric minimiser among those exists and that it solves the Euler-Lagrange equations. However, in general it is still an open problem whether minimisers solve the Euler-Lagrange equations. This has led Francfort and Sivaloganathan [24] to suggest a slight modification of the problem where then the functional $I^0(u) = \int_A \frac{1}{2} |\nabla u|^2 \, dx$ does not energetically penalise the compression of volume. This modification, in order to stay somewhat physical, then needs to impose the volume constraint $\det \nabla u > 0$ on the set of admissible functions, which, for it to be weakly closed, needs to be relaxed to $\det \nabla u \geq 0$ a.e. on the domain. They argued that minimisers of $I^0$ for $N > 0$ could not solve the Euler-Lagrange equations as they are formally equivalent to Laplace’s equation which only admits a unique solution. We carried out a more detailed analysis of this problem. Like Post and Sivaloganathan we considered rotationally symmetric maps and deriving and solving the energy-momentum equations found explicit solutions for each $N \in \mathbb{N}$. For all $N > 1$ the rotationally symmetric minimisers map what we called the hedgehog region $H$ onto the inner boundary. Thus, on this region the minimisers are degenerate and fail to solve the Euler-Lagrange equations. Furthermore, the solutions are smooth on $H$ and $A \setminus H$, but only differentiable once across the common boundary of $H$ and $A \setminus H$. Having found explicit solutions, which is a rare occasion in nonlinear elasticity, we were able to exploit the structure of the solutions to show that they minimise the energy among a large class of functions (including maps that are not rotationally symmetric). However, it was not possible to fully conclude that minimisers must be rotationally symmetric. To this end we discussed a symmetrisation approach by Sivaloganathan and
Spector \cite{58} which again only gave the partial result that a rotationally symmetric map is energetically favourable over maps for which either the radial or the angular map does not depend on the angle.

With the fresh insights from this modified problem we returned to the one investigated by Post and Sivaloganathan. In particular, although rotationally symmetric minimisers do satisfy the Euler-Lagrange equation, it could still be possible that \( \det \nabla u = 0 \) on a set of measure zero and the previous considerations indicated that \( \dot{\rho}(a) = 0 \) (and hence \( \det \nabla u(a, \theta) = \dot{\rho}(a)\rho(a)/a = 0 \)) could be plausible. Using slightly stricter conditions on the function \( h \) and the techniques of Bauman et al. \cite{13,14} we can show, however, that this is not the case. Furthermore, we show that the function \( d = \det \nabla u \) and \( z = \frac{1}{2} |\nabla u|^2 + f(\det \nabla u) \) with \( f(d) = h'(d)d - h(d) \) are monotonically increasing and decreasing respectively and that a maximum principle for the function \( \rho/r \) holds.

This research contributes to the understanding of twist maps on an annulus, in particular in the case without volume compression where an explicit solution could be found. However, some questions are still open, e.g. whether a (local) minimiser to either of the two considered problems needs to be rotationally symmetric. This is in contrast to the works by Sivaloganathan and Spector \cite{57,58} on an annulus where minimisers are rotationally symmetric, however, in their case no rotations are involved. In a broader view of course the question of uniqueness or nonuniqueness of equilibrium solutions to pure displacement boundary problem when the domain is homeomorphic to a ball still remains open.

Chapter 4. In this chapter we introduced the new semiconvexity called \( n \)-polyconvexity. We proved that for a function \( f : \mathbb{R}^{d \times D} \rightarrow \mathbb{R} \) the notions of \( (d \wedge D) \)-polyconvexity and \( 1 \)-polyconvexity coincide with polyconvexity and rank-one convexity respectively (Corollary 4.11). In this setting (i.e. extended real-valued functions) and for \( d = D \geq 3 \) we also provided examples that are \( 1 \)-polyconvex but not \( 2 \)-polyconvex, and \( 2 \)-polyconvex but not \( 3 \)-polyconvex, thereby showing that the sets of \( n \)-polyconvex functions are distinct for each \( n \leq 3 \). It is easy to see that method of construction for the example can be extended to higher dimension as well so that for all \( 1 < n < d \wedge D \) the notions of \( n \)-polyconvexity are genuinely new. The definition of \( n \)-polyconvexity of a function is based on the definition of \( n \)-polyconvexity of a function at a point \( F \in \mathbb{R}^{d \times D} \), see Definition 4.6. Note that in most cases polyconvexity and rank-one convexity are defined without such a preliminary step, although for instance Ball \cite{7} uses rank-one convexity at a point to define rank-one convexity. An integral part of unifying the concept of polyconvexity, which is convexity on the space of minors, and rank-one convexity, which is convexity
along rank-one directions, was to introduce simple rank-$n$ subspaces of $\mathbb{R}^{d \times D}$. They serve as ‘directions’ to mimic the directional dependence of rank-one convexity as well as spaces onto which we can transfer the convexity property of the function in the sense of polyconvexity. Its analogy to polyconvexity becomes particularly clear with the equivalent condition derived in Theorem 4.10 in terms of a convex representative on each affine space defined through the use of such a simple rank-$n$ subspace. Quasiconvexity, which, under additional assumptions, is necessary and sufficient for the weak lower semicontinuity of an integral functional, lies ‘in between’ poly- and rank-one convexity. Both polyconvexity and directional convexity (which includes rank-one convexity and separate convexity) are studied intensely in the literature (cf. [20] and references therein for polyconvexity or [40, 60] for directional convexity) as means to derive sufficient or necessary conditions (but not both simultaneously) for quasiconvexity. To the best of the authors knowledge the hybrid approach between polyconvexity and directional convexity that we use to define $n$-polyconvexity is the first that creates ‘new’ semiconvexities that lie in between poly- and rank-one convexity other than quasiconvexity. Although studying possible relations of $n$-polyconvexity with quasiconvexity is the main motivation for the new concept this work focussed mainly on establishing basic properties of $n$-polyconvexity (for instance, similarities to rank-one convexity can be seen by generalising the concept of recursive rank-one combinations, see Definition 4.12, and characterising $n$-polyconvexity through so-called recursive $n$-polyconvex combination, see Proposition 4.13).

As part of the basic properties of $n$-polyconvexity we turned our attention towards defining $n$-polyaffine functions where our notion of $n$-polyconvexity at a point is a key ingredient. Instead of defining an $n$-polyaffine function as a function for which both itself and its negative are $n$-polyconvex we did so by using the corresponding versions for a point $F \in \mathbb{R}^{d \times D}$, i.e. a function $f$ is $n$-polyaffine at $F$ if $f$ and $-f$ are $n$-polyconvex at $F$. Whereas the first approach which is usually used in the literature would have led to all $n$-polyaffine functions being equivalent to the usual polyaffine functions, this is not true for the latter. For fixed $F \in \mathbb{R}^{d \times D}$ and each $1 < n \leq d \wedge D$ the set of $n$-polyaffine functions at $F$ is strictly larger than the set of $(n-1)$-polyaffine functions at $F$, and only for $n = d \wedge D$ is it equal to the set of polyaffine functions. This is due to the observation that $n$-polyaffine functions at $F$ can be arbitrary outside the rank-one cone around $F$, whereas they are the restriction of a polyaffine function on the rank-one cone around $F$. Note that therefore in particular 1-polyaffine functions at a point $F \in \mathbb{R}^{d \times D}$ are a much bigger class of functions than rank-one affine functions as known in the literature. This is the result of Theorem 4.20, the proof of which required quite some technical detail which was aided by the new notation $\Lambda_{\mathcal{A}}^F$ and $\Lambda_{\mathcal{A}}^{F,k}$ introduced in (4.23) and (4.31). We believe
that the results of Theorem 4.20 obtained in this manner may not be the most efficient way of doing so and that the map \( \Lambda^F_{A,k} \) has more properties than the three (4.32), (4.33) and (4.34) that were used. In particular we want to hint at one possible variant of formula (4.24), which read

\[
\Lambda^F_{A,k} \left( \emptyset \right) = \sum_{k=0}^{n} \sum_{I_n,k \in (S^n_k)} \sum_{E \in E^k} \left( \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^{n,k} \in (S^k_l)} \Lambda^F_{E,k-l} \left( J_l^{n,k} \right) \right) U_E \left( I_n,k \right).
\]

With the definition of \( \Lambda^F_{A,k} \) we can further reduce notation of this equation, getting rid of the innermost summation to obtain

\[
\Lambda^F_{A,k} \left( \emptyset \right) = \sum_{k=0}^{n} \sum_{I_n,k \in (S^n_k)} \sum_{E \in E^k} \Lambda^F_{E,k-1} \left( S^k \right) U_E \left( I_n,k \right).
\]

Further analysis of the map \( \Lambda^F_{E,k} \) has shown that many of the terms vanish or are the same of negatives of others. (This was exploited in Step 3 in the proof of Theorem 4.20 and is an implication of property (iii) of \( \Lambda^F_{E,k} \)). For instance, those terms \( \Lambda^F_{E,k-1} \left( S^k \right) \) vanish where the \( k \)-tuple \( E = (E_1, \ldots, E_k) \in E^k \) contains any two matrices \( E_i, E_j, i \neq j \) that have their one in the same row or column (i.e. when they are rank-one connected). Hence, we only need to consider \( k \)-tuples \( E \in E^k \) for which, when summed up to \( I_E = \sum_{i=1}^{k} E_i \), the matrix \( I_E \) contains exactly \( k \) ones, each in a different row and column. If we then have a different \( k \)-tuple \( \tilde{E} \in E^k \) with the same property and such that \( I_{\tilde{E}} \) can be obtained from \( I_E \) by successively permuting any two rows or columns of \( I_E \) that contain a one, then we have that

\[
\Lambda^F_{E,k-1} \left( S^k \right) = (-1)^p \Lambda^F_{\tilde{E},k-1} \left( S^k \right),
\]

where \( p \) is the number of permutations needed to get \( I_{\tilde{E}} \) from \( I_E \). Therefore, instead of summing over \( E \in E^k \) in (6.1) we could single out only one representative for each possible selection of \( k \) rows and \( k \) columns and instead obtain a summation over corresponding terms in \( U_E \left( I_n,k \right) \). In a moments time we will write down the precise conjecture to this description, but yet again not without introducing new notation. For a matrix \( U \in \mathbb{R}^{d \times k} \)

166
with column vectors \( u_1, \ldots, u_k \in \mathbb{R}^d \), i.e.

\[
U = \begin{pmatrix} u_1 & u_2 & \ldots & u_k \end{pmatrix},
\]

we define \( \mathrm{adj}_k(U) \), the \( k \)-th adjugate of \( U \), as the vector

\[
[\mathrm{adj}_k(U)]_\alpha = (-1)^{\alpha+1} \det U_{\varphi^k_d(\alpha)},
\]

where \( U_{\varphi^k_d(\alpha)} \) is the square matrix obtained by selecting the rows \( \varphi^k_d(\alpha) \) of \( U \). Note that this corresponds to the standard definition of the \( k \)-th adjugate matrix following Dacorogna [20], see also Definition A.7 with the sole difference that we acknowledge \( \mathrm{adj}_k U \) as a vector in \( \mathbb{R}^\binom{d}{k} \) immediately rather than as a matrix in \( \mathbb{R}^{\binom{d}{k} \times 1} \) (hence the power \( \alpha + 1 \)). Then for given vectors \( u_1, \ldots, u_n \in \mathbb{R}^d \), \( v_1, \ldots, v_n \in \mathbb{R}^D \) and an index set \( I^{n,k} \in \binom{n}{k} \) we define \( \hat{U}(I^{n,k}) \in \mathbb{R}^{D \times k} \) as the matrix containing the \( k \) column vectors \( u_i \) with \( i \in I^{n,k} \) and \( \hat{V}(I^{n,k}) \in \mathbb{R}^{D \times k} \) analogously. Further let for \( \alpha \in \{1, \ldots, \binom{d}{k}\} \) and \( \beta \in \{1, \ldots, \binom{D}{k}\} \) fixed \( \mathcal{E}(\alpha, \beta) = (\mathcal{E}_1(\alpha, \beta), \ldots, \mathcal{E}_k(\alpha, \beta)) \in E^k \) be defined such that \( \mathcal{E}_i(\alpha, \beta) = e_{[\varphi^k_d(\alpha)]_i} \otimes e_{[\varphi^k_d(\beta)]_i} \). With \( \mathcal{E}(\alpha, \beta) \) defined in this manner we have that \( \sum_{i=1}^k \mathcal{E}_i(\alpha, \beta) \) is the identity matrix if the rows \( \varphi^k_d(\alpha) \) and columns \( \varphi^k_d(\beta) \) were selected. Then we claim that

\[
\Lambda_{\emptyset} = \sum_{k=0}^n \sum_{I^{n,k} \in \binom{n}{k}} \sum_{\alpha=1}^n \sum_{\beta=1}^\binom{d}{k} A^{k-1}_{\mathcal{E}(\alpha, \beta)}(S^k) \det U(I^{n,k})_{\alpha} \det V(I^{n,k})_{\beta} = \sum_{k=0}^n \sum_{I^{n,k} \in \binom{n}{k}} \sum_{\alpha=1}^n \sum_{\beta=1}^\binom{D}{k} A^{k-1}_{\mathcal{E}(\alpha, \beta)}(S^k) \det U(I^{n,k}) \otimes \det V(I^{n,k})_{\alpha \beta}.
\]

(6.2)

Although this formula does not look easier than (6.1) it should be noted that in the former we are summing over \( \mathcal{E} \in E^k \), whereas in the latter over \( \alpha \in \{1, \ldots, \binom{d}{k}\} \) and \( \beta \in \{1, \ldots, \binom{D}{k}\} \). The first sum has \( d^k \cdot D^k \) summands (albeit only \( k! \binom{d}{k} \binom{D}{k} \) nonzero ones), whereas the second only has \( \binom{k}{d} \binom{D}{k} \). In the particular case when \( k = n = d = D \) this makes \( d^{2d} \) (or \( d! \) nonzero) summands for summing over \( \mathcal{E} \in E^k \) as opposed to just one for \( \alpha = \beta = 1 \) and thus (6.2) would be a great simplification.

At this stage we proved that our definition of \( n \)-polyaffine functions at a point allows other functions than the usual polyaffine, or equivalently rank-one affine, functions. Recall
that a function $h : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{\pm \infty\}$ is $n$-polyaffine at $F \in \mathbb{R}^{d \times D}$ if and only if there exists $\beta \in \mathbb{R}^n$ such that

$$h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F) \quad (4.17)$$

for all $\tilde{F} \in C^n(F)$. Although this closely resembles the structure of polyaffine functions the major difference is that (4.17) only has to hold for $\tilde{F} \in C^n(F)$ and $h$ could assume any other value outside the rank-$n$ around $F$. For instance it could assume $\pm \infty$ in this region, thereby showing that the set of $n$-polyaffine functions at a point $F \in \mathbb{R}^{d \times D}$ is strictly smaller than the class of $(n-1)$-polyaffine functions at the same $F$ for all $1 < n \leq d \wedge D$.

In particular we obtain 1-polyaffine functions at $F$ that are more versatile than rank-one affine functions. As it is not possible to express rank-one convex but not polyconvex functions as the pointwise supremum of rank-one affine functions it is a natural question to ask whether the new flexibility of 1-polyaffine functions at a point allows us to do so. One of the main results of this work is that this is true, i.e. a finite function is rank-one convex if and only if it can be written as the pointwise supremum of 1-polyaffine functions at each point (see Theorem 4.30), i.e. $f : \mathbb{R}^{d \times D} \to \mathbb{R}$ is rank-one convex if and only if

$$f(F) = \sup \{h(F) : h \text{ is 1-polyaffine at } F \text{ and } h \leq f\}.$$ 

The assumption of finiteness of the function is crucial not only for the proof of the statement using approximating nondegenerate $T_k$ configurations, but also because Example 4.31 poses a counterexample (the example fails to be strongly 1-polyconvex at zero).

In the case of $n$-polyconvexity for $1 < n < d \wedge D$ we needed to make further assumptions to achieve a similar result. To that end we introduced the notion of strong $n$-polyconvexity, see Definition 4.34 and prove that any finite function is strongly $n$-polyconvex if and only if it can be written as the pointwise supremum of $n$-polyaffine functions at each point. However, even for finite functions we are yet unable to conclude that $n$-polyconvexity implies strong $n$-polyconvexity (the reverse implication is trivially true). Note that for finite functions strong 1-polyconvexity, 1-polyconvexity and rank-one convexity are all equivalent. Figure 6.1 summarises the relationships between the various semiconvexity notions. By $\mathbb{P}C^n$ we denote all functions that can be written as the pointwise supremum of $n$-polyaffines, i.e. $\mathbb{P}C^n = \{f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\} : f(F) = \sup \{h(F) : h \in \mathbb{P}A^n(F) \text{ and } h \leq f\}\}$ and abbreviate $\mathbb{P}C = \mathbb{P}C^{d \wedge D}$ and $\mathbb{R}C = \mathbb{P}C^1$. If were to add quasiconvexity to the figure we would include the known relationships of ‘polyconvexity implies quasiconvexity’ in general and ‘quasiconvexity implies rank-one convexity’ in the
finite case, i.e. in the style of Figure 6.1.

\[
\begin{array}{c}
p_{\text{pc}} \\ \text{str. } (d \land D) - p_{\text{pc}} \\
\text{pc} \\ (d \land D) - p_{\text{pc}} \\
\end{array}
\end{array} 
\begin{array}{c}
p_{\text{qc}} \\ \text{str. } (d \land D - 1) - p_{\text{pc}} \\
\text{qc} \\ (d \land D - 1) - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{rc}} \\ \text{str. } 2 - p_{\text{pc}} \\
\text{rc} \\ 2 - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{RC}} \\ \text{str. } 1 - p_{\text{pc}} \\
\text{RC} \\ 1 - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{PC}} \\ \text{str. } (d \land D - 1) - p_{\text{pc}} \\
\text{PC} \\ (d \land D - 1) - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{RC}} \\ \text{str. } 2 - p_{\text{pc}} \\
\text{RC} \\ 2 - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{PC}} \\ \text{str. } 1 - p_{\text{pc}} \\
\text{PC} \\ 1 - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{RC}} \\ \text{str. } 2 - p_{\text{pc}} \\
\text{RC} \\ 2 - p_{\text{pc}} \\
\end{array}
\begin{array}{c}
p_{\text{PC}} \\ \text{str. } 1 - p_{\text{pc}} \\
\text{PC} \\ 1 - p_{\text{pc}} \\
\end{array}
\]

Figure 6.1.: Relationships between the various semiconvexity notions. Grey implications hold for finite functions.

Whereas the implication of quasiconvexity to rank-one convexity does not hold for general extended real-valued functions, see Ball & Murat [9], there exist other subclasses of functions for which it does. For example Conti [19] shows that for functions \( f : \mathbb{R}^{d \times d} \rightarrow \mathbb{R} \cup \{+\infty\} \) such that \( f \) is finite on the set \( \{ F : \det F = 1 \} \) and \(+\infty\) elsewhere quasiconvexity also implies rank-one convexity. (In fact the statement is given in greater generality, but this version will suffice for the point to be made.) It would be interesting to know whether for functions of this kind rank-one convexity also implies strong rank-one convexity. However, the method of proof applied for finite functions does not carry over analogously to this setting as the approximating nondegenerate \( T_k \) configuration leave the set \( \{ F : \det F = 1 \} \). If \( F_i \in C^1(F), i = 1, \ldots, k \) such that \( F = \sum_{i=1}^{k} \lambda_i F_i \) for some \( \lambda \in \Lambda_k \) we have to prove that then \( f(F) \leq \sum_{i=1}^{k} \lambda_i f(F_i) \). We identified \( \{ F_i \}_{i=1}^{k} \) as a degenerate \( T_k \) configuration and used approximating nondegenerate \( T_k \) configurations \( \{ F_i^\varepsilon \} \) to show the inequality with \( F_i^\varepsilon \) substituted for \( F_i \) and then the limit \( \varepsilon \to 0 \) gave the desired result. However, this argument breaks down in the following example, where the approximating \( T_k \) configurations are not a subset of \( \{ F : \det F = 1 \} \) for any \( \varepsilon > 0 \). Let \( F_1 = I + \left( \frac{1}{2} \right) \otimes \left( \frac{1}{2} \right), F_2 = I + 2 \left( \frac{1}{-1} \right) \otimes \left( \frac{1}{-1} \right), F_3 = I + \left( \frac{3}{2} \right) \otimes \left( \frac{1}{2} \right) \) and \( F_4 = I + \left( \frac{0}{-6} \right) \otimes \left( \frac{1}{0} \right) \). Note that \( \det F_i = 1, F_i \in C^1(I) \), none of the \( F_i \) are rank-one connected with each other and \( I = \sum_{i=1}^{4} F_i. \) Then with the same construction we would have \( F_2^\varepsilon = F_2 + \varepsilon \left( \frac{0}{-3} \right) \otimes \left( \frac{3}{-3} \right) \) and it holds that \( \det F_2^\varepsilon > 1 \) for all \( \varepsilon > 0 \). Therefore the limit \( f(F_2^\varepsilon) \) for \( \varepsilon \to 0 \) does not exist. Although this construction does not work it may still be that there exists a different approximating nondegenerate \( T_k \) configuration that lies in \( \{ F : \det F = 1 \} \). On the other hand, if a counterexample can be found, i.e. there
exists a function \( f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\} \) with \( f(F) < +\infty \) if and only if \( \det F = 1 \) such that \( f \) is rank-one convex but not strongly 1-polyconvex, then an important question is whether quasiconvexity does also imply strong 1-polyconvexity as it would then be a stronger notion than rank-one convexity. In turn this could have implications for computing the quasiconvex envelopes functions with an incompressibility constraint as all current methods rely on computing the poly- and rank-one convex envelope and hoping that they are equal (which then also coincide with the quasiconvex envelope). If strong 1-polyconvexity is stronger than rank-one convexity for such functions, but still implied by quasiconvexity, then we should compare the polyconvex and strong 1-polyconvex envelopes instead.

Finally, returning once more to the notion of generalised semiaffine functions at a point \( F \) as we have introduced it for \( n \)-polyconvexity we raise the question whether there exists a definition of quasiaffine functions at a point that is more general than the usual notion of quasiaffine, i.e. polyaffine functions and that additionally allows the characterisation of finite quasiconvex functions as the pointwise supremum of quasiaffine functions at \( F \). If such a notion and characterisation exist then it is clear that the quasiaffine functions at a given \( F \) cannot in general be invariant under transposition of the domain, i.e. there exists a function \( h : \mathbb{R}^{d \times D} \to \mathbb{R} \) that is quasiaffine at a given \( F \in \mathbb{R}^{d \times D} \) but the function \( \hat{h} : \mathbb{R}^{D \times d} \to \mathbb{R} \) defined so that \( \hat{h}(F) = h(F^T) \) is not quasiaffine at \( F^T \). This has to be the case as it is known that quasiconvexity is not invariant under transposition. Müller [45] proved it in the finite valued case, whereas Kružík [36] provided an extended real-valued counterexample. In contrast, \( n \)-polyaffine functions at \( F \) are invariant under transposition as for \( h : \mathbb{R}^{d \times D} \to \mathbb{R} \) \( n \)-polyaffine at \( F \) we have that \( h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F) \) for some \( \beta = (\beta_1, \ldots, \beta_{d \times D}) \in \mathbb{R}^{t(d,D)} \) and for all \( \tilde{F} \in \mathcal{C}^n(F) \). Then for \( \hat{h} \) as above we have that \( \hat{h}(\tilde{F}^T) = h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F) = \langle \hat{\beta}, T(F^T) - T(F^T) \rangle + \hat{h}(F^T) \) for all \( \tilde{F}^T \in \mathcal{C}^n(F^T) \) where \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{d \times D}) \) with \( \hat{\beta}_i = \beta_i^T \). Therefore \( \hat{h} \) is \( n \)-polyaffine at \( F^T \).

Crucial for this equality to hold, although not explicit in the above, is the observation that \( \mathcal{C}^n(F)^T = \mathcal{C}^n(F^T) \). Thus, if for instance quasiaffine functions at \( F \) correspond to the same structure but for a different set, i.e. \( h \) is quasiaffine at \( F \) if and only if \( h(\tilde{F}) = \langle \beta, T(\tilde{F}) - T(F) \rangle + h(F) \) for all \( \tilde{F} \in Q(F) \) where \( Q(F) \) is a set in \( \mathbb{R}^{d \times D} \) then we would expect that \( Q(F^T) \neq Q(F)^T \).

In Section 4.3 we defined \( n \)-polyconvexity for sets. We aligned our definitions with Dacorogna [20] and concluded that a set \( E \) is \( n \)-polyconvex if and only if its characteristic function \( \chi_E \) is \( n \)-polyconvex (see Proposition 4.40). Furthermore we motivated the relevance of \( T_k \) configurations for causing differences in the intersectional and functional \( n \)-polyconvex hulls. We presented an equivalent yet different definition of standard \( T_k \)
configurations to the one used in the literature (cf. [30]), which focusses solely on the additional points rather than the rank-one directions. We then continued with an example of a 2-polyconvex set in $\mathbb{R}^{3\times 3}$ diagonal space that does not contain a $T_k$ configuration, but for which the functional 2-polyconvex hull differs from the set (as discussed in Section 4.4.2). Using the intuition from the given example we defined generalised $T_k$ configuration that account for the same effects of rank-one convexity for $n$-polyconvexity and chose the terminology of $T_k^{n-pc}$ configurations. However, it is not clear whether our choice of $T_k^{n-pc}$ configurations is wide enough or whether more than one auxiliary point for each group $\{A_{ij}\}_j$ should be considered.

Furthermore, we considered $n$-polyconvex envelopes of functions. We defined the envelope operator $P_n f$ in Definition 4.48 and showed that it can be approximated by a limiting process $P_k^n f$ similar to the lamination process that is known for rank-one convexity, see Theorem 4.49. For polyconvexity no limiting process is necessary and it holds that $P_{d\wedge D} f = P_k^n f$ for all $k \geq 0$. Intuitively speaking such a lamination process is necessary since in the computation of $P_k^n f$ for fixed $k$ and a fixed point $F$ we only consider points with difference up to rank $n$ to $F$ and there may exist points that account for a lowering of $P_k^n f(F)$ in comparison to $P_{k-1}^{n-1} f(F)$ that are themselves influenced by points that are outside the rank-$n$ cone around $F$. This is not the case for polyconvexity where the points from the whole space are considered. Similarly, we define the intersectional $n$-polyconvex hull of a set and find that, for instance, the intersectional $n$-polyconvex hull of open sets is open (Theorem 4.51) as well as a duality between it and the $n$-polyconvex envelope of the characteristic function of the set (Proposition 4.53). Moreover we also define the functional $n$-polyconvex hull and obtain, as is known for rank-one convexity and polyconvexity, that the closure of the intersectional $n$-polyconvex hull is contained in the functional $n$-polyconvex hull (Theorem 4.56). Whereas the functional polyconvex hull of a set would at most additionally contain all boundary points of the intersectional polyconvex hull of a compact set, i.e. $P_{d\wedge D}^K = P_{d\wedge D}^l K$ for compact $K$, it is known for rank-one convexity that the functional rank-one convex hull of a set can be much larger. This is the case when a $T_k$ configuration is present. Example 4.45 defines a 2-polyconvex set of eight discrete points that contains no $T_k$ configuration but for which the functional 2-polyconvex set must include additional points. We conjecture that similar configurations exist for all other $2 < n \leq d \wedge D$ as well. Finally, we conjecture that the $n$-polyconvex hull of a set can be obtained as the zero set of the $n$-polyconvex envelope of $n$-distance function of the set, see Conjecture 4.60. While for rank-one convexity and quasiconvexity the first power is sufficient we believe that the $n$-th power of the distance function must be used. This is due to a similar observation for $s$-polyconvexity.
by Šilhavý [53]. Assume 1 < n < d ∧ D. Since a set K is n-polyconvex if and only if 
K ∩ (F + V) is polyconvex for each F ∈ R^{d×D} and simple rank-n subspace V ⊆ R^{d×D}
we turn our attention to F = 0 and V = \{ F ∈ R^{d×D} : F_{ij} = 0 \text{ if } i > n \text{ or } j > n \} only.
Furthermore we restrict dist^n_K to V. It is easy to see that P_n (dist^n_K|_V) ≥ P_n dist^n_K on V.
For P_n (dist^n_K|_V) we want to find to the largest convex function g : co T(V) → R such
that dist^n_K|_V(\tilde{F}) ≥ g ◦ T(\tilde{F}) for all \tilde{F} ∈ V. Since V we have that adj_s = 0 if s > n this
is equivalent to finding the largest convex \tilde{g} : R^{d×D} × R^{(\otimes^2)}(D) × ... × R^{(\otimes^n)}(D) → R such
that \tilde{g} ◦ (\tilde{F}, \text{adj}_2 \tilde{F}, ..., \text{adj}_n \tilde{F}) ≤ dist^n_K|_V(\tilde{F}) for all \tilde{F} ∈ V. Hence we are in the setting
of s-polyconvexity in the sense of Šilhavý and his results tell us that we need to choose
p ≥ n. It remains to investigate whether p = n is sufficient and it is currently not clear
how to proceed from here.

As one of the main motivations of studying n-polyconvexity we investigate whether
or not n-polyconvexity implies quasiconvexity for any 1 ≤ n ≤ d ∧ D or vice versa.
By using Šverák’s quasiconvex functions [65] we can show that quasiconvexity does
not imply 2-polyconvexity for any d, D ≥ 2. This immediately rules out all the other
implications for quasiconvexity to n-polyconvexity for n > 2 as well. Since it is probably
a realistic expectation that n-polyconvexity does not in general imply quasiconvexity for
any n < d ∧ D either, the most efficient way would be to find a (d ∧ D − 1)-polyconvex
function that is not quasiconvex. However, at the present time we do not have any
genuinely (d ∧ D − 1)-polyconvex functions (i.e. they are (d ∧ D − 1)-polyconvex, but not
polyconvex) at our disposal that we could study with respect to quasiconvexity.

The task of finding finite n-polyconvex functions that are not (n + 1)-polyconvex is
not simple. In order to construct such examples we focussed on quadratic functions
only. We derived general conditions that assert strong n-polyconvexity of a quadratic
function in Theorem 4.63. In fact, Serre’s construction, cf. Ball [6], provides a rank-one
convex but not 2-polyconvex quadratic function. Although the method of construction
cannot be used to find a 2-polyconvex but not 3-polyconvex quadratic function it could
provide an example of a 3-polyconvex but not 4-polyconvex quadratic function if a system
of inequalities is can be proven to be contradictory. Note that, in order to conclude
that there exists a quadratic rank-one convex function that is not 2-polyconvex we
exploited the fact that the constructed function has quadratic negative growth outside
the rank-one cone around zero and so a bound by a linear term \langle α, · \rangle is impossible.
However, for 2-polyconvexity the bound \langle α, \text{adj}_2 · \rangle is itself quadratic and it may not be
possible to construct a quadratic function that behaves worse outside the rank-2 cone
than this bound. For this reason we suspect that, in the polynomial case, one must look
for functions with bigger growth. Modifying the example of Alibert, Dacorogna and

172
Marcellini, cf. [20] or Theorem A.6 in Appendix A.2 to functions $f_C : \mathbb{R}^{3 \times 3} \to \mathbb{R}$ such that

$$f_C(F) = |F|^2 \left(|F|^2 - C \text{tr} \text{cof}(F)\right)$$

we suggest to investigate whether there exists $C \in \mathbb{R}$ such that $f_C$ is 2-polyconvex but not polyconvex. A necessary condition is that $f_C$ is polyconvex on the simple rank-2 subspace $V = \text{span}\{e_i \otimes e_j : i, j \leq 2\}$. Denote by $\pi^{2 \times 2}$ the orthogonal projection of $\mathbb{R}^{3 \times 3}$ to $\mathbb{R}^{2 \times 2}$ by deleting the third row and column and by $F^{2 \times 2} = \pi^{2 \times 2}(F)$ for $F \in \mathbb{R}^{3 \times 3}$. Then $f_C(F) = f_\gamma(F^{2 \times 2})$ for all $F \in V$, where $f_\gamma$ is the function from Theorem A.6 with $C = 2\gamma$. From the theorem we know that $|C| \leq 2$ is necessary for $f_C$ to be 2-polyconvex. At least in the range $C \in (1, 2]$ (but possibly not exclusively) $f$ is not polyconvex. This can easily be seen as $f(\text{diag}(x, x, x)) = (1 - C)9x^4$ has negative polynomial growth of degree four if $C > 1$, which cannot be bounded below by a polyaffine function as any such function has polynomial growth of at most three. Therefore, there is a chance that there exist $C \in [-2, 2]$ such that $f_C$ is 2-polyconvex but not polyconvex. Alternatively one could consider the Example 4.45 since we suspect that its functional 2-polyconvex hull is not a polyconvex set. The functional 2-polyconvex hull contains the auxiliary points $J_i$ as well as the dashed curves from $A_{i1}$ through $J_i$ to $A_{i2}$ for each $i$. However, we believe that the polyconvex hull also contains more points around each $J_i$ on the set $\{F : \det F = 1\}$ other than those one this curve but yet such a point still needs to be found and whether or not it would provide an intuition for constructing a 2-polyconvex function that separates it from the original set is questionable.

Finally we showed that $n$-polyconvexity is nonlocal for all $n > 1$. This result was in itself not difficult to prove using Kristensen’s work on the nonlocality of polyconvexity in $\mathbb{R}^{2 \times 2}$ [20] but it causes some of the biggest difficulties for trying to prove Conjecture 4.33 or in trying to answer the question whether or not there exists a global representative $g : \mathbb{R}^{(d,D)} \to \mathbb{R} \cup \{+\infty\}$ such that $g$ is convex on all sets $\text{co} T(F + V)$ for all $F \in \mathbb{R}^{d \times D}$, $V$ simple rank-$n$ and $f = g \circ T$ for any $n$-polyconvex function $f : \mathbb{R}^{d \times D} \to \mathbb{R} \cup \{+\infty\}$. The difficulties arise from the nonconvexity of the sets $T(F + V)$ when $V$ is a simple rank-$n$ subspace with $n > 1$. In order to prove or disprove that a ‘global’ representative $g$ in the above sense exists it is necessary to study possible convex combinations of elements in the intersection of two simple rank-$n$ subspaces, i.e. for $X \in \text{co} T(F + V_1) \cap \text{co} T(F + V_2)$, for $V_1, V_2$ simple rank-$n$. On each coset $F + V_i$ there exists a convex representative $g_{F + V_i}$ as of the proof of Theorem 4.10, i.e. the Busemann representative. However, the question is whether $g_{F + V_1}(X) = g_{F + V_2}(X)$ for $X \in \text{co} T(F + V_1) \cap \text{co} T(F + V_2)$. 

173
We can show that this is false for general simple rank-\(n\) subspaces \(V_1\) and \(V_2\). Consider for example the simple rank-2 subspaces \(V_1 = \text{span}\{e_1 \otimes e_1, e_2 \otimes e_2\} \subseteq \mathbb{R}^{2 \times 2}\) and \(V_2 = \mathbb{R}^{2 \times 2}\). From Lemma 4.14 we can easily conclude that \(\text{co} T(V_1) = \text{diag}(\mathbb{R}^2) \times \mathbb{R}\) and \(\text{co} T(V_2) = \mathbb{R}^{\tau(2,2)}\). Now we fix \(X = (0, 1) \in \mathbb{R}^{\tau(2,2)}\). Note that \(X \in \text{co} T(V_1) \cap \text{co} T(V_2)\) with \(X = \frac{1}{2} T(F_1) + \frac{1}{2} T(F_2)\) for \(F_1 = \left( \begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix} \right)\) and \(F_2 = \left( \begin{smallmatrix} -2 & -1 \\ -1 & -1 \end{smallmatrix} \right)\). Further we define \(L = \text{aff}\{(T(F_1), T(F_2))\}\), where \(\text{aff} S\) denotes the affine hull of a set \(S\), i.e. the smallest affine space that contains \(S\). Then \(f : \mathbb{R}^{2 \times 2} \to \mathbb{R}\) defined so that \(f(F) = \text{dist}(T(F), L)\) is a polyconvex function since \(\text{dist}(-, L)\) is a convex representative of \(f\) on \(\mathbb{R}^{\tau(2,2)}\). It is easy to see that for the Busemann representative \(g_{V_2}\) on \(\text{co} T(V_2)\) it holds that \(g_{V_2}(X) = 0\).

With another simple calculation we can show that there exists \(c > 0\) such that \(f \geq c\) on \(V_1\) and this implies \(g_{V_1}(X) \geq c > g_{V_2}(X)\). Note that both the subspace \(V_1\) is not a ‘full’ simple rank-\(n\) subspace in the sense that one can add another basis element (say \(e_1 \otimes e_2\)) and the new space is still a simple rank-\(n\) subspace. Therefore, there is still the possibility that the Busemann representatives coincide for such ‘full’ simple rank-\(n\) subspaces and this needs further investigation.

**Chapter 5.** Abstract convexity has been developed from the realisation that the concept from convex analysis, namely that convex functions or sets can be characterised through simpler objects like affine functions or halfspaces, does not rely in particular on affine functions or halfspaces. Instead other ‘elementary’ building blocks may be used. Stripped to the bare minimum abstract convexity only requires a subset \(\mathcal{M} \subseteq E\) (\(E\) a lattice), which acts as the ‘elementary’ building blocks. The set \(\mathcal{M}\) then generates \(\mathcal{M}\)-convexity as described in Section 5.1. Both standard convexity and polyconvexity fit into this framework where in the case of functions it is custom to write \(W \subseteq E\) as the subset of ‘elementary’ functions instead of \(\mathcal{M}\). In the case of convexity or polyconvexity we choose \(W\) as the set of affine functions or polyaffine functions respectively. However, it is known that rank-one convexity or separate convexity (both directional convexities) do not fit into this framework as rank-one convex (or separately convex) functions need not be bounded below by simple or ‘elementary’ functions (see Tartar’s remark in the beginning of Section 5.4). On the other hand, we showed in Chapter 4 that finite rank-one convex functions (strong \(n\)-polyconvex functions) can be written as the pointwise supremum of \(1\)-polyaffine functions (\(n\)-polyaffine functions) at every point, which is somewhat reminiscent of abstract convexity methods. This shows that abstract convexity needs to be generalised further in order to include this new kind of semiconvexity. In this chapter we therefore pursued this generalisation for functions, i.e. we have the lattice \(E = \mathbb{R}^X\), by allowing the set of ‘elementary’ functions \(W \subseteq \mathbb{R}^X\) to depend on \(x \in X\),
i.e. \( W_x \subseteq \bar{\mathbb{R}}^X \) for each \( x \in X \). With such a choice of a family of subsets \( W_X = \{ W_x \}_{x \in X} \) we defined \( W_X \)-convexity. In doing so we have created a greater overlap between the semiconvexities that are typically studied within the calculus of variations and those of abstract convexity as the latter is now able to include rank-one convexity.

While concepts of the \( W_X \)-convex hull operator can still be defined analogously to classical abstract convexity, it is not true any more that the \( W_X \)-convex hull of a function can be obtained by the pointwise supremum of ‘elementary’ functions below the given one, i.e. in general we do not have that \((W_X \text{-co } f)(x) = \sup \{ w_x(x) : w_x \in W_x, w_x \leq f \}\). This is due to the fact that in general the elementary functions \( w_x \in W_x \) for some \( x \in X \) need not be \( W_X \)-convex. Nevertheless, we were able to give sense to the right hand side of the above equation by denoting it as the so-called laminate operator \( W_X \text{-lam} \) and were able to show that \( W_X \text{-co } f = \lim_{i \to +\infty} W_X \text{-lam}^i f \). This was largely inspired by analogous effects known for the computation of the rank-one convex envelope of a function as of Section 4.4.1 where potentially infinitely many lamination steps are necessary to converge to the rank-one convex envelope.

In Section 5.2.4 we investigated conjugations for \( W_X \) convexity. We were able to define the operator \( c_X^* X c_X \) and show that it corresponds to the laminate operator \( W_X \text{-lam} \). Consequently, \( c_X^* \) cannot in general be the dual \( c_X^* \) of \( c_X \) as \( c_X^* c_X \) needs to be a hull operator. Therefore, it would be desirable to have a conjugation \( c \) that decomposes the \( W_X \)-convex hull (provided \( W_x \) is closed under vertical shifts for all \( x \in X \)). We showed with \( c_0 \) that such a conjugation always exists. However, it is not very useful as the target space \( (C(W_X) \to \bar{\mathbb{R}}) \) of \( c_0 \) is vast. It is known that in classical abstract convexity this space can be reduced to just \((W \to \bar{\mathbb{R}})\) and thus we tried to define the conjugation \( c_X \) that reduces the target space in a similar way. At the end of this section we made an attempt to define a conjugation \( c_X \) such that \( c_X^* c_X = W_X \text{-co} \) but with a smaller target space than \((C(W_X) \to \bar{\mathbb{R}})\), but Example 5.25 shows that the choice is not suitable. This leaves the question whether a more useful definition of \( c_X \) can be found such that \( c_X^* c_X = W_X \text{-co} \) as it is the case for classical abstract convexity.

In Section 5.3 we considered the special case of abstract \( n \)-polyconvexity which is a natural application of our generalised theory of Abstract Convex in view of Chapter 4. Based on the equivalence between the \( n \)-polyconvexity of a set and the \( n \)-polyconvexity of its characteristic function we defined abstract \( n \)-polyconvex sets as those for which their characteristic function is abstract \( n \)-polyconvex. As expected, the abstract \( n \)-polyconvex hull of a set corresponds to the zero set of the abstract \( n \)-polyconvex envelope of the sets characteristic function as of Proposition 5.32. Furthermore we show that the abstract \( n \)-polyconvex hull is an intermediate concept between the intersectional and functional
(strong) $n$-polyconvex hull. In contrast to the functional $n$-polyconvex hull the abstract $n$-polyconvex is not able to ‘recognise’ $\mathcal{T}_k^{n\text{-pc}}$ configurations. For example, the classical $T_4$ configuration is an abstract but not functionally 1-polyconvex. On the other hand, the abstract $n$-polyconvex envelope is stronger than the $n$-polyconvex envelope in the sense that abstract $n$-polyconvex envelopes are lower semicontinuous on all sets $F + V$ for $F \in \mathbb{R}^{d\times D}$ and $V \subseteq \mathbb{R}^{d\times D}$ simple rank-$n$ (see Proposition 5.35). Note, however, that this difference only occurs in the extended real-valued case where it is easy to see that the lower semicontinuity of a function is necessary for the weak lower semicontinuity of its integral functional. For extended real-valued polyconvex functions this is commonly assured by requiring continuity of the function (in [20] this is masked under the term Carathéodory function), but abstract polyconvexity could be sufficient. Similarly we would argue that abstract 1-polyconvexity of an extended real-valued function is at least necessary for the weak lower semicontinuity of its integral functional, however, further research is necessary in this direction.

At the end of Section 5.3.3 we suggested an alternative approach to computing the rank-one convex envelope of a function. Instead of computing $P_i f$ we compute the $i$-th $\mathcal{F}^n$-biconjugate at $F \in \mathbb{R}^{d\times D}$ for each $F$. Irrespective of the clear differences between $f^{F^0,1}$ and $P_1^1$ in the extended real-valued case there may be a difference in performance of numerical implementations of both methods. Dolzmann [22] suggested a numerical algorithm for the rank-one convexification of a function that coincides with its rank-one convex envelope outside a given ball and which works on a discrete mesh $M$ of points and a discrete set of rank-one directions $D$. Without going into the details of the algorithm it can be described as follows.

1. Initialise $f^0 = f$ and $i := 0$;
2. $g = f^i$;
3. for all points $F \in M$ for all $d \in D$
   compute $g := \text{convexify}(g, F, d, M)$;
4. $f^{i+1} = g$;
5. if $\|f^{i+1} - f^i\|_{\infty} > \varepsilon$ go to 2; else stop;

where convexify($g, F, d, M$) computes the convexification of $g|_{(F + \mathbb{R}d)^\cap M}$ on the line $F + \mathbb{R}d = \{F + td : t \in \mathbb{R}\}$ intersected with the mesh $M$. Note that $f^i$ as computed in the algorithm does not necessarily correspond to $P^i f$, but both share the property that they are only working in rank-one directions. An analogue straightforward implementation of our proposed approach would change Step 3 to

3’. for all points $F \in M$ compute $g(F) := g^{FF}(F)$;
Note that in $3$ we now no longer need to go through all rank-one directions of the set $D$. However, the trade-off is that the problem of computing $g^{FF}(F)$ is higher dimensional than convexify$(g, F, d, M)$ and it would be interesting to see which of the two approaches is faster.

Last but not least we showed that directional convexity can be regarded a special case of the generalised abstract convexity theory. We defined the new concepts of directionally affine functions at a point and strongly directionally affine functions at a point. For finite functions there is a one-to-one correspondence between directionally convex functions and those that can be written as the pointwise supremum of strongly directionally affine functions. As for the case of $n$-polyconvexity it may well be that with this new viewpoint of directional convexity a number of new insights may be gained.
A. More on \(n\)-polyconvexity

A.1. Useful definitions and theorems

From convex analysis we need the following:

**Definition A.1.** Let \(k \in \mathbb{N}\). Then we define \(\Lambda_k = \{ \lambda \in \mathbb{R}^k : \lambda_i \geq 0, i = 1, \ldots, k, \sum_{i=1}^{k} \lambda_i = 1 \} \).

**Theorem A.2** (cf. Thm. 2.10, [20]). Let \(E \subseteq \mathbb{R}^n\) be closed and convex and \(x \in \partial E\). Then there exists \(\alpha \in \mathbb{R}^n, \alpha \neq 0\) such that
\[
\langle x, \alpha \rangle \leq \langle y, \alpha \rangle
\]
for all \(y \in E\).

**Theorem A.3** (cf. Thm. 2.13, [20]). Let \(E \subseteq \mathbb{R}^n\). Then
\[
\text{co } E = \{ x \in \mathbb{R}^n : x = \sum_{i=1}^{n+1} \lambda_i x_i, x_i \in E, \lambda \in \Lambda_{n+1} \}.
\]

A.2. Relations between the established semiconvexity conditions

In Section 2.2.1 we already introduced the concepts of quasiconvexity and polyconvexity. Although we did not go into it in any further detail, the construction of infimising sequences for the cases where no minimum exists and microstructure develops involves another type of convexity called rank one convexity. In the one-dimensional example from Section 2.3 we have chosen the infimising sequence so that the sections with different gradients, called laminates, are connected continuously (which is necessary as the sequence would otherwise not be in \(W^{1,4}(0,1)\)). In a higher-dimensional setting, the laminates also
have to be continuous across their interface. A necessary condition for that being possible is that \( \text{rank}(B - A) \leq 1 \) which is equivalent to the existence of \( a \in \mathbb{R} \) and \( n \in SO(3) \) s.t.

\[
B - A = a \otimes n,
\]

where \( n \) is the normal to the hyperplane where the two gradients meet. Figure A.1 shows an element of such an infimising sequence for \( \Omega \subseteq \mathbb{R}^2 \) and with the two gradients \( A, B \) being the gradients with zero energy. In the limit this lamination approximates the boundary layer.

Figure A.1: \( k \)-th element of an infimising sequence with rank one connected laminates \( A \) and \( B \) with occupying a fraction \( \lambda \) and \( 1 - \lambda \) of the domain respectively. With increasing fineness of the lamination the boundary layer thickness decreases \([44]\).

The gradient \( F = \lambda A + (1 - \lambda)B \) with \( \lambda \in [0, 1] \), which would be the gradient of the generalised solution of the relaxed problem in Theorem 2.21. The corresponding gradient Young measure accounting for the microstructure is \( \nu = \lambda \delta_A + (1 - \lambda)\delta_B \), see Theorem 2.20. The laminate pictured in Fig A.1 is called a single laminate. It is also possible to have laminates within laminates, which are then called second (for one laminate in a laminate) or higher (for more than two lamination steps) laminates. The idea of the relaxed problem is that any gradient \( F \) on the rank one line between the zero energy gradients \( A \) and \( B \) can be realised with zero energy as well. This motivates the following definition:

**Definition A.4.** A function \( f : \mathbb{R}^{n \times m} \to \mathbb{R} \) is said to be rank one convex if

\[
f(\lambda F + (1 - \lambda)G) \leq \lambda f(F) + (1 - \lambda)f(G)
\]
for every $\lambda \in [0,1]$, $F, G \in \mathbb{R}^{n \times m}$ with $\text{rank}(F - G) \leq 1$.

The the following relations between the different types of convexities can be obtained:

**Theorem A.5 ([20] Theorem 5.3]).**

(i) Let $f : \mathbb{R}^{n \times m} \to \mathbb{R}$. Then

$f \text{ convex} \Rightarrow f \text{ polyconvex} \Rightarrow f \text{ quasiconvex} \Rightarrow f \text{ rank one convex}.$

(ii) If $f : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{+\infty\}$, then

$f \text{ convex} \Rightarrow f \text{ polyconvex} \Rightarrow f \text{ rank one convex}.$

(iii) If $n = 1$ or $m = 1$, then all these notions are equivalent.

Note that similarly to swlsc in the extended-real valued case the convexity relations for quasiconvexity in the extended-real valued case are not clear. In such a setting it is known, see [9], that quasiconvexity does not imply rank-one convexity. Furthermore, the reverse implications do not generally hold. For example the following theorem holds:

**Theorem A.6 ([20] Theorem 5.51]).** Let $\gamma \in \mathbb{R}$ and let $f_\gamma : \mathbb{R}^{2 \times 2} \to \mathbb{R}$ be defined as

$$f_\gamma(F) = |F|^2(|F|^2 - 2\gamma \det F).$$

Then

- $f_\gamma$ is convex $\iff |\gamma| \leq \gamma_c = \frac{2}{3} \sqrt{2}$
- $f_\gamma$ is polyconvex $\iff |\gamma| \leq \gamma_p = 1$
- $f_\gamma$ is quasiconvex $\iff |\gamma| \leq \gamma_q > 1$ and $\gamma_q > 1$
- $f_\gamma$ is rank one convex $\iff |\gamma| \leq \gamma_r = \frac{2}{\sqrt{3}}$.

It is currently not known whether $\gamma_q < \gamma_r$. More generally, it can be shown that rank one convexity does not imply quasiconvexity if $n \geq 3$ as the famous example from Šverák [66] shows. The case $n = 2, m \geq 2$ is still open.
A.3. The polyconvex hull of the $T_4$ configuration

In Section 4.1 we consider the set of points $\{B_i\}$, $i = 1, \ldots, 4$ which, when the points are translated, rescaled and the irrelevant coordinate is removed, correspond to

$$B_1 = (-1, 3), \quad B_2 = (3, 1), \quad B_3 = (1, -3), \quad B_4 = (-3, -1).$$

These points define the diagonal matrices $A_1$ through $A_4 = \text{diag}(B_i)$. It is well known that the polyconvex hull of these points is as sketched in Fig. 4.3, however, no proof of this was found. Here we prove that this depiction is correct. Recall that the polyconvex hull $K_{pc}$ of a set $K \subseteq \mathbb{R}^{d \times D}$ can be computed via the projection $\pi$ of all minors included in the convex hull of the set $T(K)$ in $\mathbb{R}^{\tau(d,D)}$ onto its first $d \times D$ components, i.e. $K_{pc} = \pi(\text{co}(T(K)) \cap T(\mathbb{R}^{d \times D})) = \{F \in \mathbb{R}^{d \times D} : T(F) \in \text{co}(T(K))\}$, cf. [20, Thm. 7.13]. Thus, we have to find the intersection of the $C := \text{co}(\{T(A_i)\}_{i=1,\ldots,4})$ with the minors manifold $T(\mathbb{R}^{2 \times 2})$. Since the $A_i$ are diagonal matrices we may also only consider the minors manifold of the diagonal matrices $T(\text{diag}(\mathbb{R}^{2 \times 2}))$. For a diagonal matrix $A = \text{diag}(x, y)$ we have that $T(A) = (A, \text{adj}_2(A)) = (\text{diag}(x, y), xy)$ which we will identify with the vector $(x, y, xy) \in \mathbb{R}^3$. Hence, the set $C$ is a subset of $\{(x, y, d) : x, y, d \in \mathbb{R}\} = \mathbb{R}^3$ and we will continue to identify the matrices $A_i$ with $B_i$ in $\mathbb{R}^2$. Then $C$ is the polyhedron defined by the points $T(B_1) = (-1, 3, -3)$, $T(B_2) = (3, 1, 3)$, $T(B_3) = (1, -3, -3)$ and $T(B_4) = (-3, -1, 3)$ which contains the origin. The task is to find those $x, y$ for which $(x, y, xy)$ lies in the inside of this polyhedron. Therefore we first find the intersection of the manifold $\{(x, y, xy) : x, y \in \mathbb{R}\}$ with each face of $C$. For example, the face $F$ defined by $T(B_2)$, $T(B_3)$ and $T(B_4)$ can be parametrised in $x, y$ by

$$F = \left\{(x, y, d) \in \mathbb{R}^3 : d = -\frac{3}{5}x + \frac{9}{5}y + 3, \ x, y \in \mathbb{R}\right\}.$$

Setting $d = xy$ allows us to find the desired intersection, which is given by the set $\{(x, y) : y = \frac{15 - 3x}{5}\}$. The situation is drawn in Figure A.2. The area between the two curves defines all points $(x, y)$ for which $(x, y, xy)$ lie on the same side of the face $F$ as the origin, which is the side that contains the polyhedron. Repeating this for the other three faces completes the polyconvex hull as depicted in Figure 4.3.

A.4. Determinants and the minors map $T$

For polyconvexity or $n$-polyconvexity we use the minors map $T : \mathbb{R}^{d \times D} \to \mathbb{R}^{\tau(d,D)}$ with $T(F) = (F, \text{adj}_2 F, \ldots, \text{adj}_{d \land D} F)$. Here we properly define the adjugate matrices $\text{adj}_s F$
Figure A.2.: Graph of \((x, y)\) points for which \((x, y, xy)\) intersects with the face defined by \(T(B_2), T(B_3)\) and \(T(B_4)\). The thick black curve represents the part of this intersection that lies inside the convex hull of the points \(B_i\) (dashed box).

For \(s = 1, \ldots, d \wedge D\), see [20, Sec. 5.4]. To this end, for \(d, s \in \mathbb{N}\) with \(s \leq d\), let

\[
I^d_s = \{(\alpha_1, \ldots, \alpha_s) \in \mathbb{N}^s : 1 \leq \alpha_1 < \alpha_2 < \ldots < \alpha_s \leq n\}.
\]

Then we define the map \(\varphi^d_s : \{1, \ldots, \binom{d}{s}\} \to I^d_s\) as the only map that respects the backwards lexicographic order when read backwards on \(I^d_s\). For example, for \(d = 4\) and \(s = 2\) we have

\[
\varphi^4_2(1) = (3, 4), \quad \varphi^4_2(2) = (2, 4), \quad \varphi^4_2(3) = (1, 4)
\]
\[
\varphi^4_2(4) = (2, 3), \quad \varphi^4_2(5) = (1, 3), \quad \varphi^4_2(6) = (1, 2).
\]

For more details and examples refer to [20] Sec. 5.4. Then we define:

**Definition A.7.** Let \(F \in \mathbb{R}^{d \times D}\) and \(s \in \mathbb{N}\) with \(1 \leq s \leq d \wedge D\). Then \(\text{adj}_s F \in \mathbb{R}^{(\binom{d}{s}) \times (\binom{D}{s})}\) is the matrix defined by

\[
[\text{adj}_s F]_{\alpha \beta} = (-1)^{\alpha + \beta} \det \left( M_{\varphi^d_s(\alpha), \varphi^D_s(\beta)}(F) \right),
\]

for \(s = 1, \ldots, d \wedge D\), see [20, Sec. 5.4]. To this end, for \(d, s \in \mathbb{N}\) with \(s \leq d\), let
where $M_{\varphi_s^\alpha, \varphi_s^\beta}(F)$ corresponds to the $s \times s$-matrix obtained from $F$ by selecting the entries of $F$ common in the rows $\varphi_s^\alpha(\alpha)$ and columns $\varphi_s^D(\beta)$ of $F$.

In Section 4.2 we use the following result.

**Lemma A.8.** Let $F \in \mathbb{R}^{d \times D}$, $F_i \in \mathbb{R}^{d \times D}$, $i = 1, \ldots, n$ and $\lambda \in \Lambda_n$ such that

$$T(F) = \sum_{i=1}^{n} \lambda_i T(F_i).$$

Further let $A \in \mathbb{R}^{d \times D}$. Then

$$T(F + A) = \sum_{i=1}^{n} \lambda_i T(F_i + A). \quad (A.1)$$

This is a well-known property of the minors map $T$, however, we were unable to find a reference for it.

**Proof.** We will prove the validity of (A.1) for $A = a \otimes b$ only, i.e. for rank-one matrices. Repeated use of this then proves the result for general $A$.

Let $F, F_i$ and $\lambda$ as above. Proving equation (A.1) is equivalent to showing $\text{adj}_s(F + A) = \sum_{i=1}^{n} \lambda_i \text{adj}_s(F_i + A)$ or doing so for each entry of the latter. Note that each entry of $\text{adj}_s$ is a determinant of a $(s \times s)$ submatrix of its argument, i.e. $F + A$ or $F_i + A$. As a representative we choose the bottom right entry of $\text{adj}_s$, that is $[\text{adj}_s(\cdot)]_{\alpha \beta}$ with $\alpha = (s^\alpha)$, $\beta = (s^D)$. This entry contains the determinant of the $(s \times s)$ submatrix of a matrix by choosing the first $s$ rows and columns. Let $\xi$ be the matrix of the first $s$ rows and columns of $F$, $\xi_i$ of $F_i$ and $\eta$ of $A$ respectively. We then want to show that

$$\det(\xi + \eta) = \sum_{i=1}^{n} \lambda_i \det(\xi_i + \eta).$$

Note that $\eta$ is a rank-one matrix as it is obtain from the rank-one matrix $A$. We now use a result about determinants of sums, cf. [?, Prop. 5.67]. Denoting by the couple $(I, J)$ with $I$ and $J$ sets of ordered row indices such that $I \cup J = \{1, \ldots, s\}$ and $I \cap J = \emptyset$ we define for a two given matrices $\hat{\xi}, \hat{\eta} \in \mathbb{R}^{s \times s}$ the matrix

$$((\hat{\xi}^I, \hat{\eta}^J) \in \mathbb{R}^{s \times s}$$

such that the $k$-th row of $(\hat{\xi}^I, \hat{\eta}^J)$ is equal to the $k$-th row of $\hat{\xi}$ if $k \in I$ or the $k$-th row of $\hat{\eta}$ if $k \in J$. Then the result asserts that

$$\det(\hat{\xi} + \hat{\eta}) = \sum_{(I, J)} \det((\hat{\xi}^I, \hat{\eta}^J)).$$

\[184\]
Note that applying the above to the matrices $\xi$ and $\eta$, or $\xi_i$ and $\eta_i$ we are left with only
\[ \det(\xi + \eta) = \sum_{j=1}^{s} \det(\xi^j) \]
where $\xi^j$ is the matrix $\xi$ with the $j$-th row replaced with the $j$-th row of $\eta$ (and similarly for $\xi_i$). This is due to $\eta$ being rank-one convex and whenever we choose more than two rows of $\eta$ we have two linear dependent rows and hence the determinant vanishes. Denote by $\eta^j$ the $j$-th row of $\eta$. Then we can further evolve each determinant of the right hand side in the above equation. We obtain
\[ \det(\xi + \eta) = \sum_{j=1}^{s} \sum_{k=1}^{s} (-1)^{j+k} \eta_{jk} \det(\xi^j_k) \]
where $\xi^j_k$ denotes the matrix obtain from $\xi^j$ by deleting the $j$-th row and $k$-th column.
Note that $\xi^j_k$ is a $(s-1) \times (s-1)$ submatrix of $\xi$. Doing the above for $\xi_i$ and from the assumption that $\det(\xi^j_k) = \sum_{i=1}^{n} \lambda_i \det((\xi_i)^j_k)$ we find
\[ \sum_{i=1}^{n} \lambda_i \det(\xi + \eta) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{s} \sum_{k=1}^{s} (-1)^{j+k} \eta_{jk} \sum_{i=1}^{n} \lambda_i \det((\xi_i)^j_k) = \sum_{j=1}^{s} \sum_{k=1}^{s} (-1)^{j+k} \eta_{jk} \det(\xi^j_k) = \det(\xi + \eta), \]
which concludes the proof.

Furthermore, in the proof of Theorem 4.63 we need that for fixed $F \in \mathbb{R}^{d \times D}$ and $\beta \in \mathbb{R}^{\tau(d,D)}$ we can find $\tilde{\beta} \in \mathbb{R}^{\tau(d,D)}$ so that $\langle \beta, T(\tilde{F} - F) \rangle = \langle \tilde{\beta}, T(\tilde{F}) - T(F) \rangle$ for all $\tilde{F} \in \mathbb{R}^{d \times D}$, see (4.61). We prove the existence of such $\tilde{\beta}$ in the following lemma.

**Lemma A.9.** Let $F \in \mathbb{R}^{d \times D}$ and $\beta \in \mathbb{R}^{\tau(d,D)}$. Then there exists $\tilde{\beta} \in \mathbb{R}^{\tau(d,D)}$ such that
\[ \langle \beta, T(\tilde{F} - F) \rangle = \langle \tilde{\beta}, T(\tilde{F}) - T(F) \rangle \quad (A.2) \]
for all $\tilde{F} \in \mathbb{R}^{d \times D}$.

**Proof.** For $\tilde{F}$ we write $F + \sum_{x=1}^{d \wedge D} u^x \otimes v^x$. Then (A.2) for all $\tilde{F} \in \mathbb{R}^{d \times D}$ is equivalent to
\[ \langle \beta, T(\sum_{x=1}^{d \wedge D} u^x \otimes v^x) \rangle = \langle \tilde{\beta}, T(F + \sum_{x=1}^{d \wedge D} u^x \otimes v^x) - T(F) \rangle \] for all \( u^x \in \mathbb{R}^d \) and \( v^x \in \mathbb{R}^D \), \( x = 1, \ldots, d \wedge D \). Using the expansion (4.24) for \( \gamma = T \) we find that it must hold that

\[
\sum_{k=1}^{d \wedge D} \sum_{l \in \binom{[d \wedge D]}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^k \in \binom{E^k}{l}} \langle \beta, T^0_x(J_l^k) \rangle \right] U_E(I^k) = \sum_{k=1}^{d \wedge D} \sum_{l \in \binom{[d \wedge D]}{k}} \sum_{E \in E^k} \left[ \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^k \in \binom{E^k}{l}} \langle \tilde{\beta}, T^F_x(J_l^k) \rangle \right] U_E(I^k)
\]

for all \( u^x \in \mathbb{R}^d \) and \( v^x \in \mathbb{R}^D \), \( x = 1, \ldots, d \wedge D \). Note that the sole difference between the terms on the left and right side of the equality is \( T^0_x \) and \( T^F_x \) (and \( \beta, \tilde{\beta} \)). With similar observations to those in the proof of Theorem 4.20 we find that this is equivalent to the existence of \( \tilde{\beta} \in \mathbb{R}^{\tau(d,D)} \) such that

\[
\sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^k \in \binom{E^k}{l}} \langle \beta, T^0_x(J_l^k) \rangle = \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^k \in \binom{E^k}{l}} \langle \tilde{\beta}, T^F_x(J_l^k) \rangle
\]

for all \( E \in E^k, k = 1, \ldots, d \wedge D \), which is a very large linear system of equalities. However, as we have seen, the properties of \( \Lambda_x^E \) from the proof of Theorem 4.20 imply that many of these are zero, equal or negatives of each other. In fact, without providing too much detail, the linear system can be reduced to \( \tau(d,D) \times \tau(d,D) \) equations only, with \( A\beta = B\tilde{\beta} \), \( A, B \in \mathbb{R}^{\tau(d,D) \times \tau(d,D)} \). For instance, for \( k = \{1, \ldots, d \wedge D\} \) we have that \( \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^k \in \binom{E^k}{l}} \text{adj}^0_{s,E}(J_l^k) = 0 \) for all \( s \neq k \), and that for \( s = k \) only \( \binom{k}{k} \) choices of \( E \in E^k \) will give an either +1 or −1 entry in the respective position of \( \text{adj}^0_s \). Similarly, \( \sum_{l=0}^{k} (-1)^{k-l} \sum_{J_l^k \in \binom{E^k}{l}} \text{adj}^0_{s,E}(J_l^k) = 0 \) for \( s < k \), and for \( s = k \) it has the same entries as \( \text{adj}^0_s \) for each choice of \( E \in E^k \), whereas for \( s > k \) we may have nonzero entries. However, this implies that, with the appropriate choice of linear independent equations, we may choose \( A = I \) and then \( B \) will be an upper triangular matrix with the identity on the diagonal. Therefore, \( B \) is invertible and for \( \beta \in \mathbb{R}^{\tau(d,D)} \) hence there exists \( \tilde{\beta} \) such that \( B\tilde{\beta} = \beta \).

\[ \square \]

A.5. On the coefficients of \( T_k \) configurations

In Section 4.2.3 we use a degenerate \( T_k \) configuration to prove that finite rank-one convex functions can be written as the pointwise supremum of 1-polyaffine functions at each
point. In the proof we defined the auxiliary points $X_\varepsilon^i, Y_\varepsilon^i$ in (4.47) and (4.48) as

$$X_\varepsilon^i = \varepsilon \sum_{j=1}^{i} \lambda_{i} Y_{i},$$  \hspace{1cm} (4.47)

$$Y_\varepsilon^i = Y_{i} + X_\varepsilon^i.$$  \hspace{1cm} (4.48)

We claim that the $T(X_\varepsilon^i)$ are a convex combination of the $T(Y_\varepsilon^i)$ with the coefficients $\lambda_\varepsilon^i(\mu^\varepsilon)$ and that for $\varepsilon \to 0$ we recover the original coefficients $\lambda_i$ of the degenerated $T_k$ configuration. Here we investigate these claims and start by showing the first part with the help of the following proposition.

**Proposition A.10.** Let $(A_1, \ldots, A_k)$ form a $T_k$ configuration with coefficients $\lambda \in \Lambda_k$ and auxiliary points $J_1, \ldots, J_k$ as of Definition 4.46. Then $J_1, \ldots, J_k \in P_{d \land D}(\{A_1, \ldots, A_k\})$ with

$$T(J_i) = \sum_{i=1}^{k} \mu_i T(A_i)$$  \hspace{1cm} (A.3)

where $\mu \in \Lambda_k$,

$$\mu_i = \left[ (1 - \lambda_i) \prod_{j=i+1}^{k} \lambda_j \right] / \left( 1 - \prod_{i=1}^{k} \lambda_i \right)$$  \hspace{1cm} (A.4)

and where $pc$ denotes the polyconvex hull.

It is a known result that $J_1, \ldots, J_k \in P_{d \land D}(\{A_1, \ldots, A_k\})$ since it holds that $J_i \in P_{d \land D}^f(\{A_1, \ldots, A_k\}) \subseteq P_{d \land D}^f(\{A_1, \ldots, A_k\})$ and that the latter equals $P_{d \land D}^c(\{A_1, \ldots, A_k\})$ because $\{A_1, \ldots, A_k\}$ is compact. Instead, the focus of this proposition is rather on the coefficients $\mu_i$ of the convex combination (A.3) and we are not aware of a proof of this in the literature. We use the specific choice of coefficients in the proof of Theorem 4.30. For the proof of this proposition we will use the following lemma to show that the sum of the $\mu_i$ is 1.

**Lemma A.11.** Let $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$. Then it holds that

$$\sum_{i=1}^{k} (1 - \lambda_i) \prod_{j=i+1}^{k} \lambda_j = 1 - \prod_{i=1}^{k} \lambda_i.$$  \hspace{1cm} (A.5)

where $\prod_{j=i+1}^{k} \lambda_j = 1$ whenever $i + 1 > k$ by convention.
Proof. Equation (A.5) is clearly true for \( k = 1 \). For an induction, assume it holds for \( k \). Then

\[
\sum_{i=1}^{k+1} (1 - \lambda_i) \prod_{j=i+1}^{k+1} \lambda_j = (1 - \lambda_{k+1}) + \lambda_{k+1} \sum_{i=1}^{k} (1 - \lambda_i) \prod_{j=i+1}^{k} \lambda_j
\]

\[
= (1 - \lambda_{k+1}) + \lambda_{k+1} \left( 1 - \prod_{i=1}^{k} \lambda_i \right) = 1 - \prod_{i=1}^{k+1} \lambda_i
\]

and thus, (A.5) holds for all \( k \in \mathbb{N} \). \( \square \)

Proof of Proposition A.10. Since the minors operator \( T \) is linear on rank-one connections we have

\[
\begin{align*}
T(J_i) &= \lambda_i T(J_{i-1}) + (1 - \lambda_i) T(A_i), \quad i = 2, \ldots, k \\
T(J_1) &= \lambda_1 T(J_k) + (1 - \lambda_1) T(A_1)
\end{align*}
\]

W.l.o.g. we consider \( J_k \) only. Using the equations above we can subsequently substitute for each \( J_i \) and obtain

\[
T(J_k) = \lambda_k T(J_{k-1}) + (1 - \lambda_k) T(A_k)
\]

\[
= \lambda_{k-1} \lambda_k T(J_{k-2}) + (1 - \lambda_{k-1}) \lambda_k T(A_{k-1}) + (1 - \lambda_k) T(A_k)
\]

\[
\ldots
\]

\[
= \lambda_2 \ldots \lambda_{k-1} \lambda_k T(J_1) + (1 - \lambda_2) \lambda_3 \ldots \lambda_{k-1} \lambda_k T(A_2) + \ldots
\]

\[
+ (1 - \lambda_{k-1}) \lambda_k T(A_{k-1}) + (1 - \lambda_k) T(A_k)
\]

\[
= \lambda_1 \ldots \lambda_k T(J_k) + (1 - \lambda_1) \lambda_2 \ldots \lambda_k T(A_1) + (1 - \lambda_2) \lambda_3 \ldots \lambda_k T(A_2) + \ldots
\]

\[
+ (1 - \lambda_{k-1}) \lambda_k T(A_{k-1}) + (1 - \lambda_k) T(A_k)
\]

Thus, substracting \( \lambda_1 \ldots \lambda_k T(J_k) \) we find

\[
\left( 1 - \prod_{i=1}^{k} \lambda_i \right) T(J_k) = \sum_{i=1}^{k} \left( 1 - \lambda_i \right) \left( \prod_{j=i+1}^{k} \lambda_j \right) T(A_i)
\]

Thus we see that upon defining the coefficients \( \mu_i \) as in (A.4) we obtain (A.3). It is clear that \( \mu_i \geq 0 \), but furthermore they satisfy \( \sum_{i=1}^{k} \mu_i = 1 \), which becomes obvious with the help of Lemma A.11. Therefore, \( T(J_k) = \sum_{i=1}^{k} \mu_i T(A_i) \) and thus \( J_k \in P_{d \wedge D}^\cap(\{A_1, \ldots, A_k\}) \). \( \square \)
We now return to the degenerated $T_k$ configuration from Theorem 4.30. We had rank-one matrices $Y_i \in C^1(0)$, $i = 1, \ldots, k$ such that $T(F) = \sum_{i=1}^k \lambda_i T(F + Y_i)$. With (4.47) and (4.48) and the parameter $\varepsilon > 0$ we are able to transform this into a nondegenerated $T_k$ configuration with

$$X_{i+1} = \frac{1}{1 + \varepsilon \lambda_{i+1}} X_i + \frac{\varepsilon \lambda_{i+1}}{1 + \varepsilon \lambda_{i+1}} Y_{i+1} \frac{1}{1 - \mu_{i+1}}$$

see (4.49). Using Proposition [A.10] we find that the $X_i$ are in the polyconvex hull of $Y_i$ with

$$\lambda_i^\varepsilon(\mu^\varepsilon) : = \left(1 - \mu_i^\varepsilon \prod_{j=i+1}^k \mu_j^\varepsilon\right) \left(1 - \prod_{i=1}^k \mu_i^\varepsilon\right),$$

where $\mu_i^\varepsilon = 1/(1 + \varepsilon \lambda_i)$. We now want to show that $\lambda_i^\varepsilon(\mu^\varepsilon) \rightarrow \lambda_i$ as $\varepsilon \rightarrow 0$. For that consider first the term $1 - \prod_{i=1}^k \mu_i^\varepsilon$. It is

$$1 - \prod_{i=1}^k \mu_i^\varepsilon = 1 - \prod_{i=1}^k \frac{1}{1 + \varepsilon \lambda_i} = \prod_{i=1}^k \frac{1 + \varepsilon \lambda_i - 1}{\prod_{i=1}^k (1 + \varepsilon \lambda_i)}$$

and by expanding the product in the numerator $\prod_{i=1}^k (1 + \varepsilon \lambda_i) = 1 + \varepsilon(\sum_{i=1}^k \lambda_i) + O(\varepsilon^2)$ and noting that $\sum_{i=1}^k \lambda_i = 1$ we have

$$1 - \prod_{i=1}^k \mu_i^\varepsilon = \varepsilon \frac{1 + O(\varepsilon)}{\prod_{i=1}^k (1 + \varepsilon \lambda_i)}.$$

Hence we find that

$$\lambda_i^\varepsilon(\mu^\varepsilon) = \frac{\lambda_i \prod_{i=1}^k \frac{1}{1 + \varepsilon \lambda_i} \prod_{i=1}^k (1 + \varepsilon \lambda_i)}{1 + O(\varepsilon)} = \frac{\lambda_i \prod_{i=1}^k \prod_{j=1}^{i-1} (1 + \varepsilon \lambda_j)}{1 + O(\varepsilon)}$$

and it is obvious that $\lambda_i^\varepsilon(\mu^\varepsilon) \rightarrow \lambda_i$ as $\varepsilon \rightarrow 0$. 

189
Bibliography


