Stochastic Factorisations, Sandwiched Simplices and the Topology of the Space of Explanations

David Mond, Jim Smith and Duco van Straten

June 4, 2003

Abstract

We study the space of stochastic factorisations of a stochastic matrix, motivated by the statistical problem of hidden random variables. We show that this space is homeomorphic to the space of simplices sandwiched between two nested convex polyhedra, and use this geometrical model to gain some insight into its structure and topology. We show

Theorem 0.1 *Let V be a $n \times m$ stochastic matrix of rank r. If V is "small" (roughly speaking, if its column vectors are nearly parallel to one another) then the space of rank-size stochastic factorisations of V is homotopy-equivalent to the space of $(r-1)$-simplices with vertices on the $r-2$-sphere S^{r-2}.*

See §1.2 below for a brief discussion of the concept of homotopy-equivalence.

Theorem 0.2 *Let V be a $n \times m$ stochastic matrix of rank 3. The quotient by the action of the symmetric group S_3 (permuting the columns of U) of the space of factorisations $V = UW$, with U and W stochastic matrices of size $n \times 3$ and $3 \times m$ respectively, may be empty, and otherwise is homotopy-equivalent to a circle, or has k contractible connected components, where $0 \leq k \leq n+m$. When $m = n$, there are stochastic matrices V for which this upper bound is realised.*

We briefly discuss applications to statistics.

Key words: Stochastic matrices, hidden variables, topology, Morse theory

Author addresses:

David Mond
Mathematics Institute
University of Warwick
Coventry CV4 7AL
England

mond@maths.warwick.ac.uk

Jim Smith
Department of Statistics
University of Warwick
Coventry CV4 7AL
England

j.q.smith@warwick.ac.uk

Duco van Straten
Fachbereich Mathematik
Johannes Gutenberg-Universität
Staudingerweg 9
55099 Mainz
Germany

straten@mathematik.uni-mainz.de

1 Introduction

This paper is motivated by a statistical question, but uses Morse theory and singularity theory to progress towards its solution. The statistical question is explained, for non-specialists, in the paragraphs following. In Section 2 it is translated into a question in combinatorial geometry, which we attack in Sections 3 and 4 using a version of Morse theory for a certain class of piecewise smooth functions. These form the technical heart of the paper, and might be read with some interest for their geometrical content, independently of any application to statistics.

The statistical problem is concerned with conditional independence of discrete random variables. The random variables X and Y are *independent* if for all values i, j,

$$P\{X = i \;\&\; Y = j\} = P\{X = i\}P\{Y = j\};$$

they are *conditionally independent* with respect to a third, Z, if for each value k of Z,

$$P\{X = i \;\&\; Y = j \mid Z = k\} = P\{X = i \mid Z = k\}P\{Y = j \mid Z = k\},$$

where $P\{X = i \mid Z = j\}$ is the probability that $X = i$ given that $Z = j$, $P\{X = i \mid Z = j\} = P\{X = i \;\&\; Z = j\}/P\{Z = j\}$. That is, within each level set of Z, the variables X and Y are independent.

Conditional independence with respect to a third variable may sometimes be thought of as an explanation for the dependence of X and Y. Here is an example:

Random variables X and Y, with sample space (= domain) the set of all people, are defined as follows:

$$X(x) = \begin{cases} 2 & \text{if} \quad x \text{ suffers from baldness} \\ 1 & \text{if} \quad x \text{ has cropped hair} \\ 0 & \text{if} \quad x \text{ has shoulder-length hair} \end{cases}$$

and

$$Y(x) = \begin{cases} 2 & \text{if} \quad x \text{ watches more than 2 hours per week of football on TV} \\ 1 & \text{if} \quad x \text{ watches between 0 and 2 hours per week of football on TV} \\ 0 & \text{if} \quad x \text{ does not watch football on TV} \end{cases}$$

It is found that X and Y are not independent: an individual who watches football on television is much more likely to suffer baldness than one who does not. Nevertheless, this does not imply the existence of a causal link. Within each gender group (men and women), X and Y *are* independent — the variables X and Y are conditionally independent with respect to the variable "gender".

The correlation between X and Y can be explained simply by the fact that men are more likely than women both to suffer baldness and to watch football on television.

Given random variables X and Y which are found not to be independent, it is important to be able to determine whether there exists an explanatory random variable Z (distinct from X and Y) with respect to which they are conditionally independent. In general, such a variable

Z is not unique, and it is important to gain insight into the properties of the space of all possible Z — the space of possible explanations of the observed dependence.

Conditional independence of X and Y with respect to Z is equivalent to the matrix equation

$$P\{Y|X\} = P\{Z|X\}P\{Y|Z\}, \tag{1}$$

where for two discrete random variables A and B $P\{A|B\}$ is the matrix of conditional probabilities of A given B, with i,jth element

$$P\{A|B\}_{i,j} = P\{A = j \ \&\ B = i\}/P\{B = i\}.$$

All three matrices in (1) are stochastic (i.e. all entries are non-negative and each row sums to 1), and thus we are led to a study of the space of factorisations of stochastic matrices.

1.1 Introduction for Statisticians

In recent years there has been considerable interest in statistical models whose random variables exhibit conditional independence structures, particularly models which are encoded by a directed or undirected graph (see [26],[25],[14]). In this paper we examine some of the topological features of a subclass of these models, related to latent class models, where we assume that all of the random variables in the model are discrete and one is hidden. The parameters of these models are then a collection of certain conditional probabilities which need to be estimated.

When complete samples of data from vectors of all of the random variables are available, the estimation of these conditional probabilities is relatively straightforward. In particular it is well known that undirected graphical models lie in the exponential family (see e.g. Lauritzen, [14]) and it has been proved that directed acyclic graphical models lie in the curved exponential family ([8],[9]). Both these families have a helpful geometrical structure ([13]) which makes the statistical problems of estimation and model selection amenable to standard, albeit sometimes quite complex, statistical methodologies.

However, in practice it is often the case that one or more of the random variables in a graphical model remains totally unobserved in the available sample. This may be because that variable represents a hidden cause or explanation (hence the title of our paper), or because sampling it is extremely costly, or simply because sampling was performed before it was realised that the missing variable can be relevant. When a variable in a graphical model remains totally unobserved, estimation can become much more difficult. In particular it has been shown that even very simple graphical models will fall outside the curved exponential family ([8]) and have a much richer geometrical structure ([22]). Maximum likelihood estimates are rarely unique. Furthemore, preliminary estimates of simple special cases (e.g.[6],[7], [20], [21]) show that the space of maximum likelihood estimates can be the union of several disconnected regions of the parameter spaces, and that these components are not open. Different connected components will often relate to completely different explanations for the data — for examples of this see [5] — so despite its being difficult to achieve, it is vital to obtain a good understanding of the nature and extent of this fragmentation before any statistical inferences are drawn. In addition, because of the lack of closed form solutions of models with hidden variables, numerical methods are often employed to calculate the estimates of the vector of conditional probabilities defining that model. Convergence of these algorithms is often disrupted by the existence of multiple disconnected maxima and regions of very low probability with complicated shapes. Alternatively, convergence may appear to take place when it has not. So the results of this paper are pertinent to both computational and inferential issues associated with graphical models.

Each graphical model has an associated factorisation of the joint mass function over the random variables in the model where each component factor is a function of a subset of all of its discrete random wariables. Furthermore, as discussed at some length in [21], for many practical models the number of variables in each of these factors will be small, and the geometry of the space of maximum likelihood estimates will have a simple relationship with the geometry associated with each of these low dimensional subsets of variables. It is appropriate therefore to begin a study of the topology of the parameter space of graphical models with hidden variables by focussing on problems involving only a small number of random variables. In this paper we will discuss the homotopy types of the solution spaces associated with the joint distribution of three random variables Y_i with state space $\{1, \ldots, r_i\}, 1 \leq i \leq 3$, with the property that Y_3 is conditionally independent of Y_1 with respect to Y_2 and the margin on Y_1, Y_3 is extensively sampled while data on Y_2 is completely absent.

1.2 Why should statisticians be interested in topology?

Standard computer packages search for stochastic factorisations, but to our knowledge no attention has yet been paid to the question of whether the solutions obtained are unique, or on the contrary, may depend on initial guestimates. We show here that the space of stochastic factorisations of a given stochastic matrix may have many disjoint pieces ("connected components"). On each connected component the likelihood function will have at least one (local) maximum, and thus the output of an algorithm seeking an optimal factorisation may well depend on which component it is set loose in. But we believe that the behaviour of algorithms is sensitive to more subtle topological features of the space of factorisations. A great deal of information about the topology of a space X is provided by its *homology groups* $H_k(X)$, which measure the presence of "k-dimensional holes" in the space. For a brief and accessible introduction to this topic, we recommend [19]. The rank of its homology groups are a measure of the complexity of a space, and have a bearing on the behaviour of optimisation algorithms on the space. For example, any optimisation function will have many critical points (points at which it is not clear in which direction increase is possible), not all of which are maxima or minima. It is well known that if X is a manifold (as is the case for the spaces we consider here), the number of critical points of a "generic" function (that is, a function whose critical points are all non-degenerate) is bounded below by the sum of the ranks of the homology groups $H_k(X)$. ([17], §5).

Different spaces may have the same homology groups. For example, if one space can be contracted to another (as is the case, say, of the annulus and the circle), then their homology groups coincide. Two spaces which may both be contracted to the same space are said to be *homotopy-equivalent*, or to have the same *homotopy-type*. Other examples are a punctured sphere and a point, a punctured torus and a figure 8, and a twice-punctured sphere and a circle. A space is *contractible* if it can be contracted to a point. Since homotopy-equivalent spaces have the same homology, we can determine the homology of a space X by determining its homotopy-type. It is this that we set out to do.

2 Stochastic Factorisations and Cones

Let U and V be matrices with n rows, and let $L(U)$ and $L(V)$ be the subspaces of $\mathbb{R}^n$ generated by their columns. It is elementary linear algebra that V can be divided by U (i.e. there exists a matrix B such that $V = UB$) if and only if each of the column-vectors of V lies in $L(U)$, i.e. if $L(V) \subseteq L(U)$. If all the entries of B are required to be non-negative, then each of the column vectors of V must be expressible as a non-negative linear combination of the columns of U; that is, each column of V must lie in the cone $\{v \in \mathbb{R}^n : v = \sum_i \lambda_i u_i : \lambda_i \geq 0 \text{ for all } i\}$, where the u_i are the columns of U, which we denote by $C(U)$. So V is *positively* divisible by U if and only if $C(V) \subset C(U)$.

Recall that a matrix is *stochastic* if all its entries are non-negative and each row adds up to 1. Requiring that a matrix V should be *stochastically* divisible by a stochastic matrix U (i.e. $V = UB$ with B stochastic) is stronger than asking that V should be positively divisible by U. Nevertheless, we have

Proposition 2.1 *If V and U are stochastic matrices, and the number of columns of U is equal to the rank of V, then V is stochastically divisible by U if and only if $C(V) \subset C(U)$.*

This follows from 2.2 below. As a consequence of 2.1, we derive a very simple geometric condition which determines whether a given stochastic $n \times m$ matrix V of rank r admits a stochastic factorisation $V = UW$ with U of size $n \times r$ (henceforth a 'rank-size stochastic factorisation'). We also obtain a geometrical description of the space of all such factorisations of a given matrix V.

We adopt the following notation:

$$\begin{array}{lcl} \mathbb{R}_+ & = & \{x \in \mathbb{R} : x \geq 0\} \\ S_{p,q} & = & \text{set of all } p \times q \text{ stochastic matrices} \\ \widetilde{SF}_r(V) & = & \{(U, W) : r = \text{rank } V,\ U \in S_{n,r},\ W \in S_{r,m},\ \ V = UW\} \end{array}$$

Thus, $\widetilde{SF}_r(V)$ is the space of all rank-size stochastic factorisations of V. Since it is contained in the product of two spaces of matrices, which is in a natural way a Euclidean space, it makes sense to talk about smooth mappings with $\widetilde{SF}_r(V)$ as domain or range (*cf* [16]): a map from $\widetilde{SF}_r(V)$ to a manifold X is smooth if it extends locally to a smooth map on the ambient Euclidean space, and a map from X to $\widetilde{SF}_r(V)$ is smooth if it is smooth when thought of as a map into the ambient Euclidean space.

We denote by $SF_r(V)$ the quotient of $\widetilde{SF}_r(V)$ by the natural action of the symmetric group S_r, permuting the columns of the matrix U. This action is known as *aliasing* in Statistics; it corresponds to re-ordering the space of values of the random variable Z.

Lemma 2.2 *Suppose that V is an $n \times m$ stochastic matrix of rank r. If*

1. *the vectors $\hat{u}_1, \ldots, \hat{u}_r$ lie in $L(V) \cap \mathbb{R}^n_+$, and*
2. *the cone $C(\hat{u}_1, \ldots, \hat{u}_r)$ contains the cone $C(V)$,*

then

1. *there exist unique coefficients $\alpha_i > 0$ such that $\sum_i \alpha_i \hat{u}_i = \mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)^t$, and*
2. *denoting the vectors $\alpha_i \hat{u}_i$ by u_i, there exist unique $\beta_{i,j} \geq 0$ such that $v_j = \sum_i \beta_{i,j} u_i$ for each i, and such that $\sum_j \beta_{i,j} = 1$ for each j.*

That is, V factorises as the product of the stochastic matrix U with columns $u_1, \ldots, u_r$ and the stochastic matrix $B = [\beta_{i,j}]$.

Proof Since $\mathbf{1} \in C(V) \subset C(\hat{u}_1, \ldots, \hat{u}_r)$, there exist (unique) coefficients $\alpha_i > 0$ such that $\sum_i \alpha_i \hat{u}_i = \mathbf{1}$. The α_i are unique because $L(V) \subset L(\hat{u}_1, \ldots, \hat{u}_r)$, and so the $\hat{u}_i$ are linearly independent. The α_i are non-negative, by definition of $C(\hat{u}_1, \ldots, \hat{u}_r)$. If $\alpha_i = 0$ for some i then the point $\mathbf{1}$ lies on the boundary of the (simplicial) cone $\mathcal{C}(\hat{u}_1, \ldots, \hat{u}_r)$; it follows that all of the v_i must lie on the same bounding face of this cone, since $\sum_i v_i = \mathbf{1}$. This contradicts the supposition that the rank of V is r. Thus $\alpha_i \neq 0$ for all i.

Since $v_1, \ldots, v_m \in C(\hat{u}_1, \ldots, \hat{u}_r) = C(u_1, \ldots, u_r)$, there exist $\beta_{i,j} \in \mathbb{R}_+$ such that

$$\sum_i \beta_{i,j} u_j = v_j$$

for each j. Denote the matrix with $\beta_{i,j}$ in i-th row and j-th column by B. Then $V = UB$.

The matrix U defines an injective mapping $\mathbb{R}^r \to \mathbb{R}^n$, since it has rank r. As both $\mathbf{1} = V\mathbf{1} = UB\mathbf{1}$ and $\mathbf{1} = U\mathbf{1}$ it follows that $B\mathbf{1} = \mathbf{1}$, showing that B is stochastic also. □

The significance of these lemmas is that in order to decide whether a given $n \times m$ stochastic matrix V has a rank-size stochastic factorisation, we need merely look at the cone generated by its column vectors: such a factorisation exists if and only if this cone is contained in the cone generated by some r vectors (where $r = \text{rank } V$) *in the positive orthant $\mathbb{R}^n_+$ of $\mathbb{R}^n$*.

Example 2.3 Suppose $n = 4$ and V is a 4×4 stochastic matrix for which $L(V)$ is the 3-plane in $\mathbb{R}^4$ with equation $w + x - y - z = 0$. Then $L(V) \cap \mathbb{R}^4_+$ is the cone on the four vectors $(1, 0, 0, 1), (1, 0, 1, 0), (0, 1, 0, 1)$ and $(0, 1, 1, 0)$ (in fact $L(V) \cap [0, 1]^4$ is a regular octahedron).

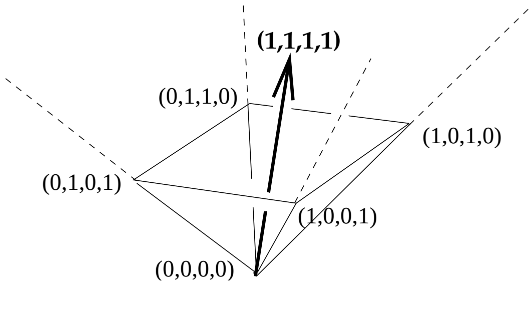

Figure 1

The slice of $L(V)\cap\mathbb{R}^4_+$ by a plane orthogonal to $\mathbf{1}$ is a square. Suppose V has column vectors $v_1,\dots,v_4$. If the rays $\mathbb{R}_+\cdot v_i$ meet the square at points close to its vertices, as shown in (i) below, it is not possible to surround them by a triangle lying entirely in the square, and therefore not possible for the cone generated by any three vectors in $L(V)\cap[0,1]^4$ to contain $C(v_1,\dots,v_4)$. Hence V has no stochastic factorisation of size 3 (i.e. of rank size). If the rays $\mathbb{R}_+\cdot v_i$ meet the square at points closer to its centre, as in (ii) below, it is possible to surround them with a triangle lying in the square; in this case the matrix V has a rank-size stochastic factorisation.

(i) non−factorisable (ii) factorisable

Figure 2

Let $Q_n\subset\mathbb{R}^n$ be the hyperplane

$$\{(x_1,\dots,x_n)\in\mathbb{R}^n:\sum_i x_i=1\}.$$

It is clear that if C and C' are cones contained in $\mathbb{R}^n_+$ then $C\subseteq C'$ if and only if $C\cap Q_n\subseteq C'\cap Q_n$.

To simplify notation, denote simply by V_s (for V_{slice}) the space $C(V)\cap Q_n$ and by W_s the space $L(V)\cap\mathbb{R}^n_+\cap Q_n$. Under the assumption that rank $V=r$, V_s and W_s are solids in the $(r-1)$-dimensional affine space $L(V)\cap Q_n$. Let $\tilde{\mathbf{\Delta}}_{V_s,W_s}$ denote the space of all ordered $(r-1)$ simplices Δ such that

$$V\subseteq\Delta\ \subseteq W.$$

By means of Lemma 2.2 we define a map $\phi:\tilde{\mathbf{\Delta}}_{V_s,W_s}\to\widetilde{SF}_r(V)$ as follows: if $\Delta=(\hat u_1,\dots,\hat u_r)\in\tilde{\mathbf{\Delta}}_{V_s,W_s}$, it follows that $C(V)\subseteq C(\Delta)$, and thus there exist unique α_i such that $U:=(\alpha_1\hat u_1,\dots,\alpha_r\hat u_r)$ is a stochastic matrix, and such that V is stochastically divisible by U. We set $\phi(\Delta)=(U,B)$, where U is constructed as we have just described, and $V=UB$.

There is an obvious map $\psi:\widetilde{SF}_r(V)\to\tilde{\mathbf{\Delta}}_{V_s,W_s}$, defined simply by mapping the stochastic factorisation (U,B) to the $r-1$-simplex with vertices $(\mathbb{R}\cdot u_1)\cap Q_n,\dots,(\mathbb{R}\cdot u_r)\cap Q_n$.

Both ϕ and ψ are smooth maps, and are mutually inverse. We have proved

Theorem 2.4 *Let V be an $n \times m$ stochastic matrix of rank r. Then $\widetilde{SF}_r(V)$ is diffeomorphic to $\tilde{\mathbf{\Delta}}_{V_s,W_s}$.* $\square$

Let $\mathbf{\Delta}_{V_s,W_s}$ denote the quotient of $\tilde{\mathbf{\Delta}}_{V_s,W_s}$ by the symmetric group action which permutes the vertices of the simplices. Our diffeomorphism $\widetilde{SF}_r(V) \simeq \tilde{\mathbf{\Delta}}_{V_s,W_s}$ is equivariant with respect to the symmetric group actions on the two spaces, and so we also have

Corollary 2.5 *With the hypotheses of 2.4, $SF_r(V)$ is homeomorphic to $\mathbf{\Delta}_{V_s,W_s}$.*

3 Topology of the space of explanations $\mathbf{\Delta}_{V_s,W_s}$

In this section we use our geometrical model $\mathbf{\Delta}_{V_s,W_s}$ for the space $SF_r(V)$ of rank-size stochastic factorisations of a given $n \times m$ stochastic matrix, modulo aliasing, to obtain information about the homotopy type of $SF_r(V)$.

Proposition 3.1 *If V is a $n \times m$ stochastic matrix of rank 2, then the space $\tilde{\mathbf{\Delta}}_{V_s,W_s}$ is diffeomorphic to the disjoint union of two rectangles. In particular V has a rank-size stochastic factorisation. The quotient $\mathbf{\Delta}_{V,W}$ is homeomorphic to a rectangle. In particular it is contractible.*

Proof The proof is essentially contained in the following diagram (in which $m = 4$).

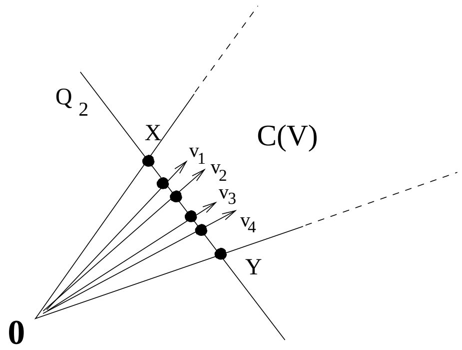

Figure 3

After re-ordering the v_i, we can arrange that $C(v_1, \ldots, v_m) = C(v_1, v_m)$, as shown. Denote by $\hat{v}_i$ the intersection of the lines $\mathbb{R}_+ \cdot v_i$ and Q_2, and by X and Y the intersection of Q_2 with the boundary of $L(V) \cap \mathbb{R}^n_+$.

It is clear that $\tilde{\mathbf{\Delta}}_{V_s,W_s}$ is diffeomorphic to the space of pairs of points $(\hat{u}_1, \hat{u}_2)$ on the line-segment XY with $\hat{u}_1$ between X and $\hat{v}_1$, and $\hat{u}_2$ between Y and $\hat{v}_m$, or vice versa. That is,

$$\tilde{\mathbf{\Delta}}_{V_s,W_s} = [X, \hat{v}_1] \times [\hat{v}_m, Y] \coprod [\hat{v}_m, Y] \times [X, \hat{v}_1].$$

$\square$

We note that this contradicts an assertion of Gilula ([10]), whose main theorem states in particular that not all stochastic matrices of rank 2 have a rank-size factorisation.

The case $r \geq 3$ is considerably more interesting.

Stochastic factorisation of stochastic matrices of rank at least 3

In the remainder of the section we drop the subindices and refer only to the geometric model, $\mathbf{\Delta}_{V,W}$, where $V \subset W$ are n-dimensional convex polyhedra contained in some Euclidean $\mathbb{R}^n$. We denote by $\mathbf{\Delta}_{V,W}$ the space of n-simplices contained in W and containing V, and by $\mathbf{\Delta}_{V,\partial W}$ the subspace of $\mathbf{\Delta}_{V,W}$ consisting of n-simplices whose vertices lie on ∂W. There is an obvious deformation-retraction from $\mathbf{\Delta}_{V,W}$ to $\mathbf{\Delta}_{V,\partial W}$: fix a point $P \in V$; for each $\Delta \in \mathbf{\Delta}_{V,W}$, simply push each vertex A of Δ along the ray PA until it meets ∂W. From now on we prefer to consider the smaller space $\mathbf{\Delta}_{V,\partial W}$. When V is small in relation to W, the homotopy-type of this space is easily described.

First, let $\mathring{\mathbf{\Delta}}_{P,\partial W}$ denote the space of n-simplices with vertices on W and containing P in their interior, and let $\mathbf{\Delta}^{\mathrm{reg}}_{S^{n-1}}$ denote the space of *regular* n-simplices with vertices on S^{n-1}. Then

Lemma 3.2

$$\mathring{\mathbf{\Delta}}_{P,\partial W} \simeq \mathbf{\Delta}^{reg}_{S^{n-1}} \simeq O(n)/S_{n+1}$$

(where $\simeq$ denotes homotopy-equivalence).

Proof *Step 1:* Let B be an n-ball with centre at P. Radial projection from P defines a homeomorphism $\partial W \to \partial B$; applying this to the vertices of simplices in $\mathring{\mathbf{\Delta}}_{P,\partial W}$ gives rise to a homeomorphism $\psi : \mathring{\mathbf{\Delta}}_{P,\partial W} \to \mathring{\mathbf{\Delta}}_{P,\partial B}$.

Step 2: We define a retraction $\mathring{\mathbf{\Delta}}_{P,\partial B} \to \mathbf{\Delta}^{\mathrm{reg}}_{\partial B}$ by means of a vector field X on $\mathring{\mathbf{\Delta}}_{P,\partial B}$. The value of such a vector field at a simplex Δ is determined by the collection $\{X_w \in T_w S^{r-1} :$ w a vertex of $\Delta\}$, which we define as follows: Let Δ_w be the face of Δ opposite w, and let L_w be the ray drawn orthogonal to Δ_w from its circumcentre and passing through P. Then X_w is the vector tangent to the unique minimal geodesic from w to the point where L_w meets the sphere, and has norm equal to the length of this geodesic. Evidently $w \in L_w$ for every vertex w if and only if Δ is a regular simplex. It follows that by flowing along the integral curves of X we retract $\mathring{\mathbf{\Delta}}_{P,\partial B}$ to $\mathbf{\Delta}^{\mathrm{reg}}_{\partial B}$. One can check that under this flow the simplices continue to contain P in their interior.

This proves the first homotopy-equivalence. The second follows from the fact that $O(n)$ acts transitively on $\mathbf{\Delta}^{\mathrm{reg}}_{\partial B}$, with isotropy the group of isometries of a regular n-simplex, isomorphic to S_{n+1}. □

Let $P \in V$ and let tV denote the dilation of V with centre P and scale factor $t \in \mathbb{R}_{\geq 0}$.

Proposition 3.3 *For convex polyhedra $V, W \subset \mathbb{R}^n$ with V contained in the interior of W, there exists $\eta > 0$ such that for $0 < t < \eta$, the inclusion $\mathbf{\Delta}_{tV,\partial W} \subset \mathring{\mathbf{\Delta}}_{P,\partial W}$ is a homotopy-equivalence.*

Together, 3.3, 3.2 and 2.5 prove Theorem 0.1. Although 3.3 is unsurprising, the proof requires a little preparation.
Define a function

$$f_P : \overset{\circ}{\mathbf{\Delta}}_{P,\partial W} \to \mathbb{R}$$

by

$$f_P(\Delta) = \sup\{t : tV \subset \Delta\} = \inf\{t : tV \cap \partial\Delta \neq \emptyset\}.$$

Remark 3.4 Evidently f_P is continuous, but its domain is a polyhedron, not everywhere smooth, and even where it is smooth (namely at those simplices Δ all of whose vertices lie in the interior of $n-1$-faces of ∂W), f_P is not always C^1. The crucial fact about it is that it is the minimum of a collection of $n+1$ functions $f_P^{(i)}$, defined by

$$f_P^{(i)}(\Delta) = \min\{t : V \cap \Delta_i \neq \emptyset\},$$

where Δ_i is the i-th face of Δ. Each $f_P^{(i)}$ in its turn is the minimum of a further collection of functions $f_P^{(i,j)}$, where

$$f_P^{(i,j)}(\Delta) = \text{unique value of } t \text{ such that } tv^{(j)} \in \Delta_i$$

where $v^{(j)}$ is the j-th *vertex* of ∂V. Since the polyhedra V and W are semi-algebraic subsets of $\mathbb{R}^n$, f_P is a semi-algebraic function (i.e. has semi-algebraic graph).

It is clear that $\overset{\circ}{\mathbf{\Delta}}_{P,\partial W}$ has a finite partition into smooth pieces ("strata") on which f_P is smooth. Thus, in particular f_P is locally Lipschitz, with respect to any reasonable choice of metric on $\overset{\circ}{\mathbf{\Delta}}_{P,\partial W}$, and it is this that we now exploit.

Observe that

$$\mathbf{\Delta}_{V,\partial W} = f_P^{-1}([1,\infty)).$$

We obtain information about the homotopy-type of $\mathbf{\Delta}_{V,\partial W}$ by doing Morse-theory with the function f_P. Since neither f_P nor its domain is smooth, this presents some difficulties, and we do not claim to overcome them all here. One might be tempted to try to use stratified Morse theory *a la* Goresky-MacPherson, but in fact even in the lowest dimensional cases the number of strata is very large (see e.g. Figure 5 below), and most contain no *topologically* critical points of f_P. Instead, on the one hand we exploit the fact that f_P is locally Lipschitz to obtain appropriate notions of regular and critical points, and draw conclusions about the local topological triviality of f_P, and on the other we make use of the Morse theory for minima described by V.I.Matov ([15], but see also [2]).

For the former, we recall that any locally Lipschitz map between smooth manifolds is almost everywhere differentiable. For such maps there is a notion of "generalised derivative": if $f : \mathbb{R}^n \to \mathbb{R}^p$ is locally Lipschitz then $\delta_{x_0} f$ is the convex hull, in $L(\mathbb{R}^n, \mathbb{R}^p)$, of the set

$$\{\lim_{k \to \infty} d_{x_k} f : f \text{ is differentiable at } x_k \text{ for all } k,\ (x_k) \to 0\}.$$

Moreover, there is a Lipschitz inverse function theorem, proved by F.H.Clarke in [3], from which a Lipschitz implicit function theorem follows:

Proposition 3.5 *Suppose that M is a C^1 manifold and $f : M \to \mathbb{R}$ is locally Lipschitz. If $0 \notin \delta_x f$ then there is a Lipschitz homeomorphism-germ $\psi : (\mathbb{R}^n, 0) \to (M, x)$ such that $f \circ \psi$ is the standard projection $(x_1, \ldots, x_n) \mapsto x_1$.* □

From this, a Lipschitz version of the Ehresmann fibration theorem follows. The step from proper submersion to locally trivial fibre bundle is more difficult, or at least less well-known, in this context than in the smooth category. Nevertheless it follows in the topological category by the "Isotopy Extension Principle"(cf [23], Corollary 6.15), and in the Lipschitz category by a theorem of Siebenmann and Sullivan ([24]).

Lemma 3.6 *Let M be a smooth manifold and $f : M \to \mathbb{R}$ be a Lipschitz function. Suppose that $f^{-1}[a, b]$ is compact, and that for all $t \in [a, b]$ and $x \in f^{-1}(t)$, $0 \notin \delta_x f$. Then for any $t_1, t_2 \in (a, b)$, $f^{-1}([a, t_1])$ and $f^{-1}[a, t_2])$ are Lipschitz homeomorphic, as are $f^{-1}([a, t_1))$ and $f^{-1}[a, t_2))$.* □

We will call points x as in Proposition 3.5 *Lipschitz regular*, and their complement *Lipschitz critical*. Although the domain $\overset{\circ}{\mathbf{\Delta}}_{P,\partial W}$ of f_P is not a smooth manifold, there is a bi-Lipschitz homeomorphism ψ from $\overset{\circ}{\mathbf{\Delta}}_{P,\partial W}$ to the smooth manifold $\overset{\circ}{\mathbf{\Delta}}_{P,\partial B}$ where B is a ball centred at P, as shown in the proof of 3.2. Moreover, $f_P \circ \psi^{-1}$ continues to be the minimum of a collection of piecewise-smooth functions. From this it follows that Sard's theorem holds:

Lemma 3.7 *The set of Lipschitz-critical values of $f_P \circ \psi^{-1}$ is contained in a semi-algebraic set of measure zero in $\mathbb{R}$.*

Proof Any polyhedron in a Euclidean space is semi-algebraic, and the map f_P is itself semi-algebraic. The graph of f_P can be embedded in a Euclidean space $\mathbb{R}^D \times \mathbb{R}$, and then given a Whitney stratification (a finite partition into manifolds, obeying certain regularity conditions, see e.g. [1]). We can identify f_P with projection onto $\mathbb{R}$. The set of critical points of f_P on each stratum is of measure 0, by Sard's theorem (see e.g. [16]), and semi-algebraic, by the Tarski-Seidenberg theorem (see e.g. [1]). By the regularity of the stratification, if Δ lies in a stratum X_α and is not a critical point of $f_{P|\ X_\alpha}$, then Δ is not a Lipschitz-critical point of f_P. Hence the set of Lipschitz critical values of f_P is contained in a finite union of semi-algebraic sets of measure zero. □

Proof of Proposition 3.3 If the set of Lipschitz critical values of f_P accumulated at 0 then by the curve selection lemma (see e.g. [18], pp 25ff) there would be an interval, containing 0, of critical values. By 3.7, this cannot happen. Hence there is some $\eta > 0$ such that $(0, \eta)$ contains no critical value of f_P. By 3.6, f_P is a locally Lipschitz-trivial fibre bundle over $(0, \eta]$, and in particular for any $t \in (0, \eta)$, the inclusion

$$\mathbf{\Delta}_{tV,\partial W} = f_P^{-1}([t, \infty)) \hookrightarrow f_P^{-1}((0, \infty)) = \overset{\circ}{\mathbf{\Delta}}_{P,\partial W}$$

is a homotopy-equivalence. □

We remark that for *large* t, $\mathbf{\Delta}_{tV,\partial W}$ is obviously empty. What can be said about $\mathbf{\Delta}_{tV,\partial W}$ for t intermediate between these extremes? To answer this we consider the Morse theory of f_P. That is, we try to describe the changes in the homotopy-type of $\mathbf{\Delta}_{tV,\partial W}$ as t passes through Lipschitz-critical values of f_P.

Motivating our analysis is the following surprising theorem due to V.I.Matov ([15]):

Theorem 3.8 *Let $f_1, \ldots, f_k$ be smooth functions in general position on the smooth manifold M^n, let $f = \min\{f_1, \ldots, f_k\}$, and let $x \in M$ be a point at which the values of all k functions coincide. Then either x is a topologically non-critical point of f, or the germ of f at x is topologically equivalent to an ordinary Morse critical point of index $\geq k-1$.* □

In fact the bound on the index becomes less surprising if we consider, for example,the case of the minimum of two functions $f_1, f_2 : \mathbb{R} \to \mathbb{R}$. A moment's thought shows that where $f_1(x) = f_2(x)$, $\min\{f_1, f_2\}$ can have a local maximum, or be topologically non-critical, but *cannot* have a local minimum. One can gain further intuition from Figures 11 and 12, below, which show, respectively, the level sets of a function $\min\{f_1, f_2, f_3\}$ in the neighbourhood of a critical point where the values of the f_i coincide, and the level sets in the neighbourhood of a Morse critical point of index 2. The topological equivalence is clear.

Remark 3.9 It is the bound on the Morse index of the critical point that is of most interest to us, since there is evidence that the following *ansatz* holds:
In calculating the homotopy-type of $\mathbf{\Delta}_{V,\partial W}$, we can assume that all Lipschitz-critical points of f_P occur at simplices Δ for which all the values $f_P^{(i)}(\Delta)$ coincide.

If Matov's theorem applied directly, from this ansatz we would deduce that if t_1 and t_2 are regular values of f_P and there is a unique critical value in (t_1, t_2) with a unique critical point lying over it, then $\mathbf{\Delta}_{t_1V,\partial W}$ is obtained from $\mathbf{\Delta}_{t_2V,\partial W}$ *by gluing in a cell of index at least n.* Since we know that for t large, $\mathbf{\Delta}_{tV,\partial W} = \emptyset$, this would place strong bounds on the homotopy-type of $\mathbf{\Delta}_{V,\partial W}$: it has the homotopy-type of a CW complex of dimension $\leq \dim \mathring{\mathbf{\Delta}}_{0,\partial W} - n = n^2 - n - 1$. See 3.12 below for an approximation to this.

For $n = 2$ (when $\mathbf{\Delta}_{V,W}$ is the space of triangles sandwiched between nested convex plane polygons), things do run according to this conjectural sequence.

Lemma 3.10 *Suppose that $n = 2$, that $\Delta \in \mathring{\mathbf{\Delta}}_{P,\partial W}$ with $f_P(\Delta) = t_0$, and that $t_0 V$ does not touch all three edges of Δ. Then Δ is not a Lipschitz-critical point of f_P.*

Proof First we fix a germ of a piecewise linear homeomorphism $\psi : (\mathring{\mathbf{\Delta}}_{P,\partial W}, \Delta) \to (\mathbb{R}^3, 0)$, as follows. Through each vertex w_i of Δ select one edge of ∂W, and consider the (bi-infinite) line containing it. The product of these three lines will be our $\mathbb{R}^3$. If at each vertex of Δ there is only one edge of ∂W, no further discussion is necessary: locally $\mathring{\mathbf{\Delta}}_{P,\partial W} \simeq \mathbb{R}^3$. If at any of the vertices of Δ two edges of ∂W meet, radial projection from P of the vertices of triangles defines a map germ $(\mathring{\mathbf{\Delta}}_{P,\partial W}, \Delta) \to (\mathbb{R}^3, 0)$ which is, by the convexity of W, a homeomorphism in the neighbourhood of Δ.

Suppose now we are in the favourable situation where $\mathring{\mathbf{\Delta}}_{P,\partial W}$ is smooth at Δ. We define a smooth path $\gamma(\lambda)$ through Δ such that

$$\frac{d(f_P \circ \gamma)}{d\lambda}(0) > 0.$$

We suppose that the edge w_1w_2 of Δ does not meet t_0V. If Δ meets the edge w_0w_i ($i = 1$ or 2) then this edge does not lie in ∂W, since we assume t_0V is contained in the interior of W. Let E_i be the edge of W containing w_i, and let $w_i(\lambda)$ be the radial projection from P to E_i of the point $(1-\lambda)w_i + \lambda w_0$.

We take $\gamma(\lambda)$ to be the triangle with vertices $w_0, w_1(\lambda), w_2(\lambda)$.

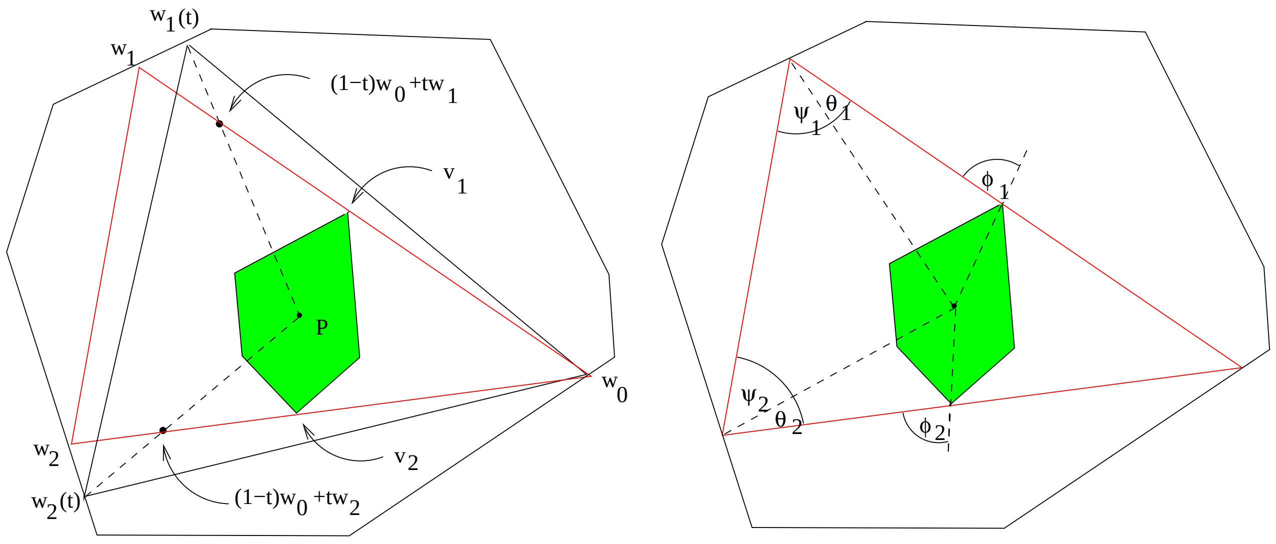

Figure 4: Deformation of Δ which increases f_P

By straightforward geometry, in the figure shown

$$\frac{d(f_P \circ \gamma)}{d\lambda}\bigg|_{\lambda=0} = \min\{\frac{\|v_i - w_i\| \sin\theta_i \sin(\psi_i + \theta_i)}{\|P - v_i\| \sin\psi_i \sin\phi_i} : i = 1, 2\} > 0.$$

If vertices $v_{i,j}, j = 1, 2$ of t_0V lie on the edges w_0w_i of Δ, this formula must be replaced by

$$\min\{\frac{\|v_i - w_i\| \sin\theta_i \sin(\psi_i + \theta_i)}{\|P - v_{i,j}\| \sin\psi_i \sin\phi_{i,j}} : i, j = 1, 2\},$$

which is still strictly positive.

If one or both of the w_i coincide with vertices of ∂W, then the situation is slightly more complicated, since the moving vertices $w_i(\lambda)$ lie on different edges of ∂W for $\lambda > 0$ and $\lambda < 0$, and thus the angle ψ_i has different values $\psi_{i,+}$ and $\psi_{i,-}$ for λ positive and negative. We then have

$$\frac{d(f_P \circ \gamma)}{d\lambda}_{\lambda=0\pm} = \min\{\frac{\|v_i - w_i\| \sin\theta_i \sin(\psi_{i,\pm} + \theta_i)}{\|P - v_{i,j}\| \sin\psi_i \sin\phi_{i,j}} : i,j = 1,2\},$$

both of which are again strictly positive. The values on $d(f_P/ \circ \lambda)/d\lambda(0)$ of the members of the generalised derivative $\delta\Delta f_P$ are contained in the closed interval in $\mathbb{R}$ bounded by the maximum and the minimum of all of these expressions. Since all are strictly positive, the conditions of the Lipschitz implicit function theorem are met, and Δ is not a critical point of f_P. □

Figure 5 shows many distinct configurations satisfying the hypotheses of 3.10; in each, the construction of the proof of 3.10 furnishes us with a path through Δ along which f_P increases. Lemma 3.10 thus enables us to deal with a large number of geometrically distinct configurations at once.

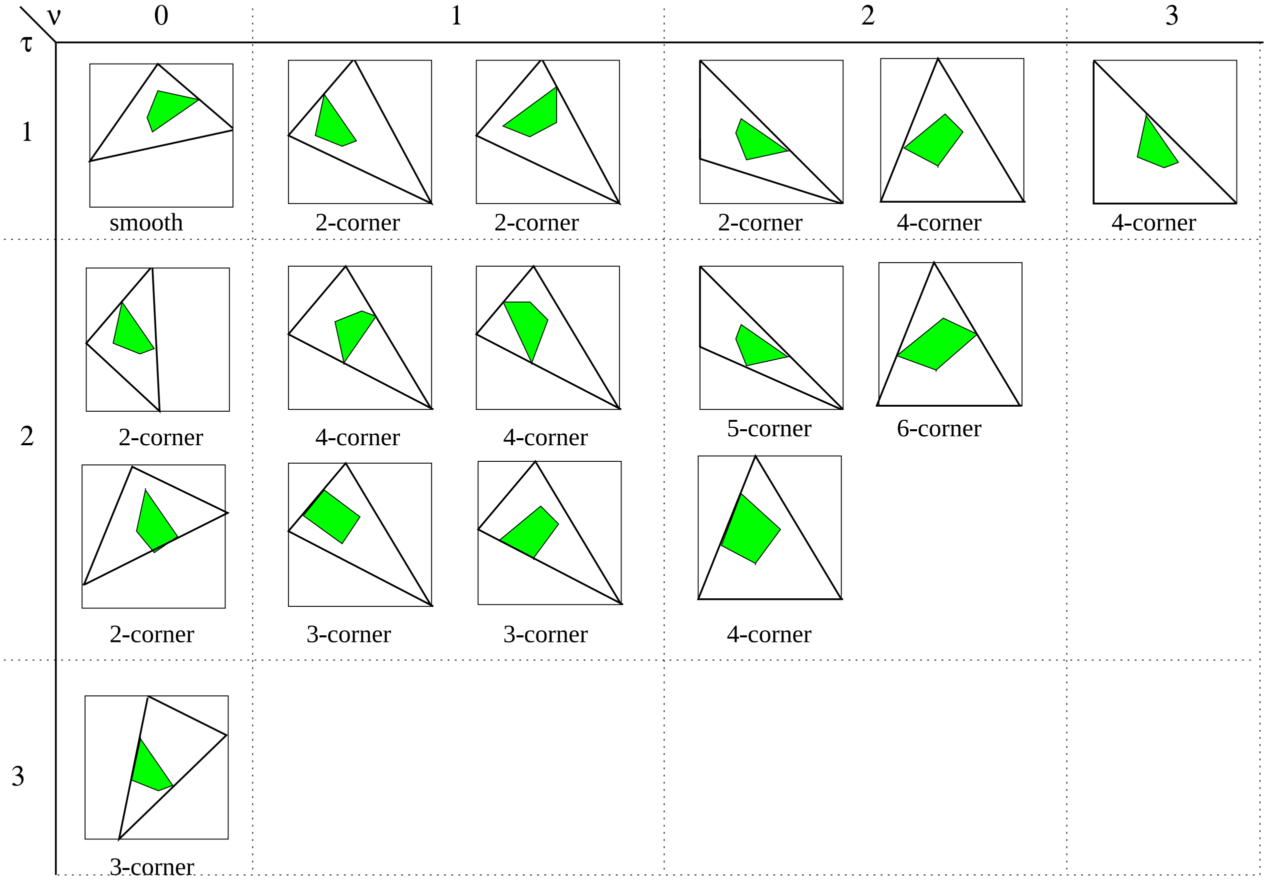

Figure 5: Regular points of f_P

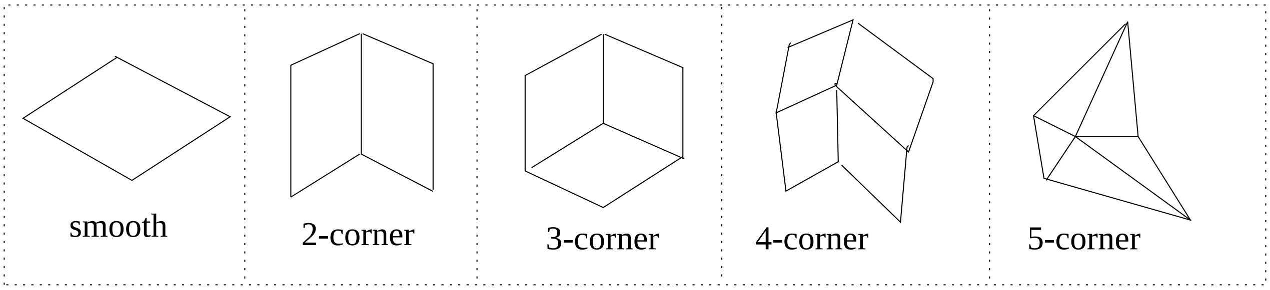

Figure 6: Level sets of regular points of f_P

One might imagine that a version of Lemma 3.10 would hold in higher dimensions, and indeed essentially the same proof works when W is smooth and strictly convex.

Lemma 3.11 *Let $W \subset \mathbb{R}^n$ be a strictly convex n-dimensional manifold, with ∂W of class C^k ($k \geq 1$). Let $V \subset int(W)$ be an n-dimensional convex polyhedron and let $P \in int(V)$. If $\Delta \in \mathring{\boldsymbol{\Delta}}_{P,\partial W}$ with $f_P(\Delta) = t_0$, and $t_0 V$ does not meet all of the $(n-1)$-dimensional faces of Δ, then Δ is not a critical point of f_P.*

Proof Let Δ_j denote the $(n-1)$-face of Δ opposite the vertex w_j. Suppose that $t_0 V \cap \Delta = \emptyset$.

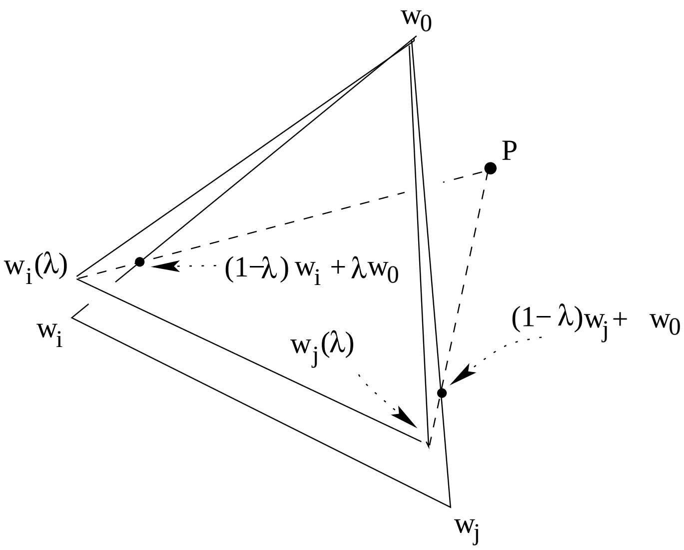

Figure 7: Deformation of simplex Δ which increases $f_P(\Delta)$

Step 1: We construct a flow on a neighbourhood of Δ in $\mathring{\boldsymbol{\Delta}}_{P,\partial W}$.

First we construct a smooth path $\Delta(\lambda)$ through Δ itself. For $i = 1, \ldots, n$, let $w_i(\lambda)$ be the radial projection to ∂W of $(1-\lambda)w_i + \lambda w_0$, let $\Delta(\lambda)$ be the simplex $(w_0, w_1(\lambda), \ldots, w_n(\lambda))$, and let $\Delta_j(\lambda)$ be the $(n-1)$-face of $\Delta(\lambda)$ opposite the vertex $w_j(\lambda)$. Note that for each simplex near Δ, the same construction also gives a path, and thus we have a flow on a neighbourhood of Δ. It has the same differentiability class as ∂W.

Step 2: We obtain a more concrete evaluation of f_P.

Call Δ_j a *contact face* of Δ if $t_0 V \cap \Delta_j \neq \emptyset$. We can extend the ordering of the vertices of Δ to an ordering of the vertices of each simplex in some neighbourhood of Δ in the space of simplices, and then on this neighbourhood define a real-valued function f_P^j whose value on a simplex is the least value of t such that tV meets the j-th face of the simplex. Then $f_P = \min\{f_P^j : j = 0, \ldots, n\}$. Since V is a polyhedron, for each contact face Δ_j there exist contact points $v_j^-, v_j^+ \in \partial V$ (not necessarily unique, not necessarily distinct) such that $tv_j^-, tv_j^+ \in \Delta_j$ and such that for small non-positive λ, $f_P^j(\Delta(\lambda))$ is the unique value of t such that $tv_j^- \in \Delta_j(\lambda)$, and similarly for small non-negative λ, $\inf\{t : tV \cap \Delta_j(\lambda) \neq \emptyset\}$ is the unique value of t such that $tv_j^+ \in \Delta_j(\lambda)$.

Denote these values by $\Delta_j(\lambda)/v_j^-$ and $\Delta_j(\lambda)/v_j^+$. Each is a C^k function of λ.

Step 3: We show that f_P has positive derivative along the flow.

Provided that for each $i \neq 0$ for which w_i belongs to a contact face, the vector $\overline{w_i w_0}$ does not lie in $T_{w_i}\partial W$, it follows that for each contact face Δ_j, $\frac{d}{d\lambda}(\Delta_j(\lambda)/v_{j^+})_{|\lambda=0}$ and $\frac{d}{d\lambda}(\Delta_j(\lambda)/v_{j^-})_{|\lambda=0}$ are both strictly positive. The condition that $\overline{w_i w_0} \notin T_{w_i}\partial W$ is, of course, guaranteed by the strict convexity of W.

There exist contact faces Δ_{j^+} and Δ_{j^-} such that for small λ, $f_P(\Delta(\lambda)) = f_P^{j^+}(\Delta(\lambda))$ for $\lambda \geq 0$ and $f_P(\Delta(\lambda)) = f_P^{j^-}(\Delta(\lambda))$ for $\lambda \leq 0$.

By the differentiability of the flow defined in Step 1, the linear maps in the generalised derivative $\delta_\Delta f_P$ take values between $\frac{d}{d\lambda}(\Delta_j(\lambda)/v_{j^+})_{|\lambda=0}$ and $\frac{d}{d\lambda}(\Delta_j(\lambda)/v_{j^-})_{|\lambda=0}$. It follows that every linear map in $\delta_\Delta f_P$ is non-singular. Thus, by the Lipschitz Implicit Function Theorem, Δ is not a critical point of f_P. $\square$

However, the corresponding statement fails in general when W is a polyhedron. In the following figure, V is tangent to Δ only on three of its faces. However, Δ is a critical point of f_P if $V \cap H \subset \Delta \cap H \subset W \cap H$ is a critical 2-dimensional configuration (see 4.3 below), where H is the plane containing the base of V.

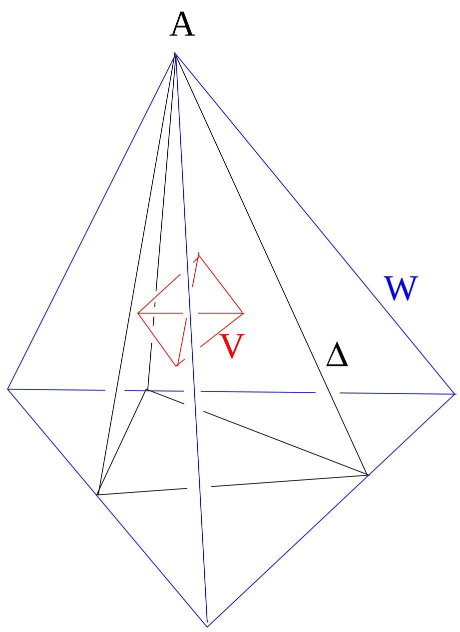

Figure 8: Counterexample to extension of 3.11 to polyhedral W

Despite this failure, we are still able to recover one of the consequences of the ansatz described in Remark 3.9:

Theorem 3.12 *Let V and W be convex n-dimensional polyhedra, with $V \subset W \subset \mathbb{R}^n$. Then $H^q(\mathbf{\Delta}_{V,\partial W};\mathbb{Z}) = 0$ for $q > n^2 - n - 1$.*

Proof By suitable choice of P we can assume that 1 is not a critical value of f_P. Let A be an approximation to W with smooth, strictly convex boundary. It can be chosen so close to W that there exist $t_1 < 1$ and $t_2 > 1$ such that $t_1 W \subset A \subset t_2 W$, and there is no critical value of f_P in $[1/t_2, 1/t_1]$. Moreover, by standard transversality argument we can suppose that the function $f_P : \mathring{\mathbf{\Delta}}_{0,\partial A} \to \mathbb{R}$ meets the requirements of Matov's theorem, so that $\mathbf{\Delta}_{V,A}$ has the homotopy-type of a CW complex of dimension $\leq n^2 - n - 1$ (cf [17], §3).

The configuration $V \subset t_i W$ can be transformed to $1/t_i V \subset W$ by dilation centred at P, so $\mathbf{\Delta}_{V,t_i W}$ is homeomorphic to $\mathbf{\Delta}_{1/t_i V,W}$. By 3.6 the inclusion $\mathbf{\Delta}_{1/t_i V,W} \subset \mathbf{\Delta}_{1/t_2,W}$ is a homotopy-equivalence and hence so is the inclusion $\mathbf{\Delta}_{V,t_1 W} \hookrightarrow \mathbf{\Delta}_{V,t_2 W}$. Since this inclusion factors through $\mathbf{\Delta}_{V,A}$, the result follows. □

4 Factorisation of stochastic matrices of rank 3

In the case of matrices of rank 3, we now obtain more detailed information.

Proposition 4.1 *Let V be an $n \times m$ stochastic matrix of rank 3. Then*
(1) $L(V) \cap Q_n \cap \mathbb{R}^n_+$ is a convex polygon with no more than n edges, and
(2) $C(V) \cap Q_n$ is a convex polygon with no more than m edges.

Proof
(1) $L(V) \cap Q_n \cap \mathbb{R}^n_+$ is a slice of a regular $n-1$-simplex by a plane, which has to meet its interior (as $\mathbf{1} \in C(V)$). This slice is a convex polygon with no more than n edges (one for each of the faces of the simplex that $L(V)$ meets).
(2) $C(V)$ is a cone generated by m vectors in the positive orthant, each of which is transverse to Q_n. $\square$

These bounds are sharp. This is obvious for the second; for the first, it is not hard to check that for each k with $3 \leq k \leq n$, there exist 2-planes containing 0 whose intersection with the regular simplex Q_n is a k-gon.

Thus, the space of rank-size stochastic factorisations of an $n \times m$ stochastic matrix of rank 3, is homeomorphic to the space of triangles contained in a given convex plane polygon with no more than n edges and containing another given convex polygon with no more than m edges.

Let ν denote the number of vertices of Δ, coinciding with vertices of W, and which bound edges of Δ which pass through a vertex of tV. For a generic $P \in V$, the codimension in $\mathbf{\Delta}_{t_0 V, \partial W}$ of the set of triangles with given values of τ and ν is $\tau + \nu$.

Lemma 4.2 *By choosing P appropriately we can ensure that*
(1) no critical triangle $\Delta \in \mathring{\mathbf{\Delta}}_{P,\partial W}$ has a vertex at more than one of the vertices of ∂W, and
(2) if $\Delta \in \mathring{\mathbf{\Delta}}_{P,\partial W}$ then $\partial\Delta$ does not pass through more than four vertices of tV for any $t \in \mathbb{R}$.

Proof (1) Finitely many lines (the diagonals) join the distinct vertices of ∂W. By choosing P appropriately we can ensure that for all values of t, at most one vertex of tV lies on any of the diagonals of ∂W. A triangle with two diagonals of ∂W among its edges is therefore not critical, since one of these edges contains no vertex of tV.
(2) If a triangle ABC passes through five vertices of tV, then two of its edges, say AB and AC, must each contain an edge of tV, and the remaining edge, BC, must contain a vertex. That is, two edges of tV, continued, meet at a point A of ∂W. This occurs only for finitely many values of t. For each such value of t, a triangle ABC containing two edges of tV is determined. For almost all P, for all these special values t the third edge BC of the triangle ABC does not then pass through any vertex of tV. $\square$

By this lemma, we can obtain a complete list of the combinatorial types of configurations (critical and non-critical) which occur for generic choice of P by considering the special case in which $m = n = 4$.

Calculations are considerably simplified if we apply a projective transformation ϕ which turns W into a square; such a transformation induces a diffeomorphism

$$\phi_* : \mathbf{\Delta}_{V,\partial W} \to \mathbf{\Delta}_{\phi(V),\partial\phi(W)}$$

and since we are at this point interested only in the topology of $\mathbf{\Delta}_{V,\partial W}$, let us therefore now assume that W is a square.

Now we consider the critical points of f_P, and examine the way that $f_P^{-1}([t, \infty))$ changes as t passes through a critical value. By Lemma 4.2, the only remaining combinatorial types generically occurring are those shown in Figure 9.

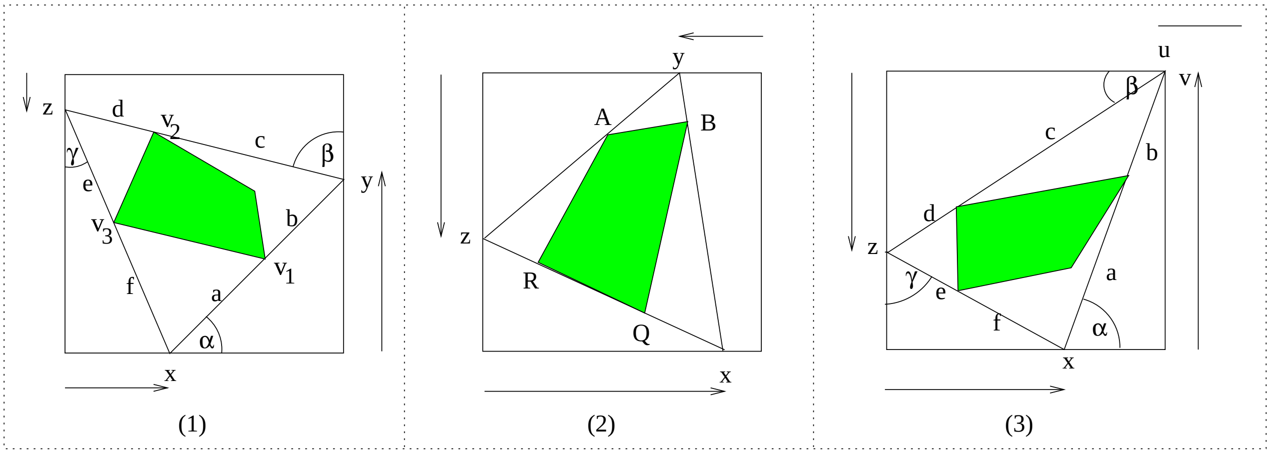

Figure 9: Critical points of f_P

Lemma 4.3 *In (1), Δ is either a 3-corner (and in particular non-critical) or a saddle.*

In (2) and (3), Δ is a 4-corner or a local maximum.

Proof In each configuration, Δ passes through three or more vertices of P. In each case, we vary the position x of one of its vertices, and try to complete a triangle in $\mathbf{\Delta}_{V,W}$ retaining the three tangencies. Despite its naiveté, this dynamic description seems to us to give the clearest understanding, and so we clarify its premise. We imagine uncoupling the two edges of the triangle Δ which meet on the bottom edge of the square, and view x as the coordinate of the lower end of the edge xy. We let w be the co-ordinate of the point where the edge zx meets the bottom edge of the square. Initially, of course, $w = x$. We imagine the edges of Δ meeting at the other two vertices as remaining coupled. Through our insistence on tritangency, by shifting x to a nearby value $x' = x + \delta x$ we define new values $y' = y + \delta y, z' = z + \delta z, w' = w + \delta w$ for the coordinates of the end points of the other edges. Of course, if $w' \neq x'$ we will no longer have a triangle in $\mathbf{\Delta}_{V,\partial W}$. However, if the new edge $z'w'$ crosses $x'y'$ inside W, then by pushing the point Q of intersection to ∂W along the ray PQ, we obtain a triangle Δ' in $\mathbf{\Delta}_{tV,\partial W}$. This slack can be used to construct a path in $\mathring{\mathbf{\Delta}}_{P,\partial W}$ along which f_P has positive derivative, and, as in the proof of 3.10, we see that Δ is not a critical point of f_P.

In general (e.g. in configurations (2) and (3) above) dw/dx does not exist; however the one-sided derivatives dw/dx_+ and dw/dx_- always exist. If both one-sided derivatives are greater than 1, f_P increases along a path through Δ in which x increases; if both are less than 1, f_P increases along a path in which x decreases. Thus Δ is a critical point precisely when

$dw/dx_- \geq 1 \geq dw/dx_+$ or when $dw/dx_- \geq 1 \geq dw/dx_+$.

(1) Here, evidently, dw/dx exists. It is equal to

$$\frac{bdf}{ace} \tan\alpha \ \tan\gamma.$$

If $dw/dx \neq 1$, Δ is a 3-corner on $\mathbf{\Delta}_{tV,W}$: the level set $f_P^{-1}(t)$ is contained in the union of smooth surfaces D_1, D_2, D_3, with D_i the set of triangles in $\mathring{\mathbf{\Delta}}_{P,\partial W}$ containing the vertex v_i to tV. These three meet in general position Δ, and locally $f_P^{-1}([t,\infty))$ is contained in the intersection of the three half-spaces D_i^- they define. It thus resembles an octant in $\mathbb{R}^3$.

If $dw/dx = 1$, then Δ is a saddle; for $d^2w/dx^2 > 0$, so that it becomes easier to complete the tritangent triangles in $\mathbf{\Delta}_{tV,W}$ as x moves from its initial value. Thus, the unique allowable infinitesimal motion in either direction in $\mathbf{\Delta}_{tV,W}$ lifts to a true motion. The three surfaces D_i meet at Δ as shown in Figure 11, that is to say, pairwise transversely, but with the curve of intersection of each two being simply tangent to the third. The small arrow on each surface D_i in Figure 11 indicates the region D_i^-. The union of the three surfaces is known in Singularity Theory as the "birth of two triple-points" -see e.g. [12].

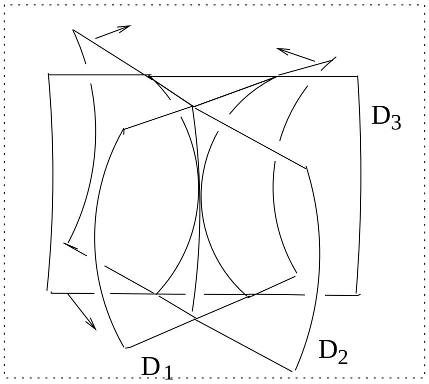

Figure 10

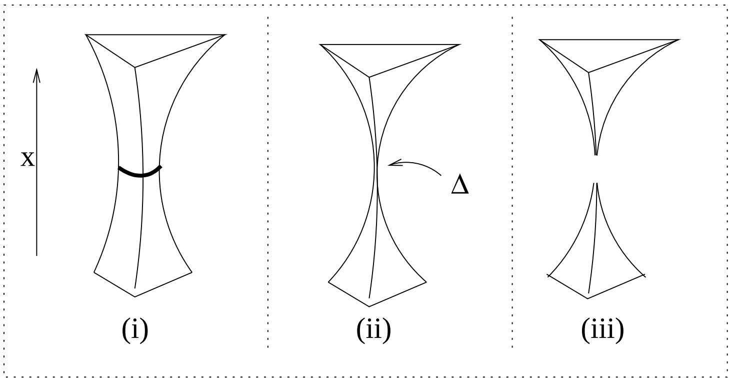

Figure 11

The sequence of pictures in Figure 11 shows the transition in the level sets of f_P as t passes through the critical value. The heavy line around the waist of the level set in Figure 11(i), which contracts to a point in Figure 11(ii), is known as a *vanishing cycle* in Singularity Theory. The vertical arrow in (i) indicates the x-coordinate axis (*cf* Figure 9 (1)).

(2) If x increases then the edge zx pivots at Q; if x decreases then it pivots at R. Therefore $dw/dx_- > dw/dx_+$. Generically, neither limit will be equal to 1 for any value of t, since this configuration occurs only for finitely many values of t.

There are thus three cases: $1 > dw/dx_-$, $dw/dx_- > 1 > dw/dx_+$ and $dw/dx_+ > 1$. Only in the second of these is Δ a critical point for f_P. In this case, Δ is isolated in the level set of f_P, and is a *local maximum* for f_P. In the first and third cases, Δ is a 4-corner. This last can be seen as follows. Denote the triangles ABQ and ABR in Figure 9(2) above by tV_1 and tV_2 respectively. Then tV is the convex hull of $tV_1 \cup tV_2$, and

$$\mathbf{\Delta}_{tV,\partial W} = \mathbf{\Delta}_{tV_1,\partial W} \cap \mathbf{\Delta}_{tV_2,\partial W}.$$

In the neighbourhood of Δ, $\mathbf{\Delta}_{tV_1,\partial W}$ is a 3-corner, diffeomorphic to a closed octant in $\mathbb{R}^3$, with bounding surfaces D_A, D_B and D_Q. By adding the vertex R to tV_1, we introduce a fourth smooth surface, D_R. In case (b), the region D_R^- meets the octant only at Δ; in the other two cases, it meets the octant in a pyramidal region with vertex at Δ.

A similar analysis can be brought to bear on Configuration (3). However, it is more revealing to use projective duality. Recall that duality of projective configurations associates to each line in projective space the corresponding point in the dual space of lines $(\mathbb{P}^2)^\vee$ and to each point in $P \in \mathbb{P}^2$ the line ℓ_P in $(\mathbb{P}^2)^\vee$ corresponding to the set of lines passing through P.
It is readily checked that configuration (3) is projectively dual to configuration (2).

Take coordinates on $\mathbb{R}^2$ with P at the origin, and include $\mathbb{R}^2$ in $\mathbb{RP}^2$ by $(x, y) \mapsto [x, y, 1]$. Let $\check{x}, \check{y}, \check{z}$ be coordinates on $(\mathbb{RP}^2)^\vee$; lines not passing through $(0, 0)$ in $\mathbb{R}^2$ become points in the finite portion $\{\check{z} \neq 0\}$ of $(\mathbb{RP}^2)^\vee$, and thus, taking affine coordinates $\check{u} = \check{x}/\check{z}, \check{v} = \check{y}/\check{z}$, points in the dual affine space $(\mathbb{R}^2)^\vee$.

Denote the dual of a configuration C by $C^\vee$. If $V \subset W \subset \mathbb{R}^2$ is a configuration of convex polygons, then so is $W^\vee \subset V^\vee \subset (\mathbb{R}^2)^\vee$. And if $V \subset T \subset W$ is a triangle, then $T^\vee$ is also a triangle and $W^\vee \subset T^\vee \subset V^\vee$. It follows that projective duality induces a bijection (in fact a diffeomorphism) $\mathbf{\Delta}_{tV,W} \to \mathbf{\Delta}_{W^\vee,(tV)^\vee}$.

The dual of the line $\ell = \{ax + by + c = 0\} \subset \mathbb{R}^2$ is the point $(a/c, b/c) \in (\mathbb{R}^2)^\vee$, and so the dual of $t\ell$ is $(a/ct, b/ct)$. Hence $(tV)^\vee = (1/t)V^\vee$. Denote by $\check{P}$ the origin of coordinates in the dual affine space. Denote by sQ the image of a point or figure under dilation with centre $\check{P}$ and scale factor s. Clearly dilation induces an isomorphism $\mathbf{\Delta}_{W^\vee,(tV)^\vee} \simeq \mathbf{\Delta}_{sW^\vee,s(tV)^\vee}$. Hence

$$\mathbf{\Delta}_{tV,W} \simeq \mathbf{\Delta}_{W^\vee,(tV)^\vee} = \mathbf{\Delta}_{W^\vee,(1/t)V^\vee} \simeq \mathbf{\Delta}_{tW^\vee,V^\vee}.$$

In particular, the topological change (if any) in $\mathbf{\Delta}_{tV,W}$ as t passes through a critical value of f_P with critical point of type (3), is the same as the change in $\mathbf{\Delta}_{tW^\vee,V^\vee}$ associated with a point of type (2), which has already been discussed. □

Remark 4.4 (1) If W is replaced by a circle, tV is a triangle and the vertices X, Y and Z of V lie on the edges BC, AC and AB of Δ, respectively, then the dynamical analysis of the proof of 4.3 shows that Δ is a critical point for f_P if and only if

$$\frac{AZ}{ZB}\frac{BX}{XC}\frac{CY}{YA} = 1.$$

This equality appears in *Ceva's Theorem* in Euclidean geometry, which asserts that it is necessary and sufficient for the lines AX, BY and CZ to be concurrent. For this reason we call Configuration (1) of 4.3 (when it is critical) a *Cevian* configuration.
(2) The sequence of level sets shown in Figure 11 should be compared with the following sequence, showing the level sets near the standard non-degenerate critical point of index 2 $f(x, y, z) = z^2 - x^2 - y^2$.

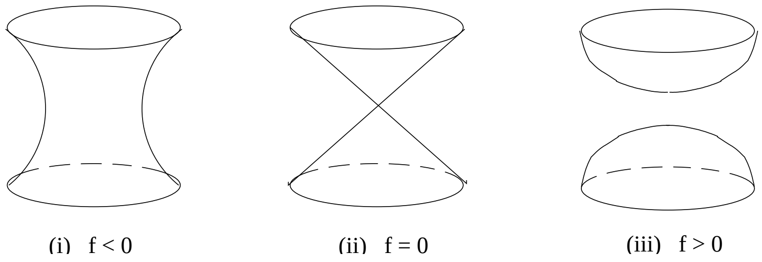

Figure 12: Level sets near non-degenerate critical point of index 2

At a critical point of type (1) the behaviour of f_P is exactly as described by Matov's theorem 3.8: f_P is locally the minimum of three *smooth* functions whose values coincide at the critical point.
(3) Under the projective duality described in the proof of 4.3, the dual of a Cevian configuration is also a Cevian configuration.
(4) The appearance of projective duality in this problem can be understood in terms of the original statistical problem, as follows. As discussed in Section 1, random variables X and Y are conditionally independent with respect to a third variable Z if and only if $P\{Y|X\} = P\{Z|X\}P\{Y|Z\}$, where $P\{A|B\}$ is the matrix of conditional probabilities of A given B. The search for explanatory random variables Z thus corresponds to the search for stochastic factorisations of the stochastic matrix $P\{X|Y\}$, and as shown in Section 2, this leads to the geometric problem of sandwiched simplices. One can equally look for stochastic factorisations of $P\{Y|X\}$; these are in a natural way dual to stochastic factorisations of $P\{X|Y\}$, and it is this duality that corresponds to projective duality in the associated geometrical problem.

Using the description of the topological transitions given in 4.3 we now compute the homotopy type of $\mathbf{\Delta}_{V,W}$. We return to the general case: V and W are once again convex polygons with p and q edges, respectively.

Lemma 4.5 *For all values of $t \geq 0$, $\mathbf{\Delta}_{V,\partial W}$ has the homotopy type of a CW complex.*

Proof This follows by the standard argument (*cf* [17]): for t large, $\mathbf{\Delta}_{tV,\partial W}$ is empty; as t diminishes, it passes through a finite number of critical values, at each of which the homotopy-type changes as described in 4.3. Up to homotopy, the effect of each change is simply to glue in a cell. □

Theorem 4.6 $\mathbf{\Delta}_{V,\partial W}$ *(and therefore $\mathbf{\Delta}_{V,W}$) is homotopy equivalent to a circle if V is small; otherwise it has a finite number of contractible connected components.*

Proof Assume for convenience that no two distinct critical points have the same critical value, and let $\varepsilon > 0$ be less than the minimum difference between consecutive critical values.

Each of the critical points of f_P is either a strict local maximum or a saddle — topologically equivalent to a non-degenerate critical point of index 2. Thus, in the case of a maximum,

$\mathbf{\Delta}_{(t-\varepsilon)V,\partial W}$ is homeomorphic to the disjoint union of $\mathbf{\Delta}_{(t+\varepsilon)V,\partial W}$ and a 3-ball, and in the case of a saddle, $\mathbf{\Delta}_{(t-\varepsilon)V,\partial W}$ is homeomorphic to the union of $\mathbf{\Delta}_{(t+\varepsilon)V,\partial W}$ and the product of a disc and an interval, $D^2 \times [0,1]$, glued in along $D^2 \times \{0,1\}$.

The conclusion follows by standard topological techniques: patch together this local description with a trivialisation of the family outside the neighbourhood of the critical point. See e.g.[17] for a description. $\square$

It remains to estimate the number of connected components of $\mathbf{\Delta}_{V,\partial W}$. For an upper bound it is necessary only to estimate how many local maxima f_P will have. Let p and q denote the number of edges of V and W respectively.

Lemma 4.7 *f_P has at most q local maxima of type (3). If $q = 3$ then f_P has no local maximum of type (3).*

Proof Suppose f_P has a local maximum of type (3) with critical value t_c and suppose that the critical triangle Δ_c has a vertex at the vertex A of W. Then for $t > t_c$ there is no triangle in $\mathbf{\Delta} tV, \partial W$ with vertex at A. $\square$

Lemma 4.8 *f_P has at most p local maxima of type (2). If $p = 3$ then f_P has no local maximum of type (2).*

Proof At a local maximum of type (2) with critical value t_c, one edge of the polygon tV is contained in an edge of the critical triangle Δ_c. For $t > t_c$, it is not possible for the corresponding edge of tV to be contained in any edge of a triangle in $\mathbf{\Delta}_{tV,\partial W}$. Therefore there is at most one local maximum of type (2) for each edge of V.

The second assertion is obvious. $\square$

We have now proved

Theorem 4.9 *$\mathbf{\Delta}_{V,\partial W}$ has no more than $p+q$ connected components. If W is a triangle, $\mathbf{\Delta}_{V,\partial W}$ has no more than p connected components. If V is a triangle, $\mathbf{\Delta}_{V,\partial W}$ has no more than q connected components. If V and W are both triangles, $\mathbf{\Delta}_{V,\partial W}$ is connected.* $\square$

This result, together with 4.6, proves Theorem 0.2, except for the statement concerning the existence of stochastic matrices realising the upper bound on the number of connected components.

5 Maximal Configurations

Let $V \subset W$ be convex polygons in the plane, with p and q edges, respectively. We call the configuration $V \subset W$ *maximal* if $\mathbf{\Delta}_{V,W}$ has the maximal possible number of connected components, namely $p + q$. We do not know if maximal configurations occur for arbitrary values of p and q; nevertheless, we can construct maximal configurations in which $p = q$. The simplest of these, for $p = 4$, is shown in Figure 13.

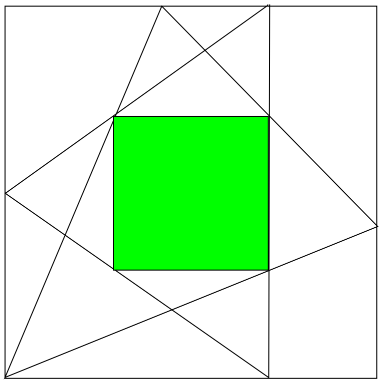

Figure 13: Configuration in which $\Delta_{V,W}$ has 8 connected components

Here V and W are concentric parallel squares, whose sides are in the ratio $\sqrt{2} - 1 : 1$. The figure shows two triangles of different types. The symmetry of the figure gives three more of each type. All of them are isolated in $\mathbf{\Delta}_{V,W}$, which thus has 8 connected components.

This configuration arises in considering the stochastic factorisations of the stochastic matrix

$$\frac{1}{2\sqrt{2}}\begin{pmatrix} \sqrt{2}-1 & 1 & 1 & \sqrt{2}-1 \\ 1 & \sqrt{2}-1 & 1 & \sqrt{2}-1 \\ 1 & \sqrt{2}-1 & \sqrt{2}-1 & 1 \\ \sqrt{2}-1 & 1 & \sqrt{2}-1 & 1 \end{pmatrix}$$

(see [5]). The reader will recognise from Figure 9 and Lemma 4.3 the two triangles shown here. The side-ratio chosen here is the only one for which either type of local maximum occurs (given that the squares are concentric and parallel). It is not a coincidence that both occur for the same ratio, for, as described in the proof of 4.3, the configurations (2) and (3) of Figure 9 are projectively dual. More precisely, if $V \subset W$ is a configuration of parallel concentric regular polygons with sides in the ratio $r : 1$, then with respect to the coordinates described in the proof of 4.3, so is $W^\vee \subset V^\vee$. And if $V \subset T \subset W$ is a triangle of one of the two types shown in Figure 9, then $W^\vee \subset T^\vee \subset V^\vee$ is a triangle of the other type. By regularity and concentricity, there is at most one ratio r_p of side-lengths for which a pair of concentric parallel regular p-gons admits a triangle of either type (and therefore of both types). Since the function f_P must have local maxima on $\mathbf{\Delta}_{P,\partial W}$ (where P is the centre of the polygons) it follows from Lemma 4.3 that such a ratio r_p does exist.

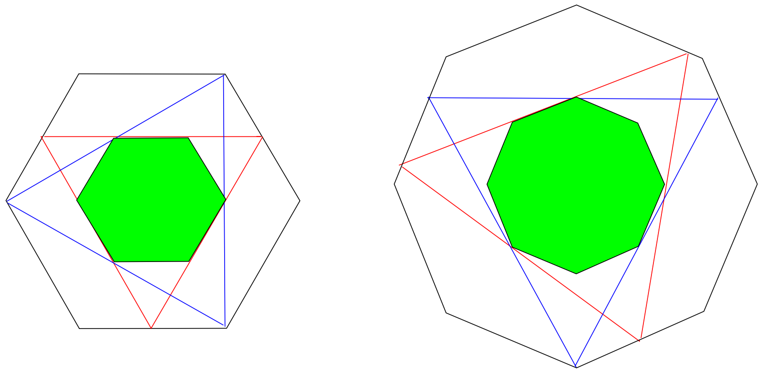

Figure 14

The rotation group of the polygons acts on each triangle of each type, and thus we get p of each, unless there is non-trivial isotropy. This occurs when p is divisible by 3: for example, when $p = 6$ there are only two triangles of each type. Nevertheless, by breaking the symmetry of the configuration by means of a perturbation so small that each of the critical points of type (2) and (3) persists, we can ensure that the coincident triangles separate from one another, and thus, once again, we obtain a maximal configuration. Here we use the fact that the "non-degenerate" or "generic" critical points described in 4.3 do persist under small perturbations.

Addendum Since submitting this paper, we have learned of the paper [4], which discusses the *non-negative rank* of a (non-negative) matrix, and contains a result equivalent to our (2.1). The non-negative rank of a non-negative matrix V is the smallest number of non-negative matrices of rank 1 which can be added together to give V. It is easy to see that the non-negative rank of V is equal to the minimal number of generators of a cone $C \subset \mathbb{R}^N_+$ containing the columns of V.

Acknowledgments We are grateful to Yuliy Baryshnikov for several essential suggestions, and to James Montaldi for a very useful discussion in which we understood the Cevian configuration described in Lemma 4.3 and Remark 4.4, and which was the key to further progress. We thank David Epstein for pointing out the work of Siebenmann and Sullivan ([24]) to us, and we thank Nick Bingham and Ian Stewart for helpful comments on earlier drafts of the paper.

References

[1] J. Bochnak, M. Coste, M.-F. Roy, *Real Algebraic Geometry*, Springer-Verlag, Berlin, 1998

[2] I.A. Bogaevsky, Metamorphoses of singularities of minimum functions, and bifurcations of shock waves of Burger's equation with vanishing viscosity, *Algebra i Analiz*, 1(4):1–16, 1989. English translation in *St. Petersburg (Leningrad) Math. J.*, **1** (1990), no. 4, 807–823.

[3] F.H. Clarke, On the inverse function theorem, *Pacific J. Math.*, 64 (1976) 97-102

[4] J. Cohen and U. Rothblum, Nonnegative rank, decompositions and factorisations of non-negative matrices, *Linear Algebra Appl.* 190 (1993), 149-168

[5] J. Croft, D. Mond and J.Q. Smith, *A systematic approach to some discrete latent variable models*, Warwick Statistics Department Report, (2000)

[6] J. Croft, J.Q. Smith, Discrete mixtures in simple Bayesian networks with hidden variables, *Journal of Computational Statistics and Data Analysis* 41 3-4, (2003) pp 539-547

[7] J. Croft, J.Q. Smith, Bayesian networks for discrete multivariate data: an algebraic inferential approach, *Journal of Multivariate Analysis*, 84, 2 (2003) pp 387-402 (2002).

[8] D. Geiger, D. Heckerman, H. King and C.Meek, *Stratified exponential families: graphical models and model selection*, Technical report MSR-TR-98-31, Microsoft Research centre, WA, USA, (1998).

[9] D. Geiger and C. Meek, Graphical models and exponential families, in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, California, (1998) 156-165

[10] Z. Gilula, Singular value decomposition of probability matrices: probabilistic aspects of latent dichotomous variables, *Biometrika* 66 (1979) 339-344

[11] M. Goresky and R. MacPherson, *Stratified Morse Theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag (1988)

[12] V.V. Goryunov, Monodromy of the image of a mapping, *Functional Analysis and Applications* 25 (1991), 174-180

[13] R.E. Kass, P.W. Vos, *Geometric Foundations of Asymptotic Inference*, Wiley and Sons, New York (1997)

[14] S.L. Lauritzen, *Graphical Models*, Oxford University Press (1996)

[15] V.I. Matov, The topological classification of germs of the maximum and minimax functions of a family of functions in general position, *Uspekhi Mat Nauk* **37** (4) (1982), 167-168 (*Russian Math. Surveys* **37** (4) (1982), 127-128)

[16] John Milnor, *Topology from the Differentiable Viewpoint*, University Press of Virginia, Charlottsville, 1990

[17] John Milnor, *Morse Theory*, Princeton University Press, (1963)

[18] John Milnor, *Singular Points of Complex Hypersurfaces*, Princeton University Press, (1968)

[19] H. Sato, *Algebraic Topology: An Intuitive Approach*, Translations of Mathematical Monographs, Volume 183, American Math. Soc., Providence, Rhode Island, 1999

[20] R. Settimi and J.Q. Smith, On the geometry of Bayesian graphical models with hidden variables, in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, California, 401-408 (1997)

[21] R. Settimi and J.Q. Smith, On the geometry and model selection of Bayesian directed graphs with isolated hidden nodes, University of Chicago Res. Rep. (2000) (submitted to Statistica Sinica)

[22] R. Settimi and J.Q. Smith, Geometry, moments and conditional independence trees with hidden variables, *Ann. Statistics,* **28** 4 (2000), 1179-1205

[23] L. Siebenmann, Deformations of homeomorphisms on stratified sets, *Comm. Math. Helvet.* 47 (1972) 123-163

[24] L. Siebenmann and D. Sullivan, On complexes that are Lipschitz manifolds, *Proceedings of the Georgia Topology Conference, Athens Ga. 1977*, J.C. Cantrell ed., Academic Press (1979), pp 503-525

[25] D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen and R.G. Cowell, *Bayesian analysis in expert systems*, Statistical Science, 8 (1993) 219-282

[26] J. Whittaker, *Graphical models in applied mathematical statistics*, Wiley, Chichester, (1990)

Short title for page headings: Stochastic factorisations