On the Critical Points of Gaussian Mixtures

by

Benjamin Wallace

A thesis submitted to the

Department of Mathematics and Statistics

in conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

July 2013

Abstract

This thesis is concerned with studying the question whether or not Gaussian mixtures have finitely many critical points. The relevance of this problem to the convergence of the mean-shift algorithm is discussed and an overview of some basic properties of the critical points of Gaussian mixtures is provided. Some previous results that are then reviewed include a reduction of this problem in the homoscedastic case and the construction of a very simple mixture with a large but finite number of critical points. A class of counterexamples is then presented that indicate that the inverse function theorem cannot be used to provide a direct solution to this problem. Finally, while the general problem is left unsolved, a proof is obtained in each of two special cases not previously seen in the literature.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Tamás Linder, who began to support and advise me during my undergraduate years and without whom completion of this thesis would not have been possible. He has not only been encouraging in my efforts but also thoughtful and critical with regards to the approaches I have taken in my research and the decisions I have made in my academic life. I believe that these qualities are of the utmost importance in any kind of mentor.

I would also like to thank my parents, who have always supported me and inspired me to pursue my aspirations. They have never asked anything of me in return and for that reason I hope to make them proud.

Contents

List of Figures

1 Introduction

The aim of this thesis is to investigate the following question: do all Gaussian mixtures have finitely many critical points? While the critical points of Gaussian mixtures have been studied in the past, the focus has usually been on finding tight bounds on modality (i.e., number of local maxima), and such bounds have only been obtained in rather restricted settings. In contrast, we seek a very loose bound on the total number of critical points (including minima and saddle points).

A solution to the problem at hand would have direct implications for the so-called "mean-shift algorithm." Thus, for motivational purposes, we begin by introducing this algorithm in Section 2.1. The mean-shift, first introduced in [8], is a mode-seeking algorithm intended for a class of multimodal probability density functions that arise in density estimation; that is, it was designed to find the local maxima of such functions. Knowledge of the locations of such maxima may be used, for instance, in certain clustering schemes (see [4]).

However, the mean-shift is not guaranteed to converge. After reviewing the conditions for convergence given in [12] and demonstrating in Section 2.2 that some basic facts from [13] can be generalized, we introduce Gaussian mixtures in Section 3.1. These form an important class of density functions for which the mean-shift will always converge provided the following conjecture holds: Gaussian mixtures cannot have an infinite number of critical points.

We continue in this section by specializing some of the results from Section 2.2 to the case of Gaussian mixtures and presenting a result of [3] that simplifies our conjecture in the case of homoscedastic mixtures. In Section 3.2, we present an interesting construction of a simple high-modality mixture from [6]. This example is instructive with regards to the unexpected behaviour that Gaussian mixtures can exhibit. However, our main focus is on the fact that, as a rather trivial consequence of a preliminary result in the same paper, this mixture satisfies our conjecture.

The rest of the thesis contains some results and constructions that we have not seen in the literature. In Section 4, we construct a simple class of Gaussian mixtures in order to show that the most direct approach to proving our conjecture (i.e., application of the inverse function theorem) may not always succeed. Unfortunately, we are unable to characterize any interesting situations in which such an approach does succeed.

Finally, in Section 5, we examine some situations under which the number of critical points of a Gaussian mixture can be seen to be finite. It is significantly easier to identify such situations when the critical points lie on a sufficiently smooth (in fact, analytic) curve. Our approach is thus to use some of the basic properties of the critical points to determine conditions under which this occurs. In particular, we show that proportional-covariance mixtures whose component means lie on a straight line and 2-component mixtures with arbitrary covariances have finitely many critical points.

2 Kernel Density Estimation and the Mean-Shift

2.1 Definitions and Convergence Criteria

The following discussion follows in the spirit of [5] and [9] but in the more general setting of [12].

Definition. Let $k : [0, \infty) \to [0, \infty)$ be a non-increasing C^1 function[1] that is not identically 0. By a *kernel* with *profile* k, we mean an integrable function $K : \mathbb{R}^d \to [0, \infty)$ of the form $K(x) = Ck(\|x\|^2)$, where C is a normalizing constant (so that K integrates to 1 over $\mathbb{R}^d$).

Note. Given k as above, define $K : \mathbb{R}^d \to [0, \infty)$ by $K(x) = Ck(\|x\|^2)$, where C is a constant. Then by Lemma C.2, K is integrable over $\mathbb{R}^d$ if and only if $k(t^2)t^{d-1}$ is integrable over $[0, \infty)$.

Definition. Let K be a kernel with profile k and let $H_1, \ldots, H_n$ be symmetric, positive-definite $d \times d$ real matrices. Given a data sample $x_1, \ldots, x_n \in \mathbb{R}^d$ drawn from a probability distribution over $\mathbb{R}^d$ with density function f, a *kernel density estimator* $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ of f with kernel K is a function of the form

$$\hat{f}(x) = \sum_{i=1}^{n} \pi_i K_i(x),$$

where (letting $|A| = |\det(A)|$ for any square matrix A)

$$K_i(x) = |H_i|^{-1/2} K(H_i^{-1/2}(x - x_i)) = C_i k\left((x - x_i)^\top H_i^{-1}(x - x_i)\right),$$

$C_i = C|H_i|^{-1/2}$, and the π_i are elements of $[0, 1]$ satisfying $\displaystyle\sum_{i=1}^{n} \pi_i = 1$.

Note.

1. Since
$$\int_{\mathbb{R}^d} K_i(x)\, \mathrm{d}x = \int_{\mathbb{R}^d} K(H_i^{-1/2}x)|H_i|^{-1/2}\, \mathrm{d}x = \int_{\mathbb{R}^d} K(x)\, \mathrm{d}x = 1,$$
the $|H_i|^{-1/2}$ factor in the above summands ensures that $\hat{f}$ is indeed a probability density.

2. The matrices H_i are sometimes referred to as the bandwidth matrices of $\hat{f}$.

[1] See Appendix C on differentiability at the endpoints of a half-open interval and the notation used below.

For the following discussion, fix a kernel density estimator $\hat{f}$ as above. For convenience, we shall define the norm $\|v\|_i^2 = v^\top H_i^{-1} v$ for $v \in \mathbb{R}^d$.

Assumption 1. Without loss of generality, we shall always assume that $0 < \pi_i < 1$ and that the x_i are distinct.

Definition. We shall refer to the local maxima of a probability density function as its *modes*.

It is often desirable to locate the modes of a density estimate $\hat{f}$. For instance, consider the problem of clustering the data points $x_1, \ldots, x_n$: generally speaking, this involves determining a partition of $\{x_1, \ldots, x_n\}$ such that points in the same partition element share similar features. One approach to clustering begins by postulating that the data has been drawn from a multimodal density function. Then the space in which the data resides can be partitioned into the regions whose points are nearest (in some sense) to the various modes; this then induces a partition of the data.

A mode of $\hat{f}$, being a critical point, must satisfy $\nabla \hat{f}(x) = 0$, where

$$\nabla \hat{f}(x) = \sum_{i=1}^{n} \pi_i \nabla K_i(x)$$

and by symmetry of the H_i,

$$\begin{aligned} \nabla K_i(x) &= C_i k'((x - x_i)^\top H_i^{-1}(x - x_i)) \left(H_i^{-1}(x - x_i) + H_i^{-\top}(x - x_i)\right) \\ &= 2C_i k'(\|x - x_i\|_i^2) H_i^{-1}(x - x_i). \end{aligned} \tag{1}$$

Assumption 2. Suppose that the derivative of the kernel profile k satisfies $k' < 0$.

By the above assumption, if we let $L_i(x) = -2C_i k'\left(\|x - x_i\|_i^2\right) > 0$, then

$$\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}$$

is invertible; we can thus write

$$
\begin{aligned}
\nabla \hat{f}(x) &= \sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}(x_i - x) \\
&= \left[\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}\right] \left[\left(\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}\right)^{-1} \left(\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1} x_i\right) - x\right] \\
&= \left[\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}\right] m(x), \qquad (2)
\end{aligned}
$$

where

$$
\begin{aligned}
m(x) &= \left(\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}\right)^{-1} \left(\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1} x_i\right) - x \\
&= \left(\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}\right)^{-1} \nabla \hat{f}(x).
\end{aligned}
$$

Thus, x is a critical point of $\hat{f}$ if and only if $m(x) = 0$; equivalently, x must be a fixed point of the map $x \mapsto x + m(x)$. This motivates the mean shift algorithm for seeking modes of $\hat{f}$: an initial value $y_1 \in \mathbb{R}^d$ is chosen and the sequence y_j is computed via the iterative algorithm

$$
y_{j+1} = y_j + m(y_j) = \left(\sum_{i=1}^{n} \pi_i L_i(y_j) H_i^{-1}\right)^{-1} \left(\sum_{i=1}^{n} \pi_i L_i(y_j) H_i^{-1} x_i\right). \qquad (3)
$$

If this sequence converges[1] to some $y \in \mathbb{R}^d$, then taking the limit in j on both sides of the first equality above yields $y = y + m(y)$ by continuity of k', so $\nabla \hat{f}(y) = 0$. In other words, if the sequence generated by the mean-shift algorithm converges, then its limit is a critical point of the kernel density estimate.

Note also that by positive-definiteness,

$$
(\nabla \hat{f}(x))^\top m(x) = (\nabla \hat{f}(x))^\top \left(\sum_{i=1}^{n} \pi_i L_i(x) H_i^{-1}\right)^{-1} \nabla \hat{f}(x) \geq 0,
$$

[1]In practice, of course, the y_j may not converge in a finite number of steps, if indeed they converge at all. Thus, one may wish to halt the algorithm once $\|y_{j+1} - y_j\|$ drops below a designated threshold.

with equality if and only if $\nabla \hat{f}(x) = 0$. Thus, $\hat{f}$ at x increases in the $m(x)$ direction.

Following incorrect proofs of convergence in [5] and [1], criteria for convergence of the mean-shift were given in the following theorem of [12]; we reproduce the proof here with some additional clarifications.

Theorem 2.1. *Let $\hat{f}$ satisfy the assumptions above and suppose that k is convex*[2] *and k' is bounded.*

(a) The sequence $\hat{f}(y_j)$ converges.

(b) If y_j converges to y, then y is a critical point of $\hat{f}$.

(c) The sequence y_j converges if the set of critical points of $\hat{f}$ is finite.

Proof. It has already been shown above that (b) holds.

To prove (a), it suffices to show that $\hat{f}(y_j)$ is monotonic, since it is bounded. Now

$$\begin{aligned}
\hat{f}(y_{j+1}) - \hat{f}(y_j) &= \sum_{i=1}^{n} \pi_i (K_i(y_{j+1}) - K_i(y_j)) \\
&= \sum_{i=1}^{n} \pi_i C_i \left(k(\|y_{j+1} - x_i\|_i^2) - k(\|y_j - x_i\|_i^2) \right) \\
&\geq \sum_{i=1}^{n} \pi_i C_i k'(\|y_j - x_i\|_i^2) \left(\|y_{j+1} - x_i\|_i^2 - \|y_j - x_i\|_i^2 \right) \qquad \text{(by convexity)} \\
&= \frac{1}{2} \sum_{i=1}^{n} \pi_i L_i(y_j) \left(\|y_j - x_i\|_i^2 - \|y_{j+1} - x_i\|_i^2 \right).
\end{aligned}$$

But

$$\begin{aligned}
\|y_j - x_i\|_i^2 - \|y_{j+1} - x_i\|_i^2 &= \|y_j - x_i\|_i^2 - \|y_{j+1} - y_j + y_j - x_i\|_i^2 \\
&= \|y_j - x_i\|_i^2 - \|y_{j+1} - y_j\|_i^2 + 2(y_{j+1} - y_j)^\top H_i^{-1}(x_i - y_j) - \|y_j - x_i\|_i^2
\end{aligned}$$

[2] See Appendix C.1.

$$
\begin{aligned}
&= 2(y_{j+1} - y_j)^\top H_i^{-1}(x_i - y_j) - \|y_{j+1} - y_j\|_i^2 \\
&\geq 2(y_{j+1} - y_j)^\top H_i^{-1}(x_i - y_j),
\end{aligned}
$$

so that

$$
\begin{aligned}
\hat{f}(y_{j+1}) - \hat{f}(y_j) &\geq \sum_{i=1}^n \pi_i L_i(y_j)(y_{j+1} - y_j)^\top H_i^{-1}(x_i - y_j) \\
&= (y_{j+1} - y_j)^\top \left[\sum_{i=1}^n \pi_i L_i(y_j) H_i^{-1} x_i - \sum_{i=1}^n \pi_i L_i(y_j) H_i^{-1} y_j \right] \\
&= (y_{j+1} - y_j)^\top \left[\sum_{i=1}^n \pi_i L_i(y_j) H_i^{-1} y_{j+1} - \sum_{i=1}^n \pi_i L_i(y_j) H_i^{-1} y_j \right] \quad \text{(by (3))} \\
&= (y_{j+1} - y_j)^\top \left(\sum_{i=1}^n \pi_i L_i(y_j) H_i^{-1} \right) (y_{j+1} - y_j) \qquad (4) \\
&\geq 0,
\end{aligned}
$$

with equality if and only if $y_j = y_{j+1}$ by positive-definiteness of the H_i and the hypotheses on k'. This proves (a).

Let us turn to the proof of (c). If $y_j = y_{j+1}$ for some j, then we are done, so suppose otherwise; this immediately implies that $\hat{f}(y_j)$ is strictly increasing and that $\nabla \hat{f}(y_j) \neq 0$ for all j. Now from (4),

$$
\hat{f}(y_{j+1}) - \hat{f}(y_j) \geq \sum_{i=1}^n \pi_i L_i(y_j) \|y_{j+1} - y_j\|_i^2 \geq 0.
$$

Thus, since $\hat{f}(y_{j+1}) - \hat{f}(y_j) \to 0$, either $\|y_{j+1} - y_j\|_i \to 0$ or

$$
\sum_{i=1}^n \pi_i L_i(y_j) \to 0.
$$

But since $k' < 0$ and k is convex, the latter is only possible if y_j gets arbitrarily far from the x_i; this would in turn imply that $\hat{f}(y_j) \to 0$, contradicting the fact that this quantity is increasing. Thus[3],

$$y_{j+1} - y_j \to 0.$$

It follows from (2) and then (3) that

$$\nabla \hat{f}(y_j) = \left(\sum_{i=1}^{n} \pi_i L_i(y_j) H_i^{-1} \right) m(y_j) = \left(\sum_{i=1}^{n} \pi_i L_i(y_j) H_i^{-1} \right) (y_{j+1} - y_j) \to 0$$

since the L_i are bounded by hypothesis.

Now suppose $\hat{f}$ has finitely many critical points $z_1, \ldots, z_N$. Since $\hat{f}$ is bounded, the set $S = \{x \in \mathbb{R}^d : \hat{f}(x) \geq \hat{f}(y_2) > 0\}$ is compact. Thus, $\|\nabla \hat{f}\|$ is bounded away from 0 on $S \setminus \bigcup_{i=1}^{N} B(z_i, \epsilon)$, where $\epsilon > 0$ is such that the $\overline{B}(z_i, \epsilon)$ are disjoint[4]. It follows from the fact that $\nabla \hat{f}(y_j) \to 0$ that for large enough j, y_j is in either S^c or $\bigcup_{i=1}^{N} B(z_i, \epsilon)$; since $\hat{f}(y_j)$ is monotonically increasing, it must be the latter: $y_j \in \bigcup_{i=1}^{N} B(z_i, \epsilon)$. That y_j is in only one of the $B(z_i, \epsilon)$ for large j then follows from the facts that these sets are separated by a positive distance and that $y_{j+1} - y_j \to 0$. □

2.2 Basic Properties of the Critical Points

The above proposition motivates an investigation of the set of critical points of various kernels. For instance, rearranging the equation $\nabla \hat{f}(x) = 0$ and using the assumptions made above yields the following.

Proposition 2.2. *A point $x \in \mathbb{R}^d$ is a critical point of $\hat{f}$ if and only if*

$$x = \left(\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1} x_i \right). \tag{5}$$

[3]Note that by the equivalence of norms (Theorem A.5), the fact that we are using $\|\cdot\|_i$ here is immaterial.

[4]Here $B(z_i, \epsilon)$ and $\overline{B}(z_i, \epsilon)$ denote the open and closed balls (respectively) of radius ϵ about z_i; see Appendix B for their definitions.

Proof. We have

$$\begin{aligned}
\nabla \hat{f}(x) = 0 &\Leftrightarrow \sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1}(x - x_i) = 0 \\
&\Leftrightarrow \left(\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1} \right) x = \sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1} x_i \\
&\Leftrightarrow x = \left(\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1} x_i \right)
\end{aligned}$$

since

$$\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2) H_i^{-1}$$

is invertible by our assumptions. □

An observation made in [13] in the case of Gaussian kernels (which we introduce in the next subsection) but that holds in our more general setting is that dividing the coefficients $\pi_i C_i k'(\|x - x_i\|_i^2)$ of both sums in (5) by $\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2)$ (which is non-zero by Assumption 2) leaves (5) unchanged. Writing,

$$\alpha_i = \alpha_i(x) = \frac{\pi_i C_i k'(\|x - x_i\|_i^2)}{\sum_{i=1}^{n} \pi_i C_i k'(\|x - x_i\|_i^2)},$$

this means that the critical points x satisfy

$$x = \left(\sum_{i=1}^{n} \alpha_i H_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i H_i^{-1} x_i \right).$$

Corollary 2.3. *The critical points of the kernel density estimate $\hat{f}$ lie in the image of the standard $(n-1)$-simplex*[5] *under the map*

$$(\alpha_1, \ldots, \alpha_n) \mapsto \left(\sum_{i=1}^{n} \alpha_i H_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i H_i^{-1} x_i \right). \tag{6}$$

Corollary 2.4. *The set S of critical points of the kernel density estimate $\hat{f}$ is finite if and only if it is discrete*[6]*.*

[5] See Appendix A.1.

[6] See Appendix B for the definition.

Proof. Clearly, if S is finite, then it is discrete. For the converse, suppose that S is discrete. By the previous corollary, compactness of the $(n-1)$-simplex, and continuity of (6), S is a subset of a compact set. Thus, if S is infinite, then it has a limit point x by the Bolzano-Weierstrass theorem (Theorem B.1). But since $\nabla \hat{f}$ is assumed to be continuous, S is closed, hence contains x; but x is not isolated in S, so this contradicts the fact that S is discrete. □

3 Gaussian Kernels and Mixtures

3.1 Definitions and Basic Properties

In this section, we specialize some of the above results and discussion to the important case of Gaussian kernels.

Definition. The *Gaussian kernel* G over $\mathbb{R}^d$ is the kernel over $\mathbb{R}^d$ with profile $g(t) = e^{-t/2}$.

Note. The profile of the Gaussian kernel satisfies all the assumptions of the previous section.

Thus, the Gaussian kernel G has the form

$$G(x) = Ce^{-x^\top x/2},$$

where, in this case, $C = (2\pi)^{-d/2}$.

Definition. The d-dimensional *Gaussian density function* with *mean* $\mu \in \mathbb{R}^d$ and *covariance matrix* $\Sigma \in \mathbb{R}^{d\times d}$ (which is required to be a positive-definite symmetric matrix) is the function $h : \mathbb{R}^d \to \mathbb{R}$ given by

$$h(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

A d-dimensional *Gaussian mixture* with n components is a convex combination of n Gaussian densities in $\mathbb{R}^d$.

Concretely, a Gaussian mixture is a probability density function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = \sum_{i=1}^{n} \pi_i f_i(x),$$

where

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)\right) = |\Sigma_i|^{-1/2} g(\|\Sigma_i^{-1/2}(x-\mu_i)\|^2),$$

where g is the Gaussian kernel profile. Thus, Gaussian mixtures arise as kernel density estimators that use the Gaussian kernel. We call the f_i the *components* of f and the μ_i and Σ_i the *component means* and *component covariances*, respectively.

Definition. A *proportional-covariance* Gaussian mixture is one whose component covariance matrices Σ_i are proportional, i.e., for which there exists a matrix Σ and constants σ_i^2 such that $\Sigma_i = \sigma_i^2 \Sigma$ for each i. Proportional-covariance mixtures with such covariance matrices include as special cases

- *homoscedastic* mixtures, in which $\sigma_i^2 = 1$ for each i and
- *isotropic* mixtures, in which $\Sigma = I$ (but the σ_i^2 may be arbitrary).

By Theorem 2.1, the mean-shift algorithm using a Gaussian kernel will always converge if the following is true.

Conjecture A. *Any Gaussian mixture has a finite set of critical points.*

One might naïvely suppose that a Gaussian mixture with n components simply has n modes. This is clearly false as a mixture in which the component means are sufficiently close to one another may only have one mode. An example of this phenomenon will be provided in Section 4. Even worse, a mixture with n components may have more than n modes. Situations in which this occurs will be discussed in Section 3.2.

It was pointed out in [3] that the modality of a homoscedastic mixture is equal to the modality of an appropriate mixture of standard Gaussians. The focus there was on local maxima, but the same proof applies to all critical points.

Theorem 3.1. *The critical points of a homoscedastic Gaussian mixture f with component covariances Σ and component means $\mu_1, \ldots, \mu_n$ are in one-to-one correspondence with the critical points of a homoscedastic isotropic mixture with component covariances I whose component means are related to the μ_i by a non-singular linear map.*

Proof. Using the spectral theorem, write $\Sigma^{-1} = U\Lambda U^\top$. Let $y : \mathbb{R}^d \to \mathbb{R}^d$ be the linear change of coordinates given by $y(x) = U\Lambda^{-1/2}x$. Then

$$
\begin{aligned}
(y(x) - \mu_i)^\top \Sigma^{-1}(y(x) - \mu_i) &= (U\Lambda^{-1/2}x - \mu_i)^\top \Sigma^{-1}(U\Lambda^{-1/2}x - \mu_i) \\
&= (U\Lambda^{-1/2}x)^\top \Sigma^{-1}(U\Lambda^{-1/2}x) - 2(U\Lambda^{-1/2}x)^\top \Sigma^{-1}\mu_i + \mu_i^\top \Sigma^{-1}\mu_i \\
&= x^\top \Lambda^{-1/2}U^\top \Sigma^{-1}U\Lambda^{-1/2}x - 2x^\top \Lambda^{-1/2}U^\top \Sigma^{-1}\mu_i + \mu_i^\top U\Lambda U^\top \mu_i \\
&= x^\top x - 2x^\top \Lambda^{1/2}U^\top \mu_i + (\Lambda^{1/2}U^\top \mu_i)^\top(\Lambda^{1/2}U^\top \mu_i) \\
&= \|x - \Lambda^{1/2}U^\top \mu_i\|^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
f(y(x)) &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\sum_{i=1}^{n} \pi_i \exp\left(-\frac{1}{2}(y(x) - \mu_i)^\top \Sigma^{-1}(y(x) - \mu_i)\right) \\
&= (2\pi)^{-d/2}|\Sigma|^{-1/2}\sum_{i=1}^{n} \pi_i \exp\left(-\frac{1}{2}\|x - \Lambda^{1/2}U^\top \mu_i\|^2\right)
\end{aligned}
$$

$$= |\Sigma|^{-1/2} g(x),$$

where

$$g(x) = (2\pi)^{-d/2} \sum_{i=1}^{n} \pi_i \exp\left(-\frac{1}{2}\|x - \Lambda^{1/2} U^\top \mu_i\|^2\right)$$

is a homoscedastic isotropic Gaussian mixture with identity covariance and component means $\Lambda^{1/2} U^\top \mu_i$. Moreover,

$$|\Sigma|^{-1/2} \nabla g(x) = \nabla(f \circ y)(x) = (\nabla y(x))^\top \nabla f(y(x)) = \Lambda^{-1/2} U^\top \nabla f(U \Lambda^{-1/2} x),$$

so since $\Lambda^{-1/2} U^\top$ is non-singular, x is a critical point of $f \circ y$, or equivalently of g, if and only if $U \Lambda^{-1/2} x$ is a critical point of f. □

Note. In the case of isotropic homoscedastic mixtures, the component covariance matrix Σ is already diagonal so we get $U = I$ and $\Lambda = \Sigma = \sigma^2 I$ above (for some σ). It follows that x is a critical point of the mixture with identity component covariances and component means $\sigma \mu_i$ if and only if $\frac{1}{\sigma} x$ is a critical point of the mixture with component covariances $\sigma^2 I$ and component means μ_i.

We thus have the following.

Corollary 3.2. *The homoscedastic case of Conjecture A is equivalent to the homoscedastic isotropic case.*

Despite this, we will discuss in the following section how even homoscedastic isotropic mixtures can exhibit non-trivial behaviour. For convenience, we state the following specializations of the results of Section 2.2 to the case of Gaussian mixtures before proceeding.

Since the derivative of the Gaussian kernel's profile $g(x) = e^{-x/2}$ is given by $g'(x) = -\frac{1}{2}g(x)$,

$$\nabla f_i(x) = -f_i(x)\Sigma_i^{-1}(x - \mu_i) \tag{7}$$

and Proposition 2.2 and Corollaries 2.3 and 2.4 reduce to the following.

Proposition 3.3. *Let $f = \sum_{i=1}^{n} \pi_i f_i$ be a Gaussian mixture, where the component f_i has mean μ_i and covariance Σ_i. Then x is a critical point of f if and only if*

$$x = \left(\sum_{i=1}^{n} \pi_i \Sigma_i^{-1} f_i(x) \right)^{-1} \left(\sum_{i=1}^{n} \pi_i \Sigma_i^{-1} \mu_i f_i(x) \right).$$

Corollary 3.4. *If x is a critical point of the Gaussian mixture f, then*

$$x = \left(\sum_{i=1}^{n} \alpha_i \Sigma_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i \Sigma_i^{-1} \mu_i \right), \tag{8}$$

for some $\alpha_i \in [0,1]$ satisfying $\sum_{i=1}^{n} \alpha_i = 1$. Thus, the critical points of f lie in the image of the standard $(n-1)$-simplex under the map

$$(\alpha_1, \ldots, \alpha_n) \mapsto \left(\sum_{i=1}^{n} \alpha_i \Sigma_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i \Sigma_i^{-1} \mu_i \right).$$

Note. The last result is stated as Theorem 1 in [13]; the authors of this paper refer to the set containing the critical points of f as the "ridgeline manifold" of f.

The proof of the following corollary is the same as that of Corollary 2.4.

Corollary 3.5. *The set of critical points of a Gaussian mixture is finite if and only if it is discrete.*

3.2 Mixtures With Strong Symmetry Assumptions

As noted in [9], Conjecture A is trivial when $d = 1$ (this follows from Corollary D.5 of the current thesis). Even more is known about the modes in this case: it was shown in [3] that a 1-dimensional Gaussian mixture with n components has at most n modes.

However, the situation is more complicated in higher dimensions. For instance, the main result in [14] states that a mere 2-component mixture in $d > 1$ dimensions can have at most $d + 1$ modes[7] and that this bound is tight; the contour plot of a 2-component mixture with 3 modes constructed in [13] is shown in Figure 1. As can be seen, this construction requires that the components of the mixture have very different covariances; indeed, Corollary 1 of [14] states that 2-component proportional-covariance mixtures can only have at most 2 modes.

One might thus hope that restrictions on the covariance matrices would improve the situation for larger numbers of components. In fact, it was conjectured in [3] that the number of modes of a homoscedastic or isotropic mixture is bounded by the number of components of the mixture. A counterexample to this conjecture was later found by the same authors and presented in [2] and then generalized and studied more deeply in [6] and [7] (we shall address some of these last results below).

Nevertheless, the results of [6] make clear the fact that their high-modality mixtures have only finitely many critical points; this is a direct consequence of their "axes lemma," which is the focus of this section due to its direct relevance to Conjecture A. First, we need the following "coordinate transformation lemma" of [6]. The reader may wish to consult

[7] It seems likely that the methods of [13] and [14] could also be used to study the number of minima and saddle points of 2-component mixtures.

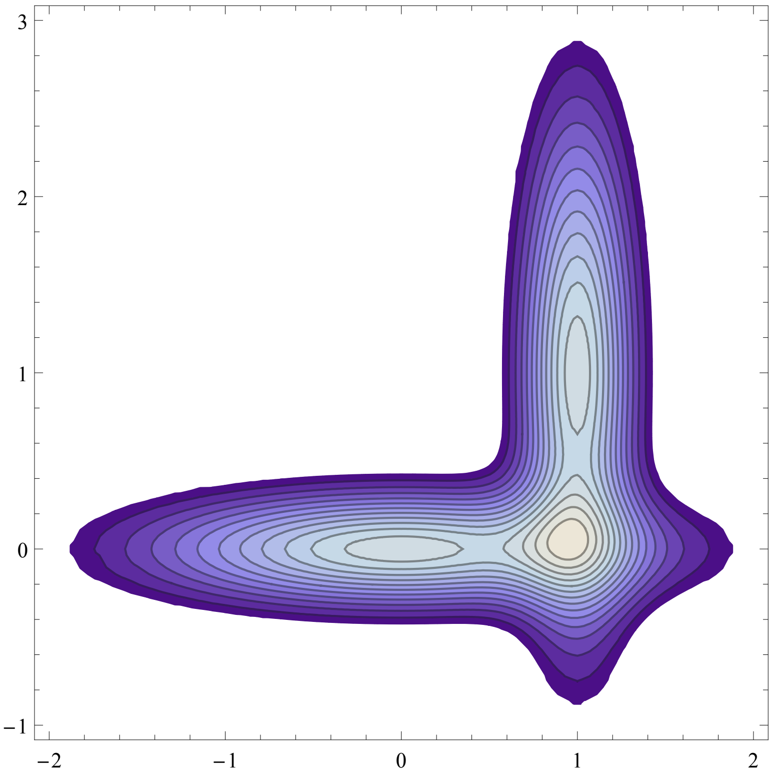

Figure 1: Contour plot of a Gaussian mixture in 2 dimensions with 2 components and 3 modes. See [13] for details.

Appendix A.1 for some terminology before proceeding.

Lemma 3.6. *Let x be an element of the scaled standard n-simplex cS_n. Then the barycentric coordinates $\alpha_i = \alpha_i(x)$ of x are given by*

$$\alpha_i = \frac{1}{n+1} + \frac{1}{2(n+1)c^2}\left(\sum_{j=1}^{n+1} \|x - ce_j\|^2 - (n+1)\|x - ce_i\|^2\right).$$

Proof. For $i \neq j$, let L_{ij} denote the 1-face of cS_n spanned by ce_i and ce_j and note that $\|ce_i - ce_j\|^2 = 2c^2$. Let $p_{ij} : cS_K \to L_{ij}$ be the orthogonal projection map onto L_{ij} (so that $x - p_{ij}(x)$ is orthogonal to $e_j - e_i$) and define $x_{ij} = x_{ij}(x) = \dfrac{1}{c\sqrt{2}}\|ce_j - p_{ij}(x)\|$. Let us show that

$$x_{ij} = \frac{1}{2} + \frac{1}{4c^2}(\|x - ce_j\|^2 - \|x - ce_i\|^2). \tag{9}$$

First note that the right-hand side depends only on x through $p_{ij}(x)$ since

$$\begin{aligned}\|x - ce_j\|^2 - \|x - ce_i\|^2 &= (\|x - p_{ij}(x)\|^2 + \|p_{ij}(x) - ce_j\|^2) - (\|x - p_{ij}(x)\|^2 + \|p_{ij}(x) - ce_i\|^2) \\ &= \|p_{ij}(x) - ce_j\|^2 - \|p_{ij}(x) - ce_i\|^2.\end{aligned}$$

It thus suffices to verify (9) in the case that $x = p_{ij}(x)$; in this case, we can write $x = (1-t)ce_j + tce_i$ for some $t \in [0,1]$. It follows that

$$\|x - ce_i\|^2 = (1-t)^2\|ce_j - ce_i\|^2 = 2(1-t)^2c^2$$

and

$$\|x - ce_j\|^2 = t^2\|ce_j - ce_i\|^2 = 2t^2c^2,$$

so that the right-hand side of (9) becomes

$$\frac{1}{2} + \frac{1}{4c^2}(\|x - ce_j\|^2 - \|x - ce_i\|^2) = \frac{1}{2} + \frac{2c^2}{4c^2}(2t-1) = t,$$

which agrees with the definition of x_{ij} in this case:

$$x_{ij} = \frac{1}{c\sqrt{2}}\|x - ce_j\| = \frac{tc\sqrt{2}}{c\sqrt{2}} = t.$$

Next let $b_i(c)$ denote the barycenter of the $(n-1)$-face of cS_n complementary to the 0-face ce_i, i.e.

$$b_i(c) = \sum_{\substack{k=1 \\ k\neq i}}^{n+1} \frac{1}{n}ce_k,$$

so that

$$\begin{aligned}\|ce_i - b_i(c)\|^2 &= \frac{n+1}{n}c^2, \qquad (10)\\ b_i(c)^\top e_j &= \frac{c}{n}, \quad (j \neq i),\end{aligned}$$

and $b_i(c)^\top e_i = 0$. Let θ be the angle between L_{ij} and the line segment L_i from ce_i to $b_i(c)$, i.e.

$$\pm\cos\theta = \frac{(v_1 - v_2)^\top (w_1 - w_2)}{\|v_1 - v_2\|\|w_1 - w_2\|}$$

for any $v_1, v_2 \in L_{ij}$ and $w_1, w_2 \in L_i$. In particular,

$$\begin{aligned}
\cos\theta &= \frac{(b_i(c) - ce_i)^\top (ce_j - ce_i)}{\|b_i(c) - ce_i\|\|ce_j - ce_i\|} \\
&= \frac{b_i(c)^\top ce_j + c^2}{c\sqrt{2}\|b_i(c) - ce_i\|} \\
&= \frac{b_i(c)^\top e_j + c}{\sqrt{2}\|b_i(c) - ce_i\|}.
\end{aligned}$$

But by (10),

$$b_i(c)^\top e_j + c = \frac{c}{n} + c = \left(\frac{n+1}{n}\right) c = \frac{\|ce_i - b_i(c)\|^2}{c},$$

so

$$\cos\theta = \frac{\|ce_i - b_i(c)\|^2}{c\sqrt{2}\|ce_i - b_i(c)\|} = \frac{\|ce_i - b_i(c)\|}{c\sqrt{2}} = \frac{c\|b_i(1) - e_i\|}{c\sqrt{2}} = \frac{\|e_i - b_i(1)\|}{\sqrt{2}},$$

where the third equality also follows from (10). Note that the final expression for $\cos\theta$ obtained above is (again by (10)) independent of i and j.

Another expression for $\cos\theta$ may be obtained when $x \in L_i$; in this case, the barycentric coordinates α_k of x for $k \neq i$ are all equal (since x is a convex combination of $b_i(c)$ and ce_i) and so we can write

$$x = \sum_{\substack{j=1 \\ j\neq i}}^{n+1} \frac{1-\alpha_i}{n} ce_j + \alpha_i ce_i = (1-\alpha_i) b_i(c) + \alpha_i ce_i.$$

It follows that

$$\begin{aligned}
\cos\theta &= \frac{(x - ce_i)^\top (p_{ij}(x) - ce_i)}{\|x - ce_i\|\|p_{ij}(x) - ce_i\|} \\
&= \frac{(x - p_{ij}(x))^\top (p_{ij}(x) - ce_i) + p_{ij}(x)^\top (p_{ij}(x) - ce_i) - ce_i^\top (p_{ij}(x) - ce_i)}{(1-\alpha_i)\|b_i(c) - ce_i\|\|p_{ij}(x) - ce_i\|}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\|p_{ij}(x) - ce_i\|}{(1-\alpha_i)\|b_i(c) - ce_i\|} \\
&= \frac{\|ce_j - ce_i\| - \|ce_j - p_{ij}(x)\|}{(1-\alpha_i)\|b_i(c) - ce_i\|} \\
&= \frac{c\sqrt{2}(1 - x_{ij})}{(1-\alpha_i)\|b_i(c) - ce_i\|} \\
&= \frac{\sqrt{2}(1 - x_{ij})}{(1-\alpha_i)\|b_i(1) - e_i\|},
\end{aligned}
$$

where the third equality follows from orthogonality of $x - p_{ij}(x)$ and $p_{ij}(x) - ce_i$, the fourth equality follows from the fact that $p_{ij}(x)$ lies on L_{ij}, and the last equality follows from (10).

Setting the two expressions for $\cos\theta$ above equal to each other and rearranging yields

$$
\|b_i(1) - e_i\|^2(1-\alpha_i) = 2(1 - x_{ij})
$$

for $x \in L_i$. Summing both sides of this equation over all $j \neq i$ yields

$$
n\|b_i(1) - e_i\|^2(1-\alpha_i) = 2n - 2\sum_{\substack{j=1 \\ j\neq i}}^{n+1} x_{ij} \tag{11}
$$

for $x \in L_i$. However, the left-hand side (hence also the right-hand side) of this equality depends on x only through its orthogonal projection $p_i(x)$ onto L_i; this follows from the fact that α_i is constant along the hyperplane P orthogonal to L_i and containing $p_i(x)$.

To see this, note that the vertices v_j of the $(n-1)$-simplex formed by intersecting P with cS_n are all equally distant from ce_i so can be written as

$$
v_j = \beta ce_i + (1-\beta)ce_j
$$

for some $\beta \in [0,1]$. Thus, $x \in P \cap cS_n$ has the form

$$
x = \sum_{j\neq i} \gamma_j v_j = \beta\left(\sum_{j\neq i}\gamma_j\right) ce_i + \sum_{j\neq i}\gamma_j(1-\beta)ce_j = \beta ce_i + \sum_{j\neq i}\gamma_j(1-\beta)ce_j,
$$

where $\sum_{j \neq i} \gamma_j = 1$. Since $\beta + \sum_{j \neq i} \gamma_j (1-\beta) = 1$, this means that the i-th barycentric coordinate of x in cS_n is given by $\alpha_i = \beta$; that is, α_i is the same for all $x \in P$.

It follows that (11) holds for all $x \in cS_n$. Simplifying this equation using our expressions for $\|b_i - e_i\|$ and x_{ij}, we get

$$\begin{aligned}(n+1)(1-\alpha_i) &= n + \frac{1}{2c^2}\left(n\|x - ce_i\|^2 - \sum_{j \neq i} \|x - ce_j\|^2\right) \\ &= n + \frac{1}{2c^2}\left((n+1)\|x - ce_i\|^2 - \sum_{j=1}^{n+1} \|x - ce_j\|^2\right).\end{aligned}$$

Rearranging this yields the desired expression for α_i. $\square$

Following [6], consider the homoscedastic isotropic $(n+1)$-component Gaussian mixture f in $\mathbb{R}^{n+1}$ with component covariances $\sigma^2 I$, component means $\mu_i = ce_i$, and weights $\pi_i = \dfrac{1}{n+1}$.

The main result of [6] is that, for appropriately chosen values of c, the mixture f can have $n+2$ modes and a number of critical points that grows exponentially in n. Here, though, we are more interested in the locations of the critical points.

Definition. An *axis* of an n-simplex S spanned by $v_1, \ldots, v_{n+1}$ is a line segment connecting a barycenter of a k-face F of S (for $k < n$) to the barycenter of the $(n-k-1)$-face of S complementary to F.

Note. Suppose x lies on an axis of an n-simplex S, i.e., suppose x can be written as a convex combination of the barycenter of a k-face of S and the barycenter of the complementary $(n-k-1)$-face. Equivalently, $k+1$ of the barycentric coordinates of x are a multiple of $\dfrac{1}{k+1}$ and the remaining $n-k$ are a multiple of $\dfrac{1}{n-k}$. Thus, x lies on an axis if and only if its barycentric coordinates take on at most two distinct values. When $S = cS_n$, this is equivalent by Lemma 3.6 to $\|x-ce_i\|$ taking on at most two distinct values as i runs through $1,\ldots,n+1$.

Theorem 3.7. *The critical points of the mixture f lie on the axes of the scaled standard n-simplex cS_n.*

Proof. The case $n=1$ follows directly from Corollary 3.4, so suppose $n \geq 2$. Moreover, by the note following Theorem 3.1, it suffices to consider a fixed value of σ^2; for simplicity, take $\sigma^2 = \dfrac{1}{2\pi}$ so that the components f_i of f have the form

$$f_i(x) = e^{-\pi\|x-\mu_i\|^2}.$$

Let x be a critical point of f, so that the barycentric coordinates of x are given by $\alpha_i = \dfrac{f_i(x)}{f(x)}$ by Proposition 3.3. Suppose by way of contradiction that x does not lie on an axis, so that by the preceding note, for some i, j, and k, we have $\|x-se_i\| < \|x-se_j\| < \|x-se_k\|$. Then by the previous lemma,

$$\alpha_i - \alpha_k = \frac{1}{2(n+1)c^2}((n+1)(\|x-ce_k\|^2 - \|x-ce_i\|^2)) = \frac{1}{2c^2}(\|x-ce_k\|^2 - \|x-ce_i\|^2).$$

Thus,

$$\frac{1}{2c^2}(\|x-ce_k\|^2 - \|x-ce_i\|^2) = \frac{f_i(x)-f_k(x)}{f(x)} = \frac{e^{-\pi\|x-ce_i\|^2} - e^{-\pi\|x-ce_k\|^2}}{f(x)}$$

and similarly,

$$\frac{1}{2c^2}(\|x-ce_k\|^2 - \|x-ce_j\|^2) = \frac{e^{-\pi\|x-ce_j\|^2} - e^{-\pi\|x-ce_k\|^2}}{f(x)}.$$

It follows that

$$-\frac{f(x)}{2c^2} = \frac{e^{-\pi\|x-ce_k\|^2} - e^{-\pi\|x-ce_i\|^2}}{\|x-ce_k\|^2 - \|x-ce_i\|^2}$$

$$= \frac{e^{-\pi\|x-ce_k\|^2} - e^{-\pi\|x-ce_j\|^2}}{\|x-ce_k\|^2 - \|x-ce_j\|^2}.$$

In other words, letting $t_l = \|x - ce_l\|^2$ for $l = i, j, k$, we have

$$\frac{g(t_k) - g(t_i)}{t_k - t_i} = \frac{g(t_k) - g(t_j)}{t_k - t_j},$$

where $t_i < t_j < t_k$ and $g(t) = e^{-\pi t}$. But this contradicts the strict convexity of g (see Theorem C.3). □

As will be shown in Section 5, the axes lemma implies that the mixture f has finitely many critical points.

A somewhat tangential line of inquiry suggested by the axes lemma is an investigation of Gaussian mixtures satisfying certain symmetry conditions. For instance, what can we say about the critical points of a homoscedastic isotropic mixture with equal weights whose component means are placed at the vertices of a regular polytope? Due to the rather opaque nature of the proof of the axes lemma (which seems to stem from its reliance upon a rather complicated change of coordinates), it is not entirely clear how to pursue an investigation of this nature. Perhaps a good place to start would with an elucidation of the role that symmetry plays in the locations of the critical points.

4 Mixtures with a Degenerate Critical Point

The most direct approach to proving Conjecture A is to use the following corollary to the inverse function theorem (Theorem C.1).

Corollary 4.1. *Let $f : U \to \mathbb{R}^d$, where $U \subseteq \mathbb{R}^d$ is open. Suppose f is C^2 about one of its critical points x. If x is non-degenerate, then it is isolated.*

Proof. The hypotheses of the inverse function theorem are satisfied with ∇f in place of f, so ∇f is injective in a neighbourhood of x. Thus, in this neighbourhood, there is no $y \neq x$ such that $\nabla f(y) = 0$. $\square$

The purpose of this section is to show that that Corollary 4.1 does not suffice to prove Conjecture A. Let $f = \sum_{i=1}^{n} \pi_i f_i$ be a Gaussian mixture, where the component f_i has mean μ_i and covariance Σ_i. Recalling (7), we compute the Hessian[8]

$$\begin{aligned} Hf_i(x) &= -\Sigma_i^{-1}(If_i(x) + (x-\mu_i)Df_i(x)) \\ &= -\Sigma_i^{-1}(If_i(x) - (x-\mu_i)(x-\mu_i)^\top \Sigma_i^{-1} f_i(x)) \\ &= \Sigma_i^{-1}((x-\mu_i)(x-\mu_i)^\top \Sigma_i^{-1} - I) f_i(x). \end{aligned}$$

Thus, the Hessian of f is

$$Hf(x) = \sum_{i=1}^{n} \pi_i \Sigma_i^{-1}((x-\mu_i)(x-\mu_i)^\top \Sigma_i^{-1} - I) f_i(x).$$

Unfortunately, it is not clear how to characterize all the situations under which $Hf(x)$ degenerates, even when we restrict our attention to the case of x a critical point. It is not too hard, however, to present a simple class of mixtures with a degenerate critical point.

For instance, consider the mixture $f = \dfrac{f_1 + f_2}{2}$ with parameters

$$n = 2, \Sigma_1 = \Sigma_2 = \sigma^2 I, \pi_1 = \pi_2 = 1/2, \mu_1 = 0, \text{ and } \mu_2 = \mu,$$

for some $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$. By Proposition 3.3, x is a critical point of f if and only if

$$x = \frac{f_2(x)\mu}{2f(x)}.$$

[8] See Appendix C for the notation used here.

Since $f_1(\mu/2) = f_2(\mu/2) = f(\mu/2)$, it is easy to see that $x = \mu/2$ is a critical point of f. Moreover, the Hessian of this mixture is

$$Hf(x) = \frac{1}{2}\sigma^{-2}(xx^\top\sigma^{-2} - I)f_1(x) + \frac{1}{2}\sigma^{-2}((x-\mu)(x-\mu)^\top\sigma^{-2} - I)f_2(x).$$

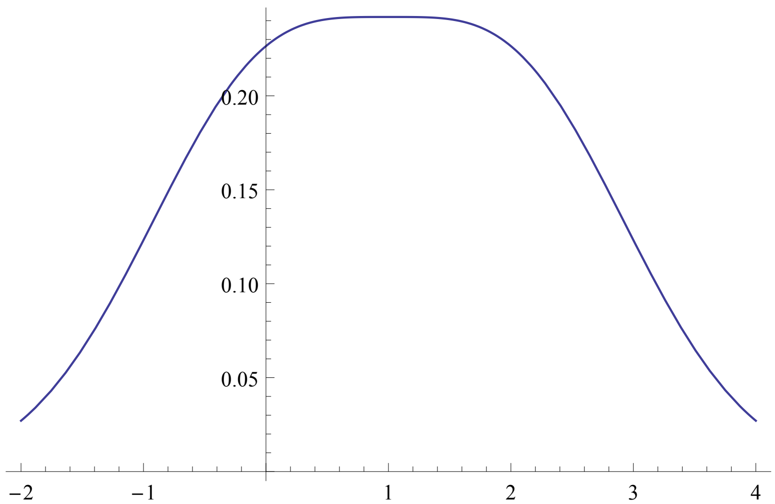

Figure 2: Gaussian mixture with a degenerate critical point at $x = 1$.

Therefore,

$$\begin{aligned}\sigma^2 Hf(\mu/2) &= \frac{1}{2}\left(\frac{1}{4}\mu\mu^\top\sigma^{-2} - I\right)f_1(\mu/2) + \frac{1}{2}\left(\frac{1}{4}\mu\mu^\top\sigma^{-2} - I\right)f_2(\mu/2)\\ &= \frac{1}{2}f(\mu/2)\left(\frac{1}{4}\mu\mu^\top\sigma^{-2} - I\right),\end{aligned}$$

so by Lemma A.6

$$\left(\frac{2\sigma^2}{f(\mu/2)}\right)^d \det Hf(\mu/2) = \det\left(\frac{1}{4}\mu\mu^\top\sigma^{-2} - I\right) = \frac{1}{4\sigma^2}\|\mu\|^2 - 1.$$

Thus, $Hf(\mu/2)$ degenerates when $\|\mu\| = 2\sigma$. Note that this is the largest value of $\|\mu\|$ for which Corollary 4(a) of [13] allows us to deduce that the mixture is unimodal. The case where $d = 1, \sigma = 1$, and $\mu = 2$ is plotted in Figure 2.

5 Critical Points Lying on an Analytic Curve

One can attempt to study the critical points of Gaussian mixtures by looking at cases in which an analytic curve passes through them. Note that given any finite set of points $x^{(1)}, \ldots, x^{(N)} \in \mathbb{R}^d$, one can fix distinct $t_1, \ldots, t_N \in \mathbb{R}$ and let $p_i : \mathbb{R} \to \mathbb{R}$ be the polynomial whose graph passes through the points $(t_1, x_i^{(1)}), \ldots, (t_n, x_i^{(N)}) \in \mathbb{R} \times \mathbb{R}$. Then the curve $p : \mathbb{R} \to \mathbb{R}^d$ whose components are the p_i has a graph passing through $(t_1, x^{(1)}), \ldots, (t_n, x^{(N)}) \in \mathbb{R} \times \mathbb{R}^d$. With regards to the critical points of a Gaussian mixture, we have the following partial converse.

Proposition 5.1. *Suppose an analytic curve $x : [a, b] \to \mathbb{R}^d$ passes through the critical points of a Gaussian mixture f. Then f has finitely many critical points as long as $f \circ x$ is non-constant.*

Proof. By hypothesis, we have

$$\{y \in \mathbb{R}^d : Df(y) = 0\} = x(\{t \in [a, b] : (Df)(x(t)) = 0\})$$

and since $(f \circ x)'(t) = (Df)(x(t))x'(t)$,

$$x(\{t : (Df)(x(t)) = 0\}) \subseteq x(\{t : (f \circ x)'(t) = 0\}) = x(S),$$

where $S = \{t \in [a, b] : (f \circ x)'(t) = 0\}$. But the composition $f \circ x : [a, b] \to \mathbb{R}$, being given by

$$f(x(t)) = \sum_{i=1}^{n} \pi_i C_i \exp\left(-\frac{1}{2} \sum_{i,j=1}^{d} a_{ij}^{(k)} (x(t) - \mu_k)_i (x(t) - \mu_k)_j\right),$$

where $\Sigma_k^{-1} = \left(a_{ij}^{(k)}\right)$, is analytic by Theorem D.2; hence, its set of critical points S is discrete as long as it is non-constant. Moreover, $(f \circ x)'$ is continuous, so S is closed. Thus, when

$f \circ x$ is non-constant, S is finite; otherwise, it would contain a limit point, contradicting the fact that it is discrete. It follows that $x(S)$ is finite and so the set of critical points of f is a subset of a finite set. $\square$

Corollary 5.2. *If the critical points of a Gaussian mixture f lie on a straight line, there are finitely many of them.*

Proof. First note that $f(x) \neq 0$ and $f(x) \to 0$ as $\|x\| \to \infty$ (i.e., f is non-constant along straight lines). Now by Corollary 3.4, the critical points of f lie in a compact set; so by hypothesis, they lie on a line segment of finite length. Thus, letting $x : [a, b] \to \mathbb{R}^d$ be a sufficiently long line segment (so that $f \circ x$ is non-constant), the result follows from the previous proposition. $\square$

Corollary 5.3. *The mixture f considered in Theorem 3.7 has finitely many critical points.*

Proof. This follows from Theorem 3.7 along with Corollary 5.2. $\square$

A simple case in which Proposition 5.1 is applicable can be found using the following preliminary result.

Lemma 5.4. *If A is a symmetric matrix, then the entries of the parameterized matrix $(I - \alpha A)^{-1}$ are analytic functions of α for all $\alpha \in \mathbb{R}$ such that $I - \alpha A$ is non-singular.*

Proof. Since A is symmetric, we can diagonalize it as $A = U\Lambda U^\top$. It follows that $I - \alpha A = U(I - \alpha\Lambda)U^\top$, so the entries of $(I - \alpha A)^{-1} = U(I - \alpha\Lambda)^{-1}U^\top$ are linear combinations (with constant coefficients) of the entries of $(I - \alpha\Lambda)^{-1}$. But the non-zero entries of this last matrix are all of the form $(1 - \alpha\lambda)^{-1}$ for eigenvalues λ of A, hence are analytic for $\alpha\lambda \neq 1$. That is, the entries of $(I - \alpha A)^{-1}$ are analytic for all α such that the eigenvalues of $I - \alpha A$ are non-zero. $\square$

Corollary 5.5. *Any 2-component Gaussian mixture has finitely many critical points.*

Proof. By Corollary 3.4 the critical points of a 2-component Gaussian mixture lie in the image of a curve of the form

$$(\alpha_1, \alpha_2) \mapsto (\alpha_1 \Sigma_1^{-1} + \alpha_2 \Sigma_2^{-1})^{-1}(\alpha_1 \Sigma_1^{-1}\mu_1 + \alpha_2 \Sigma_2^{-1}\mu_2).$$

Here, $\alpha_1, \alpha_2 \in [0, 1]$ and $\alpha_1 + \alpha_2 = 1$, so we can let $\alpha_1 = \alpha$ so that $\alpha_2 = 1 - \alpha$ to see that the critical points lie in the image of the map

$$\alpha \mapsto x(\alpha) = (\alpha \Sigma_1^{-1} + (1 - \alpha)\Sigma_2^{-1})^{-1}(\alpha \Sigma_1^{-1}\mu_1 + (1 - \alpha)\Sigma_2^{-1}\mu_2),$$

which is clearly analytic for all α such that the entries of

$$(\alpha \Sigma_1^{-1} + (1 - \alpha)\Sigma_2^{-1})^{-1}$$

are analytic.

Since

$$\alpha \Sigma_1^{-1} + (1 - \alpha)\Sigma_2^{-1}$$

is positive-definite for $\alpha \in [0, 1]$ and

$$\begin{aligned}
(\alpha \Sigma_1^{-1} + (1 - \alpha)\Sigma_2^{-1})^{-1} &= (\Sigma_2^{-1} - \alpha(\Sigma_2^{-1} - \Sigma_1^{-1}))^{-1} \\
&= ((I - \alpha(\Sigma_2^{-1} - \Sigma_1^{-1})\Sigma_2)\Sigma_2^{-1})^{-1} \\
&= \Sigma_2(I - \alpha(I - \Sigma_1^{-1}\Sigma_2))^{-1},
\end{aligned}$$

the curve x is analytic on $[0, 1]$.

Thus, either f has finitely many critical points or $f \circ x$ is constant i.e., $(f \circ x)'(\alpha) = 0$ for $\alpha \in [0, 1]$. But

$$\nabla f(x) = \pi_1 f_1(x) \Sigma_1^{-1}(\mu_1 - x) + \pi_2 f_2(x) \Sigma_2^{-1}(\mu_2 - x)$$

and $x'(\alpha)$ is given by

$$-(\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1})^{-1}(\Sigma_1^{-1} - \Sigma_2^{-1})(\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1})^{-1}(\alpha\Sigma_1^{-1}\mu_1 + (1-\alpha)\Sigma_2^{-1}\mu_2)$$
$$+ (\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2).$$

Thus,

$$\nabla f(x(0)) = \nabla f(\mu_2) = \pi_1 f_1(\mu_2)\Sigma_1^{-1}(\mu_1 - \mu_2)$$

and

$$x'(0) = \Sigma_2 \Sigma_1^{-1}(\mu_1 - \mu_2),$$

so

$$(f \circ x)'(0) = (\nabla f(x(0)))^\top x'(0) = \pi_1 f_1(\mu_2)(\mu_1 - \mu_2)^\top \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}(\mu_1 - \mu_2) \geq 0$$

with equality if and only if $\pi_1 = 0$ or $\mu_1 = \mu_2$, both of which contradict our assumptions. □

With regards to modes, this is weaker than the main result of [14], which was discussed in Section 3.2.

Another special case under which the critical points of a Gaussian mixture lie on an analytic curve can be found using the fact, noted in [13], that the critical points of a homoscedastic mixture lie in the convex hull of the mixture's component means. In fact, this holds more generally for proportional-covariance mixtures.

Proposition 5.6. *The critical points of a proportional-covariance Gaussian mixture lie in the convex hull of the mixture's component means.*

Proof. If $\Sigma_i = \sigma_i^2 \Sigma$, then (8) becomes

$$\begin{aligned} x &= \left(\sum_{i=1}^{n} \alpha_i \sigma_i^{-2} \Sigma^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i \sigma_i^{-2} \Sigma^{-1} \mu_i \right) \\ &= \left(\sum_{i=1}^{n} \alpha_i \sigma_i^{-2} \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i \sigma_i^{-2} \mu_i \right) \\ &= \sum_{i=1}^{n} \beta_i \mu_i, \end{aligned}$$

where

$$\beta_i = \frac{1}{\sum_{j=1}^{n} \alpha_j \sigma_j^{-2}} \alpha_i \sigma_i^{-2}$$

and $\alpha_i \in [0,1]$, so $0 \le \beta_i \le 1$ and $\displaystyle\sum_{i=1}^{n} \beta_i = 1$. □

Corollary 5.7. *If the component means of a proportional-covariance Gaussian mixture lie on a straight line, then the mixture has finitely many critical points.*

Proof. By Propositions 5.6, the critical points of such a mixture lie on a straight line, so the result follows from Corollary 5.2. □

Note. The above yields a simplified proof of Corollary 5.5 in the proportional-covariance case.

6 Conclusion

We have brought together a variety of results related to the conjecture that Gaussian mixtures have finitely many critical points. As discussed, this problem is motivated in large part by the convergence criteria for the the mean-shift algorithm presented in [12]. However, proving this conjecture can be regarded as a problem of more general mathematical interest due to the significance of the Gaussian density in mathematics and the importance of Gaussian mixture models in applications. Moreover, Gaussian mixtures can exhibit rather interesting behaviours, such as the high (but finite) modality of the mixtures in [6].

As an aside, we are of the opinion that the axes lemma proved there motivates a more general investigation into the critical points of Gaussian mixtures satisfying various symmetry conditions. As discussed earlier, a first step in such an investigation could involve a clarification of the proof of the axes lemma.

We have also constructed a class of Gaussian mixtures that exhibit a degenerate critical point, demonstrating that Conjecture A is not necessarily implied by the inverse function theorem. However, a precise characterization of the situations under which the critical points of a Gaussian mixture degenerate is rather elusive due to the complexity of the Hessian of such mixtures. It is interesting to note the connection between the transition to unimodality and the degeneration of the critical point in our class of examples. It could be of some interest to examine this connection more closely.

Finally, we have found some situations under which the critical points of a Gaussian mixture lie on an analytic curve and are easily seen to be finite in number. Unfortunately, the "dimensionality reduction" approach we took to prove these special cases of Conjecture A is rather hard to apply most of the time; indeed, the cases of this conjecture that we verified in this way are rather special. Moreover, a generalization of this approach that seeks out surfaces or higher-dimensional manifolds containing the critical points would likely be fruitless due to the behaviour of the zero sets of analytic functions in more than one dimension.

Though we have encountered certain difficulties in attempting to prove Conjecture A, let us remind the reader that we have only used very elementary methods thus far. In addition to the potential for future work discussed above, let us not forget the possibility of applying more sophisticated tools to this problem. A highly relevant subject in this regard is that of real analytic geometry, the study of the zero sets of real analytic functions, and we believe that [11] is an excellent resource on this subject.

References

[1] Miguel Á. Carreira-Perpiñan. Gaussian mean-shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, May 2007.

[2] Miguel Á. Carreira-Perpiñan and Christopher K. I. Williams. An isotropic Gaussian mixture can have more modes than components, Dec 2003. EDI-INF-RR-0185.

[3] Miguel Á. Carreira-Perpiñan and Christopher K. I. Williams. On the number of modes of a Gaussian mixture. *Scale Space Methods in Computer Vision*, 2695:625–640, 2003.

[4] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, August 1995.

[5] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[6] Herbert Edelsbrunner, Brittany Terese Fasy, and Günter Rote. Add isotropic Gaussian kernels at own risk. In *Proceedings of the 2012 symposium on Computational Geometry*, pages 91–100, 2012.

[7] Brittany Terese Fasy. *Modes of Gaussian Mixtures and an Inequality for the Distance Between Curves in Space*. PhD thesis, Department of Computer Science, Duke University, 2012.

[8] Keinosuke Fukunaga and Larry D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, January 1975.

[9] Youness Aliyari Ghassabeh, Tamás Linder, and Glen Takahara. On some convergence properties of the subspace constrained mean shift. *Pattern Recognition*, 46(11):3140–3147, November 2013.

[10] Kenneth Hoffman and Ray Kunze. *Linear Algebra*. Prentice-Hall, Inc., second edition, 1971.

[11] Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Boston, second edition, 2002.

[12] Xiangru Li, Zhanyi Hu, and Fuchao Wu. A note on the convergence of the mean shift. *Pattern Recognition*, 40:1756–1762, 2007.

[13] Surajit Ray and Bruce G. Lindsay. The topography of multivariate normal mixtures. *The Annals of Statistics*, 53(5):2042–2065, 2005.

[14] Surajit Ray and Dan Ren. On the upper bound of the number of modes of a multivariate normal mixture. *Journal of Multivariate Analysis*, 108:41–52, 2012.

[15] Halsey L. Royden and Patrick M. Fitzpatrick. *Real Analysis*. Pearson, fourth edition, 2010.

[16] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Company, third edition, 1976.

A Linear Algebra

Definition. A $d \times d$ matrix A is *symmetric* if it equals its own transpose: $A = A^\top$. A $d \times d$ symmetric matrix A is said to be *positive-definite* if $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^d$ with equality if and only if $x = 0$.

Note that since the operations of inverting and of transposing a matrix commute with one another, the inverse of a symmetric matrix is itself symmetric.

Definition. An invertible matrix U is said to be *orthogonal* if its transpose equals its inverse: $U^\top = U^{-1}$.

Theorem A.1 (Spectral theorem for symmetric matrices)**.** *Let A be a symmetric matrix. Then there exists an orthogonal matrix U such that $A = U\Lambda U^\top$, where Λ is a diagonal matrix whose diagonal entries are the eigenvalues of A (counting multiplicities).*

Proof. See the corollary to Theorem 18 on p. 314 of [10]. □

Theorem A.2. *A symmetric $d \times d$ matrix A is positive-definite if and only if all of its eigenvalues are positive.*

Proof. Using the spectral theorem, write $A = U^\top \Lambda U$. Then

$$x^\top A x = x^\top U^\top \Lambda U x = (Ux)^\top \Lambda (Ux) = \sum_{i=1}^{d} \lambda_i (Ux)_i^2,$$

where the λ_i are the entries of Λ (i.e., the eigenvalues of A). Since U is invertible, $x \neq 0$ if and only if $Ux \neq 0$. Thus, it is clear for such x that $x^\top A x > 0$ whenever the eigenvalues of A are positive.

Conversely, suppose $x^\top A x > 0$ for $x \neq 0$. Then letting $x = U^\top e_j$, we see that

$$0 < x^\top A x = x^\top U^\top \Lambda U x = e_j^\top \Lambda e_j = \lambda_j.$$

□

Corollary A.3. *Every positive-definite matrix A has a positive-definite square root, i.e. a positive-definite matrix B such that $B^2 = A$.*

Proof. Applying the spectral theorem to a symmetric positive-definite matrix A to get $A = U\Lambda U^\top$, it is easy to see that $B = U\Lambda^{1/2}U^\top$ is a positive-definite square root of A, where $\Lambda^{1/2}$ is the diagonal matrix whose diagonal entries are the positive square roots of the corresponding diagonal entries of Λ. □

In fact, the positive-definite square root of a positive-definite matrix A is unique; we shall content ourselves with denoting the square root obtained in the proof above by $A^{1/2}$.

Corollary A.4. *If A and B are symmetric and positive-definite, then so are A^{-1} and $aA+bB$ for any $a, b \geq 0$.*

Proof. We have already noted above that A^{-1} is symmetric. That it is positive-definite can be seen to be true by writing

$$x^\top A^{-1} x = x^\top A^{-1} A A^{-1} x = (A^{-1}x)^\top A (A^{-1}x) > 0$$

or by noting that the eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A (by the spectral theorem).

Symmetry of $aA + bB$ is obvious; positive-definiteness can be seen by writing

$$x^\top (aA + bB)x = ax^\top Ax + bx^\top Bx > 0.$$

□

Theorem A.5 (Equivalence of norms)**.** *Let $\|\cdot\|_1 : \mathbb{R}^d \to \mathbb{R}$ and $\|\cdot\|_2 : \mathbb{R}^d \to \mathbb{R}$ be two norms on $\mathbb{R}^d$. Then there exist constants $C \geq c > 0$ such that*

$$c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1$$

for every $x \in \mathbb{R}^d$. Hence, if $x_n \in \mathbb{R}^d$ is a sequence, then $\|x_n\|_1 \to 0$ if and only if $\|x_n\|_2 \to 0$.

Proof. See Theorem 4 on p. 260 of [15]. □

Lemma A.6. *If A and B are $d \times d$ invertible matrices and $u, v \in \mathbb{R}^d$, then*

$$\det(A + uv^\top B) = (1 + v^\top BA^{-1}u) \det(A).$$

Proof. First, note that

$$\left[\begin{array}{c|c} I & 0 \\ \hline v^\top & 1 \end{array}\right] \left[\begin{array}{c|c} I + uv^\top & u \\ \hline 0 & 1 \end{array}\right] \left[\begin{array}{c|c} I & 0 \\ \hline -v^\top & 1 \end{array}\right] = \left[\begin{array}{c|c} I + uv^\top & u \\ \hline v^\top + v^\top uv^\top & v^\top u + 1 \end{array}\right] \left[\begin{array}{c|c} I & 0 \\ \hline -v^\top & 1 \end{array}\right]$$
$$= \left[\begin{array}{c|c} I & u \\ \hline 0 & v^\top u + 1 \end{array}\right],$$

where the above are all $(d + 1) \times (d + 1)$ matrices (written in block form). Thus, the determinant of the last matrix must equal the product of the determinants of the first three matrices above; that is,

$$v^\top u + 1 = \det(I + uv^\top).$$

Replacing u and v above by $A^{-1}u$ and $B^\top v$ (respectively) thus yields

$$\det(A + uv^\top B) = \det(A)\det(I + (A^{-1}u)(v^\top B)) = \det(A)(1 + v^\top BA^{-1}u),$$

as required. $\square$

A.1 Simplices

For $i = 1, \dots, d$, denote by e_i the vector in $\mathbb{R}^d$ whose j-th component is 1 if $j = i$ and 0 otherwise.

Definition. An *n-simplex* S_n in $\mathbb{R}^d$ (for $d \geq n+1$) is the convex hull of any $n+1$ linearly independent vectors in $\mathbb{R}^d$; concretely, S_n has the form

$$S_n = \left\{ \sum_{i=1}^{n+1} \alpha_i v_i : \alpha_i \in [0,1], \sum_{i=1}^{n+1} \alpha_i = 1 \right\}$$

for some choice of linearly independent $v_1, \dots, v_{n+1}$. In this case, we will say that S_n is *spanned* (as a simplex) by the v_i. The *standard n-simplex* is obtained by setting $d = n+1$ and $v_i = e_i$. A *scaled* standard n-simplex is an n-simplex of the form

$$cS_n = \{cx : x \in S_n\}$$

for some $c > 0$, where S_n is the standard n-simplex.

Definition. Let S_n be the n-simplex spanned by the set of vectors $A = \{v_1, \dots, v_{n+1}\} \subseteq \mathbb{R}^d$.

1. If $x \in S_n$, then the $\alpha_i \in [0,1]$ such that $x = \sum_{i=1}^{n+1} \alpha_i v_i$ are called the *barycentric coordinates* of x (in S_n).

2. The *barycenter* of S_n is the point $b \in S_n$ whose barycentric coordinates α_i in S_n all equal $\dfrac{1}{n+1}$.

3. For $0 \leq k \leq n$, a *k-face* of S_n is a standard k-simplex (so the previous definitions apply to it) spanned by a subset of size $k+1$ of A.

4. Given a k-face F_k of S_n (for $k < n$) spanned by a subset $B \subseteq A$, the l-face *complementary* to F_k is the l-face of S_n spanned by the complementary subset $A \setminus B \subseteq A$; thus, $l = (n+1) - (k+1) - 1 = n - k - 1$.

B Topology

Definition. Let d be a positive integer. The *open ball* (or *open interval* if $d = 1$) in $\mathbb{R}^d$ of radius $r > 0$ centered at $x \in \mathbb{R}^d$ is the subset $B(x, r)$ of the form

$$B(x,r) = \{y \in \mathbb{R}^d : \|x - y\| < r\}.$$

Call a subset $S \subseteq \mathbb{R}^d$ *open* if for every $x \in S$ there exists $r > 0$ such that $B(x, r) \subseteq S$.

Definition. Let $S \subseteq \mathbb{R}^d$.

1. A point $x \in S$ is said to be *isolated* in S if for some $r > 0$ the open ball $B(x, r)$ is disjoint from $S \setminus \{x\}$. The set S is said to be *discrete* if every $x \in S$ is isolated in S.

2. A point $x \in \mathbb{R}^d$ is called a *limit point* of S if for every $r > 0$, the open ball $B(x, r)$ contains an element of S not equal to x. The *closure* of S is the union of S and its set of limit points and will be denoted by $\overline{S}$. The set S is said to be *closed* if it equals its closure (i.e., if it contains all of its limit points).

3. The set S is said to be *bounded* if $S \subseteq B(0, R)$ for some $R > 0$.

4. We shall call S *compact* if it is both closed and bounded.

Definition. We refer to the closure $\overline{B}(x,r)$ of the open ball $B(x,r)$ as the *closed ball* of radius r about x.

Concretely, we have

$$\overline{B}(x,r) = \{y \in \mathbb{R}^d : \|x - y\| \leq r\}.$$

We will make use of the following important theorem.

Theorem B.1 (Bolzano-Weierstrass)**.** *A set $S \subseteq \mathbb{R}^d$ is compact if and only if every infinite subset of S has a limit point in S.*

Proof. See Theorem 2.41 on p. 40 of [16]. □

C Calculus

Let $U \subseteq \mathbb{R}^d$ be open and let $f : U \to \mathbb{R}^m$. We denote the space of linear maps $\mathbb{R}^d \to \mathbb{R}^m$ by $L(\mathbb{R}^d, \mathbb{R}^m)$, the derivative of f by $Df : U \to L(\mathbb{R}^d, \mathbb{R}^m)$, and the k-th partial derivative of f with respect to the variables $x_{i_1}, \ldots, x_{i_k}$ (where $i_1, \ldots, i_k \in \{1, \ldots, d\}$, possibly with repetition) by $\dfrac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}}$ (we assume the reader is familiar with the definitions of these objects). We will identify the derivative of a function with its matrix representative, the *Jacobian matrix.* When $d = 1$, we usually refer to the vector $f'(x) = (Df)(x) \cdot 1 \in \mathbb{R}^m$ as the derivative of f at $x \in U \subseteq \mathbb{R}$.

Note. In the case that $d = 1$ and U is a half-open or closed interval, say $U = [a, b]$, we may still define differentiability of f at the endpoints a and b by replacing the limit in the usual definition by a one-sided limit. For instance, in this case, we would call $f : [a, b] \to \mathbb{R}^m$

differentiable at a with derivative $f'(a)$ if for all $\epsilon > 0$ there exists $h > 0$ such that

$$\|f(a+h) - f(a) - hf'(a)\| < \epsilon.$$

Definition. The function f is said to be C^k if all of its k-th partial derivatives exist and are continuous. We say that f is C^∞ if it is C^k for all positive integers k.

Definition. We call x a *critical* (or *stationary*) *point* of f if $Df(x) = 0$. We shall call a critical point of f *isolated* if it is isolated as an element of the set of critical points of f.

Theorem C.1 (Inverse function theorem)**.** *Let $m = d$, so that $f : U \to \mathbb{R}^d$. Let $x \in U$ and suppose that f is C^1 and $Df(x)$ is invertible. Then there exists an open set $X \subseteq \mathbb{R}^d$ containing x such that $f|_X : X \to \mathbb{R}^d$ is one-to-one.*

Proof. See Theorem 9.24 on p. 221 of [16]. □

In what follows, let $m = 1$ so that $f : U \to \mathbb{R}$ and $Df : U \to L(\mathbb{R}^d, \mathbb{R})$; thus, $Df(x)$ can be represented as a $1 \times n$ matrix for any $x \in U$. We denote the vector represented by the transpose of this matrix by $\nabla f(x)$.

Definition. We define the *gradient* of $f : U \to \mathbb{R}$ to be the map $\nabla f : U \to \mathbb{R}^d$, which assigns to each $x \in U$ the vector $\nabla f(x) = (Df(x))^\top$.

Note that x is a critical point of f if and only if $\nabla f(x) = 0$.

Definition. The *Hessian* of f is the derivative of the gradient of f, i.e. the map

$$Hf = D(\nabla f) : U \to L(\mathbb{R}^d, \mathbb{R}^d).$$

A critical point x of f is said to be *non-degenerate* if $Hf(x)$ is non-degenerate.

Thus, $Hf(x) : \mathbb{R}^d \to \mathbb{R}^d$ is a linear map for each $x \in U$.

We will make use of the following change of coordinates.

Lemma C.2 (Integration of spherically symmetric functions)**.** *Let* $k : [0, \infty) \to \mathbb{R}$ *and define* $K : \mathbb{R}^d \to \mathbb{R}^d$ *by* $K(x) = k(\|x\|^2)$*. Then*

$$\int_{\mathbb{R}^d} K(x)\,\mathrm{d}x = A_d \int_0^\infty k(t^2)t^{d-1}\,\mathrm{d}t$$

for some constant A_d*.*

Proof. The case $d = 1$ is trivial, so suppose $d \geq 2$ and consider the change of coordinates[9]

$$(x_1, \ldots, x_d) = T(r, \theta) = T(r, \theta_1, \ldots, \theta_{d-1})$$

given by

$$x_1(r, \theta) = r\cos\theta_1$$

$$x_i(r, \theta) = r\sin\theta_1 \ldots \sin\theta_{i-1}\cos\theta_i$$

$$x_d(r, \theta) = r\sin\theta_1 \ldots \sin\theta_{d-1},$$

where $2 \leq i \leq d-1, r \in [0, \infty), \theta_j \in [0, \pi]$ for $1 \leq j \leq d-2$ and $\theta_{d-1} \in [0, 2\pi)$.

The Jacobian matrix $DT(r, \theta)$ contains only two nonzero entries in its top row:

$$\frac{\partial x_1}{\partial r} = \cos\theta_1$$
$$\frac{\partial x_1}{\partial \theta_1} = -r\sin\theta_1.$$

Thus, by the Laplace expansion of the determinant,

$$\det DT(r, \theta) = M_{11}\cos\theta_1 - M_{12}r\sin\theta_1,$$

[9]When $d = 2$ and 3 these are simply polar and spherical coordinates, respectively.

where M_{ij} is the determinant of the $(d-1)\times(d-1)$ matrix m_{ij} obtained by eliminating the i-th row and j-th column of $DT(r,\theta)$. Now the entries of m_{11} are the partial derivatives of the x_i with respect to the θ_j, hence are of the form $rc_{ij}(\theta)$, where $c_{ij}(\theta)$ is a product of trigonometric functions (each taking a single angle θ_k as an argument); hence, M_{11} equals r^{d-1} times a sum of products of trigonometric functions. Similarly, one can observe that M_{12} equals r^{d-2} times a sum of products of such functions. It follows that $|\det DT(r,\theta)|$ is of the form $r^{d-1}|c(\theta)|$, where $|c|$ is integrable, say with integral A_d.

Therefore, using this change of coordinates we get

$$\begin{aligned}\int_{\mathbb{R}^d} k(\|x\|^2)\,\mathrm{d}x &= \int_0^{2\pi}\int_0^{\pi}\dots\int_0^{\pi}\int_0^{\infty} k(r^2)r^{d-1}|c(\theta)|\,\mathrm{d}r\mathrm{d}\theta_1\dots\mathrm{d}\theta_{d-1}\\ &= A_d\int_0^{\infty} k(r^2)r^{d-1}\,\mathrm{d}r.\end{aligned}$$

We can determine the constant A_d by considering the case where

$$k(r)=\begin{cases}1, & \text{if } r\le 1\\ 0, & \text{otherwise}\end{cases}.$$

By the above, we get

$$\mathrm{Vol}(\overline{B}(0,1))=\int_{\overline{B}(0,1)}\mathrm{d}x=\int_{\mathbb{R}^d}k(\|x\|^2)\,\mathrm{d}x=A_d\int_0^{\infty}k(r^2)r^{d-1}\,\mathrm{d}r=A_d\int_0^1 r^{d-1}\,\mathrm{d}r=\frac{A_d}{d},$$

where $\mathrm{Vol}(\overline{B}(0,1))$ is the volume of the closed unit ball $\overline{B}(0,1)$. Thus,

$$A_d = d\,\mathrm{Vol}(\overline{B}(0,1)).$$

□

C.1 Convexity

Here we follow Section 6.6 of [15].

Definition. A subset $S \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in S$ and $t \in [0, 1]$ we have $tx + (1-t)y \in S$. If S is convex, then a function $f : S \to \mathbb{R}$ is said to be *convex* (respectively, *strictly convex*) if for all $x, y \in S$ (respectively, all distinct $x, y \in S$) and $t \in [0, 1]$ (respectively, $t \in (0, 1)$) we have $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$ (respectively, $f(tx + (1-t)y) < tf(x) + (1-t)f(y)$).

Theorem C.3. *Let $U \subseteq \mathbb{R}$ be a convex open set and let $f : U \to \mathbb{R}$.*

1. *The function f is convex if and only if for all $x, y, z \in U$ with $x < y < z$,*

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(y)}{z - y};$$

   the same holds for strictly convex functions if we replace the inequality by a strict inequality.

2. *If f is differentiable, then it is convex if and only if*

$$f'(x) \leq \frac{f(x) - f(y)}{x - y}$$

   for all $x, y \in U$; the same holds for strictly convex functions if we replace the inequality by a strict inequality and require that $x \neq y$.

3. *If f is twice differentiable and $f''(x) \geq 0$ (respectively, $f''(x) > 0$) for all $x \in U$, then it is convex (respectively, strictly convex).*

D Real Analytic Functions

This section follows the first two chapters of [11].

Definition. A *power series* in d variables about $x_0 \in \mathbb{R}^d$ is an expression of the form

$$\sum_{\alpha \in \mathbb{N}^d} a_\alpha (x - x_0)^\alpha,$$

where $x \in \mathbb{R}^d$, $a_\alpha \in \mathbb{R}$ and

$$y^\alpha = \prod_{i=1}^{d} y_i^{\alpha_i}$$

for any $y \in \mathbb{R}^d$ and $\alpha \in \mathbb{N}^d$. Such a power series is said to *converge* (at x) if it converges in the ordinary sense under some ordering of $\mathbb{N}^d$.

Note that if the power series above converges at a point, it does not necessarily converge in the ordinary sense under every ordering of $\mathbb{N}^d$. For this reason, we restrict attention to the *domain of convergence* of a power series, the set of all points x for which the power series converges absolutely in some neighbourhood of x, i.e. for which there exists an $r > 0$ such that $|x - y| < r$ implies that

$$\sum_{\alpha \in \mathbb{N}^d} |a_\alpha (y - x_0)^\alpha|$$

converges; here the ordering of $\mathbb{N}^d$ is irrelevant.

Definition. Let $U \subseteq \mathbb{R}^d$ be open. A function $f : U \to \mathbb{R}$ is said to be *real analytic* if for each $x_0 \in U$, there exist $a_\alpha \in \mathbb{R}$ and $r > 0$ such that $|x - x_0| < r$ implies that

$$f(x) = \sum_{\alpha \in \mathbb{N}^d} a_\alpha (x - x_0)^\alpha.$$

Call a function $F : U \to \mathbb{R}^m$ *(real) analytic* if each of its components $U \to \mathbb{R}$ is analytic.

Several familiar facts about analytic functions on open subsets of $\mathbb{R}$ generalize to the situation at hand.

Theorem D.1. *Let f be a real analytic function on an open set $U \subseteq \mathbb{R}^d$. Then f is C^∞ in U.*

Theorem D.2. *If $f : U \to \mathbb{R}$ and $g : V \to \mathbb{R}$ are analytic, where $U, V \subseteq \mathbb{R}^d$ are open sets with non-empty intersection, then $f + g : U \cap V \to \mathbb{R}$, $fg : U \cap V \to \mathbb{R}$, and $f/g : U \cap V \cap \{x \in \mathbb{R}^d : g(x) \neq 0\} \to \mathbb{R}$ are analytic.*

A particular feature of analytic functions of one variable is the following.

Theorem D.3. *Let U be an open interval. Suppose $f, g : U \to \mathbb{R}$ are analytic and let*

$$E = \{x \in U : f(x) = g(x)\}.$$

If E contains a limit point of U, then $f(x) = g(x)$ for all $x \in U$.

Theorem D.4. *If $f : U \to \mathbb{R}$ is analytic, where $U \subseteq \mathbb{R}^d$ is open, then f is C^∞.*

Corollary D.5. *Let $U \subseteq \mathbb{R}$ be an open interval. If $f : U \to \mathbb{R}$ is a non-constant analytic function, then it has a discrete set of critical points.*

Proof. A point $x \in U$ is critical for f if it is a zero of f', which is analytic. But f is non-constant, so f' is not identically zero, hence has a discrete set of zeros. $\square$

It is well-known that the previous theorem need not hold if f or g is only required to be C^∞ in U. Similarly, it may fail for analytic functions in several variables. For instance, take $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = xy$. Then f is clearly analytic and not identically zero, but the zero set of f is the union of the x- and y-axes.