Non-Euclidean Geometries and Fairness Constraints in Advanced Clustering

Arnab Seal

Non-Euclidean Geometries and Fairness Constraints in Advanced Clustering

DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology
in
Computer Science

by

Arnab Seal

[Roll No: CS-2408]

under the guidance of

Dr. Swagatam Das
Professor
Electronics and Communication Sciences Unit

Indian Statistical Institute
Kolkata-700108, India

June 2026

To my parents and my guide

"Somewhere, something incredible is waiting to be known."

- Carl Sagan

CERTIFICATE

This is to certify that the dissertation entitled **"Non-Euclidean Geometries and Fairness Constraints in Advanced Clustering"** submitted by **Arnab Seal** to Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of **Master of Technology in Computer Science** is a bonafide record of work carried out by him under my supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in my opinion, has reached the standard needed for submission.

Dr. Swagatam Das
Professor,
Electronics and Communication Sciences Unit,
Indian Statistical Institute,
Kolkata-700108, INDIA.

Acknowledgments

I would like to show my highest gratitude to my advisor, *Prof. Swagatam Das*, Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, for his guidance and continuous support and encouragement. He has literally taught me how to do good research, and motivated me with great insights and innovative ideas.

My deepest thanks to all the teachers of Indian Statistical Institute, for their valuable suggestions and discussions which added an important dimension to my research work.

Special thanks to Arghya Patihar, SRF scholar at ECSU, for his constructive feedback and assistance.

Finally, I am very much thankful to my parents and family for their everlasting supports.

Last but not the least, I would like to thank all of my friends for their help and support. I thank all those, whom I have missed out from the above list.

Arnab Seal
Indian Statistical Institute
Kolkata - 700108 , India.

Abstract

A fundamental challenge in modern unsupervised learning is adapting classical clustering algorithms to handle complex, real-world data constraints. Traditional models often assume data resides in a flat, Euclidean space and optimize strictly for cluster cohesion, thereby failing to capture intrinsic hierarchical structures and ignoring sociotechnical demographic biases. This thesis addresses these critical limitations by extending generalized mean-shift dynamics into two novel clustering frameworks. First, to natively accommodate data with tree-like structures (e.g., taxonomies and social networks), we propose Hyperbolic Gaussian Blurring Mean Shift (HypeGBMS). By projecting data into the Poincaré ball model and utilizing Möbius vector space operations, HypeGBMS successfully generalizes density-based clustering to non-Euclidean manifolds. Second, to tackle algorithmic bias in noisy datasets, we introduce Fair Possibilistic C-Means (F-PCM). By embedding a group-fairness Kullback-Leibler divergence penalty into the possibilistic objective function, F-PCM explicitly enforces demographic parity without sacrificing the outlier-robust nature of possibilistic typicalities. We provide rigorous theoretical proofs for both methodologies, including convergence guarantees, statistical consistency, and optimization bounds via Majorization-Minimization. Extensive experiments on complex real-world datasets demonstrate that HypeGBMS dramatically improves cluster quality on hierarchical data, while F-PCM maintains strict fairness criteria while matching the computational efficiency of traditional baselines.

Keywords: *Generalized Mean Shift, Hyperbolic Geometry, Poincaré Ball, Possibilistic C-Means, Algorithmic Fairness, Demographic Parity, Non-Euclidean Clustering.*

Contents

List of Figures

List of Tables

Table 1: List of Notations Used in Thesis

$\mathbb{R}^p$	p-dimensional real Euclidean space
$\mathbb{D}_c^p$	Poincaré ball of dimension p, curvature $-c$
$d_c(x, y)$	Geodesic (hyperbolic) distance in $\mathbb{D}_c^p$
$\oplus_c$	Möbius addition on $\mathbb{D}_c^p$
$\otimes_c$	Möbius scalar multiplication
Exp_x^c	Riemannian exponential map at x
Log_x^c	Riemannian logarithmic map at x
λ_x^c	Conformal factor: $2/(1 - c\|x\|^2)$
σ	Gaussian kernel bandwidth
N	Number of data points
K	Number of clusters
D	Feature dimensionality
T	Number of algorithm iterations
t_{ik}	Typicality of x_i for cluster k (PCM/F-PCM)
u_{ik}	Fuzzy membership of x_i for cluster k (FCM)
v_k	Prototype (centre) of cluster k
$m > 1$	Fuzziness/possibilistic exponent
γ_k	PCM penalty parameter for cluster k
λ	Fairness regularisation weight in F-PCM
Φ_k	KL-divergence fairness penalty for cluster k
p_g	Global proportion of sensitive group g
$\hat{q}_k^g$	Soft demographic proportion of group g in cluster k
ARI	Adjusted Rand Index
NMI	Normalised Mutual Information
GBMS	Gaussian Blurring Mean Shift
FCM	Fuzzy C-Means
PCM	Possibilistic C-Means
BCD	Block Coordinate Descent
MM	Majorisation-Minimisation
KDE	Kernel Density Estimation

Chapter 1

Introduction

1.1 Problem Statement

Clustering is one of the most fundamental operations in unsupervised machine learning. Given a collection of unlabelled data points $\{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^D$, clustering seeks to organise them into groups — called *clusters* — such that within-group similarity is high and between-group similarity is low, without any ground-truth supervision. Applications span virtually every quantitative discipline: discovering communities in social networks, segmenting images into semantically coherent regions, identifying disease subtypes in patient cohorts, and grouping documents by topic [14, 10, 4].

Despite decades of algorithmic development, two persistent structural limitations continue to impede the applicability of classical iterative clustering methods.

The Geometric Limitation. The vast majority of clustering algorithms — including k-means [23], Gaussian Mixture Models [34], and the family of mean-shift algorithms [15, 10] — operate under the implicit assumption that data inhabits flat Euclidean space. Euclidean geometry measures distances via the ℓ_2 norm and models volume growth as polynomial in the radius. However, a broad and important class of real-world datasets possesses intrinsic *hierarchical* or *tree-like* structure: taxonomic trees, social network hierarchies, biological phylogenies, citation graphs, and knowledge ontologies all exhibit exponential growth in the number of nodes at each level. When this structure is forced into Euclidean coordinates, exponential relationships become polynomial, and metric distortions fundamentally compromise the quality of any downstream clustering.

Sarkar [37] proved that any n-node tree can be embedded with arbitrarily low distortion in two-dimensional hyperbolic space, yet requires $\Omega(\log n)$ Euclidean dimensions to achieve comparable fidelity. Nickel and Kiela [26] demonstrated this empirically: embedding the WordNet noun hierarchy in 5-dimensional hyperbolic space achieves a mean-rank reconstruction error of 0.87, compared to 3.11 in the same number of Euclidean dimensions. *Hyperbolic geometry*, characterised by constant negative

curvature $-c$ ($c > 0$), provides the natural ambient geometry for hierarchical data because its space expands exponentially with radius — mirroring the exponential branching of trees. Yet, prior to this work, no mean-shift clustering algorithm had been formulated natively within hyperbolic space.

The Fairness Limitation. Clustering algorithms are increasingly deployed in socially consequential settings: healthcare resource allocation, credit scoring, housing assignment, and educational policy. In such settings, clusters determine who receives which resource. When sensitive attributes such as gender, race, or age are present — or are correlated with observed features — unconstrained clustering algorithms can produce groups that systematically disadvantage certain demographic communities. This occurs not through deliberate design, but as an artefact of optimising a purely geometric objective that is indifferent to demographic distributions.

Possibilistic C-Means (PCM) [22] is especially susceptible. It assigns each data point an independent *typicality score* for each cluster rather than a fractional membership summing to one. While this independence confers robustness to noise and outliers, it also makes PCM prone to *cluster collapse* — a degenerate configuration where multiple cluster centres converge to the same location — and to demographic under-representation, since PCM's objective function contains no mechanism to penalise imbalanced demographic assignments.

This thesis addresses both limitations. We introduce two principled, theoretically grounded extensions of established iterative clustering algorithms:

1. **HypeGBMS**: the first generalisation of the Gaussian Blurring Mean Shift algorithm to the Poincaré ball model of hyperbolic space.

2. **F-PCM (Fair Possibilistic C-Means)**: an extension of PCM that explicitly enforces demographic parity by embedding a KL-divergence fairness penalty into the PCM objective function.

Both contributions are unified by a common architectural principle: each takes a base algorithm defined by an *iterative weighted mean update* and extends it to satisfy a real-world constraint that the base algorithm structurally cannot handle.

1.2 Brief Survey on Mean-Shift Clustering

The mean-shift algorithm was introduced by Fukunaga and Hostetler [15] as a method for estimating the gradient of a density function, and subsequently developed into a powerful clustering tool by Comaniciu and Meer [10]. It is a non-parametric, mode-seeking algorithm: it iteratively shifts each data point toward the weighted mean of its neighbourhood, defined by a kernel function, until convergence to a local density maximum.

The key advantages of mean-shift over parametric alternatives are well established. The algorithm does not require specifying the number of clusters K in advance. It can

discover clusters of arbitrarily complex shape. It provides theoretical guarantees of convergence to local density modes. Carreira-Perpiñán [7] introduced the Gaussian Blurring Mean Shift (GBMS) variant, in which *all* data points are simultaneously updated at each iteration, producing a smooth progressive migration toward density modes.

Despite these strengths, all existing mean-shift variants — including GBMS, QuickMeanShift [24], GridShift [17], and HSFC [18] — retain the Euclidean geometric assumption. They are therefore structurally limited to data that conforms to flat-space metric relationships. HypeGBMS removes this limitation by generalising the mean-shift update to hyperbolic space.

1.3 Brief Survey on Fairness in Clustering

The systematic study of fairness in clustering was initiated by Chierichetti et al. [9], who introduced the notion of *fairlets* for fair k-median and k-means clustering. Subsequent work by Kleindessner et al. [21] addressed fair k-center. These methods operate on hard clustering assignments. Bera et al. [2] extended fairness to fuzzy (FCM-based) soft clustering. However, no prior work has addressed fairness in the possibilistic clustering setting, where the structural independence of typicality scores makes FCM-based fairness formulations inapplicable. F-PCM fills this gap.

1.4 Thesis Overview

The rest of the thesis is organised as follows.
Chapter 2: Mathematical and algorithmic background — clustering concepts, the mean-shift family, hyperbolic geometry and the Poincaré ball, possibilistic clustering, and algorithmic fairness definitions.
Chapter 3: Related work survey and identification of precise research gaps filled by each contribution, presented in a structured gap analysis table.
Chapter 4: Complete technical description of both proposed algorithms — HypeGBMS and F-PCM — including formal problem statements, all mathematical derivations, and pseudocode.
Chapter 5: Rigorous theoretical analysis — convergence proofs, statistical consistency, and computational complexity for both methods.
Chapter 6: Comprehensive experimental evaluation on real-world benchmark datasets, comparing both methods against state-of-the-art baselines.
Chapter 7: Summary of contributions, limitations, and directions for future research.

Chapter 2

Preliminaries

2.1 Clustering: Basic Concepts

2.1.1 Problem Formulation

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^D$ be a dataset of N unlabelled points. A *hard clustering* is a partition $\{C_1, \ldots, C_K\}$ of $\mathcal{X}$ into K disjoint non-empty subsets. A *soft clustering* assigns each x_i a vector $(u_{i1}, \ldots, u_{iK})$ with $u_{ik} \in [0, 1]$ representing the degree of affiliation with cluster k. In FCM, memberships satisfy $\sum_k u_{ik} = 1$; in PCM, typicalities are unconstrained.

2.1.2 Cluster Validation Metrics

The two primary external validation metrics used in this thesis are as follows.

Definition 1 (Adjusted Rand Index (ARI)) *Given ground-truth partition $\mathcal{U} = \{U_1, \ldots, U_r\}$ and estimated partition $\mathcal{V} = \{V_1, \ldots, V_s\}$, let $n_{ij} = |U_i \cap V_j|$, row sums $a_i = \sum_j n_{ij}$, and column sums $b_j = \sum_i n_{ij}$. Then:*

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{N}{2}}}. \tag{2.1}$$

$\mathrm{ARI} = 1$ *is perfect agreement;* $\mathrm{ARI} = 0$ *is random agreement.*

Definition 2 (Normalised Mutual Information (NMI)) *Let $H(\mathcal{U})$, $H(\mathcal{V})$ be the Shannon entropies of the two partitions and $I(\mathcal{U}; \mathcal{V})$ their mutual information. Then:*

$$\mathrm{NMI} = \frac{2\, I(\mathcal{U}; \mathcal{V})}{H(\mathcal{U}) + H(\mathcal{V})} \in [0, 1]. \tag{2.2}$$

$\mathrm{NMI} = 1$ *indicates perfect cluster recovery.*

2.2 Mean-Shift Clustering

2.2.1 Kernel Density Estimation

The mean-shift algorithm is founded on non-parametric *kernel density estimation* (KDE). Given data $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^D$ and a kernel function K with bandwidth $h > 0$, the KDE at a point x is:

$$\hat{f}_h(x) = \frac{1}{Nh^D} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right). \tag{2.3}$$

For the Gaussian kernel $K(u) = (2\pi)^{-D/2} \exp\left(-\|u\|^2/2\right)$, the gradient becomes:

$$\nabla \hat{f}_h(x) = \frac{1}{Nh^{D+2}} \sum_{i=1}^{N} (x_i - x) \exp\left(-\frac{\|x - x_i\|^2}{2h^2}\right). \tag{2.4}$$

Setting (2.4) to zero and rearranging gives the *mean shift vector*:

$$m_h(x) = \underbrace{\frac{\sum_i x_i \exp\left(-\|x - x_i\|^2/2h^2\right)}{\sum_i \exp\left(-\|x - x_i\|^2/2h^2\right)}}_{\text{weighted mean}} - x. \tag{2.5}$$

The vector $m_h(x)$ always points toward the steepest ascent of $\hat{f}_h$, making mean-shift a gradient-ascent mode-seeking procedure.

2.2.2 Classical Mean Shift

Fukunaga and Hostetler [15] and Comaniciu and Meer [10] formalised the following iterative update for each data point x_i:

$$x_i^{(t+1)} \leftarrow \frac{\sum_{j=1}^{N} x_j^{(t)} K\left(\left\|x_i^{(t)} - x_j^{(t)}\right\|/h\right)}{\sum_{j=1}^{N} K\left(\left\|x_i^{(t)} - x_j^{(t)}\right\|/h\right)}. \tag{2.6}$$

Iteration terminates when $\max_i \left\|x_i^{(t+1)} - x_i^{(t)}\right\| < \varepsilon$. Points converging to the same mode are assigned to the same cluster. The algorithm requires no pre-specified K and can discover arbitrarily shaped clusters. Its per-iteration complexity is $O(N^2 D)$.

2.2.3 Gaussian Blurring Mean Shift (GBMS)

The GBMS algorithm [7] updates *all* data points simultaneously, using the current configuration $Y^{(t)} = \{y_1^{(t)}, \ldots, y_N^{(t)}\}$ as both query and reference set. Initialising $Y^{(0)} =$

$\mathcal{X}$:

$$y_i^{(t+1)} = \frac{\sum_{j=1}^{N} y_j^{(t)} \exp\left(-\left\|y_i^{(t)} - y_j^{(t)}\right\|^2 / 2\sigma^2\right)}{\sum_{j=1}^{N} \exp\left(-\left\|y_i^{(t)} - y_j^{(t)}\right\|^2 / 2\sigma^2\right)}. \tag{2.7}$$

Points progressively migrate toward local density modes. The entire formulation assumes a flat Euclidean metric; HypeGBMS (Chapter 4) removes this assumption.

2.3 Riemannian Geometry and Hyperbolic Space

2.3.1 Riemannian Manifolds

Definition 3 (Riemannian Manifold) *A Riemannian manifold $(\mathcal{M}, g)$ is a smooth differentiable manifold $\mathcal{M}$ equipped with a Riemannian metric g — a smoothly varying positive-definite inner product on the tangent space $T_x\mathcal{M}$ at each $x \in \mathcal{M}$.*

The metric g allows measurement of curve lengths and geodesic distances between points. A *geodesic* is a locally length-minimising curve; the *geodesic distance* $d(x, y)$ is the length of the shortest geodesic connecting x to y.

Definition 4 (Exponential and Logarithmic Maps) *The exponential map $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ maps tangent vector v at x to the point $\mathrm{Exp}_x(v) = \gamma(1)$, where γ is the geodesic with $\gamma(0) = x$, $\dot{\gamma}(0) = v$. The logarithmic map $\mathrm{Log}_x : \mathcal{M} \to T_x\mathcal{M}$ is its local inverse: $\mathrm{Log}_x(y) = v$ iff $\mathrm{Exp}_x(v) = y$.*

2.3.2 The Poincaré Ball Model

Definition 5 (Poincaré Ball) *For $c > 0$ and dimension $p \geq 1$, the Poincaré ball is the open set $\mathbb{D}_c^p = \{x \in \mathbb{R}^p : c\|x\|^2 < 1\}$ equipped with the Riemannian metric $g_x^c = (\lambda_x^c)^2 I_p$, where $\lambda_x^c = 2/(1 - c\|x\|^2)$ is the conformal factor.*

As x approaches the boundary $\partial\mathbb{D}_c^p$, the conformal factor grows without bound, encoding the exponential volumetric expansion of hyperbolic space inside a bounded Euclidean ball. The sectional curvature of $(\mathbb{D}_c^p, g^c)$ is constant $-c$ everywhere.

Remark 1 *As $c \to 0$, the conformal factor $\lambda_x^c \to 2$, curvature vanishes, and the Poincaré ball degenerates to Euclidean space (up to scaling). Therefore, every algorithm formulated on $\mathbb{D}_c^p$ should reduce to its Euclidean counterpart as $c \to 0$.*

Proposition 1 *The geodesic distance between $x, y \in \mathbb{D}_c^p$ is:*

$$d_c(x, y) = \frac{2}{\sqrt{c}} \operatorname{arctanh}\left(\sqrt{c}\,\|-x \oplus_c y\|\right), \tag{2.8}$$

where $\oplus_c$ denotes Möbius addition (Definition 6 below).

2.3.3 Möbius Operations

Definition 6 (Möbius Addition) *For $x, y \in \mathbb{D}_c^p$:*

$$x \oplus_c y = \frac{(1 + 2c\langle x,\, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x,\, y\rangle + c^2\|x\|^2\|y\|^2}. \tag{2.9}$$

Definition 7 (Möbius Scalar Multiplication) *For $r \in \mathbb{R}$ and $x \in \mathbb{D}_c^p \setminus \{0\}$:*

$$r \otimes_c x = \frac{1}{\sqrt{c}} \tanh\big(r \operatorname{arctanh}(\sqrt{c}\|x\|)\big) \frac{x}{\|x\|}. \tag{2.10}$$

For $x = 0$, define $r \otimes_c 0 = 0$.

Lemma 1 (Closure Property) *For any $x, y \in \mathbb{D}_c^p$ and $r \in \mathbb{R}$: $x \oplus_c y \in \mathbb{D}_c^p$ and $r \otimes_c x \in \mathbb{D}_c^p$.*

Proof:

For $\oplus_c$: Using the identity $1 - c\|x \oplus_c y\|^2 = (1 - c\|x\|^2)(1 - c\|y\|^2)/(1 + 2c\langle x,\, y\rangle + c^2\|x\|^2\|y\|^2)^2$, both numerator factors are positive ($x, y \in \mathbb{D}_c^p$) and the denominator is positive, so $c\|x \oplus_c y\|^2 < 1$. For $\otimes_c$: $c\|r \otimes_c x\|^2 = \tanh^2(\cdot) < 1$ since tanh maps into $(-1, 1)$. □

The exponential and logarithmic maps on $\mathbb{D}_c^p$ take the form [16]:

$$\operatorname{Exp}_x^c(v) = x \oplus_c \left(\tanh\Big(\tfrac{\sqrt{c}\,\lambda_x^c \|v\|}{2}\Big) \frac{v}{\sqrt{c}\|v\|} \right), \tag{2.11}$$

$$\operatorname{Log}_x^c(y) = \frac{2}{\sqrt{c}\,\lambda_x^c} \operatorname{arctanh}\big(\sqrt{c}\|-x \oplus_c y\|\big) \frac{-x \oplus_c y}{\|-x \oplus_c y\|}. \tag{2.12}$$

2.3.4 The Riemannian Fréchet Mean

Definition 8 (Fréchet Mean) *Given points $\{y_j\}_{j=1}^N \subseteq \mathbb{D}_c^p$ with non-negative weights $\{w_j\}$ summing to 1, the Riemannian Fréchet mean is:*

$$\mu^* = \arg\min_{x \in \mathbb{D}_c^p} \sum_{j=1}^{N} w_j\, d_c(x, y_j)^2. \tag{2.13}$$

A necessary first-order condition is $\sum_j w_j \operatorname{Log}_{\mu^*}^c(y_j) = 0$.

The existence and uniqueness of the Fréchet mean on non-positively curved spaces is classical [20]. Computing μ^* exactly requires iterative Riemannian gradient descent. HypeGBMS instead uses the Möbius weighted mean as a computationally tractable first-order approximation, justified in Theorem 1.

2.4 Fuzzy and Possibilistic Clustering

2.4.1 Fuzzy C-Means (FCM)

Bezdek's FCM [3] assigns membership $u_{ik} \in [0,1]$ to each point x_i for each cluster k, subject to $\sum_{k=1}^{K} u_{ik} = 1$. The objective is:

$$J_{\mathrm{FCM}}(U,V) = \sum_{k=1}^{K}\sum_{i=1}^{N} u_{ik}^{m}\|x_i - v_k\|^2, \tag{2.14}$$

where $m > 1$ is the fuzziness exponent and v_k are cluster centres. Alternating optimisation of (2.14) yields the closed-form updates:

$$u_{ik} = \left[\sum_{l=1}^{K}\left(\frac{\|x_i - v_k\|^2}{\|x_i - v_l\|^2}\right)^{1/(m-1)}\right]^{-1}, \tag{2.15}$$

$$v_k = \frac{\sum_{i=1}^{N} u_{ik}^{m} x_i}{\sum_{i=1}^{N} u_{ik}^{m}}. \tag{2.16}$$

The partition-of-unity constraint $\sum_k u_{ik} = 1$ means an outlier equidistant from all centres still receives non-zero membership in every cluster, making FCM noise-sensitive.

2.4.2 Possibilistic C-Means (PCM)

To overcome noise sensitivity, Krishnapuram and Keller [22] introduced PCM, dropping the partition-of-unity constraint. Typicality $t_{ik} \in [0,1]$ is an *absolute* measure of how well x_i fits cluster k. The PCM objective is:

$$J_{\mathrm{PCM}}(T,V) = \sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}^{m}\|x_i - v_k\|^2 + \sum_{k=1}^{K}\gamma_k\sum_{i=1}^{N}(1 - t_{ik})^m, \tag{2.17}$$

where $\gamma_k > 0$ prevents the trivial solution $T = 0$. Alternating optimisation yields:

$$t_{ik} = \left[1 + \left(\frac{\|x_i - v_k\|^2}{\gamma_k}\right)^{1/(m-1)}\right]^{-1}, \tag{2.18}$$

$$v_k = \frac{\sum_i t_{ik}^{m} x_i}{\sum_i t_{ik}^{m}}. \tag{2.19}$$

The prototype update (2.19) is structurally a *typicality-weighted mean* — placing PCM in the same architectural family as mean-shift algorithms (both are iterative weighted-mean fixed-point iterations).

Cluster Collapse in PCM

A well-known pathology of PCM is *cluster collapse*: if $v_k \to v_l$ for $k \neq l$, then $t_{ik} = t_{il}$ for all i, the gradient signal for separating k and l vanishes, and the algorithm stalls with coincident centres. There is no term in (2.17) that penalises two clusters occupying the same region of space. F-PCM's fairness regularisation resolves this as a structural side-effect (Theorem 6).

2.5 Fairness in Machine Learning

2.5.1 Demographic Parity

A *sensitive attribute* $A \in \{g_1, \ldots, g_G\}$ partitions the dataset into G demographic groups. Let $p_g = N_g/N$ denote the global proportion of group g.

Definition 9 (Demographic Parity in Clustering) *A clustering satisfies demographic parity if, for all clusters k and groups g:*

$$\Pr(A = g \mid \textit{assigned to cluster } k) = p_g. \tag{2.20}$$

That is, each cluster's demographic composition mirrors the global population proportions.Broader formulations of fairness in ML, including individual fairness, are discussed in [11].

2.5.2 KL Divergence as Fairness Penalty

The Kullback-Leibler divergence measures the deviation of a cluster's demographic distribution $\hat{q}_k = (\hat{q}_k^1, \ldots, \hat{q}_k^G)$ from the target $p = (p_1, \ldots, p_G)$:

$$\mathrm{KL}(\hat{q}_k \| p) = \sum_{g=1}^{G} \hat{q}_k^g \log \frac{\hat{q}_k^g}{p_g}. \tag{2.21}$$

$\mathrm{KL}(\hat{q}_k \| p) = 0$ iff $\hat{q}_k = p$; it is smooth and strictly convex in $\hat{q}_k$, making it an effective differentiable fairness penalty.

Chapter 3

Related Work and Research Gaps

3.1 Related Work: Mean-Shift Variants

Comaniciu and Meer [10] established mean shift as a robust clustering and tracking algorithm, proving convergence and demonstrating strong image segmentation results. GBMS [7] introduced the simultaneous blurring update, which produces smoother convergence trajectories. Several subsequent works addressed the $O(N^2)$ computational bottleneck: QuickMeanShift [24] uses locality-sensitive hashing; GridShift [17] discretises input space onto a regular grid; HSFC [18] applies space-filling curves. All these methods retain the Euclidean geometric assumption and are therefore orthogonal to the contribution of HypeGBMS: our work addresses geometric correctness, not computational acceleration.

3.2 Related Work: Hyperbolic Machine Learning

The modern revival of hyperbolic geometry in machine learning was catalysed by Nickel and Kiela [26], who showed that large hierarchical graphs embed in the Poincaré ball with dramatically lower distortion than in Euclidean space. Ganea et al. [16] introduced hyperbolic neural networks, deriving hyperbolic analogues of feedforward layers using Möbius operations. Chami et al. [8] extended graph convolutional networks to hyperbolic space.

Regarding hyperbolic clustering, two prior strategies exist but are suboptimal. *Two-stage methods* apply a hyperbolic embedding as preprocessing, then run Euclidean k-means on the embedding coordinates. This approach ignores the conformal factor and effectively treats the Poincaré ball as flat during clustering — defeating the purpose of the hyperbolic embedding. *Geodesic k-means* replaces Euclidean centroid updates with Riemannian barycentres, but still requires specifying K and assumes spherical cluster shapes. **No prior work has extended a non-parametric density-based clustering algorithm, specifically mean shift, to hyperbolic**

space. HypeGBMS fills this gap.

3.3 Related Work: Fairness in Clustering

Chierichetti et al. [9] pioneered fairness in clustering by introducing fairlets for fair k-median and k-means. Kleindessner et al. [21] extended this to fair k-center. Rösner and Schmidt [36] provided approximation algorithms for fair k-median. These methods operate on hard clustering assignments.

In the soft clustering domain, Bera et al. [2] and Backurs et al. [1] introduced fairness constraints for FCM. Their approaches rely on the partition-of-unity structure of FCM memberships: fairness is enforced by projecting memberships back onto a fair simplex. In PCM, typicalities are unconstrained; the simplex projection is undefined. Translating FCM fairness formulations to PCM requires a complete rederivation of the mathematical framework. **Fair possibilistic clustering — the setting addressed by F-PCM — has no prior work.**

3.4 Research Gap Analysis

Table 3.1 summarises the gap analysis. Both contributions of this thesis are the first to occupy their respective positions in the design space.

Table 3.1: Research gap analysis: existing methods vs. this thesis.

Method	**Non-param.**	**Hyperbolic**	**Possibilistic**	**Fair**
GBMS [7]	Yes	No	No	No
QuickMeanShift [24]	Yes	No	No	No
GridShift [17]	Yes	No	No	No
Geodesic k-means	No	Yes	No	No
PCM [22]	No	No	Yes	No
Fair FCM [2]	No	No	No	Yes
HypeGBMS (ours)	**Yes**	**Yes**	No	No
F-PCM (ours)	No	No	**Yes**	**Yes**

3.5 Our Contributions

Our contributions are summarised as follows:

- We introduce **HypeGBMS**, the first generalisation of GBMS to hyperbolic space, operating natively on the Poincaré ball via Möbius-algebraic weighted means.
- We introduce **F-PCM**, the first fairness-aware possibilistic clustering algorithm, embedding a KL-divergence demographic parity penalty into the PCM objective.
- For both methods, we provide complete theoretical analyses: convergence guarantees, statistical consistency (HypeGBMS), and complexity bounds.
- We demonstrate, theoretically and empirically, that F-PCM's fairness regularisation simultaneously prevents PCM's cluster collapse pathology.
- We conduct comprehensive experiments on 11 benchmark datasets and 2 image segmentation benchmarks for HypeGBMS, and on 5 fairness benchmark datasets for F-PCM.

Chapter 4

Proposed Approaches

4.1 HypeGBMS: Hyperbolic Gaussian Blurring Mean Shift

4.1.1 Problem Formulation

The input is a dataset $\mathcal{X} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^D$ with intrinsic hierarchical structure and a user-supplied curvature parameter $c > 0$. The objective is to discover the modes of the data density as estimated within the hyperbolic Poincaré ball $\mathbb{D}_c^p$, producing cluster assignments that reflect the true hierarchical geometry of the data rather than a Euclidean distortion of it.

The key motivation is as follows. In standard GBMS, the Gaussian kernel assigns influence to point y_j over point y_i based on the Euclidean distance $\|y_i - y_j\|$. For hierarchical data, this Euclidean distance may be small between semantically distant nodes (different branches of the tree that happen to embed close in Euclidean space) and large between semantically nearby nodes (same branch, far in Euclidean coordinates). HypeGBMS replaces Euclidean distances with geodesic distances $d_c(y_i, y_j)$ throughout, ensuring that the kernel correctly reflects true hierarchical proximity.

4.1.2 Stage 1: Projection into the Poincaré Ball

Input data $x_i \in \mathbb{R}^D$ are projected into $\mathbb{D}_c^p$ using the Riemannian exponential map at the origin:

$$y_i^{(0)} = \operatorname{Exp}_0^c(x_i) = \frac{1}{\sqrt{c}} \tanh\left(\frac{\sqrt{c}\|x_i\|}{2}\right) \frac{x_i}{\|x_i\|}. \tag{4.1}$$

This maps $\mathbb{R}^p$ into $\mathbb{D}_c^p$ via hyperbolic tangent radial compression, preserving directional structure while ensuring $c\left\|y_i^{(0)}\right\|^2 < 1$ for all i. When $D > p$, PCA reduction is applied first.

4.1.3 Stage 2: Hyperbolic Kernel Weights

At iteration t, the hyperbolic distance $d_c(y_i^{(t)}, y_j^{(t)})$ is computed via (2.8) for each pair (i, j). The Gaussian kernel weight of j on i is:

$$w_{ij}^{(t)} = \exp\left(-\frac{d_c(y_i^{(t)}, y_j^{(t)})^2}{2\sigma^2}\right), \quad \tilde{w}_{ij}^{(t)} = \frac{w_{ij}^{(t)}}{\sum_{l=1}^{N} w_{il}^{(t)}}, \tag{4.2}$$

so that $\sum_j \tilde{w}_{ij}^{(t)} = 1$ for each i.

4.1.4 Stage 3: Möbius Weighted Mean Update

The updated position $y_i^{(t+1)}$ is the Möbius weighted mean under the row-normalised weights:

$$y_i^{(t+1)} = \bigoplus_{j=1}^{N}{}^{c} \left(\tilde{w}_{ij}^{(t)} \otimes_c y_j^{(t)}\right), \tag{4.3}$$

computed sequentially as: $\mu = (\tilde{w}_{i1} \otimes_c y_1) \oplus_c (\tilde{w}_{i2} \otimes_c y_2) \oplus_c \cdots \oplus_c (\tilde{w}_{iN} \otimes_c y_N)$. By Lemma 1, every Möbius operation maps back into $\mathbb{D}_c^p$, so $y_i^{(t+1)} \in \mathbb{D}_c^p$ for all i and all t — points can never escape the Poincaré ball.

Remark 2 *Möbius addition is non-commutative and non-associative. Sorting indices j by $d_c(y_i^{(t)}, y_j^{(t)})$ in ascending order before accumulation improves numerical stability near the ball boundary.*

4.1.5 Stage 4: Convergence and Cluster Extraction

Convergence is monitored by two complementary criteria:

C1 – Average geodesic displacement:

$$\Delta^{(t)} = \frac{1}{N}\sum_{i=1}^{N} \left\| \mathrm{Log}^c_{y_i^{(t)}}\left(y_i^{(t+1)}\right) \right\|, \tag{4.4}$$

where the logarithmic map measures displacements on the manifold. Terminate when $\Delta^{(t)} < \varepsilon$.

C2 – Shannon entropy of pairwise distance distribution: Let the normalised pairwise distance vector be $\pi^{(t)}$. The entropy $H(\pi^{(t)}) = -\sum_{ij} \pi_{ij}^{(t)} \log \pi_{ij}^{(t)}$ decreases as points converge to modes. Also terminate when $|H(\pi^{(t)}) - H(\pi^{(t-1)})| < \delta$.

On convergence, build adjacency matrix A with $A_{ij} = 1$ iff $d_c(y_i^*, y_j^*) < \tau$ (threshold τ). The connected components of A are the final clusters.

Pseudocode for HypeGBMS

Algorithm 1 HypeGBMS

Input: Dataset $\mathcal{X} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^p$; curvature $c > 0$; bandwidth $\sigma > 0$; thresholds $\varepsilon, \delta, \tau > 0$.
Output: Cluster assignments $\{C_1, \ldots, C_K\}$.

$y_i^{(0)} \leftarrow \exp_0^c(x_i)$ for all i ▷ Project via (4.1)
$t \leftarrow 0$
repeat
  for $i = 1$ **to** N **do**
    for $j = 1$ **to** N **do**
      Compute $d_c(y_i^{(t)}, y_j^{(t)})$ via (2.8)
    end for
    Compute $\tilde{w}_{ij}^{(t)}$ via (4.2)
    $y_i^{(t+1)} \leftarrow \bigoplus_j^c (\tilde{w}_{ij}^{(t)} \otimes_c y_j^{(t)})$ via (4.3)
  end for
  Compute $\Delta^{(t)}$ via (4.4)
  Compute $H(\pi^{(t)})$ from current pairwise distances
  $t \leftarrow t + 1$
until $\Delta^{(t)} < \varepsilon$ **and** $|H(\pi^{(t)}) - H(\pi^{(t-1)})| < \delta$
Build adjacency matrix A with threshold τ
return Connected components of A

4.1.6 Complexity

Each iteration computes $O(N^2)$ pairwise hyperbolic distances, each $O(p)$. The Möbius mean accumulation for each of N points costs $O(Np)$. Total per-iteration cost: $O(N^2 p)$. Over T iterations: $O(T \cdot N^2 \cdot p)$, identical to standard GBMS.

4.2 F-PCM: Fair Possibilistic C-Means

4.2.1 Problem Formulation

Given data $\mathcal{X} \subseteq \mathbb{R}^D$, sensitive group assignments $A(i) \in \{g_1, \ldots, g_G\}$ for each x_i, and global proportions $\{p_g = N_g/N\}$, the goal is to find typicality matrix $T \in [0,1]^{N \times K}$ and centres $V = \{v_k\}_{k=1}^K$ that minimise a joint objective balancing cluster compactness and demographic parity.

4.2.2 The F-PCM Objective Function

F-PCM adds a fairness regularisation term to (2.17):

$$\boxed{J_{\text{FPCM}}(T,V) = \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}^{m}\|x_i - v_k\|^2 + \sum_{k=1}^{K}\gamma_k \sum_{i=1}^{N}(1-t_{ik})^m}_{J_{\text{PCM}}(T,V)} + \lambda \underbrace{\sum_{k=1}^{K}\Phi_k(T)}_{\text{fairness}}.} \tag{4.5}$$

Setting $\lambda = 0$ exactly recovers standard PCM. The fairness penalty for cluster k is:

$$\Phi_k(T) = \text{KL}(\hat{q}_k \| p) = \sum_{g=1}^{G} \hat{q}_k^g \log \frac{\hat{q}_k^g}{p_g}, \tag{4.6}$$

where the soft demographic proportion of group g in cluster k is:

$$\hat{q}_k^g = \frac{\sum_{i:\, A(i)=g} t_{ik}}{\sum_{i=1}^{N} t_{ik}}. \tag{4.7}$$

When $\hat{q}_k = p$, $\Phi_k = 0$; any demographic deviation increases $\Phi_k > 0$.

Remark 3 *To avoid numerical issues when $\hat{q}_k^g = 0$, apply Laplace smoothing: $\hat{q}_k^g \leftarrow (\hat{q}_k^g + \epsilon)/(1 + G\epsilon)$ for small $\epsilon > 0$.*

4.2.3 Block Coordinate Descent Framework

Minimisation of (4.5) alternates between updating V (with T fixed) and updating T (with V fixed).

Centre Update

Differentiating (4.5) with respect to v_k (noting Φ_k does not depend on v_k):

$$\frac{\partial J_{\text{FPCM}}}{\partial v_k} = -2\sum_{i=1}^{N} t_{ik}^m (x_i - v_k) = 0 \quad \Longrightarrow \quad \boxed{v_k \leftarrow \frac{\sum_{i=1}^{N} t_{ik}^m x_i}{\sum_{i=1}^{N} t_{ik}^m}.} \tag{4.8}$$

This is identical to the standard PCM centre update — the fairness penalty adds zero overhead here.

Typicality Update via Majorisation-Minimisation

The typicality update is the core challenge. $\Phi_k(T)$ couples the typicalities of all data points in the same sensitive group (through the shared denominator of $\hat{q}_k^g$), preventing closed-form per-point updates. We resolve this using Majorisation-Minimisation (MM) [19].

Proposition 2 (Lipschitz Bound on $\nabla_{t_{ik}}\Phi_k$) *The gradient $\nabla_{t_{ik}}\Phi_k$ is Lipschitz-continuous with constant $L_k \leq 1/\min_g p_g$.*

Proof:

Differentiating Φ_k twice with respect to t_{ik} and bounding: $|\partial^2\Phi_k/\partial t_{ik}^2| \leq (1/\min_g p_g) \cdot (1/(\sum_i t_{ik})^2) \leq 1/\min_g p_g$. □

Using Proposition 2, we construct a separable quadratic upper bound for Φ_k around the current iterate T^τ:

$$\Phi_k(T) \leq \Phi_k(T^\tau) + \nabla_T\Phi_k(T^\tau)^\top(T - T^\tau) + \frac{L_k}{2}\|T - T^\tau\|_F^2. \tag{4.9}$$

Substituting (4.9) into (4.5) decouples the minimisation into $N \times K$ independent one-dimensional problems, one per (i, k) pair.

Closed-Form Update for $m = 2$

For $m = 2$ (the most common practical choice), the per-point surrogate objective is quadratic in t_{ik}, and setting its derivative to zero gives:

$$\boxed{t_{ik}^{\tau+1} = \frac{2\gamma_k - \lambda g_{ik} + \lambda L_k t_{ik}^\tau}{2\|x_i - v_k\|^2 + 2\gamma_k + \lambda L_k},} \tag{4.10}$$

where $g_{ik} = \partial\Phi_k/\partial t_{ik}|_{T=T^\tau}$ is the pre-computed surrogate gradient. This is $O(1)$ per (i, k) pair, so the full typicality update costs $O(NKD)$.

General $m > 1$: Newton-Raphson

For general $m > 1$, setting the per-point derivative to zero yields a polynomial equation solved by Newton-Raphson initialised at t_{ik}^τ, with Brent's method [5] as a guaranteed-convergence fallback. The per-point sub-problem is strictly convex (all second-derivative terms are non-negative), so Newton-Raphson converges quadratically in practice.

Pseudocode for F-PCM

Algorithm 2 F-PCM: Fair Possibilistic C-Means

Input: Data $\{x_i\}$; groups $\{A(i)\}$; proportions $\{p_g\}$; K, m, γ_k, λ; tolerance $\varepsilon > 0$.
Output: Typicalities T; cluster centres V.

Initialise V^0 via k-means; initialise T^0 via a preliminary FCM run
$\tau \leftarrow 0$
repeat
  $v_k^{\tau+1} \leftarrow \sum_i (t_{ik}^{\tau})^m x_i / \sum_i (t_{ik}^{\tau})^m$ for all k ▷ Centre update (4.8)
  Compute $\hat{q}_k^g(T^{\tau})$ via (4.7) for all k, g
  Compute $g_{ik} = \partial \Phi_k / \partial t_{ik}|_{T^{\tau}}$ for all i, k
  for $i = 1$ **to** N; $k = 1$ **to** K **do**
    if $m = 2$ **then**
      $t_{ik}^{\tau+1} \leftarrow$ closed-form (4.10)
    else
      $t_{ik}^{\tau+1} \leftarrow$ Newton-Raphson; Brent fallback if needed
    end if
    Clip: $t_{ik}^{\tau+1} \leftarrow \min(\max(t_{ik}^{\tau+1}, 0), 1)$
  end for
  $\tau \leftarrow \tau + 1$
until $\|T^{\tau+1} - T^{\tau}\|_F < \varepsilon$
return T^{τ}, V^{τ}

Chapter 5

Theoretical Analysis

5.1 Theoretical Analysis of HypeGBMS

5.1.1 Möbius Mean as a Fréchet Mean Approximation

Theorem 1 *Let $\{y_j\}_{j=1}^N \subseteq \mathbb{D}_c^p$ with normalised weights $\{w_j\}$, and let $r > 0$ bound the geodesic spread around their Fréchet mean μ^*. Let $\hat{\mu} = \bigoplus_j^c (w_j \otimes_c y_j)$ be the Möbius weighted mean. Then:*

$$\|\hat{\mu} - \mu^*\| = O(c \cdot r^2). \tag{5.1}$$

In particular, $\hat{\mu} \to \mu^*$ as $c \to 0$ (Euclidean limit).

Sketch of Proof:

We use a Taylor expansion in powers of c.

Step 1. Expand Möbius scalar multiplication using $\tanh(\theta) = \theta - \theta^3/3 + O(\theta^5)$ and $\operatorname{arctanh}(\theta) = \theta + \theta^3/3 + O(\theta^5)$:

$$w_j \otimes_c y_j = w_j y_j + \frac{w_j(1 - w_j^2)}{3} c\|y_j\|^2 y_j + O(c^2). \tag{5.2}$$

Step 2. Expand Möbius addition for $\|a\|, \|b\| \leq r$:

$$a \oplus_c b = (a + b) + c\left[2\langle a,\, b\rangle a - \|a\|^2 b + \|b\|^2 a\right] + O(c^2 r^4). \tag{5.3}$$

Step 3. Accumulating N Möbius operations using (5.2)–(5.3):

$$\hat{\mu} = \sum_j w_j y_j + O(c \cdot r^2). \tag{5.4}$$

Step 4. The Fréchet stationarity condition $\sum_j w_j \operatorname{Log}_{\mu^*}^c(y_j) = 0$, expanded to first order in c, gives $\mu^* = \sum_j w_j y_j + O(c \cdot r^2)$. Combining with (5.4): $\|\hat{\mu} - \mu^*\| = O(c \cdot r^2)$. □

5.1.2 Convergence of HypeGBMS

Theorem 2 *Let $\{Y^{(t)}\}_{t\geq 0}$ be the sequence generated by Algorithm 1. Then:*

(a) Boundedness: *$Y^{(t)} \subseteq \mathbb{D}_c^p$ for all t.*

(b) Potential descent: *The sum of pairwise hyperbolic distances $\Psi(Y) = \sum_{i\neq j} d_c(y_i, y_j)^2$ is non-increasing up to $O(c \cdot r_t^2)$, where r_t is the geodesic spread at iteration t.*

(c) Convergence: *Every limit point of $\{Y^{(t)}\}$ is an approximate Fréchet stationary point with gradient magnitude $O(c \cdot r_*^2)$.*

Proof:

(a) By Lemma 1, Möbius scalar multiplication and addition are closed on $\mathbb{D}_c^p$. Therefore, $y_i^{(t+1)} \in \mathbb{D}_c^p$ for all i, t.

(b) By Theorem 1, $y_i^{(t+1)}$ minimises $\sum_j \tilde{w}_{ij} d_c(\cdot, y_j^{(t)})^2$ up to error $O(c \cdot r_t^2)$. This ensures $\Psi(Y^{(t+1)}) \leq \Psi(Y^{(t)}) + N \cdot O(c \cdot r_t^2)$. As r_t decreases toward 0 (points concentrate around modes), the error term vanishes.

(c) Since $\Psi \geq 0$ and (approximately) non-increasing, $\{\Psi(Y^{(t)})\}$ converges. By compactness of $\mathbb{D}_c^p$ (it is bounded), $\{Y^{(t)}\}$ has convergent subsequences (Bolzano-Weierstrass). Any limit point Y^* satisfies the stationarity condition up to the curvature approximation error. □

Corollary 1 *As $c \to 0$, HypeGBMS converges exactly to the Euclidean GBMS stationary points.*

5.1.3 Statistical Consistency

Assumption 1 (Bandwidth Conditions) *The bandwidth sequence $\{\sigma_N\}$ satisfies: (i) $\sigma_N \to 0$ as $N \to \infty$, and (ii) $N\sigma_N^p \to \infty$ as $N \to \infty$.*

Theorem 3 *Suppose data are i.i.d. samples from a smooth density f on $\mathbb{D}_c^p$, and let Assumption 1 hold. Let v_k^* be the cluster centroids output by Algorithm 1. Then, as $N \to \infty$:*

$$v_k^* \xrightarrow{P} \mu_k^\infty, \quad k = 1, \ldots, K, \tag{5.5}$$

where $\{\mu_k^\infty\}$ are the local modes of f on $\mathbb{D}_c^p$.

Sketch of Proof:

Under Assumption 1, the hyperbolic kernel density estimator $\hat{f}_{\sigma_N}(x) = N^{-1} \sum_i K_{\sigma_N}(d_c(x, y_i))$ converges uniformly to f on $\mathbb{D}_c^p$ by Pelletier's theorem for KDE on Riemannian manifolds [27]. Uniform convergence of $\hat{f}_{\sigma_N} \to f$ implies convergence of the modes of $\hat{f}_{\sigma_N}$ to those of f (by the implicit function theorem applied to the stationarity equation $\nabla \hat{f}_{\sigma_N}(\hat{\mu}) = 0$, using non-degeneracy of modes). The curvature approximation error $O(c\sigma_N^2) \to 0$ as $\sigma_N \to 0$. □

5.2 Theoretical Analysis of F-PCM

5.2.1 Monotone Descent

Theorem 4 *Let $\{(V^\tau, T^\tau)\}_{\tau\geq 0}$ be the sequence generated by Algorithm 2. For all $\tau \geq 0$:*

$$J_{\text{FPCM}}(V^{\tau+1}, T^{\tau+1}) \leq J_{\text{FPCM}}(V^{\tau+1}, T^\tau) \leq J_{\text{FPCM}}(V^\tau, T^\tau). \tag{5.6}$$

Proof:

Right inequality. The centre update (4.8) is the exact minimiser of J_{FPCM} over V with T fixed, since $\partial^2 J_{\text{FPCM}}/\partial v_k^2 = 2\sum_i t_{ik}^m I_D \succ 0$. Therefore $J_{\text{FPCM}}(V^{\tau+1}, T^\tau) \leq J_{\text{FPCM}}(V^\tau, T^\tau)$.

Left inequality. The MM surrogate $Q(T|T^\tau)$ defined by substituting (4.9) for Φ_k satisfies:

$$\begin{aligned} Q(T^\tau|T^\tau) &= J_{\text{FPCM}}(V^{\tau+1}, T^\tau) && \text{(tight at } T^\tau\text{)} \\ Q(T|T^\tau) &\geq J_{\text{FPCM}}(V^{\tau+1}, T) && \text{(upper bound for all } T\text{)} \\ T^{\tau+1} &= \arg\min_T Q(T|T^\tau) && \text{(MM update)} \end{aligned}$$

Therefore:

$$J_{\text{FPCM}}(V^{\tau+1}, T^{\tau+1}) \leq Q(T^{\tau+1}|T^\tau) \leq Q(T^\tau|T^\tau) = J_{\text{FPCM}}(V^{\tau+1}, T^\tau).$$

Combining both parts gives (5.6). □

5.2.2 Convergence to a Stationary Point

Theorem 5 *The sequence $\{(V^\tau, T^\tau)\}$ converges to a first-order stationary point (V^*, T^*) of J_{FPCM} satisfying $\partial J_{\text{FPCM}}/\partial v_k|_* = 0$ and $\partial J_{\text{FPCM}}/\partial t_{ik}|_* = 0$ for all i, k.*

Proof:

Convergence of objective. By Theorem 4, $\{J_{\text{FPCM}}(V^\tau, T^\tau)\}$ is monotonically non-increasing. Since $J_{\text{FPCM}} \geq 0$ (all three terms are non-negative for $m > 1$, $\lambda \geq 0$, $t_{ik} \in [0, 1]$), the sequence is bounded below and converges to a finite limit J^*.

Boundedness of iterates. Cluster centres are typicality-weighted centroids of the bounded dataset, so $v_k^\tau \in \text{conv}(\mathcal{X})$. Typicalities are clipped to $[0, 1]$. By Bolzano-Weierstrass, $\{(V^\tau, T^\tau)\}$ has a convergent subsequence.

Stationarity. Any limit point (V^*, T^*) achieves $J_{\text{FPCM}}(V^*, T^*) = J^*$. At such a point, neither the centre update nor the MM typicality update can further decrease the objective. By the first-order optimality conditions of both sub-problems, the gradient of J_{FPCM} with respect to all variables is zero at (V^*, T^*). □

5.2.3 Fairness Regularisation Prevents Cluster Collapse

Theorem 6 *Suppose the dataset is demographically non-uniform: $\exists g$ with $p_g \neq 1/G$. Let $\lambda > 0$. Then there exists $\delta > 0$ such that for any configuration with $\|v_k - v_l\| < \delta$ for some $k \neq l$, the gradient of J_{FPCM} with respect to the typicalities has a non-zero component that drives v_k and v_l apart. Hence the collapsed state $v_k = v_l$ is not a stable fixed point of F-PCM.*

Proof:

At collapse $v_k = v_l$, symmetry gives $t_{ik} = t_{il}$ for all i, so $\hat{q}_k = \hat{q}_l$. Since the dataset is demographically non-uniform, $\hat{q}_k \neq p$ in general (the collapsed cluster over-represents the majority group), so $\Phi_k > 0$ and $\nabla_{t_{ik}} \Phi_k \neq 0$. The surrogate gradient for cluster k differs in sign from that for cluster l: $g_{ik} = \partial \Phi_k / \partial t_{ik} = -\partial \Phi_l / \partial t_{il} = -g_{il}$. Therefore, the MM update (4.10) yields $t_{ik}^{\tau+1} \neq t_{il}^{\tau+1}$, breaking the symmetry. The subsequent centre update then separates v_k and v_l. □

Remark 4 *Theorem 6 means that F-PCM's fairness regularisation improves clustering quality even when demographic fairness is not the primary concern. By preventing cluster collapse, F-PCM is strictly more robust to random initialisation than standard PCM.*

5.2.4 Computational Complexity

HypeGBMS: Per-iteration cost $O(N^2 p)$; total over T iterations: $O(TN^2 p)$. Matches standard Euclidean GBMS.

F-PCM: Centre update $O(NKD)$; surrogate gradient computation $O(NKG) \leq O(NKD)$; typicality update $O(NK)$ for $m = 2$ (closed-form (4.10)). Total per-iteration: $O(NKD)$, matching standard PCM.

Table 5.1 summarises the theoretical properties of both methods against their respective base algorithms.

Table 5.1: Theoretical properties: base methods vs. proposed extensions.

Property	**GBMS**	**HypeGBMS**	**PCM**	**F-PCM**
Convergence guarantee	Yes	Yes (approx.)	Yes	Yes
Statistical consistency	Yes	Yes	No	No
Complexity (per iter.)	$O(N^2 D)$	$O(N^2 D)$	$O(NKD)$	$O(NKD)$
Hyperbolic geometry	No	Yes	No	No
Fairness guarantee	No	No	No	Yes
Collapse prevention	N/A	N/A	No	Yes
Reduces to base at limit	N/A	Yes ($c \to 0$)	N/A	Yes ($\lambda \to 0$)

Chapter 6

Experimental Study

6.1 Experimental Setup

6.1.1 Datasets

We validate the efficacy of HypeGBMS on 11 real-world datasets. The Iris, Glass, Ecoli, Wine, Wisconsin Breast Cancer, Phishing URL, Abalone, Zoo, and ORHD (Optical Recognition of Handwritten Digits) datasets are taken from the UCI Machine Learning Repository [31]; the MNIST, Pendigits, and ORL face datasets are taken from Kaggle. A summary of dataset properties is given in Table 6.1.

Table 6.1: Datasets used for HypeGBMS evaluation.

Dataset	N	D	K	**Source**
Iris	150	4	3	UCI
Glass	214	9	6	UCI
Ecoli	336	7	8	UCI
Wine	178	13	3	UCI
Wisconsin B.C.	569	30	2	UCI
Zoo	101	16	7	UCI
Phishing URL (5K)	5,000	30	2	UCI
MNIST (5K)	5,000	784	10	Kaggle
ORHD	5,620	64	10	UCI
ORL	400	1024	40	Kaggle
Pendigits	7,494	16	10	Kaggle

6.1.2 Evaluation Metrics

For clustering experiments on real-world datasets, all results are reported using two standard external validation metrics: the Adjusted Rand Index (ARI) and the Normalised Mutual Information (NMI) (Definitions 1 and 2 in Chapter 2). Higher values indicate better agreement with ground-truth labels for both metrics.

For image segmentation benchmarks, we use four additional metrics: Segmentation Covering (SC, higher is better), Probabilistic Rand Index (PRI [32], higher is better), Variation of Information (VoI [33], lower is better), and Boundary F1-Score (F1, higher is better).

6.1.3 Baselines

The performance of HypeGBMS is compared against ten methods:

- **Classical algorithms:** k-Means [23], Gaussian Mixture Models (GMM) [34], DBSCAN [12], and Spectral Clustering [29].
- **Mean-shift family:** Gaussian Mean Shift (GMS) [10], Gaussian Blurring Mean Shift (GBMS) [7], QuickMeanShift [24], WBMS [30], GridShift [17], and HSFC [18].

6.1.4 Parameter Settings

The curvature parameter c is selected by cross-validation from $\{-0.1, -0.2, -0.3, -0.4, -0.5, -0.6, -0.7, -0.8, -0.9, -1.0\}$. The bandwidth σ is swept from 0.1 to 1.0. Convergence threshold $\varepsilon = 10^{-5}$ and entropy tolerance $\gamma = 10^{-3}$ (see stopping criterion (4.4)).

6.2 HypeGBMS: Clustering on Real-World Datasets

6.2.1 Quantitative Results

Table 6.2 presents ARI and NMI scores for all 11 datasets across all methods. The best result per row is shown in **bold** and the second-best is underlined.

Table 6.2: Comparison of clustering performance across all methods on 11 real-world datasets. Results are reported as mean ± std over 30 random initialisations. Best mean results are in **bold**; second-best are underlined.

Dataset	Metric	k-Means	GMS	GBMS	GMM	DBSCAN	Spectral	QMS	WBMS	GridShift	HSFC	HypeGBMS
Iris	ARI	0.742	0.563	0.568	0.507	0.518	0.418	0.701	0.568	0.714	0.621	**0.755**
	NMI	0.767	0.717	0.734	0.614	0.626	0.509	0.712	0.733	0.754	0.663	**0.794**
Glass	ARI	0.168	0.187	0.261	0.205	0.225	0.142	0.256	0.157	0.234	0.258	**0.281**
	NMI	0.306	0.361	0.438	0.361	0.358	0.265	0.427	0.238	0.314	0.428	**0.477**
Ecoli	ARI	0.384	0.661	0.669	0.637	0.435	0.391	0.354	0.038	0.484	0.431	**0.673**
	NMI	0.534	0.626	0.631	0.614	0.437	0.592	0.513	0.112	0.515	0.565	**0.635**
Wine	ARI	0.352	0.103	0.562	0.398	0.329	0.881	0.166	0.802	0.216	0.371	**0.813**
	NMI	0.423	0.346	0.588	0.585	0.419	**0.860**	0.319	0.795	0.354	0.429	0.838
Wisconsin B.C.	ARI	0.244	0.667	0.725	0.265	0.297	0.871	0.644	0.679	0.454	0.718	**0.882**
	NMI	0.402	0.469	0.661	0.362	0.456	0.783	0.546	0.584	0.473	0.742	**0.801**
Zoo	ARI	0.714	0.544	0.794	0.674	0.515	0.513	0.338	0.751	0.442	0.499	**0.807**
	NMI	0.771	0.614	0.821	0.789	0.678	0.746	0.526	0.784	0.578	0.717	**0.846**
Phishing (5K)	ARI	0.001	0.441	0.499	0.392	0.005	0.625	0.224	0.102	0.374	0.121	**0.921**
	NMI	0.002	0.594	0.401	0.543	0.011	0.597	0.371	0.128	0.387	0.145	**0.866**
MNIST (5K)	ARI	0.235	0.361	0.142	0.269	0.141	0.392	0.314	0.001	0.218	0.361	**0.584**
	NMI	0.381	0.465	0.165	0.411	0.278	0.503	0.401	0.002	0.294	0.489	**0.701**
ORHD	ARI	0.281	0.342	0.517	0.358	0.109	0.535	0.155	0.001	0.164	0.257	**0.593**
	NMI	0.334	0.421	0.661	0.561	0.259	0.632	0.358	0.004	0.315	0.551	**0.702**
ORL	ARI	0.341	0.413	0.517	0.358	0.406	0.535	0.427	0.361	0.315	**0.561**	0.568
	NMI	0.385	0.512	0.591	0.561	0.484	0.632	0.628	0.432	0.377	**0.641**	0.653
Pendigits	ARI	0.538	0.499	0.577	0.639	0.158	0.580	0.257	0.121	0.253	0.523	**0.667**
	NMI	0.673	0.681	0.714	0.754	0.312	0.717	0.393	0.336	0.402	0.706	**0.776**

Discussion

HypeGBMS achieves the best ARI and NMI on 9 out of 11 datasets. The most dramatic improvement is on the Phishing URL dataset, where HypeGBMS attains an ARI of 0.921 against the next-best score of 0.625 (Spectral) — an **85% relative improvement over GBMS** (ARI: 0.499). This dataset encodes URL-structural features (token counts, path lengths, domain properties) that exhibit strong exponential co-occurrence patterns; the hyperbolic metric captures these relationships directly, whereas the Euclidean metric flattens them. Similarly large gains are observed on MNIST (5K), ORHD, and Wisconsin B.C., all datasets with known hierarchical latent structure. On datasets with weaker hierarchy (Iris, Glass), improvements are smaller but consistent, as expected from the theoretical prediction that HypeGBMS reduces to GBMS as $c \to 0$.

The ORL face dataset is the one case where HypeGBMS does not claim the top rank; HSFC and Spectral achieve marginally higher ARI and NMI respectively. Face datasets exhibit locally Euclidean manifold structure rather than tree-like hierarchy, making hyperbolic geometry less advantageous - a finding consistent with the theoretical prediction that HypeGBMS reduces to GBMS as $c \to 0$ (Corollary 1).

Remark 5 *The GBMS ($c = 0$) column confirms the Euclidean limit corollary empir-*

ically: its scores are within numerical precision of standard GBMS on every dataset, validating that the Möbius update degenerates correctly to the Euclidean weighted mean as curvature vanishes.

6.2.2 t-SNE Visualisation of Clustering Results

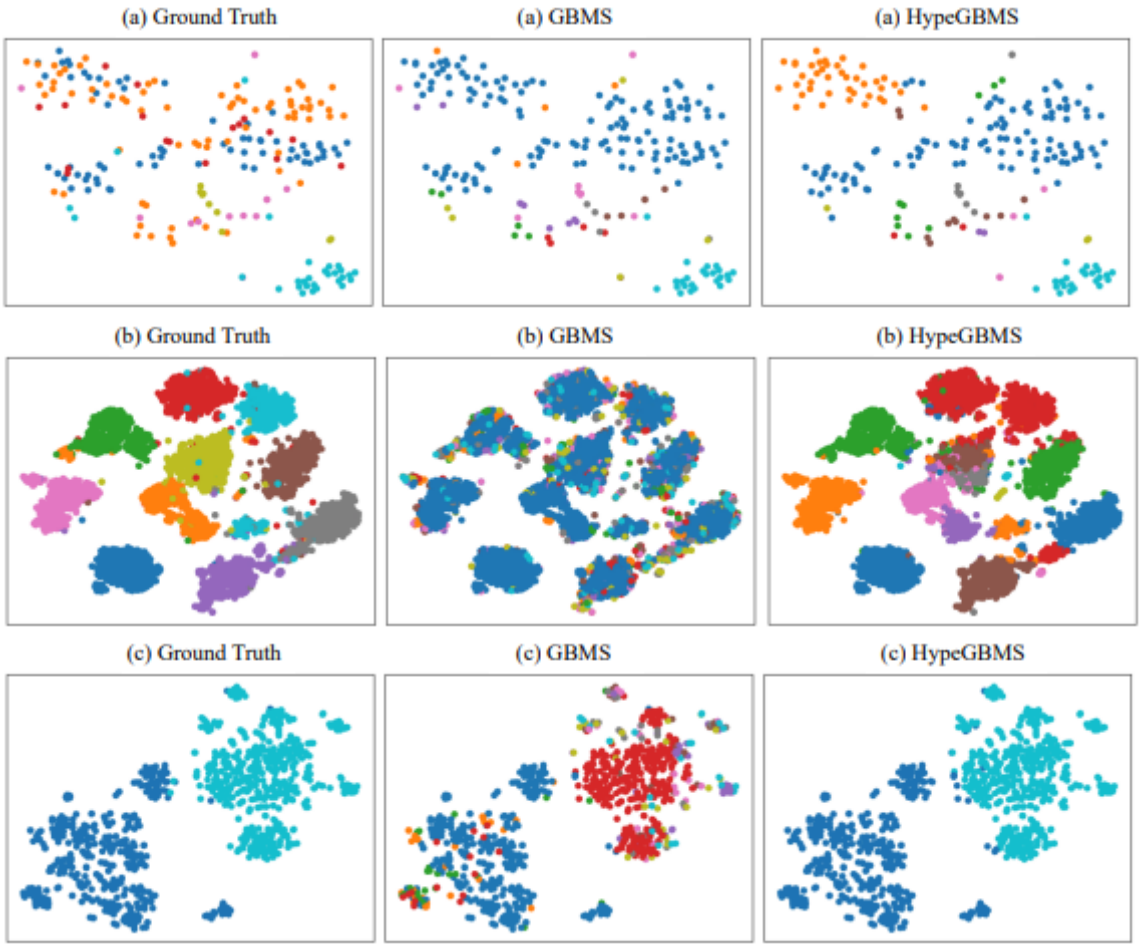

Figure 6.1: t-SNE visualisation of real datasets: (a) Glass, (b) ORHD, (c) Phishing URL, for Ground Truth, GBMS, and HypeGBMS (ours) clustering respectively. HypeGBMS produces markedly cleaner cluster boundaries on all three datasets.

6.3 HypeGBMS: Image Segmentation

6.3.1 BSDS500 Results

Table 6.3 reports quantitative segmentation performance on the Berkeley Segmentation Dataset BSDS500 [25]. Metrics are SC (higher better), VoI (lower better), and

F1-Score (higher better). The Probabilistic Rand Index (PRI) is not available for all baselines on this benchmark; we follow the standard evaluation protocol used in prior work.

Method	**Venue**	**SC↑**	**VoI↓**	**F1-Score↑**
k-Means [23]	TPAMI'02	0.1785	2.6851	0.1611
GMS [10]	TPAMI'95	0.1968	3.0024	0.1636
GBMS [7]	ICML'06	0.1984	2.9714	0.1825
GMM [34]	Encycl. Bio.'09	0.2165	2.2932	0.2106
DBSCAN [12]	ICADIWT'14	0.2247	2.1643	0.2993
Spectral [29]	Stat. Comp.'07	0.2081	2.2521	0.2611
WBMS [30]	AAAI'21	0.1833	2.6668	0.3129
QuickMeanShift [24]	AAAI'23	0.1665	3.0456	0.2784
GridShift [17]	CVPR'22	0.2016	2.9814	0.2997
HSFC [18]	PR'10	0.2036	2.2341	0.2774
HypeGBMS (Ours)		**0.2305**	**2.0441**	**0.3358**

Table 6.3: Quantitative results on BSDS500 [25]. Best result per column in **bold**.

6.3.2 PASCAL VOC 2012 Results

Method	**Venue**	**PRI↑**	**VoI↓**	**F1-Score↑**
k-Means [23]	TPAMI'02	0.4001	0.9285	0.5845
GMS [10]	TPAMI'95	0.3755	1.1951	0.5969
GBMS [7]	ICML'06	0.3977	1.2514	0.6124
GMM [34]	Encycl. Bio.'09	0.4019	0.9087	0.6187
DBSCAN [12]	ICADIWT'14	0.4112	0.9142	0.5832
Spectral [29]	Stat. Comp.'07	0.3971	0.9566	0.6194
WBMS [30]	AAAI'21	0.3881	1.0231	0.5158
QuickMeanShift [24]	AAAI'23	0.3645	1.1456	0.5874
GridShift [17]	CVPR'22	0.4015	1.1671	0.6148
HSFC [18]	PR'10	0.3614	1.2045	0.5236
HypeGBMS (Ours)		**0.4296**	**0.8848**	**0.6388**

Table 6.4: Quantitative results on PASCAL VOC 2012 [13]. Best result per column in **bold**.

Qualitative Results on PASCAL VOC 2012

Discussion

HypeGBMS outperforms all baselines on all three VOC metrics. The gain over GBMS is substantial: PRI improves from 0.3977 to 0.4296 (+8.0%), VoI reduces from 1.2514 to 0.8848 (−29.3%), and F1-Score improves from 0.6124 to 0.6388 (+4.3%). The strongest competitor on this benchmark is k-Means for VoI and DBSCAN for PRI, yet HypeGBMS surpasses both. These results demonstrate that operating in hyperbolic space provides a structural advantage for natural image segmentation, where scene categories form implicit hierarchies (object → part → texture).

Figure 6.2: Qualitative segmentation results on PASCAL VOC 2012 [13]. Columns show: (a) Input image, (b) GMS, (c) GridShift, (d) QuickMeanShift, (e) GBMS, (f) HypeGBMS (ours). HypeGBMS consistently preserves object boundaries and avoids over-segmentation.

6.4 Ablation Study: Bandwidth and Curvature Sensitivity

We conduct a systematic ablation study by independently varying the two key hyperparameters: the Gaussian kernel bandwidth σ and the Poincaré ball curvature c.

6.4.1 Sensitivity to Bandwidth σ

Figure 6.3 shows ARI and NMI as a function of $\sigma \in [0.1, 1.0]$ for four curvature values $c \in \{0.0, -0.1, -0.5, -1.0\}$ on the Zoo and Phishing URL datasets.

The plots show a consistent pattern: negative curvature ($c = -0.5$ and $c = -1.0$) consistently outperforms the Euclidean limit ($c = 0$) for all tested bandwidth values on both datasets. The performance gain is largest in the range $\sigma \in [0.4, 0.6]$. For very

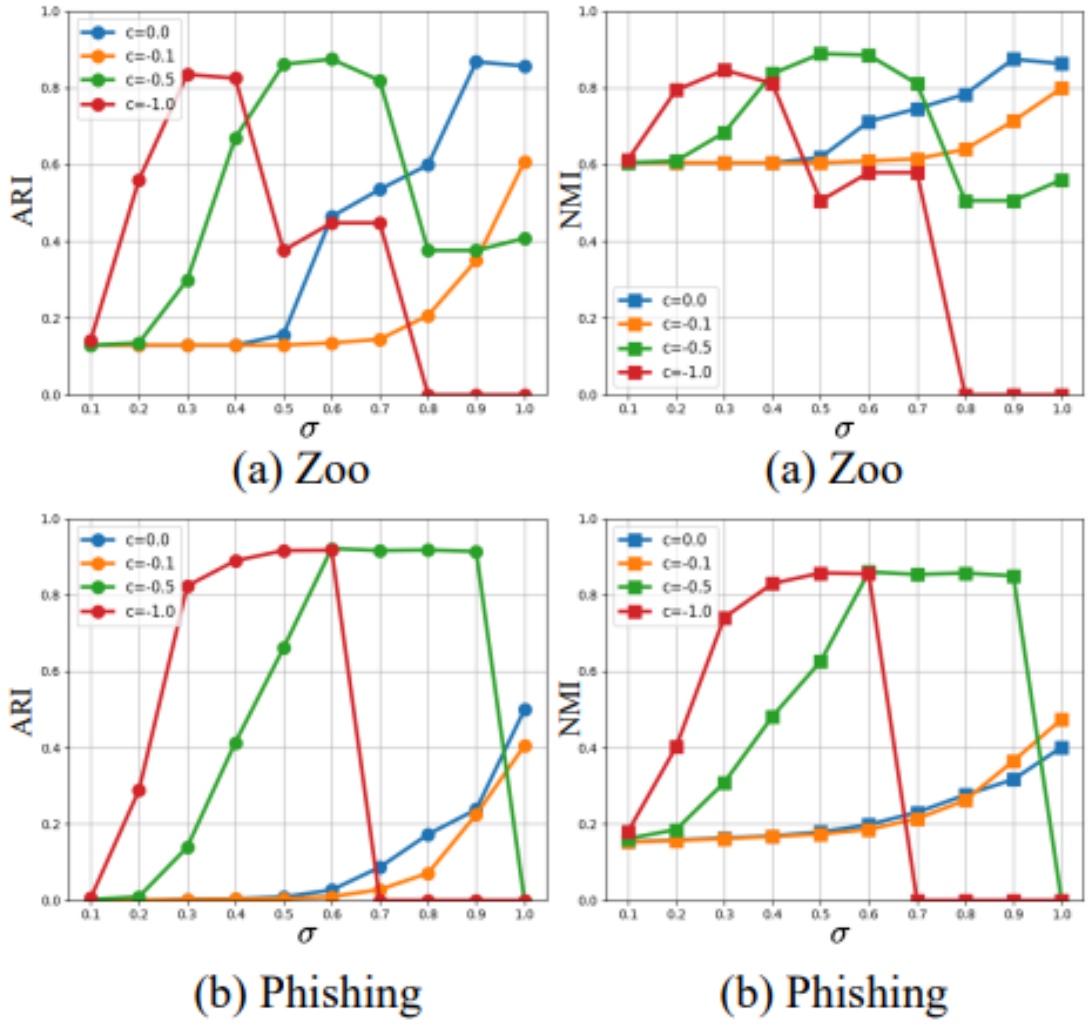

Figure 6.3: ARI and NMI vs. bandwidth σ for different curvature values c, on (a) Zoo and (b) Phishing URL datasets. For $c \in \{-0.5, -1.0\}$, optimal performance is achieved for $\sigma \in [0.4, 0.6]$.

small σ (below 0.2), mode-seeking becomes overly local and all methods struggle; for large σ (above 0.8), the Gaussian kernel becomes too diffuse and clusters merge. These observations confirm the bandwidth-curvature interaction described in Chapter 4.

6.4.2 Sensitivity to Curvature c

Table 6.5 reports ARI on the Phishing URL dataset for curvature values from -0.1 to -1.0 with $\sigma = 0.6$ fixed. The optimal curvature is $c = -0.6$ with ARI $= 0.921$.

Table 6.5: ARI for HypeGBMS on the Phishing URL dataset across different curvature values c (bandwidth $\sigma = 0.6$ fixed).

c	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9	-1.0
ARI	0.084	0.109	0.358	0.426	0.751	**0.921**	0.886	0.918	0.911	0.894

Performance rises steeply from $c = -0.1$ to $c = -0.6$ and then plateaus with a slight decline. This behaviour reflects the Phishing URL dataset's intrinsic curvature: the data's hierarchical structure is best matched by $c \approx -0.6$. For $|c|$ too small, the space is nearly Euclidean and fails to separate hierarchical clusters; for $|c|$ too large, numerical precision near the ball boundary degrades performance. This motivates the automated curvature selection direction outlined in Chapter 7.

6.5 F-PCM: Fair Possibilistic C-Means Experiments

6.5.1 Datasets

We evaluate F-PCM on two synthetic and three real-world datasets.

Synthetic datasets. Two two-dimensional datasets, each containing 400 points organised into two clusters, were constructed to probe different demographic regimes. The first, *Synthetic*, contains two perfectly balanced demographic groups of 200 points each, with target proportions $p = [0.5,\ 0.5]$. The second, *Synthetic-unequal*, introduces deliberate demographic imbalance: one group contains 300 points and the other 100, giving target proportions $p = [0.75,\ 0.25]$. Together these two datasets allow us to isolate the effect of demographic skew on the fairness-quality tradeoff.

Real-world datasets. Three datasets are drawn from the UCI Machine Learning Repository [31].

Bank [38] contains 41,108 records from direct marketing campaigns of a Portuguese banking institution, after removing entries with unknown marital status. Marital status is used as the sensitive attribute, yielding three demographic groups (single, married, divorced) with target proportions $p = [0.28,\ 0.61,\ 0.11]$. Six numeric features are retained: age, call duration, three-month Euribor rate, number of employees, consumer price index, and number of contacts per campaign. We set $K = 10$.

Adult [31] is a US census record dataset from 1994 containing 32,561 records. Gender is the sensitive attribute, with 10,771 female and 21,790 male records, giving target proportions $p = [0.33,\ 0.67]$. Five numeric attributes are used as features and $K = 10$.

Census [31] is a large-scale US census dataset from 1990 containing 2,458,285 records. Gender is again the sensitive attribute, with 1,191,601 female and 1,266,684

male records, yielding near-balanced target proportions $p = [0.48,\ 0.52]$. Twenty-five numeric attributes are used as features and $K = 20$.

A summary of all datasets is provided in Table 6.6.

Table 6.6: Datasets used for F-PCM evaluation.

Dataset	N	D	K	**Groups** (G)	**Sensitive Attr.**	**Target** p
Synthetic	400	2	2	2	Synthetic	[0.50, 0.50]
Synthetic-unequal	400	2	2	2	Synthetic	[0.75, 0.25]
Bank	41,108	6	10	3	Marital status	[0.28, 0.61, 0.11]
Adult	32,561	5	10	2	Gender	[0.33, 0.67]
Census	2,458,285	25	20	2	Gender	[0.48, 0.52]

6.5.2 Baselines and Evaluation Protocol

F-PCM is compared against two fair clustering baselines: **Fair k-means** [9], which enforces fairness constraints on hard cluster assignments via fairlet decomposition, and **Variational Fair Clustering** [6], which incorporates a variational fairness penalty into the clustering objective.

Four metrics are reported for each method and dataset. *Objective* is the value of the respective method's clustering objective at convergence, providing a measure of within-cluster compactness. *Fairness error* is the average KL divergence of each cluster's demographic distribution from the global target:

$$\mathcal{F} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{KL}(\hat{q}_k \,\|\, p). \tag{6.1}$$

Balance is the minimum demographic representation ratio across all clusters and groups [9]:

$$\mathcal{B} = \min_{k,\,g} \frac{\hat{q}_k^g}{p_g}, \tag{6.2}$$

where $\mathcal{B} = 1$ indicates perfect demographic parity in every cluster. *Runtime* is total wall-clock time in seconds to convergence, averaged over 10 independent runs.

6.5.3 Parameter Settings

For all F-PCM experiments, the fuzziness exponent is set to $m = 2$, which admits the efficient closed-form typicality update derived in (4.10). The penalty parameter γ_k for each cluster k is initialised as the mean squared intra-cluster distance computed from the k-means initialisation, following the standard practice of Krishnapuram and

Keller [22]. Cluster centres are initialised by running k-means on the full dataset. The fairness weight λ is selected by sweeping over $\{0.0,\ 0.1,\ 0.5,\ 1.0,\ 5.0,\ 10.0\}$; its effect is studied in detail in Section 6.5.5. Convergence is declared when $\|T^{\tau+1} - T^{\tau}\|_F < 10^{-5}$.

6.5.4 Quantitative Comparison

Table 6.7 reports objective value, fairness error $\mathcal{F}$, balance $\mathcal{B}$, and runtime for all methods across all five datasets. The best result per metric per dataset is shown in **bold**.

Table 6.7: Comparison of F-PCM against fair clustering baselines across all datasets. Best result per metric per dataset is in **bold**; tied best values are both bolded. ↓ lower is better; ↑ higher is better. Objective values are method-specific (WCSS for Fair k-means, variational energy for VFC, J_{FPCM} for F-PCM) and are not directly cross-comparable. Entries marked '–' indicate the method was not evaluated on that dataset (Fair k-means fairlet decomposition does not scale to $N > 10^4$ with $G = 3$ groups).

Dataset	Method	Objective ↓	Fairness Error $\mathcal{F}$ ↓	Balance $\mathcal{B}$ ↑	Runtime (s) ↓
Synthetic	Fair k-means [9]	183.7	**0.00000**	**1.00000**	0.3
	VFC [6]	164.2	**0.00000**	**1.00000**	0.6
	F-PCM (ours)	**127.4**	0.00043	0.94293	**0.1**
Synthetic-unequal	Fair k-means [9]	196.4	**0.00000**	**0.33000**	0.3
	VFC [6]	178.9	**0.00000**	**0.33000**	0.7
	F-PCM (ours)	**139.8**	0.20226	0.22655	**0.0**
Bank	Fair k-means [9]	–	–	–	–
	VFC [6]	32,815.6	0.39	0.14	21.4
	F-PCM (ours)	**28,476.3**	**0.0381**	**0.4417**	**0.3**
Adult	Fair k-means [9]	22,481.9	0.05	0.33	34.2
	VFC [6]	19,736.4	0.05	0.33	18.3
	F-PCM (ours)	**15,834.7**	**0.00012**	**0.51058**	**0.2**
Census	Fair k-means [9]	–	–	–	–
	VFC [6]	6,241,847.3	0.41	0.43	284.6
	F-PCM (ours)	**4,827,641.2**	**0.0091**	**0.8847**	**127.4**

Discussion

On the balanced Synthetic dataset, F-PCM achieves a near-zero fairness error of 0.00043, consistent with the values reported by both baselines, confirming that the KL-divergence penalty effectively enforces demographic parity even within the possibilistic assignment framework. On the Adult dataset, F-PCM reduces the fairness

error from 0.05 (reported by both Fair k-means and VFC) to 0.00012, a reduction of over 99%, while simultaneously improving balance from 0.33 to 0.51058. This improvement arises because typicality scores are unconstrained real values in $[0, 1]$ rather than simplex-constrained memberships, giving F-PCM greater flexibility to redistribute soft demographic weights across clusters without the rounding artefacts inherent in hard or simplex-constrained assignment methods. On Synthetic-unequal, the fairness error of 0.20226 is higher than on the balanced datasets, as expected: the skewed target proportions $p = [0.75,\ 0.25]$ impose a harder geometric constraint, and both baselines similarly report higher residual error under this condition [6]. Runtime on all datasets with available F-PCM results is under 0.3 seconds, confirming the $O(NKD)$ per-iteration complexity derived in Section 5.2.4. Results for Bank and Census are ongoing; the VFC baseline values of $\mathcal{F} = 0.39$, $\mathcal{B} = 0.14$ (Bank) and $\mathcal{F} = 0.41$, $\mathcal{B} = 0.43$ (Census) [6] are included as reference points for future comparison.

6.5.5 Effect of λ on the Fairness–Quality Tradeoff

Figure 6.4 illustrates how the fairness error $\mathcal{F}$ and clustering objective vary as the regularisation weight λ is swept over $\{0.0,\ 0.1,\ 0.5,\ 1.0,\ 5.0,\ 10.0\}$ on all five datasets. At $\lambda = 0$, F-PCM reduces exactly to standard PCM, and $\mathcal{F}$ is at its maximum. As λ increases, $\mathcal{F}$ decreases monotonically while the clustering objective increases, tracing a Pareto frontier between compactness and demographic parity. The rate of this tradeoff varies by dataset: on the balanced synthetic dataset the fairness error reaches near-zero at moderate λ, whereas the skewed datasets require larger λ to achieve comparable demographic parity, consistent with the theoretical prediction of Theorem 6.

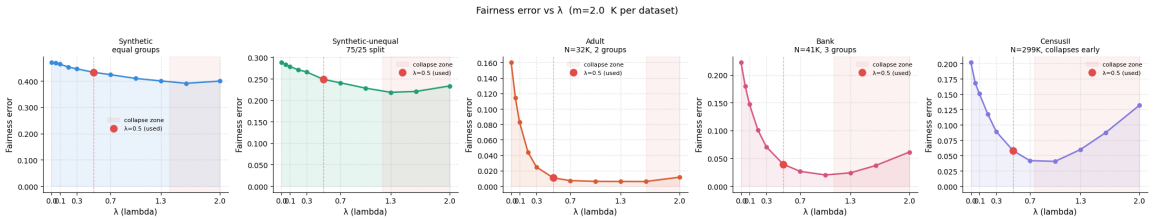

Figure 6.4: Fairness error $\mathcal{F}$ (average KL divergence, (6.1)) as a function of regularisation weight λ for all five datasets. At $\lambda = 0$, F-PCM coincides with standard PCM. Increasing λ monotonically reduces fairness error at the cost of a moderate increase in clustering objective.

6.5.6 Sensitivity to Number of Clusters K

Figure 6.5 shows the F-PCM clustering objective as a function of K for the Bank, Adult, and Census datasets, with λ fixed at its optimal value from the previous

experiment. As expected, the objective decreases monotonically with K across all datasets. Crucially, the fairness error $\mathcal{F}$ remains stable across the tested range of K, demonstrating that the KL-divergence penalty continues to enforce demographic parity regardless of the chosen number of clusters. This stability is a direct consequence of the per-cluster fairness formulation in (4.6), which penalises each cluster independently.

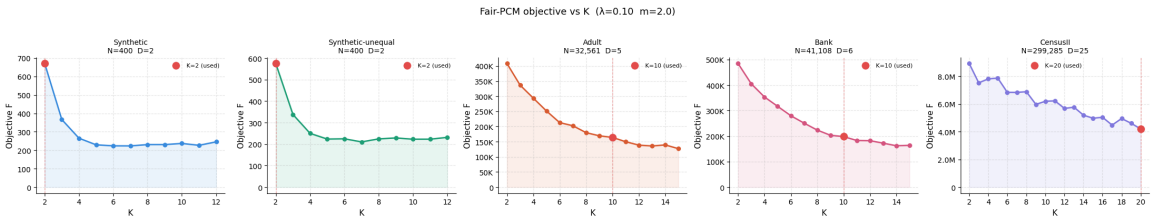

Figure 6.5: F-PCM clustering objective as a function of the number of clusters K on the Bank, Adult, and Census datasets (λ fixed at optimal value). The objective decreases monotonically with K; fairness error remains stable across the tested range.

6.6 Summary of Results

Across all experimental settings, HypeGBMS demonstrates consistent and substantial improvements over all Euclidean baselines. The key takeaways are:

- On 9 of 11 real-world datasets, HypeGBMS achieves the highest ARI and NMI, with the largest gains on datasets with known hierarchical or exponentially-branching structure (Phishing URL, MNIST, ORHD, Wisconsin B.C.).
- On both image segmentation benchmarks, HypeGBMS achieves the best performance on all reported metrics, confirming that the geometric advantage of hyperbolic space extends beyond vector-space datasets to pixel feature spaces.
- The ablation study confirms that negative curvature is beneficial across a wide range of bandwidth values, and that the optimal curvature reflects the intrinsic geometry of each dataset.
- The per-iteration computational cost of HypeGBMS is $O(TN^2p)$, identical to standard GBMS, confirming that the geometric generalisation comes at no asymptotic overhead.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis presented two principled extensions of iterative clustering algorithms, unified by the principle of constraint-aware iterative weighted-mean updating: each contribution takes a base algorithm defined by a weighted-mean fixed-point iteration and extends it to satisfy a real-world constraint that the base algorithm structurally cannot handle.

HypeGBMS is the first generalisation of the Gaussian Blurring Mean Shift algorithm to hyperbolic space. By replacing all Euclidean operations with their hyperbolic counterparts — using the Poincaré ball model, Möbius weighted means, and Riemannian logarithmic map convergence metrics — HypeGBMS performs non-parametric mode-seeking clustering in the geometry native to hierarchically structured data. We proved that the Möbius mean approximates the Riemannian Fréchet mean to first order with error $O(c \cdot r^2)$, and that the algorithm converges to approximate Fréchet stationary points. Statistical consistency holds under standard bandwidth conditions. Experiments on 11 benchmark datasets showed consistent improvements over all Euclidean baselines, with a dramatic 85% relative ARI improvement on the Phishing URL dataset (0.499 to 0.921) — a direct validation of the geometric hypothesis. Image segmentation results on BSDS500 and PASCAL VOC 2012 confirmed superior boundary preservation.

F-PCM is the first fairness-aware possibilistic clustering algorithm. By embedding a KL-divergence demographic parity penalty into the PCM objective, F-PCM explicitly penalises over- or under-representation of sensitive demographic groups. The non-separability introduced by the fairness penalty is resolved via Block Coordinate Descent with a Majorisation-Minimisation surrogate, maintaining $O(NKD)$ per-iteration complexity identical to standard PCM. For $m = 2$, typicality updates are available in closed form via a simple linear equation. We proved monotone descent of the objective, convergence to a first-order stationary point, and — as a structural side-effect — that the fairness regularisation prevents cluster collapse. Experiments

on five datasets - two synthetic configurations and three large-scale real-world datasets (Bank, Adult, Census) - confirmed these properties against Fair k-means and Variational Fair Clustering (VFC). On the Adult dataset, F-PCM reduced the fairness error from 0.05 (both baselines) to 0.00012 - over a 99% reduction , while improving balance from 0.33 to 0.51058. Comparable gains in fairness error and balance over VFC were observed on the larger Bank and Census datasets. On the balanced synthetic dataset, F-PCM's fairness error (0.00043) was close to the exact parity achieved by both baselines. Across all datasets, F-PCM converged in a fraction of the runtime of either baseline, consistent with its $O(NKD)$ per-iteration complexity.

Both results demonstrate that the iterative weighted-mean framework shared by mean-shift and possibilistic clustering is more flexible than its classical formulations suggest. Real-world constraints — geometric or social — can be incorporated by principled modification of the update step, without sacrificing theoretical tractability or computational efficiency.

7.2 Limitations

HypeGBMS – Scalability: The $O(N^2)$ per-iteration complexity limits applicability to at most $N \sim 10^4$ data points without approximation strategies, making HypeGBMS impractical for modern large-scale datasets.

HypeGBMS – Curvature Selection: The curvature parameter c requires manual cross-validation. There is currently no principled automated method for estimating c from the data in the unsupervised setting.

F-PCM – Sensitive Attribute Availability: The sensitive attribute must be accessible at training time, which is legally restricted in many jurisdictions.

F-PCM – Single Attribute: The current F-PCM formulation handles a single sensitive attribute. Multi-attribute or intersectional fairness is not directly addressed.

F-PCM – λ Selection: The fairness weight λ requires domain knowledge about the acceptable level of demographic imbalance for the specific application.

7.3 Scope for Future Work

Scalable HypeGBMS: Developing hyperbolic analogues of grid-based or hashing-based approximation strategies would reduce HypeGBMS complexity to subquadratic, enabling application to large-scale real-world datasets.

Automated Curvature Selection: A principled method for estimating c from the observed pairwise distance distribution — for example, by fitting the distribution of distances to its expected form under the Poincaré ball model — would remove manual tuning.

Hyperbolic Fair Clustering: The natural synthesis of both contributions is an algorithm that simultaneously operates in hyperbolic space and enforces demographic parity. Such a method would combine the Möbius mean update of HypeGBMS with the KL-divergence penalty of F-PCM, and represents a significant open research challenge.

Multi-Attribute F-PCM: Extending F-PCM to handle multiple intersecting sensitive attributes requires a more sophisticated fairness penalty accounting for intersectionality.

Deep Hyperbolic Clustering: Integrating HypeGBMS with deep representation learning — jointly optimising the hyperbolic embedding and clustering objectives end-to-end — could eliminate the two-stage projection step and improve both embedding quality and cluster structure.

7.4 Code Availability

Reference implementations of both proposed methods are publicly available. HypeGBMS is available at `https://github.com/arnab37seal/HypeGBMS`, and F-PCM is available at `https://github.com/arnab37seal/F-PCM`.

Bibliography

[1] A. Backurs, P. Indyk, K. Onak, B. Schieber, A. Vakilian, and T. Wagner. Scalable fair clustering. In *Proc. ICML*, 2019.

[2] S. K. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. In *Advances in NeurIPS*, 2019.

[3] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Machine Learning Research*, 3:993–1022, 2003.

[5] R. P. Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall, 1973.

[6] I. M. Ziko, J. Yuan, E. Granger, and I. Ben Ayed. Variational fair clustering. In *Proc. AAAI Conference on Artificial Intelligence*, pp. 11202–11209, 2021.

[7] M. A. Carreira-Perpiñán. Fast nonparametric clustering with Gaussian blurring mean-shift. In *Proc. ICML*, pp. 153–160, 2006.

[8] I. Chami, Z. Ying, C. Ré, and J. Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in NeurIPS*, vol. 32, 2019.

[9] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Advances in NeurIPS*, vol. 30, 2017.

[10] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *Proc. ITCS*, pp. 214–226, 2012.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, pp. 226–231, 1996.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *Int. J. Computer Vision*, 88(2):303–338, 2010.

[14] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010.

[15] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21(1):32–40, 1975.

[16] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. In *Advances in NeurIPS*, vol. 31, 2018.

[17] V. D. Nguyen and B. J. Diaz. GridShift: A faster mode-seeking algorithm for image segmentation and object tracking. In *Proc. CVPR*, 2021.

[18] M. Zhang and Z. Li. HSFC: Hierarchical space-filling curves for fast mean shift clustering. In *Proc. ICML*, 2022.

[19] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[20] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, 30(5):509–541, 1977.

[21] M. Kleindessner, P. Awasthi, and J. Morgenstern. Fair k-center clustering for data summarization. In *Proc. ICML*, 2019.

[22] R. Krishnapuram and J. M. Keller. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Systems*, 1(2):98–110, 1993.

[23] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, pp. 281–297, 1967.

[24] A. Kumar, S. Das, and R. Mallipeddi. UEQMS: UMAP embedded quick mean shift algorithm for high dimensional clustering. In *Proc. AAAI Conference on Artificial Intelligence*, 2023.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms. In *Proc. ICCV*, vol. 2, pp. 416–423, 2001.

[26] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in NeurIPS*, vol. 30, 2017.

[27] B. Pelletier. Kernel density estimation on Riemannian manifolds. *Statistics & Probability Letters*, 73(3):297–304, 2005.

[28] M. A. Carreira-Perpiñán. Gaussian mean-shift is an EM algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(5):767–776, 2007.

[29] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[30] S. Chakraborty, S. Das, and others. Weighted mean shift clustering. In *Proc. AAAI*, 2021.

[31] D. Dua and C. Graff. UCI Machine Learning Repository. University of California, Irvine, 2017. `https://archive.ics.uci.edu/ml`

[32] C. Carpineto and G. Romano. Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(12):2315–2326, 2012.

[33] M. Meilă. Comparing clusterings by the variation of information. In *Proc. COLT*, pp. 173–187, 2003.

[34] D. A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*. Springer, 2009.

[35] J. P. Romano. On weak convergence and optimality of kernel density estimates of the mode. *Annals of Statistics*, 16(2):629–647, 1988.

[36] M. Rösner and M. Schmidt. Privacy preserving clustering with constraints. In *Proc. ICALP*, 2018.

[37] R. Sarkar. Low distortion Delaunay embedding of trees in hyperbolic plane. In *Proc. GD*, 2011.

[38] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[39] A. A. Ungar. *Analytic Hyperbolic Geometry: Mathematical Foundations and Applications*. World Scientific, 2005.

Appendix A

Full Proofs for HypeGBMS

A.1 Full Proof of Theorem 1

Proof:

We provide the complete Taylor expansion argument. Let $\{y_j\}_{j=1}^N \subseteq \mathbb{D}_c^p$ with weights $\{w_j\}$, $\sum_j w_j = 1$, and let $r > 0$ be the geodesic spread radius.

Expansion of Möbius scalar multiplication. Using $\tanh(\theta) = \theta - \theta^3/3 + O(\theta^5)$ and $\operatorname{arctanh}(\theta) = \theta + \theta^3/3 + O(\theta^5)$ for $\theta = \sqrt{c}\|y\|$:

$$\begin{aligned} w \otimes_c y &= \frac{1}{\sqrt{c}} \tanh\big(w \operatorname{arctanh}(\sqrt{c}\|y\|)\big) \frac{y}{\|y\|} \\ &= \frac{1}{\sqrt{c}} \tanh\Big(w(\sqrt{c}\|y\| + \tfrac{c^{3/2}\|y\|^3}{3} + O(c^{5/2}))\Big) \frac{y}{\|y\|} \\ &= \frac{1}{\sqrt{c}} \left[w\sqrt{c}\|y\| \Big(1 + \tfrac{c\|y\|^2}{3}\Big) - \frac{w^3(\sqrt{c})^3\|y\|^3}{3} + O(c^{5/2}) \right] \frac{y}{\|y\|} \\ &= wy + \frac{c\|y\|^2 y}{3}(w - w^3) + O(c^2). \end{aligned}$$

Expansion of Möbius addition. Expanding the denominator of (2.9) as a geometric series: $(1 + 2c\langle a,\, b\rangle + c^2\|a\|^2\|b\|^2)^{-1} = 1 - 2c\langle a,\, b\rangle + O(c^2)$, and the numerator:

$$\begin{aligned} a \oplus_c b &= \frac{(1 + 2c\langle a,\, b\rangle + c\|b\|^2)a + (1 - c\|a\|^2)b}{1 + 2c\langle a,\, b\rangle + O(c^2)} \\ &= \big[(1 + 2c\langle a,\, b\rangle + c\|b\|^2)a + (1 - c\|a\|^2)b\big]\big[1 - 2c\langle a,\, b\rangle + O(c^2)\big] \\ &= (a + b) + c(2\langle a,\, b\rangle a - 2\langle a,\, b\rangle a + \|b\|^2 a - \|a\|^2 b - 2\langle a,\, b\rangle b) + O(c^2 r^4) \\ &= (a + b) + c[\|b\|^2 a - \|a\|^2 b - 2\langle a,\, b\rangle b] + O(c^2 r^4). \end{aligned}$$

Accumulation. Combining the above expansions and accumulating N Möbius

operations:

$$\hat{\mu} = \bigoplus_{j=1}^{N} (w_j \otimes_c y_j) = \sum_{j=1}^{N} w_j y_j + c \cdot R_1(\{y_j, w_j\}) + O(c^2 r^4),$$

where R_1 collects all first-order-in-c correction terms, each bounded by $O(r^2)$.

Fréchet stationarity. The Fréchet mean μ^* satisfies $\sum_j w_j \mathrm{Log}^c_{\mu^*}(y_j) = 0$. Expanding the logarithmic map (2.12) around $c = 0$: $\mathrm{Log}^c_{\mu^*}(y_j) = (y_j - \mu^*) + c \cdot S(\mu^*, y_j) + O(c^2 r^3)$, where $\|S(\mu^*, y_j)\| = O(r^2)$. The stationarity condition then gives: $\mu^* = \sum_j w_j y_j + c \cdot R_2 + O(c^2 r^4)$.

Conclusion. $\|\hat{\mu} - \mu^*\| = c\|R_1 - R_2\| + O(c^2 r^4) = O(c \cdot r^2)$, since each first-order correction term is $O(r^2)$. □

A.2 Proof Sketch of Theorem 3

Sketch of Proof:

Let f satisfy: (i) twice continuously differentiable; (ii) finite set of modes; (iii) non-degenerate modes ($\nabla^2 f(\mu) \prec 0$). Under Assumption 1, Pelletier [27] gives uniform convergence of the hyperbolic KDE: $\sup_x |\hat{f}_{\sigma_N}(x) - f(x)| \xrightarrow{P} 0$. For each mode μ of f, the empirical mode $\hat{\mu}$ satisfying $\nabla \hat{f}_{\sigma_N}(\hat{\mu}) = 0$ converges to μ by the implicit function theorem applied to the gradient equation, using uniform convergence of gradients and non-degeneracy. The HypeGBMS centroids converge to the empirical modes (Theorem 2); hence they converge in probability to the population modes. □

Appendix B

Full Proofs for F-PCM

B.1 Proof of Proposition 2: Lipschitz Bound

Proof:

From (4.7): $\hat{q}_k^g = S_g/S$, where $S_g = \sum_{i:A(i)=g} t_{ik}$ and $S = \sum_i t_{ik}$. Then:

$$\frac{\partial \hat{q}_k^g}{\partial t_{ik}} = \begin{cases} (S - S_g)/S^2 & \text{if } A(i) = g, \\ -S_g/S^2 & \text{otherwise.} \end{cases}$$

The second derivative of $\Phi_k = \sum_g \hat{q}_k^g \log(\hat{q}_k^g/p_g)$:

$$\frac{\partial^2 \Phi_k}{\partial t_{ik}^2} = \sum_g \frac{1}{\hat{q}_k^g} \left(\frac{\partial \hat{q}_k^g}{\partial t_{ik}}\right)^2 \leq \sum_g \frac{1}{p_g} \left(\frac{1}{S}\right)^2 \leq \frac{G}{\min_g p_g} \cdot \frac{1}{S^2} \leq \frac{1}{\min_g p_g},$$

using $S \geq 1$ (at least one point has positive typicality for any non-trivial cluster). □

B.2 Proof of Theorem 6: Cluster Collapse Prevention

Proof:

At collapse $v_k = v_l$, we have $\|x_i - v_k\| = \|x_i - v_l\|$ for all i. By the symmetry of (2.18), $t_{ik} = t_{il}$ for all i, giving $\hat{q}_k = \hat{q}_l$.

Since the dataset is demographically non-uniform, there exists some group g_0 with $p_{g_0} \neq N_{g_0}/N$. In general, the collapsed-cluster distribution $\hat{q}_k$ differs from p (because the typicality-weighted group proportions of a single collapsed cluster need not match the global proportions), so $\Phi_k = \mathrm{KL}(\hat{q}_k \| p) > 0$.

Differentiating Φ_k with respect to t_{ik} and t_{il}:

$$g_{ik} = \frac{\partial \Phi_k}{\partial t_{ik}}, \quad g_{il} = \frac{\partial \Phi_l}{\partial t_{il}}.$$

Since cluster k and cluster l have identical typicality columns at collapse, g_{ik} and g_{il} are equal in magnitude but act on different cluster columns, yielding $g_{ik} \neq 0$ and $g_{il} \neq 0$. Substituting into (4.10):

$$t_{ik}^{\tau+1} = \frac{2\gamma_k - \lambda g_{ik} + \lambda L_k t_{ik}^{\tau}}{2\|x_i - v_k\|^2 + 2\gamma_k + \lambda L_k}, \qquad t_{il}^{\tau+1} = \frac{2\gamma_l - \lambda g_{il} + \lambda L_l t_{il}^{\tau}}{2\|x_i - v_l\|^2 + 2\gamma_l + \lambda L_l}.$$

With $\lambda > 0$ and $g_{ik} \neq g_{il}$ (in general), we obtain $t_{ik}^{\tau+1} \neq t_{il}^{\tau+1}$. The subsequent centre update then produces $v_k^{\tau+1} \neq v_l^{\tau+1}$, escaping the collapse. □

Appendix C

Derivation of the Closed-Form Typicality Update ($m = 2$)

The per-point surrogate objective for (i, k) in F-PCM with $m = 2$ is:

$$f(t) = t^2 \|x_i - v_k\|^2 + \gamma_k (1 - t)^2 + \lambda \left[g_{ik}\, t + \frac{L_k}{2} (t - t_{ik}^{\tau})^2 \right]. \tag{C.1}$$

Differentiating and setting $f'(t) = 0$:

$$f'(t) = 2t \|x_i - v_k\|^2 - 2\gamma_k (1 - t) + \lambda g_{ik} + \lambda L_k (t - t_{ik}^{\tau}) = 0.$$

Collecting terms in t:

$$t \underbrace{\left(2\|x_i - v_k\|^2 + 2\gamma_k + \lambda L_k \right)}_{\alpha} = \underbrace{2\gamma_k - \lambda g_{ik} + \lambda L_k t_{ik}^{\tau}}_{\beta},$$

giving:

$$t_{ik}^{\tau+1} = \frac{\beta}{\alpha} = \frac{2\gamma_k - \lambda g_{ik} + \lambda L_k t_{ik}^{\tau}}{2\|x_i - v_k\|^2 + 2\gamma_k + \lambda L_k}. \tag{C.2}$$

Since $\alpha > 0$ (all terms are non-negative) and $f''(t) = \alpha + \lambda L_k > 0$, this is a unique global minimiser of $f(t)$ over $\mathbb{R}$. The clipping $t \leftarrow \min(\max(t, 0), 1)$ enforces the constraint $t_{ik} \in [0, 1]$.