A. Proofs

A.1. Proof of Theorem 3.1:

• We first show ($\Rightarrow$): let us note $\alpha = \|\mathbf{a}\|^2$, we have the following equality $\|h(g_\beta(\mathbf{a}))\|^2 = \frac{\alpha}{(\sqrt{\alpha+1}+1)^2}$ which is an increasing function in α when α is positive.

• We now show ($\Leftarrow$): We give the formulation of $g_\beta^{-1} \circ h^{-1}$:

$$\forall \mathbf{c} = (c_1, \cdots, c_d) \in \mathcal{P}^d, h^{-1}(\mathbf{c}) = \left(\sqrt{\left\| \frac{2\mathbf{c}}{1-\|\mathbf{c}\|^2} \right\|^2 + \beta}, \frac{2c_1}{1-\|\mathbf{c}\|^2}, \cdots, \frac{2c_d}{1-\|\mathbf{c}\|^2} \right) \in \mathcal{H}^{d,\beta} \tag{20}$$

$$= \left(\sqrt{\left\| \frac{2\mathbf{c}}{1-\|\mathbf{c}\|^2} \right\|^2 + \beta}, \frac{2\mathbf{c}}{1-\|\mathbf{c}\|^2} \right) \in \mathcal{H}^{d,\beta} \tag{21}$$

The vector $g_\beta^{-1}(h^{-1}(\mathbf{c}))$ is obtained by removing the first element of $h^{-1}(\mathbf{c})$:

$$\forall \mathbf{c} \in \mathcal{P}^d, g_\beta^{-1}(h^{-1}(\mathbf{c})) = \frac{2\mathbf{c}}{1-\|\mathbf{c}\|^2} \tag{22}$$

By using these definitions, we have the following implication: we recall that $\forall \mathbf{c} \in \mathcal{P}^d$, $\|\mathbf{c}\| < 1$ due to the definition of $\mathcal{P}^d$. Let us note $\gamma = \|\mathbf{c}\|^2 < 1$, we have $\|g_\beta^{-1}(h^{-1}(\mathbf{c}))\|^2 = \frac{4\gamma}{(1-\gamma)^2}$ which is an increasing function in γ on $(0, 1)$. QED

A.2. Proof of Lemma 3.2

By using the linearity of the Lorentzian inner product, we have the following equality:

$$\max_{\boldsymbol{\mu} \in \mathcal{H}^{d,\beta}} \sum_{i=1}^{n} \nu_i \langle \mathbf{x}_i, \boldsymbol{\mu} \rangle_{\mathcal{L}} = \max_{\boldsymbol{\mu} \in \mathcal{H}^{d,\beta}} \langle \sum_{i=1}^{n} \nu_i \mathbf{x}_i, \boldsymbol{\mu} \rangle_{\mathcal{L}} \tag{23}$$

Since $\boldsymbol{\mu} \in \mathcal{H}^{d,\beta}$ and $\forall \mathbf{a} \in \mathcal{H}^{d,\beta}, \mathbf{b} \in \mathcal{H}^{d,\beta}$, $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{L}}$ is maximized if $\mathbf{a} = \mathbf{b}$, Eq. (23) is then maximized by finding the scaling factor $\gamma > 0$ that satisfies $(\gamma \sum_{i=1}^{n} \nu_i \mathbf{x}_i) = \boldsymbol{\mu} \in \mathcal{H}^{d,\beta}$. We explain in the next paragraph how to construct such a γ.

For any positive time-like vector $\mathbf{a}$, the vector $\mathbf{b} = \frac{1}{|\|\mathbf{a}\|_{\mathcal{L}}|}\mathbf{a}$ satisfies $\langle \mathbf{b}, \mathbf{b} \rangle_{\mathcal{L}} = \frac{\langle \mathbf{a}, \mathbf{a} \rangle_{\mathcal{L}}}{|\langle \mathbf{a}, \mathbf{a} \rangle_{\mathcal{L}}|} = -1$. Therefore, $\beta \langle \mathbf{b}, \mathbf{b} \rangle_{\mathcal{L}} = \langle \sqrt{\beta}\mathbf{b}, \sqrt{\beta}\mathbf{b} \rangle_{\mathcal{L}} = -\beta$ (*i.e.* $\sqrt{\beta}\mathbf{b} \in \mathcal{H}^{d,\beta}$ by definition since $\sqrt{\beta}\mathbf{b}$ is positive) and we then have $\gamma = \frac{\sqrt{\beta}}{|\|\sum_{i=1}^{n} \nu_i \mathbf{x}_i\|_{\mathcal{L}}|}$. QED

A.3. Proof of Theorem 3.4

Let $\mathbf{f}_1, \cdots, \mathbf{f}_n$ be a set of at least two different points in $\mathcal{F}^d$, and $\boldsymbol{\mu}$ be the centroid of the set $g_\beta(\mathbf{f}_1), \cdots, g_\beta(\mathbf{f}_n)$ (with same weighting $\forall i, \nu_i = 1$). The squared Euclidean norm of $g_\beta^{-1}(\boldsymbol{\mu}) \in \mathcal{F}^d$ is:

$$\|g_\beta^{-1}(\boldsymbol{\mu})\|^2 = \frac{\beta}{-\|\sum_{i=1}^{n} g_\beta(\mathbf{f}_i)\|_{\mathcal{L}}^2} \|\sum_{i=1}^{n} \mathbf{f}_i\|^2 \tag{24}$$

Since $\|\sum_{i=1}^{n} \mathbf{f}_i\|^2$ does not depend on β and the factor $\frac{\beta}{-\|\sum_{i=1}^{n} g_\beta(\mathbf{f}_i)\|_{\mathcal{L}}^2}$ is positive for $\beta > 0$, we only need to show that it is a increasing function in β, or equivalently, that its reciprocal is decreasing in β.

The reciprocal of $\frac{\beta}{-\| \sum_{i=1}^n g_\beta(\mathbf{f}_i)\|_{\mathcal{L}}^2}$ is:

$$\frac{-\| \sum_{i=1}^n g_\beta(\mathbf{f}_i)\|_{\mathcal{L}}^2}{\beta} = \frac{-[\sum_{i=1}^n \|g_\beta(\mathbf{f}_i)\|_{\mathcal{L}}^2 + 2\sum_{i=1}^n \sum_{j>i}^n \langle g_\beta(\mathbf{f}_i), g_\beta(\mathbf{f}_j)\rangle_{\mathcal{L}}]}{\beta} \tag{25}$$

$$= n - \frac{2}{\beta}\sum_{i=1}^n \sum_{j>i}^n \langle g_\beta(\mathbf{f}_i), g_\beta(\mathbf{f}_j)\rangle_{\mathcal{L}} \tag{26}$$

$$= n + \frac{2}{\beta}\sum_{i=1}^n \sum_{j>i}^n \langle \sqrt{\|\mathbf{f}_i\|^2 + \beta}, \sqrt{\|\mathbf{f}_j\|^2 + \beta}\rangle - \langle \mathbf{f}_i, \mathbf{f}_j\rangle \tag{27}$$

We then need to show that the function

$$\frac{1}{\beta}\left[\langle \sqrt{\|\mathbf{f}_i\|^2 + \beta}, \sqrt{\|\mathbf{f}_j\|^2 + \beta}\rangle - \langle \mathbf{f}_i, \mathbf{f}_j\rangle\right] \tag{28}$$

is decreasing in β if $\mathbf{f}_i \neq \mathbf{f}_j$. For this purpose, We study the sign of its gradient. The gradient of Eq. (28) *w.r.t.* β is written:

$$\frac{2\langle \mathbf{f}_i, \mathbf{f}_j\rangle \sqrt{\|\mathbf{f}_i\|^2 + \beta}\sqrt{\|\mathbf{f}_j\|^2 + \beta} - \beta(\|\mathbf{f}_i\|^2 + \|\mathbf{f}_j\|^2) - 2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2}{2\beta^2 \sqrt{\|\mathbf{f}_i\|^2 + \beta}\sqrt{\|\mathbf{f}_j\|^2 + \beta}} \tag{29}$$

The denominator is positive. If $\langle \mathbf{f}_i, \mathbf{f}_j\rangle$ is negative, then Eq. (29) is negative. If $\langle \mathbf{f}_i, \mathbf{f}_j\rangle = 0$, then Eq. (29) is 0 if $\mathbf{f}_i = \mathbf{f}_j = \mathbf{0}$, and negative otherwise. Otherwise, the Cauchy-Schwarz inequality is used to prove that Eq. (29) is negative. In other words, we need to prove that (assuming that $\langle \mathbf{f}_i, \mathbf{f}_j\rangle$ is positive):

$$2\langle \mathbf{f}_i, \mathbf{f}_j\rangle \sqrt{\|\mathbf{f}_i\|^2 + \beta}\sqrt{\|\mathbf{f}_j\|^2 + \beta} - \beta(\|\mathbf{f}_i\|^2 + \|\mathbf{f}_j\|^2) - 2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 < 0 \tag{30}$$

$$\iff 2\langle \mathbf{f}_i, \mathbf{f}_j\rangle \sqrt{\|\mathbf{f}_i\|^2 + \beta}\sqrt{\|\mathbf{f}_j\|^2 + \beta} < \beta(\|\mathbf{f}_i\|^2 + \|\mathbf{f}_j\|^2) + 2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 \tag{31}$$

$$\iff (2\langle \mathbf{f}_i, \mathbf{f}_j\rangle \sqrt{\|\mathbf{f}_i\|^2 + \beta}\sqrt{\|\mathbf{f}_j\|^2 + \beta})^2 < [\beta(\|\mathbf{f}_i\|^2 + \|\mathbf{f}_j\|^2) + 2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2]^2 \tag{32}$$

$$\iff 4\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2(\|\mathbf{f}_i\|^2 + \beta)(\|\mathbf{f}_j\|^2 + \beta) = 4\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2(\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 + \beta\|\mathbf{f}_i\|^2 + \beta\|\mathbf{f}_j\|^2 + \beta^2) \tag{33}$$

$$< 4\|\mathbf{f}_i\|^4\|\mathbf{f}_j\|^4 + 4\beta\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2(\|\mathbf{f}_i\|^2 + \|\mathbf{f}_j\|^2) + \beta^2\|\mathbf{f}_i\|^4 + \beta^2\|\mathbf{f}_j\|^4 + 2\beta^2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 \tag{34}$$

Since, by the Cauchy-Schwarz inequality, we have:

$$\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2 \leq \|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 \tag{35}$$

we then have (term by term):

$$4\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 \leq 4\|\mathbf{f}_i\|^4\|\mathbf{f}_j\|^4 \tag{36}$$

$$4\beta\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2\|\mathbf{f}_i\|^2 \leq 4\beta\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2\|\mathbf{f}_i\|^2 \tag{37}$$

$$4\beta\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2\|\mathbf{f}_j\|^2 \leq 4\beta\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2\|\mathbf{f}_j\|^2 \tag{38}$$

$$2\beta^2\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2 \leq 2\beta^2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 \tag{39}$$

$$2\beta^2\langle \mathbf{f}_i, \mathbf{f}_j\rangle^2 \leq 2\beta^2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 \leq \beta^2\|\mathbf{f}_i\|^4 + \beta^2\|\mathbf{f}_j\|^4 \tag{40}$$

Eq. (40) is explained by $\beta^2\|\mathbf{f}_i\|^4 + \beta^2\|\mathbf{f}_j\|^4 - 2\beta^2\|\mathbf{f}_i\|^2\|\mathbf{f}_j\|^2 = \beta^2(\|\mathbf{f}_i\|^2 - \|\mathbf{f}_j\|^2)^2 \geq 0$.

The right part of Eq. (40) is then an equality iff $\|\mathbf{f}_i\|^2 = \|\mathbf{f}_j\|^2$, and the Cauchy-Schwarz relation is an equality iff $\mathbf{f}_i$ and $\mathbf{f}_j$ are linearly dependent. Both of these conditions are satisfied iff $\mathbf{f}_i = \mathbf{f}_j$ (assuming that $\langle \mathbf{f}_i, \mathbf{f}_j\rangle$ is positive). However, we assume that there are at least two different points $\mathbf{f}_i$ and $\mathbf{f}_j$ such that $\mathbf{f}_i \neq \mathbf{f}_j$, which then implies that Eq. (29) is negative for this choice of $\mathbf{f}_i$ and $\mathbf{f}_j$ (*i.e.* when $\mathbf{f}_i \neq \mathbf{f}_j$).

This proves that the Euclidean norm in $\mathcal{F}^d$ of the centroid of different points decreases as $\beta > 0$ decreases. From Section A.1, the Euclidean norm of a point in $\mathcal{P}^d$ increases as the Euclidean norm of its projection in $\mathcal{F}^d$ increases, which completes the proof.

A.4. Proof of Theorem 3.5

The proof of Theorem 3.5 is as follows:

$$\arg\min_{\mathbf{b}\in\mathcal{B}} \sum_{i=1}^{n} \nu_i \mathsf{d}^2_{\mathcal{L}}(\mathbf{x}_i, \mathbf{b}) = \arg\min_{\mathbf{b}\in\mathcal{B}} \sum_{i=1}^{n} \nu_i \|\mathbf{x}_i - \mathbf{b}\|^2_{\mathcal{L}} = \arg\min_{\mathbf{b}\in\mathcal{B}} -2\beta \sum_{i=1}^{n} \nu_i - 2\sum_{i=1}^{n} \nu_i \langle \mathbf{x}_i, \mathbf{b}\rangle_{\mathcal{L}} \tag{41}$$

$$= \arg\min_{\mathbf{b}\in\mathcal{B}} \; -\langle \sum_{i=1}^{n} \nu_i \mathbf{x}_i, \mathbf{b}\rangle_{\mathcal{L}} = \arg\min_{\mathbf{b}\in\mathcal{B}} \; -\langle \sqrt{\beta} \frac{\sum_{i=1}^{n} \nu_i \mathbf{x}_i}{|\|\sum_{i=1}^{n} \nu_i \mathbf{x}_i\|_{\mathcal{L}}|}, \mathbf{b}\rangle_{\mathcal{L}} \tag{42}$$

$$= \arg\min_{\mathbf{b}\in\mathcal{B}} \; \mathsf{d}^2_{\mathcal{L}}(\boldsymbol{\mu}, \mathbf{b}) \tag{43}$$

A.5. Proof that the Poincaré metric is not differentiable when the distance is zero

We now show that the Poincaré metric in Eq. (3) is not differentiable when the distance is zero (*i.e.* $\mathbf{c} = \mathbf{d}$). We take the example when the dimensionality is 2 although the proof can be extended to any dimension.

Let us note $\mathbf{c} = (x, c_2)$ and $\mathbf{d} = (d_1, d_2)$ the considered points in Eq. (3) when the dimensionality is 2. We consider that our variable is x (i.e. the first element of $\mathbf{c}$). We also note $h = x - d_1$. Note that we could equivalently consider that our variables are the other elements of $\mathbf{c}$ or $\mathbf{d}$. Since we assume that $\mathbf{c} = \mathbf{d}$, we have $c_2 = d_2$ and we study the behavior of x when it tends to d_1 or equivalently when h tends to 0.

When $c_2 = d_2$, Eq. (3) can be written as:

$$cosh^{-1}(f(h)) = log(f(h) + \sqrt{f^2(h) - 1}) \tag{44}$$

where

$$f(h) = 1 + 2\frac{h^2}{(1 - (d_1 + h)^2 - d_2^2)(1 - d_1^2 - d_2^2)} = 1 + 2\frac{h^2}{\alpha} \tag{45}$$

where we note $\alpha = (1 - (d_1 + h)^2 - d_2^2)(1 - d_1^2 - d_2^2) > 0$ which is positive since the points $\mathbf{c}$ and $\mathbf{d}$ are constrained to be in the interior of the unit disk.

For all $h \in \mathbb{R}$, $cosh^{-1}(f(h))$ is then nonnegative.

We also have $f(0) = 1$ and $cosh^{-1}(f(0)) = 0$.

We now study two one-sided limits. The first one is the right-sided limit:

$$\lim_{h\to 0^+} \frac{cosh^{-1}(f(h)) - cosh^{-1}(f(0))}{h} = \lim_{h\to 0^+} \frac{cosh^{-1}(f(h))}{h} = \lim_{h\to 0^+} (cosh^{-1}(f(h)))' \tag{46}$$

which is nonnegative since $cosh^{-1}(f(h))$ is nonnegative and h is positive. The second one is the left-sided limit:

$$\lim_{h\to 0^-} \frac{cosh^{-1}(f(h)) - cosh^{-1}(f(0))}{h} = \lim_{h\to 0^-} \frac{cosh^{-1}(f(h))}{h} = \lim_{h\to 0^-} (cosh^{-1}(f(h)))' \tag{47}$$

which is nonpositive since $cosh^{-1}(f(h))$ is nonnegative and h is negative. We show in the following that these two limits are different, which implies that the function is not differentiable when $\mathbf{c} = \mathbf{d}$. We then study Eq. (46) and show that it is nonzero.

The derivative of $cosh^{-1}$ *w.r.t.* $z > 1$ is:

$$(cosh^{-1})'(z) = \frac{1}{\sqrt{z^2 - 1}} \tag{48}$$

The derivative of f *w.r.t.* h is:

$$f'(h) = \frac{4h(d_1 h + d_2^2 + d_1^2 - 1)}{(d_2^2 + d_1^2 - 1)(h^2 + 2d_1 h + d_2^2 + d_1^2 - 1)^2} \tag{49}$$

The derivative of $cosh^{-1}(f)$ *w.r.t.* h is then $f'(h)(cosh^{-1})'(f(h))$ which can be written:

$$f'(h)(cosh^{-1})'(f(h)) = \frac{4h(d_1h + d_2^2 + d_1^2 - 1)}{(d_2^2 + d_1^2 - 1)(h^2 + 2d_1h + d_2^2 + d_1^2 - 1)^2\sqrt{(1 + 2\frac{h^2}{(1-(d_1+h)^2-d_2^2)(1-d_1^2-d_2^2)})^2 - 1}} \tag{50}$$

$$= \frac{4h(d_1h + d_2^2 + d_1^2 - 1)}{(d_2^2 + d_1^2 - 1)(h^2 + 2d_1h + d_2^2 + d_1^2 - 1)^2\sqrt{\left(2\frac{h^2}{(1-(d_1+h)^2-d_2^2)(1-d_1^2-d_2^2)}\right)\left(2\frac{h^2}{(1-(d_1+h)^2-d_2^2)(1-d_1^2-d_2^2)} + 2\right)}} \tag{51}$$

$$= \frac{2h(d_1h + d_2^2 + d_1^2 - 1)}{(h^2 + 2d_1h + d_2^2 + d_1^2 - 1)\sqrt{h^2\,(h^2 + (1 - (d_1 + h)^2 - d_2^2)(1 - d_1^2 - d_2^2))}} \tag{52}$$

From the above equation, one can see that the right sided limit L is:

$$L := \lim_{h\to 0^+}(cosh^{-1}(f(h)))' = \lim_{h\to 0^+}\frac{2(d_1h + d_2^2 + d_1^2 - 1)}{(h^2 + 2d_1h + d_2^2 + d_1^2 - 1)\sqrt{(h^2 + (1 - (d_1 + h)^2 - d_2^2)(1 - d_1^2 - d_2^2))}} \tag{53}$$

$$= \frac{2}{1 - d_2^2 - d_1^2} = \frac{2}{1 - \|\mathbf{d}\|^2} \tag{54}$$

which is nonzero by definition of the domain of $\mathbf{d} = (d_1, d_2)$. For instance, when $\mathbf{d} = 0$, we have $L = 2$.

A similar proof can show that the left-sided limit is:

$$\lim_{h\to 0^-}(cosh^{-1}(f(h)))' = -L \tag{55}$$

The fact that these two one-sided limits are different shows that Eq. (3) is not differentiable when $\mathbf{c} = \mathbf{d}$.

Moreover, it is worth noting that L increases as the ℓ_2-norm of $\mathbf{d}$ increases.

A.6. Comparison of optimizers

We experimentally compared the convergence and retrieval performance of different optimizers to minimize $L_{\mathcal{D}}$ (defined in Eq. (13)) for the ACM dataset when using the squared Lorentzian distance with $\beta = 0.01$. The optimizers we compared are:

- the standard SGD with momentum
- the standard SGD without momentum
- *Outer-product approximation of the natural gradient* (Ollivier, 2015)
- *Quasi-diagonal approximation of the natural gradient* (Ollivier, 2015)

Fig. 4 illustrates the training loss, the mean rank and mean average precision performance of the different optimizers as a function of the number of epochs. The standard SGD obtains better retrieval performance than NG approximation methods although its convergence is slightly worse. Standard SGD with momentum is simple to implement, does not require storing gradients in a matrix, obtains good retrieval performance and shows faster convergence than SGD without momentum. We therefore choose this optimizer.

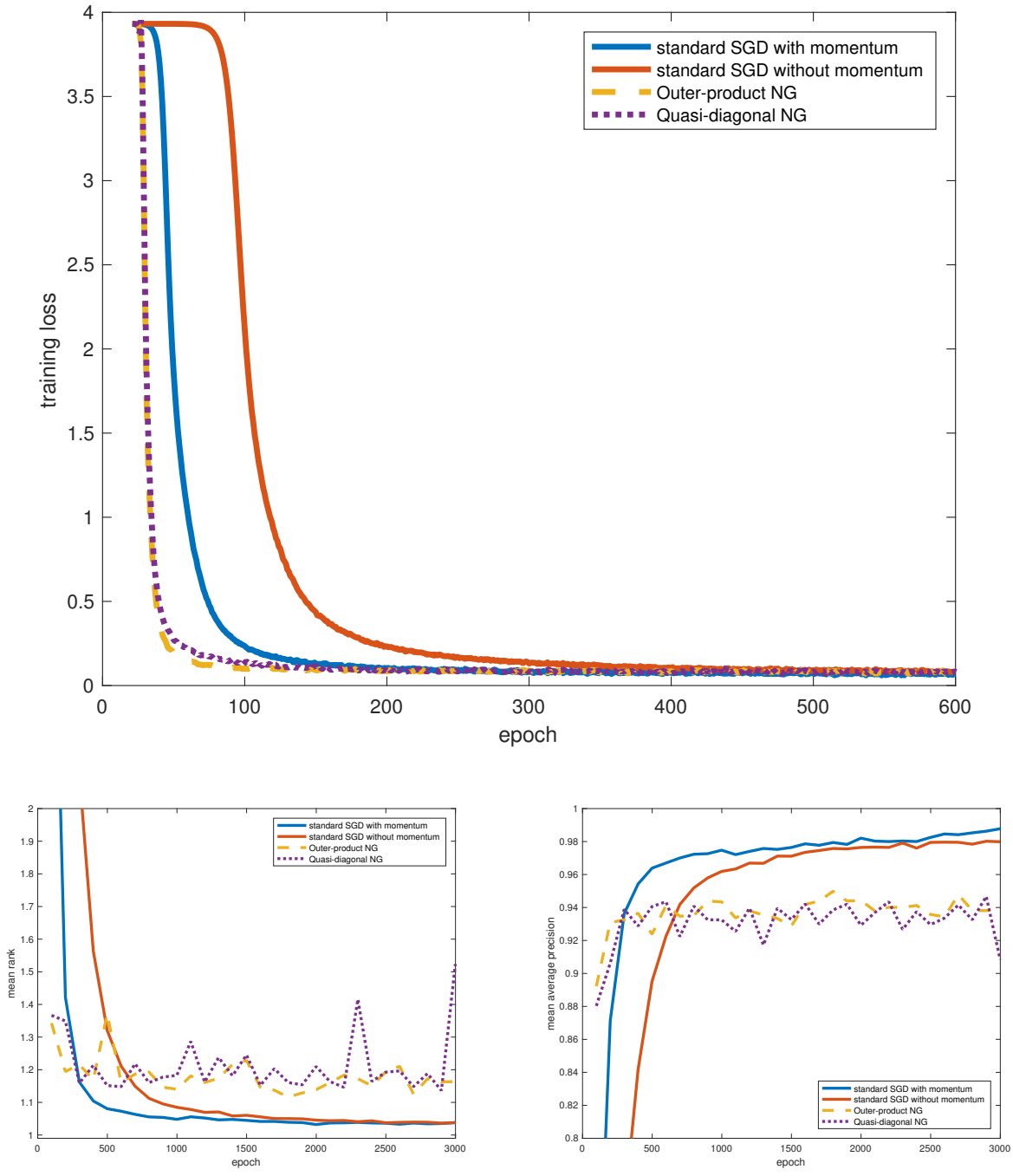


Figure 4. Comparison of four different solvers for the ACM dataset for different evaluation metrics as a function of the number of epochs.

Table 3. Percentage of node representations that have smaller Euclidean norm than their children in the tree when $\lambda = 0$

Method	Wordnet Nouns	Wordnet Verbs	EuroVoc	ACM	MeSH
Ours ($\beta = 0.01$)	97.2%	96.3%	98.6%	98.3%	97.7%
Ours ($\beta = 0.1$)	97.9%	96.0%	98.4%	98.0%	97.5%
Ours ($\beta = 1$)	94.3%	92.0%	98.3%	94.3%	93.2%
Poincaré distance	61.4%	52.5%	81.6%	99.9%	48.3%

A.7. Study of the percentage of node representations that have smaller Euclidean norm than their children

We report in Table 3 the percentage of nodes that have a Euclidean norm greater than their parent in the tree for the Poincaré distance and for the Lorentzian distance for different values of β and when $\lambda = 0$.

For the Lorentzian distance, more than 90% of pairs (parent,child) satisfy the desired order of Euclidean norms. The percentage increases with smaller values of β, this illustrates our point on the impact on the Euclidean norm of the center of mass.

We tried to run the same baseline with the Poincaré distance but it was very unstable to optimize. It returns in general a worse performance than the Lorentzian distance except for the ACM dataset which is the smallest and easiest dataset.

A.8. Details on the experiments of Section 4.3

The convolutional neural network that we adopt consist of 4 convolutional layers, each followed by batch normalization and ReLU activation function. 2 additional fully connected layers are applied after the convolution to obtain the feature vector per image.

The dimensionality of the learned representations d, and $\beta = 0.1$ for the Lorentzian distance. We plot the representations learned with this network when $d = 2$ in Fig. 5. The shape of the space in the Euclidean case is due to the relu activation functions. One can observe that semantically similar superclasses are already close to each other (*e.g.* "Vehicle 1" and "Vehicle 2", or "Large man-made outdoor things" and "Large natural outdoor scenes"). However, the small dimensionality of the space makes it hard to separate the categories.

For the experiment in Fig. 3, we chose $\alpha = 5$, and used batch size $n = 600$ for both chosen distances so that the model takes the superclass information into account more than the class information. Many extracted subtrees are similar between both distances.

However, some differences are the following:

• *Insects* are closer to *flowers* with the Lorentzian distance. This is explained by the fact that bees (known for their role in pollination) are in the insect superclass.

• *Reptiles* (containing aquatic animals such as crocodiles, turtles (and dinosaurs?)) are close to fish and aquatic mammals.

• *Large natural outdoor scenes* and *large man-made outdoor things* are closer to *tree* with the Lorentzian distance.

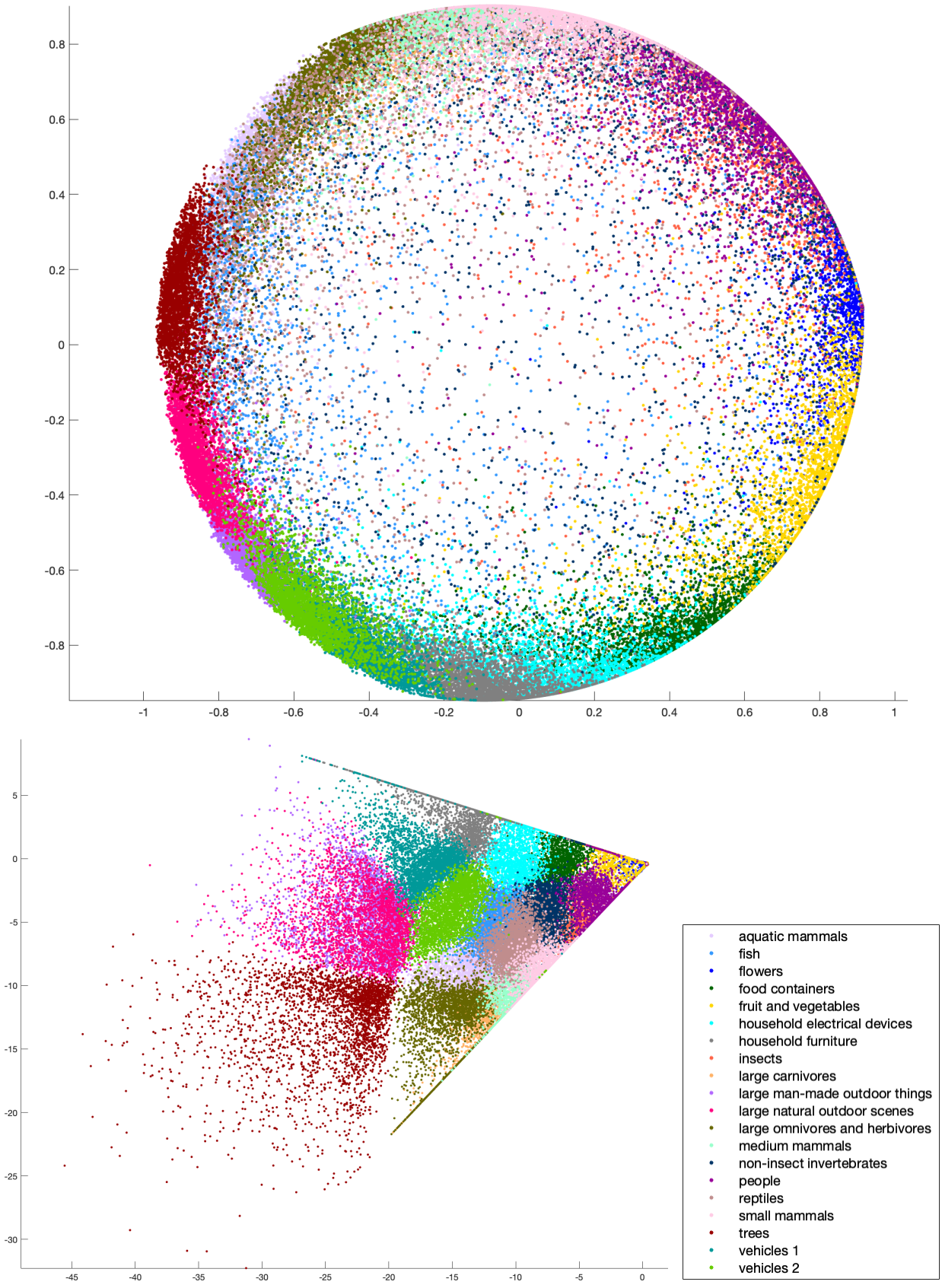
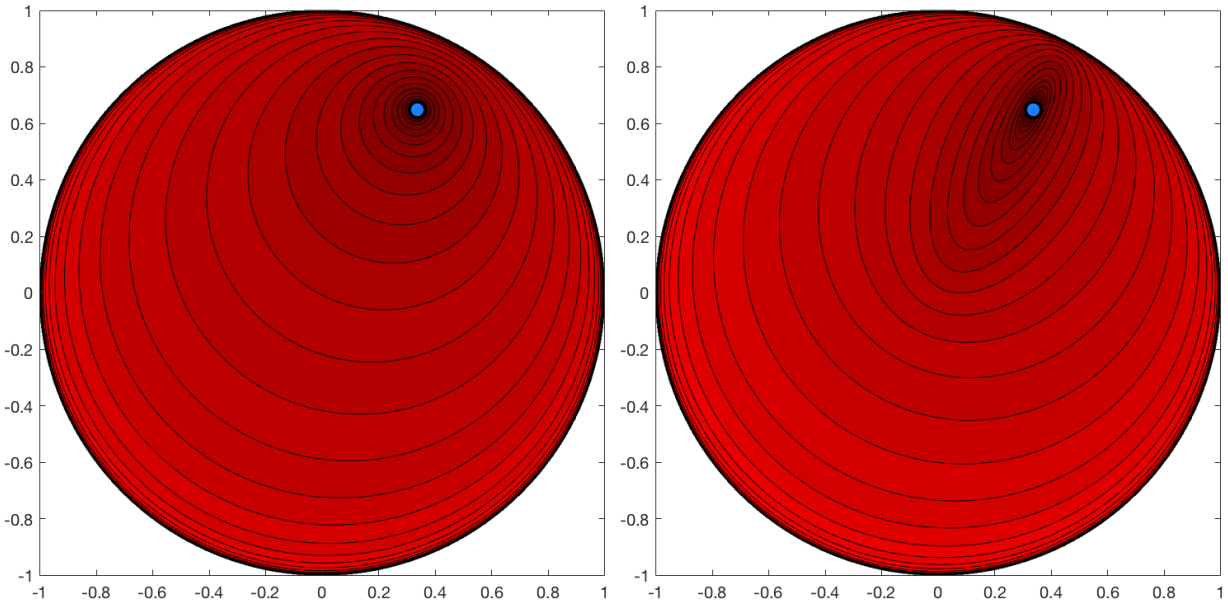
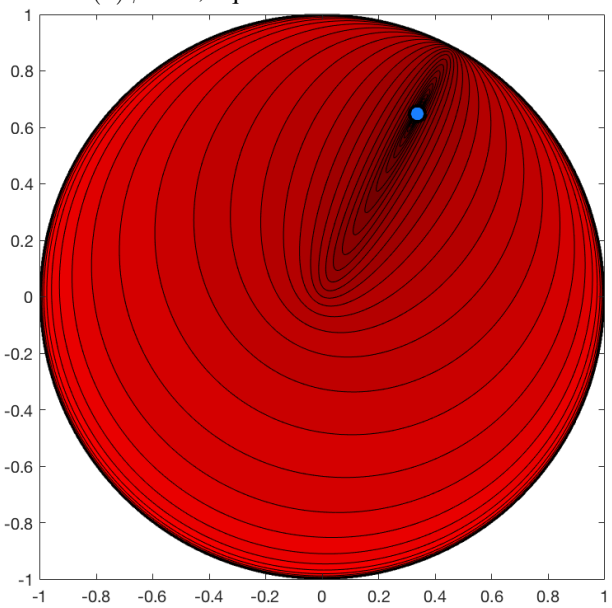


Figure 5. CIFAR-100 representations learned with the neural network when the output dimensionality of the neural network is 2. (top) The chosen distance is the squared Lorentzian distance. (bottom) The chosen distance is the squared Euclidean distance.

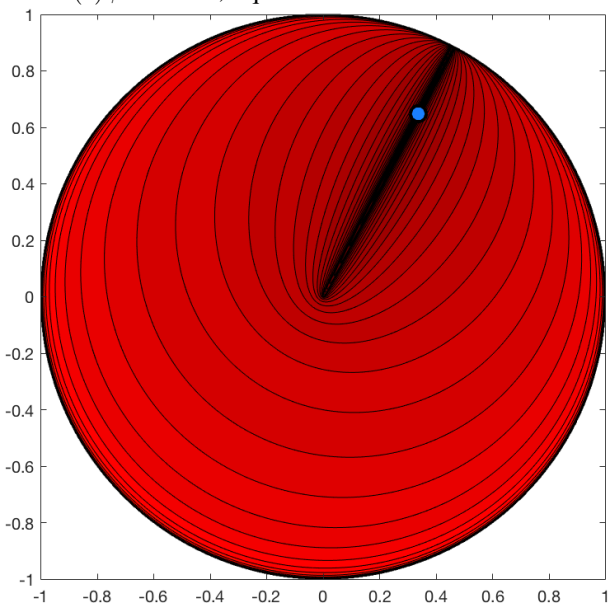


(a) $\beta = 1$, Squared Lorentzian distance

(b) $\beta = 10^{-1}$, Squared Lorentzian distance

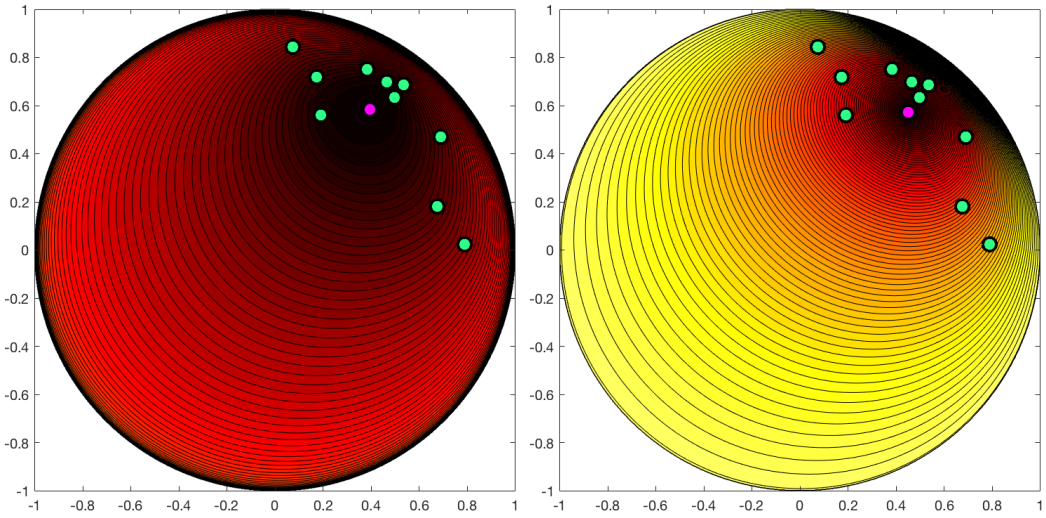


(c) $\beta = 10^{-2}$, Squared Lorentzian distance



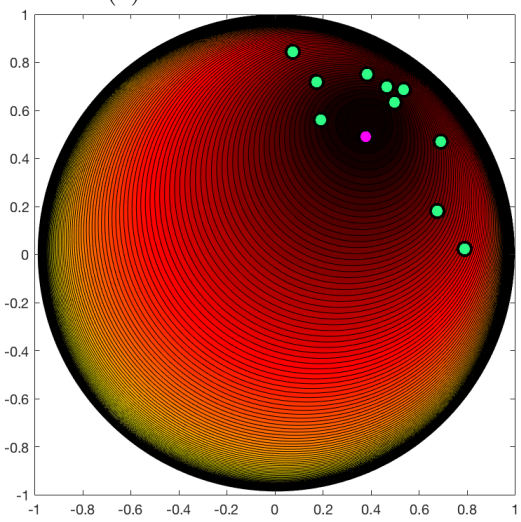
(d) $\beta = 10^{-4}$, Squared Lorentzian distance

Figure 6. Level set of distances to the point in purple depending on the parameter β. The distance tends to be smaller along the radius that contains the point as explained in Section 3.2.

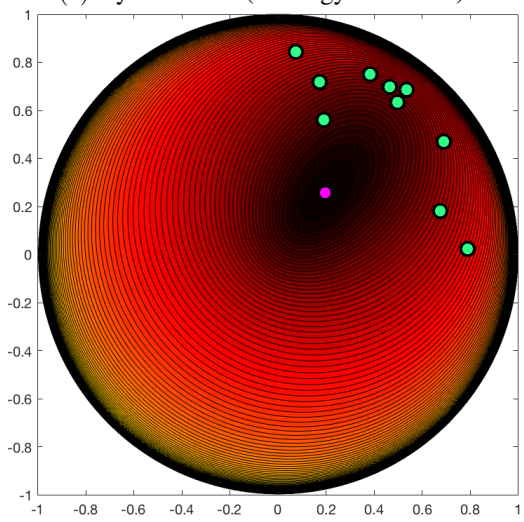


(a) Poincaré distance metric

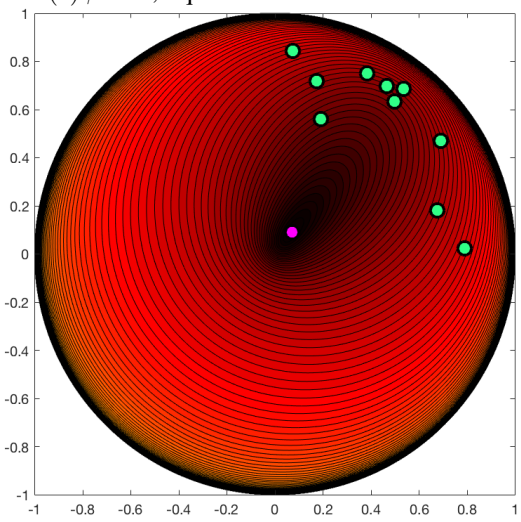
(b) Gyrodistance (to the gyrocentroid)



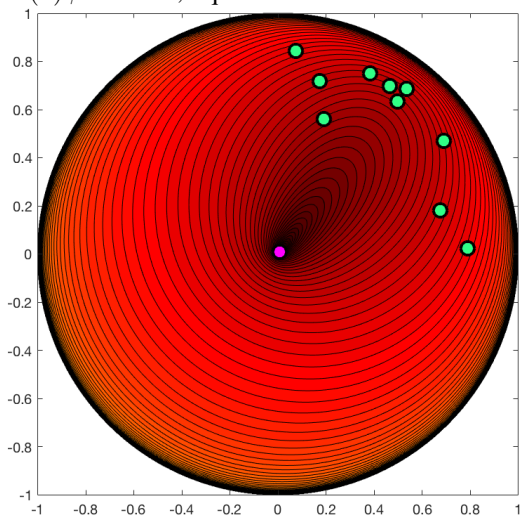
(c) $\beta = 1$, Squared Lorentzian distance



(d) $\beta = 10^{-1}$, Squared Lorentzian distance



(e) $\beta = 10^{-2}$, Squared Lorentzian distance



(f) $\beta = 10^{-4}$, Squared Lorentzian distance

Figure 7. 10 examples are represented in green in a Poincaré ball. Their centroid *w.r.t.* the chosen distance function is in magenta and the level sets represent the sum of the distances between the current point and the training examples. The centroid *w.r.t.* the squared Lorentzian distance can be calculated in closed-form. Smaller values of β induce smaller Euclidean norms of the centroid.