Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions

Boris Muzellec  
CREST, ENSAE  
boris.muzellec@ensae.fr

Marco Cuturi  
Google Brain and CREST, ENSAE  
cuturi@google.com

Abstract

Embedding complex objects as vectors in low dimensional spaces is a longstanding problem in machine learning. We propose in this work an extension of that approach, which consists in embedding objects as elliptical probability distributions, namely distributions whose densities have elliptical level sets. We endow these measures with the 2-Wasserstein metric, with two important benefits: (i) For such measures, the squared 2-Wasserstein metric has a closed form, equal to a weighted sum of the squared Euclidean distance between means and the squared Bures metric between covariance matrices. The latter is a Riemannian metric between positive semi-definite matrices, which turns out to be Euclidean on a suitable factor representation of such matrices, which is valid on the entire geodesic between these matrices. (ii) The 2-Wasserstein distance boils down to the usual Euclidean metric when comparing Diracs, and therefore provides a natural framework to extend point embeddings. We show that for these reasons Wasserstein elliptical embeddings are more intuitive and yield tools that are better behaved numerically than the alternative choice of Gaussian embeddings with the Kullback-Leibler divergence. In particular, and unlike previous work based on the KL geometry, we learn elliptical distributions that are not necessarily diagonal. We demonstrate the advantages of elliptical embeddings by using them for visualization, to compute embeddings of words, and to reflect entailment or hypernymy.

1 Introduction

One of the holy grails of machine learning is to compute meaningful low-dimensional embeddings for high-dimensional complex data. That ability has recently proved crucial to tackle more advanced tasks, such as for instance: inference on texts using word embeddings [Mikolov et al., 2013b, Pennington et al., 2014, Bojanowski et al., 2017], improved image understanding [Norouzi et al., 2014], representations for nodes in large graphs [Grover and Leskovec, 2016].

Such embeddings have been traditionally recovered by seeking isometric embeddings in lower dimensional Euclidean spaces, as studied in [Johnson and Lindenstrauss, 1984, Bourgain, 1985]. Given $n$ input points $x_1, \ldots, x_n$, one seeks as many embeddings $y_1, \ldots, y_n$ in a target space $\mathcal{Y} = \mathbb{R}^d$ whose pairwise distances $\|y_i - y_j\|_2$ do not depart too much from the original distances $d_X(x_i, x_j)$ in the input space. Note that when $d$ is restricted to be 2 or 3, these embeddings $(y_i)$ provide a useful way to visualize the entire dataset. Starting with metric multidimensional scaling (mMDS) [De Leeuw, 1977, Borg and Groenen, 2005], several approaches have refined this intuition [Tenenbaum et al., 2000, Roweis and Saul, 2000, Hinton and Roweis, 2003, Maaten and Hinton, 2008]. More general criteria, such as reconstruction error [Hinton and Salakhutdinov, 2006, Kingma and Welling, 2014]; co-occurrence [Globerson et al., 2007]; or relational knowledge, be it in metric learning [Weinberger and Saul, 2009] or between words [Mikolov et al., 2013b] can be used to obtain vector embeddings. In such cases, distances $\|y_i - y_j\|_2$ between embeddings, or alternatively their dot-products $\langle y_i, y_j \rangle$
must comply with sophisticated desiderata. Naturally, more general and flexible approaches in which the embedding space $\mathcal{Y}$ needs not be Euclidean can be considered, for instance in generalized MDS on the sphere [Maron et al., 2010], on surfaces [Bronstein et al., 2006], in spaces of trees [Bădouin et al., 2007, Fakcharoenphol et al., 2003] or, more recently, computed in the Poincaré hyperbolic space [Nickel and Kiela, 2017].

**Probabilistic Embeddings.** Our work belongs to a recent trend, pioneered by Vilnis and McCallum, who proposed to embed data points as *probability measures* in $\mathbb{R}^d$ [2015], and therefore generalize point embeddings. Indeed, point embeddings can be regarded as a very particular—and degenerate—case of probabilistic embedding, in which the uncertainty is infinitely concentrated on a single point (a Dirac). Probability measures can be more spread-out, or event multimodal, and provide therefore an opportunity for additional flexibility. Naturally, such an opportunity can only be exploited by defining a metric, divergence or dot-product on the space (or a subspace thereof) of probability measures. Vilnis and McCaullum proposed to embed words as *Gaussians* endowed either with the Kullback-Leibler (KL) divergence or the expected likelihood kernel [Jebara et al., 2004]. The Kullback-Leibler and expected likelihood kernel on measures have, however, an important drawback: these geometries do not coincide with the usual Euclidean metric between point embeddings when the variances of these Gaussians collapse. Indeed, the KL divergence and the $\ell_2$ distance between two Gaussians diverges to $\infty$ or saturates when the variances of these Gaussians become small. To avoid numerical instabilities arising from this degeneracy, Vilnis and McCaullum must restrict their work to diagonal covariance matrices.

**Contributions.** We propose in this work a new framework for probabilistic embeddings, in which point embeddings are seamlessly handled as a particular case. We consider arbitrary families of elliptical distributions, which subsume Gaussians, and also include uniform elliptical distributions, which are arguably easier to visualize because of their compact support. Our approach uses the 2-Wasserstein distance to compare elliptical distributions. The latter can handle degenerate measures, and both its value and its gradients admit closed forms [Gelbrich, 1990], either in their natural Riemannian formulation, as well as in a more amenable local Euclidean parameterization. We provide numerical tools to carry out the computation of elliptical embeddings in different scenarios, both to optimize them with respect to metric requirements (as is done in multidimensional scaling) or with respect to dot-products (as shown in our applications to word embeddings for entailment, similarity and hypernymy tasks) for which we introduce a proxy using a polarization identity.

**Notations** $S^d_{++}$ (resp. $S^d_+$) is the set of positive (resp. semi-)definite $d \times d$ matrices. For two vectors $x, y \in \mathbb{R}^d$ and a matrix $M \in S^d$, we write the Mahalanobis norm induced by $M$ as $\|x - c\|_M^2 = (x - c)^T M (x - c)$ and $|M|$ for $\text{det}(M)$. For $V$ an affine subspace of dimension $m$ of $\mathbb{R}^d$, $\lambda_V$ is the Lebesgue measure on that subspace. $M^T$ is the pseudo inverse of $M$.

## 2 The Geometry of Elliptical Distributions in the Wasserstein Space

We recall in this section basic facts about elliptical distributions in $\mathbb{R}^d$. We adopt a general formulation that can handle measures supported on subspaces of $\mathbb{R}^d$ as well as Dirac (point) measures. That level of generality is needed to provide a seamless connection with usual vector embeddings, seen in the context of this paper as Dirac masses. We recall results from the literature showing that the squared 2-Wasserstein distance between two distributions from the same family of elliptical distributions is equal to the squared Euclidean distance between their means plus the squared Bures metric between their scale parameter scaled by a suitable constant.

**Elliptically Contoured Densities.** In their simplest form, elliptical distributions can be seen as generalizations of Gaussian multivariate densities in $\mathbb{R}^d$: their level sets describe concentric ellipsoids, shaped following a scale parameter $C \in S^d_{++}$, and centered around a mean parameter $c \in \mathbb{R}^d$ [Cambanis et al., 1981]. The density at a point $x$ of such distributions is $f(||x - c||_{C^{-1}})/\sqrt{|C|}$ where the generator function $f$ is such that $\int_{\mathbb{R}^d} f(||x||^2)dx = 1$. Gaussians are recovered with $f = g, g(\cdot) \propto e^{-\frac{1}{2}}$ while uniform distributions on full rank ellipsoids result from $f = u, u(\cdot) \propto 1$. Because the norm induced by $C^{-1}$ appears in formulas above, the scale parameter $C$ must have full rank for these definitions to be meaningful. Cases where $C$ does not have full rank can however
appear when a probability measure is supported on an affine subspace\(^1\) of \(\mathbb{R}^d\), such as lines in \(\mathbb{R}^2\), or even possibly a space of null dimension when the measure is supported on a single point (a Dirac measure), in which case its scale parameter \(C\) is 0. We provide in what follows a more general approach to handle these degenerate cases.

**Elliptical Distributions.** To lift this limitation, several reformulations of elliptical distributions have been proposed to handle degenerate scale matrices \(C\) of rank \(rk\ C < d\). Gelbrich [1990, Theorem 2.4] defines elliptical distributions as measures with a density w.r.t. the Lebesgue measure of dimension \(rk\ C\), in the affine space \(c + \text{Im} \ C\), where the image of \(C\) is \(\text{Im} \ C \defeq \{Cx, x \in \mathbb{R}^d\}\). This approach is intuitive, in that it reduces to describing densities in their relevant subspace. A more elegant approach uses the parameterization provided by characteristic functions [Cambanis et al., 1981, Fang et al., 1990]. In a nutshell, recall that the characteristic function of a multivariate Gaussian is equal to \(\phi(t) = e^{it^TCt}\) where, as in the paragraph above, \(\varrho() = e^{-|t|^2}/2\). A natural generalization to consider other elliptical distributions is therefore to consider for \(g\) other functions \(h\) of positive type [Ushakov, 1999, Theo.1.8.9], such as the indicator function \(u\) above, and still apply them to the same argument \(t^TCt\). Such functions are called characteristic generators and fully determine, along with a mean \(c\) and a scale parameter \(C\), an elliptical measure. This parameterization does not require the scale parameter \(C\) to be invertible, and therefore allows to define probability distributions that do not have necessarily a density w.r.t. to the Lebesgue measure in \(\mathbb{R}^d\). Both constructions are relatively complex, and we refer the interested reader to these references for a rigorous treatment.

**Rank Deficient Elliptical Distributions and their Variances.** For the purpose of this work, we will only require the following result: the variance of an elliptical measure is equal to its scale parameter \(C\) multiplied by a scalar that only depends on its characteristic generator. Indeed, given a mean vector \(\mu\) and a scale parameter \(C\) to be the measure with characteristic function \(t \mapsto e^{it^TCt}\). In that case, one can show that the covariance matrix of \(\mu_h, C\) is equal to its scale parameter \(C\) times a constant \(\tau_\varrho\) that only depends on \(h\), namely

\[
\text{var}(\mu_h, C) = \tau_\varrho C.
\]

For Gaussians, the scale parameter \(C\) and its covariance matrix coincide, that is \(\tau_\varrho = 1\). For uniform elliptical distributions, one has \(\tau_\varrho = 1/(d + 2)\): the covariance of a uniform distribution on the volume \(\{c + Cx, x \in \mathbb{R}^d, \|x\| = 1\}\), such as those represented in Figure 1, is equal to \(C/(d + 2)\).

The 2-Wasserstein Bures Metric A natural metric for elliptical distributions arises from optimal transport (OT) theory. We refer interested readers to [Santambrogio, 2015, Peyré and Cuturi, 2018] for exhaustive surveys on OT. Recall that for two arbitrary probability measures \(\mu, \nu \in \mathcal{P}(\mathbb{R}^d)\), their squared 2-Wasserstein distance is equal to

\[
W^2_2(\mu, \nu) \defeq \inf_{X \sim \mu, Y \sim \nu} \mathbb{E}\| X - Y \|_2^2.
\]

This formula rarely has a closed form. However, in the footsteps of Dowson and Landau [1982] who proved it for Gaussians, Gelbrich [1990] showed that for \(\alpha \defeq \mu_{h, a, A}\) and \(\beta \defeq \mu_{h, b, B}\) in the same family \(\mathcal{P}_h = \{\mu_{h, c, C}, c \in \mathbb{R}^d, C \in S^d_+\}\), one has

\[
W^2_2(\alpha, \beta) = \|a - b\|^2_2 + 2\mathbb{B}^2(\text{var} \alpha, \text{var} \beta) = \|a - b\|^2_2 + \tau_h \mathbb{B}^2(A, B) ,
\]

\(^1\)For instance, the random variable \(Y\) in \(\mathbb{R}^d\) obtained by duplicating the same normal random variable \(X\) in \(\mathbb{R}\), \(Y = [X, X]\), is supported on a line in \(\mathbb{R}^2\) and has no density w.r.t. the Lebesgue measure in \(\mathbb{R}^2\).
where $\mathcal{B}^2$ is the (squared) Bures metric on $S^d_{+}$, proposed in quantum information geometry [1969] and studied recently in [Bhatia et al., 2018, Malagò et al., 2018],

$$\mathcal{B}^2(X, Y) \overset{\text{def}}{=} \text{Tr}(X + Y - 2(X^{\frac{1}{2}}YX^{\frac{1}{2}})^{\frac{1}{2}}).$$

The factor $\tau_h$ next to the rightmost term $\mathcal{B}^2$ in (2) arises from homogeneity of $\mathcal{B}^2$ in its arguments (3), which is leveraged using the identity in (1).

A few remarks (i) When both scale matrices $A = \text{diag} \, d_A$ and $B = \text{diag} \, d_B$ are diagonal, $W_2^2(\alpha, \beta)$ is the sum of two terms: the usual squared Euclidean distance between their means, plus $\tau_h$ times the squared Hellinger metric between the diagonals $d_A, d_B$: $J^2(d_A, d_B) \overset{\text{def}}{=} \|\sqrt{d_A} - \sqrt{d_B}\|^2$.

(ii) The distance $W_2$ between two Diracs $\delta_a, \delta_b$ is equal to the usual distance between vectors $\|a - b\|_2$. (iii) The squared distance $W_2^2$ between a Dirac $\delta_a$ and a measure $\mu_{h,b,B}$ in $P_h$ reduces to $\|a - b\|^2 + \tau_h \text{Tr}B$. The distance between a point and an ellipsoid distribution therefore always increases as the scale parameter of the latter increases. Although this point makes sense from the quadratic viewpoint of $W_2^2$ (in which the quadratic contribution $\|a - x\|_2^2$ of points $x$ in the ellipsoid that stand further away from $a$ than $b$ will dominate that brought by points $x$ that are closer, see Figure 3) this may be counterintuitive for applications to visualization, an issue that will be addressed in Section 4. (iv) The $W_2$ distance between two elliptical distributions in the same family $P_h$ is always finite, no matter how degenerate they are. This is illustrated in Figure 1 in which a uniform measure $\mu_{a,A}$ is shown to be exactly equidistant to four other uniform elliptical measures, some of which are degenerate. However, as can be hinted by the simple example of the Hellinger metric, that distance may not be differentiable for degenerate measures (in the same sense that $(\sqrt{x} - \sqrt{y})^2$ is defined at $x = 0$ but not differentiable w.r.t $x$). (v) Although we focus in this paper on uniform elliptical distributions, not only because they are easier to plot and visualize, but also because the other elliptical family simply amounts to changing the constant $\tau_h$ next to the Bures metric in (2). Alternatively, increasing (or tuning) that parameter $\tau_h$ simply amounts to considering elliptical distributions with increasingly heavier tails.

3 Optimizing over the Space of Elliptical Embeddings

Our goal in this paper is to use the set of elliptical distributions endowed with the $W_2$ distance as an embedding space. To optimize objective functions involving $W_2$ terms, we study in this section several parameterizations of the parameters of elliptical distributions. Location parameters only appear in the computation of $W_2$ through their Euclidean metric, and offer therefore no particular challenge. Scale parameters are more tricky to handle since they are constrained to lie in $S^d_{+}$. Rather than keeping track of scale parameters, we advocate optimizing directly on factors (square roots) of such parameters, which results in simple Euclidean (unconstrained) updates reviewed below.

Geodesics for Elliptical Distributions When $A$ and $B$ have full rank, the geodesic from $\alpha$ to $\beta$ is a curve of measures in the same family of elliptic distributions, characterized by location and scale parameters $c(t), C(t)$, where

$$c(t) = (1 - t)a + tb; \quad C(t) = ((1 - t)I + tT^{AB}) A ((1 - t)I + tT^{AB})^T,$$

and where the matrix $T^{AB}$ is such that $x \rightarrow T^{AB}(x - a) + b$ is the so-called Brenier optimal transportation map [1987] from $\alpha$ to $\beta$, given in closed form as,

$$T^{AB} \overset{\text{def}}{=} A^{-\frac{1}{2}} (A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}} A^{-\frac{1}{2}},$$

and is the unique matrix such that $B = T^{AB} A T^{AB}$ [Peyré and Cuturi, 2018, Remark 2.30]. When $A$ is degenerate, such a curve still exists as long as $\text{Im} \, B \subset \text{Im} \, A$, in which case the expression above is still valid using pseudo-inverse square roots $A^{1/2}$ in place of the usual inverse square-root.

Differentiability in Riemannian Parameterization Scale parameters are restricted to lie on the cone $S^d_{+}$. For such problems, it is well known that a direct gradient-and-project based optimization on scale parameters would prove too expensive. A natural remedy to this issue is to perform manifold optimization [Absil et al., 2009]. Indeed, as in any Riemannian manifold, the Riemannian gradient $\frac{1}{2}d^2(x, y)$ is given by $-\log_y x$ [Lee, 1997]. Using the expressions of the $\exp$ and $\log$ given in [Malagò et al., 2018], we can show that minimizing $\frac{1}{2}\mathcal{B}^2(A, B)$ using Riemannian gradient descent corresponds to making updates of the form, with step length $\eta$

$$A' = ((1 - \eta)I + \eta T^{AB}) A ((1 - \eta)I + \eta T^{AB})^T.$$
When $0 \leq \eta \leq 1$, this corresponds to considering a new point $A'$ closer to $B$ along the Bures geodesic between $A$ and $B$. When $\eta$ is negative or larger than 1, $A'$ no longer lies on this geodesic but is guaranteed to remain PSD, as can be seen from (6). Figure 2 shows a $W_2$ geodesic between two measures $\mu_0$ and $\mu_1$, as well as its extrapolation following exactly the formula given in (4). That figure illustrates that $\mu_t$ is not necessarily geodesic outside of the boundaries $[0, 1]$ w.r.t. three relevant measures, because its metric derivative is smaller than 1 [Ambrosio et al., 2006, Theorem 1.1.2]. When negative steps are taken (for instance when the $W_2^2$ distance needs to be increased), this lack of geodesicity has proved difficult to handle numerically for a simple reason: such updates may lead to degenerate scale parameters $A'$, as illustrated around time $t = 1.5$ of the curve in Figure 2. Another obvious drawback of Riemannian approaches is that they are not as well studied as simpler non-constrained Euclidean problems, for which a plethora of optimization techniques are available. This observations motivates an alternative Euclidean parameterization, detailed in the next paragraph.

\[ \nabla_{L} \frac{1}{2} \mathcal{B}^2(A, B) = (I - T_{AB}) L, \quad \text{with updates} \quad L' = ((1 - \eta)I + \eta T_{AB}) L. \quad (7) \]

**Links between Euclidean and Riemannian Parameterization**  

A canonical way to handle a PSD constraint for $A$ is to rewrite it in factor form $A = LL^T$. In the particular case of the Bures metric, we show that this simple parametrization comes without losing the geometric interest of manifold optimization, while benefitting from simpler additive updates. Indeed, one can (see supplementary material) that the gradient of the squared Bures metric has the following gradient:

\[ \nabla_{L} \frac{1}{2} \mathcal{B}^2(A, B) = (I - T_{AB}) L, \quad \text{with updates} \quad L' = ((1 - \eta)I + \eta T_{AB}) L. \quad (7) \]

When considering a general loss function $\mathcal{L}$ that take as arguments squared Bures distances, one can also show that $\mathcal{L}$ is geodesically convex w.r.t. to scale matrices $A$ if and only if it is convex in the usual sense with respect to $L$, where $A = LL^T$. Write now $L_B = T_{AB}L$. One can recover that $L_B L_B^T = B$. Therefore, expanding the expression $\mathcal{B}^2$ for the right term below we obtain

\[ \mathcal{B}^2(A, B) = \mathcal{B}^2(LL^T, L_B L_B^T) = \mathcal{B}^2(LL^T, T_{AB}L(T_{AB}L)^T) = \|L - T_{AB}L\|^2_F. \]

Indeed, the Bures distance simply reduces to the Frobenius distance between two factors of $A$ and $B$. However these factors need to be carefully chosen: given $L$ for $A$, the factor for $B$ must be computed according to an optimal transport map $T_{AB}$.

**Polarization between Elliptical Distributions**  

Some of the applications we consider, such as the estimation of word embeddings, are inherently based on dot-products. By analogy with the polarization identity, $\langle x, y \rangle = (\|x - 0\|^2 + \|y - 0\|^2 - \|x - y\|^2)/2$, we define a Wasserstein-Bures pseudo-dot-product, where $\delta_0 = \mu_{a_{x,y}}$ is the Dirac mass at $0$,

\[
[\mu_{a, A} : \mu_{b, B}] \overset{\text{def}}{=} \frac{1}{2} \left( W_2^2(\mu_{a, A}, \delta_0) + W_2^2(\mu_{b, B}, \delta_0) - W_2^2(\mu_{a, A}, \mu_{b, B}) \right) = \langle a, b \rangle + \text{Tr} \left( A^2 B A^2 B \right)^{1/2}.
\]
Note that $\langle \cdot , \cdot \rangle$ is not an actual inner product since the Bures metric is not Hilbertian, unless we restrict ourselves to diagonal covariance matrices, in which case it is the inner product between $(\mathbf{a}, \sqrt{d_\mathbf{A}})$ and $(\mathbf{b}, \sqrt{d_\mathbf{B}})$. We use $[\mu_{\mathbf{A}} : \mu_{\mathbf{B}}]$ as a similarity measure which has, however, some regularity: one can show that when $\mathbf{a}, \mathbf{b}$ are constrained to have equal norms and $\mathbf{A}$ and $\mathbf{B}$ equal traces, then $[\mu_{\mathbf{A}} : \mu_{\mathbf{B}}]$ is maximal when $\mathbf{a} = \mathbf{b}$ and $\mathbf{A} = \mathbf{B}$. Differentiating all three terms in that sum, the gradient of this pseudo dot-product w.r.t $\mathbf{A}$ reduces to $\nabla_{\mathbf{A}} [\mu_{\mathbf{A}} : \mu_{\mathbf{B}}] = \mathbf{T}^{\mathbf{AB}}$.

**Computational Aspects** The computational bottleneck of gradient-based Bures optimization lies in the matrix square roots and inverse square roots operations that arise when instantiating transport maps $\mathbf{T}$ as in (5). A naive method using eigenvector decomposition is far too time-consuming, and there is not yet, to the best of our knowledge, a straightforward way to perform it in batches on a GPU. We propose to use Newton-Schulz iterations (Algorithm 1, see [Higham, 2008, Ch. 6]) to approximate these root computations. These iterations producing both a root and an inverse root approximation, and relying exclusively on matrix-matrix multiplications, stream efficiently on GPUs.

Another problem lies in the fact that numerous roots and inverse-roots are required to form map $\mathbf{T}$. To solve this, we exploit an alternative formula for $\mathbf{T}^{\mathbf{AB}}$ (proof in the supplementary material):

$$\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}} = \mathbf{B}^{\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{B}^{-\frac{1}{2}}. \tag{8}$$

In a gradient update, both the loss and the gradient of the metric are needed. In our case, we can use the matrix roots computed during loss evaluation and leverage the identity above to compute on a budget the gradients with respect to either scale matrices $\mathbf{A}$ and $\mathbf{B}$. Indeed, a naive computation of $\nabla_{\mathbf{A}} \mathcal{W}_2^2(\mathbf{A}, \mathbf{B})$ and $\nabla_{\mathbf{B}} \mathcal{W}_2^2(\mathbf{A}, \mathbf{B})$ would require the knowledge of 6 roots:

$$\mathbf{A}^{-\frac{1}{2}}, \mathbf{B}^{-\frac{1}{2}}, \left(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}\right)^{\frac{1}{2}}, \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}}\right)^{\frac{1}{2}}, \mathbf{A}^{-\frac{1}{2}}, \text{ and } \mathbf{B}^{-\frac{1}{2}}$$

to compute the following transport maps

$$\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}, \mathbf{T}^{\mathbf{BA}} = \mathbf{B}^{-\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{B}^{-\frac{1}{2}},$$

namely four matrix roots and two matrix inverse roots. We can avoid computing those six matrices using identity (8) and limit ourselves to two runs of Algorithm 1, to obtain the same quantities as

$$\{Y_1 \overset{\text{def}}{=} \mathbf{A}^{\frac{1}{2}}, Z_1 \overset{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}, \{Y_2 \overset{\text{def}}{=} \left(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}\right)^{\frac{1}{2}}, Z_2 \overset{\text{def}}{=} \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}}\right)^{\frac{1}{2}}\}$$

$$\mathbf{T}^{\mathbf{AB}} = Z_1 Y_2 Z_1, \mathbf{T}^{\mathbf{BA}} = Y_1 Z_2 Y_1.$$

When computing the gradients of $n \times m$ squared Wasserstein distances $\mathcal{W}_2^2(\alpha_i, \beta_j)$ in parallel, one only needs to run $n$ Newton-Schulz algorithms (in parallel) to compute matrices $(Y_i, Z_i)_{i \leq n}$, and then $n \times m$ Newton-Schulz algorithms to recover cross matrices $Y_{ij}^{\mathcal{W}_2^2}, Z_{ij}^{\mathcal{W}_2^2}$. On the other hand, using an automatic differentation framework would require an additional backward computation of the same complexity as the forward pass evaluating computation of the roots and inverse roots, hence requiring roughly twice as many operations per batch.

**Avoiding Rank Deficiency at Optimization Time** Although $\mathcal{W}_2^2(\mathbf{A}, \mathbf{B})$ is defined for rank deficient matrices $\mathbf{A}$ and $\mathbf{B}$, it is not differentiable with respect to these matrices if they are rank deficient. Indeed, as mentioned earlier, this can be compared to the non-differentiability of the Hellinger metric, $(\sqrt{x} - \sqrt{y})^2$ when $x$ or $y$ becomes 0, at which point which becomes not differentiable. If $\text{Im} \mathbf{B} \not\subset \text{Im} \mathbf{A}$, which is notably the case if $\text{rk} \mathbf{B} > \text{rk} \mathbf{A}$, then $\nabla_{\mathbf{A}} \mathcal{W}_2^2(\mathbf{A}, \mathbf{B})$ no longer exists. However, even in that case, $\nabla_{\mathbf{B}} \mathcal{W}_2^2(\mathbf{A}, \mathbf{B})$ exists iff $\text{Im} \mathbf{A} \subset \text{Im} \mathbf{B}$. Since it would be cumbersome to account for these subtleties in a large scale optimization setting, we propose to add a small common regularization term to all the factor products considered for our embeddings, and set $\mathbf{A}_\epsilon = \mathbf{L} \mathbf{L}^T + \epsilon \mathbf{I}$ were $\epsilon > 0$ is a hyperparameter. This ensures that all matrices are full rank, and thus that all gradients exist. Most importantly, all our derivations still hold with this regularization, and can be shown to leave the method to compute the gradients w.r.t $\mathbf{L}$ unchanged, namely remain equal to $(\mathbf{I} - \mathbf{T}^{\mathbf{A}, \mathbf{B}}) \mathbf{L}$.

---

**Algorithm 1 Newton-Schulz**

**Input:** PSD matrix $\mathbf{A}$, $\epsilon > 0$

\[ \mathbf{Y} \leftarrow \mathbf{A} / \| \mathbf{A} + \epsilon \mathbf{I} \| \mathbf{I}, \mathbf{Z} \leftarrow \mathbf{I} \]

while not converged do

\[ \mathbf{T} \leftarrow (3 \mathbf{I} - \mathbf{Z} \mathbf{Y}) / 2, \mathbf{Y} \leftarrow \mathbf{Y} \mathbf{T}, \mathbf{Z} \leftarrow \mathbf{T} \mathbf{Z} \]

end while

\[ \mathbf{Y} \leftarrow \sqrt{(1 + \epsilon) \| \mathbf{A} \| \mathbf{Y}}, \mathbf{Z} \leftarrow \frac{\mathbf{Z}}{\sqrt{(1 + \epsilon) \| \mathbf{A} \|}} \]

**Output:** square root $\mathbf{Y}$, inverse square root $\mathbf{Z}$
4 Experiments

We discuss in this section several applications of elliptical embeddings. We first consider a simple mMDS type visualization task, in which elliptical distributions in $d = 2$ are used to embed isometrically points in high dimension. We argue that for such purposes, a more natural way to visualize ellipses is to use their precision matrices. This is due to the fact that the human eye somewhat acts in the opposite direction to the Bures metric, as discussed in Figure 3. We follow with more advanced experiments in which we consider the task of computing word embeddings on large corpora as a testing ground, and equal or improve on the state-of-the-art.

Figure 3: (left) three points on the plane. (middle) isometric elliptic embedding with the Bures metric: ellipses of a given color have the same respective distances as points on the left. Although the mechanics of optimal transport indicate that the blue ellipsoid is far from the two others, in agreement with the left plot, the human eye tends to focus on those areas that overlap (below the ellipsoid center) rather than those far away areas (north-east area) that contribute more significantly to the $W_2$ distance. (right) the precision matrix visualization, obtained by considering ellipses with the same axes but inverted eigenvalues, agree better with intuition, since they emphasize that overlap and extension of the ellipse means on the contrary that those axis contribute less to the increase of the metric.

Figure 4: Toy experiment: visualization of a dataset of 10 PISA scores for 35 countries in the OECD. (left) MDS embeddings of these countries on the plane (right) elliptical embeddings on the plane using the precision visualization discussed in Figure 3. The normalized stress with standard MDS is 0.62. The stress with elliptical embeddings is close to $5e-3$ after 1000 gradient iterations, with random initializations for scale matrices (following a Standard Wishart with 4 degrees of freedom) and initial means located on the MDS solution.

Visualizing Datasets Using Ellipsoids Multidimensional scaling [De Leeuw, 1977] aims at embedding points $x_1, \ldots, x_n$ in a finite metric space in a lower dimensional one by minimizing the stress $\sum_{i,j} (\|x_i - x_j\| - \|y_i - y_j\|)^2$. In our case, this translates to the minimization of $L_{\text{MDS}}(a_1, \ldots, a_n, A_1, \ldots, A_n) = \sum_{i,j} (\|x_i - x_j\| - W_2(\mu_{a_i}, A_i, \mu_{a_j}, A_j))^2$. This objective can be crudely minimized with a simple gradient descent approach operating on factors as advocated in Section 3, as illustrated in a toy example carried out using data from OECD’s PISA study.

Word Embeddings The skipgram model [Mikolov et al., 2013a] computes word embeddings in a vector space by maximizing the log-probability of observing surrounding context words given an input central word. Vilnis and McCallum [2015] extended this approach to diagonal Gaussian embeddings using an energy whose overall principles we adopt here, adapted to elliptical distributions with full covariance matrices in the 2-Wasserstein space. For every word $w$, we consider an input (as a word) and an output (as a context) representation as an elliptical measure, denoted respectively $\mu_w$ and $\nu_w$, both parameterized by a location vector and a scale parameter (stored in factor form).

http://pisadataexplorer.oecd.org/ide/idepisa/
Given a set $\mathcal{R}$ of positive word/context pairs of words $(w, c)$, and for each input word a set $N(w)$ of $n$ negative context words sampled randomly, we adapt Vilnis and McCallum’s loss function to the $\mathcal{W}_{2}^{2}$ distance to minimize the following hinge loss:

$$\sum_{(w, c) \in \mathcal{R}} \left[ M - [\mu_{w} : \nu_{c}] + \frac{1}{n} \sum_{c' \in N(w)} [\mu_{w} : \nu_{c'}] \right]^{+},$$

where $M > 0$ is a margin parameter. We train our embeddings on the concatenated ukWaC and WaCkypedia corpora [Baroni et al., 2009], consisting of about 3 billion tokens, on which we keep only the tokens appearing more than 100 times in the text (for a total number of 261583 different words). We train our embeddings using adagrad [Duchi et al., 2011], sampling one negative context per positive context and, in order to prevent the norms of the embeddings to be too highly correlated with the corresponding word frequencies (see Figure in supplementary material), we use two distinct sets of embeddings for the input and context words.

We compare our full elliptical to diagonal Gaussian embeddings trained using the methods described in [Vilnis and McCallum, 2015] on a collection of similarity datasets by computing the Spearman rank correlation between the similarity scores provided in the data and the scores we compute based on our embeddings. Note that these results are obtained using context ($\nu_{w}$) rather than input ($\mu_{w}$) embeddings. For a fair comparison across methods, we set dimensions by ensuring that the number of free parameters remains the same: because of the symmetry in the covariance matrix, elliptical embeddings in dimension $d$ have $d + d(d + 1)/2$ free parameters ($d$ for the means, $d(d + 1)/2$ for the covariance matrices), as compared with $2d$ for diagonal Gaussians. For elliptical embeddings, we use the common practice of using some form of normalized quantity (a cosine) rather than the direct dot product. We implement this here by computing the mean of two cosine terms, each corresponding separately to mean and covariance contributions:

$$\mathbb{S}_{\mathbb{G}}[\mu_{a, A}, \mu_{b, B}] := \frac{\langle a, b \rangle}{\|a\| \|b\|} + \frac{\text{Tr} (A^{1/2}BA^{1/2})^{1/2}}{\sqrt{\text{Tr}A \text{Tr}B}}$$

Using this similarity measure rather than the Wasserstein-Bures dot product is motivated by the fact that the norms of the embeddings show some dependency with word frequencies (see figures in supplementary) and become dominant when comparing words with different frequencies scales. An alternative could have been obtained by normalizing the Wasserstein-Bures dot product in a more standard way that pools together means and covariances. However, as discussed in the supplementary material, this choice makes it harder to deal with the variations in scale of the means and covariances, therefore decreasing performance.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>W2G/45/C</th>
<th>Ell/12/CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimLex</td>
<td>25.09</td>
<td>24.09</td>
</tr>
<tr>
<td>WordSim</td>
<td>53.45</td>
<td>66.02</td>
</tr>
<tr>
<td>WordSim-R</td>
<td>61.70</td>
<td>71.07</td>
</tr>
<tr>
<td>WordSim-S</td>
<td>48.99</td>
<td>60.58</td>
</tr>
<tr>
<td>MEN</td>
<td>65.16</td>
<td>65.58</td>
</tr>
<tr>
<td>MC</td>
<td>59.48</td>
<td>65.95</td>
</tr>
<tr>
<td>RG</td>
<td>69.77</td>
<td>65.58</td>
</tr>
<tr>
<td>YP</td>
<td>37.18</td>
<td>25.14</td>
</tr>
<tr>
<td>MT-287</td>
<td>61.72</td>
<td>59.53</td>
</tr>
<tr>
<td>MT-771</td>
<td>57.63</td>
<td>56.78</td>
</tr>
<tr>
<td>RW</td>
<td>40.14</td>
<td>29.04</td>
</tr>
</tbody>
</table>
We also evaluate our embeddings on the Entailment dataset ([Baroni et al., 2012]), on which we obtain results roughly comparable to those of [Vilnis and McCallum, 2015]. Note that contrary to the similarity experiments, in this framework using the (unsymmetrical) KL divergence makes sense and possibly gives an advantage, as it is possible to choose the order of the arguments in the KL divergence between the entailment and entailed words.

**Hypernymy** In this experiment, we use the framework of [Nickel and Kiela, 2017] on hypernymy relationships to test our embeddings. A word A is said to be a hypernym of a word B if any B is a type of A, e.g. any dog is a type of mammal, thus constituting a tree-like structure on nouns. The WORDNET dataset [Miller, 1995] features a transitive closure of 743,241 hypernymy relations on 82,115 distinct nouns, which we consider as an undirected graph of relations R. Similarly to the skipgram model, for each noun u we sample a fixed number n of negative examples and store them in set N(u) to optimize the following loss: 

\[
\sum_{(u,v) \in R} \log \frac{e^{d(u,v)}}{e^{d(u,v)} + \sum_{v' \in N(u)} e^{d(u,v')}}.
\]

We train the model using SGD with only one set of embeddings. The embeddings are then evaluated on a link reconstruction task: we embed the full tree and rank the similarity of each positive hypernym pair \((u, v)\) among all negative pairs \((u, v')\) and compute the mean rank thus achieved as well as the mean average precision (MAP), using the Wasserstein-Bures dot product as the similarity measure. Elliptical embeddings consistently outperform Poincare embeddings for dimensions above a small threshold, as shown in Figure 6, which confirms our intuition that the addition of a notion of variance or uncertainty to point embeddings allows for a richer and more significant representation of words.

**Conclusion** We have proposed to use the space of elliptical distributions endowed with the \(W_2\) metric to embed complex objects. This latest iteration of probabilistic embeddings, in which a point an object is represented as a probability measure, can consider elliptical measures (including Gaussians) with arbitrary covariance matrices. Using the \(W_2\) metric we can provide a natural and seamless generalization of point embeddings in \(\mathbb{R}^d\). Each embedding is described with a location \(c\) and a scale \(C\) parameter, the latter being represented in practice using a factor matrix \(L\), where \(C\) is recovered as \(LL^T\). The visualization part of work is still subject to open questions. One may seek a different method than that proposed here using precision matrices, and ask whether one can include more advanced constraints on these embeddings, such as inclusions or the presence (or absence) of intersections across ellipses. Handling multimodality using mixtures of Gaussians could be pursued. In that case a natural upper bound on the \(W_2\) distance can be computed by solving the OT problem between these mixtures of Gaussians using a simpler proxy: consider them as discrete measures putting Dirac masses in the space of Gaussians endowed with the \(W_2\) metric as a ground cost, and use the optimal cost of that proxy as an upper bound of their Wasserstein distance. Finally, note that the set of elliptical measures \(\mu_{c,C}\) endowed with the Bures metric can also be interpreted, given that \(C = LL^T, L \in \mathbb{R}^{d \times k}\), and writing \(\bar{I}_i = I_i - \bar{I}\) for the centered column vectors of \(L\), as a discrete point cloud \((c + \frac{1}{\sqrt{k}}\bar{I}_i)\), endowed with a \(W_2\) metric only looking at their first and second order moments. These \(k\) points, whose mean and covariance matrix match \(c\) and \(C\), can therefore fully characterize the geometric properties of the distribution \(\mu_{c,C}\), and may provide a simple form of multimodal embedding.
References


