Information-Geometric Set Embeddings (IGSE): From Sets to Probability Distributions

Ke Sun
CSIRO Data61
Sydney, Australia
sunk@ieee.org

Frank Nielsen
Sony Computer Science Laboratories, Inc.
Tokyo, Japan
Frank.Nielsen@acm.org

Abstract

This letter introduces an abstract learning problem called the "set embedding": The objective is to map sets into probability distributions so as to lose less information. We relate set union and intersection operations with corresponding interpolations of probability distributions. We also demonstrate a preliminary solution with experimental results on toy set embedding examples.

Keywords: discrete to continuous embeddings, statistical manifold, Gaussian manifold, mixture & exponential centroids, information divergence.

1 Set Embedding

The problem called *set embedding* is described as follows: We are given a collection $\mathcal{O} \subset 2^A$ of subsets (i.e., a *family of subsets*) of an *implicitly* given set A, where 2^A denotes the power set of A. We aim to derive a faithful numerical representation of all elements of $\mathcal{O}$ by mapping $\mathcal{O}$ to a continuous space $\mathcal{M}$, so that the images of subsets can approximately preserve the relationships among the subset elements of $\mathcal{O}$. See fig. 1 for an illustration of the general concept, where "statistical manifold" is a continuous space to be introduced latter.

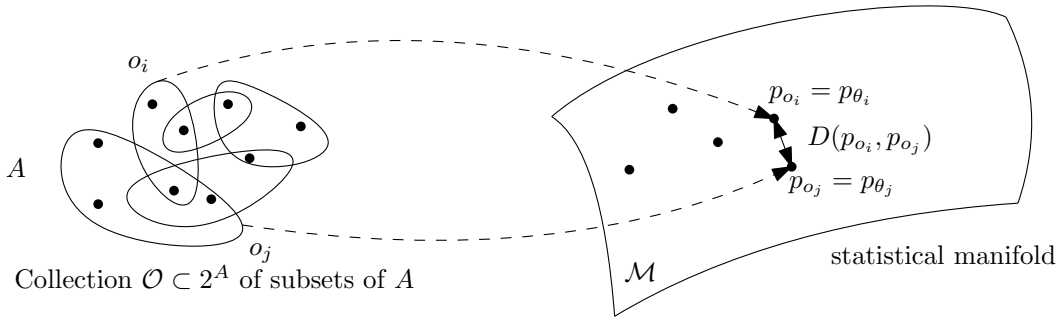


Figure 1: Framework of information-geometric set embedding.

This problem is interesting as a collection of subsets is a very basic algebraic structure, and set embedding gives a uniform and continuous representation on a non-uniform and possibly discrete inputs. Notice that we aim to embed the elements of $\mathcal{O}$ that are subsets rather than elements of A (thus we only need to know A implicitly and can deal with infinite spaces). One can construct singleton subsets to solve the later *element embedding* problem. By definition, set embedding should

To be presented at Sets & Partitions Workshop at 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.
Codes available at `https://informationgeometry.xyz/IGSE`

be *permutation-invariant* because the function input are subsets whose elements are not ordered. In practice, this invariance can be implemented by special neural networks like DeepSets [21].

Set embedding is a broad problem and is closely related to *word and sentence embeddings* [14, 7] (e.g., word2vec with vector space operations) and *graph embedding* [18, 9] (e.g., node2vec). In the problems of word and sentence embeddings, a sentence can be regarded as an *ordered multiset*. They can be extended to embedding partially ordered sets [2]. In graph embedding, a graph is a collection of edges, which is a collection of sets of two nodes. In the case of hypergraphs [22], an edge can be any non-empty set of vertices. Set embedding is a more general concept than *hypergraph embedding*, because the vertex set in hypergraphs is usually finite, while in sets the number of elements can be infinite. The difference with previous work is also in terms of context: set embedding can consider the "metric" properties of A (*e.g.* the volume of subsets) that is often not considered by graph embedding. It can even deal with general *topological spaces*, where the collection $\mathcal{O}$ of open subsets is closed under certain operations. The derived continuous representations can be useful in subsequent downstream machine learning tasks. Set embedding can also be useful for generating visualizations for researchers working on set theory and theoretical computer science, and for making (intuitive) illustrations in textbooks. If the given subsets are all finite, set embedding performs unsupervised multiple instance learning [4].

2 Information-Geometric Set Embedding

We further constrain the problem setting to *information-geometric set embedding* (IGSE), where the embedding target space $\mathcal{M}$ is a statistical manifold [1], *i.e.* a potentially curved space of probability distributions. Hence, the set embedding problem reduces to find for each subset $X \in \mathcal{O}$ a corresponding probability distribution $p_X \in \mathcal{M}$. Our considerations are listed as follows.

- First, a statistical manifold is a generalization of a flat Euclidean space. For example, consider the 2D space of uni-variate Gaussian distributions with the coordinate frame (μ, σ), where μ is the mean and σ is the standard deviation. The Fisher-Rao Riemannian geometry induced by the Fisher metric is of type hyperbolic like for any other location-scale family. Notice that any subspace of constant standard deviation $\{(\mu, \sigma) : \sigma = \sigma_0\}$ is isometric to the real line [8]. Therefore IGSE generalizes real vector embeddings.
- Second, the cardinality $|X|$ of a set X naturally corresponds to the entropy $H(X)$ of a probability distribution p_X, as they both measures the *uncertainty* in drawing a random element. The most striking example is Hartley's entropy [10] defined by $H_{\mathrm{Hartley}}(X) = \log |X|$, and thus $|X| = \exp\left(H_{\mathrm{Hartley}}(X)\right)$.
- Third, informally, a distribution is a "soft" region of space characterized by its support, which bears some similarity to the concept of a set.
- Fourth, basic set operations like union and intersection roughly corresponds to interpolating distributions. A pair of distributions p_A and p_B in an exponential family can have two different types of centroids: their m-*centroid* c_m, which is the linear centroid in the expectation parameters η, or their e-*centroid* c_e, the linear centroid in the natural parameters θ satisfying $c_{\mathrm{e}} \propto \sqrt{p_A p_B}$. Another type of interpolation is taking the mixture model $p_{\mathrm{mix}} = \frac{1}{2}(p_A + p_B)$, which is generally outside of the exponential family (but stays inside for mixture families instead of exponential families). These interpolating methods correspond to two basic set operations: the union and the intersection. See *zero-forcing* (intersection) and *zero-avoiding* (union) properties of left-sided/right-sided Kullback-Leibler centroids [17] (see also [15] for properties with respect to α-divergences). We have $\mathrm{support}(p_{\mathrm{mix}}) \approx \mathrm{support}(A) \cup \mathrm{support}(B)$, and $\mathrm{support}(c_e) \approx \mathrm{support}(A) \cap \mathrm{support}(B)$, where $\mathrm{support}(\cdot)$ denote the *"effective support,"* where the probability density is sufficiently large. See fig. 2 for a toy example of various interpolation schemes of two Gaussians.

We assume that the σ-algebra $(A, 2^A)$ is associated with a probability measure μ, so that the uniform distribution can be defined. We propose the following axioms that an IGSE method should (approximately) satisfy:

① $\forall X \in \mathcal{O}$, the entropy $H(p_X)$ of $p_X \in \mathcal{M}$ is a monotonically increasing function of $H(U_X)$, where U_X means the uniform distribution over the elements of the subset $X \subset A$;

② For (X_1, X_2) a random pair of subsets in $\mathcal{O}^2$, the statistic $D(p_{X_1} : p_{X_2})$ shall be positively correlated to $D(U_{X_1} : U_{X_2})$, where D is an *information divergence* [1] measuring dissimilarities between distributions.

Informally, by the first condition, the capacity or uncertainty is preserved by IGSE. By the second condition, the proximity of any pair of subsets is preserved. That means, subsets with a large overlap are embedded close by, and subsets with little or zero overlap are embedded far away.

Obviously, there exists a trivial set embedding satisfying both conditions: the uniform distribution U_X. However, the uniform distribution is usually not a compact representation and therefore is less useful: To describe the uniform distribution, one needs to describe all elements in the subset X. (This will require to consider the target domain to be $\Delta_{|A|}$, the $|A|$-dimensional standard simplex parameterized by $|A| - 1$ parameters.) Instead, we constrain the target domain $\mathcal{M}$ to a parametric family of distributions with much less parameters, *e.g.* the space of Gaussian distributions, so as to derive a compact uniform numerical representation (similar to the idea of dimensionality reduction).

If $\mathcal{O}$ has a finite cardinality, IGSE is reduced to the problem of embedding histograms. We define the *atomic subsets* w.r.t. $\mathcal{O}$ as $A_1, \cdots, A_m \in 2^A$, so that

- $\forall i \in [m]$, $A_i \neq \emptyset$; if $i \neq j$, $A_i \cap A_j = \emptyset$; $\cup_{i=1}^m A_i = \cup_{X \in \mathcal{O}} X$;
- $\forall X \in \mathcal{O}, \forall i \in [m]$, $A_i \subset X$ or $A_i \cap X = \emptyset$.

The first condition means the atomic subsets form a *partition* of $\cup_{X \in \mathcal{O}} X$. By the second condition, if an atomic subset has overlap with $X \in \mathcal{O}$, then it must be a subset of X. For example, if $\mathcal{O} = \{X_1, X_2\}$, $X_1 - X_2 \neq \emptyset$, $X_2 - X_1 \neq \emptyset$, and $X_1 \cap X_2 \neq \emptyset$, then the atomic subsets are $X_1 - X_2$, $X_2 - X_1$, and $X_1 \cap X_2$. It is straightforward to prove that the set of atom subsets w.r.t. a finite $\mathcal{O}$ is finite by mathematical induction. One can recursively compute the atomic subsets of $\mathcal{O}$ as follows: First, choose any $X \in \mathcal{O}$ and compute the atomic subsets $A_1, \cdots, A_n$ of $\mathcal{O} - X$. Then, the atomic subsets of $\mathcal{O}$ is given by all non-empty subsets in $X - \cup_{i=1}^n A_i$, $A_1 - X, \cdots, A_n - X$, $A_1 \cap X, \cdots, A_n \cap X$. In the worse case, the number of atomic subsets grows exponentially w.r.t. $|\mathcal{O}|$. The concept of atomic subsets yields an equivalence relation in 2^{2^A}, which contains all families of subsets: If $\mathcal{O}_1$ and $\mathcal{O}_2$ induce the same atomic subsets, then we denote $\mathcal{O}_1 \sim_{\text{atom}} \mathcal{O}_2$. It means that the closures of $\mathcal{O}_1$ and $\mathcal{O}_2$ under basic set operations (intersection, union, subtraction) are the same. We propose the following invariance that an IGSE method should (try to) satisfy

③ If $\mathcal{O}_1 \sim_{\text{atom}} \mathcal{O}_2$, then their IGSE should be consistent, in the sense that in both embeddings p_X is the same for any $X \in \mathcal{O}_1 \cap \mathcal{O}_2$.

Then, $\forall X \in \mathcal{O}$, the associated U_X can be defined as a histogram over the atomic subsets. We iterate over the set of all atomic subsets, and select the ones which satisfy $A_i \subset X$, then U_X is a mixture distribution

$$U_X = \frac{1}{Z} \sum_{i : A_i \subset X} V(A_i) U_{A_i}, \tag{1}$$

where $V(A_i)$ is the volume of A_i w.r.t. the base measure μ, and Z is the partition function. Therefore, the problem is reduced to embedding histograms as other families of probability distributions. This

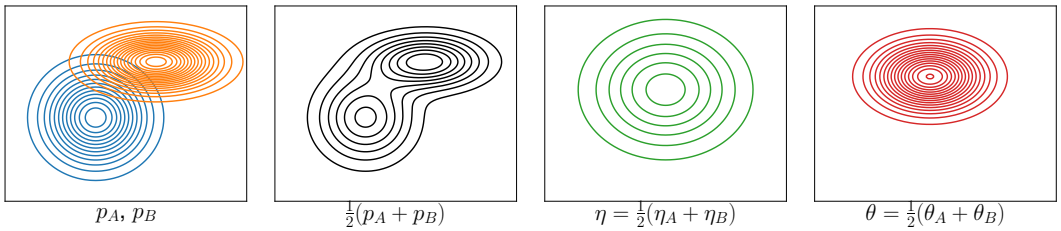


Figure 2: From left to right: (1) Two Gaussian distributions; (2) their mixture; (3) Their m-centroid (center of mass in the expectation parameters; mixture centroid tending to cover the support of the mixture); (4) Their e-centroid (center of mass in the natural parameters; exponential centroid tending to cover the mode of the mixture).

is different from information-geometric dimensionality reduction [5, 19] and distribution regression [20] in that both the source domain and the target domain is a statistical manifold. Distribution regression [20] is a supervised learning problem and learns a mapping from a distribution, given implicitly by a set of random samples, to a real valued response. Our problem setting is also different from information geometric kernel density estimation (IGKDE; chapter 4 [19]) or embedding graph nodes into probability distributions [3].

Notice that we can also associate to each element a of A an elementary Dirac probability distribution $p_a(x) = \delta_a(x) = \delta(x - a)$ (1 iff $x = a$ and 0 otherwise), where the sample space of the distribution is A. Then we view a subset $O \in \mathcal{O}$ as a mixture distribution of Diracs ("empirical distribution" of the subset): $p_O(x) = \frac{1}{|O|} \sum_{a \in O} \delta_a(x)$, where $|\cdot|$ denote the cardinality of subset O. Thus we reinterpret Eq. 1 as a decomposition of a set mixture distributions into *atomic subset mixtures*.

Set embedding is different from graph drawing of Venn diagrams [6] that is a topological subset embedding.

3 A Preliminary Solution

We describe a preliminary solution to solve IGSE through the Kullback-Leibler (KL) divergence and the Jensen-Shannon (JS) divergence on the Gaussian manifold, *i.e*. the space of Gaussian distributions. Our purpose is not a systematical empirical study, but to get some intuitions, and to show that IGSE is easy to implement in practice.

By definition, the KL divergence is $\mathrm{KL}(p : q) = \int p(\boldsymbol{x}) \left[\log p(\boldsymbol{x}) - \log q(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x}$. It can be infinite if the input U_{X_1} and U_{X_2} have different supports. We therefore use the *damped KL* divergence

$$\mathrm{KL}_\epsilon(p : q) = \max\left(\int p(\boldsymbol{x}) \left[\log p(\boldsymbol{x}) - \log\left(q(\boldsymbol{x}) + \epsilon\right)\right] \mathrm{d}\boldsymbol{x},\ 0\right), \tag{2}$$

where $\epsilon > 0$ is a small constant, to measure the distances between the input uniform distributions. For positive measures $\tilde{p}$ and $\tilde{q}$, we can define the extended Kullback-Leibler divergence as

$$\mathrm{KL}_+(\tilde{p} : \tilde{q}) = \mathrm{KL}(\tilde{p} : \tilde{q}) + \tilde{q} - \tilde{p}. \tag{3}$$

Thus $\mathrm{KL}_\epsilon(p : q) = \mathrm{KL}(p : q) + \epsilon \int \mu(x) \mathrm{d}\boldsymbol{x}$, where the integral is calculated on the union of the support distributions.

For the output Gaussian distributions, their KL divergence is always well defined if their covariance matrices have full rank. Denote a pair of multivariate Gaussian distributions as $G_1(\boldsymbol{x}\,;\, \boldsymbol{\mu}_1, \mathrm{diag}(\boldsymbol{\sigma}_1))$ and $G_2(\boldsymbol{x}\,;\, \boldsymbol{\mu}_2, \mathrm{diag}(\boldsymbol{\sigma}_2))$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the means, $\mathrm{diag}(\boldsymbol{\sigma}_1)$ and $\mathrm{diag}(\boldsymbol{\sigma}_2)$ are the covariance matrices, and $\mathrm{diag}(\cdot)$ denotes the diagonal matrix constructed using the given diagonal entries. We have

$$\mathrm{KL}(G_1 : G_2) = \sum_{j=1}^{d} \left[\log \sigma_2^{(j)} - \log \sigma_1^{(j)} + \frac{1}{2} \frac{(\sigma_1^{(j)})^2 + (\mu_1^{(j)} - \mu_2^{(j)})^2}{(\sigma_2^{(j)})^2}\right] - \frac{d}{2}, \tag{4}$$

where $\sigma_1^{(j)}$ denotes the j'th entry of the vector $\boldsymbol{\sigma}_1$, and $d = \dim(\boldsymbol{x})$. On the other hand, the Jensen-Shannon (JS) divergence [16] is

$$\mathrm{JS}(p : q) = \frac{1}{2}\mathrm{KL}\left(p(\boldsymbol{x}) : \frac{p(\boldsymbol{x}) + q(\boldsymbol{x})}{2}\right) + \frac{1}{2}\mathrm{KL}\left(q(\boldsymbol{x}) : \frac{p(\boldsymbol{x}) + q(\boldsymbol{x})}{2}\right), \tag{5}$$

which is bounded in the range $[0, 1]$ (when using base-2 logarithms) and can naturally handle the case when p and q have different supports. To compute the JS divergence between two embedding points, we need to solve the KL divergence between a Gaussian distribution and a Gaussian mixture of two components. This KL divergence, on the RHS of eq. (5), by definition is an integration, which can be approximated by Monte-Carlo sampling techniques and the reparameterisation trick [13]. Given G_1

Table 1: Five different families of subsets

$\mathcal{O}_1$	{A} , {B} , {C} , {A,B} , {B,C} , {C, A} , {A,B,C}
$\mathcal{O}_2$	{A,B} , {B,C} , {A} , {B} , {C}
$\mathcal{O}_3$	{A,B,C,D,E,F} , {B,C,D} , {C,D,E} , {A}
$\mathcal{O}_4$	{A,B,C,D,E,F,G} , {A,B,C,D,E} , {A,B,C} , {A}
$\mathcal{O}_5$	{A,B,C} , {B,C,D} , {C,D,E} , {D,E,A} , {E,A,B}

and G_2, we have the approximation

$$\mathrm{KL}\left(G_1(\boldsymbol{x}) \,:\, \frac{G_1(\boldsymbol{x})+G_2(\boldsymbol{x})}{2}\right) \approx -\sum_{j=1}^{d}\log\sigma_1^{(j)} - \frac{d}{2} + \log 2$$
$$-\frac{1}{K}\sum_{i=1}^{K}\log\left[\exp\left(\sum_{j=1}^{d}\left(-\log\sigma_1^{(j)} - \frac{1}{2(\sigma_1^{(j)})^2}(x_i^{(j)}-\mu_1^{(j)})^2\right)\right)\right.$$
$$\left.+\exp\left(\sum_{j=1}^{d}\left(-\log\sigma_2^{(j)} - \frac{1}{2(\sigma_2^{(j)})^2}(x_i^{(j)}-\mu_2^{(j)})^2\right)\right)\right], \tag{6}$$

where $\{\boldsymbol{x}_i\}_{i=1}^K$ are i.i.d. samples drawn from $G_1(\boldsymbol{x})$. At the limit $K\to\infty$, the approximation becomes accurate (i.e., consistent). In summary, the divergence for all pairs in $\mathcal{O}$ and for all pairs of embedding distributions can therefore be computed.

Given $\mathcal{O}$, we first augment it with the union of

$$\{X_1\cap X_2 \,:\, X_1, X_2\in\mathcal{O}^2\},$$
$$\{X_1\cup X_2 \,:\, X_1, X_2\in\mathcal{O}^2\},$$
$$\{X_1 - X_2 \,:\, X_1, X_2\in\mathcal{O}^2\},$$
$$\{X_2 - X_1 \,:\, X_1, X_2\in\mathcal{O}^2\},$$

or a random subset of the union, so that our IGSE has certain invariance w.r.t. set operations, as stated in axiom ③. Note that it is hard to accurately satisfy ③, because there is an exponentially large number of $\mathcal{O}'$ which satisfies $\mathcal{O}\sim_{\text{atom}}\mathcal{O}'$.

Then, one can implement the IGSE through minimizing the *stress function*:

$$\text{(axiom ②)}\qquad \sum_{(X_1,X_2)\in\mathcal{O}^2}\|D(U_{X_1}:U_{X_2}) - aD(G_1:G_2)\|_2^2, \tag{7}$$

with an auto-differentiation framework, where G_i is the Gaussian distribution associated with X_i, and $a\in\Re^+$ is a free parameter. In order to satisfying our axiom ①, we further constrain the covariance matrix of G_i to be $\mathrm{diag}(\boldsymbol{\sigma}_i)$, where

$$\text{(axiom ①)}\qquad \log\sigma_i^{(j)} = \tau^{(j)} + \log V_i,$$

where $\boldsymbol{\tau}\in\Re^d$ are free parameters, and V_i is the volume or the number of elements in X_i. Intuitively, the larger the input subsets, the larger the variance of the embedding distributions. To avoid this reparameterisation of $\boldsymbol{\sigma}_i$, an alternative method is to simply initialize the free parameter $\sigma_i^{(j)}$ with V_i. The computational complexity of the stress function is $O(Kd|\mathcal{O}|^2)$, which can be further reduced by random sampling of the pairs $(X_1, X_2)\in\mathcal{O}^2$.

As a toy example, we embed discrete subsets into 2D Gaussian distributions with diagonal covariance matrices. In order to minimize the stress function in eq. (7), we apply batch gradient descent using the Adam optimizer [12] based on a constant learning rate. We use the set cardinality to initialize the covariance matrix. See table 1 for five families of subsets. See fig. 3 for their corresponding 2D Gaussian embeddings. For the first dataset $\mathcal{O}_1$, both embeddings based on the KL and JS divergences

can faithfully present the power set of $\{A, B, C\}$. For $\mathcal{O}_2$, both embeddings reflect the relationships between the given subsets. For example, G_4 is in the middle of G_1 and G_2, showing the relationship $\{\mathrm{B}\} = \{\mathrm{A}, \mathrm{B}\} \cap \{\mathrm{B}, \mathrm{C}\}$. For $\mathcal{O}_3$, both embeddings are similar and informative w.r.t. the input subsets. For example, {B,C,D} (embedded into G_2) and {C,D,E} (embedded into G_3) have a overlap. For $\mathcal{O}_4$, JS appears better than KL, as it show a series of Gaussian distributions with decreasing variance, and roughly contained in one another. For $\mathcal{O}_5$, JS also appears better because it shows a circular structure of the given subsets. Overall, our toy IGSE method based on two different divergences can intuitively represent a given family of subsets, where JS divergence seems to perform better as it can naturally handle distributions with different support.

If the input contains not only a family $\mathcal{O}$ of subsets, but also the features of the set elements, one should consider using deep neural networks which are designed to be permutation invariant [21, 11]. These networks provide a *parametric mapping* between the subsets and their embedding images (distributions). This is different from the above non-parametric approach, where the probability distributions are free parameters to be learned. In this case, the neural network output should be a distribution satisfying our axiom ①, and the cost function should be designed to satisfy our axiom ②. It is also possible to use such networks for the general case by feeding one-hot vectors as the input features.

References

[1] S. Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, 2016.

[2] M. A. Bautista, A. Sanakoyeu, and B. Ommer. Deep unsupervised similarity learning using partially ordered sets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1923–1932, 2017.

[3] Aleksandar Bojchevski and Stephan Günnemann. Deep Gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *ICLR*, 2018.

[4] Marc-Andr Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning. *Pattern Recogn.*, 77(C):329–353, 2018.

[5] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero,III. Information-geometric dimensionality reduction. *IEEE Signal Processing Magazine*, 28(2):89–99, 2011.

[6] Stirling Chow and Frank Ruskey. Drawing area-proportional Venn and Euler diagrams. In *International Symposium on Graph Drawing*, pages 466–477. Springer, 2003.

[7] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.

[8] Sueli I. R. Costa, Sandra A. Santos, and João E. Strapasson. Fisher information distance: a geometrical reading, 2012. arXiv:1210.2354 [stat.ME].

[9] A. Grover and J. Leskovec. Node2Vec: Scalable feature learning for networks. In *KDD*, pages 855–864, 2016.

[10] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, 7(3):535–563, 1928.

[11] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[15] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[16] Frank Nielsen. On the Jensen-Shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

[17] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

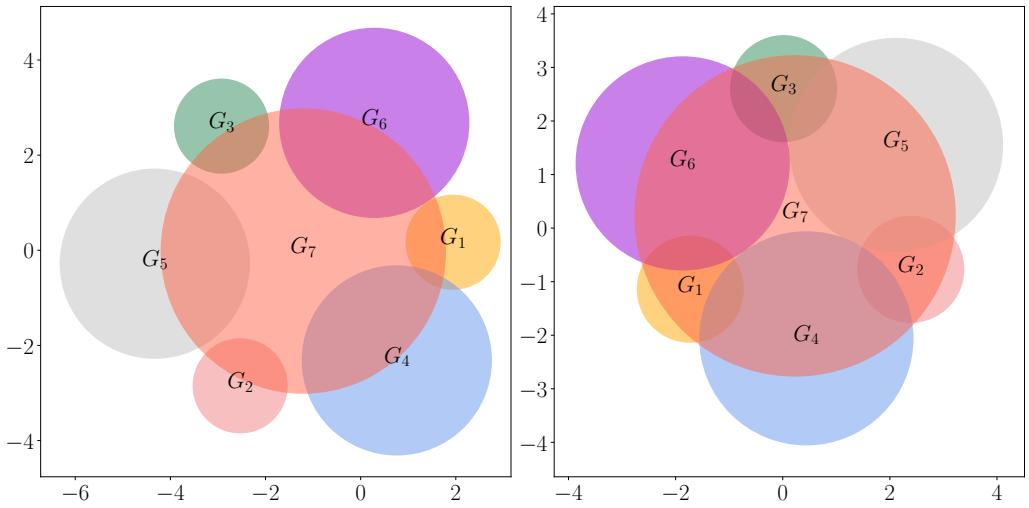
[18] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online learning of social representations. In *KDD*, pages 701–710, 2014.

[19] Ke Sun. *Information Geometry and Data Manifold Representations*. PhD thesis, Université de Genève, 2015.

[20] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

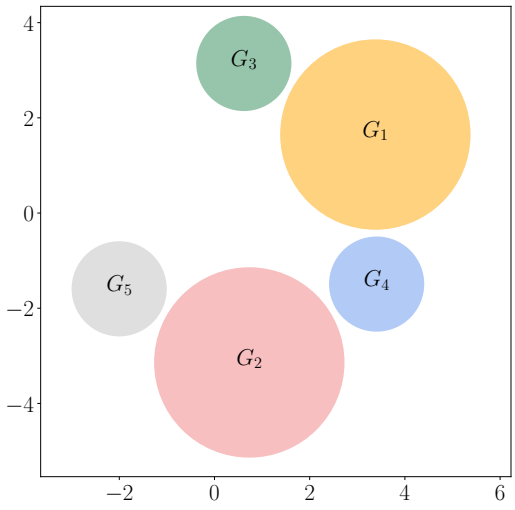
[21] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc., 2017.

[22] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *NIPS 19*, pages 1601–1608. MIT Press, 2007.

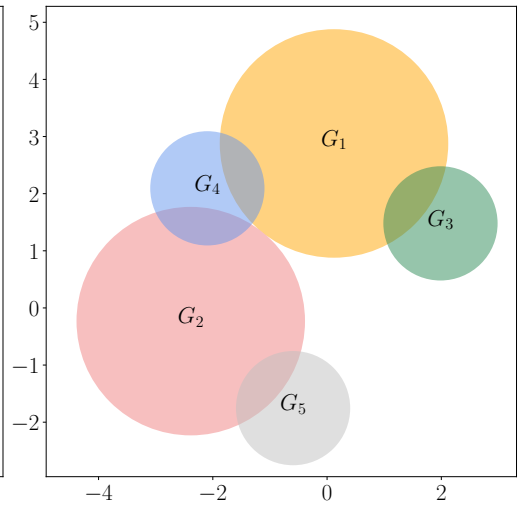


(a) KL ($\mathcal{O}_1$)

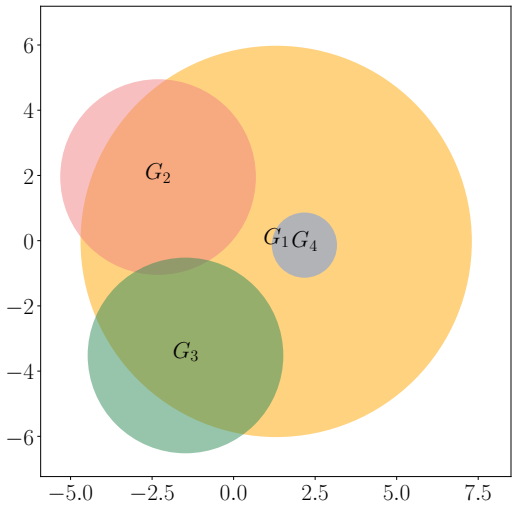
(b) JS ($\mathcal{O}_1$)



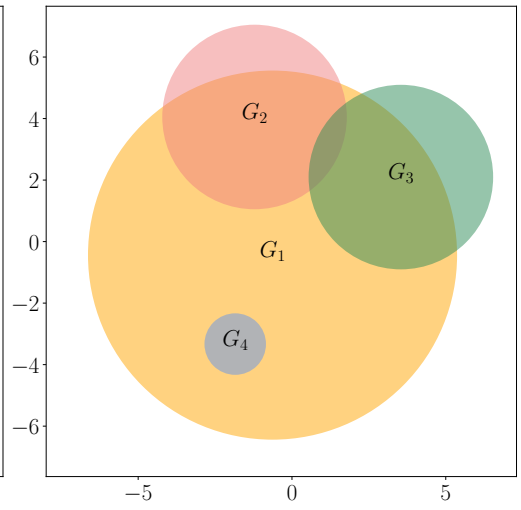
(c) KL ($\mathcal{O}_2$)



(d) JS ($\mathcal{O}_2$)

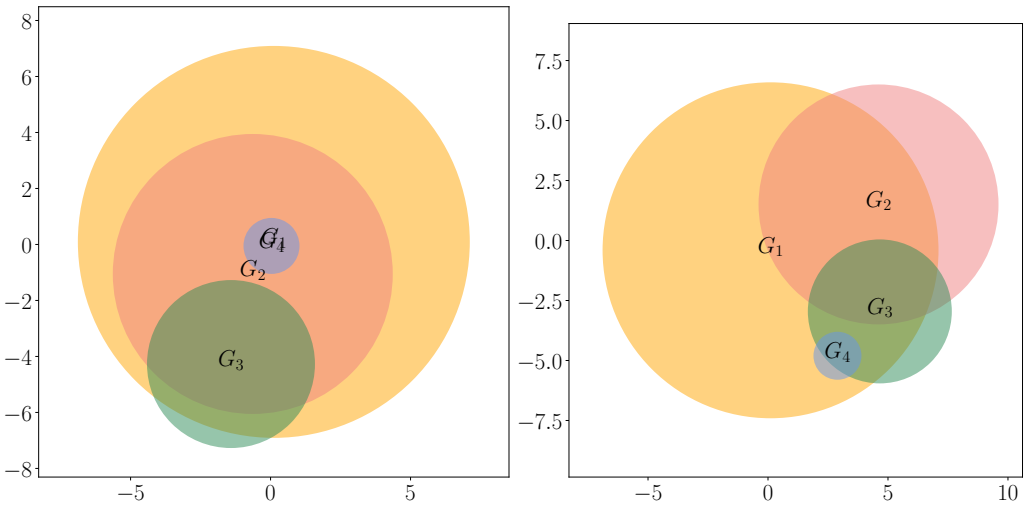


(e) KL ($\mathcal{O}_3$)

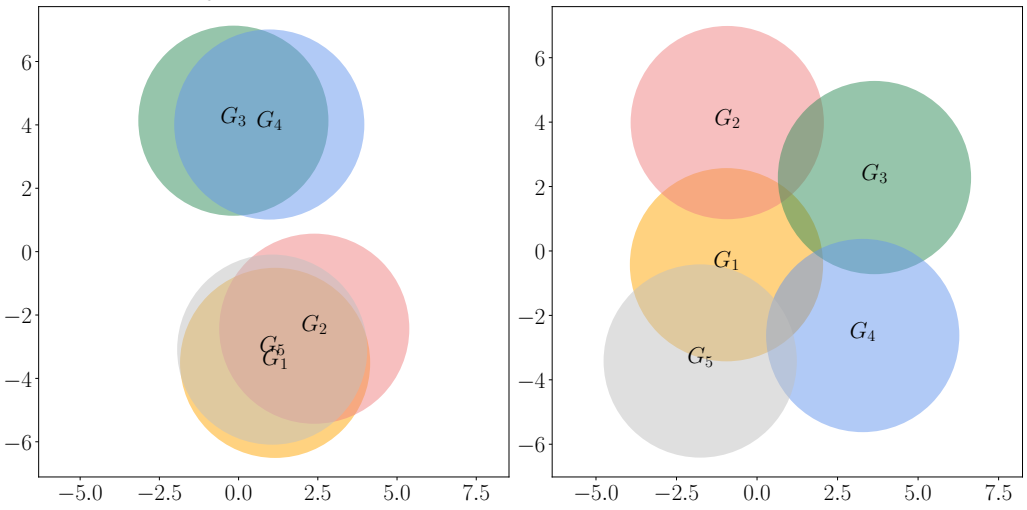


(f) JS ($\mathcal{O}_3$)

Figure 3: Embedding $\mathcal{O}_1$ (top), $\mathcal{O}_2$ (middle), and $\mathcal{O}_3$ (bottom) into 2D Gaussian distributions based on KL divergence (left) and JS divergence (right). Each colored ellipse represents a 2D Gaussian distribution. It shows the 2D region defined by mean$\pm$std.



(g) KL ($\mathcal{O}_4$) (h) JS ($\mathcal{O}_4$)



(i) KL ($\mathcal{O}_5$) (j) JS ($\mathcal{O}_5$)

Figure 3 (Cont.): Embedding $\mathcal{O}_4$ and $\mathcal{O}_5$ into 2D Gaussian distributions based on KL divergence (left) and JS divergence (right).