Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2006-036 May 16, 2006
CBCL-261

Learning using the Born Rule

Lior Wolf

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Learning using the Born Rule

Lior Wolf

The McGovern Institute for Brain Research
Massachusetts Institute of Technology
Cambridge, MA
liorwolf@mit.edu

Abstract

In Quantum Mechanics the transition from a deterministic description to a probabilistic one is done using a simple rule termed the Born rule. This rule states that the probability of an outcome (a) given a state (Ψ) is the square of their inner products ($(a^\top \Psi)^2$).

In this paper, we unravel a new probabilistic justification for popular algebraic algorithms, based on the Born rule. These algorithms include two-class and multiple-class spectral clustering, and algorithms based on Euclidean distances.

1 Introduction

In this work we form a connection between spectral theory and probability theory. Spectral theory is a powerful tool for studying matrices and graphs and is

often used in machine learning. However, the connections to the statistical tools employed in this field, are ad-hoc and domain specific.

Consider, for example, an affinity matrix $A = [a_{ij}]$ where a_{ij} is a similarity between points i and j in some dataset. One can obtain an effective clustering of the data points by simply thresholding the first eigenvector of this affinity matrix [Perona and Freeman, 1998]. Most previously suggested justifications of spectral clustering will fail to explain this: the view of spectral clustering as an approximation of the normalized minimal graph cut problem [Shi and Malik, 2000] requires the use of the normalized Laplacian; the view of spectral clustering as the infinite limit of a stochastic walk on a graph [Meila and Shi, 2000] requires a stochastic matrix; other approaches can only explain this when the affinity matrix is approximately block diagonal [Ng et al., 2001].

In this work we connect the spectral theory and probability theory by utilizing the basic probability rule of Quantum Mechanics. We show that by modeling class membership using the Born rule one attains the above spectral clustering algorithm as the two class clustering algorithm. Moreover, let $v_1 = [v_1(1), v_1(2), ...]^\top$ be the first eigenvector of A, with an eigenvalue of λ_1, then according to this model the probability of point j to belong to the dominant cluster is given by $\lambda_1 v_1(j)^2$. This result is detailed in Sec. 3, as well as a justification to one of the most popular multiple-class spectral-clustering algorithms.

In Quantum Mechanics the Born rule is usually taken as one of the axioms. However, this rule has well-established foundations. Gleason's theorem [Gleason, 1957] states that the Born rule is the only consistent probability distribution for

a Hilbert space structure. Wootters [1981] was the first to note the intriguing observation that by using the Born rule as a probability rule, the natural Euclidean metric on a Hilbert space coincides with a natural notion of a statistical distance. Recently, attempts were made to derive this rule with even fewer assumptions by showing it to be the only possible rule that allows certain invariants in the quantum model. We will briefly review these justifications in Sec. 2.1.

Physics based methods have been used in solving learning and optimization problems for a long time. Examples include simulated annealing, the use of statistical mechanic models in neural networks, and heat equations based kernels. However, we would like to stress that all that is stated here has grounds that are independent of any physical model. From a statistics point of view, our paper could be viewed as a description of learning methods that use the Born rule as a plug-in estimator [Devroye et al., 1996]. Spectral clustering algorithms can then be seen as posterior based clustering algorithms [Sinkkonen et al., 2002].

1.1 Previous Justifications to Spectral Algorithms

We would like to motivate our approach by considering some of the previous justifications available for algebraic learning algorithms. This short review of previous approaches would serve us in demonstrating two points: (I) Justifications to spectral algorithms are highly sought after, especially probabilistic ones; (II) The existing justifications have limitations.

The first explanation, given in [Perona and Freeman, 1998], is a compression type of argument. The justification for the use of the first eigenvector of the affinity

matrix was that out of all unit norm vectors it is the one that gives rise to the rank-1 matrix which is most similar to that matrix. This explanation is simple, but does not give any insight outside the scope of algebraic estimation.

Another explanation provided by Perona and Freeman [1998] is based on the block diagonal case. This perspective is also the basis of the perturbation theory analysis given in [Ng et al., 2001], and it is based on the fact that an ideal affinity matrix, once the columns and the rows have been reordered, is block diagonal. While the block diagonal case is very appealing, the explanations based on it are limited to matrices that are block diagonal up to a first order approximation. Most affinity matrices are not of this type.

Nowadays, the dominant view with regards to the effectiveness of spectral clustering is as an approximation of the normalized minimal graph cut problem [Shi and Malik, 2000]. Given a division of the set of vertices of a weighted graph V into two sets S and R, the normalized graph cut score is defined as $\frac{cut(S,R)}{assoc(S)} + \frac{cut(S,R)}{assoc(R)}$, where $cut(S, R)$ is the sum of weights of all edges that connect an element from S with an element from R, and $assoc(S)$ is the sum of all weights of the edges that have an end in the set S.

Let D be the diagonal matrix defined as $D_{ii} = \sum_j A_{ij}$, the Laplacian of A is defined as $L = D - A$. The minimization for the above graph cut score can be written as the minimization of an expression of the form $\frac{y^\top Ly}{y^\top Dy}$, such that y is a set indicator vector with elements that are either 1 or a value of $-b$ (see below) under the constrain that the sum of $y^\top D$ is zero. b is a variable that depends (linearly) on the sets, but this is mostly ignored in the literature we are aware of.

The above optimization problem is relaxed to the minimization of $\frac{y^\top Ly}{y^\top Dy}$ where y is a real valued vector, which is algebraically equivalent to finding the leading eigenvector $N = D^{-1/2}AD^{-1/2}$. The major advantage of the normalized cut explanation is that it explains why the "additively normalized affinity matrix" L and the "multiplicatively normalized affinity matrix" N tend to give better clustering results in practice than the original matrix A. However, this is not always the case, and this approach fails to explain the success of working directly with the original affinity matrix. In addition, this explanation is not probabilistic in nature, and it is based on a crude relaxation of an integer programming problem.

An attempt to give a probabilistic explanation to the (binary) Ncut criterion is given in [Meila and Shi, 2000] where spectral clustering is viewed as the infinite limit of a stochastic walk on a graph [Meila and Shi, 2000]. This explanation is based upon the fact that the first eigenvector of a stochastic matrix encodes the probabilities of being the nodes of the associated graph after an infinite number of steps (the stationary distribution). The arguments given in this framework are based on the observation that disjoint clusters are clusters for which the probability of transitions from one cluster to the other, while starting at the stationary distribution, is small. The above explanation, similar to the previous explanations, does not provide an effective multiple class algorithm (compared to the NJW algorithms [Verma and Meila, 2003]).

The efforts to provide sound explanations to algebraic algorithms are wide spread. In sec. 4 we will provide a justification to the simple yet effective Relevant Component Analysis (RCA) algorithm [Bar-Hillel et al., 2005] designed to

learn a distance matrix from partially labeled data. The RCA algorithm is conceptually straightforward as it simply finds a distance that minimizes the Euclidean distances of points that are known to belong to the same class. In order to further justify it, Bar-Hillel et al. derived an information theoretical explanation as well as a Gaussian maximum likelihood explanations. Here we provide another justification.

It is a major goal of this work that algebraic algorithms, such as RCA, which employ Euclidean distances, would be considered justified based on the Born rule, without a need to provide other probabilistic models for the sake of justification.

2 The Quantum Probability Model

The quantum probability model takes place in a Hilbert space $\mathcal{H}$ of finite or infinite dimension (The results in this paper hold for both Hilbert and real vector spaces). A *state* is represented by a positive definite linear mapping (a matrix ρ) from this space to itself, with a trace of 1, i.e $\forall \Psi \in \mathcal{H} \;\; \Psi^\top \rho \Psi \geq 0 \; , \;\; Tr(\rho) = 1$. Such ρ is self adjoint ($\rho^\top = \rho$, where throughout this paper $\top$ means complex conjugate and transpose) and is called a density matrix.

Since ρ is self adjoint, its eigenvectors Φ_i are orthonormal ($\Phi_i^\top \Phi_j = \delta_{ij}$), and since it is positive definite its eigenvalues p_i are real and positive $p_i \geq 0$. The trace of a matrix is equal to the sum of its eigenvalues, therefore $\sum_i p_i = 1$.

The equality $\rho = \sum_i p_i \Phi_i \Phi_i^\top$ is interpreted as "the system is in state Φ_i with probability p_i". The state ρ is called pure if $\exists i$ s.t $p_i = 1$. In this case, $\rho = \Psi \Psi^\top$

for some normalized *state vector* Ψ, and the system is said to be in state Ψ. Note that the representation of a mixed (not pure) state as a mixture of states with probabilities is not unique.

A *measurement* M with an outcome x in some set $\mathcal{X}$ is represented by a collection of positive definite matrices $\{m_x\}_{x\in\mathcal{X}}$ such that $\sum_{x\in\mathcal{X}} m_x = 1$ (1 being the identity matrix in $\mathcal{H}$). Applying a measurement M to state ρ produces outcome x with probability

$$p_x(\rho) = Tr(\rho m_x). \tag{1}$$

Eq. 1 is the *Born rule*. Most quantum models deal with a more restrictive type of measurement called the *von Neumann* measurement, which involves a set of projection operators $m_a = aa^\top$ for which $a^\top a' = \delta_{aa'}$. As before $\sum_{a\in M} aa^\top = 1$. For this type of measurement the Born rule takes a simpler form: $p_a(\rho) = Tr(\rho aa^\top) = Tr(a^\top \rho a) = a^\top \rho a$. Assuming ρ is a pure state $\rho = \Psi\Psi^\top$, this can be simplified further to:

$$p_a(\rho) = (a^\top \Psi)^2 \quad . \tag{2}$$

Most of the derivations below require the recovery of the parameters of unknown distributions. Therefore, in order to reduce the number of parameters, we use the simpler model (Eq. 2), in which a measurement is simply given as an orthogonal basis. However, some applications and future extensions might require the more general form.

2.1 Whence the Born rule?

The Born rule has an extremely simple form that is convenient to handle, but why should it be considered justified as a probabilistic model for a broad spectrum of data? In quantum physics the Born rule is one of the axioms, and it is essential as a link between deterministic dynamics of the states and the probabilistic outcomes. It turns out that this rule cannot be replaced with other rules, as any other probabilistic rule would not be consistent. Next we will briefly describe three existing approaches for deriving the Born rule, and interpret their relation to learning problems.

Assigning probabilities when learning with vectors. Gleason's theorem [Gleason, 1957] derives the Born rule by assuming the Hilbert-space structure of *observables* (what we are trying to measure). In this structure, each orthonormal basis of $\mathcal{H}$ corresponds to a mutually exclusive set of results of a given measurement.

Theorem 1 (Gleason's theorem) *Let $\mathcal{H}$ be a Hilbert space of a dimension greater than 2. Let f be a function mapping the one-dimensional projections on $\mathcal{H}$ to the interval [0,1] such that for each orthonormal basis $\{\Psi_k\}$ of $\mathcal{H}$ it happens that $\sum_k f(\Psi_k \Psi_k^\top) = 1$. Then there exists a unique density matrix ρ such that for each $\Psi \in \mathcal{H} \quad f(\Psi\Psi^\top) = \Psi^\top \rho \Psi$.*

By this theorem the Born rule is the only rule that can assign probabilities to all the measurements of a state (each measurement is given by a resolution of the identity matrix), such that these probabilities rely only on the density matrix of

that specific state.

To be more concrete: we are given examples (data points) as vectors, and are asked to assign them to clusters or classes. Naturally, we choose to represent each class with a single vector (in the simplest case). We do not want the clusters to overlap. Otherwise, in almost any optimization framework the most prominent cluster would appear repeatedly. The simplest constraint to add is an orthogonality constraint on the unknown vector representations of the clusters. Using this simple model and simple constraint, a question arises: what probability rule can be used? Gleason's theorem suggests that under the constraint that all the probabilities sum to one there is only one choice.

Gleason's theorem is very simple and appealing, and it is very powerful as a justification for the use of the Born rule. However, the assumptions of Gleason's theorem are somewhat restrictive. They assume that the algorithm that assigns probabilities to measurements on a state has to assign probabilities to all possible von Neumann measurements. Some Quantum Mechanic approaches, as well as our use of the Born rule in machine learning, do not require that probabilities are defined for all resolutions of the identity.

Axiomatic approaches. Recently a new line of inquiry has emerged that tries to justify the Born rule from the axioms of the decision theory. The first work in this direction was done by Deutsch [1999]. Lately, Saunders [2004] has shown that the Born rule is the only probability rule that is invariant to the specific representation of several models, and new results keep on getting published. The full discussion of this is omitted.

Statistical approach. Another approach for justifying the Born rule, different from the axiomatic one above, is the approach taken by Wootters [1981]. Wootters defines a natural notion of statistical distance between two states that is based on *statistical distinguishability*. Assume a measurement with N outcomes. For some state, the probability of each outcome is given by p_i. By the central limit theorem, two states are distinguishable if for large n: $\sqrt{n}/2 \sum_{i=1}^{N} (\delta p_i)^2/p_i > 1$, where δp_i is the difference in their probabilities for outcome i. If they are indistinguishable, then the frequencies of outcomes we get from the two states by repeating the measurement are not larger than the typical variation for one of the states. The distance between states Ψ and Ψ' is simply the length of the shortest chain of states that connects the two states such that every two states in this chain are indistinguishable. Wootters showed that the Born rule could be derived by assuming that the statistical distance between states is proportional to the angle between their vector representation.

The implication of this result is far reaching: We often use the distance between two given vectors as their similarity measure. We would sometimes be interested in deriving relations which are more complex than just distances. Statistics and probabilities are a natural way to represent such relations, and the Born rule is a probabilistic rule which conforms with the natural metric between vectors.

Other previous work. In contrast to some connections between Spectral Clustering and (classical) Random Walks, we would like to explicitly state that this work has little to do with current studies in Quantum Computation, and in par-

ticular with Quantum Random Walks, which deal with random walks for which the current state can be in superposition. We would also like to state that this work has little to do with the line of work on Quantum Mechanics as a Complex Probability Theory [Youssef, 1994]. Quantum Mechanics Probabilities is of course a very active field of research. There is, however, no previous work to our knowledge that relates spectral machine learning algorithms with the QM probability model.

In the next section, we use the Born rule to derive the most popular variants of spectral clustering. Interestingly, Horn and Gottlieb [2002] have developed a clustering method based on Quantum Mechanics concepts. Their method, however, is based on the Schrödinger potential equation and is different from ours. To avoid confusion, we use algebraic concepts and not physics-based concepts. The probabilities will be derived from the Born rule directly, without giving it QM interpretations.

3 Spectral Clustering

Spectral clustering methods are becoming extremely popular, e.g., for image segmentation. We show that these methods can be seen as model based segmentation, where the underlying model in the Born probability model. Spectral clustering belong to the family of kernel based algorithms [Schölkopf and Smola, 2002]. For these family of algorithms, the similarities between every two data points is given by a positive definite kernel function. Since the kernel is positive definite, it can

be factored and the data points can be viewed as vectors in some vector space. We will represent points in such kernel spaces with upper Greek letters. Care should be taken never to require the explicit computation of vectors or matrices in those implicit vector spaces, as these could be of infinite dimension. Below this dimension is marked as d.

Two class clustering. We are given a set of n norm-1 input samples $\{\Phi_j\}_{j=1}^n$ that we would like to cluster into k clusters. We model the probability of a point Φ to belong to cluster i as $p(i|\Phi) = a_i^\top \Phi\Phi^\top a_i, \quad$ where $a_i^\top a_j = \delta_{ij}$. i.e., we use our input vectors as state vectors and use von Neumann measurement to determine cluster membership.

A natural score to maximize is the sum over all of the k clusters of the posterior probabilities given our input points. For every data point this sum would be 1 if this point if fully modeled by the k clusters, however, since $k < \min(d, n)$ we can expect this sum to be lower than 1. Hence, in order to concentrate as much probability as possible in the first k posteriors, we maximize

$$L(\{a_i\}_{i=1}^k) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k} p(i|\Phi_j) = \sum_{i=1}^{k} a_i^\top \Gamma_\Phi a_i \quad , \tag{3}$$

where $\Gamma_\Phi = \frac{1}{n}\sum_{j=1}^n \Phi_j\Phi_j^\top$.

$L(\{a_i\}_{i=1}^k)$ is maximized when $\{a_i\}_{i=1}^k$ are taken to be the first k eigenvectors of Γ_Φ [Coope and Renaud, 2000]. This maximization is unique up to multiplication with a $k \times k$ unitary matrix.

Let $\phi = [\Phi_1|\Phi_2|...|\Phi_n]$. Similar to other kernel methods, it is useful to under-

stand the relationships between the eigenvectors of $\phi\phi^\top = n\Gamma_\Phi$ and the eigenvectors of $\phi^\top\phi$. For this end, consider the singular value decomposition of the matrix $\phi = USV^\top$, where the columns of U are the eigenvectors of $\phi\phi^\top$, the matrix S is a diagonal matrix constructed by the sorted square roots of the eigenvalues of $\phi\phi^\top$, and the matrix V contains the eigenvectors of $\phi^\top\phi$. Let u_i (v_i) denote the ith column of U (V) and let s_i denote the ith element along the diagonal of S. For all i not exceeding the dimensions of ϕ we have: $s_i v_i = \phi^\top u_i$.

If we are only interested in bipartite clustering we can just use the first eigenvector and get $p(1|\Phi_j) = (u_1^\top \Phi_j)^2 = (s_1 v_1(j))^2, \quad v_i(j)$ being the jth element of the vector v_i. Hence, the probability of belonging to the first cluster is proportional to the first eigenvector of the data's affinity matrix $\phi^\top\phi$. A bipartite clustering would be attained by thresholding this value. This approach agrees with most spectral clustering methods for bipartite clustering [1].

The use of the Born rule for deriving the clustering algorithm is more than just a justification for the same algorithm used by many. In addition to binary class memberships, we receive the probability for each point to belong to each of the two clusters. However, as in many spectral clustering approaches, the situation is less appealing when considering more than two clusters.

[1] Some spectral clustering methods [Perona and Freeman, 1998] use a threshold on the value of $v_1(j)$ and not on its square. However, by the Perron-Frobenius theorem for kernels with non-negative values (e.g Gaussian kernels) the first eigenvector of the affinity matrix is a same sign vector.

3.1 Multiple class clustering

In the multiple class case the fact that the solution of the maximization problem above (Eq. 3) is given only up to a multiplication with a unitary matrix induces an inherent ambiguity. Given that Eq. 3 above is maximized by a set of orthogonal vectors $[a_1, a_2, ..., a_k]$, the final clustering should be the same for all other sets orthogonal vectors of the form $[a_1, a_2, ..., a_k]\, O$, where O is some $k \times k$ orthogonal matrix, as they maximize that score function as well (an alternative is to require additional conditions that make the solution unique).

Using a probability model, the most natural way to perform clustering would be to apply *model based clustering*. Following the spirit of [Zhong and Ghosh, 2004], we base our model-based clustering on affinities between every two data points that are derived from the probability model. To do so, we are interested not in the probability of the cluster membership given the data point, but in the probability of the data point being generated by a given cluster [Zhong and Ghosh, 2004], i.e., two points are likely to belong to the same cluster if they have the same profile of cluster membership.

This is given by the Bayes rule as $p(\Phi|i) \cong \frac{p(i|\Phi)p(\Phi)}{p(i)}$. We normalize the cluster membership probabilities such that the probability of generating each point by the k clusters is one – otherwise the scale produced by the probability of generating each point would harm the clustering. For example, all points with small $p(\Phi)$ would end up in one cluster. Each data point is represented by a vector q of k elements where element i is $q(i) = \frac{p(i|\Phi)/p(i)}{\sum_{j=1}^{k} p(j|\Phi)/p(j)}$.

The probability of cluster i membership given the point Φ_j is given by: $p(i|\Phi_j) =$

$(u_i^\top \Phi_j)^2 = (s_i v_i(j))^2$. The prior probability of cluster i is estimated by $p(i) = \int p(i,\Phi)\sigma(\Phi) = \int p(i|\Phi_j)p(\Phi_j)\sigma(\Phi) \cong \frac{1}{n}\sum_{j=1}^{n} p(i|\Phi_j) = u_i^T \Gamma_\Phi u_i = s_i^2$.

Hence the elements of the vector q_j, which represent a normalized version of the probabilities of the point Φ_j to be generated from the k clusters, are estimated as: $q_j(i) = \frac{(s_i v_i(j))^2/s_i^2}{\sum_{l=1}^{k}(s_l v_l(j))^2/s_l^2} = \frac{v_i(j)^2}{\sum_{l=1}^{k} v_i(j)^2}$.

Ignoring the problem of an unknown unitary transformation ("rotation"), we can compare two such vectors by using any affinity between probabilities. The most common choices would be the Kullback-Leibler (KL) divergence and the Hellinger distance [Devroye, 1987]. We prefer using the Hellinger distance since it is symmetric, and would lead to simpler solutions. The similarity between distributions, used in the Hellinger distance is simply *affinity*$(q_j, q_{j'}) = \sum_i \sqrt{q_j(i)}\sqrt{q_{j'}(i)}$. However, this affinity is not invariant to a unitary transformation of V. Below, we will discuss an algorithm that is invariant to the unitary transformation, and relate it to the Hellinger distances between those model-based distributions.

The NJW algorithm [Ng et al., 2001] is one of the most popular spectral clustering algorithms. It is very successful in practice [Verma and Meila, 2003], and was proven to be optimal in some ideal cases. The NJW algorithm considers the vectors $r_j(i) = \frac{v_i(j)}{(\sum_{l=1}^{k} v_l(j)^2)^{1/2}}$. The vectors r_j differ from the point-wise square root of the vectors q_j above, in that the numerator can have both positive and negative values. The Hellinger affinity above would be different from the dot product $r_j^\top r_{j'}$ if for some coordinate i the sign of $r_j(i)$ is different from the sign of $r_{j'}(i)$. In this case the dot product will always be lower than the Hellinger affinity.

The NJW algorithm clusters the r_j representation of the data points using k-

means clustering in $\mathcal{R}^k$. The NJW therefore finds clusters ($C_i, i = 1..k$) and cluster centers (c_i) as to minimize the sum of squared distances $\sum_{i=1}^{k} \sum_{j \in C_i} ||c_i - r_j||_2^2$. This clustering measure is invariant to the choice of basis of the space spanned by the first k eigenvectors of $\phi^\top \phi$. To see this it is enough to consider distances of the form $||(\sum_i \alpha_i r_i) - r_j||_2^2$, since the c_i's chosen by the k-means algorithm are linear combinations of the vectors r_i in a specific cluster. Assume that the subspace spanned by the first k eigenvectors of $\phi^\top \phi$ is rotated by some $k \times k$ unitary transformation O. Then each vector r_j would be rotated by O and become Or_j - there is no need to renormalize since O preserves norms. O preserves distances as well, and the α_i's are not dependent of O, so $||(\sum_i \alpha_i r_i) - r_j||_2^2 = ||(\sum_i \alpha_i O r_i) - O r_j||_2^2$.

For simplicity, allow us to remove the cluster center, which serves as a proxy in the k-means algorithm, and search for the cluster assignments which minimize: $\sum_{i=1}^{k} \sum_{j,j' \in C_i} ||r_j - r_{j'}||_2^2$. The Hellinger distance $D^2(q_j, q_{j'}) = \sum_i (\sqrt{q_j(i)} - \sqrt{q_{j'}(i)})^2$ is always lower than the $L2$ distance between r_j and $r_{j'}$ and hence, the above criteria bounds from above the natural clustering score: $\sum_{i=1}^{k} \sum_{j,j' \in C_i} D^2(q_j, q_{j'})^2$. Note that for most cases this bound is tight since points in the same cluster have similar vector representations.

3.2 Clustering Points of Any Norm

Above, we assumed that the norm of each input point is 1. Spectral clustering is usually done using Gaussian or heat kernels [Ng et al., 2001, Belkin and Niyogi, 2003]. For these kernels the norm of each input point in the feature space (the vector space in which the kernel is just a dot product [Schölkopf and Smola, 2002]).

A related remark is that in our model the cluster membership probability of the point Φ is the same as of the point $-\Phi$. In our model, it might be more appropriate to consider *rays* of the form $c\Phi$, c being some scalar. The norm-1 vector is just the representation of this ray. This distinction between rays and norm-1 vectors does not make much difference when considering kernels with non-negative values – all the vectors in the feature space are aligned to have a positive dot product.

In many situations, spectral clustering methods are applied to data points which do not have a norm of one, for example, when using the linear kernel or the polynomial kernel. In such cases, points which are of higher norm have a more significant effect on the resulting clustering. Therefore, it is natural to relate the norm of a point to the magnitude of effect it has on the final solution.

In order to incorporate the norm of the points into our probabilistic framework, we assume that all points belong to a sphere of radius R, i.e., all points in our data set have a norm smaller than R. This is a common assumption when analyzing kernel algorithms theoretically. Without loss of generality we assume that $R = 1$, otherwise, we can divide every point by R, or equivalently we can multiply the kernel function by R^{-2}.

The norm of the point Φ can be then viewed as a uniform prior on all the observations arising from this point. We suggest using the prior $\hat{p}(\Phi) = (\Phi^\top \Phi) = ||\Phi||^2$. The probability of any outcome of the clustering observation will be weighted by this prior. In probability $\hat{p}(\Phi)$ we get one of the outcomes, and in probability $1 - \hat{p}(\Phi)$ we get no outcome at all. This prior allows us to decompose every point to a norm one point times it's norm, or in other words, each point Φ is

viewed as the point $\Phi/||\Phi||$, with an importance proportional to the square of the norm of the point.

This point based prior allows us to apply the Born rule as is to points of any norm. This holds since [2] $\bar{p}(a_i|\Phi) = p(\Phi)p(a_i|\Phi/||\Phi||) = (\Phi^\top\Phi)\frac{(a_i^\top\Phi)^2}{(\Phi^\top\Phi)} = (a_i^\top\Phi)^2$. Therefore, the form of the probability rule for points of any norm is the same as the Born rule (which is originally limited to norm-1 vectors). The implication is that by using this prior, all our results above hold for vectors of any norm.

Note that the point based prior is self consistent in the following sense: the sum over any resolution of the identity $\{a_i\}$ of the probabilities $\bar{p}(a_i|\Phi)$ is just the prior $p(\Phi)$:

$$\sum_i \bar{p}(a_i|\Phi) = \sum_i (a_i^\top\Phi)^2 = \Phi^\top \sum_i (a_i a_i^\top)\Phi = \Phi^\top\Phi$$

3.3 Normalizing the affinity matrix

So far, the affinity matrix was not normalized, i.e., we did not use the Laplacian $L = D - A$ or the "normalized Laplacian" $N = D^{-1/2}AD^{-1/2}$ (A being our kernel $A = \phi^\top\phi$, and D being the diagonal matrix whose entries contain the sum of the rows of A). These modified kernel matrices show preferable performance in many practical situations compared to the original kernel matrix (although this is not always the case). The normalized Laplacian N is required by the views of spectral clustering presented in [Shi and Malik, 2000, Meila and Shi, 2000,

[2] We abuse the notation a bit so that the relation between $p(i|\Phi)$ and a_i would be more clear.

Belkin and Niyogi, 2003], however, we choose to view these strictly as applying normalization to the affinity matrices. This is the practitioner's point of view expressed in Y. Weiss' tutorial on spectral clustering given at NIPS 2002.

Normalization is usually done so that the magnitude of the individual effect of different data points on the final solution would be similar. For example, when solving a homogeneous system of linear equations using the least squares loos function, the norm of each equation (represented as row of the matrix A in the system $Ab = 0$) determines the magnitude of its effect on the final solution. Therefore, it is a common practice to normalize each row of the matrix A to have a norm of 1 before solving for b. Put differently, the distance of each equation from zero becomes uniform.

In out framework, we cluster using the Hellinger affinities, therefore the normalization scheme should make sure that all of our points have the same Hellinger affinity to a reasonable anchor point. We choose to have this anchor point at the mean posterior distribution, where the mean is defined based on the simple observation that the Hellinger affinities define a positive definite kernel. In the feature space associated with this kernel each probability is represented by the point wise square root of its value. Since the Hellinger affinity is linear in this feature space, we can replace the affinity of the mean with the mean of the affinities.

Given a set of data points $\{\Phi_i\}_{i=1}^n$ and any von-Neumann measurement on our data points (an orthogonal basis of the vector space where the kernel function is a dot product) $\{u_i\}_{i=1}^d$, the posterior distributions of the outcomes are given as $p(i|\Phi_j) = (\Phi_j^\top u_i)^2$. The Hellinger affinity between the posterior distribution of a

point Ψ and the mean posterior distribution is given by $1/n \sum_i \sum_j |(\Psi^\top u_i)||(\Phi_j^\top u_i)|$. This expression is bounded from below by $1/n \sum_i (\Psi^\top \Phi_j)$ (it is important to bound from below in order not to normalize too much). For every point in the dataset Φ_i, this is exactly the value d_i/n where $d_i = \sum_j (\Phi_i^\top \Phi_j)$ is sum of the $i-th$ row of the kernel matrix $\phi^\top \phi$, where $\phi = [\Phi_1|\Phi_2|...|\Phi_n]$. Note that this sum is dominated by contributions from points Φ_j which are similar to Φ_i. For such points we can expect the above bound to be tight.

To summarize the above finding: kernels that have a constant row sum are considered to be normalized. Consider the Laplacian $L = D - A$. The Laplacian has a row sum of zero, however, by construction, the most relevant information is stored in its lower magnitude eigenvectors, and does not fit our framework as is. Still, it is a common practice in spectral clustering to consider the matrix $\hat{L} = d_{max} I + A - D$ instead, where $d_{max} = \max_i d_i$. This matrix is positive definite, has the relevant information in its largest eigenvectors, and has the same eigenvectors as the laplacian (but in reverse order). Note that the row sum of $\hat{L}$ is constant and equals d_{max}.

Consider now the multiplicative normalization $N = D^{-1/2} A D^{-1/2}$. The row sum of this matrix is not constant. However, as noted by Zass and Shashua [2005], the matrix N is closer to be double stochastic than the matrix A, and can be seen as the first iteration of a process $A^{k+1} \leftarrow D^{-1/2} A^k D^{-1/2}$ that converges to a double stochastic matrix. Hence, from our perspective, the matrix N is more normalized version of A. The reason why N might be preferable to $\hat{L}$ is that N is closer to the original A, making it more loyal to the true underlying similarities between

the data points.

3.4 Out of sample clustering

Our approach to spectral clustering is model-based, therefore it is very natural to apply it to any new point without re-clustering. This is in contrast to the original NJW algorithm and to related algorithms such as the Laplacian Eigenmaps [Belkin and Niyogi, 2003] algorithm. To overcome this problem, a term of regularization was recently added to the latter [Belkin et al., 2004]. A more general treatment was proposed by Bengio et al. [2004] who suggested the use of the Nyström formula in order to estimate the approximated eigenvector values for the new points. The model based approach we propose leads to exactly the same solution.

In order to apply the clustering algorithm to a new point Ψ, we need to compute the model based posterior probabilities associated with it. These probabilities can be expressed as functions of probabilities of the form $(u_i^\top \Psi)^2$, where u_i is an eigenvector of the input data's covariance matrix. Since the points in our kernel space can have a very high or even infinite dimension, we cannot compute the eigenvectors u_i directly. However, we can use the equality used in kernel PCA: $u_i = (1/s_i)\phi v_i$ (as before, v_i are eigenvectors of the kernel $\phi^\top \phi$, s_i^2 are eigenvalues). In order to compute the dot product $u_i^\top \Psi = (1/s_i) v_i^\top \phi^\top \Psi$ we only need to know the elements of the vector $\phi^\top \Psi$. This vector contains the value of the kernel function computed between every element in our training set (the columns of ϕ) and the new example Ψ.

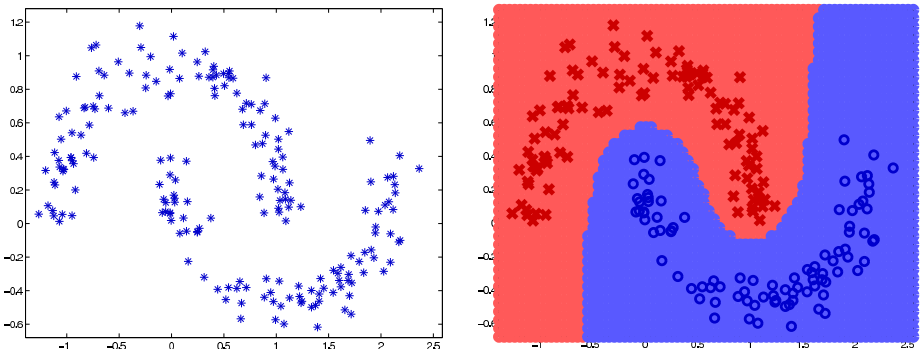

Figure 1: This figure demonstrates out-of-sample classification using the Born rule spectral clustering scheme. (left) the original training dataset of the two moons example of [Belkin et al., 2004]. (right) the resulting clustering together with an out-of-sample classification of the nearby 2D plane.

The above out-of-sample classification scheme is demonstrated in Fig. 1. We used the two moons example of [Belkin et al., 2004], and a Gaussian kernel similar to the one used there.

4 Euclidean-distance based Algorithms

The RCA algorithm [Bar-Hillel et al., 2005] is a very effective algorithm for learning Mahalanobis metrics from partially labeled data. Given groups of data points that are known to belong to the same class, the algorithm finds a metric in which the distances within these groups are minimized.

Let Φ_{ji} be data point i in group j, $j = 1...g$. Let the number of data points in group j be n_j and let the total number of points be $n = \sum_j n_j$. Let $\bar{\Phi}_j = 1/n_j \sum_{i=1}^{n_j} \Phi_{ji}$ be the mean of the points in group j. Note that even though this view was not presented in [Bar-Hillel et al., 2005], the data points can be given

either directly or through a kernel function (the full discussion on how to compute the RCA transform in kernel spaces is outside the scope of this work). The RCA algorithm produces the Mahalanobis metric B which optimizes the following optimization problem:

$$\min_{B} \sum_{j=1}^{g} \sum_{i=1}^{n_j} ||\Phi_{ji} - \bar{\Phi}_j||_B^2 \quad s.t. \ |B| \geq 1 \ , \tag{4}$$

where $||\Psi||_B^2 = \Psi^\top B \Psi$ and $|B|$ denotes the determinant of B. The constrain on the determinant is required, since the lower the magnitude of the matrix B, the lower the score is. The optimization of that problem is given up to scale by the matrix $\hat{C}^{-1}$, where $\hat{C}$ is the matrix which contains the sum of all the covariance matrices of the g groups.

When working in high dimensions, with a limited number of training points, regularization is necessary. Bar-Hillel et al. propose to reduces the dimension of the data as part of the RCA procedure. An alternative regularization technique would be to add the requirement that all vectors that differ only by a small amount in one coordinate of the feature space would belong to the same class. This form of regularization amounts to adding the identity matrix times a constant to the matrix $\hat{C}$ prior to inverting it. This simple regularization is compatible with the explanation below (it simply adds pairs of virtual points), and can be incorporated into the kernel version of RCA.

Our view of Relevant Component Analysis is as follows. First, we view the optimization problem is Eq. 4 as an approximation of the following optimization

problem

$$\min_{B} \max_{j,i} ||\Phi_{ji} - \bar{\Phi}_j||_B \quad s.t. \ |B| \geq 1 \ ,$$

which is much harder to optimize for. This view is based on the fact that the sum of squares is dominated by the largest values in the sum. The entity $||\Phi_{ji} - \bar{\Phi}_j||_B$ is then interpreted as the statistical distinguishability between a point and its group's center.

Recall that Wootters [1981] showed that the underlying probability distribution behind the Euclidean distance is the Born rule. His analysis was based on his own (very intuitive) definition of statistical distinguishability and on the "best case scenario" in which the measurement is selected to emphasize the differences between the two vectors. The discussion below applies to every measurement, and employs the Hellinger distance as the statistical distinguishability.

Below we assume that all data points have norms which are bounded by 1, otherwise we normalize all of the points. Let Φ and Ψ be two points of dimension d (possible infinite) and Let $\{u_i\}_{i=1}^{d}$ be any von-Neumann measurement on our data points (an orthogonal basis of the vector space where the kernel function is a dot product). The two posterior distributions $p(i|\Phi) = (u_i^\top \Phi)^2$ and $p(i|\Psi) = (u_i^\top \Psi)^2$ have a Hellinger distance of $\sqrt{\sum_i (|(u_i^\top \Phi)| - |(u_i^\top \Psi)|)^2}$. This distance is bounded from above by $\sqrt{\sum_i ((u_i^\top \Phi) - (u_i^\top \Psi))^2} = ||U^\top \Phi - U^\top \Psi|| = ||\Phi - \Psi||$, where U is a $d \times d$ matrix with u_i as columns.

The Born rule view of the Euclidean distance we have suggested is that it is an upper bound to the statistical distance between the posterior probabilities

associated with any possible measurement. This bound is expected to be tight for nearby points.

5 Conclusions

The spectral clustering techniques are an extremely popular family of methods in unsupervised learning. Still, the explanations that exist to their success are partial. In this work we used Gleason's theorem to show that for the very natural model in which all data-points are represented as vectors, and in which the clusters are also represented by unit vectors, the Born rule is the appropriate probability rule. Using the Born rule, we were able to derive and interpret statistically the most successful spectral clustering practices, and to extend the framework to out-of-sample clustering.

Moreover, we provided a simple result in the flavor of Wootters' result, linking Euclidean distances and statistical distinguishability. These statistical views of the $L2$ metric can serve as probabilistic justifications to the common use of this metric in learning, adding to the understanding of this practice.

Understandably, our proposed model cannot be "the right model", excluding other models from being valid. Still, we have shown that it is well founded, consistent with common practices, useful in motivating those practices and maybe even elegant.

References

Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *JMLR*, 6(Jun):937–965, 2005.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

M. Belkin, P. Niyogi, , and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. Technical Report TR-2004-06, Uni. Chicago, IL, 2004.

Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004.

I.D. Coope and P.F. Renaud. Trace inequalities with applications to orthogonal regression and matrix nearness problems. Technical Report UCDMS 2000/17, Uni. Canterbury, NZ, 2000.

D. Deutsch. Quantum theory of probability and decisions. In *Proc. of the Roy. Soc. of London A455*, pages 3129–3137, 1999.

L. Devroye. *A Course in Density Estimation*. Birkhauser Boston Inc., 1987. ISBN 0-8176-3365-0.

L. Devroye, L. Györfi, and G. Lugosi. *A Probababilistic Theory of Pattern Recognition*. Springer, 1996.

A. M. Gleason. Measures on the closed subspaces of a hilbert space. *J. Math. & Mechanics*, 6:885–893, 1957.

D. Horn and A. Gottlieb. Algorithms for data clustering in pattern recognition problems based on quantum mechanics. *Physical Review Letters*, 88(1):885–893, 2002.

Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *NIPS*, pages 873–879, 2000.

A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

P. Perona and W.T. Freeman. A factorization approach to grouping. In *ECCV*, 1998.

S. Saunders. Derivation of the born rule from operational assumptions. In *Proceedings of the Royal Society A*, volume 460, pages 1–18, 2004.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

Janne Sinkkonen, Samuel Kaski, and Janne Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In *ECML '02: Proceedings of*

the 13th European Conference on Machine Learning, pages 418–430, London, UK, 2002. Springer-Verlag. ISBN 3-540-44036-4.

D. Verma and M. Meila. A comparison of spectral clustering algorithms. Technical Report UW CSE 03-05-01, Michigan State Uni., 2003.

W.K. Wootters. Statistical distance and hilbert space. *Physical Review D.*, 23: 357–362, 1981.

S. Youssef. Quantum mechanics as complex probability theory. *Modern Physics Letters*, A(9), 1994.

R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *International Conference on Computer Vision*, 2005.

S. Zhong and J. Ghosh. A unified framework for model-based clustering. *JMLR*, 4(6):1001–1037, 2004.