A Free Energy Based Approach for Distance Metric Learning

Sho Inaba
Computational Sciences PhD program
University of Massachusetts Boston
Boston, USA
Sho.Inaba001@umb.edu

Carl T. Fakhry
Department of Computer Science
University of Massachusetts Boston
Boston, USA
carlfakhry@hotmail.com

Rahul V. Kulkarni
Department of Physics
University of Massachusetts Boston
Boston, USA
Rahul.Kulkarni@umb.edu

Kourosh Zarringhalam*
Department of Mathematics
University of Massachusetts Boston
Boston, USA
kourosh.zarringhalam@umb.edu

ABSTRACT

We present a reformulation of the distance metric learning problem as a penalized optimization problem, with a penalty term corresponding to the von Neumann entropy of the distance metric. This formulation leads to a mapping to statistical mechanics such that the metric learning optimization problem becomes equivalent to free energy minimization. Correspondingly, our approach leads to an analytical solution of the optimization problem based on the Boltzmann distribution. The mapping established in this work suggests new approaches for dimensionality reduction and provides insights into determination of optimal parameters for the penalty term. Furthermore, we demonstrate that the metric projects the data onto direction of maximum dissimilarity with optimal and tunable separation between classes and thus the transformation can be used for high dimensional data visualization, classification, and clustering tasks. We benchmark our method against previous distance learning methods and provide an efficient implementation in an R package available to download at: https://github.com/kouroshz/fenn

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; **Classification and regression trees**; Model verification and validation; • **Mathematics of computing** → *Convex optimization.*

KEYWORDS

distance metric learning, dimensionality reduction, high-dimensional data visualization

ACM Reference Format:
Sho Inaba, Carl T. Fakhry, Rahul V. Kulkarni, and Kourosh Zarringhalam. 2019. A Free Energy Based Approach for Distance Metric Learning. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330975

1 INTRODUCTION

The main objective of distance metric learning is to 'optimally' transform the data in order to bring similar points closer to each other while keeping the distance between the dissimilar points bounded away from zero. Distance metric learning can be traced back to the works in [2, 8, 11, 12]. However the formulation in [26] is generally regarded as the modern origin of the approach. In [26], the transformation is identified by minimizing the Mahalanobis distance between similar pairs with the constraint that the distance between the dissimilar pairs is bounded below by a constant. Since its original introduction, several researchers have developed a wealth of algorithms for distance metric learning with different flavors. Distance learning methods are designed to improve the performance of learning algorithms such as clustering with k-means [1, 5, 16, 26] or classification with k-NN [9, 17, 24, 25]. While the intended applications of these methods can be diverse, the overall schemes show considerable overlap. We refer the reader to [3], [14] and [22] for comprehensive surveys on distance metric learning.

As noted by the authors in [26], if the same metric is used to measure the distance between points in each class, the optimal solution will project the data onto a line, which is not suitable for most of the intended applications. Modifications to the original problem are typically applied to avoid such 'degenerate' solutions. For instance in [26], the authors propose using a transformation (square root) of the metric in the constraint part of the problem. Similarly, in [27], the authors propose a method called DML-eig, which also applies modifications to the metric. Another popular approach is the Large Margin Nearest Neighbor (LMNN) developed in [24, 25, 27], wherein local information is used to learn a global transformation matrix. In [5] and [23], authors propose information theoretic methods for metric learning.

In this paper, we reformulate the metric learning optimization problem in a way that does not involve modifications to the metric and keeps the metric consistent for measuring distances between points, regardless of class. To avoid trivial solutions, we propose to add a penalty term that is based on the Von Neumann entropy. Our choice of the entropy-based penalty term is motivated by a natural connection to the mathematical formalism used for Quantum

[*] To whom correspondence should be addressed.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3330975

Statistical Mechanics (QSM). With this formulation, we develop a mapping of the optimization problem to minimization of the Helmholtz free energy in QSM. Establishing such a mapping allows for approaches and insights from statistical mechanics to be used for solving the optimization problem and for other applications. Importantly, we provide an analytical solution for the optimization problem and therefore, unlike many widely used distance learning algorithms, our method does not rely on gradient-descent based solvers, which are well known for their high sensitivity to scaling and learning rates. The analytical approach also avoids costly computations such as eigenvalue decomposition in each iteration. Further, we demonstrate that our analytical solution identifies the most discriminant features that result in maximum separation between classes. We also establish a connection between our analytical solution and multi-class Linear Discriminant Analysis (LDA) and show that our approach can be used to make improvements to multi-class LDA by providing tunable scaling that leads to controlled separation of classes, while the connection to QSM provides physical intuition for setting the scaling for optimal separation.

In the following sections, we establish the theoretical basis for our method. The primary focus of this paper is on the theoretical insights and generalizations based on the mappings established to statistical mechanics and the geometrical aspects of the metric. To illustrate the effectiveness of the proposed approach, we apply it for supervised classification of several traditional UCI datasets and include comparisons with results from previous distance metric learning methods. We close the paper with a brief conclusion and discussion of future directions.

2 METRIC LEARNING MODEL AND RELATION TO OTHER METHODS

Let $\{(x_i, \ell(x_i))\}_{i=1}^n$ be a set of labeled training data with sample points $x_i = (x_{i1}, \cdots, x_{ip})^T \in \mathbb{R}^p$ and their labels $\ell(x_i)$. Let $\mathcal{S} = \{(x,y)|\ell(x) = \ell(y)\}$ and $\mathcal{D} = \{(x,y)|\ell(x) \neq \ell(y)\}$ be the sets of similar and dissimilar training examples respectively. Let S_+^d denote the cone of positive semi-definite matrices. Our goal is to find an optimal (pseudo) metric that directly minimizes the distance between similar pairs while keeping the dissimilar pairs apart. As a starting point, consider the optimization problem originally proposed in [26]:

$$\begin{aligned} \min_{M \in S_+^d} & \sum_{(x,y)\in\mathcal{S}} (x-y)^T M (x-y) \\ \text{s.t.} & \sum_{(x,y)\in\mathcal{D}} (x-y)^T M (x-y) \geq 1. \end{aligned} \tag{1}$$

Here $(x-y)^T M(x-y) = d_M^2(x,y)$ is a pseudo metric. As noted by the authors in [26], the solution of this problem projects the data onto a line. To avoid this, Xing et al. [26] propose solving the following alternative problem:

$$\begin{aligned} \max_{M \in S_+^d} & \sum_{(x,y)\in\mathcal{D}} d_M(x,y) \\ \text{s.t.} & \sum_{(x,y)\in\mathcal{S}} d_M^2(x,y) \leq 1. \end{aligned} \tag{2}$$

Several other reformulations of this problem have been proposed. Most closely related to our method is the formulation proposed by Ying and Li [27]:

$$\begin{aligned} \max_{M \in S_+^d} & \min_{(i,j)\in\mathcal{D}} d_M^2(x,y) \\ \text{s.t.} & \sum_{(x,y)\in\mathcal{S}} d_M^2(x,y) \leq 1. \end{aligned} \tag{3}$$

Ying and Li show that the problem in (3) can be reformulated as a generalized eigenvalue optimization, and subsequently show that LMNN [24, 25] can also be formulated in a similar fashion as an eigenvalue optimization problem. Here we present an alternative approach to recast the original problem in (1) as an eigenvalue optimization problem with the important distinction that our formulation has an analytic solution. We will also show that our method has a natural mapping to statistical mechanics, thereby providing a bridge between a class of distance metric learning problems to a class of well established physical problems. Moreover, we will provide a geometric interpretation of our method and demonstrate that the learned metric identifies directions of maximum separation between classes and generalizes LDA.

3 VON NEUMANN ENTROPY PENALIZED DISTANCE METRIC LEARNING

To begin, we first normalize the original problem with the number of classes and numbers of samples in each class. This normalization helps to avoid problems based on combining unbalanced datasets and it provides a consistent framework for our results regardless of the number of classes in the training data. Let N be the number of classes and let n_x be the number of samples in the class containing the sample point x. Then, we consider the normalized problem:

$$\begin{aligned} \min_{M \in S_+^d} \frac{1}{N} & \sum_{(x,y)\in\mathcal{S}} \frac{1}{n_x^2} (x-y)^T M (x-y) \\ \text{s.t.} \frac{1}{N(N-1)} & \sum_{(x,y)\in\mathcal{D}} \frac{1}{n_x n_y} (x-y)^T M (x-y) \geq 1. \end{aligned} \tag{4}$$

Next, we introduce some notation and reformulate our optimization problem. Let $\tau = (x,y)$ represent a pair of samples and let $X_\tau = (x-y)(x-y)^T$ be the un-normalized projection onto $x-y$. Then, we have $(x-y)^T M(x-y) = tr(X_\tau^T M) = \langle X_\tau, M\rangle$. Define

$$X_{\mathcal{D}} = \frac{1}{N(N-1)} \sum_{\tau\in\mathcal{D}} \frac{1}{n_x n_y} X_\tau, \tag{5}$$

and

$$X_{\mathcal{S}} = \frac{1}{N} \sum_{\tau\in\mathcal{S}} \frac{1}{n_x^2} X_\tau. \tag{6}$$

Without loss of generality we may assume that $X_{\mathcal{D}}$ is a positive definite matrix ($X_{\mathcal{D}}$ is symmetric positive semi-definite and if needed, we may replace $X_{\mathcal{D}}$ by $X_{\mathcal{D}} + \delta I$ for a small value of δ to make it a positive definite matrix). We now reformulate the optimization problem in (4), following the same approach as in [27].

Lemma 3.1. *The inequality constraint in 4 can be replaced with an equality constraint to obtain the following optimization problem:*

$$\min_{M \in S_+^d} \langle X_{\mathcal{S}}, M\rangle \quad \textit{s.t.} \quad \langle X_{\mathcal{D}}, M\rangle = 1. \tag{7}$$

PROOF. Let M^* be the optimal solution of problem (4) and define $M_0^* = \frac{M^*}{\langle X_{\mathcal{D}}, M^* \rangle}$. Then $\langle X_{\mathcal{D}}, M_0^* \rangle = 1$. Moreover $\langle X_{\mathcal{S}}, M_0^* \rangle = \frac{1}{\langle X_{\mathcal{D}}, M^* \rangle} \langle X_{\mathcal{S}}, M^* \rangle \leq \langle X_{\mathcal{S}}, M^* \rangle$. Hence M_0^* is also an optimal solution. □

For a given positive semi-definite matrix M, let $S = X_{\mathcal{D}}^{1/2} M X_{\mathcal{D}}^{1/2}$ and let $\tilde{x} = X_{\mathcal{D}}^{-1/2} x$. Define

$$\tilde{X}_\tau = (\tilde{x} - \tilde{y})(\tilde{x} - \tilde{y})^T = \left(X_{\mathcal{D}}^{-1/2}(x - y)\right)\left(X_{\mathcal{D}}^{-1/2}(x - y)\right)^T, \quad (8)$$

$$\tilde{X}_{\mathcal{S}} = \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} \tilde{X}_\tau. \quad (9)$$

We have the following theorem.

THEOREM 3.2. *The constrained optimization problem in* (7) *can be reformulated as follows:*

$$\min_{S \in \mathcal{P}} \langle \tilde{X}_{\mathcal{S}}, S \rangle, \quad (10)$$

where $\mathcal{P}$ is the spectrahedron $\mathcal{P} = \{S \in S_+^d | tr(S) = 1\}$.

PROOF. Note that

$$\tilde{X}_\tau = (X_{\mathcal{D}}^{-1/2}(x - y))(X_{\mathcal{D}}^{-1/2}(x - y))^T = X_{\mathcal{D}}^{-1/2} X_\tau X_{\mathcal{D}}^{-1/2}. \quad (11)$$

It is straightforward to show that $\langle \tilde{X}_{\mathcal{S}}, S \rangle = \langle X_{\mathcal{S}}, M \rangle$ and that $\langle X_{\mathcal{D}}, M \rangle = tr(X_{\mathcal{D}}^{1/2} M X_{\mathcal{D}}^{1/2}) = tr(S)$. Hence $\langle X_{\mathcal{D}}, M \rangle = 1$ if and only if $tr(S) = 1$ and we obtain a simple optimization problem. □

As mentioned before, the optimal solution of the above problem will project the data points onto a line. As such, we propose the following smoothed optimization problem that will avoid such solutions:

$$\begin{aligned}
&\min_{S \in \mathcal{P}} \langle \tilde{X}_{\mathcal{S}}, S \rangle - \mu H(S) = \\
&\min_{S \in \mathcal{P}} tr(\tilde{X}_{\mathcal{S}} S) - \mu H(S) = \\
&\min_{S \in \mathcal{P}} \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} (\tilde{x} - \tilde{y})^T S (\tilde{x} - \tilde{y}) - \mu H(S), \quad (12)
\end{aligned}$$

where $H(S)$ is the Von Neumann entropy of S and is defined by $H(S) = -\sum_{i=1}^p \lambda_i \log(\lambda_i)$ and μ is a smoothing parameter. Here the λ_i's are the eigenvalues of the matrix S. Note that since $S \in S_+^d$ and $tr(S) = 1$, this will necessarily impose the constraint that $\lambda_i > 0$ and $\sum_{i=1}^p \lambda_i = 1$.

3.1 Analytical solution of the optimization problem

Optimization problems for distance metric learning are typically addressed using numerical approaches. However, given our use of the Von Neumann Entropy $H(S)$ as a smoothing term, an analytical expression for the solution of the optimization problem can be derived. In the following, we present two approaches for obtaining the analytical solution. One approach involves a direct solution of the optimization problem using Lagrange multipliers, whereas the other method involves a mapping to statistical mechanics. Both approaches are instructive and lead to insights for further analysis as presented below.

3.1.1 Analytic solution using Lagrange multipliers. The constraint $S \in \mathcal{P}$ is equivalent to S symmetric, $\lambda_k \geq 0$, and $\sum_{k=1}^p \lambda_k = 1$. Define the Lagrangian

$$\begin{aligned}
\mathcal{L}(S) = \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} (\tilde{x} - \tilde{y})^T S (\tilde{x} - \tilde{y}) + \mu \sum_{k=1}^p \lambda_k \log(\lambda_k) \\
+ \xi \Big(\sum_{k=1}^p \lambda_k - 1\Big),
\end{aligned}$$

where ξ is the Lagrange multiplier. We will make use of the following lemma.

LEMMA 3.3. *The derivative of $H(S)$ with respect to S is given by the following expression:*

$$\nabla_S [H(S)] = \sum_{k=1}^p \nabla_S [\lambda_k \log(\lambda_k)] = \sum_{k=1}^p (1 + \log(\lambda_k)) u_k u_k^T$$

where the u_k's are the eigenvectors of $\tilde{X}_{\mathcal{S}}$.

PROOF. Let $s = s_{ij}$ denote the (i, j)-th entry of S and let $S = U \Lambda U^T$ be the spectral decomposition of S. Then,

$$\frac{\partial \Lambda}{\partial s} = U^T \frac{\partial S}{\partial s} U - U^T \frac{\partial U}{\partial s} \Lambda - \Lambda \frac{\partial U^T}{\partial s} U \quad (13)$$

Note that $\left[U^T \frac{\partial S}{\partial s} U\right]_{xy} = \left[U_{ix} U_{iy}\right]$, and hence $\frac{\partial \lambda_k}{\partial s} = U_{ik} U_{jk} = \left[u_k u_k^T\right]_{ij}$ and $\nabla_S \lambda_k = u_k u_k^T$. Examining the diagonal entries of each sides of the equation (13) and observing that the last two terms have diagonal elements equal to 0, we get that:

$$\begin{aligned}
\nabla_S \left[\sum_{k=1}^p \lambda_k \log(\lambda_k)\right] &= \sum_{k=1}^p \nabla_S [\lambda_k \log(\lambda_k)] \\
&= \sum_{k=1}^p (1 + \log(\lambda_k)) u_k u_k^T.
\end{aligned}$$

□

THEOREM 3.4. *The solution S to the optimization problem in* (10) *is given by $S = \sum_{i=1}^p \lambda_i u_i u_i^T$ where λ_i is the i-th eigenvalue given by:*

$$\lambda_i = \frac{e^{-\left(u_i^T \tilde{X}_{\mathcal{S}} u_i\right)/\mu}}{\sum_{i=1}^p e^{-\left(u_i^T \tilde{X}_{\mathcal{S}} u_i\right)/\mu}}, \quad (14)$$

and its corresponding eigenvector u_i is the i-th eigenvector of $\tilde{X}_{\mathcal{S}}$.

PROOF. First, we differentiate the first part of $\mathcal{L}(S)$ with respect to S, we get that

$$\begin{aligned}
\nabla_S \left[\frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} (\tilde{x} - \tilde{y})^T S (\tilde{x} - \tilde{y})\right] &= \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} \nabla_S \left[(\tilde{x} - \tilde{y})^T S (\tilde{x} - \tilde{y})\right] \\
&= \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} (\tilde{x} - \tilde{y})(\tilde{x} - \tilde{y})^T \\
&= \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} \tilde{X}_\tau = \tilde{X}_{\mathcal{S}}.
\end{aligned}$$

Next, using the computed derivative for $H(S)$ in Lemma 2 we get:

$$\nabla_S \mathcal{L}(S) = \tilde{X}_{\mathcal{S}} + \mu \sum_{k=1}^{p} (1 + \log(\lambda_k)) u_k u_k^T + \xi \sum_{k=1}^{p} u_k u_k^T .$$

Setting the Lagrangian equal to 0 and multiplying each side by u_i, we get that $\tilde{X}_{\mathcal{S}} u_i = -\mu(1 + \log(\lambda_i))u_i - \xi u_i$. In particular, this means that the u_i's are eigenvectors of the symmetric matrix $\tilde{X}_{\mathcal{S}}$. Moreover, we have that $u_i^T \tilde{X}_{\mathcal{S}} u_i = (-\xi - \mu) - \mu \log(\lambda_i)$. Solving this equation for λ_i and applying the constraint $\sum_{k=1}^{p} \lambda_k = 1$, we arrive at the following solution for λ_i

$$\lambda_i = \frac{e^{-\left(u_i^T \tilde{X}_{\mathcal{S}} u_i\right)/\mu}}{\sum_{i=1}^{p} e^{-\left(u_i^T \tilde{X}_{\mathcal{S}} u_i\right)/\mu}}. \tag{15}$$

□

In other words, the previous theorem shows that the optimal solution of the eigenvalues is given by the Boltzmann distribution from which S can be constructed using the spectral decomposition as $S = \sum_{i=1}^{p} \lambda_i u_i u_i^T$. The optimal distance between the points in the transformed space is then calculated by

$$\begin{aligned} d_S^2(\tilde{x}, \tilde{y}) &= (\tilde{x} - \tilde{y})^T S (\tilde{x} - \tilde{y}) \\ &= [S^{1/2}(\tilde{x} - \tilde{y})]^T [S^{1/2}(\tilde{x} - \tilde{y})] \\ &= [S^{1/2} X_D^{-1/2}(x - y)]^T [S^{1/2} X_D^{-1/2}(x - y)] \end{aligned} \tag{16}$$

where $S^{1/2} = \sum_i \lambda_i^{1/2} u_i u_i^T$. Hence the optimal distance is obtained by first transforming the data points by $S^{1/2} X_D^{-1/2}$ and then computing the Euclidean distance in the transformed space. Note that S, $S^{1/2}$ and $\tilde{X}_{\mathcal{S}}$ share common eigenvectors. In the next section, we will give a physical interpretation for these quantities.

3.1.2 Analytical solution using the mapping to statistical mechanics. The appearance of the Boltzmann distribution strongly indicates that the optimization problem can be mapped on to a problem in statistical mechanics. Establishing such a mapping may suggest analytical solutions for a broader class of optimization problems that are currently addressed numerically. In the following, we establish such a mapping for the distance metric optimization problem. We outline the mapping from the original optimization problem to the formalism for systems studied in quantum statistical mechanics (QSM). Consider an *ensemble* of systems in the same macroscopic or thermodynamic state. In QSM, the central quantity of interest is the density matrix ρ, which contains information about the ensemble probabilities assigned to different microstates of the system in equilibrium. For a system in equilibrium at fixed temperature T, the density matrix can be obtained by minimizing the Helmholtz Free energy [13] which is given by

$$F[\rho] = tr(\rho \hat{\mathcal{H}}) - TH[\rho], \tag{17}$$

where the matrix $\hat{\mathcal{H}}$ represents the Hamiltonian of the system and $H[\rho]$ is the Von Neumann entropy. Note that we have set the Boltzmann constant $k_B = 1$. The eigenvectors of $\hat{\mathcal{H}}$ form a complete basis set and the eigenvalues correspond to allowed energy values for microstates of the system.

To connect with the original optimization problem, we make the identification $\hat{\mathcal{H}} = \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} (\tilde{x} - \tilde{y})(\tilde{x} - \tilde{y})^T = \tilde{X}_{\mathcal{S}}$. Correspondingly, the free energy $F[\rho] = tr(\rho \hat{\mathcal{H}}) - TH[\rho]$ can be expressed as:

$$F[\rho] = tr\left(\frac{\rho}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} (\tilde{x} - \tilde{y})(\tilde{x} - \tilde{y})^T \right) - TH[\rho]. \tag{18}$$

With the above mapping, minimization of the free energy is seen to be identical to the original optimization problem (Eq. (12)) with the identification $\rho = S$. Note that one of the properties of the density matrix is that $tr(\rho) = 1$, consistent with the constraint on S in the original problem. The density matrix ρ that minimizes the free energy can be derived using a unitary transformation to the basis set that diagonalizes the Hamiltonian $\hat{\mathcal{H}}$. In this basis, the density matrix ρ is also a diagonal matrix and its eigenvalues (λ_i) are given by the Boltzmann distribution, as obtained in the previous section. Correspondingly, the mapping developed can be summarized as follows: There is a one-to-one correspondence between the distance metric learning problem in Eq. (12) and the problem in QSM given in Eq. (18) i.e:

$$\min_{\rho} \; tr(\rho \hat{\mathcal{H}}) - TH[\rho] \iff \min_{S \in \mathcal{P}} tr(S \tilde{X}_{\mathcal{S}}) - \mu H(S).$$

3.2 Geometric interpretation

As we saw in the previous section, the analytical solution involves the eigenvalues and eigenvectors of the Hamiltonian matrix derived from the data, namely $\tilde{X}_{\mathcal{S}}$. In this section, we develop geometric interpretations for the terms $X_{\mathcal{S}}, X_{\mathcal{D}}$ and $\tilde{X}_{\mathcal{S}}$ to gain further insight into the geometry of the transformation. Let m_i be the mean of the i-th cluster, let C_i denote the covariance matrix of i-th cluster, and let C_m denote the covariance matrix of cluster centroids. Then,

$$\begin{aligned} X_{\mathcal{S}} &= \frac{1}{N} \sum_{\tau \in \mathcal{S}} \frac{1}{n_x^2} X_\tau \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_x^2} \sum_{j=1}^{n_x} \sum_{k=1}^{n_x} (x_j x_j^T - x_j x_k^T) \\ &= \frac{1}{N} \sum_{i=1}^{N} (E[x_i x_i^T] - m_i m_i^T) = \frac{1}{N} \sum_{i=1}^{N} C_i \end{aligned}$$

This implies that $X_{\mathcal{S}}$ is the 'mean' covariance representing an estimated covariance for all clusters. Similarly, it can be shown that $X_{\mathcal{D}} = X_{\mathcal{S}} + \frac{n}{n-1} C_m$. The combination of the mean covariance matrix $X_{\mathcal{S}}$ and the covariance matrix of the centroids C_m in $X_{\mathcal{D}}$ leads to the following theorem.

Theorem 3.5. *Let $u_1, u_2, \cdots, u_p$ be the eigenvectors of $\tilde{X}_{\mathcal{S}}$ with corresponding eigenvalue $\sigma_1, \sigma_2, \cdots, \sigma_p$ sorted in increasing order. Let $d = \min\{p, N-1\}$. If $X_{\mathcal{S}}$ is nonsingular, then $\sigma_1 \le \sigma_2 \le \cdots \le \sigma_d < 1$ and $\sigma_{d+1} = \sigma_{d+2} = \cdots = \sigma_p = 1$.*

Proof.

$$\begin{aligned} \tilde{X}_{\mathcal{S}} &= X_{\mathcal{D}}^{-\frac{1}{2}} X_{\mathcal{S}} X_{\mathcal{D}}^{-\frac{1}{2}} = X_{\mathcal{D}}^{-\frac{1}{2}} (X_{\mathcal{D}} - \frac{n}{n-1} C_m) X_{\mathcal{D}}^{-\frac{1}{2}} \\ &= I_p - \frac{n}{n-1} X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}} . \end{aligned}$$

Figure 1: Figure shows the subspace spanned by the scaled covariance matrix $X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}}$ of cluster centroids in the tilde transformed space $\tilde{x} = X_{\mathcal{D}}^{-\frac{1}{2}} x$ along with its orthogonal complement. The span of the scaled covariance matrix coincides with the $N - 1$ dimensional subspace containing the N cluster centroids (gray arrows in the indicated hyperplane). This space corresponds to the most informative projection for cluster separation. The Null space of the scaled covariance matrix (green arrow) is non-informative for cluster separation.

Since $X_{\mathcal{S}}$ is nonsingular, $X_{\mathcal{D}} = X_{\mathcal{S}} + \frac{n}{n-1} C_m$ is also nonsingular and hence, $rank(X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}}) = rank(C_m) = d$. This yields the following spectral decomposition:

$$\frac{n}{n-1} X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}} = \begin{pmatrix} U_d & U_{p-d} \end{pmatrix} \begin{pmatrix} \Lambda_d & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_d^T \\ U_{p-d}^T \end{pmatrix},$$

where U_d and U_{p-d} are eigenvector matrices of $X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}}$ corresponding to positive and 0 eigenvalues respectively and Λ_d represents a positive diagonal matrix with positives eigenvalues on the diagonal. Then,

$$\tilde{X}_{\mathcal{S}} = \begin{pmatrix} U_d & U_{p-d} \end{pmatrix} \begin{pmatrix} I_d - \Lambda_d & 0 \\ 0 & I_{p-d} \end{pmatrix} \begin{pmatrix} U_d^T \\ U_{p-d}^T \end{pmatrix}.$$

Therefore, $\sigma_i = (I_d - \Lambda_d)_{ii} < 1$ for $i \le d$ and $\sigma_i = 1$ for $i > d$. □

As can be seen from the proof of the Theorem 3.5, the eigenvectors and eigenvalues of the Hamiltonian matrix $\tilde{X}_{\mathcal{S}}$ can be split into two categories: (1) a set corresponding to the span of $X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}}$, which can be viewed as a scaling of the covariance matrix C_m with scaled eigenvalues $0 < \sigma_i < 1$ and (2) a set corresponding to the Null space of $X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}}$ with eigenvalues exactly 1.

In our approach, the transformation $\tilde{x} = X_{\mathcal{D}}^{-\frac{1}{2}} x$ is a scaling that is applied to the data points *a priori*. We refer to this space as the *tilde transformed space*. The span of the scaled covariance matrix $X_{\mathcal{D}}^{-\frac{1}{2}} C_m X_{\mathcal{D}}^{-\frac{1}{2}}$ coincides with the $N - 1$ dimensional subspace containing the N cluster centroids in the tilde transformed space. This space corresponds to the most informative projection for cluster separation, while the Null space of the scaled covariance matrix is uninformative in this regard (Fig. 1). Hence, the spectrum of $\tilde{X}_{\mathcal{S}}$ can be used to identify informative and non-informative directions. This process can be used to reduce the dimension and visualize high-dimensional datasets.

3.3 Optimal value of the tuning parameter μ

To illustrate the physical interpretation of our approach for tuning μ, we first consider the case of a single label dataset. In this case, $X_{\mathcal{D}}$ is not defined and, as shown in the preceding sections, we have $\tilde{X}_{\mathcal{S}} = X_{\mathcal{S}} = C$. Thus the Hamiltonian of the system is simply the covariance matrix obtained from the data and the principal features correspond to the low energy eigenstates. Now recall that the tuning parameter of our optimization problem μ corresponds to temperature T in the mapping to statistical mechanics and the free energy is given by $F[\rho] = \langle E \rangle - \mu H[\rho]$, where $\langle E \rangle = tr(\rho \hat{\mathcal{H}})$ is the average energy. In the low temperature limit ($\mu \longrightarrow 0$), free energy minimization is equivalent to minimizing the average energy of the system. Clearly, the average energy is minimized by setting the eigenvalues of the the density matrix ρ such that $\lambda_i = 1$ for the lowest energy eigenstate and $\lambda_i = 0$ for all the other eigenstates. This corresponds to projecting the data onto a line in the limit $\mu \longrightarrow 0$. On the other hand, in the limit $\mu \longrightarrow \infty$, the free energy is minimized by maximizing the entropy of the system, which corresponds to having equal probabilities for all the eigenstates, i.e. $\lambda_i = \frac{1}{p}$. The mapping to statistical mechanics suggests a criterion for choosing the optimal μ between these two extremes as discussed below.

A natural approach is to choose the parameter μ by maximizing the corresponding Fisher information, given that the maximum of the Fisher information corresponds to minimum variance in the estimate of μ based on the Cramer-Rao inequality [19]. The Fisher information for the parameter μ (obtained using the Boltzmann distribution derived above) is given by $\mathcal{I}(\mu) = \frac{1}{\mu^2}(\langle E^2 \rangle - \langle E \rangle^2)$ [4]. The connections to statistical mechanics provide further insight into the corresponding physical significance as outlined below. In the equilibrium state at temperature μ, let us denote the eigenstates of the Hamiltonian ($\tilde{X}_{\mathcal{S}}$) by E_i. The energy of the system is a random variable with corresponding probability distribution given by the Boltzmann distribution

$$\lambda_i = \frac{e^{-E_i/\mu}}{\sum_{i=1}^{p} e^{-E_i/\mu}}, \tag{19}$$

Using the above, it is straightforward to show that the equilibrium free energy is given by $F = -\mu \log(\sum_{i=1}^{p} e^{-E_i/\mu})$ and the average energy can be obtained from the free energy using $\langle E \rangle = -\mu^2 \dfrac{\partial(F/\mu)}{\partial \mu}$. Furthermore, the heat capacity of the system (defined as $C = \dfrac{\partial \langle E \rangle}{\partial \mu}$)

are bounded between 0 and 1 and add up to 1, which naturally solves the boundary problem of LDA. Second, the eigenvalues λ_is of S are obtained by optimizing the metric to enhance separation between classes (Eq. (12)) and hence provide more optimal scaling in the projected space, which can increase the performance of k-NN type classifiers. Finally, the scalings λ_is are tunable by varying the temperature parameter μ, which can vary from the boundary solution $\lambda_1 = 1, \lambda_2 = \cdots = \lambda_p = 0$ when $\mu \longrightarrow 0$ to equal scaling $\lambda_1 = \cdots = \lambda_p = 1/p$ when $\mu \longrightarrow \infty$ with the optimal value obtained by maximizing the Fisher Information. This can be advantageous for scalable high-dimensional data visualization. Fig. 3 illustrates this property of solution in visualizing high-dimensional data. In Fig. 3, data points from the Wine dataset are transformed and projected onto the first two SLDA components for various values of μ. The data consists of 3 categories in $\mathbb{R}^{13}$. There are a total of 2 informative components (section 3.2). By varying μ, the weights of the SLDA components can be controlled in a principled way. As can be seen, in the extreme case $\mu_{min} \longrightarrow 0$, the data is projected onto the first component, corresponding to $\lambda_1 = 1, \lambda_2 = 0$. Using μ_{opt}, obtained by maximizing Fisher Information, the weights are approximately $\lambda_1 \approx 0.7$ and $\lambda_2 \approx 0.3$, while $\mu_{\max} \longrightarrow \infty$ results in equal weights $\lambda_1 = \lambda_2 = 0.5$. For comparison, the projection onto principal components using standard PCA is depicted as well in Fig. 3.

3.5 Application to supervised learning

As in other metric learning models, training data can be used to obtain the optimal metric and to classify new examples in the optimally transformed space using k-NN or similar types of algorithms. The learning may be performed in a 'global' manner where all of the training data is utilized in learning the optimal space. Alternatively, to take advantage of local structures, one may use portions of the data to learn an ensemble of 'locally' optimal transform spaces. New examples can then be classified in each space and the final class label can be decided in an ensemble fashion. While our method does have an analytic solution and can be used efficiently to estimate a local metric, we restrict our attention to the global metric.

Our algorithm is straightforward. We first compute the within class $(C_1, \cdots C_N)$ and between class C_m covariance matrices. Next, the transformations $X_{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{N} C_i$, $X_{\mathcal{D}} = X_{\mathcal{S}} + \frac{n}{n-1}C_m$, and $\tilde{X}_{\mathcal{S}} = X_{\mathcal{D}}^{-\frac{1}{2}} X_{\mathcal{S}} X_{\mathcal{D}}^{-\frac{1}{2}}$ are calculated. We then calculate the Fisher Information $\mathcal{I}(\mu)$ for a grid of μ values spanning the range of eigenvalues of $\tilde{X}_{\mathcal{S}}$ and calculate the optimal value by identifying the local maxima of $\mathcal{I}(\mu)$ and an optional cross-validation. The final transformation $S^{1/2} = \sum_i \lambda_i^{1/2} u_i u_i^T$ is then be calculated in order to map the original data to the optimally transformed space. Since the distance metric is determined by free energy minimization, we name our method Free Energy Nearest Neighbor (FENN). The algorithm for FENN is shown in Algorithm 1. The time complexity of each step is indicated in the algorithms.

4 BENCHMARKS

We benchmarked our method on 10 UCI datasets [15] based on previous work (the information for each dataset can be seen in table 1) and we compare our performance to 8 other distance learning

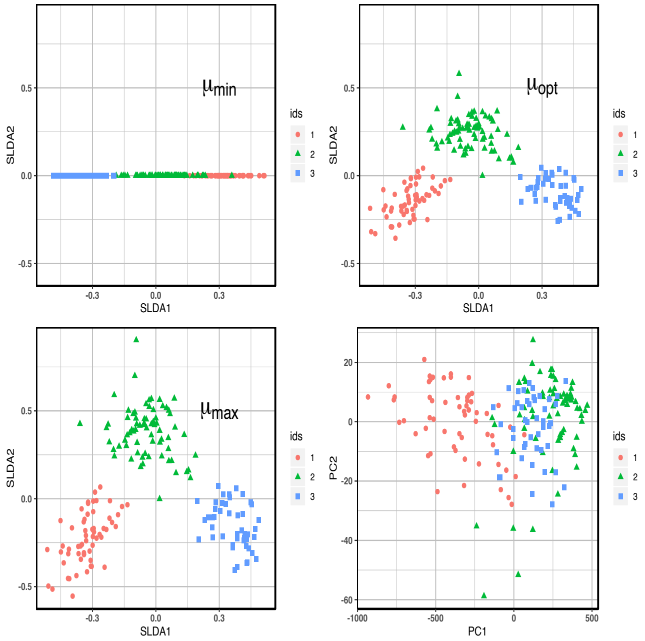

Figure 3: High dimensional data visualization. Wine data set, consisting of 3 classes in $\mathbb{R}^{13}$ are projected onto SLDA components using the indicated μ values. As $\mu \longrightarrow 0$ (μ_{min}) the first SLDA component has maximum weight of $\lambda_1 = 1$ while the second component has the weight $\lambda_2 = 0$. The weights for optimal μ value (μ_{opt}) are approximately $\lambda_1 \approx 0.7$ and $\lambda_2 \approx 0.3$. As $\mu \longrightarrow \infty$ (μ_{max}), both components have equal weight $\lambda_1 = \lambda_2 = 0.5$. The bottom right corner plot is the projection of the data onto PCA components.

```
1. Compute the matrices $X_{\mathcal{D}}^{1/2}$ and $\tilde{X}_{\mathcal{S}}$; $O((n+N)p^2 + p^3)$.
2. Compute the (energies) eigenvalues $E_i$ and eigenvectors $u_i$
   $(1 \le i \le d)$ for $\tilde{X}_{\mathcal{S}}$; $O(p^3)$.
3. Generate a grid of possible values for $\mu$, and compute $I(\mu)$
   for each $\mu$; $O(n_g p)$, where $n_g$ is a number of grid points.
4. Restrict the search grid to $(\mu_{opt} - N_\mu, \mu_{opt} + N_\mu)$ around
   $\mu_{opt}$; $O(1)$.
5. For each $\mu$ in $(\mu_{opt} - N_\mu, \mu_{opt} + N_\mu)$:
        Use $\mu$ to compute the $\lambda_i$ values $(1 \le i \le d)$; $O(d)$.
        Compute $S^{1/2} = \sum_i \lambda_i^{1/2} u_i u_i^T$, and transform the
        data using $S^{1/2} X_{\mathcal{D}}^{-1/2}$; $O(dp^2)$.
        Perform a 10 fold cross validation using $k$-NN on the
        new space and record the accuracy; $O((k+p)n)$.
6. Pick the $\mu$ with the best accuracy and compute the
   corresponding $\lambda_i$ values $(1 \le i \le d)$; $O(d)$.
7. Compute $S^{1/2} = \sum_i \lambda_i^{1/2} u_i u_i^T$, and transform the data using
   $S^{1/2} X_{\mathcal{D}}^{-1/2}$; $O(dp^2)$.
8. Predict using $k$-NN on the transformed data; $O((k+p)n')$
```

Algorithm 1: Pseudo-code for Free Energy Nearest Neighbor (FENN) and the complexity. The complexity of a whole algorithm is $O(p^3 + np^2)$ since $N, d \le \min\{n, p\}$

Table 1: Dataset Information

	bscale	glass	ionosphere	tictactoe	image segmentation	iris	wine	wdbc	car	waveform
dimensions	4	9	34	9	19	4	13	30	6	21
number of examples	625	214	351	958	2310	150	178	569	1728	5000
number of classes	3	6	2	2	7	3	3	2	4	3

Table 2: Results of 10-fold Cross Validation

	DANN	i-DANN	ADAMENN	iADAMENN	LFDA	NCA	Xing	LMNN	FENN
bscale	**0.955**	0.904	0.765	0.773	0.851	0.856	0.938	0.867	0.947
glass	0.682	0.575	0.664	0.668	0.664	0.678	0.673	0.668	**0.71**
ionosphere	0.84	0.798	0.84	0.843	0.838	0.843	0.84	**0.889**	0.846
tictactoe	0.857	0.816	0.811	0.811	0.841	0.832	0.837	0.846	**0.9**
image segmentation	0.963	0.878	0.877	0.877	0.958	0.92	0.96	0.959	**0.975**
iris	0.953	0.933	0.947	0.94	0.953	0.953	**0.96**	0.947	**0.96**
wine	0.978	0.944	0.955	0.938	0.916	0.781	0.848	0.978	**0.994**
wdbc	0.94	0.916	0.963	0.967	0.944	0.91	0.933	**0.97**	0.967
car	0.97	0.942	0.84	0.811	0.978	0.958	0.981	**0.987**	0.986
waveform	0.83	0.822	0.744	0.686	0.828	0.846	0.766	0.815	**0.849**

methods. We implemented our algorithm in an R package fenn (https://github.com/kouroshz/fenn). We use the R shogun package to run experiments for LMNN.

In addition, we implemented the 7 other widely used distance learning methods in a separate R package DistanceLearning [7] (https://github.com/carltonyfakhry/DistanceLearning), which we use to benchmark the remaining methods. The parameters for training all the methods were set as suggested in each method's original paper. In each trial, the training sets and the test sets were created by data preprocessing in our package. Then, training and testing methods from various packages were called for the same preprocessed datasets. LFDA [20, 21], NCA [9], Xing's method [26], LMNN [24, 25] and FENN are global metric methods therefore they learn a single transformation of the data. After the transformation is learned for each global method, we apply k-NN classification with $k = 5$. DANN, i-DANN [11], ADAMENN and i-ADAMENN [6] are local methods and generally take longer to train. We consider a local neighborhood ($\mu_{opt} - 4 \times p, \mu_{opt} + 4 \times p$) around μ_{opt} (a total of $6 \times p$ μ values where generated) for performing the cross validation in step 5 of Algorithm 1. Table 2 summarized the accuracy of the methods in a 10 fold cross validation experiment. In this setting, FENN performs as well or better than previous methods in 6 out of 10 datasets while it is very close in performance to the best method for the remaining 4 datasets.

5 CONCLUSION

In this paper we introduced a new formalization of distance metric learning inspired by a mapping to systems studied in quantum statistical mechanics. Our formalization applies the same metric when measuring the distance in both classes of similar and dissimilar points, while controlling the collapse of dimensions by enforcing a bound based on the Von Neumann entropy of the optimal metric. This is in contrast to several other distance metric learning methods, where collapse of dimensionality is avoided by applying difference metrics to measure the distance between similar and dissimilar points. Importantly, our formulation results in an analytical solution of the problem that avoids costly numerical approximations that are typically utilized in distance metric learning. Further, the quantum statistical mechanics interpretation of the problem allows for physical interpretation of the optimal solution and provides insights into selecting values for the entropy tuning parameter. We provide theoretical justification for optimality of the projected space in terms of class separation and show how the geometry and the tuning of the entropy parameter can be utilized for high-dimensional data visualization. Moreover, we establish a connection between our learned metric and LDA and demonstrate that our distance metric formulation generalizes LDA, while addressing some of the shortcomings. In future work, we plan to apply FENN with sparse covariance analysis and kernelization and further investigate analogous physical models for multi-class systems.

While the emphasis of the current work is on theoretical insights, the experimental results obtained also show the effectiveness of the proposed approach. Our approach establishes the connection between a class of optimization problems in distance metric learning to concepts explored in statistical mechanics and preexisting classification methods. In particular, the scaling provided by our method results in more optimal, controlled, separation of classes while still obtaining maximum discriminant directions. Our mapping to quantum statistical mechanics provides further insight into the geometrical aspects of the problem and may suggest analytical solutions for a broader class of optimization problems that are currently addressed numerically. In our opinion, further insights that will be gained building on the mappings established in this work are likely to lead to important developments in the field. Finally, to facilitate usage, we provide an efficient implementation of our algorithm along with implementation of 7 other widely used distance learning methods in R packages available to download at https://github.com/kouroshz/fenn and https://github.com/carltonyfakhry/DistanceLearning.

ACKNOWLEDGMENTS

The authors would like to thank Robert for the bagels and explaining CMYK and color spaces. The authors also wish to thank Kei, Sanae, and Kaede, for drawing a 3d picture displaying informative and uninformative features. The authors acknowledge funding support from the National Institutes of Health through grant 3U54CA156734-05S3 (as part of the University of Massachusetts Boston/Dana Farber-Harvard Cancer Center U54 partnership).

REFERENCES

[1] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. 2003. Learning Distance Functions Using Equivalence Relations. In *In Proceedings of the Twentieth International Conference on Machine Learning*. 11–18.

[2] Jonathan Baxter and Peter L. Bartlett. 1998. The Canonical Distortion Measure in Feature Space and 1-NN Classification. In *Advances in Neural Information Processing Systems 10*, M. I. Jordan, M. J. Kearns, and S. A. Solla (Eds.). MIT Press, 245–251. http://papers.nips.cc/paper/1357-the-canonical-distortion-measure-in-feature-space-and-1-nn-classification.pdf

[3] AurÃľlien Bellet, Amaury Habrard, and Marc Sebban. 2013. A Survey on Metric Learning for Feature Vectors and Structured Data. *arXiv:1306.6709 [cs, stat]* (June 2013). http://arxiv.org/abs/1306.6709 arXiv: 1306.6709.

[4] Gavin E Crooks. 2011. *Fisher information and statistical mechanics.* Technical Report. Citeseer.

[5] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. 2007. Information-theoretic Metric Learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*. ACM, New York, NYM, 209–216.⊠0⊠0⊠1⊠69.234⊠447.916.095⊠Tm⊠(https://doi.org/10.1145/1273496.1273496523⊠.95])Tj⊠.984⊠0⊠0⊠1⊠69.23602⊠0⊠0⊠1⊠69.234⊠45