Estimating and computing density based distance metrics

Sajama SAJAMA@UCSD.EDU
Alon Orlitsky ALON@UCSD.EDU
University of california at San Diego

Abstract

Density-based distance metrics have applications in semi-supervised learning, nonlinear interpolation and clustering. We consider density-based metrics induced by Riemannian manifold structures and estimate them using kernel density estimators for the underlying data distribution. We lower bound the rate of convergence of these plug-in path-length estimates and hence of the metric, as the sample size increases. We present an upper bound on the rate of convergence of all estimators of the metric. We also show that the metric can be consistently computed using the shortest path algorithm on a suitably constructed graph on the data samples and lower bound the convergence rate of the computation error. We present experiments illustrating the use of the metrics for semi-supervised classification and non-linear interpolation.

1. Introduction

Learning algorithms rely on a notion of similarity between data points to make inferences. When data is in $\mathbb{R}^d$, the standard similarity measure used is Euclidean distance. Recently, several methods have been proposed for incorporating into algorithms the intuition that two data points are similar to each other if they are connected by a high density region. This concept of similarity measure has been shown in experiments to lead to significant improvement in classification accuracy when applied to semi-supervised learning with very few labeled points, see, for example, (Blum & Chawla, 2001; Corduneanu & Jaakkola, 2003; Bousquet et al., 2004) and the references therein.

As an example consider a data sample shown in Figure 1. Assuming that data points of one class are likely to belong to a common high density region, we are looking for a measure of similarity according to which point 2 is closer to point 3 than to point 1. One way to formalize this intuitive notion of similarity is by specifying a density-based Riemannian manifold structure on $\mathbb{R}^d$ which assigns different lengths to path segments based on the data density at their location. Semi-supervised learning based on Riemannian metrics has been considered by (Vincent & Bengio, 2003; Bousquet et al., 2004).

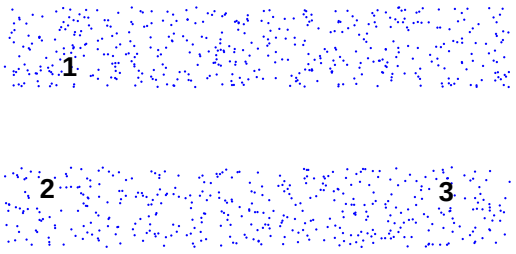

Figure 1. A simple example illustrating notion of similarity used in semi-supervised learning

Shortest paths according to such density-based distance (DBD) metrics have been proposed for non-linear interpolation of speech and images (Saul & Jordan, 1997; Bregler & Omohundro, 1995). DBD metrics could also be used for clustering when the notion of clusters is of 'connected regions' of high density separated by 'boundaries' of low density (Vincent & Bengio, 2003). (Lebanon, 2003) proposes picking a Riemannian metric from a parametric set of metrics based on an objective function which encourages metrics which reduce path lengths for paths passing through high density regions.

While DBD metrics have been considered before, their convergence rates and approximation errors have not been studied to our knowledge. Specifying a Riemannian structure on $\mathbb{R}^d$ gives us a notion of distance between data points. Since the 'true' data density is not known a priori, we show (for certain relevant

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

Riemannian metrics) that the path length values estimated from the finite data sample converge to the path lengths according to the true data density.

Computing the Riemannian distance involves the variational problem of minimizing the Riemannian length over all paths between two points. This computation problem has been extensively studied (Sethian, 1999) and finds applications in computational geometry, fluid mechanics, computer vision and material science. These methods involve building a grid whose size is exponential in the dimension of $\mathbb{R}^d$. This is inconvenient for the learning scenario where the data dimension is usually high. It is therefore necessary to consider grids based on data points, in which case the computational complexity grows at a rate polynomial in sample size n. Graph-based distance approximations have been proposed for computing the minimum distance (Vincent & Bengio, 2003; Bousquet et al., 2004), but these papers do not provide bounds on approximation error. We show that under certain graph construction and edge weighting, the approximation error goes to zero as the sample size increases. We presents experiments using these metrics for semi-supervised learning and non-linear interpolation.

2. Problem setup and definitions

We are given a set of d-dimensional data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ which are independent identically distributed samples drawn from a probability density function $f(\mathbf{x})$. To define DBD metrics, we use a modified definition of the path length of any path γ that depends on the density $f(\mathbf{x})$ and a suitably chosen function $g : \mathbb{R}^+ \to \mathbb{R}^+$,

$$\Gamma(\gamma; f) \doteq \int_{t=0}^{LE(\gamma)} g(f(\mathbf{x})) \left| \frac{d\gamma(t)}{dt} \right|_2 dt$$

where $|.|_2$ is the L_2 norm on $\mathbb{R}^d$. We can assume, without loss of generality, that all paths are parametrized to have unit speed according to the standard Euclidean metric on $\mathbb{R}^d$ and hence that $LE(\gamma)$ = Euclidean length of curve γ and $\left| \frac{d\gamma(t)}{dt} \right|_2 = 1$. The DBD distance between two points $\mathbf{x}'$ and $\mathbf{x}''$ is defined to be

$$d(\mathbf{x}', \mathbf{x}''; f) = \inf_{\gamma} \{\Gamma(\gamma; f)\} \tag{1}$$

where γ varies over the set of all paths from $\mathbf{x}'$ to $\mathbf{x}''$. The function g is assumed to have the following properties

[Ag1] g is monotonically decreasing function

[Ag2] $\inf_y g(y) > 0$

[Ag3] g has bounded first and second derivatives

[Ag4] $g(y) = 1 \quad \forall\, y \leq \alpha$

Properties [Ag1,Ag2] are assumed in order to make the resulting DBD metric suitable for our purposes. Specifically, [Ag1] ensures that paths passing through high density regions have shorter length and [Ag2] ensures that no path segment has zero length. Properties [Ag1] and [Ag3] are useful in bounding estimation error and [Ag4] is useful in proving the computation error bound.

This DBD metric can be thought of as being induced by a corresponding Riemannian structure. To specify a Riemannian manifold structure on $\mathbb{R}^d$ we need to specify the inner product on the space of tangent vectors at each point in $\mathbb{R}^d$. For $\mathbb{R}^d$ the tangent space at each point is just a copy of $\mathbb{R}^d$ itself. Hence the Riemannian structure at each point is determined by specifying the inner product between the d orthonormal unit vectors which span $\mathbb{R}^d$, i.e., $< \mathbf{e}_i, \mathbf{e}_j > \quad \forall\, i, j = 1, \ldots, d$.

$$< \mathbf{e}_i, \mathbf{e}_j > = \; g(f(\mathbf{x}))\delta_{ij} \tag{2}$$

where,

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

3. Estimating DBD metrics

Let us denote by $\mathcal{W}_s$, the set of functions which have s or more continuous derivatives. We assume that $f(\mathbf{x})$ has the following properties —

[Af1] $f(\mathbf{x}) \in \mathcal{W}_s$

[Af2] $f(\mathbf{x})$ has bounded support

[Af3] $\exists\;\; C_1$ such that $|| \bigtriangledown f|| \leq C_1$

The smoothness parameter s measures the complexity of the class of underlying distributions. Given that $f(\mathbf{x})$ belongs to $\mathcal{W}_s$, we base the density estimate on a one-dimensional kernel with the following properties

$k(x) = k(-x) \qquad \int k(x)dx = 1$

$\sup_{-\infty<x<\infty} |k(x)| \leq A < \infty$

$\int x^m k(x)dx = 0,\; m = 1, \ldots, s-1$

$0 \neq \int x^s k(x)dx < \infty$

We then estimate the density to be

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n\, h_n^d} \sum_{i=1}^{n} \mathbf{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n} \right)$$

where h_n is the *width* parameter of the kernel which is chosen to be a function of sample size n and $\mathbf{K}(\mathbf{x}) = \Pi_{j=1}^{d} k(x_j)$.

To characterize the estimators of the Riemannian metric, we use the definitions of upper and lower bounds on rate of convergence of estimators proposed in (Stone, 1980).

Definition 1. *A convergence rate r is* achievable *if there is a sequence $\{\hat{\Gamma}_n(\gamma)\}$ of estimators such that*

$$\lim_{c\to\infty} \limsup_{n} \sup_{f\in\mathcal{W}_s} P_f(|\hat{\Gamma}_n(\gamma) - \Gamma(\gamma; f)| > cn^{-r}) = 0$$

Definition 2. *A rate $r > 0$ is an* upper bound *to the rate of convergence if for every sequence $\hat{\Gamma}_n(\gamma)$ of estimators of $\Gamma(\gamma; f)$,*

$$\liminf_{n} \sup_{f\in\mathcal{W}_s} P_f(|\hat{\Gamma}_n(\gamma) - \Gamma(\gamma; f)| > cn^{-r}) > 0 \qquad \forall c > 0 \tag{3}$$

$$\lim_{c\to 0} \liminf_{n} \sup_{f\in\mathcal{W}_s} P_f(|\hat{\Gamma}_n(\gamma) - \Gamma(\gamma; f)| > cn^{-r}) = 1 \tag{4}$$

For statements in probability about random variables T_n, Q_n, whose distributions may depend on $f(\mathbf{x})$, we will use the notation $T_n = \mathcal{O}(Q_n)$ when $\lim_{c\to\infty} \limsup_n \sup_{f\in\mathcal{W}_s} P(|T_n| > c|Q_n|) = 0$. For bounding the error in estimating the distance between any pair of points we borrow from the proof techniques used in (Stone, 1980; Goldstein & Messer, 1992). We use the following two lemma about well known, e.g., (Nadaraya, 1989), properties of the kernel density estimators.

Lemma 1 (Bias of the kernel density estimator). *Let $\mu = (\mu_1, \ldots, \mu_d)$ be a d-dimensional vector, $|\mu| = \sum_{j=1}^{d} \mu_j$, $\mu! = \mu_1! \ldots \mu_d!$, $\mathbf{u}^\mu = u_1^{\mu_1} \ldots u_d^{\mu_d}$ and $D^\mu = \frac{\partial^{\mu_1}}{\partial u_1^{\mu_1}} \ldots \frac{\partial^{\mu_d}}{\partial u_d^{\mu_d}}$. Then, $\forall \mathbf{x}$, the bias*

$$E[\hat{f}_n(\mathbf{x})] - f(\mathbf{x}) = s h_n^s \int_{\mathbf{u}\in\mathbb{R}^d} F(\mathbf{u}, \mathbf{x}) \mathbf{K}(\mathbf{u}) d\mathbf{u},$$

where

$$F(\mathbf{u}, \mathbf{x}) = \sum_{|\mu|=s} \frac{\mathbf{u}^\mu}{\mu!} \int_{T=0}^{1} (1-T)^{s-1} D^\mu f(\mathbf{x} + T\mathbf{u}) dT.$$

Lemma 2 (Variance of the kernel density estimator). *$\forall \mathbf{x}$, $\forall \epsilon \geq 0$, for sufficiently large n, the variance*

$$E\left[\hat{f}_n(\mathbf{x}) - E\hat{f}_n(\mathbf{x})\right]^2 \leq \frac{(1+\epsilon) f(x)}{n h_n^d} \int_{\mathbf{u}\in\mathbb{R}^d} \mathbf{K}^2(\mathbf{u}) d\mathbf{u}.$$

Theorem 1 (Achievability). *Uniformly over all pairs of points $\mathbf{x}'$ and $\mathbf{x}'' \in \mathbb{R}^d$ and paths $\gamma(t)$ of length $LE(\gamma)$ joining them, the plug-in estimator*

$$\hat{\Gamma}_n(\gamma) = \Gamma(\gamma; \hat{f}_n)$$

that uses the kernel density estimator $\hat{f}_n$, achieves the rate of convergence $r = min(\frac{s}{s+d}, \frac{1}{2})$ where the width of the kernel density estimators $h_n = \Theta(\frac{1}{n^{\frac{1}{s+d}}})$.

Proof. We begin by defining the derivative T of the functional $\Gamma(\gamma; f)$ with respect to changes $\delta f(\mathbf{x})$ in $f(\mathbf{x})$ to be

$$T(\delta f; f) \doteq \int_{t=0}^{LE(\gamma)} g'(f(\gamma(t))) \delta f(\gamma(t)) \left|\frac{d\gamma(t)}{dt}\right|_2 dt$$

Hence, we can write

$$\begin{aligned} &|\Gamma(\gamma; \hat{f}_n) - \Gamma(\gamma; f) - T(\hat{f}_n - f; f)| \\ &= \left| \int_{t=0}^{LE(\gamma)} \left[g(\hat{f}_n) - g(f) - (\hat{f}_n - f) g'(f) \right] \left|\frac{d\gamma(t)}{dt}\right|_2 dt \right|, \end{aligned}$$

where f and $\hat{f}_n$ are evaluated at $\gamma(t)$. By a proof similar to intermediate value theorem, we know that $g(y + \delta y) - g(y) - \delta y\ g'(y) = \frac{g''(\beta)}{2!}\ \delta y^2$ for some β in the domain of g. Hence,

$$\begin{aligned} &|\Gamma(\gamma; \hat{f}_n) - \Gamma(\gamma; f) - T(\hat{f}_n - f; f)| \\ &\leq \mathrm{C} \int_{t=0}^{LE(\gamma)} \{\hat{f}_n(\gamma(t)) - f(\gamma(t))\}^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt. \end{aligned}$$

Therefore,

$$\begin{aligned} &|\Gamma(\gamma; \hat{f}_n) - \Gamma(\gamma; f)| \\ &\leq |T(\hat{f}_n - E\hat{f}_n; f)| + |T(E\hat{f}_n - f; f)| \\ &+ \left| \mathrm{C} \int_{t=0}^{LE(\gamma)} \{\hat{f}_n(\gamma(t)) - f(\gamma(t))\}^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt \right|. \end{aligned}$$

We now bound each of these three terms in turn. The first term,

$$\begin{aligned} T(\hat{f}_n - E\hat{f}_n; f) = \frac{1}{n} \sum_{i=1}^{n} T \Bigg(\Bigg\{ &\frac{1}{h_n^d} \mathbf{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \\ &- \int \mathbf{K}(\mathbf{u}) f(\mathbf{x} - h_n \mathbf{u}) d\mathbf{u} \Bigg\}; f \Bigg), \end{aligned}$$

is a sum of n, zero mean random variables. The variance of each of these variables is upper bounded by

$$\leq \frac{(1+\epsilon_1) L}{n\ h_n^d} \left(\max_{\beta} g'(\beta) \right)^2 (\max f(\mathbf{x})) \int_{\mathbb{R}^d} \mathbf{K}^2(\mathbf{u}) d\mathbf{u}.$$

This follows from Cauchy-Schwartz inequality, Fubini's theorem and for sufficiently large n by Lemma 2. The constant L is the maximum manifold distance between any two points in the support set of $f(\mathbf{x})$ which is bounded by the maximum Euclidean length between

any two such points. Since the variance of each of these random variables is bounded and since $T(\hat{f}_n - E\hat{f}_n; f)$ is the sum of n of these variables, we can conclude that $T(\hat{f}_n - E\hat{f}_n; f) = \mathcal{O}\left(1/n^{\frac{1}{2}}\right)$.

The second term $T(E\hat{f}_n - f; f)$ can be bounded in terms of the partial derivatives of $f(\mathbf{x})$ as $T(E\hat{f}_n - f; f) = \mathcal{O}(h_n^s)$. This follows from Lemma 1 and the uniform continuity of $D^\mu f$ and holds for sufficiently large n.

The third term, $\frac{1}{2}\left(\max_\beta |g''(\beta)|\right) \int_{t=0}^{LE(\gamma)} \{\hat{f}_n(\gamma(t)) - f(\gamma(t))\}^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt$ can be bounded by bounding the expectation of $\int_{t=0}^{LE(\gamma)} \{\hat{f}_n(\gamma(t)) - f(\gamma(t))\}^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt$ and using Markov's inequality and Lemma 2 and 1.

$$
\begin{aligned}
&E\left[\int_{t=0}^{LE(\gamma)} \{\hat{f}_n(\gamma(t)) - f(\gamma(t))\}^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt\right] \\
&= \int_{t=0}^{LE(\gamma)} (E\hat{f}_n - f)^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt \\
&\qquad + \int_{t=0}^{LE(\gamma)} E(\hat{f}_n - E\hat{f}_n)^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt \\
&= \mathcal{O}(\frac{1}{nh_n^d}) + \mathcal{O}(h_n^{2s})
\end{aligned}
$$

Collecting the three terms and assuming that $h_n = \Theta(1/n^{\frac{1}{s+d}})$, we conclude

$$
|\Gamma(\gamma; \hat{f}_n) - \Gamma(\gamma; f)| = \mathcal{O}(\frac{1}{n^{\frac{1}{2}}} + \frac{1}{n^{\frac{s}{s+d}}}).
$$

□

Theorem 2 (Upper bound). *No estimator of the DBD metric can converge at a rate faster than $r = \frac{1}{2}$.*

Proof. To prove this result, we show that there is a density function $f(\mathbf{x})$ and a shortest path between two points γ for which $\hat{\Gamma}(\gamma)$ cannot converge to $\Gamma(\gamma; f)$ faster than the rate r, irrespective of which estimator is used to obtain $\hat{\Gamma}(\gamma)$. This is accomplished by first demonstrating a sequence of density functions f_n which converge to f fast enough that they cannot be distinguished from f using n samples $X_1, \ldots, X_n$ with arbitrarily high accuracy. The technique, termed 'the classification argument' was used in (Stone, 1980).

Consider a density function $f_0(\mathbf{x})$ with the property that the set $\{\mathbf{x} : f_0(\mathbf{x}) > \alpha\}$ contains an open ball in $\mathbb{R}^d$ over which $f_0(\mathbf{x})$ is constant. Let γ be any line segment contained in this open ball, let $\mathbf{x}_p$ be any point in the relative interior of γ and let $\mathbf{x}_0$ be any point in the ball which does not lie on the path γ. Since $f_0(\mathbf{x})$ is constant over the ball, any line segment including γ is the shortest path between its two end points. Let ψ be a non-negative, infinitely differentiable C^∞ function with compact support (for an example called 'the blimp' see (Strichartz, 1995)). Define

$$
w_n(\mathbf{x}) \doteq \delta N n^{-\frac{1}{2}} \left\{\psi(\mathbf{x} - \mathbf{x}_p) - b_n \psi(\mathbf{x} - \mathbf{x}_0)\right\}.
$$

Here, b_n is chosen such that $\int w_n f_0 d\mathbf{x} = 0$. We define a sequence of densities $f_n = f_0(1 + w_n)$. From the assumption [Ag3] that g is a monotonically decreasing function and from the definition of f_n, it follows that the straight line γ is the shortest path between its end points under the Riemannian metric specified by $f_n \ \forall \ n$. Since b_n is a constant, it remains bounded as $n \longrightarrow \infty$. Now by the classification argument of (Stone, 1980), to prove our result it is sufficient to show the following two inequalities,

$$
\limsup_n \ n \, E_{f_0} w_n^2(X) < \infty \tag{5}
$$

$$
\frac{\Gamma(\gamma; f_n) - \Gamma(\gamma; f_0)}{2} = \Omega\left\{\delta N \left(n^{-\frac{1}{2}}\right)\right\}. \tag{6}
$$

To bound the difference between f_0 and f_n, note that

$$
\begin{aligned}
&nE_{f_0} w_n^2(X) \quad = \\
&\frac{n\delta^2 N^2}{n} \int f_0(\mathbf{x}) \left\{\psi(\mathbf{y}) - b_n \psi(\mathbf{y} + n^{\alpha_2}(\mathbf{x}_p - \mathbf{x}_0))\right\}^2 d\mathbf{x} \\
&< \infty
\end{aligned}
$$

Now we bound $\Gamma(\gamma; f_n) - \Gamma(\gamma; f_0) = T(f_n - f_0; f_0) + \mathcal{O}(\int_{t=0}^{LE(\gamma)} (f_n - f_0)^2 \left|\frac{d\gamma(t)}{dt}\right|_2 dt) = \Omega(\delta N n^{-\frac{1}{2}}) + \mathcal{O}(n^{-1})$. □

4. Computing DBD metrics

In Section 3, we analyzed the effect of using an estimate of the density function in place of the density function itself. However, even if the density were known, computing the Riemannian metric between two points is not an easy task. This is a variational minimization problem since the distance is defined as the infimum of path lengths over all paths joining the points (Equation 1). Several methods including those that propose using density-based metrics have used graphs for semi-supervised learning. However, they do not show that the proposed graph construction methods lead to a consistent distance measure, i.e., they do not show convergence of the graph shortest path length to the Riemannian metric with increasing sample size. We show that the rate at which the shortest distance on a suitably constructed graph G approaches the DBD metric, is lower bounded by $1/2d$. In the proof, we use some techniques from Isomap (Tenenbaum et al., 2000).

We first describe our method for constructing the graph and assigning weights to the graph edges. Recall ([Ag1] in Section 2) that $g(y) = 1 \quad \forall\ y \leq \alpha$. Let $C_p(\alpha) \doteq \{\mathbf{x} : \hat{f}(\mathbf{x}) \geq \alpha\}$ and let $C_p(\alpha;\epsilon) \doteq \bigcup_{\mathbf{x}\in C_p(\alpha)} B(\mathbf{x},\epsilon)$ where $B(\mathbf{x},\epsilon)$ is a d-dimensional ball of radius ϵ centered at $\mathbf{x}$.

A point $\mathbf{x} \in \mathbb{R}^d$ is *high density* if $\mathbf{x} \in C_p(\alpha;\epsilon)$. A maximal connected set of high-density points is a *high-density component*. Since the density $f(\mathbf{x})$ integrates to one, it can be shown that there will be only finitely many high-density components and hence $C_p(\alpha;\epsilon)$ will be partitioned into finitely many high-density components $R_1,\ldots,R_k$. Note that these are high density components with respect to the estimated distribution $\hat{f}$ and not the 'true' distribution $f(\mathbf{x})$. $C_p(\alpha;\epsilon)$ is being defined as a way to mollify the difficult properties of $C_p(\alpha)$ which can have complex boundaries (e.g., dendrils defined in (Blum & Chawla, 2001)) and can have an infinite number of disjoint, maximally connected components.

The graph G is then defined as follows. Its vertices are the observed data points $\mathbf{x}_1,\ldots,\mathbf{x}_n$. Two nodes $\mathbf{x}_i$, $\mathbf{x}_j$ are connected if at least one of the following holds:

1. The Euclidean distance between two nodes is at most ϵ. The weight of such an edge is $w(\mathbf{x}_i,\mathbf{x}_j) = g(f(\overline{\mathbf{x}_i+\mathbf{x}_j}/2))|\mathbf{x}_i-\mathbf{x}_j|_2$.
2. At least one of the nodes is high-density, they are at least ϵ apart and the straight line joining the two nodes leaves $C_p(\alpha;\epsilon)$. The weight of such an edge is $w(\mathbf{x}_i,\mathbf{x}_j) = |\mathbf{x}_i-\mathbf{x}_j|_2$.

We use three distance metrics between data points $\mathbf{x}$ and $\mathbf{y}$, namely, the DBD metric

$$d_M(\mathbf{x},\mathbf{y}) \doteq d(\mathbf{x},\mathbf{y};\hat{f}) = \inf_{\gamma}\{\Gamma(\gamma;\hat{f})\},$$

the graph distance

$$d_G(\mathbf{x},\mathbf{y}) \doteq \min_{P}\left(w(\mathbf{x}_0,\mathbf{x}_1)+\ldots+w(\mathbf{x}_{p-1},\mathbf{x}_p)\right),$$

and an intermediate distance

$$d_S(\mathbf{x},\mathbf{y}) \doteq \min_{P}\left(d_M(\mathbf{x}_0,\mathbf{x}_1)+\ldots+d_M(\mathbf{x}_{p-1},\mathbf{x}_p)\right)$$

where $P = (\mathbf{x}_0,\ldots,\mathbf{x}_p)$ varies over all paths along the edges of G connecting $\mathbf{x} = \mathbf{x}_0$ to $\mathbf{y} = \mathbf{x}_p$.

To lower bound the rate of convergence of the shortest path along graph G to the DBD metric, we bound the difference between the graph distance and DBD metric in Theorem 3. For this purpose we show the DBD metric and the intermediate distance are close to each other in Lemma 3. Lemma 4 and 5 state that the graph and intermediate distances are close (proof is omitted for lack of space).

Lemma 3 (Bounding the difference between DBD metric and intermediate distance). *If* $\forall \mathbf{x} \in C_p(\alpha;2\epsilon)$ $\exists$ *some data point* $\mathbf{x}_i$ *for which* $d_M(\mathbf{x},\mathbf{x}_i) \leq \delta$ *and if* $4\delta < \epsilon$, *then* $\forall$ *pairs of data points* $\mathbf{x}$ *and* $\mathbf{y}$,

$$d_M(\mathbf{x},\mathbf{y}) \leq d_S(\mathbf{x},\mathbf{y}) \leq \left(1+\frac{6\delta}{\epsilon}+\frac{8\delta^2}{\epsilon^2}\right) d_M(\mathbf{x},\mathbf{y})$$

Proof. The first inequality $d_M(\mathbf{x},\mathbf{y}) \leq d_S(\mathbf{x},\mathbf{y})$ is true by the definition of d_M and d_S. Let γ be any piecewise-smooth path connecting $\mathbf{x}$ to $\mathbf{y}$ with length l. If we are able to find a path from $\mathbf{x}$ to $\mathbf{y}$ along edges of G whose length $d_M(\mathbf{x}_0,\mathbf{x}_1)+\ldots+d_M(\mathbf{x}_{p-1},\mathbf{x}_p)$ is less than $\left(1+\frac{6\delta}{\epsilon}+\frac{8\delta^2}{\epsilon^2}\right) d_M(\mathbf{x},\mathbf{y})$, then the right hand inequality would follow by taking infimum over γ.

Note that it is sufficient to consider only those γ for which contiguous segments outside $C_p(\alpha;\epsilon)$ are straight lines. This is because, given any γ without this property, we can define a path γ' such that the length of γ' is less than the length of γ by just replacing wiggly segment of γ outside $C_p(\alpha;\epsilon)$ by straight lines (recall that the density-based Riemannian metric has been defined to be constant Euclidean in the region outside $C_p(\alpha)$). We consider different cases based on the regions the path γ passes through.

Case (a) : γ *is wholly contained in one of the sub-regions* R_k *of* $C_p(\alpha)$.
We use an argument similar to the one used in Isomap (Tenenbaum et al., 2000). If $l \leq \epsilon - 2\delta$, then $\mathbf{x}$,$\mathbf{y}$ are connected by an edge which we can use as our path through the graph. If $l > \epsilon - 2\delta$, we write $l = l_0 + (l_1 + l_1 + \ldots + l_1) + l_0$ where $l_1 = \epsilon - 2\delta$ and $\frac{\epsilon-2\delta}{2} \leq l_0 \leq \epsilon - 2\delta$. Now, cut up the arc γ into pieces in accordance with this decomposition giving a sequence of points $r_0 = \mathbf{x}, r_1, \ldots, r_p = \mathbf{y}$. Each point r_i lies within a distance δ of a sample point $\mathbf{x}_i$

$$d_M(\mathbf{x}_i,\mathbf{x}_{i+1})$$
$$\leq d_M(\mathbf{x}_i,r_i) + d_M(r_i,r_{i+1}) + d_M(r_{i+1},\mathbf{x}_{i+1}) \leq \frac{l_1\epsilon}{\epsilon-2\delta}$$
$$\&\quad d_M(\mathbf{x},\mathbf{x}_1) \leq l_0\frac{\epsilon}{\epsilon-2\delta} \quad \&\quad d_M(\mathbf{x}_{p-1},\mathbf{y}) \leq l_0\frac{\epsilon}{\epsilon-2\delta}$$

Since $l_0\frac{\epsilon}{\epsilon-2\delta} \leq \epsilon$, we find that each edge has manifold length $\leq \epsilon$ and hence belongs to G. Hence,

$$d_S(\mathbf{x},\mathbf{y}) \leq l\frac{\epsilon}{\epsilon-2\delta} < l\left(1+\frac{4\delta}{\epsilon}\right)$$

Case (b) : *All segments of γ that lie outside $C_p(\alpha;\epsilon)$ have length $\geq \epsilon - 2\delta$.*
We consider the case when both the initial and final points, $\mathbf{x}$ and $\mathbf{y}$ lie in $C_p(\alpha;\epsilon)$. The case when one or both of the end-points lies outside can be similarly handled. We divide the path γ into $2k+1$ sections, where k is the number of times γ goes outside $C_p(\alpha;\epsilon)$ i.e., — $\mathbf{x}\ldots r_{o1}\ldots r_{p1}\ldots r_{o2}\ldots r_{p2}\ldots r_{ok}\ldots r_{pk}\ldots \mathbf{x}_p$ where the sections $r_{oi} - -r_{pi}$ lie outside $C_p(\alpha)$. The d_S and d_M lengths of the interior segments are related exactly as in Case (a) and hence we can write

$$d_S(\mathbf{x}_0,\mathbf{x}_p) \leq \frac{\epsilon}{\epsilon-2\delta}\left\{d_M(\mathbf{x}_0,r_{o1}) + d_M(r_{p1},r_{o2}) + \ldots \right.$$
$$\left. + d_M(r_{pk},\mathbf{x}_p)\right\} + \left\{2\delta + d_M(r_{o1},r_{p1})\right\} + \ldots$$
$$+ \left\{2\delta + d_M(r_{ok},r_{pk})\right\}.$$

Since each outside segment has a minimum length $\epsilon - 2\delta$, $d_M(\mathbf{x},\mathbf{y}) \geq (\epsilon-2\delta)k$. Hence $2\delta k \leq \frac{2\delta}{\epsilon-2\delta} d_M(\mathbf{x},\mathbf{y})$ and

$$d_S(\mathbf{x},\mathbf{y}) \leq \left(1 + \frac{6\delta}{\epsilon} + \frac{8\delta^2}{\epsilon^2}\right) d_M(\mathbf{x},\mathbf{y}).$$

□

Lemma 4 (Bounding the difference between intermediate and graph distances - 1). *For all pairs of data points $\mathbf{x}_i$,$\mathbf{x}_j$ connected by an edge in G with $|\mathbf{x}_i - \mathbf{x}_j|_2 \leq \epsilon$,*

$$(1-\lambda_1) d_G(\mathbf{x}_i,\mathbf{x}_j) \leq d_S(\mathbf{x}_i,\mathbf{x}_j) \leq (1+\lambda_1) d_G(\mathbf{x}_i,\mathbf{x}_j)$$

where

$$\lambda_1 = 2\frac{\max|\bigtriangledown_{\mathbf{x}} g(f(\mathbf{x}))|_2 \epsilon}{\min_{\mathbf{x}} g(f(\mathbf{x}))}$$

Lemma 5 (Bounding the difference between intermediate and graph distances - 2). *For all pairs of data points $\mathbf{x}_i$,$\mathbf{x}_j$ connected by an edge in G with $|\mathbf{x}_i - \mathbf{x}_j|_2 > \epsilon$,*

$$(1-\lambda_2) d_G(\mathbf{x}_i,\mathbf{x}_j) \leq d_S(\mathbf{x}_i,\mathbf{x}_j) \leq (1+\lambda_2) d_G(\mathbf{x}_i,\mathbf{x}_j)$$

where

$$\lambda_2 = \frac{2\delta^2 \max|\bigtriangledown g(f(\mathbf{x}))|_2}{\epsilon}$$

Theorem 3 (Lower bound on the rate at which graph distance approaches DBD metric). *$\forall \zeta < 1/2d$, a computing error (uniform over all pairs of points $\mathbf{x}$, $\mathbf{y}$) of*

$$(1-\lambda) d_M(\mathbf{x},\mathbf{y}) \leq d_G(\mathbf{x},\mathbf{y}) \leq (1+\lambda) d_M(\mathbf{x},\mathbf{y})$$

with $\lambda = \mathcal{O}(n^{-\zeta})$ can be achieved for sufficiently large data sample n.

Proof. We show that the shortest path along the graph is within λ of the DBD metric, by considering two cases based on the properties of the shortest path. We define a new graph G_2 on the data points which contains only a subset of the edges in G. G_2 contains all edges in G where $|\mathbf{x}_i - \mathbf{x}_j|_2 \leq \epsilon$. In addition, it contains edges in G that leave $C_p(\alpha;\epsilon)$ and whose endpoints, $\mathbf{x}_i$ and $\mathbf{x}_j$, lie within δ of the boundary of $C_p(\alpha;\epsilon)$. Note that G_2 is sufficient to approximate all shortest paths between data points. However, it is difficult to compute/generate G_2 and hence we define a more dense graph G with the property that the extra edges are most likely not going to be used in the shortest path unless they form a good approximation to the shortest path along G_2.

Case (a) : *The shortest path along G lies entirely within the subset $G2$.*
Using the theorem from (Gine & Guillou, 2002), we can conclude that our choice in Section 3 of kernel width, $h_n = \frac{1}{n^{\frac{1}{s+d}}}$ and other properties assumed about $f(\mathbf{x})$ ensure that

$$\max_{\mathbf{x}} |f_n(\mathbf{x}) - f(\mathbf{x})| = \mathcal{O}\left(\sqrt{\left(\frac{\log(n)}{n^{\frac{s}{s+d}}}\right)}\right)$$

This means that for sufficiently large n, $\forall$ points $\mathbf{y}$ in $C_p(\alpha;2\epsilon)$ have the property that $f(\mathbf{y}) \geq \alpha - \alpha_1$ for arbitrarily small α_1. Using this fact and the δ-sampling condition (Tenenbaum et al., 2000), we know that the requirement for Theorem 3 is satisfied when $n = \Omega\left(\left(\frac{1}{\delta}\right)^d \log\frac{1}{\delta}\right)$. This condition is satisfied with a choice of $\zeta < 1/2d$ and letting $\delta = \Theta(n^{-2\zeta})$ and $\epsilon = \Theta(n^{-\zeta})$. Let $\lambda_3 = max(\lambda_1,\lambda_2)$, where λ_1 and λ_2 are defined in Lemma 4 and 5 respectively. Hence we can use Lemma 3, 4 and 5 to conclude that

$$(1-2\lambda_3) d_M(\mathbf{x},\mathbf{y}) \leq d_G(\mathbf{x},\mathbf{y})$$
$$\leq (1+2\lambda_3)\left(1 + \frac{6\delta}{\epsilon} + \frac{8\delta^2}{\epsilon^2}\right) d_M(\mathbf{x},\mathbf{y})$$

which implies that

$$(1-\lambda_4) d_M(\mathbf{x},\mathbf{y}) \leq d_G(\mathbf{x},\mathbf{y}) \leq (1+\lambda_4) d_M(\mathbf{x},\mathbf{y}),$$

where $\lambda_4 = \mathcal{O}(\epsilon + \frac{\delta}{\epsilon}) = \mathcal{O}(n^{-\zeta})$.

Case (a): *The shortest path, P, along G uses some edges that are not part of $G2$.*
Consider any edge E connecting $\mathbf{x}_l$ and $\mathbf{x}_m$ in the shortest path along G that is not in G_2. We will show that there is a path through G_2 that can closely approximate this edge E and hence this shortest path. Consider the case when only one section near the end point $\mathbf{x}_m$ is more than δ in $C_p(\alpha;\epsilon)$. The case when

more sections of E are in $C_p(\alpha;\epsilon)$ can be similarly proved. Consider the boundary point r_b where the straight line starting at $\mathbf{x}_m$ toward $\mathbf{x}_l$ first touches the edge of $C_p(\alpha;\epsilon)$. By the δ-sampling condition, there is a data point $\mathbf{x}_k$ within δ of r_b. Consider the path consisting of the edge $\mathbf{x}_l$–$\mathbf{x}_k$ and the shortest path, P_2, between $\mathbf{x}_k$ and $\mathbf{x}_m$ through those edges of G that connect nodes within ϵ of one another. Let d'_{G2} be the length of a path that follows P except when it comes to edges not in G_2 in which case it follows paths P_2 constructed to pass through G_2. Let d_{G2} be the length of shortest path along graph G_2. From proof of case (a), we know that

$$(1-\lambda_4)d_M(\mathbf{x},\mathbf{y}) \le d_{G2}(\mathbf{x},\mathbf{y}) \le (1+\lambda_4)d_M(\mathbf{x},\mathbf{y}),$$

$$d_G(\mathbf{x},\mathbf{y}) \le d'_{G2}(\mathbf{x},\mathbf{y}) \le (1+\lambda_4)d_G(\mathbf{x},\mathbf{y})$$

$$\text{and} \quad d_G(\mathbf{x},\mathbf{y}) \le d_{G2}(\mathbf{x},\mathbf{y}) \le d'_{G2}(\mathbf{x},\mathbf{y}).$$

Hence,

$$(1-2\lambda_4)d_M(\mathbf{x},\mathbf{y}) \le d_G(\mathbf{x},\mathbf{y}) \le (1+\lambda_4)d_M(\mathbf{x},\mathbf{y})$$

□

5. Experiments

Non-linear interpolation : In density-based interpolation, given two points, our task is to find an interpolating path that passes through regions of space to which the modelled density $f(\mathbf{x})$ assigns high probability (Saul & Jordan, 1997). Given a sample of points $\mathbf{x}_1,\ldots,\mathbf{x}_n$, we can find an approximation to such a path by computing the weighted graph G as described in Section 4 and tracing the shortest path between the two points through the graph. We illustrate this using a simple example where data is drawn from a single spherical Gaussian distribution with mean at $(0,0)$ and variance one in each direction. The shortest path according to a DBD metric with $g = \exp(-(f(\mathbf{x})-\alpha)/(f_{\max}-\alpha))$ and based on 1000 data samples drawn from the Gaussian distribution is shown in Figure 2.

Semi-supervised learning : DBD metrics could be used for semi-supervised classification in several ways. We compare the DBD metric 1-nearest neighbor (1-NN) method with the standard 1-NN and randomized min-cut (Blum et al., 2004) classifiers. For the DBD based 1-NN implementation, we choose the function g to fall exponentially with increase in density beyond α which in turn is chosen to be smaller than the estimated density at all sample points. The randomized min-cut method involves averaging over several min-cuts obtained by randomly changing the graph weights. It is suggested in (Blum et al., 2004)

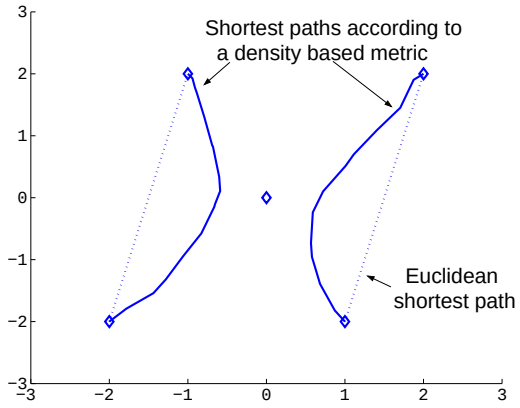

Figure 2. Density-based non-linear interpolation using 1000 iid samples drawn from a spherical, unit variance, zero mean Gaussian distribution.

that those min-cuts which lead to a very unbalanced classification are to be rejected. However, there is no clear way to choose this cut-off ratio. For the results presented here we choose the cutoff to be slightly less than the 'true class ratio' which is the ratio of the membership of the classes in the dataset. We present classification accuracy results on data from the UCI machine learning repository, summarized in Table 1.

Table 1. Description of data sets used for classification.

DATA SET	DATA DIMENSION	DATA SET SIZE	CLASS RATIO
ADULT	6	1000	0.30
ABALONE - 9 VS 13	7	892	0.29
ABALONE - 5 VS 9	7	804	0.17
DIGITS - 1 VS 2	256	2200	1.00

Figure 3 shows the percentage of accurate classification for labeled set sizes ranging from 2 to 20. For each labeled set size, the results shown are averaged over fifty randomly chosen labeled sets. DBD based 1-NN outperformed the standard 1-NN for small labeled set sizes on adult and abalone (classes 9 vs 13) data and showed no improvement in the case of abalone (5 vs 9). For the digits data, the DBD based 1-NN performed worse than standard 1-NN, probably because of the difficulty of density estimation in very high dimensions. Interestingly, of the two abalone examples, randomized min-cut outperformed both NN algorithms in one case and under-performed the NN algorithms in the other case.

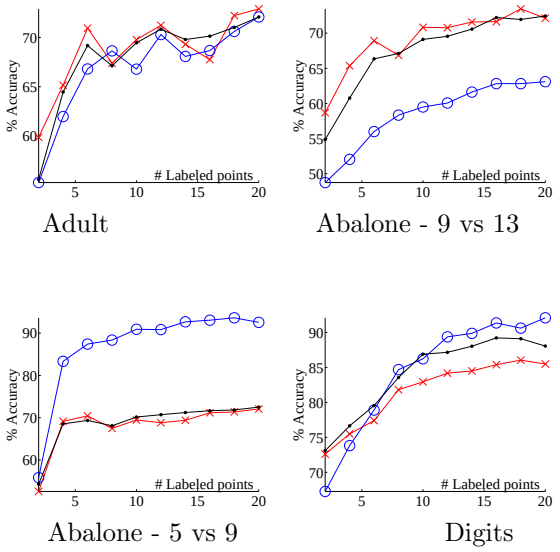

Figure 3. Classification results comparing 1-NN ('.'), DBD based 1-NN ('x') and randomized min-cut ('o') algorithms

6. Conclusions and future work

We have shown that density-based distance metrics which satisfy certain properties can be estimated consistently using an estimator obtained by plugging in the kernel density estimate of the data distribution. In terms of s, a smoothness parameter that corresponds to how many times data density is known to be differentiable and d, the data dimension, we showed that the rate of convergence of such an estimator is $\min(\frac{s}{s+d}, \frac{1}{2})$ and that rate $\frac{1}{2}$ is the fastest possible for any estimator. This contains both good and bad news. The knowledge that we have consistent estimation is useful when applying the method to voluminous data (e.g., web pages). Though the plug-in estimator achieves the best possible rate of convergence for sufficiently smooth density ($s \geq d$), our results do not guarantee that it does as well when density cannot be assumed to be very smooth. We also presented a graph construction which enables consistent computation of DBD metrics in time polynomial in sample size. We demonstrated the use of these metrics on artificial data and on data from the UCI repository.

There are several promising directions for future work. Study of alternative graph construction and weighting methods for more accurate and efficient computation will be of practical value. More experiments are needed to understand the utility of these distance measures for interpolation and semi-supervised classification. It will be interesting to develop algorithms for active learning under such prior assumption of similarity.

Acknowledgments

We thank Sanjoy Dasgupta and Thomas John for several helpful discussions.

References

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. *ICML 18*.

Blum, A., Lafferty, J. D., & Mugizi Robert Rwebangira, R. R. (2004). Semi-supervised learning using randomized mincuts. *Proc. 21th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Bousquet, O., Chapelle, O., & Hein, M. (2004). Measure based regularization. *NIPS 16*.

Bregler, C., & Omohundro, S. (1995). Nonlinear image interpolation using manifold learning. *NIPS 7*.

Corduneanu, A., & Jaakkola, T. (2003). On information regularization. *UAI 19*.

Gine, E., & Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *38*, 907–921.

Goldstein, L., & Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *The annals of statistics*, *20*, 1306–1328.

Lebanon, G. (2003). Learning riemannian metrics. *UAI 19*.

Nadaraya, E. (1989). *Nonparametric estimation of probability densities and regression curves*.

Saul, L. K., & Jordan, M. I. (1997). A variational principle for model-based morphing. *NIPS 9* (pp. 267–273).

Sethian, J. A. (1999). *Level set methods and fast marching methods*. Cambridge University Press.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of statistics*, *8*, 1348–1360.

Strichartz, R. (1995). *The way of analysis*. Jones and Bartlett.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.

Vincent, P., & Bengio, Y. (2003). Density-sensitive metrics and kernels. *Workshop on Advances in Machine Learning*.