Appendix to "Theoretical analysis of LLE based on its weighting step" published in the Journal of Computational and Graphical Statistics

Yair Goldberg and Ya'acov Ritov

Supplementary Proofs

S1 Proof of Lemma 3.1

Write $w_i = \sum_{m=d+1}^{K} a_m u_m = U_2 a$. The Lagrangian of the problem can be written as

$$L(a,\lambda) = \frac{1}{2} a' {U_2}' U_2 a + \lambda(\mathbf{1}' U_2 a - 1)\,.$$

Taking derivatives with respect to both a and λ, we obtain

$$\begin{aligned}\frac{\partial L}{\partial a} &= {U_2}' U_2 a - \lambda {U_2}'\mathbf{1} = a - \lambda {U_2}'\mathbf{1}\,,\\ \frac{\partial L}{\partial \lambda} &= \mathbf{1}' U_2 a - 1\,.\end{aligned}$$

Hence we obtain that $a = \frac{{U_2}'\mathbf{1}}{\mathbf{1}' U_2 {U_2}'\mathbf{1}}$.

S2 Proof of Theorem 5.1

The proof of Theorem 5.1 consists of two steps. First, we find a representation of the vector $\tilde{w}_i$, the weight vector of the perturbed neighborhood; see (14). Then we bound the distance between $\tilde{w}_i$ and w_i, the weight vector of the original neighborhood.

We start with some notations. For every matrix A, let $\lambda_j(A)$ be the j-th singular value of A. Note that $\|A\|_2 = \lambda_1(A)$. In this notation, we have $\lambda_j^i = \lambda_j(X_i)$. Denote by $T = X_i{}'X_i$ and $\widetilde{T} = \widetilde{X}_i'\widetilde{X}_i = T + \varepsilon(X_i{}'E_i + E_i{}'X_i) + \varepsilon^2 E_i{}'E_i$. Using the decomposition of (4), we may write $T = UL^2U'$ and $\widetilde{T} = \widetilde{U}\widetilde{L}^2\widetilde{U}'$. Note that $\lambda_j(T) = \lambda_j(X_i)^2$. Define U_2 and $\widetilde{U}_2$ to be the $K \times (K-d)$ matrices of the left-singular vectors corresponding to the lowest singular values, as in (4).

Note that by assumption, $\lambda_1(E_i) = 1$; hence, $\lambda_1(X_i{}'E_i) \leq \lambda_1^i \leq 1$. By Corollary 8.1-3 of Golub and Loan (1983),

$$\lambda_i(T) - 3\varepsilon \leq \lambda_i(\widetilde{T}) \leq \lambda_i(T) + 3\varepsilon\,. \tag{10}$$

Let $\delta = \lambda_d(T) - \lambda_{d+1}(T) - \varepsilon$. By Theorem 8.1-7 of Golub and Loan (1983), there is a $d \times (K-d)$ matrix Q such that $\|Q\|_2 \leq \frac{6\varepsilon}{\delta}$ and such that the columns of $\widehat{U}_2 = (U_2 + U_1Q)(I + Q'Q)^{-1/2}$ are an orthogonal basis for an invariant subspace of $\widetilde{T}$. We want to show that $\widehat{U}_2$ and $\widetilde{U}_2$ span the same subspaces. To prove this, we bound the largest singular value of $\|\widehat{U}_2'\widetilde{T}\widehat{U}_2\|_2$, and the result follows from (10).

First, note that

$$1 - \frac{6\varepsilon}{\delta} < \lambda_j\left((I + Q'Q)^{-1/2}\right) < 1 + \frac{6\varepsilon}{\delta}\,. \tag{11}$$

Hence,

$$\begin{aligned}
\left\|\widehat{U}_2'\widetilde{T}\widehat{U}_2\right\|_2 &= \left\|(I + Q'Q)^{-1/2}(U_2 + U_1Q)'\widetilde{T}(U_2 + U_1Q)(I + Q'Q)^{-1/2}\right\|_2 \\
&\leq \left(1 + \frac{6\varepsilon\lambda_1^i}{\delta}\right)^2 \left(\left\|U_2'\widetilde{T}U_2\right\|_2 + 2\left\|U_2'\widetilde{T}U_1Q\right\|_2 + \left\|Q'U_1'\widetilde{T}U_1Q\right\|_2\right) \\
&\leq \left(1 + \frac{6\varepsilon}{\delta}\right)^2 \left((\lambda_{d+1}(T) + 3\varepsilon) + \frac{(6\varepsilon)^2}{\delta} + \left(\frac{6\varepsilon}{\delta}\right)^2 (1 + 3\varepsilon)\right)
\end{aligned} \tag{12}$$

We now obtain some bounds on the size of ε. By the theorem assumption we have $\varepsilon < \frac{(\lambda_d^i)^4}{72}$. Since Assumption (A1) holds, we may assume that $\lambda_{d+1}(T) < \frac{\lambda_d(T)}{72}$. Recall that $\delta = \lambda_d(T) - \lambda_{d+1}(T) - \varepsilon$ and that $(\lambda_d^i)^2 = \lambda_d(T)$. Isolating ε we obtain that $\varepsilon < \frac{\lambda_d(T)\delta}{60}$. Similarly, we can show that $\varepsilon < \frac{\delta^2}{60}$. We also have that $\varepsilon < \frac{\lambda_d(T)}{72}$, since by assumption $\lambda_d(T) < 1$, and similarly,

$\varepsilon < \frac{\delta}{60}$. Summarizing, we have

$$\varepsilon < \min\left(\frac{\delta}{60}, \frac{\lambda_d(T)}{72}, \frac{\lambda_d(T)\delta}{60}, \frac{\delta^2}{60}\right) . \tag{13}$$

We are now ready to bound the expression in (12). We have that $(1+\frac{6\varepsilon}{\delta}) < \frac{11}{10}$ since $\varepsilon < \frac{\delta}{60}$; $\lambda_{d+1}(T) < \frac{\lambda_d(T)}{72}$ by assumption; $3\varepsilon < \frac{\lambda_d(T)}{24}$ since $\varepsilon < \frac{\lambda_d(T)}{72}$; $\frac{(6\varepsilon)^2}{\delta} < \frac{\lambda_d(T)}{120}$ since $\varepsilon < \frac{\delta}{60}$ and also $\varepsilon < \frac{\lambda_d(T)}{72}$; $\frac{(6\varepsilon)^2}{\delta^2} < \frac{\lambda_d(T)}{100}$ since $\varepsilon < \frac{\lambda_d(T)\delta}{60}$ and $\varepsilon < \frac{\delta}{60}$; $118\frac{\varepsilon^3}{\delta^2} < \frac{\lambda_d(T)}{1000}$ since $\varepsilon < \frac{\delta}{60}$ and $\varepsilon < \frac{\lambda_d(T)}{72}$. Combining all these bounds, we obtain that

$$\left\|\widehat{U}_2'\widetilde{T}\widehat{U}_2\right\|_2 < \frac{\lambda_d(T)}{10} < \lambda_d(T) - 3\varepsilon .$$

Hence, by (10) we have that $\left\|\widehat{U}_2'\widetilde{T}\widehat{U}_2\right\|_2 < \lambda_d(\widetilde{T})$. Since $\widehat{U}_2$ spans a subspace of $K-d$ dimension, it must span the subspace with the $K-d$ vectors with lowest singular values of $\widetilde{T}$. In other words, $\widehat{U}_2$ spans the same subspace as $\widetilde{U}_2$ or, equivalently, $\widehat{U}_2\widehat{U}_2' = \widetilde{U}_2\widetilde{U}_2'$. Summarizing, we obtain that

$$\tilde{w}_i = \frac{\widehat{U}_2\widehat{U}_2'\mathbf{1}}{\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1}} . \tag{14}$$

We are now ready to bound the difference between w_i and $\tilde{w}_i$.

$$\begin{aligned} \|w_i - \tilde{w}_i\|^2 &= \left\|\frac{U_2{U_2}'\mathbf{1}}{\mathbf{1}'U_2{U_2}'\mathbf{1}} - \frac{\widetilde{U}_2\widehat{U}_2'\mathbf{1}}{\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1}}\right\|^2 \\ &= \frac{1}{\mathbf{1}'U_2{U_2}'\mathbf{1}} - 2\frac{\mathbf{1}'U_2{U_2}'\widehat{U}_2\widehat{U}_2'\mathbf{1}}{\mathbf{1}'U_2{U_2}'\mathbf{1}\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1}} + \frac{1}{\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1}} \\ &= \frac{\mathbf{1}'(U_2-\widehat{U}_2)(U_2-\widehat{U}_2)'\mathbf{1}}{\mathbf{1}'U_2{U_2}'\mathbf{1}\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1}} \end{aligned}$$

We use Assumption (A2) to obtain a bound on $\mathbf{1}'U_2{U_2}'\mathbf{1}$. Denote the projection of the normalized vector $\frac{1}{\sqrt{K}}\mathbf{1}$ on the basis $\{u_j\}$ by $p_j = \frac{1}{\sqrt{K}}\mathbf{1}'u_i$. We have that

$$\|\mu_i\|^2 = \frac{1}{K}\left\|\frac{1}{\sqrt{K}}1'U_1L_1\right\|^2 = \frac{1}{K}\sum_{j=1}^{d}\left(p_j\lambda_j^i\right)^2 .$$

By Assumption (A2), $\|\mu_i\|^2 < \frac{\alpha}{K}\left(\lambda_d^i\right)^2$. Hence $\sum_{j=1}^{d} p_j^2 < \alpha$. Since $\sum_{j=1}^{K} p_j^2 = 1$, we have that

$$\sum_{j=d+1}^{K} p_j^2 = \frac{1}{K}\mathbf{1}' U_2 {U_2}' \mathbf{1} > 1 - \alpha\,. \tag{15}$$

Similarly, we obtain a bound on $\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1}$.

$$\begin{aligned}
\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1} &\geq \left\|(I+Q'Q)^{-1/2}U_2'\mathbf{1}\right\|^2 - 2\left|\mathbf{1}'U_1 Q(I+Q'Q)^{-1}U_2'\mathbf{1}\right| \\
&\geq (1-\frac{6\varepsilon}{\delta})^2 K(1-\alpha) - 2K\frac{6\varepsilon}{\delta}(1+\frac{6\varepsilon}{\delta})^2(1-\alpha)^{1/2} \\
&\geq \frac{9K(1-\alpha)}{10} - 12K\frac{\varepsilon}{\delta}\left(\frac{11}{10}\right)^2(1-\alpha)^{1/2}\,,
\end{aligned}$$

where we used $\varepsilon < \frac{\delta}{60}$. Since by assumption $\varepsilon < \frac{\lambda_d(T)\sqrt{(1-\alpha)}}{72}$, and using the facts that $\lambda_{d+1}(T) < \frac{\lambda_d(T)}{72}$ and $\varepsilon < \frac{\lambda_d(T)}{72}$, we obtain that $\varepsilon < \frac{\delta\sqrt{(1-\alpha)}}{60}$. Hence, $\mathbf{1}'\widehat{U}_2\widehat{U}_2'\mathbf{1} \geq \frac{K(1-\alpha)}{2}$.

Finally, we obtain a bound on $\mathbf{1}'(U_2-\widehat{U}_2)(U_2-\widehat{U}_2)'\mathbf{1}$.

$$\begin{aligned}
\left\|U_2 - \widehat{U}_2\right\|_2 &= \left\|U_2(I-(I+Q'Q)^{-1/2}) + U_1 Q(I+Q'Q)^{-1/2}\right\|_2 \\
&\leq \|U_2\|_2\left\|I-(I+Q'Q)^{-1/2}\right\|_2 + \|U_1\|_2\|Q\|_2\left\|(I+Q'Q)^{-1/2}\right\|_2 \\
&\leq \frac{6\varepsilon}{\delta} + \frac{6\varepsilon}{\delta}(1+\frac{6\varepsilon}{\delta}) = \frac{6\varepsilon}{\delta}(2+\frac{6\varepsilon}{\delta})\,,
\end{aligned}$$

where the last inequality follows from (11), the fact that for any eigenvector v of $(I+Q'Q)^{-1/2}$ with eigenvalue λ_v, v is also eigenvector of $I-(I+Q'Q)^{-1/2}$ with eigenvalue $1-\lambda_v$, and the fact that $\|A\|_2 = 1$ for every matrix A with orthonormal columns (see Golub and Loan, 1983). Consequently,

$$\left\|(U_2-\widehat{U}_2)'\mathbf{1}\right\|_2 \leq K\frac{6\varepsilon}{\delta}\left(2+\frac{6\varepsilon}{\delta}\right) < \frac{13K\varepsilon}{\delta}\,,$$

where we used $\varepsilon < \frac{\delta}{60}$.

Combining these results, we have that

$$\|w_i - \tilde{w}_i\| < \frac{(13K\varepsilon)/\delta}{(K(1-\alpha))/\sqrt{2}} < \frac{20\varepsilon}{\lambda_d(T)(1-\alpha)}\,,$$

where we used $\frac{21}{20\lambda_d(T)} > \frac{1}{\delta}$.

S3 Proof of Theorem 5.2

Since $\Phi(Z) = \sum_{i=1}^{n} \left\| \sum_j w_{ij}(z_j - z_i) \right\|^2$, we bound each summand separately in order to obtain a global bound.

Let the induced neighbors of $z_i = f^{-1}(x_i)$ be defined by $(\tau_1, \ldots, \tau_K) = (f^{-1}(\eta_1), \ldots, f^{-1}(\eta_K))$. Note that apriori, it is not clear that τ_j are neighbors of z_i. Let J be the Jacobian of the function f at z_i. Since f is a conformal mapping, $J'J = c(z_i)I$, for some positive $c : \Omega \to \mathbb{R}$. Using first-order approximation we have that $\eta_j - x_i = J(\tau_j - z_i) + \mathcal{O}\left(\|\tau_j - z_i\|^2\right)$. Hence, for w_i we have that

$$\sum_{j=1}^{K} w_{ij}(\tau_j - z_i) = \sum_{j=1}^{K} w_{ij} J'(\eta_j - x_i) + \mathcal{O}\left(\max_j \|\tau_j - z_i\|^2\right) . \tag{16}$$

Thus we have that

$$\left\| \sum_{j=1}^{K} w_{ij}(\tau_j - z_i) \right\|^2 = \left\| \sum_{j=1}^{K} w_{ij} J'(\eta_j - x_i) \right\|^2 + \left\| \sum_{j=1}^{K} w_{ij} J'(\eta_j - x_i) \right\| \mathcal{O}\left(\max_j \|\tau_j - z_i\|^2\right) . \tag{17}$$

We bound $\left\| \sum_{j=1}^{K} w_{ij} J'(\eta_j - x_i) \right\|$ for the vector w_i that minimizes (5). Note that by (4), $\sum_{j=1}^{K} w_{ij} J'(\eta_j - x_i) = w_i' X_i^P J + w_i' U_2 L_2 V_2' J$. However, by construction $w_i' X_i^P = 0$. Hence

$$\left\| \sum_{j=1}^{K} w_{ij} J'(\eta_j - x_i) \right\| = \|w_i' U_2 L_2 V_2' J\| \leq \|w_i\| \, \|U_2 L_2 {V_2}' J\|_2 \leq \frac{\|w_i\| \, \lambda_{d+1}^i}{\sqrt{c(z_i)}} ,$$

where we used the facts that $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$ for a any matrix A, and that $\|A\|_2 = 1$ for a matrix A with orthonormal columns (for both claims, see Golub and Loan, 1983, Section 2). Substituting in (17), we obtain that

$$\left\| \sum_{j=1}^{K} w_{ij}(\tau_j - z_i) \right\|^2 \leq \frac{\|w_i\|^2 \, (\lambda_{d+1}^i)^2}{c(z_i)} + \|w_i\| \, \lambda_{d+1}^i \mathcal{O}\left(\max_j \|\tau_j - z_i\|^2\right) .$$

Since Assumption (A2) holds, it follows from (15) that $\|w_i\|^2 = \frac{1}{\mathbf{1}' U_2 {U_2}' \mathbf{1}} < \frac{1}{K(1-\alpha)}$.

As f is a conformal mapping, we have that $c_{\min} \|\tau_j - z_i\| \leq d_{\mathcal{M}}(\eta_j, x_i)$, where $d_{\mathcal{M}}$ is the geodesic metric and $c_{\min} > 0$ is the minimum of the scale function $c(z)$ that measures the scaling change of f at z. The minimum $c_{\min}$ is attained as Ω is compact. The last inequality holds true since the geodesic distance $d_{\mathcal{M}}(\eta_j, x_i)$ is equal to the integral over $c(z)$ for some path between τ_j and z_i.

The sample is assumed to be dense; hence $\|\tau_j - x_i\| < s_0$, where s_0 is the *minimum branch separation* (see Section 5). Using Lemma 3 of Bernstein et al. (2000), we conclude that

$$\|\tau_j - z_i\| \leq \frac{1}{c_{\min}} d_{\mathcal{M}}(\eta_j, x_i) < \frac{\pi}{2c_{\min}} \|\eta_j - x_i\| \, . \tag{18}$$

Since Assumption (A1) holds, and

$$r(i)^2 = \max_j \|\eta_j - x_i\|^2 \geq \frac{1}{K}\sum_{j=1}^{K} \|\eta_j - x_i\|^2 = \|X_i\|_F^2 = \frac{1}{K}\sum_{j=1}^{K} (\lambda_j^i)^2 \geq \frac{d}{K}(\lambda_d^i)^2 \, ,$$

we have that $\lambda_{d+1} \ll r(i)$. Hence $\left\|\sum_{j=1}^{K} w_{ij}(\tau_j - z_i)\right\|^2 = \lambda_{d+1}^i \mathcal{O}\left(r(i)^2\right)$.

S4 Proof of Theorem 5.3

Before we start the proof, we need some additional notation. We say that $a_n = \boldsymbol{O}_p(c_n)$ if $a_n = o_p(c_n n^{\alpha})$ for any $\alpha > 0$ (and typically, but not necessarily, $c_n = o_p(a_n)$). We say that $a_n = \boldsymbol{\Omega}_p(c_n)$ if both $a_n = \boldsymbol{O}_p(c_n)$ and $c_n = \boldsymbol{O}_p(a_n)$. That is, if a_n and c_n are equal up to a slowly varying factor.

Let $N_i = \{j : \|x_j - x_i\| < r\} \equiv \{i_1, \ldots, i_{K_i}\}$ where $K_i = |N_i|$ is the size of x_i's neighborhood. Let the embedding function $e_i : \mathbb{R}^{K_i} \to \mathbb{R}^n$ be defined as $e_i(v) = \sum_{k=1}^{K_i} v_i \mathrm{e}_{i_k}$ where e_j is the j-th member of the standard basis of $\mathbb{R}^n$. When e_i is applied to a matrix, it is understood that it is applied to each of its columns.

Note that for a given i, K_i is a binomial random variable, with parameter n and $\int_{\|x - x_i\| < r} g(x)dx$, where g is the sampling density. Thus, $EK_i = \mathcal{O}(nr^d)$, and $K_i = \mathcal{O}_p(nr^d)$. Since g is bounded from above and away from zero, and no more than n means are considered, $K_i = \boldsymbol{\Omega}_p(nr^d)$ uniformly. That is, both $\max_i K_i = \boldsymbol{\Omega}_p(nr^d)$ and $\min_i K_i = \boldsymbol{\Omega}_p(nr^d)$. Similarly, all convergence statements below are regarding $\mathcal{O}_p(n)$ means and hold uniformly over all neighborhoods (and hence a slowly varying factor is needed in their statement).

We are now ready to start the proof. Let $x_i = f(z_i)$ and assume that $\text{dist}(x_i, f(\partial\Omega)) > r$. Let X_i be the neighborhood of x_i, and note that that the rows of X_i are drawn from a continuous bounded density, and thus are asymptotically uniformly spread on $B(x_i, r) \cap f(\Omega)$. Let $U_{i1} = X_i V_{i1} L_{i1}^{-1} \in \mathbb{R}^{K_i \times d}$ be the neighborhood X_i after projection on the first d-directions and rescaling, where U_{i1}, L_{i1}, and V_{i1} are defined as in (4). Note that the columns of U_{i1} are of norm 1 in $\mathbb{R}^{K_i}$, and hence its rows are uniformly distributed on a ball of $\mathbb{R}^d$ of radius $\boldsymbol{O}_p(1/\sqrt{K_i})$ up to some errors due to the stochastic distribution of the points, the curvature of the manifold, and the change in the density. These errors are $\boldsymbol{\Omega}_p(K_i^{-1/2})$, $\boldsymbol{O}_p(1/K_i)$ (the difference between the projection on the tangent and geodesic distance within a ball with radius scaled to $\mathcal{O}_p(1/\sqrt{K_i})$), and $\boldsymbol{O}_p(1/K_i)$ (since the distribution is uniform up to a linear $\boldsymbol{O}_p(1/\sqrt{K_i})$ term), respectively.

We now characterize the weight vector w_i for any inner point z_i. Recall that by Lemma 3.1,

$$w_i = \frac{(I - U_{i1}U_{i1}')\mathbf{1}}{\mathbf{1}'(I - U_{i1}U_{i1}')\mathbf{1}}$$

where $\mathbf{1}$ is the vector of ones of length K_i and I is the $K_i \times K_i$ identity matrix. Let $\{U_{i1}^{(1)}, \ldots, U_{i1}^{(d)}\}$ be the d columns of U_{i1}. Note that up to an $\mathcal{O}(1/K_i)$ error, the points of $U_{in}^{(m)}$, $m = 1, \ldots, d$ are a projection of points that are uniformly distributed in a d-dimensional ball of radius $\boldsymbol{\Omega}_p(1/\sqrt{K_i})$, and thus are distributed according to some symmetric distribution on a segment of length $\boldsymbol{\Omega}_p(2/\sqrt{K_i})$. By the symmetry and the size of the error, $\mathbf{1}'U_{i1}^{(m)} = \boldsymbol{O}_p(1)$ and hence also $\mathbf{1}'(U_{i1}U_{i1}')\mathbf{1} = \boldsymbol{O}_p(1)$ and the components of $U_{i1}U_{i1}'\mathbf{1}$ are of magnitude $\boldsymbol{O}_p(1/\sqrt{K_i})$. Since $\mathbf{1}'I\mathbf{1} = K_i$, we conclude that

$$w_{ij} = \begin{cases} 1/K_i + \boldsymbol{O}_p(K^{-3/2}) & \|x_j - x_i\| < r \\ 0 & \text{otherwise} \end{cases} . \tag{19}$$

We would like to compare the embedding in $\mathbb{R}^n$ of weight vectors of two close-by points x_i and x_j, such that $\|x_i - x_j\| < \rho$. Note that the minimal number of points within a ball of radius ρ centered on one of the observations is increasing to infinity with probability converging to 1, yet it is a small fraction of the number of observations within the radius r balls, and that adjacent neighborhoods mostly overlap: $\max_{i,j:\|x_i - x_j\| < \rho} |N_i \ominus N_j|/|N_i| = \boldsymbol{O}_p(\rho/r)$, where $\ominus$ denotes the symmetric difference (the cardinality of the symmetric difference is bounded by the number of points in the shell between the spheres with radius $r-\rho$ and $r+\rho$). We conclude

$$\begin{aligned}\max_{\{j:\|x_i-x_j\|<\rho\}} \|e_i(w_i)-e_j(w_j)\|^2 &\le \sum_{k\in N_i\cap N_j} ((e_i(w_i)_k - e_j(w_j)_k))^2 \\ &\quad + \sum_{k\in N_i\ominus N_j} (e_i(w_i)_k - e_j(w_j)_k)^2 \\ &= \boldsymbol{O}_p(K\cdot K^{-3} + K\rho/r\cdot K^{-2}) = \boldsymbol{O}_p(\rho/(rK))\,. \end{aligned} \tag{20}$$

Next, recall that the embedding $Y_n = \{y_1,\dots,y_n\}$ is given by the $2,\dots,d+1$ lowest eigenvectors of $I-M \equiv (I-W)'(I-W)$, where

$$Y_{(m)i} = (1-\lambda_m)^{-1}\sum_{k=1}^n M_{ik}Y_{(m)k} \tag{21}$$

(see Saul and Roweis, 2003, Section 4). We would like to show that the matrix M inherits the continuity property from W. In other words, whenever $\|x_i-x_j\|<\rho$ we have

$$\sum_{k=1}^n |M_{ik}-M_{jk}|^2 = \boldsymbol{O}_p(\rho/(rK)) = \boldsymbol{O}_p(\rho n^{-1}r^{-(d+1)})\,. \tag{22}$$

Indeed, let $\|x_i - x_j\| < \rho$ such that $\mathrm{dist}(z_i, \partial\Omega) > 2r + \rho$. Then

$$\begin{aligned}
\sum_{k=1}^n |M_{ik} - M_{jk}|^2 &= \sum_{k=1}^n \Big((W_{ik} - W_{jk}) + (W_{ki} - W_{kj}) - \sum_{s=1}^n W_{sk}(W_{si} - W_{sj}) \Big)^2 \\
&\le 3 \sum_{k=1}^n \Big((W_{ik} - W_{jk})^2 + (W_{ki} - W_{kj})^2 + (\sum_{s=1}^n W_{sk}(W_{si} - W_{sj}))^2 \Big) \\
&= \boldsymbol{O}_p(\rho/(rK)) + 3 \sum_{k \in N_i \cap N_j} (e_k(w_k)_i - e_k(w_k)_j)^2 + \sum_{k \in N_i \ominus N_j} (e_k(w_k)_i - e_k(w_k)_j)^2 \\
&\quad + 3 \sum_{k=1}^n \sum_{s=1}^n \sum_{t=1}^n W_{sk} W_{tk} (W_{si} - W_{sj})(W_{ti} - W_{tj}) \\
&= \boldsymbol{O}_p(\rho/(rK)) + 3 \sum_{t,s \in N_i \cap N_j} (W_{si} - W_{sj})(W_{ti} - W_{tj}) \sum_{k \in N_s \cap N_t} W_{sk} W_{tk} \\
&\quad + 3 \sum_{t,s \in N_i \ominus N_j} (W_{si} - W_{sj})(W_{ti} - W_{tj}) \sum_{k \in N_s \cap N_t} W_{sk} W_{tk} \\
&\quad + 3 \sum_{s \in N_i \cap N_j; t \in N_i \ominus N_j} (W_{si} - W_{sj})(W_{ti} - W_{tj}) \sum_{k \in N_s \cap N_t} W_{sk} W_{tk} \\
&\equiv \boldsymbol{O}_p(\rho/rK) + (A) + (B) + (C)
\end{aligned}$$

Recall that by assumption $\mathrm{dist}(z_i, \partial\Omega) > 2r + \rho$, and hence for every $s \in N_i \cup N_j$, the respective distance of x_s and $f(\partial\Omega)$ is at least r (see (18)), thus we can use the bound (20). We now bound the expressions (A), (B), and (C). Note that $\sum_{k \in N_s \cap N_t} W_{sk} W_{tk} = \boldsymbol{O}_p(K^{-1})$, and that for $s, t \in N_i \cap N_j$, both $(W_{si} - W_{sj})$ and $(W_{ti} - W_{tj})$ equal $\boldsymbol{O}_p(K^{-3/2})$. Since there are less then K_i^2 pairs s, t in $N_i \cap N_j$, we conclude that $(A) = \boldsymbol{O}_p(K^{-2})$.

For (B), note that there are $\boldsymbol{O}_p((K\rho/r)^2)$ pairs of points $t, s \in N_i \ominus N_j$, and that for these points, $(W_{si} - W_{sj})$ and $(W_{ti} - W_{tj})$ are $1/K + \boldsymbol{O}_p(K^{-3/2})$. We conclude that $(B) = \boldsymbol{O}_p(K^{-1}(\rho/r)^2)$. Similarly, it can be shown that $(C) = \boldsymbol{O}_p(K^{-1}\rho/r)$. Summarizing we obtain (22).

Denote the columns of the embedding Y by $\{Y^{(1)}, \ldots, Y^{(d)}\}$ and similarly for the pre-image Z. Recall that $\frac{1}{n} Y^{(m)\prime} Y^{(m)} = 1$, and that $\|(I - W)Y^{(m)}\| = Y^{(m)\prime} M Y^{(m)}$ minimizes the norm $\|(I - W)v\|$ over all vectors v such that $n^{-1} v'v = 1$ which are not in the span of $\{\mathbf{1}, Y^{(1)}, \ldots, Y^{(m-1)}\}$. On the other hand, by Theorem 5.2, there are d normalized vectors, namely $Z^{(1)}, \ldots, Z^{(d)} \in \mathbb{R}^n$, and $\zeta_n \xrightarrow{p} 0$, such that $\|(I - W)Z^{(m)}\| < \zeta_n$. Therefore, $I - M$ has at least $d + 1$ eigenvalues (including 0) less than ζ_n. Since

$(I-M)Y^{(m)} = \lambda_m Y^{(m)}$ for $|\lambda_m| < \zeta_n$, we obtain that

$$Y_i^{(m)} = (1-\lambda_m)^{-1} \sum_{k=1}^{n} M_{ik} Y_k^{(m)} \tag{23}$$

Let $\|x_i - x_j\| < \rho$, then

$$\begin{aligned} \left(Y_i^{(m)} - Y_j^{(m)}\right)^2 &= (1-\lambda_m)^{-2} \Big(\sum_{k=1}^{n} (M_{ik} - M_{jk}) Y_k^{(m)} \Big)^2 \\ &\leq (1-\lambda_m)^{-2} \Big(\sum_{\{k: M_{ik} \neq 0\}} (M_{ik} - M_{jk})^2 \Big) \sum_{k=1}^{n} \left(Y_k^{(m)}\right)^2 \\ &= \boldsymbol{O}_p(\rho / r K_i) \cdot n = \boldsymbol{O}_p(\rho / r^{d+1}) \,, \end{aligned} \tag{24}$$

where the first inequality follows from application of Cauchy-Schwarz, and the equalities in the third line follow from (22), the assumptions on ρ, and the fact that $\|Y\|^2 = n$.

Using Lemma 3 of Bernstein et al. (2000), we have

$$\|\eta_j - x_i\| \leq d_{\mathcal{M}}(\eta_j, x_i) \leq c_{\max} \|\tau_j - z_i\| \ . \tag{25}$$

Thus, if $\|z_i - z_j\| < \rho_o$ then $\|x_i - x_j\| < \rho_o / c_{\max} \equiv \rho$. If $nr^{d(d+1+\eta)} \to \infty$, we can take $\rho = r^{d+1+\eta}$ (note that $n\rho^d \to \infty$) and (9) holds.

Now sum (24) over all points within $2r$ from the boundary. Since $Y_i^{(m)}$ is included in $\boldsymbol{O}_p(K_i)$ terms, we obtain for any $\rho \ll r$:

$$\frac{1}{n} \sum_{\{i: \mathrm{dist(x_i, \partial\Omega) > 2r + \rho}\}} \max_{\{j: \|x_i - x_j\| < \rho\}} (Y_i^{(m)} - Y_j^{(m)})^2 \leq \boldsymbol{O}_p(\rho/r) \frac{1}{n} \sum_{k=1}^{n} \left(Y_k^{(m)}\right)^2 = \boldsymbol{O}_p(\rho/r) \,,$$

and (8) holds.

S5 Proof of Theorem 5.4

Consider first the local description of the curve. Let z_i be the pre-image of the i-th point. Since the curve f can be reparameterized, without loss of generality, we assume that the mapping is isometric. We also assume that $z_1 \leq \cdots \leq z_n$. Thus $z_j - z_i$ is the geodesic distance between x_i and x_j along the curve. Let ξ_{ij} be the projection of the j-th point on the tangent line at

x_i, and let $r_i = \mathcal{O}_p(K/n)$ be the radius of the i-th neighborhood. Since the curvature is bounded, difference between the arc length and its projection is of order r_i^3, or $\xi_{ij} \le |z_j - z_i| = \xi_{ij} + \mathcal{O}_p((K/n)^3)$, uniformly (see, for example, Belkin, 2003, Lemma 4.2.1). By construction $\sum_j w_{ij}\xi_{ij} = 0$ while $\sum_j |w_{ij}| = \mathcal{O}_p(1)$.

Looking more closely at the description of each point by its neighbors and at the relation to the curvature of the curve, we have for any point i (including both inner points and boundary points) that

$$\sum_j w_{ij}(z_j - z_i) = \sum_j w_{ij}\xi_{ij} + \boldsymbol{O}_p\big((K/n)^3\big) = \boldsymbol{O}_p\big((K/n)^3\big)\,. \tag{26}$$

This result can be strengthened for inner points. Since the conditions of Theorem 5.3 hold, we have for all $j \in N_i$, $W_{ij} = 1/2K + \boldsymbol{O}_p(K^{-3/2})$. Hence,

$$\begin{aligned}\sum_j w_{ij}(z_j - z_i) &= \sum_j w_{ij}(z_j - z_i - \xi_{ij}) \\ &= \frac{1}{2K}\sum_j (z_j - z_i - \xi_{ij}) + \boldsymbol{O}_p\big(K^{-1/2}(K/n)^3\big) \\ &= \boldsymbol{O}_p\big(K^{-1/2}(K/n)^3\big),\end{aligned} \tag{27}$$

where we used the fact that $(z_j - z_i - \xi_{ij}) = \mathcal{O}_p((K/n)^3)$. Since all but $2K$ are inner points, we obtain by combining (26) and (27) that

$$\|(I - W)Z\|^2 = \boldsymbol{O}_p\big(nK^5/n^6 + KK^6/n^6\big)\,. \tag{28}$$

We would like to bound $\|(I - W)Y\|$. Note that

$$\begin{aligned}\|(I - W)Y\| &= n^{1/2}\min\{\|(I - W)\xi\| : \mathbf{1}'\xi = 0, \|\xi\|^2 = 1\} \\ &\le n^{1/2}\|(I - W)Z\|/\|Z\| = \boldsymbol{O}_p(K^{7/2}/n^3).\end{aligned} \tag{29}$$

Here we used (28), and the fact that $\|Z\|^2 = (1 + \mathcal{O}_p(n^{1/2}))n$. As a result we also obtain that the second smallest eigenvalue of $M \equiv (I - W)'(I - W)$ is $\lambda = \mathcal{O}_p\big((K/n)^7\big)$ (recall that the smallest is zero, see Saul and Roweis, 2003, page 17).

Write $Y = WY + e$, and note that by (29), $\|e\| = \boldsymbol{O}_p(K^{7/2}/n^3)$ (note that by definition $\|Y\| = \sqrt{n}$). Iterating this equation we obtain

$$Y = W^m Y + (I + W + \cdots + W^{m-1})e, \quad m = 1, 2, \ldots \tag{30}$$

The sum of the entries of the rows of W are all 1. All rows i except the first and last K rows are both positive and very close to $1/(2K)$ over the indices $i-K,\dots,i+K$. Hence $W_{i\cdot}$ can be considered as a probability vector, with weights that are close to uniform on $i-K,\dots,i+K$ and 0 otherwise. To be more exact, this distribution has mean $i+\boldsymbol{O}_p(K^{-1/2})$ and standard deviation $\bigl(1+\mathbb{o}_p(1)\bigr)K$. The inner rows of $W^m_{ik}=\sum_j W_{ij}W^{m-1}_{jk}$ are convolutions of these distributions. Hence all but the first and last mK rows of W^m become closer and closer to Gaussian distribution with standard deviation $\sqrt{m}K$ centered on the diagonal. In particular $0\le W^m_{ij}=\boldsymbol{O}_p(m^{-1/2}K^{-1})$ for all $mK<i<n-mK$. By (19), $W_{ij}=\bigl(1+\boldsymbol{O}_p(K^{-1/2})\bigr)\bar W_{ij}$, where $\bar W_{ij}$ is either 0 or $(2K)^{-1}$. But W^m_{ij} is a sum of products of positive terms, which are entries of $\bar W$ up to a factor of $\bigl(1+\boldsymbol{O}_p(K^{-1/2})\bigr)^m$. Hence the inner entries of W^m are close to those of $\bar W^m$, i.e., are close to convolutions of uniform distribution vectors, up to the factor of $\bigl(1+\boldsymbol{O}_p(K^{-1/2})\bigr)^m$. In other words

$$\begin{aligned} W^m_{ij}&=\bigl(1+\boldsymbol{O}_p(mK^{-1/2})\bigr)\bar W^m_{ij}, \quad mK<i,j<n-mK\\ \max_j |W^m_{ij}|&=\boldsymbol{O}_p(m^{-1/2}K^{-1}), \quad mK<i<n-mK\,. \end{aligned} \tag{31}$$

Hence, for $0<|j-i|\le K$:

$$\begin{aligned} |W^m_{ik}-W^m_{jk}|&\le|\bar W^m_{ik}-\bar W^m_{jk}|+\boldsymbol{O}_p(mK^{-1/2})(\bar W^m_{ik}+\bar W^m_{jk}).\\ &=\boldsymbol{O}_p(|j-i|m^{-1/2}K^{-1}+mK^{-1/2})(\bar W^m_{ik}+\bar W^m_{jk})\\ &=\boldsymbol{O}_p(m^{-1/2}+mK^{-1/2})(\tilde W^m_{ik})\,, \end{aligned} \tag{32}$$

where $\tilde W_{ik}=2\bar W^m_{ik}+(\bar W^m_{jk}-\bar W^m_{ik})$. Note that $\bigl|\tilde W_{ik}\bigr|\le 2\bigl|\bar W^m_{ik}\bigr|+\sup_{l\le k}\bigl|\bar W^m_{lk}-\bar W^m_{ik}\bigr|$, and thus the distribution of $\tilde W_{ik}$ can be approximated by twice the sum of the normal density plus the absolute value of its derivative divided by $\sqrt{m}$. In particular $\sum_k|\tilde W_{ik}|=O(1)$ and $\max_k|\tilde W_{ik}|=O(m^{-1/2}K^{-1})$. On the other hand, for every $\ell=1,\dots,m-1$, and for every $i=mK+1,\dots,n-mK-1$,

$$\begin{aligned} |(W^\ell e)_i|^2&=\Bigl(\sum_j W^\ell_{ij}e_j\Bigr)^2\\ &\le\sum_j (W^\ell_{ij})^2\|e\|^2 \qquad\qquad \text{(by Cauchy-Schwarz)}\\ &\le\boldsymbol{O}_p((\ell^{-1/2}K^{-1}K^7/n^6)=\boldsymbol{O}_p(\ell^{-1/2}K^6/n^6)\,, \end{aligned}$$

where the last inequality holds since for each inner point

$$\sum_j (W^\ell_{ij})^2 \le \max_j |W^\ell_{ij}| \sum_j |W^\ell_{ij}|| \\ = \boldsymbol{O}_p(\ell^{-1/2}K^{-1})\,. \tag{33}$$

Summing this expression for $\ell = 1, \dots, m-1$ we obtain from (30)

$$\max_{mK<i<n-mK} \|y_i - \sum_j W^m_{ij} y_j\| = \mathcal{O}_p(m^{3/4}K^3/n^3).$$

Using (32), (33), and the fact that $\|Y\|^2 = n$, we obtain that for every $Km < i < n - mK$, $i < j < i + 2K$:

$$\begin{aligned}
|y_j - y_i|^2 &\le \big|\sum_k (W^m_{ik} - W^m_{jk}) y_k\big|^2 + \boldsymbol{O}_p(m^{3/2}K^6/n^6) \\
&\le \sum_k (W^m_{ik} - W^m_{jk})^2 \sum_k |y_k|^2 + \boldsymbol{O}_p(m^{3/2}K^6/n^6) \\
&\le n\boldsymbol{O}_p(m^{-1} + m^2K^{-1}) \sum_k (\tilde{W}^m_{ik})^2 + \boldsymbol{O}_p(m^{3/2}K^6/n^6) \quad \text{(using } \|Y\|^2 = n \text{ and (32))} \\
&= \boldsymbol{O}_p(\frac{n}{m^{3/2}K} + \frac{nm^{3/2}}{K^2} + \frac{m^{3/2}K^6}{n^6}) \qquad \text{(by (33))} \\
&= \boldsymbol{O}_p\big((K/n)^{1/2}\big),
\end{aligned} \tag{34}$$

by taking m to n/K divided by a slowly varying function. We conclude that the maximal difference $|y_j - y_i|$ for two interior neighboring points, at most $2K$ points apart, converges to 0.

Write now $M = I - W - H$, where $H = W' - W'W$. Note that $\sum_j H_{ij} = \sum_j W_{ji} - \sum_j \sum_k W_{ki} W_{kj} = \sum_j W_{ji} - \sum_k W_{ki} = 0$, $W_{ji} \approx 1/2K$ if $|j-i| < K$ and 0 otherwise, while $\sum_k W_{ki} W_{kj} \approx (|j-i| - 2K)/2K^2$ for points x_j with $|j - i| < 2K$. Hence H essentially computes the Hessian of Y. Formally,

$$H_{ij} = H^0_{ij} + \boldsymbol{O}_p(K^{-3/2}) \tag{35}$$

by (19), where

$$H^0_{ij} = \begin{cases} |j-i|/2K^2 & |j-i| < K\,, \\ (|j-i| - 2K)/2K^2 & K < |j-i| < 2K\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Now,

$$\begin{aligned}
\lambda n^{1/2} &= \lambda\|Y\| && \text{(by the normalization of } Y)\\
&= \|(I-M)Y\| && \text{(being eigenvalue)}\\
&= \|(I-W)Y - HY\| && \text{(definition of } H)\\
&\geq \|HY\| - \|(I-W)Y\| = \|HY\| - (\lambda n)^{1/2} && \text{(triangular inequality)}\,.
\end{aligned}$$

Hence, since $\lambda = \mathcal{O}_p\big((K/n)^6\big)$ (see discussion below (29)),

$$\|T\| \leq 2(\lambda n)^{1/2} = \mathcal{O}_p\big(K^{7/2}/n^3\big), \tag{36}$$

where $T = HY$. Define

$$\begin{aligned}
A_i^+ &= \sum_{j=1}^{K} \frac{j}{K^2} y_{i+j}\\
A_i^- &= \sum_{j=1}^{K} \frac{j}{K^2} y_{i-j},
\end{aligned} \tag{37}$$

and note that $\sum_j H_{ij}^0 y_j = -A_{i-K}^- + A_{i-K}^+ + A_{i+K}^- - A_{i+K}^+$. We combine (34) with (35) and then (36), noting that $\sum_j H_{ij}^0 = \sum_j H_{ij} = 0$:

$$\begin{aligned}
(A_{i-K}^+ - A_{i-K}^-) &= (A_{i+K}^+ - A_{i+K}^-) + \sum_j H_{ij}^0 y_j\\
&= (A_{i+K}^+ - A_{i+K}^-) + \sum_j H_{ij}^0 (y_j - y_i)\\
&= (A_{i+K}^+ - A_{i+K}^-) + \sum_j H_{ij} (y_j - y_i) + \boldsymbol{O}_p\big(K \times K^{-3/2} \times (K/n)^{1/4}\big)\\
&= (A_{i+K}^+ - A_{i+K}^-) + \sum_j H_{ij} y_j + \boldsymbol{O}_p(n^{-1/4} K^{-1/4})
\end{aligned}$$

Iterating this equation $\nu < (n-i)/2K$ times, we obtain

$$(A_{i-K}^+ - A_{i-K}^-) = (A_{i+(2\nu-1)K}^+ - A_{i+(2\nu-1)K}^-) + \sum_{m=0}^{\nu} T_{i+mK} + \boldsymbol{O}_p\big(n^{3/4}/K^{5/4}\big), \tag{38}$$

Rewriting (38) and denoting $k = 2\nu$ we obtain

$$(A^+_{i+j} - A^-_{i+j}) = (A^+_{k+j} - A^-_{k+j}) + \sum_{m=1}^{\nu+1} T_{i+j+mK} + \boldsymbol{O}_p(n^{3/4}/K^{5/4}). \tag{39}$$

Note that by Cauchy-Schwarz, the bound on ν, and (36):

$$|\sum_{j=1}^{\ell}\sum_{m=1}^{\nu+1} T_{i+j+mK}| \le (\nu+1)\sum_{j=1}^{n} |T_j| \le \frac{n}{K} n^{1/2}\|T\| = \mathcal{O}_p(K^{5/2}/n^{3/2}) \tag{40}$$

Now, careful examination of $\sum_{j=1}^{\ell}(A^+_{i+j} - A^-_{i+j})$ shows that this sum depends only on the values of the ys near the edges of the range:

$$\sum_{j=1}^{\ell}(A^+_{i+j} - A^-_{i+j}) = \sum V_k y_{i+\ell+k} - \sum V_k y_{i+k} \tag{41}$$

where $V_k \ge 0$, V_k is supported on $-K, \dots, K$, $V_k \approx k^2/2K^2$, and hence $\sum_k V_k \approx K/3$. By (34) we obtain that $\sum_j V_j y_{i+j} = y_i \sum_j V_j + \boldsymbol{O}_p(n^{-1/4}K^{3/4})$. Summing both sides of (39) over $j = 1, \dots, \ell$, dividing by K, and then using (40), (41), and the continuity of y as given in (34) yields

$$\begin{aligned} y_{i+\ell} - y_i &= y_{k+\ell} - y_k + \boldsymbol{O}_p(\ell n^{3/4}/K^{9/4} + K^{3/2}/n^{3/2} + (K/n)^{1/4}) \\ &= y_{k+\ell} - y_k + o_p(1). \end{aligned} \tag{42}$$

References

M. Belkin. *Problems of Learning on Manifolds.* PhD thesis, The University of Chicago, 2003.

M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University., 2000.

G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, Maryland, 1983.

L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low-dimensional manifolds. *J. Mach. Learn. Res.*, 4:119–155, 2003. ISSN 1533-7928.