Appendix A

A.1 Spectral Decomposition and Euclidean Distances in Diffusion Space

Here we describe some of the technical details for how a symmetric operator $\widetilde{A}$, the stochastic differential operator A and its adjoint (the Markov operator) A^* are related, and how these relations lead to different normalization schemes for the corresponding eigenvectors. (For ease of notation, we have omitted the subindex ε, since we here consider a fixed $\varepsilon > 0$.) We also show that the diffusion metric corresponds to a weighted Euclidean distance in the embedding space induced by the diffusion map.

Suppose that P is a probability measure with a compact support $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X}$ be a similarity function that is symmetric, continuous, and positivity-preserving, i.e. $k(x, y) > 0$ for all $x, y \in \mathcal{X}$. For simplicity, we assume in addition that k is positive semi-definite, i.e. for all bounded functions f on $\mathcal{X}$, $\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) f(x) f(y) dP(x) dP(y) \geq 0$. Consider two different normalization schemes of k:

$$\begin{aligned} \widetilde{a}(x,y) &= \frac{k(x,y)}{\sqrt{\rho(x)}\sqrt{p(y)}} && \text{(symmetric)} \\ a(x,y) &= \frac{k(x,y)}{\rho(x)} && \text{(stochastic)} \end{aligned}$$

where $\rho(x) = \int k(x, y) dP(y)$.

Define the *symmetric* integral operator $\widetilde{A}$ by

$$\widetilde{A} f(x) = \int_{\mathcal{X}} \widetilde{a}(x, y) f(y) dP(y).$$

Under the stated conditions, $k(x, y)$ is an L^2-kernel. It follows that $\widetilde{A}$ is a self-adjoint compact operator. The eigenvalues $\{\lambda_\ell\}_{\ell \geq 0}$ of $\widetilde{A}$ are real and the associated eigenfunctions $\{v_\ell\}_{\ell \geq 0}$ form an orthonormal basis of $L^2(\mathcal{X}; dP)$. According to Mercer's theorem, we have the spectral decomposition

$$\widetilde{a}(x, y) = \sum_{\ell \geq 0} \lambda_\ell v_\ell(x) v_\ell(y), \tag{1}$$

where the series on the right converges uniformly and absolutely to $\widetilde{a}(x, y)$.

Now consider the integral operator A and its adjoint (the Markov operator) A^*:

$$\begin{aligned} Af(x) &= \int_{\mathcal{X}} a(x,y)f(y)dP(y) \\ A^*f(x) &= \int_{\mathcal{X}} f(y)a(y,x)dP(y), \end{aligned}$$

where $\langle Af, g\rangle_{L^2(\mathcal{X};dP)} = \langle f, A^*g\rangle_{L^2(\mathcal{X};dP)}$. Let $s(x) = \rho(x)/\int \rho(y)dP(y)$. If $\widetilde{A}v_\ell = \lambda_\ell v_\ell$, then we have the corresponding eigenvalue equations

$$A\psi_\ell = \lambda_\ell \psi_\ell, \quad \text{where} \quad \psi_\ell(x) = v_\ell(x)/\sqrt{s(x)} \tag{2}$$

$$A^*\varphi_\ell = \lambda_\ell \varphi_\ell, \quad \text{where} \quad \varphi_\ell(x) = v_\ell(x)\sqrt{s(x)}. \tag{3}$$

Moreover, if $\{v_\ell\}_{\ell\geq 0}$ is an orthonormal basis of $L^2(\mathcal{X};dP)$, then the sets $\{\psi_\ell\}_{\ell\geq 0}$ and $\{\varphi_\ell\}_{\ell\geq 0}$ form orthonormal bases of the *weighted* L^2-spaces $L^2(\mathcal{X};sdP)$ and $L^2(\mathcal{X};dP/s)$, respectively. The operator A preserves constant functions, i.e. $A1 = 1$. One can also show that the matrix norm $\|\widetilde{A}\| = \sup_{f\in L^2(\mathcal{X};dP)} \frac{\|\widetilde{A}f\|}{\|f\|} = 1$. Thus, the eigenvalue $\lambda_0 = 1$ is the largest eigenvalue of the operators A and A^*. The corresponding eigenvector of A is $\psi_0 = 1$, and the corresponding eigenvector of A^* is $\varphi_0 = s$.

From Eq. 1, it follows that $a(x,y) = \sum_{\ell\geq 0} \lambda_\ell \psi_\ell(x)\varphi_\ell(y)$,where $\|\varphi_\ell\|_{L^2(\mathcal{X};dP/s)} = \|\psi_\ell\|_{L^2(\mathcal{X};sdP)} = 1$ for all $\ell \geq 0$, and $\langle \varphi_k, \psi_\ell\rangle_{L^2(\mathcal{X};dP)} = 0$ for $k \neq \ell$. More generally, if $a_m(x,y)$ is the kernel of the m^{th} iterate A^m, where m is a positive integer, then

$$a_m(x,y) = \sum_{\ell\geq 0} \lambda_\ell^m \psi_\ell(x)\varphi_\ell(y). \tag{4}$$

We define a one-parametric family of diffusion distances between points x and z according to

$$D_m^2(x,z) \equiv \|a_m(x,\cdot) - a_m(z,\cdot)\|^2_{L^2(\mathcal{X};dP/s)}, \tag{5}$$

where the parameter m determines the scale of the analysis. The diffusion metric measures the rate of connectivity between points on a data set. It will be small if there are many paths of lengths less than or equal to $2m$ between the two points, and it will be large if the number of connections is small. One can see this clearly by expanding the expression in Eq. 5 so that

$$D_m^2(x,z) = \frac{a_{2m}(x,x)}{s(x)} + \frac{a_{2m}(z,z)}{s(z)} - \left(\frac{a_{2m}(x,z)}{s(z)} + \frac{a_{2m}(z,x)}{s(x)} \right). \tag{6}$$

The quantity $D_m^2(x,z)$ is small when the transition probability densities $a_{2m}(x,z)$ and $a_{2m}(z,x)$ are large.

Finally, we look for an embedding where Euclidean distances reflect the above diffusion metric. The biorthogonal decomposition in Eq. 4 can be viewed as an orthogonal expansion of the functions $a_m(x,\cdot)$ with respect to the orthonormal basis $\{\varphi_\ell\}_{\ell \geq 0}$ of $L^2(\mathcal{X}; dP/s)$; the expansion coefficients are given by $\{\lambda_\ell^m \psi_\ell(x)\}_{\ell \geq 0}$. Hence,

$$D_m^2(x,z) = \sum_{\ell \geq 0} (\lambda_\ell^m \psi_\ell(x) - \lambda_\ell^m \psi_\ell(z))^2 = \|\Psi_m(x) - \Psi_m(z)\|^2,$$

where $\Psi_m : x \mapsto (\lambda_1^m \psi_1(x), \lambda_2^m \psi_2(x), \ldots)$ is the diffusion map of the data at time step m.

A.2 Proofs

Proof of Theorem 1. Recall that $\mathcal{F}$ is the set of uniformly bounded, three times differentiable functions with uniformly bounded derivatives whose gradients vanish at the boundary. From Theorem 2 below, we have that

$$\|A_t(\varepsilon_n, \widehat{P}_n) - \mathbf{A}_t\| = (O_P(\gamma_n) + O(\varepsilon_n)) \cdot \rho(t)$$

where $\gamma_n = \sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{(d+4)/2}}}$. Hence,

$$\begin{aligned} \|A_t(\varepsilon_n, q, \widehat{P}_n) - \mathbf{A}_t\| &\leq \|A_t(\varepsilon_n, q, \widehat{P}_n) - A_t(\varepsilon_n, \widehat{P}_n)\| + \|A_t(\varepsilon_n, \widehat{P}_n) - \mathbf{A}_t\| \\ &= \|\sum_{\ell=q+1}^{\infty} \widehat{\lambda}_{\varepsilon_n,\ell}^{t/\varepsilon_n} \widehat{\Pi}_{\varepsilon_n,\ell}\| + (O_P(\gamma_n) + O(\varepsilon_n)) \cdot \rho(t) \\ &\leq \sum_{\ell=q+1}^{\infty} \widehat{\lambda}_{\varepsilon_n,\ell}^{t/\varepsilon_n} + (O_P(\gamma_n) + O(\varepsilon_n)) \cdot \rho(t). \end{aligned}$$

Now we bound the first sum. Note that,

$$\sup_\ell |\widehat{\nu}^2_{\varepsilon_n,\ell} - \nu^2_{\varepsilon_n,\ell}| = \sup_\ell \frac{|\widehat{\lambda}_{\varepsilon_n,\ell} - \lambda_{\varepsilon_n,\ell}|}{\varepsilon_n} \leq \frac{\|\widehat{A}_{\varepsilon_n} - A_{\varepsilon_n}\|}{\varepsilon_n} = O_P(\gamma_n).$$

By a Taylor series expansion, $G_{\varepsilon_n} f = \mathbf{G} f + O(\varepsilon_n)$ uniformly for $f \in \mathcal{F}$. (This is the same calculation used to compute the bias in kernel regression. See also, Giné and Koltchinskii (2006) and Singer (2006)). So,

$$\sup_\ell |\nu^2_{\varepsilon_n,\ell} - \nu^2_\ell| \leq \|G_{\varepsilon_n} - \mathbf{G}\| = O(\varepsilon_n).$$

Therefore,

$$\begin{aligned}
\sum_{\ell=q+1}^\infty \widehat{\lambda}^{t/\varepsilon_n}_{\varepsilon_n,\ell} &= \sum_{\ell=q+1}^\infty (1 - \varepsilon_n \widehat{\nu}^2_{\varepsilon_n,\ell})^{t/\varepsilon_n} \\
&= \sum_{q+1}^\infty \exp\left\{ \frac{t}{\varepsilon_n} \log(1 - \varepsilon_n \widehat{\nu}^2_{\varepsilon_n,\ell}) \right\} \\
&= \sum_{\ell=q+1}^\infty \exp\left\{ \frac{t}{\varepsilon_n} \log(1 - \varepsilon_n [O_P(\gamma_n) + O(\varepsilon_n) + \nu^2_\ell]) \right\} \\
&= (1 + O_P(\gamma_n) + O(\varepsilon_n)) \sum_{\ell=q+1}^\infty e^{-\nu^2_\ell t}.
\end{aligned}$$

The result follows.

Proof of Theorem 2. Recall that $A_t(\varepsilon_n, \widehat{P}_n) = e^{t(\widehat{A}_{\varepsilon_n} - I)/\varepsilon_n}$. From Lemma 1 below, $\|A_\varepsilon - \widehat{A}_\varepsilon\| = \alpha(\varepsilon)$ where $\alpha(\varepsilon) = O_P\left(\sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{d/2}}}\right)$. Hence,

$$\frac{\widehat{A}_\varepsilon - I}{\varepsilon} = \frac{\widehat{A}_\varepsilon - A_\varepsilon}{\varepsilon} + \frac{A_\varepsilon - I}{\varepsilon} = \mathbf{G} + O(\varepsilon) + \text{Rem}$$

where $||\text{Rem}|| = \alpha(\epsilon)/\epsilon$ and so

$$A_t(\varepsilon, \widehat{P}_n) = \mathbf{A}_t e^{t(\widehat{A}_\varepsilon - A_\varepsilon + O(\varepsilon^2))/\varepsilon} = \mathbf{A}_t \Big[I + t(\widehat{A}_\varepsilon - A_\varepsilon + O(\varepsilon^2))/\varepsilon + o(t(\widehat{A}_\varepsilon - A_\varepsilon + O(\varepsilon^2)))/\varepsilon) \Big]$$

Therefore,

$$\begin{aligned}\|\mathbf{A}_t - A_t(\varepsilon, \widehat{P}_n)\| &= \|\mathbf{A}_t\| \left(O_P(\alpha/\epsilon) + O(\varepsilon)\right) \\ &\leq \left(O_P(\gamma) + O(\varepsilon)\right) \sum_{\ell=1}^{\infty} e^{-\nu_\ell^2 t} \;\square\end{aligned}$$

Lemma 1 *Let* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^{d/2}/\log(1/\varepsilon_n) \to \infty$. *Then* $\|A_\varepsilon - \widehat{A}_\varepsilon\| = \alpha_n$ *where* $\alpha_n = O_P\left(\sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{d/2}}}\right)$.

Proof. Uniformly, for all $f \in \mathcal{F}$, and all x in the support of P,

$$|A_\varepsilon f(x) - \widehat{A}_\varepsilon f(x)| \leq |A_\varepsilon f(x) - \widetilde{A}_\varepsilon f(x)| + |\widetilde{A}_\varepsilon f(x) - \widehat{A}_\varepsilon f(x)|$$

where $\widetilde{A}_\varepsilon f(x) = \int \widehat{a}_\varepsilon(x,y) f(y) dP(y)$. From Giné and Guillou (2002),

$$\sup_x \frac{|\widehat{p}_\varepsilon(x) - p_\varepsilon(x)|}{|\widehat{p}_\varepsilon(x) p_\varepsilon(x)|} = O_P(\alpha_n).$$

Hence,

$$\begin{aligned}|A_\varepsilon f(x) - \widetilde{A}_\varepsilon f(x)| &\leq \frac{|\widehat{p}_\varepsilon(x) - p_\varepsilon(x)|}{|\widehat{p}_\varepsilon(x) p_\varepsilon(x)|} \int |f(y)| k_\varepsilon(x,y) dP(y) \\ &= O_P(\alpha_n) \int |f(y)| k_\varepsilon(x,y) dP(y) \\ &= O_P(\alpha_n).\end{aligned}$$

Next, we bound $\widetilde{A}_\varepsilon f(x) - \widehat{A}_\varepsilon f(x)$. We have

$$\begin{aligned}\widetilde{A}_\varepsilon f(x) - \widehat{A}_\varepsilon f(x) &= \int f(y) \widehat{a}_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y)) \\ &= \frac{1}{p(x) + o_P(1)} \int f(y) k_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y)).\end{aligned}$$

Now, expand $f(y) = f(x) + r_n(y)$ where $r_n(y) = (y-x)^T \nabla f(u_y)$ and u_y is between y and x. So,

$$\int f(y) k_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y)) = f(x) \int k_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y)) + \int r_n(y) k_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y)).$$

By an application of Talagrand's inequality to each term, as in Theorem 5.1 of Giné and Koltchinskii (2006), we have

$$\int f(y)k_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y)) = O_P(\alpha_n).$$

Thus, $\sup_{f\in\mathcal{F}} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_\infty = O_P(\alpha_n)$. This also holds uniformly over $\{f \in \mathcal{F} : \|f\| = 1\}$. Moreover, $\|\widehat{A}_\varepsilon f - A_\varepsilon f\|_2 \le C\|\widehat{A}_\varepsilon f - A_\varepsilon f\|_\infty$ for some C since P has compact support. Hence,

$$\sup_{f\in\mathcal{F}} \frac{\|\widehat{A}_\varepsilon f - A_\varepsilon f\|_2}{\|f\|} = \sup_{f\in\mathcal{F},\|f\|=1} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_2 = O_P(\alpha_n)\Box$$

Proof of Theorem 3. Let $A_n = \{|\psi_1(X)| \le \delta_n\}$. Then

$$A_n^c \bigcap \Big\{\widehat{H}(X) \neq H(X)\Big\} \quad \text{implies that} \quad \Big\{|\widehat{\psi}_{\varepsilon,1}(X) - \psi_1(X)| > \delta_n\Big\}.$$

Also, $\sup_x |\psi_1(x) - \psi_{\varepsilon,1}(x)| \le c\varepsilon_n$ for some $c > 0$. Hence,

$$\begin{aligned}
\mathbb{P}\left(\widehat{H}(X) \neq H(X)\right) &= \mathbb{P}\left(\widehat{H}(X) \neq H(X), A_n\right) + \mathbb{P}\left(\widehat{H}(X) \neq H(X), A_n^c\right) \\
&\le \mathbb{P}(A_n) + \mathbb{P}\left(\widehat{H}(X) \neq H(X), A_n^c\right) \\
&\le C\delta_n^\alpha + \mathbb{P}\left(|\psi_1(X) - \widehat{\psi}_{\varepsilon,1}(X)| > \delta_n\right) \\
&\le C\delta_n^\alpha + \mathbb{P}\left((|\psi_1(X) - \psi_{\varepsilon,1}(X)| + |\psi_{\varepsilon,1}(X) - \widehat{\psi}_{\varepsilon,1}(X)|) > \delta_n\right) \\
&\le C\delta_n^\alpha + \mathbb{P}\left(|\widehat{\psi}_{\varepsilon,1}(X) - \psi_{\varepsilon,1}(X)| > \delta_n - c\varepsilon_n\right) \\
&\le C\delta_n^\alpha + \frac{\mathbb{E}\|\widehat{\psi}_{\varepsilon,1}(X) - \psi_{\varepsilon,1}(X)\|}{\delta_n - c\varepsilon_n} \\
&\le C\delta_n^\alpha + O_P\left(\sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{(d+4)/2}}}\right)\frac{1}{\delta_n - c\varepsilon_n}
\end{aligned}$$

Set $\delta = 2c\varepsilon_n$ and $\varepsilon_n = n^{-2/(4\alpha+d+8)}$ and so $\mathbb{P}\left(\widehat{H}(X) \neq H(X)\right) \le n^{-\frac{2\alpha}{4\alpha+8+d}}$ $\Box$

References

Giné, E. and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann Inst. H. Poincar 38*, 907–921.

Giné, E. and V. Koltchinskii (2006). Empirical graph Laplacian approximation of Laplace-Beltrami operators: Large sample results. In *High Dimensional Probability: Proceedings of the Fourth International Conference*, IMS Lecture Notes, pp. 1–22.

Singer, A. (2006). From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis 21*, 128–134.