Operator convergence of diffusion maps and the bistochastic normalisation

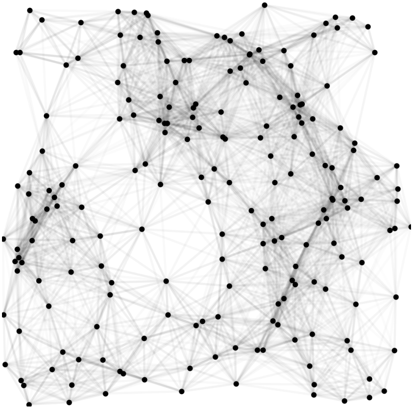
Caroline Wormell

Joint work with Sebastian Reich

28th September, 2021

Introduction

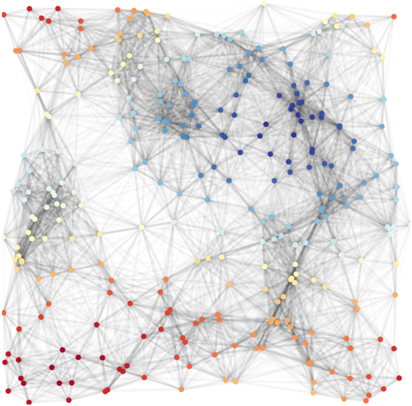
Diffusion maps: on a random point sample, create Markov process approximating a (continuous-time) diffusion.



Introduction

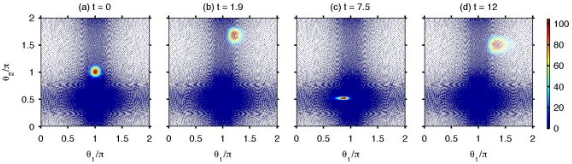
Things you can do with this diffusion:

- Eigendata of generator, which is a Laplacian:
  - Dimensionality reduction via intrinsic coordinates
  - Data clustering



Introduction

- Non-parametric forecasting (data obtained from a time series)
- Approximation of more complex operators (e.g. Berry '18)



Giannakis '19

Diffusion maps

- Input: M points $x^i \sim \rho$ abs. cts on hypertorus domain $\mathbb{D} = (\mathbb{R}/L\mathbb{Z})^d$ (e.g.).
- Construct $M \times M$ kernel matrix K

$$K_{ij} = \frac{1}{M} g_\epsilon(x^i - x^j)$$

  where g_ϵ is Gaussian kernel of *variance* ϵ.
- With appropriate weight vectors u and $v := 1/(Ku)$, construct Markov matrix

$$P = \operatorname{diag} v \; K \operatorname{diag} u$$

- As $M \to \infty$ and $\epsilon \to 0$ appropriately, P is approximation of $e^{\epsilon \mathcal{L}}$ where

$$\mathcal{L} = \frac{1}{2}\Delta + \nabla \log p \cdot \nabla \phi$$

Diffusion maps: convergence rates

Expect in general:

$$\left\| f(P^{t/\epsilon}) - f(e^{t\mathcal{L}}) \right\| = \mathcal{O}\Big(\underbrace{M^{-\frac{1}{2}} \epsilon^{-\frac{d}{4}-\frac{1}{2}} \log(\cdots)^{\cdots}}_{\text{"variance error"}} + \underbrace{\epsilon^{\theta}}_{\text{"bias error"}} \Big)$$

Know rigorously this works for

- f = pointwise evaluation of functions (von Luxburg *et al.* '08)
- f = eigendata of graph Laplacian (Calder and Trillos '20)

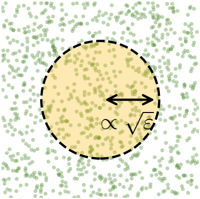


Figure: Effective support of g_ϵ contains $\mathcal{O}(M\epsilon^{d/2})$ data points.

Questions

Some mysteries:

1. How does matrix operator K acting on a random point cloud converge *as an operator* to a continuous kernel? (At the rate seen in practice?)
2. What is the (best possible) exponent in the bias error? How can we best choose weight vectors?

Kernel operator interpolation

The following operator on $C^0(\mathbb{D})$ matches kernel matrix K at sample points:

$$\mathcal{K}_\epsilon^M \phi = \sum_{i=1}^{M} \tfrac{1}{M} g_\epsilon(\cdot - x^i)\phi(x^i) = g_\epsilon * [\rho^M \phi]$$

As $M \to \infty$, expect $\mathcal{K}_\epsilon^M$ to converge to continuous kernel operator

$$\mathcal{K}_\epsilon \phi := \int_{\mathbb{D}} g_\epsilon(\cdot - x)\phi(x)\rho \, \mathrm{d}x = g_\epsilon * [\rho\phi],$$

ideally in some Banach space $\mathcal{B}_\epsilon \Subset C^0$.

Kernel operator interpolation

Because $g_\epsilon = g_{\epsilon/2} * g_{\epsilon/2}$ we can try for

$$\mathcal{K}_\epsilon^M - \mathcal{K}_\epsilon = \underbrace{g_{\epsilon/2} \quad *}_{\text{unif. bd. } C^0 \to \mathcal{B}_\epsilon} \quad \underbrace{(\mathcal{K}_{\epsilon/2}^M - \mathcal{K}_{\epsilon/2})}_{\text{small } \mathcal{B}_\epsilon \to C^0}$$

$$= \text{small } \mathcal{B}_\epsilon \to \mathcal{B}_\epsilon$$

Choice of $\mathcal{B}_\epsilon$

As $\epsilon \to 0$, convolution by $g_{\epsilon/2} * \phi \to \phi$, so we expect $\mathcal{B}_0 = C^0$.
Let the complex domain

$$\mathbb{D}_\epsilon = \mathbb{D} + B_{\mathbb{C}}(\sqrt{\epsilon/2}).$$

A scale of function spaces with very good regularity is

$$\mathcal{B}_\epsilon(\mathbb{D}) := \{\text{ct's analytic functions on } \mathbb{D}_\epsilon\}$$

endowed with sup norm.
This is good because

$$\|g_{\epsilon/2} * \phi\|_{\mathcal{B}_\epsilon} = \|g_{\epsilon/2}\|_{L^1(\partial\mathbb{D}_\epsilon)}\|\phi\|_{C^0} = e^{1/2}\|\phi\|_{C^0}$$

which gives us a uniformly bounded norm $C^0 \to \mathcal{B}_\epsilon$.

Kernel operator interpolation

Want to show that, up to log terms,

$$\delta := \|\mathcal{K}_{\epsilon/2}\phi - \mathcal{K}_{\epsilon/2}^M\phi\|_{\mathcal{B}_\epsilon \to C^0} \approx \text{pointwise error} = \mathcal{O}(M^{-1/2}\epsilon^{-d/4})$$

Recall we know that* for fixed ϕ and x,

$$\text{pointwise error} = \left|(\mathcal{K}_\epsilon\phi - \mathcal{K}_\epsilon^M\phi)(x)\right| \leq \frac{C\epsilon^{-d/4}}{M^{1/2}}|\mathcal{N}(0,1)|.$$

How to extend efficiently to uniform bounds for all $\phi \in \mathcal{B}_\epsilon$, $x \in \mathbb{D}$?

* except for large deviations

Kernel operator interpolation

Want to show that, up to log terms,

$$\delta := \|\mathcal{K}_{\epsilon/2}\phi - \mathcal{K}^M_{\epsilon/2}\phi\|_{\mathcal{B}_\epsilon \to C^0} \approx \text{pointwise error} = \mathcal{O}(M^{-1/2}\epsilon^{-d/4})$$

Recall we know that* for fixed ϕ and x,

$$\text{pointwise error} = \left|(\mathcal{K}_\epsilon\phi - \mathcal{K}^M_\epsilon\phi)(x)\right| \leq \frac{C\epsilon^{-d/4}}{M^{1/2}}|\mathcal{N}(0,1)|.$$

How to extend efficiently to uniform bounds for all $\phi \in \mathcal{B}_\epsilon$? Say,

$$\sup_{\|\phi\|_{\mathcal{B}_\epsilon}=1} \left|(\mathcal{K}_\epsilon\phi - \mathcal{K}^M_\epsilon\phi)(x)\right| \sim \frac{C\epsilon^{-d/4}}{M^{1/2}} \times \text{log terms}$$

* except for large deviations

Naive idea (Glivenko-Cantelli)

We have (bad) a priori estimate

$$\|\mathcal{K}_\epsilon - \mathcal{K}_\epsilon^M\|_{C^0} \leq \|K_\epsilon\|_{C^0} + \|\mathcal{K}_\epsilon^M\|_{C^0} \leq 2 \sup g_\epsilon = C\epsilon^{-d/2}.$$

The unit ball in $\mathcal{B}_\epsilon$ is compact in C^0, so we can cover the unit ball with a finite number of C^0 balls, i.e. there is a collection of $\#(\mathcal{B}_\epsilon, \xi)$ functions ϕ_n so that every ϕ with $\|\phi\|_{\mathcal{B}_\epsilon} \leq 1$ has $\|\phi_n - \phi\| \leq \xi$ for some n.

Naive idea (Glivenko-Cantelli)

Maximising over the ϕ_n,

$$\sup_n \left|(\mathcal{K}_\epsilon \phi_n - \mathcal{K}_\epsilon^M \phi_n)(x)\right| \leq \frac{C\epsilon^{-d/4}}{M^{1/2}} \mathcal{N}_{\#(\mathcal{B}_\epsilon,\xi)},$$

where the maximum absolute value of T (non-ind.) standard normal distributions is $\mathcal{N}_T = \mathcal{O}(\sqrt{\log T})$. Thus,

$$\sup_{\|\phi\|_{\mathcal{B}_\epsilon}=1} \left|(\mathcal{K}_\epsilon \phi - \mathcal{K}_\epsilon^M \phi)(x)\right| \leq \frac{C\epsilon^{-d/4}}{M^{1/2}} \mathcal{N}_{\#(\mathcal{B}_\epsilon,\xi)} + C\epsilon^{-d/2}\xi.$$

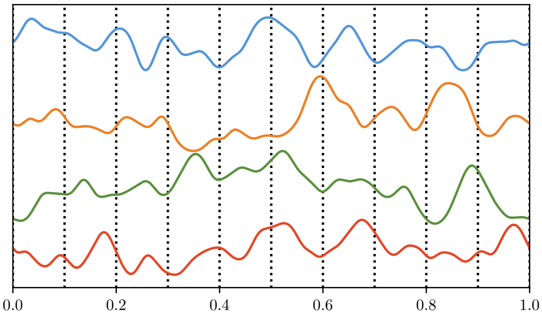
Want $\sqrt{\log \#(\mathcal{B}_\epsilon, \xi)}$ to grow sub-polynomially with $\epsilon, \xi \to 0$.

Naive idea (Glivenko-Cantelli)

In practice, if $X \subset \mathbb{R}^d$ is a hypercube of length L then

$$\log \#(C^0(X), \mathcal{B}_\epsilon(X), \xi) = \mathcal{O}\left((L\epsilon^{-1/2} \log \xi^{-1})^d\right)$$

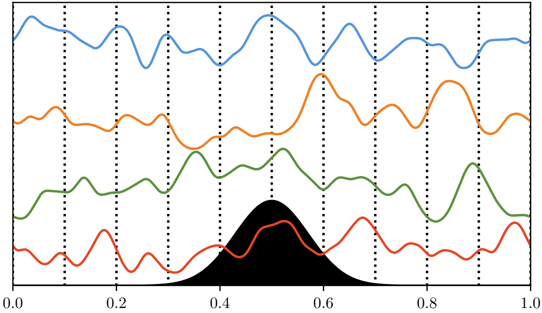
This gives us problems when $\epsilon^{1/2} \ll \operatorname{diam} \mathbb{D}$.



Local Glivenko-Cantelli

However, we only see ϕ on a small part of the domain!

$$(g_{\epsilon/2} * \psi)(x) = g_{\epsilon/2} * (\mathbb{1}_{B(x, l\sqrt{\epsilon})}\psi) + \mathcal{O}(e^{-Cl^2})\|\psi\|_{L^1}.$$



Local Glivenko-Cantelli

We really just want a set of radius $l\sqrt{\epsilon}$, where l grows logarithmically.

$\mathcal{B}_\epsilon^{x,l} := \{\text{bd. analytic functions on } B_{\mathbb{R}}(x, l\sqrt{\epsilon}) + B_{\mathbb{C}}(0, \sqrt{\epsilon}/2)\} \supset \mathcal{B}_\epsilon.$

with

$$\log \#(\mathcal{B}_\epsilon^{x,l}, \xi) = \mathcal{O}\left((l \log \xi^{-1})^d\right)$$

and we are in business:

$$\begin{aligned}\sup_{\|\phi\|_{\mathcal{B}_\epsilon}=1} \left|(\mathcal{K}_\epsilon\phi - \mathcal{K}_\epsilon^M\phi)(x)\right| &\leq \frac{C\epsilon^{-d/4}}{M^{1/2}}\mathcal{N}_{\#(\mathcal{B}_\epsilon^{x,l},\xi)} + C\epsilon^{-d/2}\xi + Ce^{-Cl^2} \\ &= \mathcal{O}\left(\epsilon^{-d/4}M^{-1/2}(\log M\epsilon^{-1})^{d-1/2}\right)\end{aligned}$$

Local Glivenko-Cantelli

We can use an easier compactness argument to extend to a supremum over all x, giving

$$\begin{aligned}\delta := \left\|(\mathcal{K}_{\epsilon/2}\phi - \mathcal{K}^{M}_{\epsilon/2})\phi\right\|_{\mathcal{B}_\epsilon \to C^0} &= \mathcal{O}\left(\epsilon^{-d/4} M^{-1/2} (\log M\epsilon^{-1})^{d-1/2}\right) \\ &= \text{appropriately small}\end{aligned}$$

All sample-based errors are then controlled by δ!

Local Glivenko-Cantelli

In particular, recall the Markov matrix

$$P = \operatorname{diag} v \;\; K \operatorname{diag} u$$

Our weight vectors u, $v = 1/Kv$ are interpolated by functions U_ϵ^M, $V_\epsilon^M = 1/\mathcal{K}_\epsilon^M U_\epsilon^M$:

$$\mathcal{P}_\epsilon^M = (\mathcal{K}_\epsilon^M U_\epsilon^M)^{-1} \, \mathcal{K}_\epsilon^M \, U_\epsilon^M$$

For any reasonable way to choose u, our operator will converge to a continuum limit:

$$\|\mathcal{P}_\epsilon^M - \mathcal{P}_\epsilon\|_{\mathcal{B}_\epsilon} \leq C\delta$$

for $\delta < \delta_0$.

Comments

Result: convergence of spectral data, complex operator problems, etc. at near-pointwise rates.

- Requires very smooth kernel with exponentially decaying tails.
- Will generalise nicely to curved manifolds!
- Argument not based on Markov normalisation.
- Specialisation to Markov kernels would improve by $\mathcal{O}(\epsilon^{1/2})$ factor (Singer '06, Calder and Trillos '20).

Bias error analysis

Our weight vectors u, v are interpolated by functions $U_\epsilon^M, V_\epsilon^M$ which converge to U_ϵ, V_ϵ as $M \to \infty$.
Have infinite limit

$$\mathcal{P}_\epsilon \phi = V_\epsilon \mathcal{K}_\epsilon [U_\epsilon \phi].$$

Want to show that as $\epsilon \to 0$

$$\mathcal{P}_\epsilon \to e^{\epsilon \mathcal{L}}.$$

Bias error analysis

Our weight vectors u, v are interpolated by functions $U_\epsilon^M, V_\epsilon^M$ which converge to U_ϵ, V_ϵ as $M \to \infty$.
Have infinite limit

$$\mathcal{P}_\epsilon \phi = V_\epsilon \mathcal{K}_\epsilon [U_\epsilon \phi].$$

Want to show that as $\epsilon \to 0$

$$\mathcal{P}_\epsilon^{t/\epsilon} \to e^{t\mathcal{L}}.$$

Bias error analysis

Know $\mathcal{L}$ is generator of SDE for invariant density p

$$\mathrm{d}X = -\frac{1}{2}\nabla p \, \mathrm{d}t + \mathrm{d}W_t$$

We can study $\mathcal{P}_\epsilon^{t/\epsilon}$ as the evolution operator of a (time-varying) SDE.

Bias error: SDE formulation

Let

$$e^{s_t} = g_t * (\rho U_\epsilon) = e^{t\Delta/2}(\rho U_\epsilon).$$

Then $\rho U_\epsilon = e^{s_0}$ and $V_\epsilon = e^{-s_\epsilon}$.

$$\mathcal{P}_\epsilon \phi := V_\epsilon g_\epsilon \star (\rho U_\epsilon \phi) = e^{-s_\epsilon} e^{\epsilon\Delta/2} e^{s_0} \phi$$

is time-ϵ operator of forward equation of SDE

$$\mathrm{d}X_t = -\nabla s_t \, \mathrm{d}t + \mathrm{d}W_t$$

So $\mathcal{P}_\epsilon^{t/\epsilon}$ is the time-t operator of

$$\mathrm{d}X_t = \underbrace{-\nabla s_{\epsilon\{t/\epsilon\}}}_{\text{fast, periodic}} \, \mathrm{d}t + \mathrm{d}W_t$$

Bias error: SDE formulation

Time-average with $\mathcal{O}(t\epsilon^2)$ error:

$$\mathrm{d}X_t \approx -\nabla \bar{s}\,\mathrm{d}t + \mathrm{d}W_t$$

$$\begin{aligned}\bar{s} &= \tfrac{1}{\epsilon}\int_0^{\epsilon} s_t\,\mathrm{d}t \\ &= \tfrac{1}{2}(s_0 + s_\epsilon) + \mathcal{O}(\epsilon^2) \\ &= \tfrac{1}{2}\log(\rho U_\epsilon / V_\epsilon) + \mathcal{O}(\epsilon^2)\end{aligned}$$

Bias error: SDE formulation

Time-average with $\mathcal{O}(t\epsilon^2)$ error:

$$\mathrm{d}X_t \approx -\nabla \bar{s}\,\mathrm{d}t + \mathrm{d}W_t$$

$$\begin{aligned}\bar{s} &= \frac{1}{\epsilon}\int_0^{\epsilon} s_t\,\mathrm{d}t \\ &= \frac{1}{2}(s_0 + s_\epsilon) + \mathcal{O}(\epsilon^2) \\ &= \underbrace{\frac{1}{2}\log(\rho U_\epsilon / V_\epsilon)}_{\text{want } = \frac{1}{2}\log p} + \mathcal{O}(\epsilon^2)\end{aligned}$$

Bias error: SDE formulation

Time-average with $\mathcal{O}(t\epsilon^2)$ error:

$$\mathrm{d}X_t \approx -\nabla \bar{s}\,\mathrm{d}t + \mathrm{d}W_t$$

$$\begin{aligned}\bar{s} &= \frac{1}{\epsilon}\int_0^{\epsilon} s_t\,\mathrm{d}t \\ &= \frac{1}{2}(s_0 + s_\epsilon) + \mathcal{O}(\epsilon^2) \\ &= \underbrace{\frac{1}{2}\log(\rho U_\epsilon / V_\epsilon)}_{\text{want } = \frac{1}{2}\log p} + \mathcal{O}(\epsilon^2)\end{aligned}$$

- Typically we fit $e^{s_0}/\rho = U_\epsilon \approx p^{1/2}/\rho$. Since $s_\epsilon = s_0 + \mathcal{O}(\epsilon)$, get $\mathcal{O}(\epsilon)$ error (for $\rho \in C^{3/2+\alpha}$).

Bias error: SDE formulation

Time-average with $\mathcal{O}(t\epsilon^2)$ error:

$$\mathrm{d}X_t \approx -\nabla \bar{s}\,\mathrm{d}t + \mathrm{d}W_t$$

$$\begin{aligned}\bar{s} &= \frac{1}{\epsilon}\int_0^\epsilon s_t\,\mathrm{d}t \\ &= \frac{1}{2}(s_0 + s_\epsilon) + \mathcal{O}(\epsilon^2) \\ &= \underbrace{\frac{1}{2}\log(\rho U_\epsilon / V_\epsilon)}_{\text{want } = \frac{1}{2}\log p} + \mathcal{O}(\epsilon^2)\end{aligned}$$

- Typically we fit $e^{s_0}/\rho = U_\epsilon \approx p^{1/2}/\rho$. Since $s_\epsilon = s_0 + \mathcal{O}(\epsilon)$, get $\mathcal{O}(\epsilon)$ error (for $\rho \in C^{3/2+\alpha}$).
- Optimally accurate approximation is $\mathcal{O}(\epsilon^2)$, obtained via fitting weight ratio: $U_\epsilon / V_\epsilon = p/\rho$.

Sinkhorn problem

Since by Markov constraint $V = 1/(\mathcal{K}U)$, this means solving symmetric Sinkhorn problem for U:

$$U \times (\mathcal{K}U) = p/\rho.$$

- Only need $\rho, p \in C^{2+\alpha}$ for $\mathcal{O}(\epsilon^2)$ eigendata convergence.
- Fast iterative algorithm to compute U.

Sinkhorn problem

Since by Markov constraint $V = 1/(\mathcal{K}U)$, this means solving symmetric Sinkhorn problem for U:

$$U \times (\mathcal{K}U) = p/\rho.$$

- Only need $\rho, p \in C^{2+\alpha}$ for $\mathcal{O}(\epsilon^2)$ eigendata convergence.
- Fast iterative algorithm to compute U.

In paper: $p = \rho$, i.e. $\mathcal{L}$ generates Langevin diffusion on ρ.

- P symmetric ($U = V$)
- P bistochastic (i.e. gives reversible Markov chain)

.

Comments

- In practice variance error $\mathcal{O}(M^{-1/2}\epsilon^{-d/4-1/2})$ will dominate bias error $\mathcal{O}(\epsilon^2)$!
- Expect convergence speed-up to work for symmetric kernels with correct 4th moments
- Only expect $\mathcal{O}(\epsilon)$ convergence on curved domains

Conclusions

In a narrow setting we prove operator convergence that:

- Implies spectral convergence and many other things
- Retains near-pointwise convergence rates for variance error
- Establishes optimal weights/convergence rates for bias error

Some extensions possible to more general settings!

Wormell, Caroline L., and Sebastian Reich. "Spectral convergence of diffusion maps: improved error bounds and an alternative normalisation." *SIAM Journal of Numerical Analysis* 59(3) (2021) 1687–1734