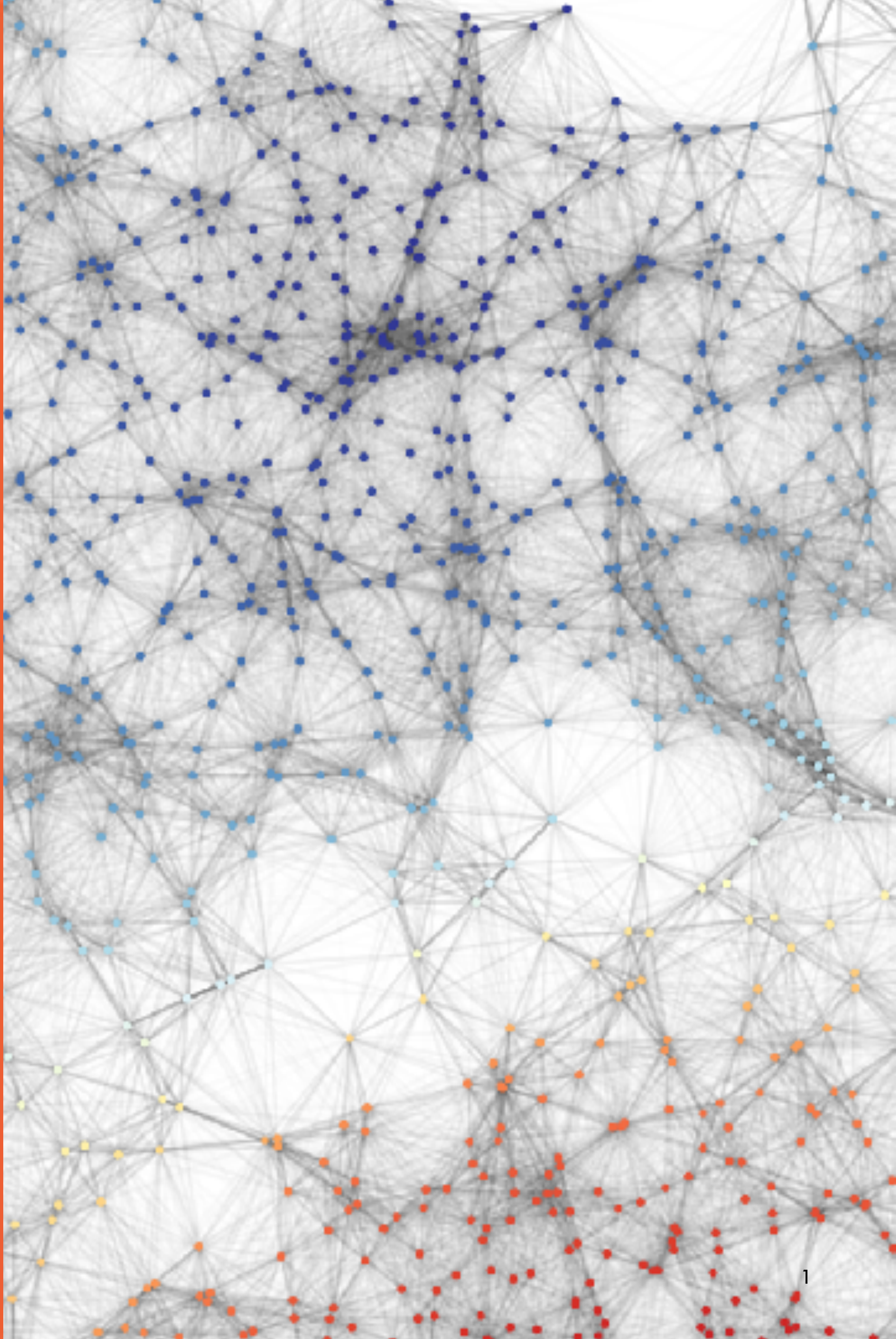
Spectral convergence of diffusion maps

Caroline Wormell
The University of Sydney

Joint work with Sebastian Reich,
Universität Potsdam



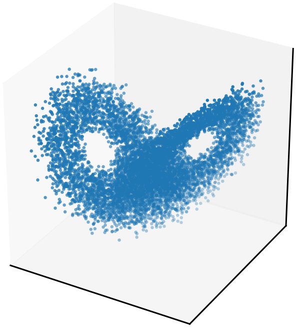
Introduction

- Take a data sample of M points $x^i \sim \rho$ from an unknown (sub-)manifold $\mathbb{D}$.
- **Aim:** do computations on (or visualise) the manifold using this data.
- **Natural object:** Laplace-Beltrami operator (weighted by p)

$$\mathscr{L}\phi := \frac{1}{2p}\nabla \cdot (p\nabla\phi) = \frac{1}{2}\Delta\phi + \frac{1}{2}\nabla \log p \cdot \nabla \log \phi$$

  - This is the generator of the gradient diffusion

$$\mathrm{d}X^t = \frac{1}{2}\nabla \log p(X^t)\,\mathrm{d}t + \mathrm{d}W^t$$



Diffusion maps algorithm

We approximate the semigroup $e^{\varepsilon\mathscr{L}}$ (transition kernel of a biased random walk), on the data $\{x^i\}_{i=1,\ldots,M}$:

- Start with a kernel matrix K:

$$K := \{g_\varepsilon(x^i - x^j)\}_{i,j=1,\ldots,M}$$

Gaussian kernel of variance ε

- *Bias* towards certain points, encoded by weight vector u
- Normalise to a Markov matrix by $v := 1/(Ku)$:

$$P := \mathrm{diag}(v)\, K\, \mathrm{diag}(u).$$

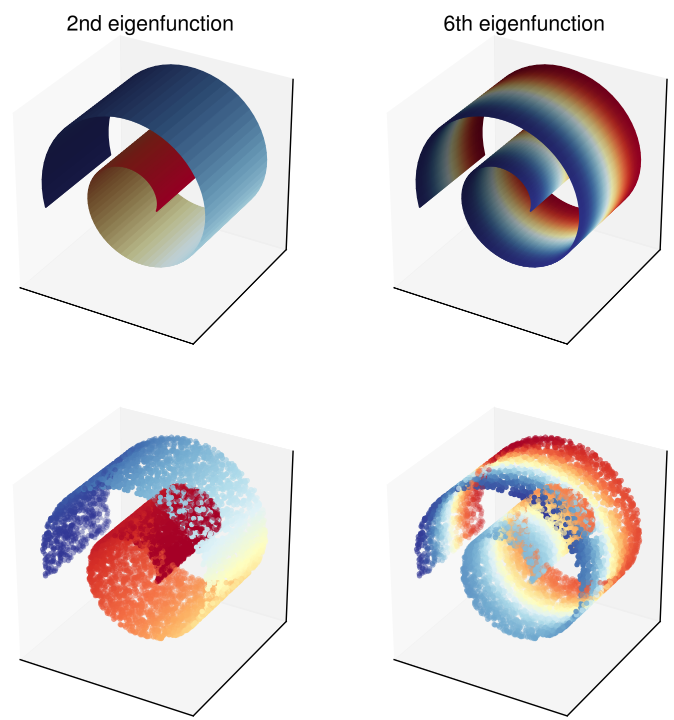
  - The invariant measure of the process is $u/v \approx \rho u^2$ ($\approx p$).

Standard weights

- Standard weights are powers of kernel density estimates of ρ:
$$u = (K\mathbf{1})^{-\alpha}.$$
- These converge to the following family of diffusions:
$$\mathscr{L}\phi = \frac{1}{2}\Delta\phi + (1-\alpha)\nabla\log\rho\cdot\nabla\phi$$
  - $\alpha = 0$ is standard graph Laplacian normalisation
  - $\alpha = 1/2$ is Langevin diffusion on ρ
  - $\alpha = 1$ is standard diffusion independent of ρ
- Other weights also possible
  - Will discuss Sinkhorn weights later…

What do we get from L-B operators?

- Eigenfunctions of Markov operators define intrinsic coordinates on manifold (Coifman '06)
- The first few eigenfunctions are usually enough to faithfully represent $\mathbb{D}$ in a low-dimensional ambient space

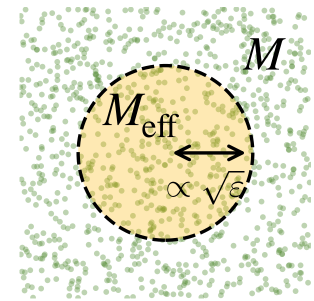


Why do this?

- Mesh-free PDE solving (Vaughn *et al*. '19, Jiang & Harlim '20)
- Compression of other operators via projection onto Laplacian eigenbasis.
  - *E.g*. Perron-Frobenius operators for forecasting: Berry *et al*. '15, Giannakis '19

Convergence of diffusion maps

$$\boxed{M_{\text{eff}} = O(M\varepsilon^{d/2})}$$



Weighted matrix P
$= $ finite-rank operator $\mathscr{P}_\varepsilon^M$

"Variance" error
$M \to \infty$

Pointwise
convergence
$\varepsilon \to 0, \varepsilon^{-2} M_{\text{eff}} \to \infty$

Weighted kernel
operator $\mathscr{P}_\varepsilon$
(spatial continuum limit)

Laplace-Beltrami
semigroup $e^{\varepsilon \mathscr{L}}$

"Bias" error
$\varepsilon \to 0$

Pointwise convergence results

Pointwise error bounds are well known:

$$(\mathscr{P}_\varepsilon^M \phi)(x) - (e^{\varepsilon \mathscr{L}} \phi)(x)$$

- The **bias error** (Singer '06):

$$|(\mathscr{P}_\varepsilon \phi)(x) - (e^{\varepsilon \mathscr{L}} \phi)(x)| = O(\varepsilon^2 \|\phi\|_{\mathrm{Lip}})$$

- If u does not depend on the x^i (e.g. $\alpha = 0$), the **variance error** is just a CLT estimate

$$|(\mathscr{P}_\varepsilon^M \phi)(x) - (\mathscr{P}_\varepsilon \phi)(x)| = O(M_{\mathrm{eff}}^{-1/2} \|\phi\|_\infty)$$

Spectral convergence results

From timestep, expect magnification of pointwise error by ε^{-1}

Expect pointwise convergence rates hold for spectral data, but…

Bias error estimates are typically $L^2 \to L^2$ error: $O(\varepsilon^{1/2})$.

Variance error estimates:

- Via compact embedding of Glivenko-Cantelli classes
  - Establish qualitative convergence, with bad rates
  - e.g. Shi (2015): variance error $= M_{\text{eff}}^{-1/2}\varepsilon^{-3d/4-3}$.
- Optimal transport results
  - Garcia Trillos *et al.* (2019): OT rate $O(M_{\text{eff}}^{-1/d+o(1)})$
  - Calder and Garcia Trillos (2019): bootstraps off previous results using central limit theorem + Rayleigh quotients
    - Issue of recursively applying CLT

How to prove pointwise convergence rates hold for spectral data, for a broad range of problems?

Structure of talk

- Variance error: local embedding estimates
- Bias error: PDE operator theory
- Sinkhorn weights: a nice application of the tools

NB: for simplicity, we will make our manifold a flat torus:
$\mathbb{D} = (\mathbb{R}/\mathbb{Z})^d$…

Interpolating the matrix

How does the matrix P relate to the functional operator $e^{\varepsilon\mathscr{L}}$?

If $z = (\phi(x^i))_{i=1,\ldots,M}$ for some function ϕ, then

$$(Kz)_i = \frac{1}{M}\sum_{j=1}^{M} g_\varepsilon(x^i - x^j)\phi(x^j) =: \mathscr{K}_\varepsilon^M(x^i)$$

So, there is a natural way to interpolate:

- Standard weight $u = (K1)^{-\alpha}$ is $U_\varepsilon^M = (\mathscr{K}_\varepsilon^M 1)^{-\alpha}$
- Left-hand weight $v = 1/(Ku)$ is $V_\varepsilon^M = 1/(\mathscr{K}_\varepsilon^M U_\varepsilon^M)$
- Weighted matrix $P = \mathrm{diag}(v)\,K\,\mathrm{diag}(u)$ is $\mathscr{P}_\varepsilon^M = V_\varepsilon^M \mathscr{K}_\varepsilon^M U_\varepsilon^M$.

Variance error

Kernel matrix K can be interpolated on functions:

$$\mathcal{K}_\varepsilon^M \phi(x) = \frac{1}{M}\sum_{i=1}^{M} g_\varepsilon(x - x^i)\phi(x^i) = \mathscr{C}_\varepsilon[\rho^M \phi],$$

convolution by g_ε Sample measure

The continuum limit is then

$$\mathcal{K}_\varepsilon \phi(x) = \int_{\mathbb{D}} g_\varepsilon(x - y)\phi(y)\rho(y)\,\mathrm{d}y = \mathscr{C}_\varepsilon[\rho\phi].$$

$\mathcal{K}_\varepsilon^M \phi(x)$ is just an empirical mean with expectation $\mathcal{K}_\varepsilon \phi(x)$, so CLT results give pointwise convergence:

$$\mathbb{P}\left[\,|\,\mathcal{K}_\varepsilon^M \phi(x) - \mathcal{K}_\varepsilon \phi(x)\,| > c\|\phi\|_\infty\right] \leq C_0 e^{-C_1 M \varepsilon^{-d/2} c^2}.$$

How can we extend this to operator convergence?

Variance error: central limit theorem

We can extend using compactness/covering arguments.

- *Function norm*: $\mathbb{D}$ can be covered by $O(\xi^{-d})$ balls of radius ξ, $\mathbb{P}\left[\|\mathscr{K}_\varepsilon^M \phi - \mathscr{K}_\varepsilon \phi\|_{C^0} > (c + \xi \operatorname{Lip} g_\varepsilon)\|\phi\|_{C^0}\right] \leq O(\xi^{-d} e^{-C_1 M \varepsilon^{-d/2} c^2})$.
- *Operator norm*: harder. Need ϕ to be in a function space embedding (very) compactly into C^0.
  - This function space should contain $\operatorname{im} \mathscr{K}_\varepsilon^M, \operatorname{im} \mathscr{K}_\varepsilon$.

Variance error: Hardy spaces

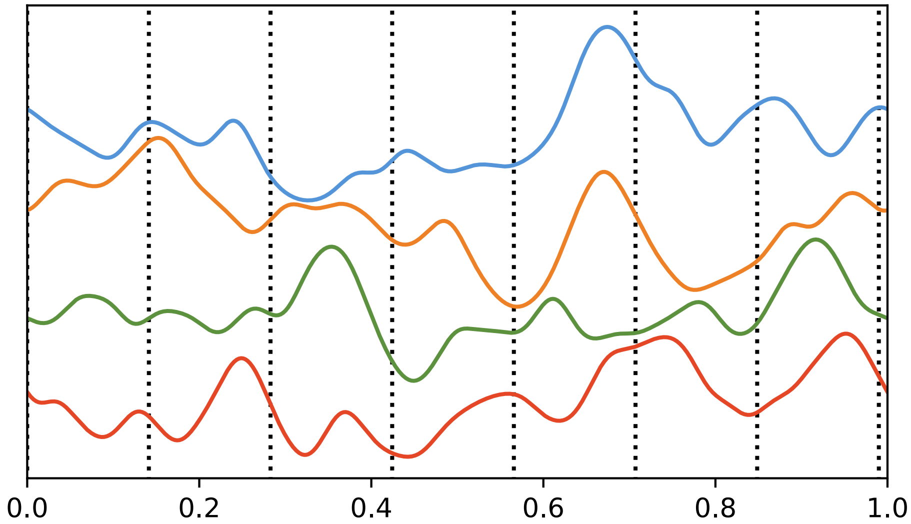
We will choose our "strong" function spaces as the scale of Hardy spaces

$$H^\infty_\varepsilon := \left\{ \phi \in C^0(\mathbb{D}_\varepsilon) : \phi \text{ analytic on } \operatorname{int} \mathbb{D}_\varepsilon \right\},$$

where $\mathbb{D}_\varepsilon$ is the complex $\sqrt{\varepsilon}$-fattening of $\mathbb{D}$.

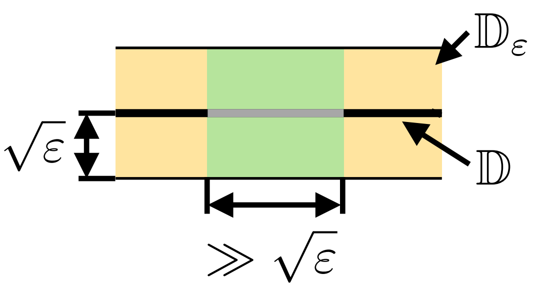
These spaces are useful because

$\|\mathscr{C}_\varepsilon\|_{C^0 \to H^\infty_\varepsilon} = O(1)$.



Variance error: Local embedding

- The H^∞_ε unit ball can only be covered using $O(e^{\varepsilon^{-d/2}(\log \xi)^d})$ C^0 balls of radius ξ. OK for $\varepsilon = 1\ldots$
- Fortunately, our operator $\mathscr{K}^M_\varepsilon$ is very localised.
  - $\mathscr{K}^M_\varepsilon \phi(x)$ mostly depends on ϕ in an $O(\sqrt{\varepsilon})$-neighbourhood of x.
  - H^∞_ε on this neighbourhood has nice covering numbers
  - Upshot:



$$\mathbb{P}\left[\|\mathscr{K}^M_\varepsilon - \mathscr{K}_\varepsilon\|_{H^\infty_\varepsilon \to C^0} > c\right] \le e^{C_2(\log c + \log \varepsilon^{-1})^{2d+1} - C_1 M \varepsilon^{-d/2} c^2}.$$

Variance error: Norm convergence

We can use the divisibility of the Gaussian kernel so

$$\mathscr{K}_\varepsilon^M = \mathscr{C}_\varepsilon \rho^M = \mathscr{C}_{\varepsilon/2} \mathscr{K}_{\varepsilon/2}^M,$$

and that $\|\mathscr{C}_\varepsilon\|_{C^0 \to H_\varepsilon^\infty} = O(1)$ to show

$$\|\mathscr{K}_\varepsilon^M - \mathscr{K}_\varepsilon\|_{H_\varepsilon^\infty \to H_\varepsilon^\infty} = O(1) \times \delta,$$

where

$$\delta := \|\mathscr{K}_{\varepsilon/2}^M - \mathscr{K}_{\varepsilon/2}\|_{H_\varepsilon^\infty \to C^0} = O(M^{-1/2}\varepsilon^{-d/4} \times \text{log terms}).$$

From here we only need to use δ to think about our error bounds.

Variance error: weighted operator

We should now consider the convergence as $M \to \infty$ of

$$U_\varepsilon^M := (\mathscr{K}_\varepsilon^M 1)^{-\alpha}$$

and

$$V_\varepsilon^M := (\mathscr{K}_\varepsilon^M U_\varepsilon^M)^{-1}.$$

Fortunately, norm convergence gives us

$$\|U_\varepsilon^M - U_\varepsilon\|_{H_\varepsilon^\infty}, \|V_\varepsilon^M - V_\varepsilon\|_{H_\varepsilon^\infty} = O(\delta)$$

So with $\mathscr{P}_\varepsilon^M := V_\varepsilon^M \mathscr{K}_\varepsilon^M U_\varepsilon^M$,

$$\|\mathscr{P}_\varepsilon^M - \mathscr{P}_\varepsilon\|_{H_\varepsilon^\infty} = O(\delta).$$

Bias error: PDE limit

Now need to compare $\mathscr{P}_\varepsilon$ and $e^{\varepsilon \mathscr{L}}$

Bias error: PDE limit

Now need to compare $\mathscr{P}_\varepsilon^{\mathbf{n}}$ and $e^{\varepsilon \mathbf{n} \mathscr{L}}$ for $n = O(\varepsilon^{-1})$

These are both Markov operators, and since

$$\mathscr{P}_\varepsilon = V_\varepsilon \mathscr{K}_\varepsilon U_\varepsilon = \frac{1}{e^{\varepsilon\Delta/2}[\rho U_\varepsilon]} e^{\varepsilon \Delta/2} \rho U_\varepsilon,$$

they are $0 \to n\varepsilon$ evolution operators of the PDEs

$$\partial_t \phi^t = \mathscr{L}\phi^t = \frac{1}{2}\Delta\phi^t + (1-\alpha)\nabla \log \rho \cdot \nabla \phi^t$$

and

$\varepsilon\{t/\varepsilon\}$

$$\partial_t \phi^t = \frac{1}{2}\Delta \phi^t + \nabla \log e^{\{t\}_\varepsilon \Delta/2}[\rho U_\varepsilon] \cdot \nabla \phi^t.$$

Because $\rho U_\varepsilon = \rho^{1-\alpha} + O(\varepsilon)$, the two drift terms are $O(\varepsilon)$-close, and we get an error

$$\|\mathscr{P}_\varepsilon^n - e^{n\varepsilon\mathscr{L}}\|_{C^{2+\beta} \to C^0} \leq O(\varepsilon)$$

for $n = O(\varepsilon^{-1})$. Playing around with negative Sobolev spaces means this works for low regularity $\rho \in C^{3/2+\beta}$.

Spectral convergence

Look at spectral projections on nth powers of operators for $n = O(\varepsilon^{-1})$.

- Gaussian and Schauder estimates give us uniform bounds on $\|e^{n\varepsilon\mathscr{L}}, \mathscr{P}_\varepsilon^n\|_{L^p\to C^k}$ for any k, p.
- A priori bounds on norm of resolvent $R(e^{n\varepsilon\mathscr{L}}, \lambda)$ in $L^2(U_\varepsilon/V_\varepsilon)$ by orthogonality.
- Use our operator convergence estimates to get estimates on resolvent error from $C^{2+\beta} \to C^0$.
- Use that spectral projection of operator $\mathscr{A}$ onto eigenvalues in $B(\lambda, r)$ is

$$\Pi = \frac{1}{2\pi i}\int_{C(\lambda,r)} R(\mathscr{A}, z)\,\mathrm{d}z.$$

Note: could use Rayleigh quotients instead

Spectral convergence

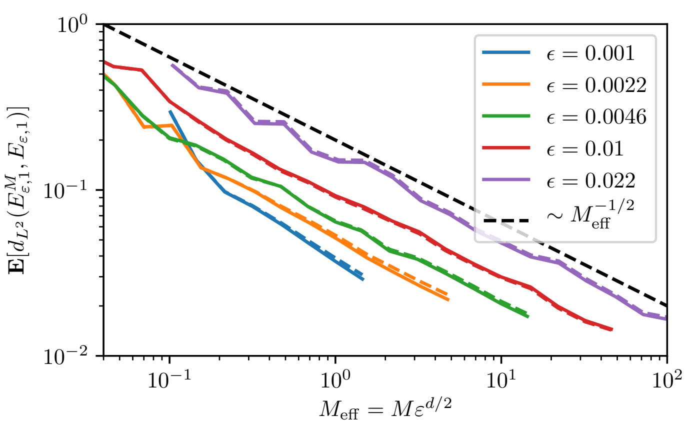
- This gives us that eigenvalues and eigenvectors (in C^0) of $\mathscr{P}_\varepsilon^M$ converge to those of $e^{\varepsilon\mathscr{L}}$ as

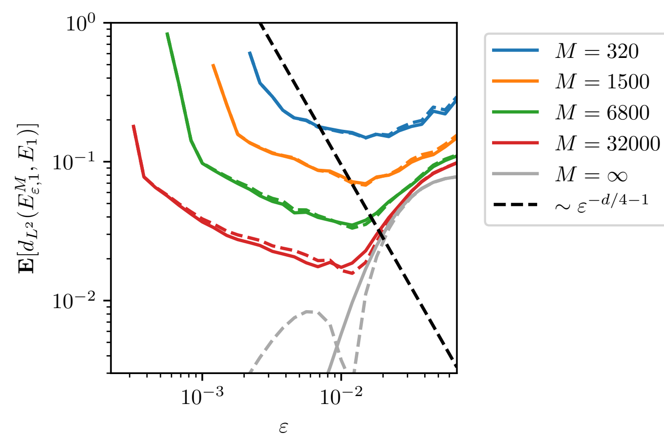
$$O(M^{-1/2}\varepsilon^{-d/4-1} + \varepsilon)$$

Pointwise variance error $\times\, \varepsilon^{-1}$

Pointwise bias error $\times\, \varepsilon^{-1}$

- The bias error is optimal but the variance error for spectral data can be improved by ϵ^s for some s small





Sinkhorn weights

The $M \to \infty$ operator convergence theory lets us work with all sorts of interesting particle discretisation problems.

What about the diffusion maps normalisation $u = v$?

$$P = \mathrm{diag}(u)\, K\, \mathrm{diag}(u).$$

- P is symmetric, so eigenfunctions are orthogonal
- Total integral and constant functions are preserved
- Other nice properties?

Sinkhorn weights

The weight u solves the *Sinkhorn problem*

$$u \times (Ku) \equiv 1.$$

In function space:

$$U_\varepsilon^M \times \mathscr{K}_\varepsilon^M[U_\varepsilon^M] \equiv 1.$$

We can link this to continuum limit U_ε just by using the implicit function theorem.

In particular, for all ε small and $\delta \leq C_2$,

$$\|U_\varepsilon^M - U_\varepsilon\|_{H_\varepsilon^\infty} \leq C_3 \delta.$$

Sinkhorn weights: improved bias error

But what do the U_ε look like?

From the Sinkhorn problem

$$U_\varepsilon \times \mathscr{K}_\varepsilon U_\varepsilon = U_\varepsilon \times \mathscr{C}_\varepsilon[\rho U_\varepsilon] \equiv 1,$$

we expect

$$U_\varepsilon = \rho^{-1/2} + O(\varepsilon).$$

In particular, as $\varepsilon \to 0$ we expect convergence to a Langevin diffusion ($\alpha = 1/2$).

Proof: write Sinkhorn iteration in the log-domain as a rapidly-oscillating nonlinear PDE and average; $\log \rho^{1/2} U_\varepsilon$ is contained in a limit cycle.

Sinkhorn weights: improved bias error

We again have that $e^{\varepsilon n \mathscr{L}}$ and $\mathscr{P}_\varepsilon^n$ are respectively $0 \to n\varepsilon$ evolution operators of the PDEs

$$\partial_t \phi^t = \mathscr{L}\phi^t = \frac{1}{2}\Delta\phi^t + \frac{1}{2}\nabla \log \rho \cdot \nabla \phi^t$$

and

$$\partial_t \phi^t = \frac{1}{2}\Delta\phi^t + \nabla \log e^{\{t\}_\varepsilon \Delta/2}[\rho U_\varepsilon] \cdot \nabla \phi^t.$$

We can approximate $\mathscr{P}_\varepsilon^n$ to $O(\varepsilon^2)$ by averaging over the drift term:

$$\partial_t \phi^t \approx \frac{1}{2}\Delta\phi^t + \nabla \bar{w}_\varepsilon \cdot \nabla \phi^t$$

where

$$\bar{w}_\varepsilon = \varepsilon^{-1} \int_0^\varepsilon \log e^{t\Delta/2}[\rho U_\varepsilon] \, \mathrm{d}t$$

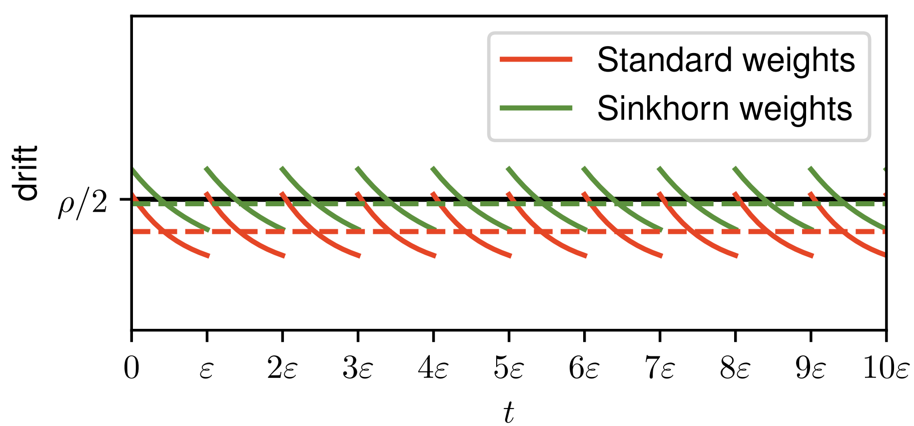
Sinkhorn weights: improved bias error

We have

$$
\begin{aligned}
\bar{w}_\varepsilon &= \varepsilon^{-1} \int_0^\varepsilon \log e^{t\Delta/2}[\rho U_\varepsilon] \, \mathrm{d}t \\
&\approx \tfrac{1}{2}(\log e^{\varepsilon\Delta/2}[\rho U_\varepsilon] + \log \rho U_\varepsilon) \\
&= \tfrac{1}{2}(\log U_\varepsilon^{-1} + \log U_\varepsilon + \log \rho) \\
&= \tfrac{1}{2} \log \rho
\end{aligned}
$$

Trapezoidal rule

So the averaging comes out of the symmetry of the operator!



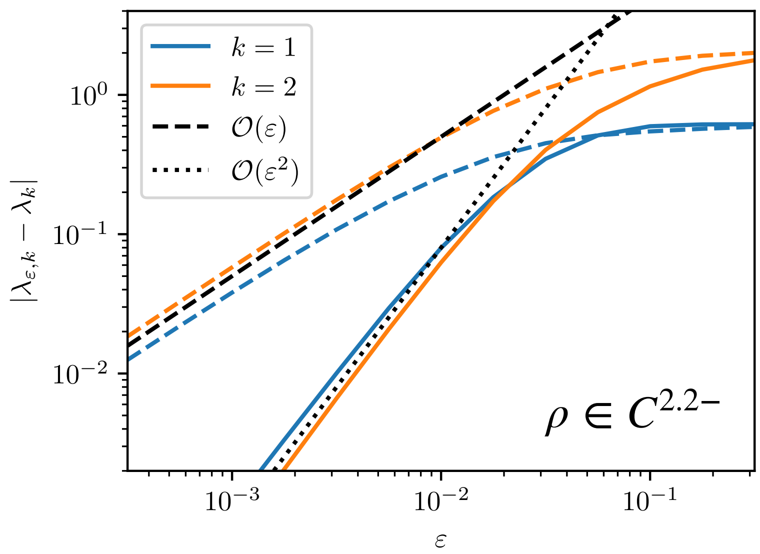
Sinkhorn weights: improved bias error

Get an error

$$\|\mathscr{P}_\varepsilon^n - e^{n\varepsilon\mathscr{L}}\|_{C^{3+\beta}\to C^0} \leq O(\varepsilon^2)$$

for $n = O(\varepsilon^{-1})$. This is the best possible asymptotic rate for weighted operators.

Using negative Sobolev spaces means this works for low regularity $\rho \in C^{2+\beta}$.



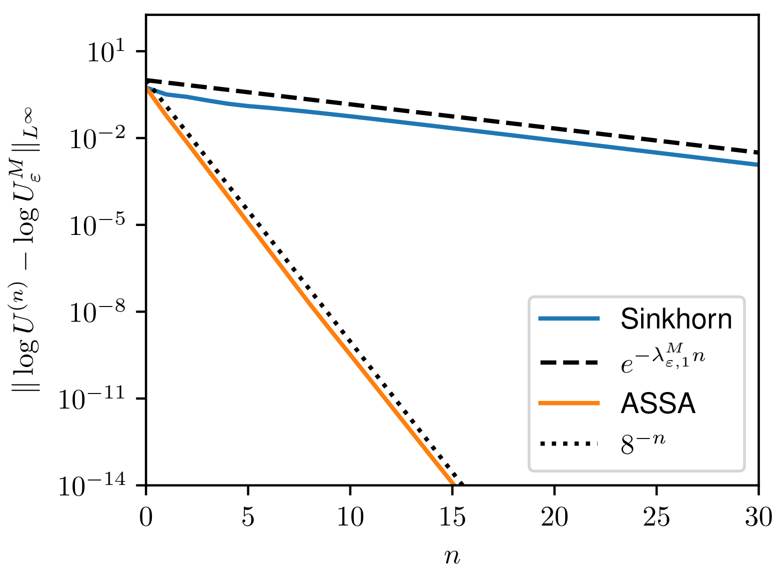
Accelerated Sinkhorn algorithm

How to actually calculate the Sinkhorn weights?

- Standard Sinkhorn iteration: use that u is the fixed point of $u^{(n+1)} = 1/(Ku^{(n)})$.
  - Jacobian about the fixed point is conjugate to $-P$ (weighted matrix)
  - Convergence is $\sim \lambda_1^n$, where $\lambda_1 = 1 - O(\varepsilon)$ is second eigenvalue. (Slow)

Accelerated Sinkhorn algorithm

- Instead:
  - Sinkhorn step: $u^{(n+1/3)} = 1/(Ku^{(n)})$
  - Sinkhorn step: $u^{(n+2/3)} = 1/(Ku^{(n+1/3)})$
  - Geometric mean: $u^{(n+1)} = \sqrt{u^{(n+1/3)}u^{(n+2/3)}}$
- The Jacobian of this algorithm is $P(1-P)/2$. Because $\sigma(P) \subseteq [0,1]$, this converges $\sim 8^{-n}$.



Conclusion

- Near-optimal bounds on spectral convergence rates *(for Gaussian kernels, on flat domains)*.
- Broadly applicable theoretical techniques for convergence of kernel methods
- Sinkhorn normalisation for diffusion maps works, and is the best choice for Langevin dynamics

Wormell, C.L. and Reich, S., Spectral convergence of diffusion maps: improved error bounds and an alternative normalisation (2020). **arXiv**:2006.02037