# Approximation rates matter: sharp estimates for Target Measure Diffusion Maps and applications to molecular dynamics

Shashank Sule (UMD), Luke Evans (Flatiron Institute), Maria Cameron (UMD)

## Target Measure Diffusion Maps

**Goal**: Approximate infinitesimal generator $\mathcal{L}$ for the time-reversible dynamics governed by the SDE

(1) $dX_t = \beta^{-1}\nabla \log(\mu)(x)dt + \sqrt{2\beta^{-1}}dW$

We recall that the generator is given by

(2) $\mathcal{L} = \beta^{-1}(\Delta - \nabla\log(\mu)\cdot\nabla)$.

**Input**: Dataset $\mathcal{X} \subset \mathcal{M} \sim d\rho$, target measure $\mu$, kernel bandwidth $\epsilon$, and kernel matrix

(3) $[K_\epsilon]_{ij} = \exp(\epsilon^{-1}\|x_i - x_j\|_2^2) := k_\epsilon(x_i, x_j)$

**Step 1 (Kernel Density estimation):**

(4) $[M]_{ij} = \delta_{ij}\mu(x_i)$ $\quad D_\epsilon = n^{-1}\mathrm{diag}(K_\epsilon \mathbf{1})$

**Step 2 (Target measure renormalization)**:

(5) $K_{\epsilon,\mu} = K_\epsilon D_\epsilon^{-1} M^{1/2}$

**Step 3/Output: Generator**:

$D_{\epsilon,\mu} = n^{-1}\mathrm{diag}(K_{\epsilon,\mu}\mathbf{1}), P_{\epsilon,\mu} = D_{\epsilon,\mu}^{-1}K_{\epsilon,\mu}$

(6) $L_{\epsilon,\mu}^{(n)} = \dfrac{I - P_{\epsilon,\mu}}{\epsilon}$

**Banisch et al. [1] prove that $4L_{\epsilon,\mu}$ is a Monte Carlo approximation of $\mathcal{L}$:**

(7) $4\beta^{-1}L_{\epsilon,\mu}^{(n)}f(x) \to \mathcal{L}f(x) + O(\epsilon)$ as $n \to \infty$

## Computing transition rates

Let $A, B \subset \Omega$ be disjoint closed subsets of $\mathcal{M}$ and $N_{AB}(T)$ the number of transitions from A to B up to time T. Then using transition path theory (TPT) [3] the transition rate $\nu_{AB}$ is given by

$$\nu_{AB} = \lim_{T\to\infty} T^{-1}N_{AB}(T) = \int_\Omega \|\nabla q\|_2^2 \mu(x)\, d\mathrm{vol}(x)$$

Here $\Omega = \mathcal{M}\setminus(A\cup B)$ and q is the committor function defined by the *committor BVP*:

(8) $\mathcal{L}q = 0, \quad q\,|_{\partial A} = 0, \quad q\,|_{\partial B} = 1$

**TMD maps can be used as a meshless algorithm for numerically solving the committor problem** since $L_{\epsilon,\mu}$ is a discrete approximation to $\mathcal{L}$:

$$L_{\epsilon,\mu}q_{TMD}(x) = 0 \quad q_{TMD}\,|_{A\cup B} = \mathbb{1}_B$$

## References


[1] Ralf Banisch, Zofia Trstanova, Andreas Bittracher, Stefan Klus, and Peter Koltai. "Diffusion maps tailored to arbitrary non-degenerate Itô processes". In: *Applied and Computational Harmonic Analysis* 48.1 (2020), pp. 242–265.

[2] Ronald R Coifman and Stéphane Lafon. "Diffusion maps". In: *Applied and computational harmonic analysis* 21.1 (2006), pp. 5–30.

[3] Eric Vanden-Eijnden and Weinan E. "Towards a theory of transition paths". In: *Journal of statistical physics* 123.3 (2006), pp. 503–523.


## Sharp estimates for Target Measure Diffusion Maps show that convergence is faster when approximating the generator of an overdamped Langevin diffusion on its committor function with a quasi-uniform sampling density:

Let the i.i.d samples $\mathcal{X}(n)$, bandwidth $\epsilon$, and target measure $\mu$ be used for constructing the TMD map generator $L_{\epsilon,\mu}^{(n)}$. Then for $f \in C^2(\mathcal{M})$ and large enough n and small enough $\epsilon$, with probability greater than $1 - 2n^{-3}$ we have:

$$|4\beta^{-1}L_{\epsilon,\mu}^{(n)}f(x) - \mathcal{L}f(x)| = \frac{\epsilon}{(2\pi)^{d/4}}\sqrt{\frac{\log n}{\rho(x_i)n\epsilon^{4+d/2}}}\left(2\|\nabla_{\mathcal{M}}f(x_i)\|\epsilon^{1/2} + 11f(x_i)\right) + \epsilon\left(\mathcal{B}_1[f,\mu] + \mathcal{B}_2[f,\mu,\rho] + \mathcal{B}_3[f,\mu,\rho]\right)$$

The expressions for $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ are given by:

$$\mathcal{B}_2[f,\mu,\rho] := -\frac{1}{16}\left(2\nabla f\cdot\nabla\left(\mu^{1/2}\frac{\Delta\rho}{\rho}\right) + \left(\mu^{1/2}\frac{\Delta\rho}{\rho}\right)f\right)$$

**Cancels when $\rho$ is quasi-uniform**

$$\mathcal{B}_3[f,\mu,\rho] := \frac{1}{16}\left[\frac{\Delta(\mu^{1/2})}{\mu^{1/2}} - \left(\frac{\Delta\rho}{\rho} - \omega\right)\right]\left[f\frac{\Delta(\mu^{1/2})}{\mu^{1/2}} - \frac{\Delta(\mu^{1/2}f)}{\mu^{1/2}}\right]$$

**Cancels when $\mathcal{L}f(x) = 0$**

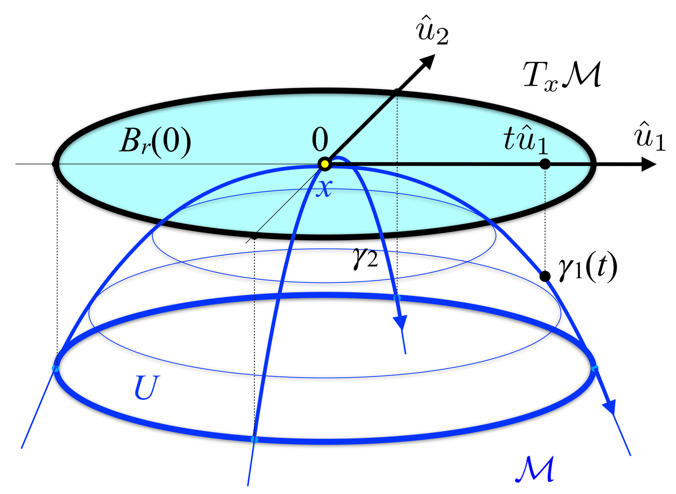
$$\mathcal{B}_1[f,\mu] := \frac{1}{4}\left[\mathcal{Q}\left(f\mu^{1/2}\right) - f\mathcal{Q}(\mu^{1/2})\right] + \frac{\epsilon}{16}\left(2\nabla f\cdot\nabla\left(\mu^{1/2}\omega\right) + (\mu^{1/2}\omega)\Delta f\right)$$

**Reduces when $\mathcal{M}$ is flat**

The key strategy to obtaining the prefactors is the expansion of integrals of the form

$$\mathcal{K}_\epsilon\rho(x) := \int_{\mathcal{M}} k_\epsilon(x,y)\rho(y)\,dy$$

We use up to *fourth order* Taylor expansions of $\mathcal{K}_\epsilon$ in normal coordinates [2]:



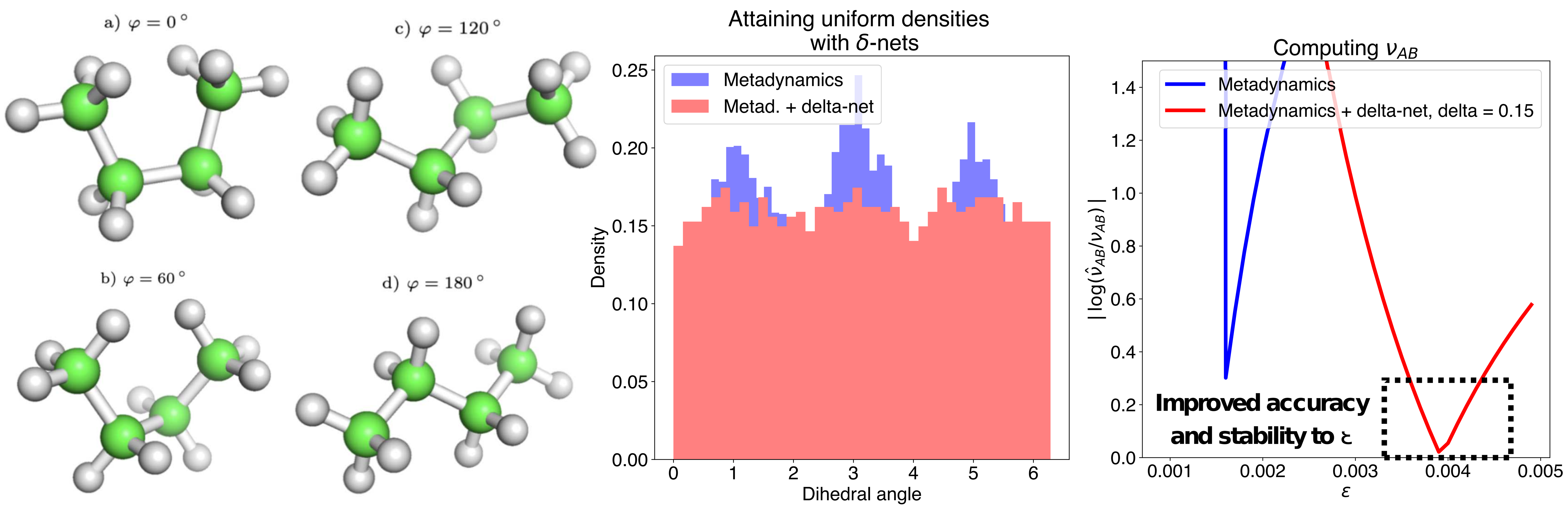

$$\|s(u) - s(0)\|_2^2 = \|u\|_2^2 + \mathcal{C}_x^1(u), \mathcal{C}_x^1(u) = O(\|u\|_2^4)$$
$$dy(u) = 1 + \mathcal{C}_x^2(u), \mathcal{C}_x^2(u) = O(\|u\|_2^2)$$

Using the **discrete maximum principle** we prove an error estimate for solutions to Dirichlet BVP's:

**Corollary.** There exists a constant $C(\mathcal{M})$ such that with probability greater than $1 - 2n^{-2}$,

$$\|q_{TMD}(x_i) - q(x_i)\|_\infty \le C\|4\beta^{-1}L_{\epsilon,\mu}^{(n)}q - \mathcal{L}q\|_\infty.$$

## When less is more: Removing points to enhance spatial uniformity ($\delta$-nets) improves the robustness of diffusion-map based transition rates for conformational changes in molecules!




Computing $\nu_{AB}$ for $C_4H_{10}$ with TMD map in 12-D,

$A = \theta^{-1}(B_{0.2}(\pi))$,

$B = \theta^{-1}(B_{0.1}(\pi/3) \sqcup B_{0.1}(5\pi/3))$

| Data | $\varepsilon^*$ | $\nu_{AB}$ |
|---|---|---|
| Metad. | 0.0016 | 0.0109 |
| Metad + delta-net | 0.0039 | 0.0112 |
| True value | N/A | 0.0114 |

**Figure: (Left to right)** The butane molecule $C_4H_{10}$ can be effectively coarse—grained with the dihedral angle in the carbon backbone; $C_4H_{10}$ is stable around $\theta \approx \pi$ and metastable around $\theta \approx \pi/3, 5\pi/3$ and the intermediate angles can be sampled using metadynamics, with $\delta$—net in $\mathbb{R}^{12}$ further uniformizing the distribution; computing $\nu_{AB}$ using TMD map shows that indeed this uniformization aids in improving the estimation of the transition rate and stabilizing it to choice of the bandwidth.