Characterizing long timescale phenomena with trajectory data

Jonathan Weare

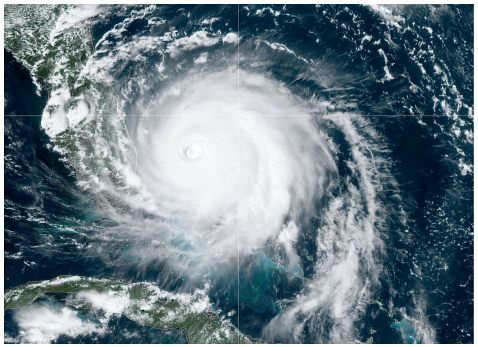
Courant Institute, New York University

with

Aaron Dinner, Douglas Dow, Dimitrios Giannakis, John Strahan, Erik Thiede, and Rob Webber

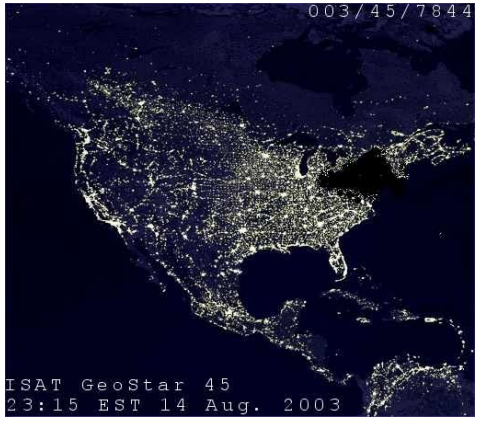
April 23, 2020

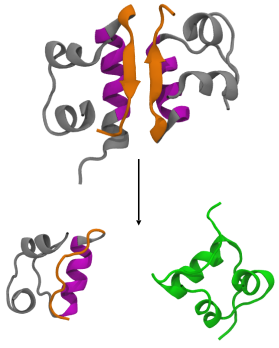
The most interesting events often occur very infrequently



The most intense (and damaging) tropical cyclones occur about once a decade, but processes (e.g. gravity waves) on the timescale of seconds to minutes must be resolved in numerical integration of weather models.

Major blackouts not (counting those due to major storms) are extremely rare in the US. The last one occurred in 2003. Within minutes millions (eventually 55 million in the US and Canada) were affected.



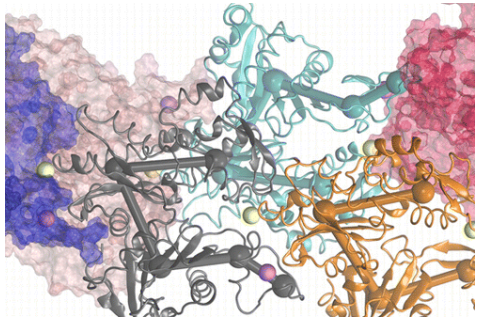


Disassociation of the insulin dimer has important therapeutic implications. It occurs roughly on the microsecond to millescond timescale. That's up to 12 orders of magnitude longer than the time scale of bond vibrations ($10^{-15}s$)

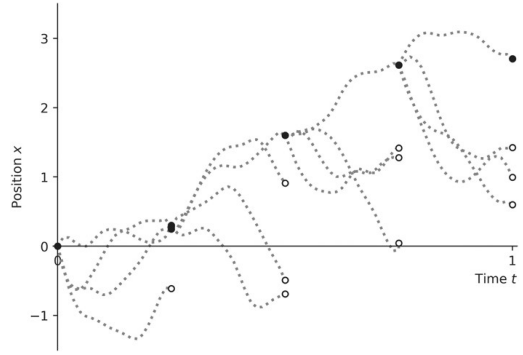
Three common approaches to interrogating long timescale processes:

1. **Direct simulation**: Find (or build) a really fast computer (Anton2 shown here) and integrate for as long as you can.

2. **Coarse graining**: Build a cheaper but less accurate model that you can run to very long timescales. (image from the Voth group)



3. **Rare event simulation**: Try to "trick" the model into undergoing the event of interest quickly while maintaining the ability to estimate unbiased statistics.



However you generate the data...

It needs to be processed it for understanding of the long timescale phenomena

However you generate the data...

It needs to be processed it for understanding of the long timescale phenomena

In this talk I:

1. Report on our analysis of a well established dynamic spectral estimation approach that approximates the slowest decorrelating features of the system.

2. Describe a family of methods that we have developed that uses short trajectories to compute statistics describing a specified long timescale event.

The Variational Approach to Conformational Dynamics

Rob Webber, Erik Thiede, Douglas Dow, Aaron Dinner

VAC estimates eigenvalues and eigenspaces of the transition operator $\mathcal{T}^\tau$ with action

$$\mathcal{T}^\tau f(x) = E\left[f(X_\tau) \,|\, X_0 = x\right]$$

VAC assumes that X_t has unique ergodic probability measure μ and that $\mathcal{T}^\tau : L^2(\mu) \to L^2(\mu)$ is self-adjoint.

Our analysis assumes quasi-compactness:

$$\mathcal{T}^\tau = \sum_{i=1}^{r} e^{-\sigma_i \tau} \mathrm{proj}[\eta_i] + \mathcal{R}^\tau \quad \text{with} \quad \|\mathcal{R}^\tau\|_2 \leq e^{-\sigma_{r+1}\tau}$$

where $1 = e^{-\sigma_1 \tau} > e^{-\sigma_2 \tau} \geq \cdots \geq e^{-\sigma_{r+1}\tau}$ and where $\eta_1, \eta_2, \ldots, \eta_r$ are eigenfunctions. $\eta_1 \equiv 1$.

VAC estimates $\mathrm{span}_{i \leq k}\{\eta_i\}$

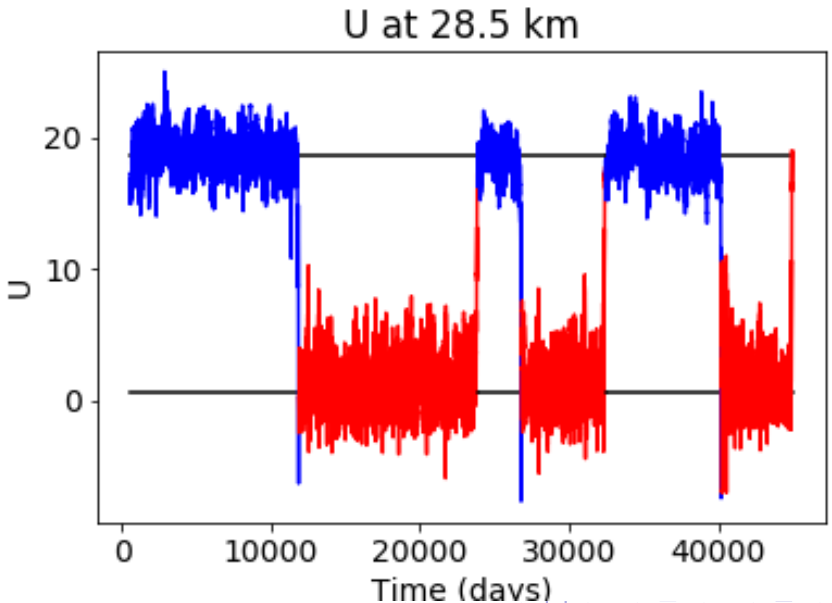
the most slowly decorrelating functions of the system.

If η belongs to the linear span of $\eta_2, \ldots, \eta_k$, then

$$\mathrm{corr}_\mu \left[\eta\left(X_0\right), \eta\left(X_\tau\right)\right] = \frac{\left\langle \eta, \mathcal{T}^\tau \eta \right\rangle_\mu}{\left\langle \eta, \eta \right\rangle_\mu} \geq e^{-\sigma_k \tau}.$$

If u is orthogonal to $\eta_1, \ldots, \eta_k$ then,

$$\mathrm{corr}_\mu \left[u\left(X_0\right), u\left(X_\tau\right)\right] = \frac{\left\langle u, \mathcal{T}^\tau u \right\rangle_\mu}{\left\langle u, u \right\rangle_\mu} \leq e^{-\sigma_{k+1} \tau}.$$



VAC workflow...

1. Generate samples of X_0 from μ and then X_τ given X_0.
2. Choose a set of basis functions $\phi_1, \phi_2, \ldots, \phi_n$.
3. Use samples to build estimates

$$\hat{C}_{ij}(0) \approx C_{ij}(0) = \langle \phi_i, \phi_j \rangle_\mu = E[\phi_i(X_0)\phi_j(X_0)]$$
$$\hat{C}_{ij}(\tau) \approx C_{ij}(\tau) = \langle \phi_i, \mathcal{T}^\tau \phi_j \rangle_\mu = E[\phi_i(X_0)\phi_j(X_\tau)]$$

4. Solve for the eigenpairs $(\hat{\lambda}_i^\tau, \hat{v}^i)$ of $\hat{C}(0)^{-1}\hat{C}(\tau)$.
5. Return approximate eigenfunctions $\hat{\gamma}_i^\tau = \sum_j \hat{v}_j^i \phi_j$.

Most common variants:

Markov State Models (MSM)s: choose a basis of indicator functions on a partition of space (usually found by clustering the data).

Time-lagged Independent Component Analysis (TICA): choose the coordinate axes as a basis.

VAC workflow...

1. Generate samples of X_0 from μ and then X_τ given X_0.
2. Choose a set of basis functions $\phi_1, \phi_2, \ldots, \phi_n$.
3. Use samples to build estimates

$$\hat{C}_{ij}(0) \approx C_{ij}(0) = \langle \phi_i, \phi_j \rangle_\mu = E[\phi_i(X_0)\phi_j(X_0)]$$
$$\hat{C}_{ij}(\tau) \approx C_{ij}(\tau) = \langle \phi_i, \mathcal{T}^\tau \phi_j \rangle_\mu = E[\phi_i(X_0)\phi_j(X_\tau)]$$

4. Solve for the eigenpairs $(\hat{\lambda}_i^\tau, \hat{v}^i)$ of $\hat{C}(0)^{-1}\hat{C}(\tau)$.
5. Return approximate eigenfunctions $\hat{\gamma}_i^\tau = \sum_j \hat{v}_j^i \phi_j$.

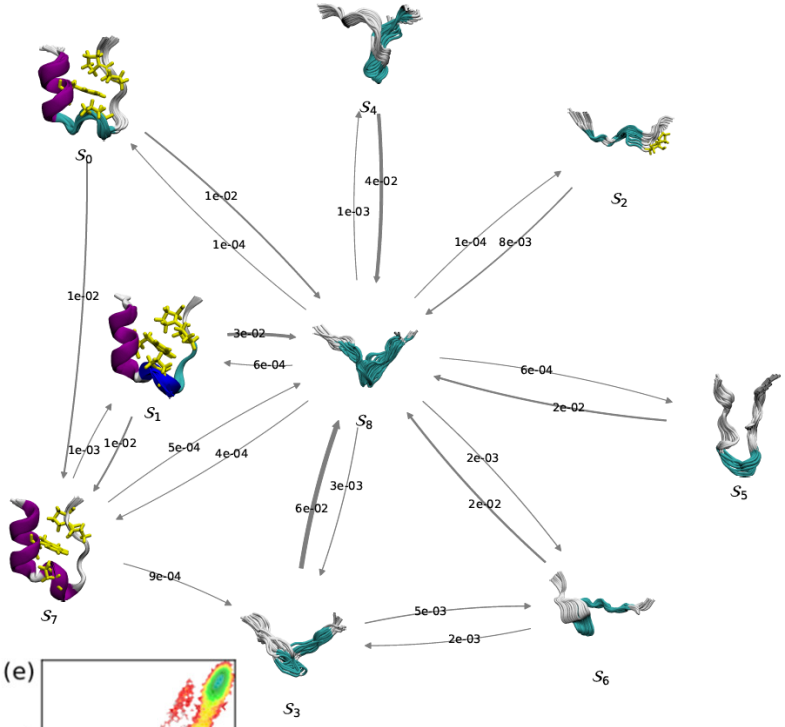
Most common variants:

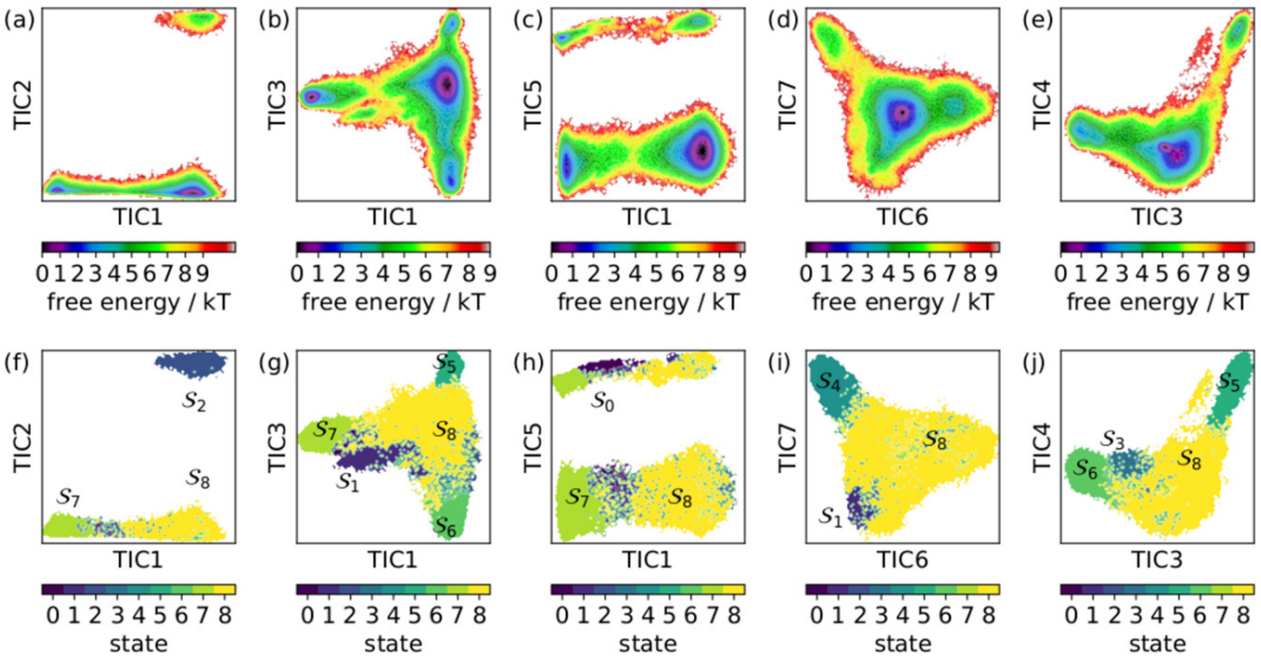
Markov State Models (MSM)s: choose a basis of indicator functions on a partition of space (usually found by clustering the data).

Time-lagged Independent Component Analysis (TICA): choose the coordinate axes as a basis.

Trp cage folding

A well studied fast folding mini-protein. This study used a long trajectory with between 12 and 31 folding/unfolding events generated on Anton. We'll see this example again later.





Sidky et al. 2019

When should we trust VAC?

We divide the VAC error into two contributions:

1. **Approximation error:** If $\hat{C} = C$ then VAC approximate eigenfunctions are $\gamma_i^\tau = \sum_j v_j^i \phi_j$ where (λ_j, v^j) are eigenpairs of $C(0)^{-1}C(\tau)$. How big is

$$dist_{\mathrm{F}}\left(\mathrm{span}_{i\leq k}\{\gamma_i^\tau\}, \mathrm{span}_{i\leq k}\{\eta_i\}\right)?$$

2. **Estimation error:** In practice we use sampled data to build the estimate $\hat{C}$ of C. How big is

$$dist_{\mathrm{F}}\left(\mathrm{span}_{i\leq k}\{\hat{\gamma}_i^\tau\}, \mathrm{span}_{i\leq k}\{\gamma_i^\tau\}\right)?$$

We use the projection distance $dist_{\mathrm{F}}(\mathcal{U}, \mathcal{W}) = \left\|\mathrm{proj}\left[\mathcal{W}^\perp\right]\mathrm{proj}\left[\mathcal{U}\right]\right\|_{\mathrm{F}}$ between subspaces of $L^2(\mu)$.

Approximation error ($\hat{C} = C$)

Natural to apply existing bounds for Rayleigh-Ritz method.

Let $\Phi = \text{span}_{i \leq n}\{\phi_i\}$ and assume $1 \in \Phi$. Then

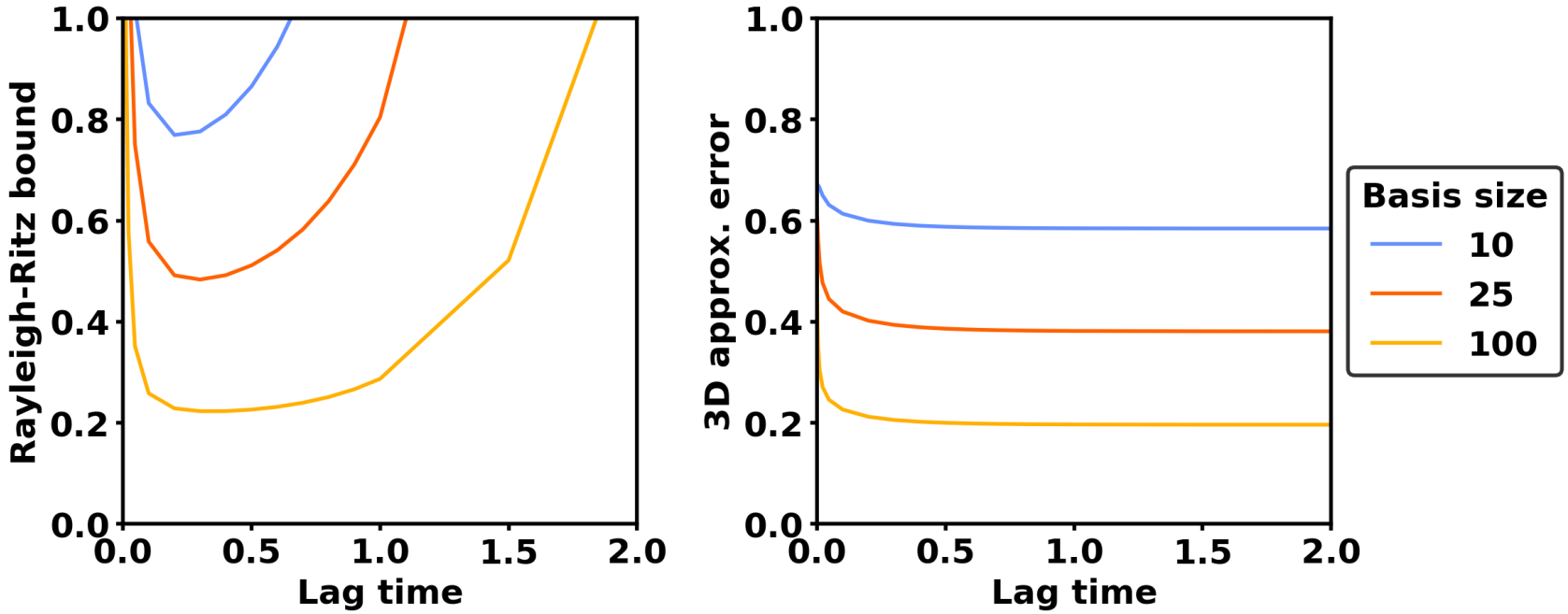
$$1 \leq \frac{dist_{\text{F}}^2\left(\text{span}_{i \leq k}\{\gamma_i^\tau\}, \text{span}_{i \leq k}\{\eta_i\}\right)}{dist_{\text{F}}^2\left(\text{span}_{i \leq k}\{\eta_i\}, \Phi\right)} \leq 1 + \frac{\|\text{proj}[\Phi^\perp]\mathcal{T}^\tau \text{proj}[\Phi]\|_2^2}{|e^{-\sigma_k \tau} - \lambda_{k+1}^\tau|^2}$$

provided that $e^{-\sigma_k \tau} > \lambda_{k+1}^\tau$. As long as $\sigma_k < \sigma_{k+1}$

$$\text{span}_{i \leq k}\{\gamma_i^\tau\} \to \text{span}_{i \leq k}\{\eta_i\} \quad \text{as} \quad \text{proj}_\Phi[\eta_i] \to \eta_i \text{ for } i \leq k$$

RR bounds used to prove λ_i^τ converge in [Djurdjevac, Sarich, and Schütte (2012)].

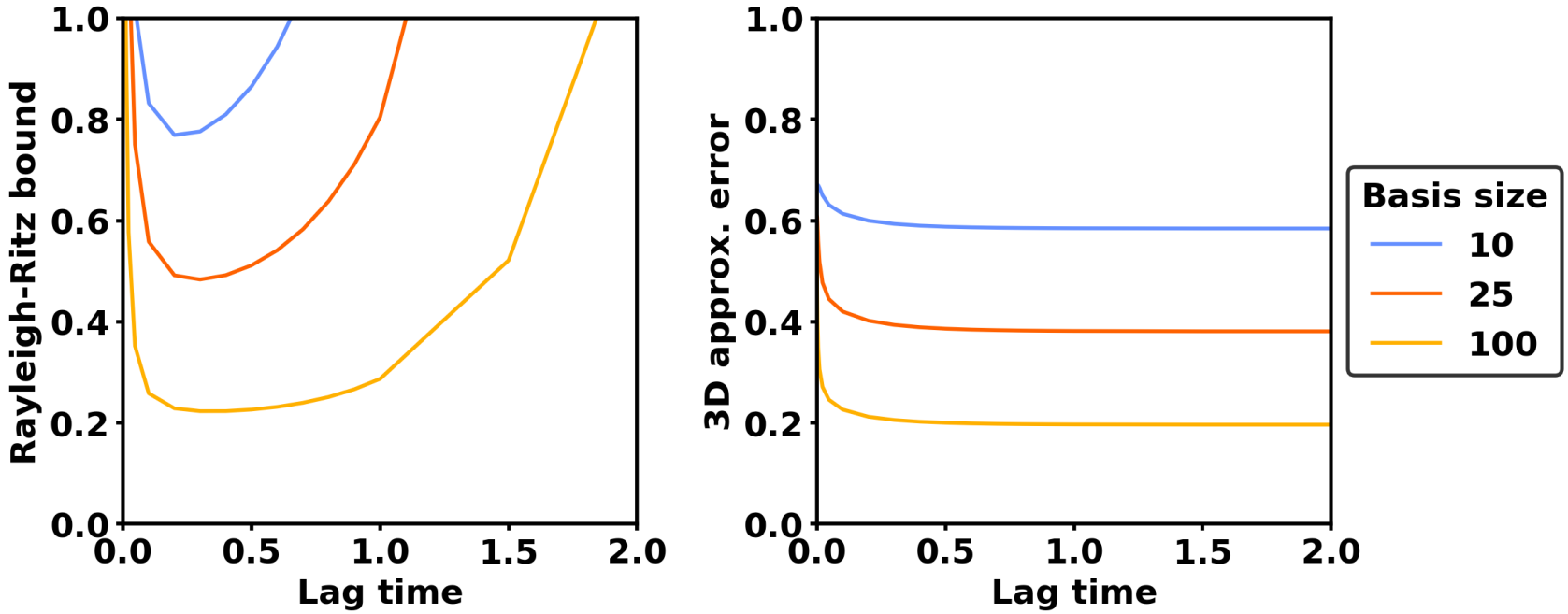
But what happens when we increase τ?



1*d* Ornstein-Uhlenbeck process with MSM (indicator) basis. Approximation error in $\mathrm{span}\{\eta_1, \eta_2, \eta_3\}$.

We need a more detailed approximation error bound.

But what happens when we increase τ?



1*d* Ornstein-Uhlenbeck process with MSM (indicator) basis. Approximation error in $\mathrm{span}\{\eta_1, \eta_2, \eta_3\}$.

We need a more detailed approximation error bound.

Approximation error dependence on τ

Going beyond the Rayleigh-Ritz bounds we prove:

Provided that σ_i is isolated,

$$\frac{\lambda_i^\tau}{e^{-\sigma_i \tau}} \to c_i \quad \text{as} \quad \tau \to \infty$$

where c_i is independent of τ.

As long as $\sigma_k < \sigma_{k+1}$

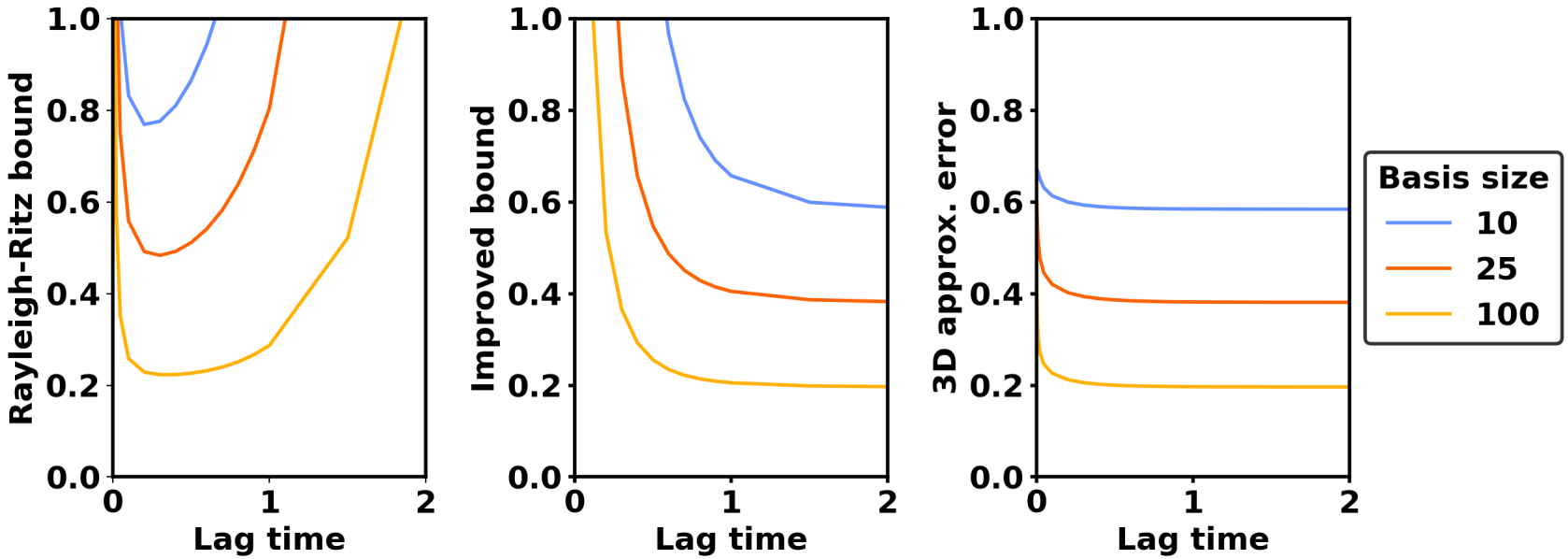
$$\text{span}_{i \le k}\{\gamma_i^\tau\} \to \text{proj}[\Phi]\text{span}_{i \le k}\{\eta_i\} \quad \text{as} \quad \tau \to \infty$$

and convergence is exponentially fast.

Provided that $e^{-\sigma_{k+1}\tau} < \lambda_k^\tau$

$$\frac{dist_{\text{F}}^2\left(\text{span}_{i \le k}\{\gamma_i^\tau\}, \text{span}_{i \le k}\{\eta_i\}\right)}{dist_{\text{F}}^2\left(\text{span}_{i \le k}\{\eta_i\}, \Phi\right)} \le 1 + \frac{1}{4}\left|\frac{e^{-\sigma_{k+1}\tau}}{\lambda_k^\tau - e^{-\sigma_{k+1}\tau}}\right|^2.$$

The new bound is sharp for large τ



1*d* Ornstein-Uhlenbeck process with MSM (indicator) basis. Approximation error in $\mathrm{span}\{\eta_1, \eta_2, \eta_3\}$.

Approximation error gets better (not worse) as τ increases.

A precise asymptotic formula for estimation error

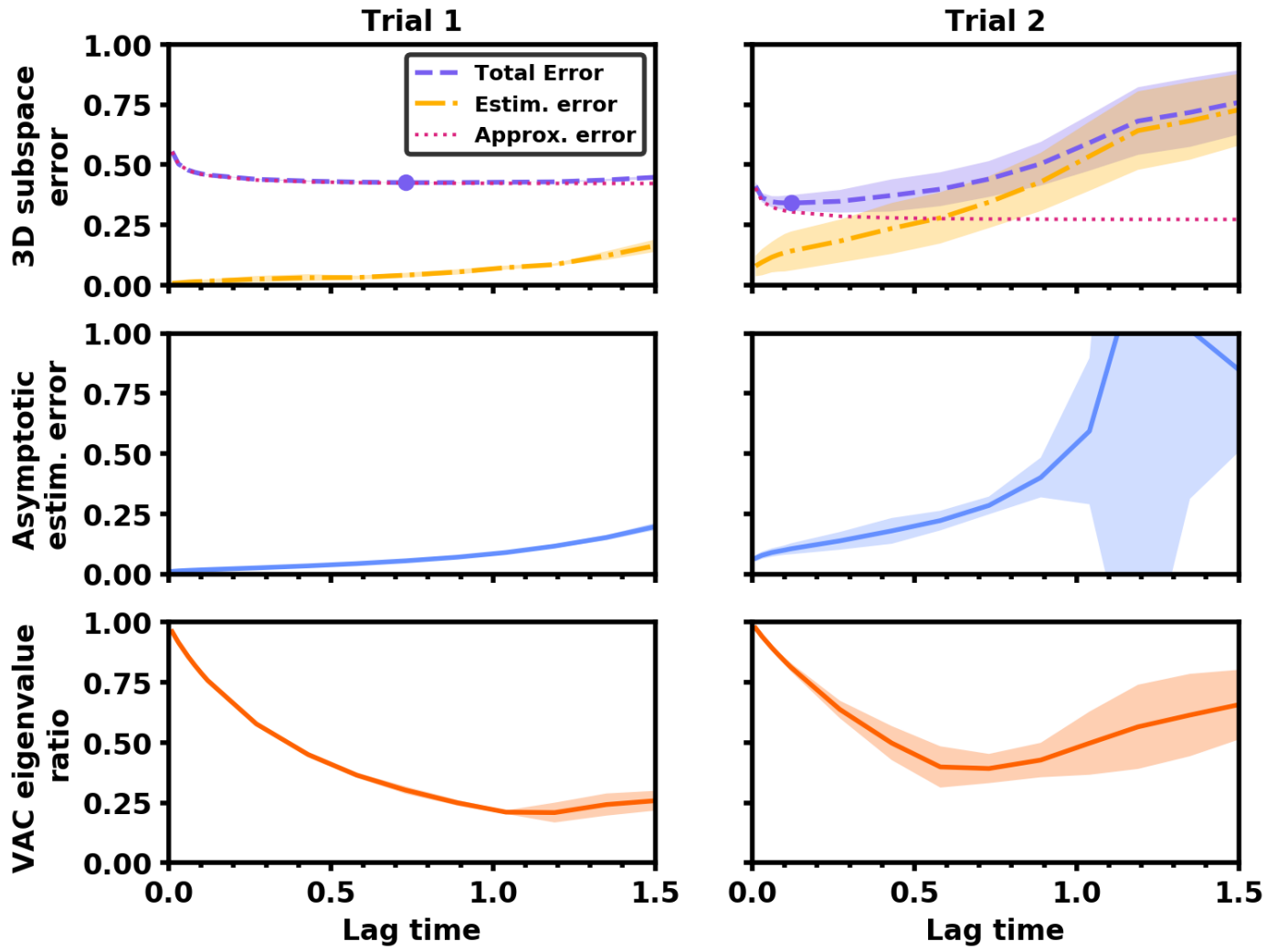
The estimation error can be expressed as

$$\begin{aligned} &\mathit{dist}_{\mathrm{F}}\left(\operatorname{span}_{i\leq k}\hat{\gamma}_i^{\tau}, \operatorname{span}_{i\leq k}\gamma_i^{\tau}\right)^2 \\ &\qquad = \sum_{i=k+1}^{n}\sum_{j=1}^{k}\left|\frac{v^i(\tau)^{\mathrm{T}}\left[\hat{C}(\tau)-\lambda_j^{\tau}\hat{C}(0)\right]v^j(\tau)}{\lambda_i^{\tau}-\lambda_j^{\tau}}\right|^2(1+o(1)) \end{aligned}$$

in the limit as $\hat{C}(\tau)\to C(\tau)$ and $\hat{C}(0)\to C(0)$.

Estimation error is small when $\hat{C}$ is close to C.

But expect estimation error to be big when $\lambda_k^{\tau}-\lambda_{k+1}^{\tau}$ is small.



1d Ornstein-Uhlenbeck process with MSM (indicator) basis. Approximation error in $\mathrm{span}\{\eta_1, \eta_2, \eta_3\}$.
Trial 1: $n = 20$, trajectory length = 10000.
Trial 2: $n = 50$, trajectory length = 500

As τ increases approximation error decreases, but estimation error increases.

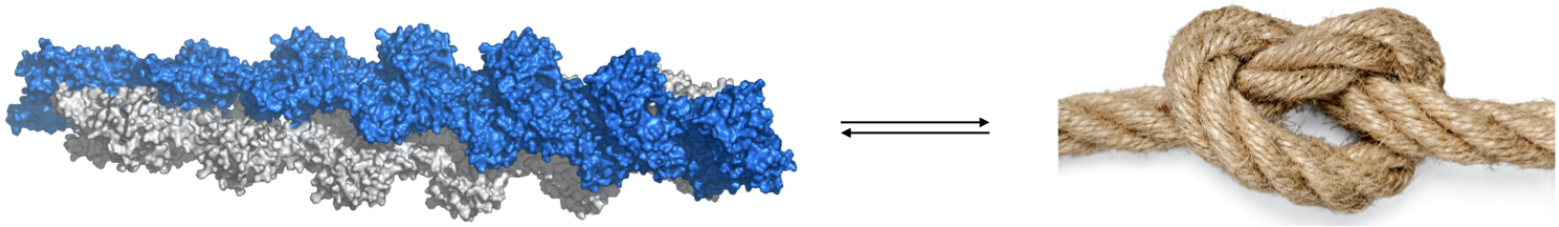
VAC summary

- New convergence bounds for VAC eigenfunctions.
- New understanding of the role of lag time.
- New diagnostic tools to help choose lag time.

Dynamic Galerkin Approximation (DGA)

Erik Thiede, John Strahan, Dimitrios Giannakis, Aaron Dinner

The truely longest timescale pocesses are often physically irrelevant



If we have a specific event in mind we should compute quantities specific to that event.

E.g. to predict the event that we reach $X_t \in B$ before $X_t \in A$ starting from $X_0 = x$ we should compute

$$q_+(x) = P(T_B < T_A \,|\, X_0 = x)$$

the "committor function." (T_A is the first time $X_t \in A$)

DGA computes conditional expectations

We'll want to incorporate a domain D:

$$\mathcal{T}^\tau f(x) = E[f(X_{\tau \wedge T}) \,|\, X_0 = x]$$

where T is the first time X_t exits D.

DGA estimates functions of the form:

$$u(x) = E\left[g(X_T) + \int_0^T h(X_s) ds \,\middle|\, X_0 = x \right]$$

E.g. for $u = q_+$ plug in $D = A \cup B$, $h = 0$, and $g(x) = \begin{cases} 1, & x \in \partial B \\ 0, & x \in \partial A \end{cases}$

Elaborations of the basic setup including e.g. a potential term and time dependence are straightforward (in principle)

DGA relies on the Feynman-Kac relation

$$u - \mathcal{T}^\tau u = \int_0^\tau \mathcal{T}^s[h\mathbf{1}_D]ds \text{ on } D \quad \text{and} \quad u = g \text{ on } \partial D$$

to avoid generating long trajectories.

If $u = \psi + w$ for a "guess" ψ satisfying the BCs and

$$r^\tau = \mathcal{T}^\tau \psi - \psi + \int_0^\tau \mathcal{T}^s[h\mathbf{1}_D]ds$$

then we can solve

$$w - \mathcal{T}^\tau w = r^\tau \text{ on } D \quad \text{and} \quad w = 0 \text{ on } \partial D$$

for w.

DGA workflow...

1. Generate samples of $X_0 \in D$ from μ and then $X_{\tau \wedge T}$ given X_0.
2. Choose a guess function ψ satisfying the BCs.
3. Choose a set of basis functions $\phi_1, \phi_2, \ldots, \phi_n$ satisfying homogenous BCs.
4. Use samples to build estimates

$$\hat{C}_{ij}(0) \approx C_{ij}(0) = \langle \phi_i, \phi_j \rangle_\mu = E[\phi_i(X_0)\phi_j(X_0)]$$
$$\hat{C}_{ij}(\tau) \approx C_{ij}(\tau) = \langle \phi_i, \mathcal{T}^\tau \phi_j \rangle_\mu = E[\phi_i(X_0)\phi_j(X_{\tau \wedge T})]$$
$$\hat{b}_i(\tau) \approx b_i(\tau) = \langle \phi_i, r^\tau \rangle_\mu$$
$$= E\left[\phi_i(X_0)\left(\psi(X_{\tau \wedge T}) - \psi(X_0) + \int_0^{\tau \wedge T} h(X_s)ds\right)\right]$$

5. Solve the linear system $-(\hat{C}(\tau) - \hat{C}(0))\hat{v} = \hat{b}$.
6. Return approximate conditional expectation $\hat{u}^\tau = \psi + \sum_j \hat{v}_j \phi_j$.

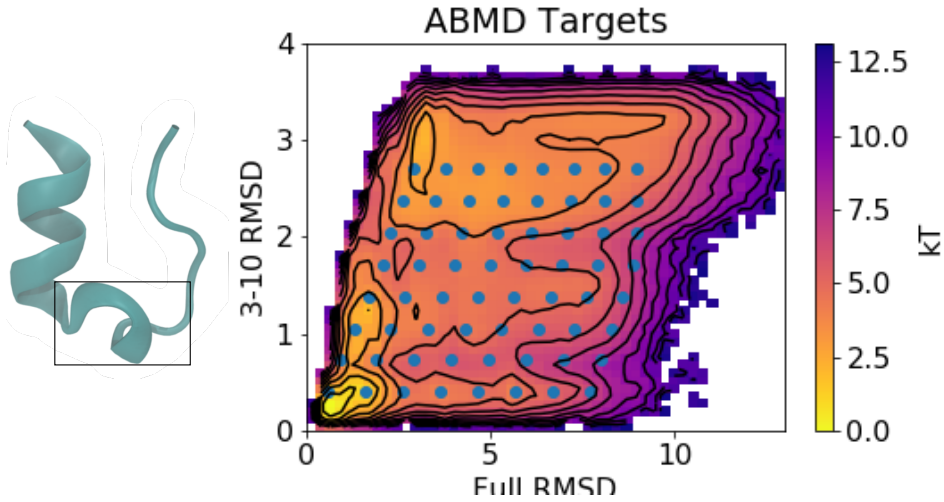
μ does not have to be the stationary measure.

Similar steps allow solution of equations involving the adjoint of $\mathcal{T}^\tau$.

Back to Trp cage folding

To evaluate DGA (and other methods) we produced a large database of samples of (X_0, X_τ).

We choose μ so that its marginal distribution in 2 variables is approximately uniform to make sure those variables are "well sampled."

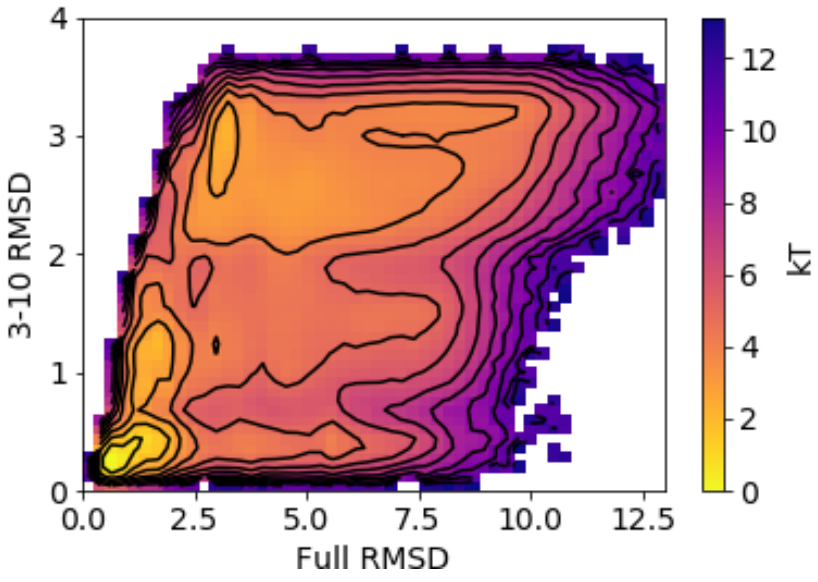


Our data set contains only short (17.5 nanosecond) trajectory fragments and zero folding events.

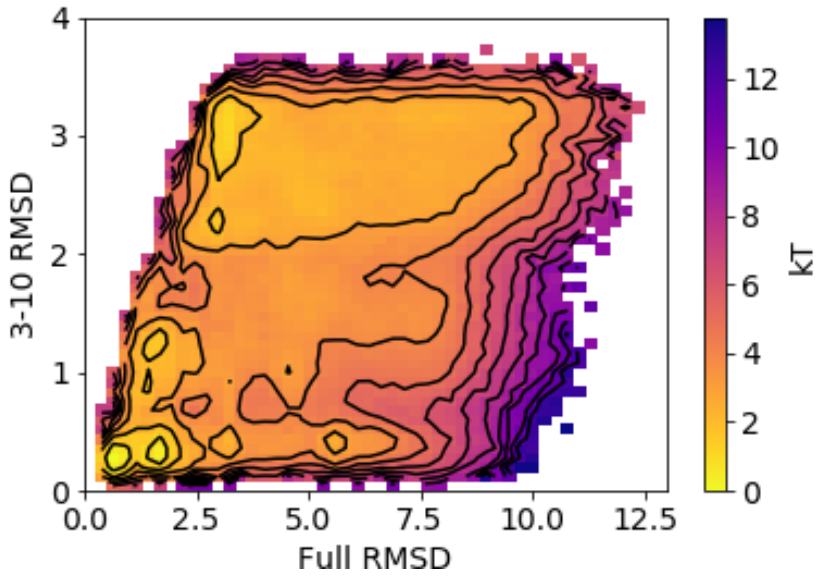
Sum of our trajectories is about 15.5 microseconds compared to about 208 for the Anton data set.

Validating the DGA stationary distribution

Left: DGA free energy



Right: REUS free energy



The change of measure ρ from the sampling measure μ to the stationary measure π is available by solving an equation involving the μ-adjoint of $\mathcal{T}^\tau$:

$$(\mathcal{T}^\tau)^\dagger_\mu \rho = \rho$$

Trp cage committor

$q_+(x) = P(T_{\text{folds}} < T_{\text{unfolds}} \,|\, X_0 = x)$

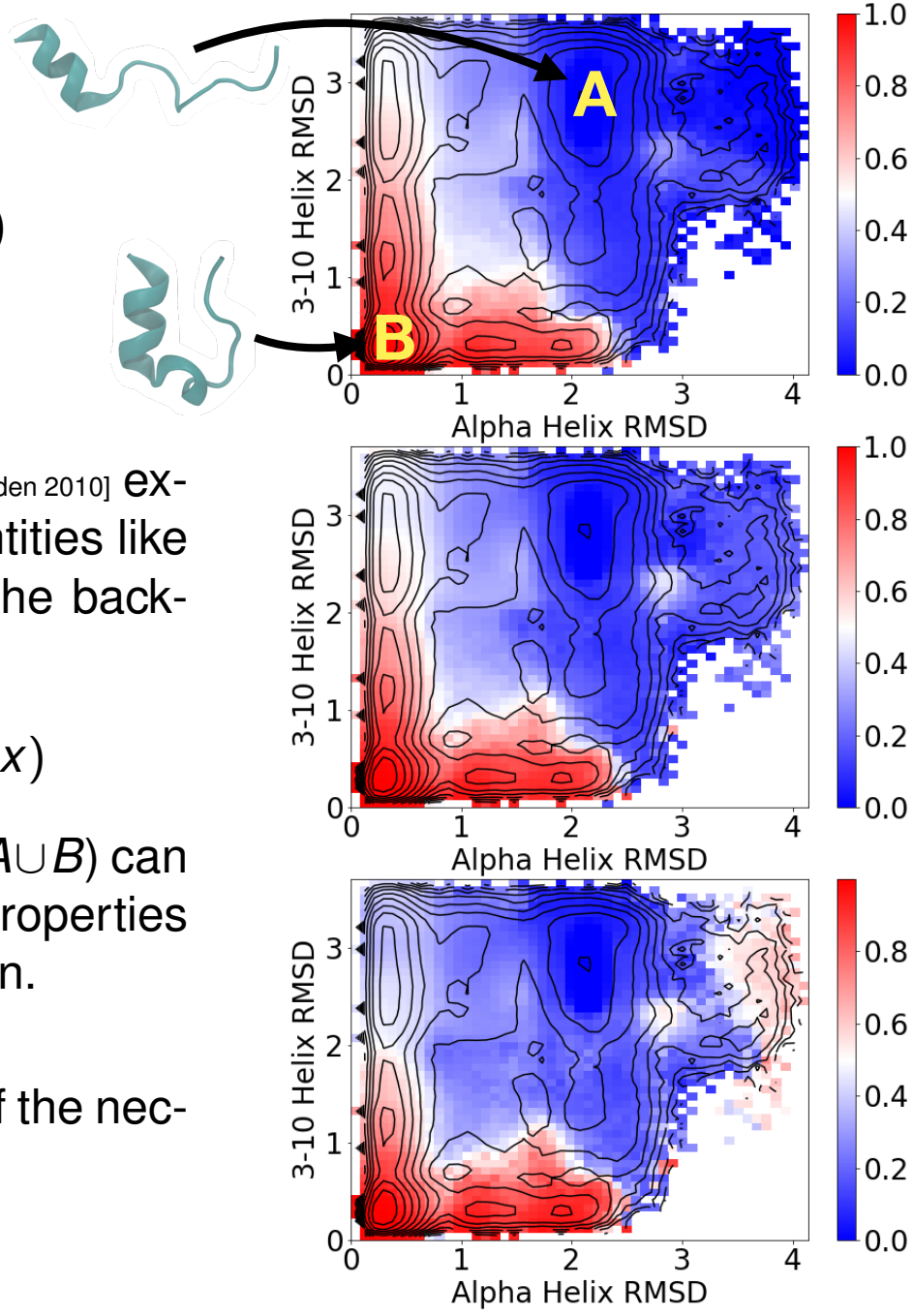
from DGA with different τ.

Transition Path Theory [E and Vanden-Eijnden 2010] explains how q_+ and additional quantities like the stationary distribution π and the backward committor

$$q_-(x) = P(X_{T^-} \in A \,|\, X_0 = x)$$

($T^- < 0$ is the last time X_t was in $A \cup B$) can be combined to characterize key properties of the steady state A to B transition.

DGA can be used to compute all of the necessary quantities

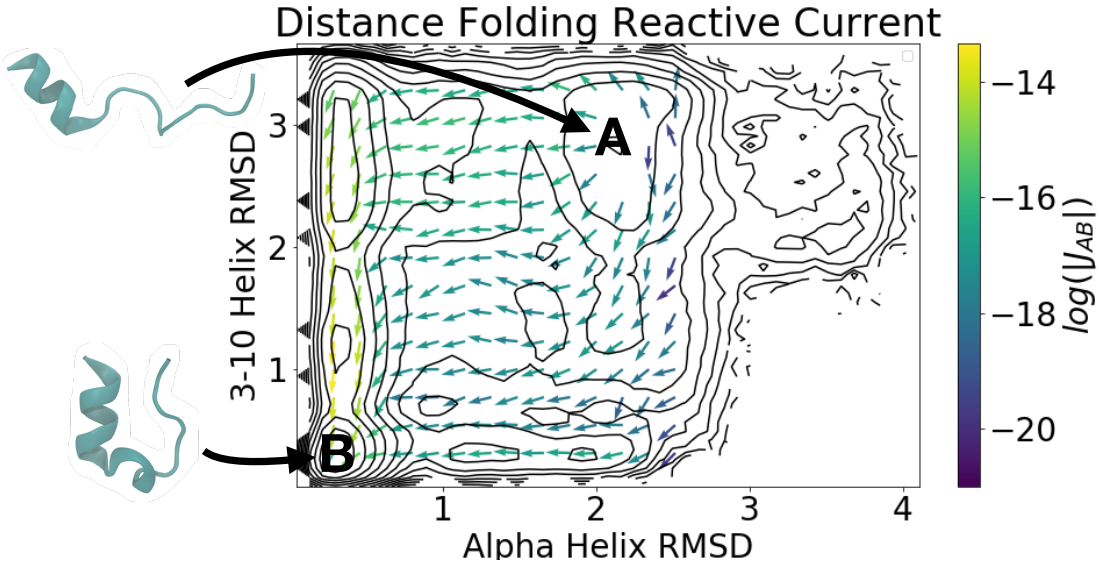


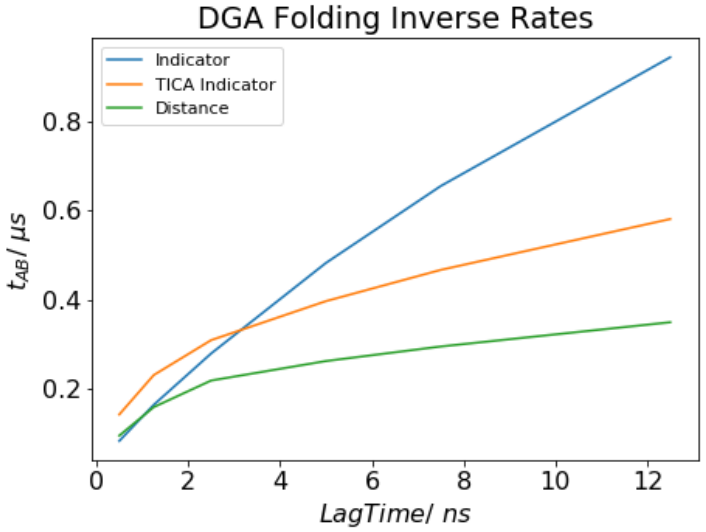
DGA+TPT

J_{AB} is a probability current of trajectories traversing from A to B.

$$R_{AB} = \int_C J_{AB} \cdot n_C d\sigma_C$$

is the number of transitions from A to B per unit time.



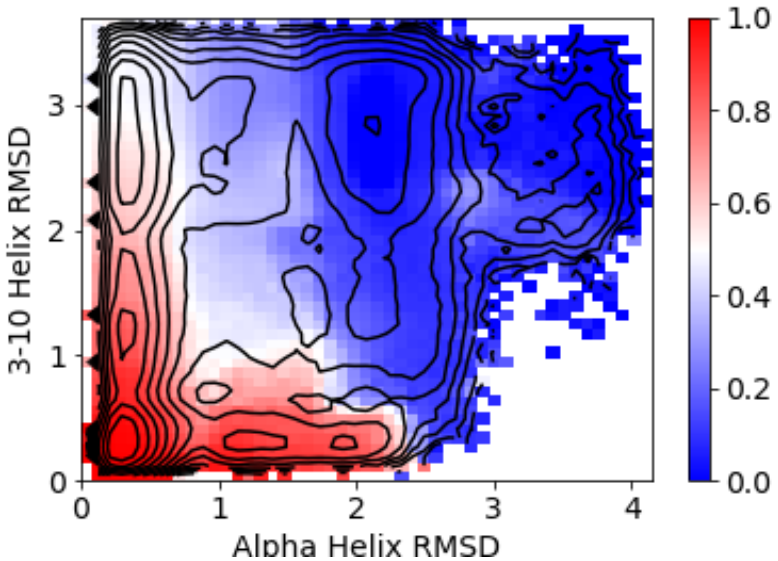


We see strong dependence on lag time and basis choice when we compute the forward rate from A to B

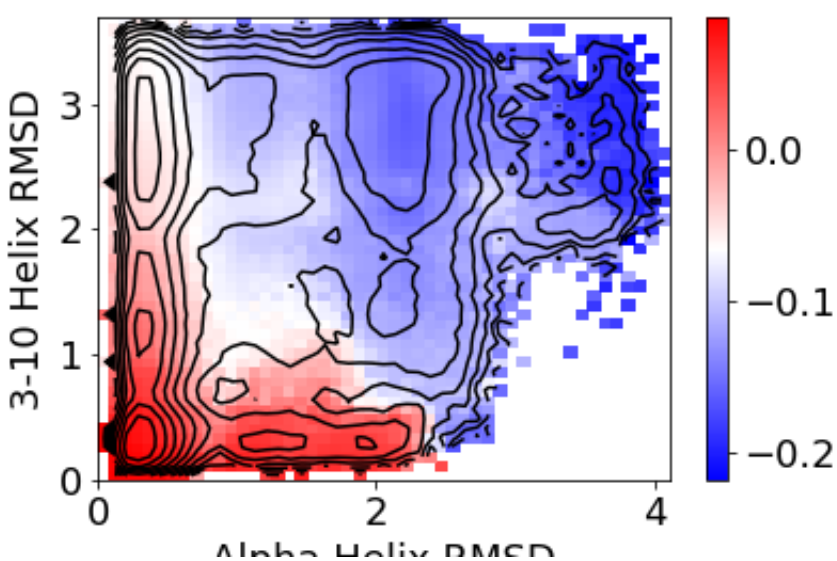
$$k_{AB} = \frac{R_{AB}}{\int q_-(x)\pi(dx)}$$

Full circle

Left: DGA committor



Right: Top TICA eigenvector



Summary and future directions

We've provided a much more complete understanding of the error properties of VAC and specifically how they depend on the lag time τ.

Repurposing the basic components of VAC we've introduced DGA, a family of estimators of conditional expectations specific to the event of interest.

We've produced a large data set of short molecular dynamics trajectories for the trp cage mini-protein to benchmark DGA performance.

Using our analysis of VAC as a roadmap we will study DGA's error properties.

We will continue development of DGA, e.g. by incorporating more flexible solution representation.