Notes on the Theory of Markov Models for Conformational Dynamics

Robert T. McGibbon
Stanford University

I. OVERVIEW

These notes were delivered as a seminar (chalk talk) at the Pande Group lab meeting on February 12, 2016. They introduce some aspects theory of reversible Markov chains in a continuous state space, with an orientation towards models of the classical conformational dynamics of molecular systems.

The topics discussed include the propagator and transfer operator, their associated eigenvalues and eigenfunctions and the physical interpretation thereof, the variational theorem for conformational dynamics, and the estimation of these eigenfunctions by time-structure independent components analysis (tICA) and Markov state models (MSMs).

II. THE SETTING

Take a discrete-time stochastic process, X_t, for $t = \{0, 1, 2, \ldots\}$, in some general (possibly Euclidean, but it doesn't really matter) space, $X_t \in \Omega$, like $\Omega = \mathbb{R}^{3N}$, where N is the number of atoms. The rule for the dynamics that we should have in mind is some Langevin / Smoluchoswki diffusion sampled at a finite interval/timestep, or thermostated Hamiltonian dynamics sampled at a finite interval/timestep, but we won't be too specific here.

But simply saying that our processes is *any* discrete-time stochastic process is too general to really make any progress, so we're going to enforce three constraints.

- The stochastic process is assumed to be Markov:

  This means, roughly, that knowing the entire past history doesn't help you any more than knowing where it was last timestep.

  $$\mathbb{P}(X_{t+1}|X_t, X_{t-1}, X_{t-2}, \ldots) = \mathbb{P}(X_{t+1}|X_t) \tag{1}$$

  The function $\mathbb{P}(X_{t+1}|X_t)$ is important, so we'll call this expression "transition density kernel" and define a little shorthand for it.

  $$p(x, y) \equiv \mathbb{P}(X_t = y|X_{t-1} = x) \tag{2}$$

  Note that we are here assuming that the function $p(x, y)$ doesn't change as a function of t. This is called time-homogeneity of the process.

- The process is ergodic. This means that there are no dynamically disconnected parts of phase space. As $t \to \infty$, every state will be visited infinitely often. And the total fraction of an infinitely long trajectory that the walker spends in any voxel dx is $\mu(x)dx$. This is called the stationary distribution, and it's the equilibrium distribution of some thermodynamic ensemble that the system is sampling (NpT, NVT, etc). For MD at constant temperature, it's the Boltzmann distribution.

- Furthermore, the assume that the stochastic processes is *reversible* with respect to a measure $\mu(x)$. The condition for reversibility

  $$\mu(x) \cdot p(x, y) = \mu(y) \cdot p(y, x) \tag{3}$$

  Reversibility is a sort of *symmetry* between the forward and backward directions of time. It's equivalent to the statement that

  $$\mathbb{P}(X_t = y \text{ and } X_{t-1} = x) = \mathbb{P}(X_t = y \text{ and } X_{t-1} = x) \tag{4}$$

  which you can interpret as saying that any path is equally likely to occur in either in the forward or backward direction.

These assumptions are rigorously valid for the processes we simulate with molecular dynamics. The Markov property really comes directly from Newton's equations of motion.

III. ENSEMBLES

The transition density kernel $p(x, \cdot)$ gives conditional distribution over where the process will advance to in the future, but it requires knowing exactly what the current position is, x. Generally, we're not going to know exactly where the particle is at any time. Even if were to *start* by knowing exactly where the particle is, after one step it would have some spread.

So, if we have some belief of where the particle is at time t, and we know the rule for the dynamics $p(x, y)$, what should our belief be about where the particle will be one step forward in the future? How do we update (propagate) our belief?

Or, for an equivalent view, we might have an ensemble of indistinguishable copies of the system, and want to propagate the ensemble forward in time. Describe our current belief (or ensemble) using a probability distribution

$$p_t(x) \tag{5}$$

This probability distribution is zero or positive everywhere on Ω, and always needs to be normalized to 1.

$$\int_\Omega dx\ p_t(x) = 1. \tag{6}$$

Great. Okay, at time $t+1$, our new distribution will be

$$p_{t+1}(y) = \int_\Omega dx\ p_t(x) \cdot p(x,y) \tag{7}$$

This should make sense. We're taking our distribution over space, and for each little element of probability mass, propagating it forward in time according to the transition density kernel.

What's going on here? We're taking a function over Ω, $p_t(x)$, and performing some operation on it that returns a new function over Ω, $p_{t+1}(x)$.

An object that transforms one function into another function is called an operator. So we can say, $p_t(x)$ is being "operated on". Write this in a **shorthand** notation as

$$p_{t+1} = \mathcal{Q} \circ p_t \tag{8}$$

IV. PROPAGATOR

We're going to call $\mathcal{Q}$ the "propagator". What **properties** does the propagator have? Can we use any of these properties to make a **simple model** of the propagator?

First, remember that $p(x,y)$ had a kind of symmetry with $p(y,x)$. It's not exactly symmetry, but reversibility is pretty close.

So $\mathcal{Q}$ should have a kind of symmetry too. How to write this for operators? This is called being **self-adjoint**. The general property we want is that for any two functions f and g that exist in this space,

$$\langle f | \mathcal{Q} \circ g \rangle = \langle \mathcal{Q} \circ f | g \rangle \tag{9}$$

For some *inner product*, $\langle \cdot | \cdot \rangle$. Maybe it's not clear that this equation actually means that $\mathcal{Q}$ is symmetric, so let's show that.

First, note the following conceptual mapping between operators and Hilbert spaces and regular linear algebra matrices and vector spaces. Basically, the following concepts are equivalent to one another.

- Matrix → Operator
- Vector → Function
- Dot product → Inner product
- Matrix vector product → Operator application.

With a matrix, you might verify that it is symmetric by looking at the i, j and j, i elements. But for operators, the only way we're allowed to interact with the operator is by applying it to a function. So how would you verify if a matrix is symmetric if you couldn't actually check the elements, and all you can do is matrix-vector product (operator applications), and dot products (inner products).

$$\vec{v}^T A \vec{w} = \sum_{ij} \vec{v}_i A_{ij} \vec{w}_j \tag{10}$$

$$(A\vec{v})^T \vec{w} = \sum_{ij} \vec{v}_i A_{ji} \vec{w}_j \tag{11}$$

If these two things are going to be equal, B_{ij} has got to be equal to A_{ij}. This is the same idea for operators in how the definition of self-adjoint works. The stuff in the transpose is the bra, and the stuff on the right is the ket.

Okay, so back to the propagator. I'm going to show the answer first, which is that in order for $\mathcal{Q}$ to be self-ajoint, we need to use a special weighted inner product.

The weighted inner product we want when working with $\mathcal{Q}$ is $\langle f | g \rangle_{\mu^{-1}}$. This inner product is defined as

$$\langle f | g \rangle_{\mu^{-1}} = \int_\Omega dx\ \frac{f(x) \cdot g(x)}{\mu(x)} \tag{12}$$

Now, lets verify the "self-adjointness" of $\mathcal{Q}$ according to this inner product. First take the left hand side of the equation defining "self-adjointness":

$$\langle f | \mathcal{Q} \circ g \rangle_{\mu^{-1}} = \left\langle f \middle| \int_\Omega dx\ g(x) p(x,y) \right\rangle_{\mu^{-1}} \tag{13}$$

$$= \int_\Omega dy\ \frac{f(y) \cdot \left(\int_\Omega dx\ g(x) p(x,y) \right)}{\mu(y)} \tag{14}$$

$$= \int_{\Omega \times \Omega} dx dy\ \frac{f(y) g(x) p(x,y)}{\mu(y)} \tag{15}$$

Similarly, for the other side of the equarion, we have

$$\langle \mathcal{Q} \circ f | g \rangle_{\mu^{-1}} = \left\langle \int_\Omega dy\ f(y) p(y,x) \middle| g \right\rangle_{\mu^{-1}} \tag{16}$$

$$= \int_\Omega dx \frac{\left(\int_\Omega dy\ f(y) p(y,x) \right) g(x)}{\mu(x)} \tag{17}$$

$$= \int_{\Omega \times \Omega} dx dy\ \frac{f(y) g(x) p(y,x)}{\mu(x)} \tag{18}$$

All the terms match up here as long as we can do a necessary flip from $p(x,y)$ to $p(y,x)$.

$$\frac{p(x,y)}{\mu(y)} = \frac{p(y,x)}{\mu(x)} \implies \mu(x) p(x,y) = \mu(y) p(y,x) \tag{19}$$

which is exactly our detailed balance condition.

Why is this important? Well, self-adjoint operator has real eigenvalues, real eigenvectors, a complete orthonormal basis, and all of this other stuff that's going to be really important for any of the theory to work.

V. TRANSFER OPERATOR

There's a trick used in the definition of MSMs and tICA, which is define another operator called the transfer operator in such a way that it is self-adjoint with a different norm.

It's a little bit clever, but once you see how it works you'll get the hang of it. Instead of always considering the "propagation" of p_t to p_{t+1}, define a new quantity called u_t.

$$u_t(x) = \frac{p_t(x)}{\mu(x)} \tag{20}$$

We've been assuming that you know $\mu(x)$, so at any point if you have an ensemble p_t you can always just rescale it then you have u_t, so they're pretty much equivalent.

So, if your current (scaled) belief about the particle's location is u_t, how does this get updated by one step of dynamics. Well, you can basically round-trip to it to p_t, propagate $p_t \to p_{t+1}$, and then rescale it to back out u_t.

$$u_{t+1}(y) = [\mathcal{T} \circ u_t](y) = \frac{1}{\mu(y)} \int_\Omega dx \ \left(u_t(x)\mu(x)\right) p(x,y) \tag{21}$$

And that's the definition of the transfer operator, $\mathcal{T}$. It's the operator that evolves the function u_t forward to u_{t+1}. Why is this helpful? Well, $\mathcal{T}$ is self-adjoint in an easier way, without a special norm

$$\langle \mathcal{T} \circ f | g \rangle = \langle f, \mathcal{T} \circ g \rangle \tag{22}$$

Let's go through the reason. For the left hand side,

$$\langle \mathcal{T} \circ f | g \rangle = \left\langle \left(\frac{1}{\mu(y)} \int_\Omega dx \ f(x)\mu(x)p(x,y) \right) \middle| g \right\rangle \tag{23}$$

$$= \int_{\Omega \times \Omega} dxdy \ \frac{f(x)g(y)p(x,y)}{\mu(y)} \tag{24}$$

And for the right hand side,

$$\langle f | \mathcal{T} \circ g \rangle = \left\langle f \middle| \left(\frac{1}{\mu(x)} \int_\Omega dy \ g(y)\mu(y)p(y,x) \right) \right\rangle \tag{25}$$

$$= \int_{\Omega \times \Omega} dxdy \ \frac{f(x)g(y)p(y,x)}{\mu(x)} \tag{26}$$

Again, we should see the same trick where reversibility of the transition kernel makes these equal.

VI. EIGENFUNCTIONS

Okay, we've got these two basically-equivalent operators. And we've seen one property they share. They basically describe everything about the dynamics. How can you build a numerical model of these guys? They're infinite dimensional, so we're going to have to cut some corners. How should they be approximated?

First key property that is implied by self-adjoint is that they have a complete basis of orthonormal eigenfunctions. Let's unpack that statement.

First, for the propagator. An eigenfunction is defined a function that satisfied the following equation:

$$\mathcal{Q} \circ \phi_i = \lambda_i \phi_i \tag{27}$$

This is a pretty excellent property, because in general computing $\mathcal{Q} \circ f$ is hard – you have to integrate over all space and such. But this means that if your f happens to be one of the eigenfunctions, all that $\mathcal{Q} \circ f$ does is just **rescale** it.

Also, there are a countably infinite number of eigenfunctions. And, they are orthogonal and normalized. That means (remember the inner product) that for all integers i, j.

$$\langle \phi_i | \phi_j \rangle_{\mu^{-1}} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{28}$$

The proof of this is pretty easy, and it uses the self-adjointness.

$$\lambda_j \langle \phi_i | \phi_j \rangle_{\mu^{-1}} = \langle \phi_i | \mathcal{Q} \circ \phi_j \rangle_{\mu^{-1}} \tag{29}$$

$$= \langle \mathcal{Q} \circ \phi_i | \phi_j \rangle_{\mu^{-1}} \tag{30}$$

$$= \lambda_i \langle \phi_i | \phi_j \rangle_{\mu^{-1}} \tag{31}$$

So $(\lambda_i - \lambda_j)\langle \phi_i | \phi_j \rangle_{\mu^{-1}} = 0$.

This means that basically the eigenfunctions are **perpendicular**. We should be able to guess one of these eigenfunctions. It's μ. Why? Because if you're current system is at equilibrium, if you propagate it forward by one step, it should still be at equilibrium. That means the eigenvalue should be 1. How do we check this mathematically?

$$[\mathcal{Q} \circ \mu](y) = \int_\Omega dx \ \mu(x)p(x,y) \tag{32}$$

Use the detailed balance trick to change this to

$$[\mathcal{Q} \circ \mu](y) = \int_\Omega dx \ \mu(y)p(y,x) \tag{33}$$

And then we can pull out $\mu(y)$ since we're integrating over x, and get

$$[\mathcal{Q} \circ \mu](y) = \mu(y) \int_\Omega dx; p(y,x) \tag{34}$$

$$= 1 \cdot \mu(y) \tag{35}$$

So μ is an eigenvector with associated eigenvalue 1, as a consequence of detailed balance, and the fact that $p(y, \cdot)$ is a proper probability distribution. These properties imply more stuff too. The Perron-Frobenius theorem implies that all of the eigenvectors other than the first need to be between -1 and 1 too.

For the transfer operator, we have something pretty similar. Let's derive it:

$$\mathcal{Q}\phi_i = \lambda_i \phi_i = \int_\Omega dx\ \phi_i(x) p(x, y) \tag{36}$$

Now, lets define a new set of functions $\psi_i(x) \equiv \phi_i(x)/\mu(x)$, so that the equation above can be written as

$$\lambda_i \psi_i(y)\mu(y) = \int_\Omega dx\ \psi_i(x)\mu(x)p(x, y) \tag{37}$$

If you move the factor of $\mu(y)$ to the denominator on the r.h.s, then this is the definition of the transfer operator, so we have

$$\lambda_i \psi_i = \mathcal{T} \circ \psi_i \tag{38}$$

This means that the ψ are eigenfunctions of $\mathcal{T}$, and that it has the same eigenvalues. The normalization condition for ψ_i is that for any i, j,

$$\langle \psi_i | \psi_j \rangle_\mu = \delta_{ij} \tag{39}$$

Note that here the measure contains a μ to the positive 1, not negative one power. Prove this to yourself by rewriting the normalization condition for the ϕ_is that we proved before in terms of ψ_is.

VII. SPECTRAL DECOMPOSITION

Let's use the eigenfunctions to write out a more explicit formula for $\mathcal{Q} \circ p$. For any initial distribution p, we write it in the **eigenvector basis** as some

$$p(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \tag{40}$$

The scalar values as are expansion coefficients – we've decomposed our function into components along each of the eigenfunctions, like writing a vector with an x components, y component, etc. How would we calculate these expansion coefficients? The orthonormality of the eigenfunctions makes this easy.

$$\langle p | \phi_j \rangle_{\mu^{-1}} = \left\langle \sum_{i=1}^{\infty} a_i \phi_i(x) \middle| \phi_j \right\rangle_{\mu^{-1}} = a_j \tag{41}$$

Now, we want

$$p_{t+1} = [\mathcal{Q} \circ p](x) = \mathcal{Q} \circ \left(\sum_{i=1}^{\infty} a_i \phi_i(x) \right) \tag{42}$$

$$= \sum_{i=1}^{\infty} a_i \left(\mathcal{Q} \circ \phi_i(x) \right) \tag{43}$$

$$= \sum_{i=1}^{\infty} a_i \lambda_i \phi_i(x) \tag{44}$$

$$p_{t+1}(x) = [\mathcal{Q} \circ p](x) = \sum_{i=1}^{\infty} \lambda_i \langle p | \phi \rangle_{\mu^{-1}} \phi(x) \tag{45}$$

This is important. You should understand each of the terms in this equation, and concept of a projection onto the eigenbasis. Question: what is the corresponding equation for $\mathcal{Q} \circ \mathcal{Q} \circ p$? How would you do the spectral decomposition of $\mathcal{T}$?

VIII. TIMESCALES

Let's say we start from an almost-pure state, $p_0(x) = \mu(x) + \phi_j(x)$. What would p_t look like for large t later?

We apply the propagator t times to p_0. The operator is linear, so we consider it acting on each term individually. Because μ is an eigenfunction with unit eigenvalue, it sails through application of the propagator unscathed. But in each step, ϕ_j gets multiplied by λ_j when we apply $\mathcal{Q}$, so we get

$$p_t(x) = \mu(x) + \lambda_j^t \phi_j(x) \tag{46}$$

Remember λ is between -1 and 1, and so this equation is telling us how quickly the distribution gets *damped* to equilibrium. It's always going to happen, since $\lambda_i^t \to 0$ as $t \to \infty$, but the bigger eigenvalues will last longer. If the molecular system as m metastable states, the propagator is going to have $m-1$ eigenvalues that are very very close to 1, so these will be the only ones that don't get damped too much for moderate to long t, while the rest that are closer to zero will get driven to zero more quickly by these powers of t.

If you focus your attention on positive eigenvalues, define the time constant, τ_i so that

$$\lambda_i = e^{-1/\tau_i} \tag{47}$$

Then, we can write this expression for $p_t(x)$ so that we see the concept of exponential decay with a certain time constant more clearly.

$$p_t(x) = \mu(x) + e^{-t/\tau_i} \phi_j(x) \tag{48}$$

So each of these eigenfunctions is like a functional perturbation away from equilibrium, and they all decay their way to zero, but at different exponential time constants. In this simple example our initial distribution was only "perturbed" from equilibrium along 1 direction, that's why we only had one term.

IX. EXPERIMENTS

Consider an idealized pump-probe experiment that does the following.

- First, preparate an ensemble in some non-equilibrium initial distribution, $p_0(x)$.
- Start the "clock", and measure, as a function of time, the evolution of some property of the system as it relaxes back towards equilibrium.

Call the experimental observable $f(x) : \Omega \to \mathbb{R}$. It basically reports on the current state of the system. This might be a variable like the distance between two atoms in the system or the IR frequencies or something.

We can write down an expression for the measured signal f_t, explicitly in terms of the eigenfunctions of $\mathcal{Q}$ and the associated timescales.

$$\begin{aligned} f_t &= \langle f \,|\, p_t \rangle && (49)\\ &= \langle f | \mathcal{Q}^t \circ p_0 \rangle && (50)\\ &= \left\langle f \middle| \sum_{i=1}^{\infty} e^{-t/\tau_i} \langle p_0 | \phi_i \rangle_{\mu^{-1}} \phi_i \right\rangle && (51)\\ &= \sum_{i=1}^{\infty} e^{-t/\tau_i} \langle p_0 | \phi_i \rangle_{\mu^{-1}} \langle f | \phi_i \rangle && (52) \end{aligned}$$

This is kind of an important result, because of the specificity of the functional form and the fact that the setup is so general. The only time-dependent term here is the exponential. So this is called multiexponential kinetics. All pump-probe experiments like this should have this signature. Each eigenfunction contribute a term to the observed signal, with an amplitude that contains terms which measure

- How much of the initial distribution is along this eigenfunction?
- How strongly does this eigenfunction couple to the observable?

X. CORRELATION FUNCTIONS

Let's say we have some arbitrary scalar observable $f(X_t) : \Omega \to \mathbb{R}$, and we want to know how rapidly this measurement changes over time. Assume that f has been shifted and scaled so that it has mean zero and variance 1.

The 1-step autocorrelation of any zero-mean unit-variance scalar timeseries is defined as

$$\text{acf}(f) = \mathbb{E}\left[f(x_t) \cdot f(x_{t+1})\right] \tag{53}$$

This average over time can be related to an average over space (ergodic theorem), so we can also calculate this quantity from the propagator / transfer operator.

$$\mathbb{E}\left[f_t \cdot f_{t+1}\right] = \int_{\Omega\times\Omega} dxdy\; \mu(x)\; p(x,y) f(x) f(y) \tag{54}$$

Because this is an expectation value we take over pairs of structures separated by 1 timestep. So the first structure, X_t, is distributed according to the equilibrium distribution – thats $\mu(x)$ – and $p(x,y)$ is the conditional distribution of the structure at 1 step ahead, given the one at time t.

Now, convince yourself that this expectation value can be written in terms of operators and inner products as

$$\mathbb{E}\left[f_t \cdot f_{t+1}\right] = \langle f | \mathcal{T} \circ f \rangle_\mu \tag{55}$$

When you check this, make sure we're not missing any factors of μ. They're very easy to lose.

Once you're satisfied with that, expand f in the ψ basis, and use this basis expansion to express the autocorrelation function

$$f = \sum_{i=1}^{\infty} a_i \psi_i \tag{56}$$

You can rewrite the autocorrelation function as

$$\begin{aligned} \mathbb{E}\left[f_t \cdot f_{t+1}\right] &= \langle f | \mathcal{T} | f \rangle_\mu && (57)\\ &= \left\langle \sum_{i=2}^{\infty} a_i \psi_i \middle| \mathcal{T} \middle| \sum_{i=2}^{\infty} a_i \psi_i \right\rangle_\mu && (58)\\ &= \sum_{i=1} \lambda_i a_i^2 && (59)\\ & && (60) \end{aligned}$$

If f is equal to some particular ψ_j (i.e. $a_j = 1$ and all other $a_i = 0$), then this implies that the autocorrelation function of f is just equal to its corresponding eigenvalue.

XI. VARIATIONAL THEOREM

Because we asserted before that the signal has unit variance, we have the following normalization condition of the function f. Basically, because the eigenfunctions are normalized, if the signal has variance 1 then the expansion coefficients a need to be scaled properly to make this true. The property is

$$\begin{aligned} 1 &= \text{Var}\left[f(x_t)\right] && (61)\\ &= \int_\Omega \mu(x) f(x)^2 = \langle f | f \rangle_\mu && (62)\\ &= \left\langle \sum_{i=1}^{\infty} a_i \psi_i \middle| \sum_{j=1}^{\infty} a_j \psi_j \right\rangle_\mu && (63)\\ &= \sum_{i=2}^{\infty} a_i^2 && (64) \end{aligned}$$

Do you see why the last step is true? It requires the linearity of the inner products to pull out the summations and then the fact that all of the terms with $i \neq j$ are zero because of the orthonormality of the eigenfunctions.

Recall also that we assumed f had zero mean. This means that a_1 must be zero, because the first transfer operator eigenfunction is $\psi_1(x) \equiv \phi_1(x)/\mu(x) = \mu(x)/\mu(x) = 1$. So for this function's expansion, $a_1 = 1$ so we can just start the sum at 2.

Also, we know all the eigenvalues $\lambda_2, \lambda_3, \ldots$ can be no larger than λ_2, since they're ordered and we put them in decreasing order (maybe we didn't specify that before, but we do now). So our expression for the correlation function from before is bounded.

$$\begin{aligned}\mathbb{E}\left[f_t \cdot f_{t+1}\right] &= \sum_{i=2} \lambda_i a_i^2 \\ &\leq \sum_{i=2} \lambda_2 a_i^2 \\ &\leq \lambda_2 \left(\sum_{i=2} a_i^2\right) \\ &\leq \lambda_2\end{aligned}$$

This is a pretty important equation. It says that the autocorrelation function of any possible observable is bounded by the eigenvalue. And the function with the largest autocorrelation constant *is* the leading eigenfunction.

TICA is just the obvious algorithm to exploit this. You can try to approximate the eigenfunction by varying a linear *ansatz* function to try to maximize this expression.

$$\psi_2 = \underset{f}{\text{argmax}}\ \mathbb{E}\left[f_t \cdot f_{t+1}\right] \tag{65}$$

$$\tag{66}$$

$$\text{such that} \tag{67}$$

$$\mathbb{E}[f(x_t)] = 0, \tag{68}$$

$$\text{Var}[f(x_t)] = 1 \tag{69}$$

The higher you get this objective, the closer you are to ψ_2. This is basically the same idea as in quantum chemistry where you vary the expansion coefficients of an ansatz wavefunction to minimize the energy. The difference is just that our variational theorem bounds the autocorrelation from above, whereas the one in quantum chemistry bounds the energy from below.

XII. MARKOV STATE MODELS

A Markov state model is a *model* for the full dynamics where we partition the space into a set of finite states the jump process of the observed trajectory projected onto these discrete states. Although molecular dynamics in full continuous phase space Ω is Markovian by construction, the dynamics of any projection of X_t is generally not Markovian.

Assume you have some states S_i that are non-overlapping and partition the phase space. Associated with each state, S_i define a function (alert: this is a clever choice with a special weighting that is going to turn out to be very convenient...)

$$\chi_i = \frac{1}{\sqrt{\pi_i}} \mathbf{1}_{S_i}(x) = \begin{cases} \frac{1}{\sqrt{\pi_i}} & x \in S_i \\ 0 & x \notin S_i \end{cases} \tag{70}$$

where π_i is the stationary weight of S_i,

$$\pi_i = \int_{x \in S_i} dx\ \mu(x) \tag{71}$$

Now, just for kicks, lets see if we can come up with some simple or physically meaningful expression for $\langle \chi_j | \mathcal{T} \circ \chi_i \rangle_\mu$ that we could actually estimate from many short MD trajectories.

Why did I pick that expression to try to calculate? Well, Eq. (55) implies that these are the type of matrix elements that are related to the correlation function that we might want to maximize.

$$C_{ij} = \langle \chi_j | \mathcal{T} \circ \chi_i \rangle_\mu \tag{72}$$

$$= \int_{\Omega \times \Omega} dxdy\ \chi_j(y) \mu(x) p(x,y) \chi_i(x) \tag{73}$$

$$= \int_{\{x \in S_i \times y \in S_j\}} dxdy\ \frac{1}{\sqrt{\pi_i \pi_j}} \mu(x) p(x,y) \tag{74}$$

$$= \frac{1}{\sqrt{\pi_i \pi_j}} \mathbb{P}\left[X_{t+1} \in S_j \text{ and } X_t \in S_i\right] \tag{75}$$

$$= \sqrt{\frac{\pi_i}{\pi_j}} \cdot \mathbb{P}(X_{t+1} \in S_j | X_t \in S_i) \tag{76}$$

$$= \sqrt{\frac{\pi_i}{\pi_j}} \cdot T_{ij} \tag{77}$$

We can re-write this as a matrix expression for the whole matrix C with two diagonal matrices for the factors of π.

$$C = D(\pi^{1/2})\ T\ D(\pi^{-1/2}) \tag{78}$$

We want to do a variational optimization of some arbitrary function $f(x) = \sum_i^N a_i \chi_i(x)$. When we maximize the correlation function of f with respect to the expansion coefficient's we'll find some "optimal" coefficients a^*, and this will yield an estimator for the the transfer operator's ψ_2:

$$\psi_2 \approx \tilde{\psi}_2 = \sum_{i=1}^N a_i^* \chi_i \tag{79}$$

We're still going to need that f have unit variance.

$$1 = \mathbb{E}(f^2) = \langle f | f \rangle_\mu = \left\langle \sum_{i=1}^N a_i \chi_i \middle| \sum_{j=1}^N a_j \chi_j \right\rangle_\mu = \sum a_i^2 \tag{80}$$

The χ aren't eigenfunctions or anything special, but the reason this comes out so simply is that in any place where χ_i is nonzero, χ_j *is* zero because they're defined based on non-overlapping states. And for $i = j$, $\langle \chi_i | \chi_i \rangle_\mu = \int_{S_i} dx\ \mu(x) \pi_i^{-1} = 1$. So that means that a needs to be unit norm.

Okay, now compute the correlation function of f.

$$\mathbb{E}\left[f_t \cdot f_{t+1}\right] = \langle f | \mathcal{T} \circ f \rangle_\mu \tag{81}$$

$$= \left\langle \sum_i^N a_i \chi_i \middle| \mathcal{T} \circ \sum_i^N a_i \chi_i \right\rangle_\mu \tag{82}$$

$$= \sum_{ij} a_i a_j \langle \chi_j | \mathcal{T} \circ \chi_i \rangle_\mu \tag{83}$$

$$= \sum_{ij} a_i a_j C_{ij} = a^T C a \tag{84}$$

Now maximize this as a function of a, remembering the constraint that a be unit norm. That defines the following optimization problem.

$$\begin{aligned} &\underset{a}{\operatorname{argmax}}\, a^T C a \\ &\text{s.t. } a^T a = 1 \end{aligned}$$

This is actually a really standard problem. This functional form is called the Rayleigh quotient. The unique maximizer is the first eigenvector of C, and the value of the objective at the maximum is the eigenvalue. To see why, just assume that a is an eigenvector. Then, $Ca = \lambda a$, so you can see that the value of the objective would be λ. And so the maximum value of that optimization problem is given by the eigenvector of C associated with the largest eigenvalue.

So we know that the optimized a^* is the dominant eigenvector,

$$C a^* = \lambda a^* \tag{85}$$

Recall that we can write C in terms of the transition probability matrix T. Plugging that in to the eigenvalue equation for C, we get

$$D(\pi^{1/2})\ T\ D(\pi^{-1/2}) a = \lambda a \tag{86}$$

Now left-multiply both sides by $D(\pi^{-1/2})$, giving

$$T\ D(\pi^{-1/2}) a = \lambda D(\pi^{-1/2}) a \tag{87}$$

Reading this off, it is an eigenvalue/eigenvector equation for T. We can see that $v \equiv D(\pi^{-1/2})a$ must be an eigenvector of of T, with eigenvalue λ. So if you know v (because you ran MSMBuilder/EMMA to estimate the transition matrix T from a trajectory and found the largest eigenvector of that matrix), then you also know the elements of a.

$$a_i^* = \sqrt{\pi} v_i \tag{88}$$

Okay, we're basically done now. Recall that the whole goal was to variationally optimize a function in this specific basis set of indicators, $\psi_2 \approx \tilde{\psi}_2 = \sum_{i=1}^N a_i^* \chi_i$, so as approximate the first eigenfunction of the transfer operator. And we used this specific basis set because the necessary matrix elements could be expressed as transition probabilities, so we actually have a way to compute everything by first computing T, then finding the largest eigenvector of T, and then scaling that eigenvector by $\sqrt{\pi}$ to get the optimized coefficients a_i.

So with this explicit expression for the a_i, we have

$$\tilde{\psi}_2 = \sum_{i=1}^N \sqrt{\pi} v_i \chi_i \tag{89}$$

Plug in the definition of χ_i from before, and the square roots of π_i cancel and you get

$$\tilde{\psi}_2 = \sum_{i=1}^N v_i \mathbf{1}_{S_i}(x) \tag{90}$$

A. Interpretation of the last equation

Okay, this is the final answer. The whole point of this is to prove in a very rigorous way that the eigenvectors of a discrete-state MSM are the variationally-optimial approximations to the true eigenvectors of the continuous-state transfer operator. They are the best discrete approximation that you can make in the specific choice of basis, which is basically a sum of step functions.

You can interpret an discrete-state MSM as a way of estimating the dominant eigenfunctions of a continuous-space Markov transfer operator. And the eigenvectors you get can be interpreted as step-function-based approximations to these continuous operator eigenfunctions. Basically, it's the same as the way you might make a histogram (built out of steps) to approximate some continuous probability distribution.

B. Comparing Markov state models

This also gives you a very rigorous way of comparing two different Markov models for the same stochastic process that use different state decompositions. Assuming that the amount of data is finite so you don't have any problems with statistics, you *still* make some error in approximating the continuous eigenfunctions with these step functions. And according to the interpretation of the variational theorem as the measure of the *quality* of an ansatz eigenfunction, the state decomposition that leads to a Markov model with a larger leading dynamical eigenvalue is the better state decomposition.