



October 22, 2010

Information Geometry (Part 1)

John Baez

Information geometry is the study of 'statistical manifolds', which are spaces where each point is a hypothesis about some state of affairs. In statistics a hypothesis amounts to a probability distribution, but we'll also be looking at the quantum version of a probability distribution, which is called a 'mixed state'. Every statistical manifold comes with a way of measuring distances and angles, called the [Fisher information metric](#). In the first seven articles in this series, I'll try to figure out what this metric really means. The formula for it is simple enough, but when I first saw it, it seemed quite mysterious.

A good place to start is this interesting paper:

- Gavin E. Crooks, [Measuring thermodynamic length](#).

which was pointed out by [John Furey](#) in a discussion about [entropy and uncertainty](#).

The idea here should work for either classical or quantum statistical mechanics. The paper describes the classical version, so just for a change of pace let me describe the quantum version.

First a lightning review of [quantum statistical mechanics](#). Suppose you have a quantum system with some Hilbert space. When you know as much as possible about your system, then you describe it by a unit vector in this Hilbert space, and you say your system is in a [pure state](#). Sometimes people just call a pure state a 'state'. But that can be confusing, because in statistical mechanics you also need more general 'mixed states' where you *don't* know as much as possible. A mixed state is described by a [density matrix](#), meaning a [positive operator](#) ρ with [trace](#) equal to 1:

$$\mathrm{tr}(\rho) = 1$$

The idea is that any observable is described by a self-adjoint operator A , and the expected value of this observable in the mixed state ρ is

$$\langle A \rangle = \mathrm{tr}(\rho A)$$

The **entropy** of a mixed state is defined by

$$S(\rho) = -\mathrm{tr}(\rho \ln \rho)$$

where we take the logarithm of the density matrix just by taking the log of each of its eigenvalues, while keeping the same eigenvectors. This formula for entropy should remind you of the one that Gibbs and Shannon used — the one I explained [a while back](#).

Back then I told you about the 'Gibbs ensemble': the mixed state that maximizes entropy subject to the constraint that some observables have a given value. We can do the same thing in quantum mechanics, and we can even do it for a bunch of observables at once. Suppose we have some observables X_1, \dots, X_n and we want to find the mixed state ρ that maximizes entropy subject to these constraints:

$$\langle X_i \rangle = x_i$$

for some numbers x_i . Then a little exercise in [Lagrange multipliers](#) shows that the answer is the **Gibbs state**:

$$\rho = \frac{1}{Z} \exp(-\lambda_1 X_1 + \dots + \lambda_n X_n)$$

Huh? 🤔

This answer needs some explanation. First of all, the numbers $\lambda_1, \dots, \lambda_n$ are called Lagrange multipliers. You have to choose them right to get

$$\langle X_i \rangle = x_i$$

So, in favorable cases, they will be functions of the numbers x_i . And when you're really lucky, you can solve for the numbers x_i in terms of the numbers λ_i . We call λ_i the **conjugate variable** of the observable X_i . For example, the conjugate variable of energy is inverse temperature!

Second of all, we take the exponential of a self-adjoint operator just as we took the logarithm of a density matrix: just take the exponential of each eigenvalue.

(At least this works when our self-adjoint operator has only eigenvalues in its spectrum, not any continuous spectrum. Otherwise we need to get serious and use the [functional calculus](#). Luckily, if your system's Hilbert space is finite-dimensional, you can ignore this parenthetical remark!)

But third: what's that number Z ? It begins life as a humble normalizing factor. Its job is to make sure ρ has trace equal to 1:

$$Z = \text{tr}(\exp(-\lambda_1 X_1 + \dots + \lambda_n X_n))$$

However, once you get going, it becomes incredibly important! It's called the **partition function** of your system.

As an example of what it's good for, it turns out you can compute the numbers x_i as follows:

$$x_i = -\frac{\partial}{\partial \lambda_i} \ln Z$$

In other words, you can compute the expected values of the observables X_i by differentiating the log of the partition function:

$$\langle X_i \rangle = -\frac{\partial}{\partial \lambda_i} \ln Z$$

Or in still other words,

$$\langle X_i \rangle = -\frac{1}{Z} \frac{\partial Z}{\partial \lambda_i}$$

To believe this you just have to take the equations I've given you so far and mess around — there's really no substitute for doing it yourself. I've done it fifty times, and every time I feel smarter.

But we can go further: after the '[expected value](#)' or 'mean' of an observable comes its [variance](#), which is the square of its standard deviation:

$$(\Delta A)^2 = \langle A^2 \rangle - \langle A \rangle^2$$

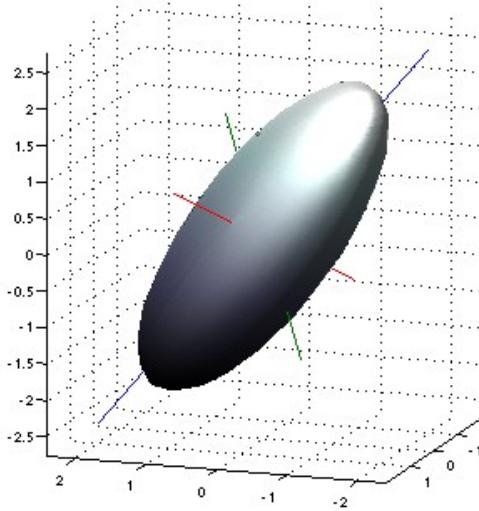
This measures the size of fluctuations around the mean. And in the Gibbs state, we can compute the variance of the observable X_i as the *second* derivative of the log of the partition function:

$$\langle X_i^2 \rangle - \langle X_i \rangle^2 = \frac{\partial^2}{\partial^2 \lambda_i} \ln Z$$

Again: calculate and see.

But when we've got lots of observables, there's something better than the variance of each one. There's the [covariance matrix](#) of the whole lot of them! Each observable X_i fluctuates around its mean value x_i ... but these fluctuations are not independent! They're *correlated*, and the covariance matrix says how.

All this is very visual, at least for me. If you imagine the fluctuations as forming a blurry patch near the point (x_1, \dots, x_n) , this patch will be ellipsoidal in shape, at least when all our random fluctuations are Gaussian. And then the *shape* of this ellipsoid is precisely captured by the covariance matrix! In particular, the eigenvectors of the covariance matrix will point along the principal axes of this ellipsoid, and the eigenvalues will say how stretched out the ellipsoid is in each direction!



To understand the covariance matrix, it may help to start by rewriting the variance of a single observable A as

$$(\Delta A)^2 = \langle (A - \langle A \rangle)^2 \rangle$$

That's a lot of angle brackets, but the meaning should be clear. First we look at the difference between our observable and its mean value, namely

$$A - \langle A \rangle$$

Then we square this, to get something that's big and positive whenever our observable is far from its mean. Then we take the mean value of the *that*, to get an idea of how far our observable is from the mean *on average*.

We can use the same trick to define the covariance of a bunch of observables X_i . We get an $n \times n$ matrix called the **covariance matrix**, whose entry in the i th row and j th column is

$$\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

If you think about it, you can see that this will measure correlations in the fluctuations of your observables.

An interesting difference between classical and quantum mechanics shows up here. In classical mechanics the covariance matrix is always symmetric — but not in quantum mechanics! You see, in classical mechanics, whenever we have two observables A and B , we have

$$\langle AB \rangle = \langle BA \rangle$$

since observables commute. But in quantum mechanics this is not true! For example, consider the position q and momentum p of a particle. We have

$$qp = pq + i$$

so taking expectation values we get

$$\langle qp \rangle = \langle pq \rangle + i$$

So, it's easy to get a non-symmetric covariance matrix when our observables X_i don't commute. However, the *real part* of the covariance matrix is symmetric, even in quantum mechanics. So let's define

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

You can check that the matrix entries here are the second derivatives of the partition function:

$$g_{ij} = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \ln Z$$

And now for the cool part: this is where information geometry comes in! Suppose that for any choice of values x_i we have a Gibbs state with

$$\langle X_i \rangle = x_i$$

Then for each point

$$x = (x_1, \dots, x_n) \in \mathbf{R}^n$$

we have a matrix

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \ln Z$$

And this matrix is not only symmetric, it's also [positive](#). And when it's [positive definite](#) we can think of it as an inner product on the [tangent space](#) of the point x . In other words, we get a [Riemannian metric](#) on \mathbf{R}^n . This is called the [Fisher information metric](#).

I hope you can see through the jargon to the simple idea. We've got a space. Each point x in this space describes the maximum-entropy state of a quantum system for which our observables have specified mean values. But in each of these states, the observables are random variables. They don't just sit at their mean value, they fluctuate! You can picture these fluctuations as forming a little smeared-out blob in our space. To a first approximation, this blob is an ellipsoid. And if we think of this ellipsoid as a 'unit ball', it gives us a standard for measuring the *length* of any little vector sticking out of our point. In other words, we've got a Riemannian metric: *the Fisher information metric!*

Now if you look at the Wikipedia article you'll see a more general but to me somewhat [scarier definition](#) of the Fisher information metric. This applies whenever we've got a manifold whose points label *arbitrary* mixed states of a system. But Crooks shows this definition reduces to his — the one I just described — when our manifold is \mathbf{R}^n and it's parametrizing Gibbs states in the way we've just seen.

More precisely: both Crooks and the Wikipedia article describe the classical story, but it parallels the quantum story I've been telling... and I think the quantum version is well-known. I believe the quantum version of the Fisher information metric is sometimes called the [Bures metric](#), though I'm a bit confused about what the Bures metric actually is.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



October 23, 2010

Information Geometry (Part 2)

John Baez

Last time I provided some background to this paper:

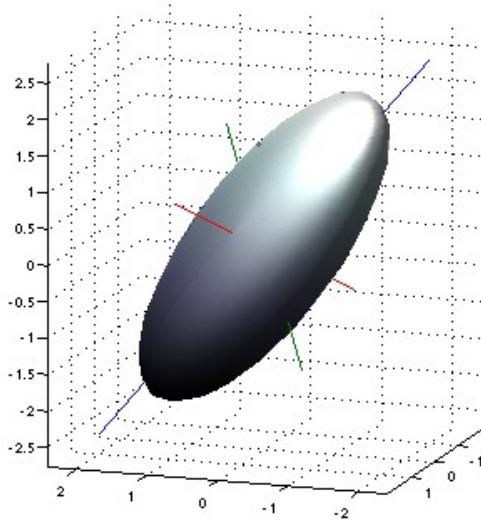
- Gavin E. Crooks, [Measuring thermodynamic length](#).

Now I'll tell you a bit about what it actually says!

Remember the story so far: we've got a physical system that's in a state of maximum entropy. I didn't emphasize this yet, but that happens whenever our system is in [thermodynamic equilibrium](#). An example would be a box of gas inside a piston. Suppose you choose any number for the energy of the gas and any number for its volume. Then there's a unique state of the gas that maximizes its entropy, given the constraint that *on average*, its energy and volume have the values you've chosen. And this describes what the gas will be like in equilibrium!

Remember, by 'state' I mean *mixed* state: it's a probabilistic description. And I say the energy and volume have chosen values *on average* because there will be random fluctuations. Indeed, if you look carefully at the head of the piston, you'll see it quivering: the volume of the gas only equals the volume you've specified *on average*. Same for the energy.

More generally: imagine picking any list of numbers, and finding the maximum entropy state where some chosen observables have these numbers as their average values. Then there will be fluctuations in the values of these observables — thermal fluctuations, but also possibly quantum fluctuations. So, you'll get a probability distribution on the space of possible values of your chosen observables. You should visualize this probability distribution as a little fuzzy cloud centered at the average value!



To a first approximation, this cloud will be shaped like a little ellipsoid. And if you can pick the average value of your observables to be whatever you'll like, you'll get lots of little ellipsoids this way, one centered at each point.

And the cool idea is to imagine the space of possible values of your observables as having a weirdly warped geometry, such that *relative to this geometry, these ellipsoids are actually spheres*.

This weirdly warped geometry is an example of an '[information geometry](#)': a geometry that's defined using the concept of information. This shouldn't be surprising: after all, we're talking about maximum entropy, and [entropy is related to information](#). But I want to gradually make this idea more precise. Bring on the math!

We've got a bunch of observables X_1, \dots, X_n , and we're assuming that for any list of numbers x_1, \dots, x_n , the system has a unique maximal-entropy state ρ for which the expected value of the observable X_i is x_i :

$$\langle X_i \rangle = x_i$$

This state ρ is called the **Gibbs state** and I told you how to find it when it exists. In fact it may not exist for *every* list of numbers x_1, \dots, x_n , but we'll be perfectly happy if it does for all choices of

$$x = (x_1, \dots, x_n)$$

lying in some open subset of \mathbf{R}^n

By the way, I should really call this Gibbs state $\rho(x)$ or something to indicate how it depends on x , but I won't usually do that. I expect some intelligence on your part!

Now at each point x we can define a [covariance matrix](#)

$$\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

If we take its real part, we get a symmetric matrix:

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

It's also nonnegative — that's easy to see, since the variance of a probability distribution can't be negative. When we're lucky this matrix will be [positive definite](#). When we're even luckier, it will depend smoothly on x . In this case, g_{ij} will define a [Riemannian metric](#) on our open set.

So far this is all review of [last time](#). Sorry: I seem to have reached the age where I can't say anything interesting without warming up for about 15 minutes first. It's like when my mom tells me about an exciting event that happened to her: she starts by saying "Well, I woke up, and it was cloudy out..."

But now I want to give you an explicit formula for the metric g_{ij} , and then *rewrite it* in a way that'll work even when the state ρ is *not* a maximal-entropy state. And this formula will then be the general definition of the '[Fisher information metric](#)' (if we're doing classical mechanics), or a quantum version thereof (if we're doing quantum mechanics).

Crooks does the classical case — so let's do the quantum case, okay? Last time I claimed that in the quantum case, our maximum-entropy state is the **Gibbs state**

$$\rho = \frac{1}{Z} e^{-\lambda^i X_i}$$

where λ^i are the 'conjugate variables' of the observables X_i , we're using the [Einstein summation convention](#) to sum over repeated indices that show up once upstairs and once downstairs, and Z is the **partition function**

$$Z = \text{tr}(e^{-\lambda^i X_i})$$

(To be honest: last time I wrote the indices on the conjugate variables λ^i as subscripts rather than superscripts, because I didn't want some poor schlep out there to think that $\lambda^1, \dots, \lambda^n$ were the powers of some number λ . But now I'm assuming you're all grown up and ready to juggle indices! We're doing Riemannian geometry, after all.)

Also last time I claimed that it's tremendously fun and enlightening to take the derivative of the logarithm of Z . The reason is that it gives you the mean values of your observables:

$$\langle X_i \rangle = -\frac{\partial}{\partial \lambda^i} \ln Z$$

But now let's take the derivative of the logarithm of ρ . Remember, ρ is an operator — in fact a [density matrix](#). But we can take its logarithm as explained last time, and the usual rules apply, so starting from

$$\rho = \frac{1}{Z} e^{-\lambda^i X_i}$$

we get

$$\ln \rho = -\lambda^i X_i - \ln Z$$

Next, let's differentiate both sides with respect to λ^i . Why? Well, from what I just said, you should be itching to differentiate $\ln Z$. So let's give in to that temptation:

$$\frac{\partial}{\partial \lambda^i} \ln \rho = -X_i + \langle X_i \rangle$$

Hey! Now we've got a formula for the 'fluctuation' of the observable X_i — that is, how much it differs from its mean value:

$$X_i - \langle X_i \rangle = -\frac{\partial \ln \rho}{\partial \lambda^i}$$

This is incredibly cool! I should have learned this formula decades ago, but somehow I just bumped into it now. I knew of course that $\ln \rho$ shows up in the formula for the **entropy**:

$$S(\rho) = \text{tr}(\rho \ln \rho)$$

But I never had the brains to think about $\ln \rho$ all by itself. So I'm really excited to discover that it's an interesting entity in its own right — and fun to differentiate, just like $\ln Z$.

Now we get our cool formula for g_{ij} . Remember, it's defined by

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

But now that we know

$$X_i - \langle X_i \rangle = -\frac{\partial \ln \rho}{\partial \lambda^i}$$

we get the formula we were looking for:

$$g_{ij} = \text{Re} \left\langle \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j} \right\rangle$$

Beautiful, eh? And of course the expected value of any observable A in the state ρ is

$$\langle A \rangle = \text{tr}(\rho A)$$

so we can also write the covariance matrix like this:

$$g_{ij} = \text{Re} \text{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j})$$

Lo and behold! This formula makes sense whenever ρ is *any* density matrix depending smoothly on some parameters λ^i . We don't need it to be a Gibbs state! So, we can work more generally.

Indeed, whenever we have *any* smooth function from a manifold to the space of density matrices for some Hilbert space, we can define g_{ij} by the above formula! And when it's positive definite, we get a Riemannian metric on our manifold: the **Bures information metric**.

The classical analogue is the somewhat more well-known 'Fisher information metric'. When we go from quantum to classical, operators become functions and traces become integrals. There's nothing complex anymore, so taking the real part becomes unnecessary. So the Fisher information metric looks like this:

$$g_{ij} = \int_{\Omega} p(\omega) \frac{\partial \ln p(\omega)}{\partial \lambda^i} \frac{\partial \ln p(\omega)}{\partial \lambda^j} d\omega$$

Here I'm assuming we've got a smooth function p from some manifold M to the space of probability distributions on some measure space $(\Omega, d\omega)$. Working in local coordinates λ^i on our manifold M , the above formula defines a Riemannian metric on M , at least when g_{ij} is positive definite. And that's the **Fisher information metric**!

Crooks says more: he describes an experiment that would let you measure the length of a path with respect to the Fisher information metric — at least in the case where the state $\rho(x)$ is the Gibbs state with $\langle X_i \rangle = x_i$. And that explains why he calls it 'thermodynamic length'.

There's a lot more to say about this, and also about another question: *What use is the Fisher information metric in the general case where the states $\rho(x)$ aren't Gibbs states?*

But it's dinnertime, so I'll stop here.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez

baez@math.removethis.ucr.andthis.edu

[home](#)





October 25, 2010

Information Geometry (Part 3)

John Baez

So far in this series of posts I've been explaining a paper by Gavin Crooks. Now I want to go ahead and explain a little research of my own.

I'm not claiming my results are new—indeed I have no idea whether they are, and I'd like to hear from any experts who might know. I'm just claiming that this is some work I did last weekend.

People sometimes worry that if they explain their ideas before publishing them, someone will 'steal' them. But I think this overestimates the value of ideas, at least in esoteric fields like mathematical physics. The problem is not people stealing your ideas: the hard part is *giving them away*. And let's face it, people in love with math and physics will do research unless you actively stop them. I'm reminded of this scene from the [Marx Brothers movie](#) where Harpo and Chico, playing wandering musicians, walk into a hotel and offer to play:

Groucho: What do you fellows get an hour?

Chico: Oh, for playing we getta ten dollars an hour.

Groucho: I see...What do you get for not playing?

Chico: Twelve dollars an hour.

Groucho: Well, clip me off a piece of that.

Chico: Now, for rehearsing we make special rate. Thatsa fifteen dollars an hour.

Groucho: That's for rehearsing?

Chico: Thatsa for rehearsing.

Groucho: And what do you get for not rehearsing?

Chico: You couldn't afford it.

So, I'm just rehearsing in public here—but I of course I hope to write a paper about this stuff someday, once I get enough material.

Remember where we were. We had considered a manifold—let's finally give it a name, say M —that parametrizes Gibbs states of some physical system. By [Gibbs state](#), I mean a state that maximizes entropy subject to constraints on the expected values of some observables. And we had seen that in favorable cases, we get a Riemannian metric on M ! It looks like this:

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

where X_i are our observables, and the angle bracket means 'expected value'.

All this applies to both classical or quantum mechanics. Crooks wrote down a beautiful formula for this metric in the classical case. But since I'm at the Centre for *Quantum* Technologies, not the Centre for Classical Technologies, I redid his calculation in the quantum case. The big difference is that in quantum mechanics, observables don't commute! But in the calculations I did, that didn't seem to matter much—mainly because I took a lot of traces, which imposes a kind of commutativity:

$$\text{tr}(AB) = \text{tr}(BA)$$

In fact, if I'd wanted to show off, I could have done the classical and quantum cases simultaneously by replacing all operators by elements of any [von Neumann algebra](#) equipped with a [trace](#). Don't worry about this much: it's just a general formalism for treating classical and quantum mechanics on an equal footing. One example is the algebra of bounded operators on a Hilbert space, with the usual concept of trace. Then we're doing quantum mechanics as usual. But another example is the algebra of suitably nice functions on a suitably nice space, where taking the trace of a function means *integrating* it. And then we're doing classical mechanics!

For example, I showed you how to derive a beautiful formula for the metric I wrote down a minute ago:

$$g_{ij} = \text{Re} \text{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j})$$

But if we want to do the classical version, we can say *Hey, presto!* and write it down like this:

$$g_{ij} = \int_{\Omega} p(\omega) \frac{\partial \ln p(\omega)}{\partial \lambda^i} \frac{\partial \ln p(\omega)}{\partial \lambda^j} d\omega$$

What did I do just now? I changed the trace to an integral over some space Ω . I rewrote ρ as p to make you think 'probability distribution'. And I don't need to take the real part anymore, since everything is already real when we're doing classical mechanics. Now this metric is the [Fisher information metric](#) that statisticians know and love!

In what follows, I'll keep talking about the quantum case, but in the back of my mind I'll be using von Neumann algebras, so everything will apply to the classical case too.

So what am I going to do? I'm going to fix a big problem with the story I've told so far.

Here's the problem: so far we've only studied a special case of the Fisher information metric. We've been assuming our states are Gibbs states, parametrized by the expectation values of some observables X_1, \dots, X_n . Our manifold M was really just some open subset of \mathbf{R}^n : a point in here was a list of expectation values.

But people like to work a lot more generally. We could look at *any* smooth function ρ from a smooth manifold M to the set of density matrices for some quantum system. We can still write down the metric

$$g_{ij} = \text{Re} \text{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j})$$

in this more general situation. Nobody can stop us! But it would be better if we could *derive* this formula, as before, starting from a formula like the one we had before:

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

The challenge is that now we don't have observables X_i to start with. All we have is a smooth function ρ from some manifold to some set of states. How can we pull observables out of thin air?

Well, you may remember that last time we had

$$\rho = \frac{1}{Z} e^{-\lambda^i X_i}$$

where λ^i were some functions on our manifold and

$$Z = \text{tr}(e^{-\lambda^i X_i})$$

was the [partition function](#). Let's copy this idea.

So, we'll start with our density matrix ρ , but then write it as

$$\rho = \frac{1}{Z} e^{-A}$$

where A is some self-adjoint operator and

$$Z = \text{tr}(e^{-A})$$

(Note that A , like ρ , is really an operator-valued function on M . So, I should write something like $A(x)$ to denote its value at a particular point $x \in M$, but I won't usually do that. As usual, I expect some intelligence on your part!)

Now we can repeat some calculations I did last time. As before, let's take the logarithm of ρ :

$$\ln \rho = -A - \ln Z$$

and then differentiate it. Suppose λ^i are local coordinates near some point of M . Then

$$\frac{\partial}{\partial \lambda^i} \ln \rho = -\frac{\partial}{\partial \lambda^i} A - \frac{1}{Z} \frac{\partial}{\partial \lambda^i} Z$$

Last time we had nice formulas for both terms on the right-hand side above. To get similar formulas now, let's define operators

$$X_i = \frac{\partial}{\partial \lambda^i} A$$

This gives a nice name to the first term on the right-hand side above. What about the second term? We can calculate it out:

$$\frac{1}{Z} \frac{\partial}{\partial \lambda^i} Z = \frac{1}{Z} \frac{\partial}{\partial \lambda^i} \text{tr}(e^{-A}) = \frac{1}{Z} \text{tr}\left(\frac{\partial}{\partial \lambda^i} e^{-A}\right) = -\frac{1}{Z} \text{tr}\left(e^{-A} \frac{\partial}{\partial \lambda^i} A\right)$$

where in the last step we use the chain rule. Next, use the definition of ρ and X_i , and get:

$$\frac{1}{Z} \frac{\partial}{\partial \lambda^i} Z = -\text{tr}(\rho X_i) = -\langle X_i \rangle$$

This is just what we got last time! Ain't it fun to calculate when it all works out so nicely?

So, putting both terms together, we see

$$\frac{\partial}{\partial \lambda^i} \ln \rho = -X_i + \langle X_i \rangle$$

or better:

$$X_i - \langle X_i \rangle = -\frac{\partial}{\partial \lambda^i} \ln \rho$$

This is a nice formula for the 'fluctuation' of the observables X_i , meaning how much they differ from their expected values. And it looks exactly like the formula we had last time! The difference is that last time we *started out* assuming we had a bunch of observables, X_i , and defined ρ to be the state maximizing the entropy subject to constraints on the expectation values of all these observables. Now we're starting with ρ and working backwards.

From here on out, it's easy. As before, we can define g_{ij} to be the real part of the covariance matrix:

$$g_{ij} = \operatorname{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

Using the formula

$$X_i - \langle X_i \rangle = -\frac{\partial}{\partial \lambda^i} \ln \rho$$

we get

$$g_{ij} = \operatorname{Re} \left\langle \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j} \right\rangle$$

or

$$g_{ij} = \operatorname{Re} \operatorname{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j})$$

Voilà!

When this matrix is positive definite at every point, we get a Riemannian metric on M . Last time I said this is what people call the '[Bures metric](#)'—though frankly, now that I examine the formulas, I'm not so sure. But in the classical case, it's called the Fisher information metric.

Differential geometers like to use ∂_i as a shorthand for $\frac{\partial}{\partial \lambda^i}$, so they'd write down our metric in a prettier way:

$$g_{ij} = \operatorname{Re} \operatorname{tr}(\rho \partial_i(\ln \rho) \partial_j(\ln \rho))$$

Differential geometers like coordinate-free formulas, so let's also give a coordinate-free formula for our metric. Suppose $x \in M$ is a point in our manifold, and suppose v, w are tangent vectors to this point. Then

$$g(v, w) = \operatorname{Re} \langle v(\ln \rho) w(\ln \rho) \rangle = \operatorname{Re} \operatorname{tr}(\rho v(\ln \rho) w(\ln \rho))$$

Here $\ln \rho$ is a smooth operator-valued function on M , and $v(\ln \rho)$ means the derivative of this function in the v direction at the point x .

So, this is all very nice. To conclude, two more points: a technical one, and a more important philosophical one.

First, the technical point. When I said ρ could be *any* smooth function from a smooth manifold to some set of states, I was actually lying. That's an important pedagogical technique: the brazen lie.

We can't really take the logarithm of *every* density matrix. Remember, we take the log of a density matrix by taking the log of all its eigenvalues. These eigenvalues are ≥ 0 , but if one of them is zero, we're in trouble! The logarithm of zero is undefined.

On the other hand, there's no problem taking the logarithm of our density-matrix-valued function ρ when it's [positive definite](#) at each point of M . You see, a density matrix is positive definite iff its eigenvalues are all > 0 . In this case it has a unique self-adjoint logarithm.

So, we must assume ρ is positive definite. But what's the physical significance of this 'positive definiteness' condition? Well, any density matrix can be diagonalized using some orthonormal basis. It can then be seen as probabilistic mixture—not a quantum superposition!—of pure states taken from this basis. Its eigenvalues are the probabilities of finding the mixed state to be in one of these pure states. So, saying that all its eigenvalues are all > 0 amounts to saying that all the pure states in this orthonormal basis show up with *nonzero* probability! Intuitively, this means our mixed state is 'really mixed'. For example, it can't be a pure state. In math jargon, it means our mixed state is in the *interior* of the convex set of mixed states.

Second, the philosophical point. Instead of starting with the density matrix ρ , I took A as fundamental. But different choices of A give the same ρ . After all,

$$\rho = \frac{1}{Z} e^{-A}$$

where we cleverly divide by the normalization factor

$$Z = \text{tr}(e^{-A})$$

to get $\text{tr } \rho = 1$. So, if we multiply e^{-A} by any positive constant, or indeed any positive *function* on our manifold M , ρ will remain unchanged!

So we have added a little extra information when switching from ρ to A . You can think of this as 'gauge freedom', because I'm saying we can do any transformation like

$$A \mapsto A + f$$

where

$$f : M \rightarrow \mathbf{R}$$

is a smooth function. This doesn't change ρ , so arguably it doesn't change the 'physics' of what I'm doing. It *does* change Z . It also changes the observables

$$X_i = \frac{\partial}{\partial \lambda^i} A$$

But it doesn't change their 'fluctuations'

$$X_i - \langle X_i \rangle$$

so it doesn't change the metric g_{ij} .

This gauge freedom is interesting, and I want to understand it better. It's related to something very simple yet mysterious. In statistical mechanics the partition function Z begins life as 'just a normalizing factor'. If you change the physics so that Z gets multiplied by some number, the Gibbs state doesn't change. But then the partition function takes on an incredibly significant role as something whose logarithm you differentiate to get lots of physically interesting information! So in some sense the partition function doesn't matter much... but *changes* in the partition function matter a lot.

This is just like the split personality of phases in quantum mechanics. On the one hand they 'don't matter': you can multiply a unit vector by any phase and the pure state it defines doesn't change. But on the other hand, *changes* in phase matter a lot.

Indeed the analogy here is quite deep: it's the analogy between probabilities in statistical mechanics and amplitudes in quantum mechanics, the analogy between $\exp(-\beta H)$ in statistical mechanics and $\exp(-itH/h)$ in quantum mechanics, and so on. This is part of a bigger story about 'rigs' which I told back in the [Winter 2007 quantum gravity seminar](#), especially in [week13](#). So, it's fun to see it showing up yet again... even though I don't completely understand it here.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



October 29, 2010

Information Geometry (Part 4)

John Baez

Before moving on, I'd like to clear up an important point, which had me confused for a while. A [Riemannian metric](#) must be symmetric:

$$g_{ij} = g_{ji}$$

When I first started thinking about this stuff, I defined the Fisher information metric to be the so-called '[covariance matrix](#)':

$$g_{ij} = \langle (X_i - \langle X_i \rangle) (X_j - \langle X_j \rangle) \rangle$$

where X_i are some observable-valued functions on a manifold M , and the angle brackets mean "expectation value", computed using a mixed state ρ that also depends on the point in M .

The covariance matrix is symmetric in classical mechanics, since then observables commute, so:

$$\langle AB \rangle = \langle BA \rangle$$

But it's not symmetric in quantum mechanics! After all, suppose q is the position operator for a particle, and p is the momentum operator. Then according to Heisenberg

$$qp = pq + i$$

in units where Planck's constant is 1. Taking expectation values, we get:

$$\langle qp \rangle = \langle pq \rangle + i$$

and in particular:

$$\langle qp \rangle \neq \langle pq \rangle$$

We can use this to get examples where g_{ij} is not symmetric.

However, it turns out that the *real part* of the covariance matrix is symmetric, even in quantum mechanics—and that's what we should use as our Fisher information metric.

Why is the real part of the covariance matrix symmetric, even in quantum mechanics? Well, suppose ρ is any density matrix, and A and B are any observables. Then by definition

$$\langle AB \rangle = \text{tr}(\rho AB)$$

so taking the complex conjugate of both sides

$$\langle AB \rangle^* = \text{tr}(\rho AB)^* = \text{tr}((\rho AB)^*) = \text{tr}(B^* A^* \rho^*)$$

where I'm using an asterisk both for the complex conjugate of a number and the adjoint of an operator. But our observables are self-adjoint, and so is our density matrix, so we get

$$\mathrm{tr}(B^*A^*\rho^*) = \mathrm{tr}(BA\rho) = \mathrm{tr}(\rho BA) = \langle BA \rangle$$

where in the second step we used the cyclic property of the trace. In short:

$$\langle AB \rangle^* = \langle BA \rangle$$

If we take real parts, we get something symmetric:

$$\mathrm{Re}\langle AB \rangle = \mathrm{Re}\langle BA \rangle$$

So, if we redefine the Fisher information metric to be the *real part* of the covariance matrix:

$$g_{ij} = \mathrm{Re}\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

then it's symmetric, as it should be.

Last time I mentioned a general setup using von Neumann algebras, that handles the classical and quantum situations simultaneously. That applies here! Taking the real part has no effect in classical mechanics, so we don't need it there—but it doesn't hurt, either.

Taking the real part never has any effect when $i = j$, either, since the expected value of the *square* of an observable is a nonnegative number:

$$\langle (X_i - \langle X_i \rangle)^2 \rangle \geq 0$$

This has two nice consequences.

First, we get

$$g_{ii} = \langle (X_i - \langle X_i \rangle)^2 \rangle \geq 0$$

and since this is true in *any* coordinate system, our would-be metric g is indeed nonnegative. It'll be an honest Riemannian metric whenever it's positive definite.

Second, suppose we're working in the special case discussed in [Part 2](#), where our manifold is an open subset of \mathbf{R}^n , and ρ at the point $x \in \mathbf{R}^n$ is the Gibbs state with $\langle X_i \rangle = x_i$. Then all the usual rules of statistical mechanics apply. So, we can compute the variance of the observable X_i using the partition function Z :

$$\langle (X_i - \langle X_i \rangle)^2 \rangle = \frac{\partial^2}{\partial \lambda_i^2} \ln Z$$

In other words,

$$g_{ii} = \frac{\partial^2}{\partial \lambda_i^2} \ln Z$$

But since this is true in *any* coordinate system, we must have

$$g_{ij} = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \ln Z$$

(Here I'm using a little math trick: two symmetric bilinear forms whose diagonal entries agree in *any* basis must be equal. We've already seen that the left side is symmetric, and the right side is symmetric by a famous fact about mixed partial derivatives.)

However, I'm pretty sure this cute formula

$$g_{ij} = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \ln Z$$

only holds in the special case I'm talking about now, where points in \mathbf{R}^n are parametrizing Gibbs states in the obvious way. In general we must use

$$g_{ij} = \operatorname{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

or equivalently,

$$g_{ij} = \operatorname{Re} \operatorname{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda_i} \frac{\partial \ln \rho}{\partial \lambda_j})$$

Okay. So much for cleaning up Last Week's Mess. Here's something new. We've seen that whenever A and B are observables (that is, self-adjoint),

$$\langle AB \rangle^* = \langle BA \rangle$$

We got something symmetric by taking the real part:

$$\operatorname{Re} \langle AB \rangle = \operatorname{Re} \langle BA \rangle$$

Indeed,

$$\operatorname{Re} \langle AB \rangle = \frac{1}{2} \langle AB + BA \rangle$$

But by the same reasoning, we get something *antisymmetric* by taking the *imaginary* part:

$$\operatorname{Im} \langle AB \rangle = -\operatorname{Im} \langle BA \rangle$$

and indeed,

$$\operatorname{Im} \langle AB \rangle = \frac{1}{2i} \langle AB - BA \rangle$$

Commutators like $AB - BA$ are important in quantum mechanics, so maybe we shouldn't just throw out the imaginary part of the covariance matrix in our desperate search for a Riemannian metric! Besides the symmetric tensor on our manifold M :

$$g_{ij} = \operatorname{Re} \operatorname{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda_i} \frac{\partial \ln \rho}{\partial \lambda_j})$$

we can also define a skew-symmetric tensor:

$$\omega_{ij} = \operatorname{Im} \operatorname{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda_i} \frac{\partial \ln \rho}{\partial \lambda_j})$$

This will vanish in the classical case, but not in the quantum case!

If you've studied enough geometry, you should now be reminded of things like 'Kähler manifolds' and 'almost Kähler manifolds'. A [Kähler manifold](#) is a manifold that's equipped with a symmetric tensor g and a skew-symmetric tensor ω which fit together in the best possible way. An [almost Kähler manifold](#) is something similar, but not quite as nice. We should probably see examples of these arising in information geometry! And that could be pretty interesting.

But in general, if we start with any old manifold M together with a function ρ taking values in mixed states, we seem to be making M into something even less nice. It gets a symmetric bilinear form g on each tangent space, and a skew-symmetric bilinear form ω , and they vary smoothly from point to point... but they might be degenerate, and I don't see any reason for them to 'fit together' in the nice way we need for a Kähler or almost Kähler manifold.

However, I still think something interesting might be going on here. For one thing, there are *other* situations in physics where a space of states is equipped with a symmetric g and a skew-symmetric ω . They show up in 'dissipative mechanics'—the study of systems whose entropy increases.

To conclude, let me remind you of some things I said in [week295](#) of This Week's Finds. This is a huge digression from information geometry, but I'd like to lay out the puzzle pieces in public view, in case it helps anyone get some good ideas.

I wrote:

- Hans Christian Öttinger, *Beyond Equilibrium Thermodynamics*, Wiley, 2005.

I thank Arnold Neumaier for pointing out this book! It considers a fascinating generalization of Hamiltonian mechanics that applies to systems with dissipation: for example, electrical circuits with resistors, or mechanical systems with friction.

In ordinary Hamiltonian mechanics the space of states is a manifold and time evolution is a flow on this manifold determined by a smooth function called the Hamiltonian, which describes the *energy* of any state. In this generalization the space of states is still a manifold, but now time evolution is determined by two smooth functions: the energy and the *entropy*! In ordinary Hamiltonian mechanics, energy is automatically conserved. In this generalization that's also true, but energy can go into the form of heat... and entropy automatically *increases*!

Mathematically, the idea goes like this. We start with a Poisson manifold, but in addition to the skew-symmetric Poisson bracket $\{F,G\}$ of smooth functions on some manifold, we also have a symmetric bilinear bracket $[F,G]$ obeying the Leibniz law

$$[F,GH] = [F,G]H + G[F,H]$$

and this positivity condition:

$$[F,F] \geq 0$$

The time evolution of any function is given by a generalization of Hamilton's equations:

$$dF/dt = \{H,F\} + [S,F]$$

where H is a function called the "energy" or "Hamiltonian", and S is a function called the "entropy". The first term on the right is the usual one. The new second term describes dissipation: as we shall see, it pushes the state towards increasing entropy.

If we require that

$$[H,F] = \{S,F\} = 0$$

for every function F , then we get conservation of energy, as usual in Hamiltonian mechanics:

$$dH/dt = \{H,H\} + [S,H] = 0$$

But we also get the second law of thermodynamics:

$$dS/dt = \{H,S\} + [S,S] \geq 0$$

Entropy always increases!

Öttinger calls this framework "GENERIC"—an annoying acronym for "General Equation for the NonEquilibrium Reversible-Irreversible Coupling". There are lots of papers about it. But I'm wondering if any geometers have looked into it!

If we didn't need the equations $[H,F] = \{S,F\} = 0$, we could easily get the necessary brackets starting with a Kähler manifold. The imaginary part of the Kähler structure is a symplectic structure, say ω , so we can define

$$\{F,G\} = \omega(dF,dG)$$

as usual to get Poisson brackets. The real part of the Kähler structure is a Riemannian structure, say g , so we can define

$$[F,G] = g(dF,dG)$$

This satisfies

$$[F,GH] = [F,G]H + G[F,H]$$

and

$$[F,F] \geq 0$$

Don't be fooled: this stuff is not rocket science. In particular, the inequality above has a simple meaning: when we move in the direction of the gradient of F , the function F increases. So adding the second term to Hamilton's equations has the effect of pushing the system towards increasing entropy.

Note that I'm being a tad unorthodox by letting ω and g eat cotangent vectors instead of tangent vectors—but that's no big deal. The big deal is this: if we start with a Kähler manifold and define brackets this way, we don't get $[H,F] = 0$ or $\{S,F\} = 0$ for all functions F unless H and S are constant! That's no good for applications to physics. To get around this problem, we would need to consider some sort of *degenerate* Kähler structure—one where ω and g are degenerate bilinear forms on the cotangent space.

Has anyone thought about such things? They remind me a little of "Dirac structures" and "generalized complex geometry"—but I don't know enough about those subjects to know if they're relevant here.

This GENERIC framework suggests that energy and entropy should be viewed as two parts of a single entity—maybe even its real and imaginary parts! And that in turn reminds me of other strange things, like the idea of using complex-valued Hamiltonians to describe dissipative systems, or the idea of "inverse temperature as imaginary time". I can't tell yet if there's a big idea lurking here, or just a mess....

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



November 2, 2010

Information Geometry (Part 5)

John Baez

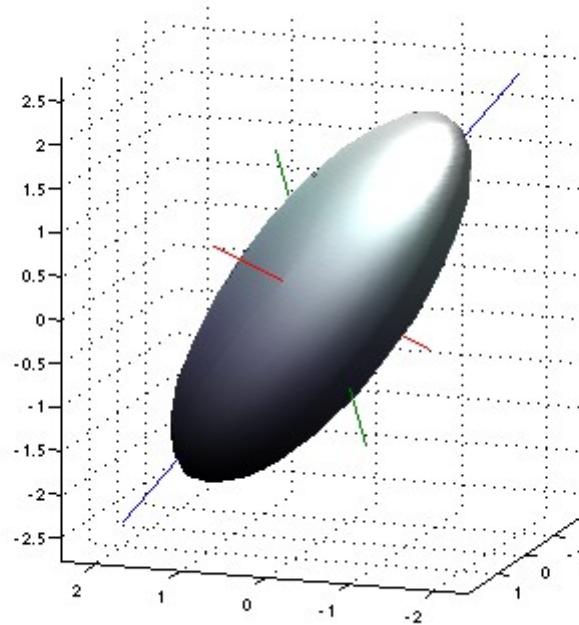
I'm trying to understand the Fisher information metric and how it's related to [Ottinger's formalism](#) for 'dissipative mechanics' — that is, mechanics including friction. They involve similar physics, and they involve similar math, but it's not quite clear how they fit together.

I think it will help to do an example. The harmonic oscillator is a trusty workhorse throughout physics, so let's do that.

So: suppose you have a rock hanging on a spring, and it can bounce up and down. Suppose it's in thermal equilibrium with its environment. It will wiggle up and down ever so slightly, thanks to thermal fluctuations. The hotter it is, the more it wiggles. These vibrations are random, so its position and momentum at any given moment can be treated as random variables.

If we take quantum mechanics into account, there's an extra source of randomness: *quantum* fluctuations. Now there will be fluctuations even at zero temperature. Ultimately this is due to the uncertainty principle. Indeed, if you know the position for sure, you can't know the momentum at all!

Let's see how the position, momentum and energy of our rock will fluctuate given that we know all three of these quantities *on average*. The fluctuations will form a little fuzzy blob, roughly ellipsoidal in shape, in the 3-dimensional space whose coordinates are position, momentum and energy:



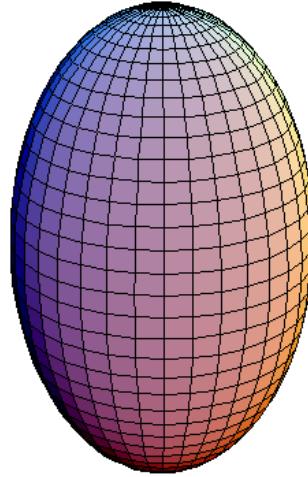
Yeah, I know you're sick of this picture, but this time it's for real: I want to calculate what this ellipsoid actually looks like! I'm not promising I'll do it — I may get stuck, or bored — but at least I'll *try*.

Before I start the calculation, let's guess the answer. A harmonic oscillator has a position q and momentum p , and its energy is

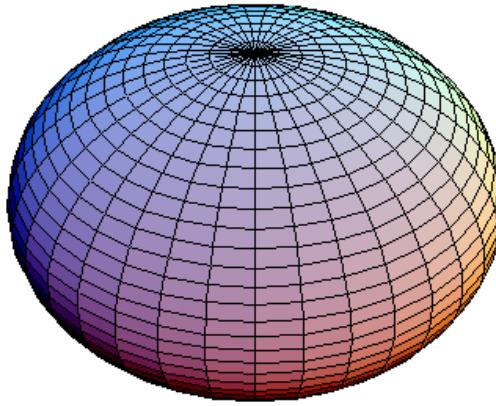
$$H = \frac{1}{2}(q^2 + p^2)$$

Here I'm working in units where lots of things equal 1, to keep things simple.

You'll notice that this energy has rotational symmetry in the position-momentum plane. This is ultimately what makes the harmonic oscillator such a beloved physical system. So, we might naively guess that our little ellipsoid will have rotational symmetry as well, like this:



or this:



Here I'm using the x and y coordinates for position and momentum, while the z coordinate stands for energy. So in these examples the position and momentum fluctuations are the same size, while the energy fluctuations, drawn in the vertical direction, might be bigger or smaller.

Unfortunately, this guess really is naive. After all, there are *lots* of these ellipsoids, one centered at each point in position-momentum-energy space. Remember the rules of the game! You give me any point in this space. I take the coordinates of this point as the *mean* values of position, momentum and energy, and I find the maximum-entropy state with these mean values. Then I work out the fluctuations in this state, and draw them as an ellipsoid.

If you pick a point where position and momentum have mean value zero, you haven't broken the rotational symmetry of the problem. So, my ellipsoid must be rotationally symmetric. But if you pick some other mean value for position and momentum, all bets are off!

Fortunately, this naive guess is actually right: *all* the ellipsoids are rotationally symmetric — even the ones centered at nonzero values of position and momentum! We'll see why soon. And if you've been following this series of posts, you'll know what this implies: the "Fisher information metric" g on position-momentum-energy space has rotational symmetry about any vertical axis. (Again, I'm using the vertical direction for energy.) So, if we slice this space with any horizontal plane, the metric on this plane must be the plane's usual metric times a constant:

$$g = \text{constant} (dq^2 + dp^2)$$

Why? Because only the usual metric on the plane, or any multiple of it, has ordinary rotations around every point as symmetries.

So, roughly speaking, we're recovering the 'obvious' geometry of the position-momentum plane from the Fisher information metric. *We're recovering 'ordinary' geometry from information geometry!*

But this should not be terribly surprising, since we used the harmonic oscillator Hamiltonian

$$H = \frac{1}{2}(q^2 + p^2)$$

as an input to our game. It's mainly just a confirmation that things are working as we'd hope.

There's more, though. [Last time](#) I realized that because observables in quantum mechanics don't commute, the Fisher information metric has a curious skew-symmetric partner called ω . So, we should also study this in our example. And when we do, we'll see that restricted to any horizontal plane in position-momentum-energy space, we get

$$\omega = \text{constant} (dq dp - dp dq)$$

This looks like a mutant version of the Fisher information metric

$$g = \text{constant} (dq^2 + dp^2)$$

and if you know your geometry, you'll know it's the usual '[symplectic structure](#)' on the position-energy plane — at least, times some constant.

All this is very reminiscent of Öttinger's work on dissipative mechanics. But we'll also see something else: while the constant in g depends on the energy — that is, on which horizontal plane we take — the constant in ω does not!

Why? It's perfectly sensible. The metric g on our horizontal plane keeps track of fluctuations in position and momentum. Thermal fluctuations get bigger when it's hotter — and to boost the average energy of our oscillator, we must heat it up. So, as we increase the energy, moving our horizontal plane further up in position-momentum-energy space, the metric on the plane gets bigger! In other words, our ellipsoids get a fat cross-section at high energies.

On the other hand, the symplectic structure ω arises from the fact that position q and momentum p don't commute in quantum mechanics. They obey Heisenberg's '[canonical commutation relation](#)':

$$qp - pq = i$$

This relation doesn't involve energy, so ω will be the same on every horizontal plane. And it turns out this relation implies

$$\omega = \text{constant} (dq dp - dp dq)$$

for some constant we'll compute later.

Okay, that's the basic idea. Now let's actually do some computations. For starters, let's see why all our ellipsoids have rotational symmetry!

To do this, we need to understand a bit about the mixed state ρ that maximizes entropy given certain mean values of position, momentum and energy. So, let's choose the numbers we want for these mean values (also known as 'expected values' or 'expectation values'):

$$\langle H \rangle = E$$

$$\langle q \rangle = q_0$$

$$\langle p \rangle = p_0$$

I hope this isn't too confusing: H, p, q are our observables which are operators, while E, p_0, q_0 are the mean values we have chosen for them. The state ρ depends on E, p_0 and q_0 .

We're doing quantum mechanics, so position q and momentum p are both self-adjoint operators on the Hilbert space $L^2(\mathbf{R})$:

$$(q\psi)(x) = x\psi(x)$$

$$(p\psi)(x) = -i\frac{d\psi}{dx}(x)$$

Indeed all our observables, including the Hamiltonian

$$H = \frac{1}{2}(p^2 + q^2)$$

are self-adjoint operators on this Hilbert space, and the state ρ is a [density matrix](#) on this space, meaning a positive self-adjoint operator with trace 1.

Now: how do we compute ρ ? It's a [Lagrange multiplier](#) problem: maximizing some function given some constraints. And it's well-known that when you solve this problem, you get

$$\rho = \frac{1}{Z}e^{-(\lambda^1 q + \lambda^2 p + \lambda^3 H)}$$

where $\lambda^1, \lambda^2, \lambda^3$ are three numbers we yet have to find, and Z is a normalizing factor called the [partition function](#):

$$Z = \text{tr}(e^{-(\lambda^1 q + \lambda^2 p + \lambda^3 H)})$$

Now let's look at a special case. If we choose $\lambda^1 = \lambda^2 = 0$, we're back a simpler and more famous problem, namely maximizing entropy subject to a constraint only on energy! The solution is then

$$\rho = \frac{1}{Z}e^{-\beta H}, \quad Z = \text{tr}(e^{-\beta H})$$

Here I'm using the letter β instead of λ^3 because this is traditional. This quantity has an important physical meaning! It's the *reciprocal of temperature* in units where Boltzmann's constant is 1.

Anyway, back to our special case! In this special case it's easy to explicitly calculate ρ and Z . Indeed, people have known how ever since [Planck](#) put the 'quantum' in quantum mechanics! He figured out how black-body radiation works. A box of hot radiation is just a big bunch of harmonic oscillators in thermal equilibrium. You can work out its partition function by multiplying the partition function of each one.

So, it would be great to reduce our general problem to this special case. To do this, let's rewrite

$$Z = \text{tr}(e^{-(\lambda^1 q + \lambda^2 p + \lambda^3 H)})$$

in terms of some new variables, like this:

$$\rho = \frac{1}{Z} e^{-\beta(H - fq - gp)}$$

where now

$$Z = \text{tr}(e^{-\beta(H - fq - gp)})$$

Think about it! Now our problem is just like an oscillator with a modified Hamiltonian

$$H' = H - fq - gp$$

What does this mean, physically? Well, if you push on something with a force f , its potential energy will pick up a term $-fq$. So, the first two terms are just the Hamiltonian for a harmonic oscillator *with an extra force pushing on it!*

I don't know a nice interpretation for the $-gp$ term. We could say that besides the extra force equal to f , we also have an extra 'gorce' equal to g . I don't know what that means. Luckily, I don't need to! Mathematically, our whole problem is invariant under rotations in the position-momentum plane, so whatever works for q must also work for p .

Now here's the cool part. We can complete the square:

$$\begin{aligned} H' &= \frac{1}{2}(q^2 + p^2) - fq - gp \\ &= \frac{1}{2}(q^2 - 2fq + f^2) + \frac{1}{2}(p^2 - 2gp + g^2) - \frac{1}{2}(g^2 + f^2) \\ &= \frac{1}{2}((q - f)^2 + (p - g)^2) - \frac{1}{2}(g^2 + f^2) \end{aligned}$$

so if we define 'translated' position and momentum operators:

$$q' = q - f, \quad p' = p - g$$

we have

$$H' = \frac{1}{2}(q'^2 + p'^2) - \frac{1}{2}(g^2 + f^2)$$

So: apart from a constant, H' is just the harmonic oscillator Hamiltonian in terms of 'translated' position and momentum operators!

In other words: we're studying a strange variant of the harmonic oscillator, where we are pushing on it with an extra force and also an extra 'orce'. But this strange variant is *exactly the same as the usual harmonic oscillator*, except that we're working in translated coordinates on position-momentum space, and subtracting a constant from the Hamiltonian.

These are pretty minor differences. So, we've succeeded in reducing our problem to the problem of a harmonic oscillator in thermal equilibrium at some temperature!

This makes it easy to calculate

$$Z = \text{tr}(e^{-\beta(H-fq-gp)}) = \text{tr}(e^{-\beta H'})$$

By our formula for H' , this is just

$$Z = e^{\frac{1}{2}(g^2+f^2)} \text{tr}(e^{-\frac{1}{2}(q'^2+p'^2)})$$

And the second factor here equals the partition function for the good old harmonic oscillator:

$$Z = e^{\frac{1}{2}(g^2+f^2)} \text{tr}(e^{-\beta H})$$

So now we're back to a textbook problem. The eigenvalues of the [harmonic oscillator Hamiltonian](#) are

$$n + \frac{1}{2}$$

where

$$n = 0, 1, 2, 3, \dots$$

So, the eigenvalues of $e^{-\beta H}$ are just

$$e^{-\beta(n+\frac{1}{2})}$$

and to take the trace of this operator, we sum up these eigenvalues:

$$\text{tr}(e^{-\beta H}) = \sum_{n=0}^{\infty} e^{-\beta(n+\frac{1}{2})} = \frac{e^{-\beta/2}}{1 - e^{-\beta}}$$

So:

$$Z = e^{\frac{1}{2}(g^2+f^2)} \frac{e^{-\beta/2}}{1 - e^{-\beta}}$$

We can now compute the Fisher information metric using this formula:

$$g_{ij} = \frac{\partial^2}{\partial \lambda^i \partial \lambda^j} \ln Z$$

if we remember how our new variables are related to the λ^i :

$$\lambda^1 = \beta f, \quad \lambda^2 = \beta g, \quad \lambda^3 = \beta$$

It's just calculus! But I'm feeling a bit tired, so I'll leave this pleasure to you.

For now, I'd rather go back to our basic intuition about how the Fisher information metric describes fluctuations of observables. Mathematically, this means it's the real part of the covariance matrix

$$g_{ij} = \text{Re} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

where for us

$$X_1 = q, \quad X_2 = p, \quad X_3 = E$$

Here we are taking expected values using the mixed state ρ . We've seen this mixed state is just like the maximum-entropy state of a harmonic oscillator at fixed temperature — except for two caveats: we're working in translated coordinates on position-momentum space, and subtracting a constant from the Hamiltonian. But neither of these two caveats affects the fluctuations $(X_i - \langle X_i \rangle)$ or the covariance matrix.

So, as indeed we've already seen, g_{ij} has rotational symmetry in the 1-2 plane. Thus, we'll completely know it once we know $g_{11} = g_{22}$ and g_{33} ; the other components are zero for symmetry reasons. g_{11} will equal the variance of position for a harmonic oscillator at a given temperature, while g_{33} will equal the variance of its energy. We can work these out or look them up.

I won't do that now: I'm after insight, not formulas. For physical reasons, it's obvious that g_{11} must diminish with diminishing energy — but not go to zero. Why? Well, as the temperature approaches zero, a harmonic oscillator in thermal equilibrium approaches its state of least energy: the so-called '[ground state](#)'. In its ground state, the standard deviations of position and momentum are as small as allowed by the Heisenberg uncertainty principle:

$$\Delta p \Delta q \geq \frac{1}{2}$$

and they're equal, so

$$\$ \$ g_{11} = (\Delta q)^2 = \frac{1}{2} \$.$$

That's enough about the metric. Now, what about the metric's skew-symmetric partner? This is:

$$\omega_{ij} = \text{Im} \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

Last time we saw that ω is all about expected values of commutators:

$$\omega_{ij} = \frac{1}{2i} \langle [X_i, X_j] \rangle$$

and this makes it easy to compute. For example,

$$[X_1, X_2] = qp - pq = i$$

so

$$\$ \$ \omega_{12} = \frac{1}{2} \$. Of course$$

$$\omega_{11} = \omega_{22} = 0$$

by skew-symmetry, so we know the restriction of ω to any horizontal plane. We can also work out other components, like ω_{13} , but I don't want to. I'd rather just state this:

Summary: Restricted to any horizontal plane in the position-momentum-energy space, the Fisher information metric for the harmonic oscillator is

$$g = \text{constant}(dq_0^2 + dp_0^2)$$

with a constant depending on the temperature, equalling $\frac{1}{2}$ in the zero-temperature limit, and increasing as the temperature rises. Restricted to the same plane, the Fisher information metric's skew-symmetric partner is

$$\omega = \frac{1}{2} dq_0 \wedge dp_0$$

(Remember, the mean values q_0, p_0, E_0 are the coordinates on position-momentum-energy space. We could also use coordinates f, g, β or f, g and temperature. In the chatty intro to this article you saw formulas like those above but without the subscripts; that's before I got serious about using q and p to mean *operators*.)

And now for the moral. Actually I have two: a physics moral and a math moral.

First, what is the physical meaning of g or ω when restricted to a plane of constant E_0 , or if you prefer, a plane of constant temperature?

Physics Moral: Restricted to a constant-temperature plane, g is the covariance matrix for our observables. It is temperature-dependent. In the zero-temperature limit, the thermal fluctuations go away and g depends only on quantum fluctuations in the ground state. On the other hand, ω restricted to a constant-temperature plane describes Heisenberg uncertainty relations for noncommuting observables. In our example, it is temperature-independent.

Second, what does this have to do with [Kähler geometry?](#) Remember, the complex plane has a *complex-valued* metric on it, called a Kähler structure. Its real part is a [Riemannian metric](#), and its imaginary part is a [symplectic structure](#). We can think of the the complex plane as the position-momentum plane for a point particle. Then the symplectic structure is the basic ingredient needed for [Hamiltonian mechanics](#), while the Riemannian structure is the basic ingredient needed for the harmonic oscillator Hamiltonian.

Math Moral: In the example we considered, ω restricted to a constant-temperature plane is equal to $\frac{1}{2}$ the usual symplectic structure on the complex plane. On the other hand, g restricted to a constant-temperature plane is a multiple of the usual Riemannian metric on the complex plane — but this multiple is $\frac{1}{2}$ *only when the temperature is zero!* So, only at temperature zero are g and ω the real and imaginary parts of a Kähler structure.

It will be interesting to see how much of this stuff is true more generally. The harmonic oscillator is much nicer than your average physical system, so it can be misleading, but I think *some* of the morals we've seen here can be generalized.

Some other time I may say more about how all this is related to [Öttinger's formalism](#), but the quick point is that he too has mixed states, and a symmetric g , and a skew-symmetric ω . So it's nice to see if they match up in an example.

Finally, two footnotes on terminology:

B: In fact, this quantity $\beta = 1/kT$ is so important it deserves a better name than 'reciprocal of temperature'. How about 'coolness'? An important lesson from statistical mechanics is that coolness is more fundamental than temperature. This makes some facts more plausible. For example, if you say "you can never reach absolute zero," it sounds very odd, since you can get as close as you like, and it's even possible to get [negative temperatures](#) — but temperature zero remains tantalizingly out of reach. But "you can never attain infinite coolness" — now that makes sense.

Gorce: I apologize to Richard Feynman for stealing the word '[gorce](#)' and using it a different way. Does anyone have a good intuition for what's going on when you apply my sort of 'gorce' to a point particle? You need to think

about velocity-dependent potentials, of that I'm sure. In the presence of a velocity-dependent potential, momentum is *not* just mass times velocity. Which is good: if it were, we could never have a system where the mean value of both q and p stayed constant over time!

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



January 21, 2011

Information Geometry (Part 6)

John Baez

So far, my thread on information geometry hasn't said much about *information*. It's time to remedy that.

I've been telling you about the Fisher information metric. In statistics this is nice a way to define a 'distance' between two probability distributions. But it also has a quantum version.

So far I've showed you how to define the Fisher information metric in three equivalent ways. I also showed that in the quantum case, the Fisher information metric is the real part of a complex-valued thing. The imaginary part is related to the uncertainty principle.

But there's yet another way to define the Fisher information metric, which really involves *information*.

To explain this, I need to start with the idea of 'information gain', or 'relative entropy'. And it looks like I should do a whole post on this.

So:

Suppose that Ω is a [measure space](#) — that is, a space you can do integrals over. By a **probability distribution** on Ω , I'll mean a nonnegative function

$$p : \Omega \rightarrow \mathbf{R}$$

whose integral is 1. Here $d\omega$ is my name for the measure on Ω . Physicists might call Ω the 'phase space' of some classical system, but probability theorists might call it a space of 'events'. Today I'll use the probability theorist's language. The idea here is that

$$\int_A p(\omega) d\omega$$

gives the probability that when an event happens, it'll be one in the subset $A \subseteq \Omega$. That's why we want

$$p \geq 0$$

Probabilities are supposed to be nonnegative. And that's also why we want

$$\int_{\Omega} p(\omega) d\omega = 1$$

This says that the probability of *some* event happening is 1.

Now, suppose we have two probability distributions on Ω , say p and q . The [information gain](#) as we go from q to p is

$$S(p, q) = \int_{\Omega} p(\omega) \log\left(\frac{p(\omega)}{q(\omega)}\right) d\omega$$

We also call this the entropy of p **relative to** q . It says how much information you learn if you discover that the probability distribution of an event is p , if before you had thought it was q .

I like relative entropy because it's related to the [Bayesian interpretation of probability](#). The idea here is that you can't really 'observe' probabilities as frequencies of events, except in some unattainable limit where you repeat an experiment over and over infinitely many times. Instead, you start with some hypothesis about how likely things are: a probability distribution called the **prior**. Then you update this using [Bayes' rule](#) when you gain new information. The updated probability distribution — your new improved hypothesis — is called the **posterior**.

And if you don't do the updating right, you need a swift kick in the posterior!

So, we can think of q as the prior probability distribution, and p as the posterior. Then $S(p, q)$ measures the *amount of information* that caused you to change your views.

For example, suppose you're flipping a coin, so your set of events is just

$$\Omega = \{\text{heads, tails}\}$$

In this case all the integrals are just sums with two terms. Suppose your prior assumption is that the coin is fair. Then

$$q(\text{heads}) = 1/2, \quad q(\text{tails}) = 1/2$$

But then suppose someone you trust comes up and says "Sorry, that's a trick coin: it always comes up heads!" So you update our probability distribution and get this posterior:

$$p(\text{heads}) = 1, \quad p(\text{tails}) = 0$$

How much information have you gained? Or in other words, what's the relative entropy? It's this:

$$S(p, q) = \int_{\Omega} p(\omega) \log\left(\frac{p(\omega)}{q(\omega)}\right) d\omega = 1 \cdot \log\left(\frac{1}{1/2}\right) + 0 \cdot \log\left(\frac{0}{1/2}\right) = 1$$

Here I'm doing the logarithm in base 2, and you're supposed to know that in this game $0 \log 0 = 0$.

So: you've learned *one bit of information!*

That's supposed to make perfect sense. On the other hand, the reverse scenario takes a bit more thought.

You start out feeling sure that the coin always lands heads up. Then someone you trust says "No, that's a perfectly fair coin." If you work out the amount of information you learned this time, you'll see it's *infinite*.

Why is that?

The reason is that something that you thought was impossible — the coin landing tails up — turned out to be possible. In this game, it counts as infinitely shocking to learn something like that, so the information gain is infinite. If you hadn't been so darn sure of yourself — if you had just believed that the coin *almost always* landed heads up — your information gain would be large but finite.

The Bayesian philosophy is built into the concept of information gain, because information gain depends on two things: the prior and the posterior. And that's just as it should be: *you can only say how much you learned if you know what you believed beforehand!*

You might say that information gain depends on *three* things: p , q and the measure $d\omega$. And you'd be right! Unfortunately, the notation $S(p, q)$ is a bit misleading. Information gain really *does* depend on just two things,

but these things are not p and q : they're $p(\omega)d\omega$ and $q(\omega)d\omega$. These are called [probability measures](#), and they're ultimately more important than the probability distributions p and q .

To see this, take our information gain:

$$\int_{\Omega} p(\omega) \log\left(\frac{p(\omega)}{q(\omega)}\right) d\omega$$

and juggle it ever so slightly to get this:

$$\int_{\Omega} \log\left(\frac{p(\omega)d\omega}{q(\omega)d\omega}\right) p(\omega)d\omega$$

Clearly this depends only on $p(\omega)d\omega$ and $q(\omega)d\omega$. Indeed, it's good to work directly with these probability measures and give them short names, like

$$d\mu = p(\omega)d\omega$$

$$d\nu = q(\omega)d\omega$$

Then the formula for information gain looks more slick:

$$\int_{\Omega} \log\left(\frac{d\mu}{d\nu}\right) d\mu$$

And by the way, in case you're wondering, the d here doesn't actually mean much: we're just so brainwashed into wanting a dx in our integrals that people often use $d\mu$ for a measure even though the simpler notation μ might be more logical. So, the function

$$\frac{d\mu}{d\nu}$$

is really just a ratio of probability measures, but people call it a [Radon-Nikodym derivative](#), because it looks like a derivative (and in some important examples it actually is). So, if I were talking to myself, I could have shortened this blog entry immensely by working with directly probability measures, leaving out the d 's, and saying:

Suppose μ and ν are probability measures; then the entropy of μ relative to ν , or information gain, is
 $S(\mu, \nu) = \int_{\Omega} \log\left(\frac{\mu}{\nu}\right) \mu$.

But I'm under the impression that people are actually reading this stuff, and that most of you are happier with functions than measures. So, I decided to start with

$$S(p, q) = \int_{\Omega} p(\omega) \log\left(\frac{p(\omega)}{q(\omega)}\right) d\omega$$

and then gradually work my way up to the more sophisticated way to think about relative entropy! But having gotten that off my chest, now I'll revert to the original naive way.

As a warmup for next time, let me pose a question. How much is this quantity

$$S(p, q) = \int_{\Omega} p(\omega) \log\left(\frac{p(\omega)}{q(\omega)}\right) d\omega$$

like a *distance* between probability distributions? A distance function, or [metric](#), is supposed to satisfy some axioms. Alas, relative entropy satisfies some of these, but not the most interesting one!

- If you've got a metric, the distance between points should always be nonnegative. Indeed, this holds:

$$S(p, q) \geq 0$$

So, we never learn a negative amount when we update our prior, at least according to this definition. It's a fun exercise to prove this inequality, at least if you know some tricks involving inequalities and convex functions — otherwise it might be hard.

- If you've got a metric, the distance between two points should only be zero if they're really the same point. In fact,

$$S(p, q) = 0$$

if and only if

$$pd\omega = qd\omega$$

It's possible to have $pd\omega = qd\omega$ even if $p \neq q$, because $d\omega$ can be zero somewhere. But this is just more evidence that we should really be talking about the probability measure $pd\omega$ instead of the probability distribution p . If we do that, we're okay so far!

- If you've got a metric, the distance from your first point to your second point is the same as the distance from the second to the first. Alas,

$$S(p, q) \neq S(q, p)$$

in general. We already saw this in our example of the flipped coin. This is a slight bummer, but I could live with it, since [Lawvere has already shown](#) that it's wise to generalize the concept of metric by dropping this axiom.

- If you've got a metric, it obeys the [triangle inequality](#). This is the really interesting axiom, and alas, this too fails. Later we'll see why.

So, relative entropy does a fairly miserable job of acting like a distance function. People call it a [divergence](#). In fact, they often call it the **Kullback-Leibler divergence**. I don't like that, because 'the Kullback-Leibler divergence' doesn't really explain the idea: it sounds more like the title of a bad spy novel, sort of like *The Eiger Sanction* only worse. 'Relative entropy', on the other hand, makes a lot of sense if you understand entropy. And 'information gain' makes sense if you understand information.

Anyway: how can we save this miserable attempt to get a distance function on the space of probability distributions? *Simple: take its matrix of second derivatives and use that to define a Riemannian metric g_{ij} .* This Riemannian metric in turn defines a metric of the more elementary sort we've been discussing today.

And this Riemannian metric is the Fisher information metric I've been talking about all along!

I'll give you the details next time.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



March 2, 2011

Information Geometry (Part 7)

John Baez

Today, I want to describe how the [Fisher information metric](#) is related to [relative entropy](#). I've explained both these concepts separately (click the links for details); now I want to put them together.

But first, let me explain what this whole series of blog posts is about. Information geometry, obviously! But what's that?

Information geometry is the geometry of 'statistical manifolds'. Let me explain that concept twice: first vaguely, and then precisely.

Vaguely speaking, a statistical manifold is a [manifold](#) whose points are hypotheses about some situation. For example, suppose you have a coin. You could have various hypotheses about what happens when you flip it. For example: you could hypothesize that the coin will land heads up with probability x , where x is any number between 0 and 1. This makes the interval $[0, 1]$ into a statistical manifold. Technically this is a manifold *with boundary*, but that's okay.

Or, you could have various hypotheses about the IQ's of American politicians. For example: you could hypothesize that they're distributed according to a Gaussian probability distribution with mean x and standard deviation y . This makes the space of pairs (x, y) into a statistical manifold. Of course we require $y \geq 0$, which gives us a manifold with boundary. We might also want to assume $x \geq 0$, which would give us a manifold *with corners*, but that's okay too. We're going to be pretty relaxed about what counts as a 'manifold' here.

If we have a manifold whose points are hypotheses about some situation, we say the manifold 'parametrizes' these hypotheses. So, the concept of statistical manifold is fundamental to the subject known as [parametric statistics](#).

Parametric statistics is a huge subject! You could say that information geometry is the application of geometry to this subject.

But now let me go ahead and make the idea of 'statistical manifold' more precise. There's a classical and a quantum version of this idea. I'm working at the [Centre of Quantum Technologies](#), so I'm being paid to be quantum—but today I'm in a classical mood, so I'll only describe the classical version. Let's say a **classical statistical manifold** is a smooth function p from a manifold M to the space of probability distributions on some measure space Ω .

We should think of Ω as a space of **events**. In our first example, it's just $\{H, T\}$: we flip a coin and it lands either heads up or tails up. In our second it's \mathbf{R} : we measure the IQ of an American politician and get some real number.

We should think of M as a space of **hypotheses**. For each point $x \in M$, we have a probability distribution p_x on Ω . This is hypothesis about the events in question: for example "when I flip the coin, there's 55% chance that it will land heads up", or "when I measure the IQ of an American politician, the answer will be distributed according to a Gaussian with mean 0 and standard deviation 100."

Now, suppose someone hands you a classical statistical manifold (M, p) . Each point in M is a hypothesis. Apparently some hypotheses are more similar than others. It would be nice to make this precise. So, you might like to define a metric on M that says how 'far apart' two hypotheses are. People know lots of ways to do this; the challenge is to find ways that have clear meanings.

Last time I explained the concept of [relative entropy](#). Suppose we have two probability distributions on Ω , say p and q . Then the **entropy of p relative to q** is the amount of information you gain when you start with the hypothesis q but then discover that you should switch to the new improved hypothesis p . It equals:

$$\int_{\Omega} \frac{p}{q} \ln \left(\frac{p}{q} \right) q d\omega$$

You could try to use this to define a distance between points x and y in our statistical manifold, like this:

$$S(x, y) = \int_{\Omega} \frac{p_x}{p_y} \ln \left(\frac{p_x}{p_y} \right) p_y d\omega$$

This is definitely an important function. Unfortunately, as I explained [last time](#), it doesn't obey the axioms that a distance function should! Worst of all, it doesn't obey the triangle inequality.

Can we 'fix' it? Yes, we can! And when we do, we get the Fisher information metric, which is actually a *Riemannian* metric on M . Suppose we put local coordinates on some patch of M containing the point x . Then the **Fisher information metric** is given by:

$$g_{ij}(x) = \int_{\Omega} \partial_i(\ln p_x) \partial_j(\ln p_x) p_x d\omega$$

You can think of my whole series of articles so far as an attempt to understand this funny-looking formula. I've shown how to get it from a few different starting-points, most recently back in [Part 3](#). But now let's get it starting from relative entropy!

Fix any point in our statistical manifold and choose local coordinates for which this point is the origin, 0. The amount of information we gain if move to some other point x is the relative entropy $S(x, 0)$. But what's this like when x is really close to 0? We can imagine doing a Taylor series expansion of $S(x, 0)$ to answer this question.

Surprisingly, to first order the answer is always zero! Mathematically:

$$\partial_i S(x, 0)|_{x=0} = 0$$

In plain English: if you change your mind slightly, you learn a negligible amount — *not* an amount proportional to how much you changed your mind.

This must have some profound significance. I wish I knew what. Could it mean that people are reluctant to change their minds except in big jumps?

Anyway, if you think about it, this fact makes it obvious that $S(x, y)$ can't obey the triangle inequality. $S(x, y)$ could be pretty big, but if we draw a curve from x and y , and mark n closely spaced points x_i on this curve, then $S(x_{i+1}, x_i)$ is zero to first order, so it must be of order $1/n^2$, so if the triangle inequality were true we'd have

$$S(x, y) \leq \sum_i S(x_{i+1}, x_i) \leq \text{const } n \cdot \frac{1}{n^2}$$

for all n , which is a contradiction.

In plain English: if you change your mind in one big jump, the amount of information you gain is more than the sum of the amounts you'd gain if you change your mind in lots of little steps! This seems pretty darn strange, but the paper I mentioned in [part 1](#) helps:

- Gavin E. Crooks, [Measuring thermodynamic length](#).

You'll see he takes a curve and chops it into lots of little pieces as I just did, and explains what's going on.

Okay, so what about second order? What's

$$\partial_i \partial_j S(x, 0)|_{x=0}?$$

Well, this is the punchline of this blog post: *it's the Fisher information metric*:

$$\partial_i \partial_j S(x, 0)|_{x=0} = g_{ij}$$

And since the Fisher information metric is a [Riemannian metric](#), we can then apply [the usual recipe](#) and define distances in a way that obeys the triangle inequality. Crooks calls this distance **thermodynamic length** in the special case that he considers, and he explains its physical meaning.

Now let me prove that

$$\partial_i S(x, 0)|_{x=0} = 0$$

and

$$\partial_i \partial_j S(x, 0)|_{x=0} = g_{ij}$$

This can be somewhat tedious if you do it by straightforwardly grinding it out—I know, I did it. So let me show you a better way, which requires more conceptual acrobatics but less brute force.

The trick is to work with the **universal** statistical manifold for the measure space Ω . Namely, we take M to be the space of *all* probability distributions on Ω ! This is typically an *infinite-dimensional* manifold, but that's okay: we're being relaxed about what counts as a manifold here. In this case, we don't need to write p_x for the probability distribution corresponding to the point $x \in M$. In this case, a point of M just *is* a probability distribution on Ω , so we'll just call it p .

If we can prove the formulas for this universal example, they'll automatically follow for every other example, by abstract nonsense. Why? Because *any* statistical manifold with measure space Ω is the same as a manifold with a smooth map to the *universal* statistical manifold! So, geometrical structures on the universal one ['pull back'](#) to give structures on all the rest. The Fisher information metric and the function S can be defined as pullbacks in this way! So, to study them, we can just study the universal example.

(If you're familiar with 'classifying spaces for bundles' or other sorts of 'classifying spaces', all this should seem awfully familiar. It's a standard math trick.)

So, let's prove that

$$\partial_i S(x, 0)|_{x=0} = 0$$

by proving it in the universal example. Given any probability distribution q , and taking a nearby probability distribution p , we can write

$$\frac{p}{q} = 1 + f$$

where f is some small function. We only need to show that $S(p, q)$ is zero to first order in f . And this is pretty easy. By definition:

$$S(p, q) = \int_{\Omega} \frac{p}{q} \ln \left(\frac{p}{q} \right) q d\omega$$

or in other words,

$$S(p, q) = \int_{\Omega} (1+f) \ln(1+f) q d\omega$$

We can calculate this to first order in f and show we get zero. But let's actually work it out to second order, since we'll need that later:

$$\ln(1+f) = f - \frac{1}{2}f^2 + \dots$$

so

$$(1+f) \ln(1+f) = f + \frac{1}{2}f^2 + \dots$$

so

$$\begin{aligned} S(p, q) &= \int_{\Omega} (1+f) \ln(1+f) q d\omega \\ &= \int_{\Omega} f q d\omega + \frac{1}{2} \int_{\Omega} f^2 q d\omega + \dots \end{aligned}$$

Why does this vanish to first order in f ? It's because p and q are both probability distributions and $p/q = 1+f$, so

$$\int_{\Omega} (1+f) q d\omega = \int_{\Omega} p d\omega = 1$$

but also

$$\int_{\Omega} q d\omega = 1$$

so subtracting we see

$$\int_{\Omega} f q d\omega = 0$$

So, $S(p, q)$ vanishes to first order in f . *Voilà!*

Next let's prove the more interesting formula:

$$\partial_i \partial_j S(x, 0)|_{x=0} = g_{ij}$$

which relates relative entropy to the Fisher information metric. Since both sides are symmetric matrices, it suffices to show their diagonal entries agree in any coordinate system:

$$\partial_i^2 S(x, 0)|_{x=0} = g_{ii}$$

Devoted followers of this series of posts will note that I keep using this trick, which takes advantage of the [polarization identity](#).

To prove

$$\partial_i^2 S(x, 0)|_{x=0} = g_{ii}$$

it's enough to consider the universal example. We take the origin to be some probability distribution q and take x to be a nearby probability distribution p which is pushed a tiny bit in the i th coordinate direction. As before we write $p/q = 1 + f$. We look at the second-order term in our formula for $S(p, q)$:

$$\frac{1}{2} \int_{\Omega} f^2 q d\omega$$

Using the usual second-order Taylor's formula, which has a $\frac{1}{2}$ built into it, we can say

$$\partial_i^2 S(x, 0)|_{x=0} = \int_{\Omega} f^2 q d\omega$$

On the other hand, our formula for the Fisher information metric gives

$$g_{ii} = \int_{\Omega} \partial_i \ln p \partial_i \ln p q d\omega \Big|_{p=q}$$

The right hand sides of the last two formulas look awfully similar! And indeed they agree, because we can show that

$$\partial_i \ln p|_{p=q} = f$$

How? Well, we assumed that p is what we get by taking q and pushing it a little bit in the i th coordinate direction; we have also written that little change as

$$p/q = 1 + f$$

for some small function f . So,

$$\partial_i(p/q) = f$$

and thus:

$$\partial_i p = f q$$

and thus:

$$\partial_i \ln p = \frac{\partial_i p}{p} = \frac{f q}{p}$$

so

$$\partial_i \ln p|_{p=q} = f$$

as desired.

This argument may seem a little hand-wavy and nonrigorous, with words like 'a little bit'. If you're used to taking arguments involving infinitesimal changes and translating them into calculus (or differential geometry), it should

make sense. If it doesn't, I apologize. It's easy to make it more rigorous, but only at the cost of more annoying notation, which doesn't seem good in a blog post.

Boring technicalities

If you're actually the kind of person who reads a section called 'boring technicalities', I'll admit to you that my calculations don't make sense if the integrals diverge, or we're dividing by zero in the ratio p/q . To avoid these problems, here's what we should do. Fix a [σ-finite](#) measure space $(\Omega, d\omega)$. Then, define the **universal statistical manifold** to be the space $P(\Omega, d\omega)$ consisting of all probability measures that are [equivalent](#) to $d\omega$, in the usual sense of measure theory. By [Radon-Nikodym](#), we can write any such measure as $qd\omega$ where $q \in L^1(\Omega, d\omega)$. Moreover, given two of these guys, say $pd\omega$ and $qd\omega$, they are [absolutely continuous](#) with respect to each other, so we can write

$$pd\omega = \frac{p}{q} qd\omega$$

where the ratio p/q is well-defined almost everywhere and lies in $L^1(\Omega, qd\omega)$. This is enough to guarantee that we're never dividing by zero, and I think it's enough to make sure all my integrals converge.

We do still need to make $P(\Omega, d\omega)$ into some sort of infinite-dimensional manifold, to justify all the derivatives. There are various ways to approach this issue, all of which start from the fact that $L^1(\Omega, d\omega)$ is a [Banach space](#), which is about the nicest sort of infinite-dimensional manifold one could imagine. Sitting in $L^1(\Omega, d\omega)$ is the hyperplane consisting of functions q with

$$\int_{\Omega} qd\omega = 1$$

and this is a [Banach manifold](#). To get $P(\Omega, d\omega)$ we need to take a subspace of that hyperplane. If this subspace were open then $P(\Omega, d\omega)$ would be a Banach manifold in its own right. I haven't checked this yet, for various reasons.

For one thing, there's a nice theory of '[diffeological spaces](#)', which generalize manifolds. Every Banach manifold is a diffeological space, and every subset of a diffeological space is again a diffeological space. For many purposes we don't need our 'statistical manifolds' to be manifolds: diffeological spaces will do just fine. This is one reason why I'm being pretty relaxed here about what counts as a 'manifold'.

For another, I know that people have worked out a lot of this stuff, so I can just look things up when I need to. And so can you! This book is a good place to start:

- Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin and Henry P. Wynn, *Algebraic and Geometric Methods in Statistics*, Cambridge U. Press, Cambridge, 2009.

I find the chapters by Raymond Streater especially congenial. For the technical issue I'm talking about now it's worth reading section 14.2, "Manifolds modelled by Orlicz spaces", which tackles the problem of constructing a universal statistical manifold in a more sophisticated way than I've just done. And in chapter 15, "The Banach manifold of quantum states", he tackles the quantum version!

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



May 26, 2011

Information Geometry (Part 8)

John Baez

Now this series on information geometry will take an unexpected turn toward 'green mathematics'. Lately I've been talking about relative entropy. Now I'll say how this concept shows up in the study of evolution!

That's an unexpected turn to me, at least. I learned of this connection just two days ago in a conversation with [Marc Harper](#), a mathematician who is a postdoc in bioinformatics at UCLA, working with my friend [Chris Lee](#). I was visiting Chris for a couple of days after attending the thesis defenses of some grad students of mine who just finished up at U.C. Riverside. Marc came by and told me about this paper:

- Marc Harper, [Information geometry and evolutionary game theory](#).

and now I can't resist telling you.

First of all: what does information theory have to do with biology? Let me start with a very general answer: biology is different from physics because biological systems are packed with information you can't afford to ignore.

Physicists love to think about systems that take only a little information to describe. So when they get a system that takes a *lot* of information to describe, they use a trick called 'statistical mechanics', where you try to ignore most of this information and focus on a few especially important variables. For example, if you hand a physicist a box of gas, they'll try to avoid thinking about the state of each atom, and instead focus on a few macroscopic quantities like the volume and total energy. Ironically, the mathematical concept of information arose first here—although they didn't call it information back then; they called it 'entropy'. The entropy of a box of gas is precisely the amount of information you've decided to forget when you play this trick of focusing on the macroscopic variables. Amazingly, remembering just this—the sheer *amount* of information you've forgotten—can be extremely useful... at least for the systems physicists like best.

But biological systems are different. They store lots of information (for example in DNA), transmit lots of information (for example in the form of biochemical signals), and collect a lot of information from their environment. And this information isn't uninteresting 'noise', like the positions of atoms in a gas. The details really matter. Thus, we need to keep track of lots of information to have a chance of understanding any particular biological system.

So, part of doing biology is developing new ways to think about physical systems that contain lots of *relevant* information. This is why physicists consider biology 'messy'. It's also why biology and computers go hand in hand in the subject called 'bioinformatics'. There's no avoiding this: in fact, it will probably force us to *automate the scientific method!* That's what Chris Lee and Marc Harper are really working on:

- Chris Lee, [General information metrics for automated experiment planning](#), presentation in the UCLA Chemistry & Biochemistry Department faculty luncheon series, 2 May 2011.

But more about that some other day. Let me instead give *another* answer to the question of what information theory has to do with biology.

There's an analogy between evolution and the scientific method. Simply put, life is an experiment to see what works; natural selection weeds out the bad guesses, and over time the better guesses predominate. This process transfers information from the world to the 'experimenter': the species that's doing the evolving, or the scientist. Indeed, the only way the experimenter can get information is by making guesses that can be wrong.

All this is simple enough, but the nice thing is that we can make it more precise.

On the one hand, there's a simple model of the scientific method called '[Bayesian inference](#)'. Assume there's a set of mutually exclusive alternatives: possible ways the world can be. And suppose we start with a '[prior probability distribution](#)': a preconceived notion of how probable each alternative is. Say we do an experiment and get a result that depends on which alternative is true. We can work out how likely this result was given our prior, and—using a marvelously simple formula called [Bayes' rule](#)—we can use this to update our prior and obtain a new improved probability distribution, called the '[posterior probability distribution](#)'.

On the other hand, suppose we have a species with several different possible genotypes. A population of this species will start with some number of organisms with each genotype. So, we get a probability distribution saying how likely it is that an organism has any given genotype. These genotypes are our 'mutually exclusive alternatives', and this probability distribution is our 'prior'. Suppose each generation the organisms have some expected number of offspring that depends on their genotype. Mathematically, it turns out this is just like updating our prior using Bayes' rule! The result is a new probability distribution of genotypes: the 'posterior'.

I learned about this from Chris Lee on the 19th of December, 2006. In my [diary](#) that day, I wrote:

The analogy is mathematically precise, and fascinating. In rough terms, it says that *the process of natural selection resembles the process of Bayesian inference*. A population of organisms can be thought of as having various 'hypotheses' about how to survive—each hypothesis corresponding to a different [allele](#). (Roughly, an allele is one of several alternative versions of a gene.) In each successive generation, the process of natural selection modifies the proportion of organisms having each hypothesis, according to Bayes' rule!

Now let's be more precise:

[Bayes' rule](#) says if we start with a 'prior probability' for some hypothesis to be true, divide it by the probability that some observation is made, then multiply by the 'conditional probability' that this observation will be made given that the hypothesis is true, we'll get the 'posterior probability' that the hypothesis is true *given that the observation is made*.

Formally, the exact same equation shows up in population genetics! In fact, Chris showed it to me—it's equation 9.2 on page 30 of this book:

- R. Bürger, *The Mathematical Theory of Selection, Recombination and Mutation*, section I.9: Selection at a single locus, Wiley, 2000.

But, now all the terms in the equation have different meanings!

Now, instead of a 'prior probability' for a hypothesis to be true, we have the frequency of occurrence of some [allele](#) in some generation of a population. Instead of the probability that we make some observation, we have the expected number of offspring of an organism. Instead of the 'conditional probability' of making the observation, we have the expected number of offspring of an organism *given* that it has this allele. And, instead of the 'posterior probability' of our hypothesis, we have the frequency of occurrence of that allele in the next generation.

(Here we are assuming, for simplicity, an asexually reproducing 'haploid' population - that is, one with just a single set of chromosomes.)

This is a great idea—Chris felt sure someone must have already had it. A natural context would be research on [genetic programming](#), a machine learning technique that uses an evolutionary algorithm to optimize a population of computer programs according to a fitness landscape determined by their ability to perform a given task. Since there has also been a lot of work on Bayesian approaches to machine learning, surely someone has noticed their mathematical relationship?

I see that [Amy Perfors](#) found these ideas as new and exciting as I did. But I still can't believe Chris was the first to clearly formulate them, so I'd still like to know who did.

Marc Harper actually went to work with Chris after reading that diary entry of mine. By now he's gone a lot further with this analogy by focusing on the role of *information*. As we keep updating our prior using Bayes' rule, we should be gaining information about the real world. This idea has been made very precise in the theory of '[machine learning](#)'. Similarly, as a population evolves through natural selection, it should be gaining information about its environment.

I've been talking about Bayesian updating as a discrete-time process: something that happens once each generation for our population. That's fine and dandy, definitely worth studying, but Marc's paper focuses on a continuous-time version called the '[replicator equation](#)'. It goes like this. Let X be the set of alternative genotypes. For each $i \in X$, let P_i be the number of organisms that have the i th genotype at time t . Say that

$$\frac{dP_i}{dt} = f_i P_i$$

where f_i is the **fitness** of the i th genotype. Let p_i be the probability that at time t , a randomly chosen organism will have the i th genotype:

$$p_i = \frac{P_i}{\sum_{i \in X} P_i}$$

Then a little calculus gives the **replicator equation**:

$$\frac{dp_i}{dt} = (f_i - \langle f \rangle) p_i$$

where

$$\langle f \rangle = \sum_{i \in X} f_i p_i$$

is the **mean fitness** of the organisms. So, the fraction of organisms of the i th type grows at a rate proportional to the fitness of that type *minus the mean fitness*. It ain't enough to be good: you gotta be better than average.

Note that all this works not just when each fitness f_i is a mere number, but also when it's a function of the whole list of probabilities p_i . That's good, because in the real world, the fitness of one kind of bug may depend on the fraction of bugs of various kinds.

But what does all this have to do with *information*?

Marc's paper has a lot to say about this! But just to give you a taste, here's a simple fact involving relative entropy, which was first discovered by Ethan Atkin. Suppose evolution as described by the replicator equation brings the whole list of probabilities p_i —let's call this list p —closer and closer to some stable equilibrium, say q . Then if a couple of technical conditions hold, the entropy of q relative to p keeps decreasing, and approaches zero.

Remember what I told you about [relative entropy](#). In Bayesian inference, the entropy q relative to p is how much information we gain if we start with p as our prior and then do an experiment that pushes us to the posterior q . So, in simple rough terms: ***as it approaches a stable equilibrium, the amount of information a species has left to learn keeps dropping, and goes to zero!***

I won't fill in the precise details, because I bet you're tired already. You can find them in Section 3.5, which is called "Kullback-Leibler Divergence is a Lyapunov function for the Replicator Dynamic". If you know all the buzzwords here, you'll be in buzzword heaven now. 'Kullback-Leibler divergence' is just another term for relative entropy. 'Lyapunov function' means that it keeps dropping and goes to zero. And the 'replicator dynamic' is the replicator equation I described above.

Perhaps next time I'll say more about this stuff. For now, I just hope you see why it makes me so happy.

First, it uses information geometry to make precise the sense in which evolution is a process of acquiring information. That's very cool. We're looking at a simplified model—the replicator equation—but doubtless this is just the beginning of a very long story that keeps getting deeper as we move to less simplified models.

Second, if you read my summary of [Chris Canning's talks on evolutionary game theory](#), you'll see everything I just said meshes nicely with that. He was taking the fitness f_i to be

$$f_i = \sum_{j \in X} A_{ij} p_j$$

where the **payoff matrix** A_{ij} describes the 'winnings' of an organism with the i th genotype when it meets an organism with the j th genotype. This gives a particularly nice special case of the replicator equation.

Third, this particularly nice special case happens to be the [rate equation](#) for a certain stochastic Petri net. So, we've succeeded in connecting the 'diagram theory' discussion to the 'information geometry' discussion! This has all sort of implications, which will take quite a while to explore.

As the saying goes, in mathematics:

Everything sufficiently beautiful is connected to all other beautiful things.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2011 John Baez

baez@math.removethis.ucr.andthis.edu

[home](#)

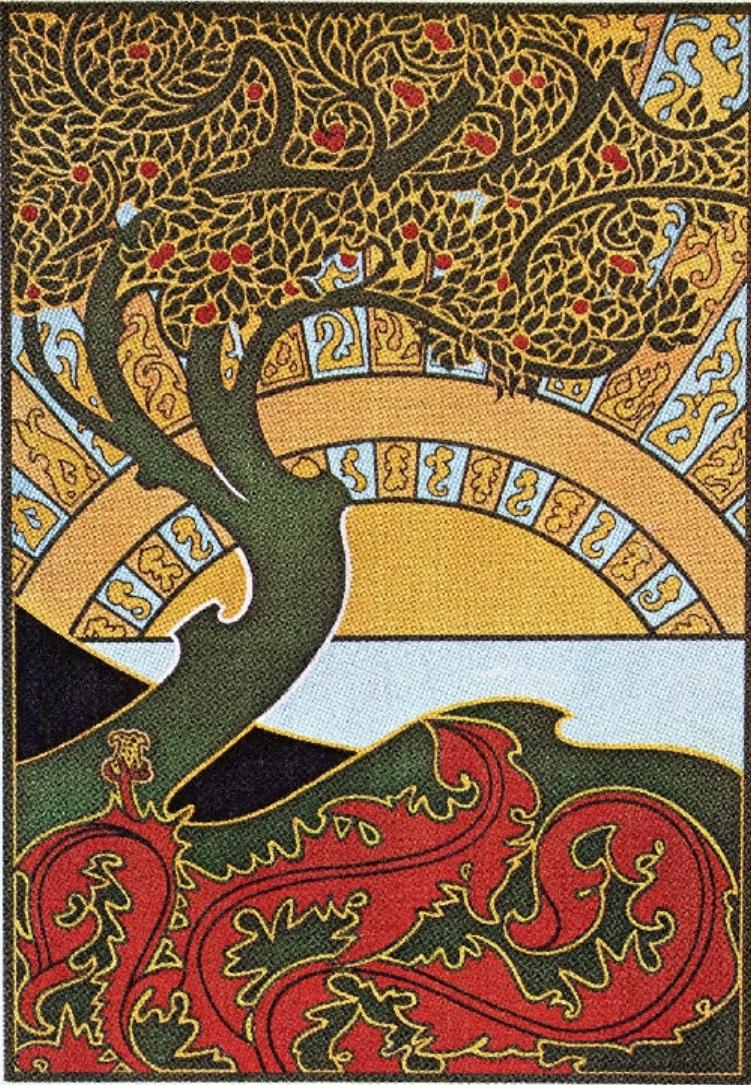




June 1, 2012

Information Geometry (Part 9)

John Baez



www.crm.cat/CBIO :: Scientific queries : Tom Leinster : tom.leinster@glasgow.ac.uk :: Administrative queries : Nlhorst@crm.cat

Committee :: Benjamin Allen : Silvia Cusimano : Tom Leinster (coordinator) : Richard Reeve : John Wooliams

CRM

THE MATHEMATICS OF BIODIVERSITY
Centre de Recerca Matemàtica, Barcelona
2-6 July 2012

Speakers (provisional) :: Benjamin Allen : John Baez : Michael Bonsall : Anne Chao
Christina Cobbold : Glenn De'ath : Elizabeth Gillet : Hans-Rolf Gregorius : Lou Jost
Tom Leinster : Alison Mather : Louise Matthews : Hans Metz : Sandrine Pavoine
Richard Reeve : Carlo Ricotta : William Sherwin : John Wooliams

BBSRC

It's time to continue this [information geometry](#) series, because I've promised to give the following talk at a [conference on the mathematics of biodiversity](#) in early July... and I still need to do some of the research! 😊

Diversity, information geometry and learning

As is well known, some measures of biodiversity are formally identical to measures of information developed by Shannon and others. Furthermore, Marc Harper has shown that the replicator equation in evolutionary game theory is formally identical to a process of Bayesian inference, which is studied in the field of machine learning using ideas from information geometry. Thus, in this simple model, a population of organisms can be thought of as a 'hypothesis' about how to survive, and natural selection acts to update this hypothesis according to Bayes' rule. The question thus arises to what extent natural changes in biodiversity can be usefully seen as analogous to a form of learning. However, some of the same mathematical structures arise in the study of chemical reaction networks, where the increase of entropy, or more precisely decrease of free energy, is not usually considered a form of 'learning'. We report on some preliminary work on these issues.

So, let's dive in! To some extent I'll be explaining these two papers:

- Marc Harper, [Information geometry and evolutionary game theory](#).
- Marc Harper, [The replicator equation as an inference dynamic](#).

However, I hope to bring in some more ideas from physics, the study of biodiversity, and the theory of stochastic Petri nets, also known as chemical reaction networks. So, this series may start to overlap with my [network theory](#) posts. We'll see. We won't get far today: for now, I just want to review and expand on what we did [last time](#).

The replicator equation

The [replicator equation](#) is a simplified model of how populations change. Suppose we have n types of self-replicating entity. I'll call these entities **replicators**. I'll call the types of replicators **species**, but they don't need to be species in the biological sense. For example, the replicators could be genes, and the types could be [alleles](#). Or the replicators could be restaurants, and the types could be restaurant chains.

Let $P_i(t)$, or just P_i for short, be the population of the i th species at time t . Then the replicator equation says

$$\frac{dP_i}{dt} = f_i(P_1, \dots, P_n) P_i$$

So, the population P_i changes at a rate proportional to P_i , but the 'constant of proportionality' need not be constant: it can be any smooth function f_i of the populations of all the species. We call $f_i(P_1, \dots, P_n)$ the [fitness](#) of the i th species.

Of course this model is absurdly general, while still leaving out lots of important effects, like the spatial variation of populations, or the ability for the population of some species to start at zero and become nonzero—which happens thanks to mutation. Nonetheless this model is worth taking a good look at.

Using the magic of vectors we can write

$$P = (P_1, \dots, P_n)$$

and

$$f(P) = (f_1(P), \dots, f_n(P))$$

This lets us write the replicator equation a wee bit more tersely as

$$\frac{dP}{dt} = f(P)P$$

where on the right I'm multiplying vectors componentwise, the way your teachers tried to brainwash you into never doing:

$$f(P)P = (f(P)_1 P_1, \dots, f(P)_n P_n)$$

In other words, I'm thinking of P and $f(P)$ as functions on the set $\{1, \dots, n\}$ and multiplying them pointwise. This will be a nice way of thinking if we want to replace this finite set by some more general space.

Why would we want to do that? Well, we might be studying lizards with different length tails, and we might find it convenient to think of the set of possible tail lengths as the half-line $[0, \infty)$ instead of a finite set.

Or, just to get started, we might want to study the pathetically simple case where $f(P)$ doesn't depend on P . Then we just have a fixed function f and a time-dependent function P obeying

$$\frac{dP}{dt} = fP$$

If we're physicists, we might write P more suggestively as ψ and write the operator multiplying by f as $-H$. Then our equation becomes

$$\frac{d\psi}{dt} = -H\psi$$

This looks a lot like Schrödinger's equation, but since there's no factor of $\sqrt{-1}$, and ψ is real-valued, it's more like the heat equation or the 'master equation', the basic equation of stochastic mechanics.

For an explanation of Schrödinger's equation and the master equation, try [Part 12](#) of the network theory series. In that post I didn't include a minus sign in front of the H . That's no big deal: it's just a different convention than the one I want today. A more serious issue is that in stochastic mechanics, ψ stands for a *probability distribution*. This suggests that we should get probabilities into the game somehow.

The replicator equation in terms of probabilities

Luckily, that's exactly what people usually do! Instead of talking about the population P_i of the i th species, they talk about the *probability* p_i that one of our organisms will belong to the i th species. This amounts to normalizing our populations:

$$p_i = \frac{P_i}{\sum_j P_j}$$

Don't you love it when notations work out well? Our big Population P_i has gotten normalized to give little probability p_i .

How do these probabilities p_i change with time? Now is the moment for that least loved rule of elementary calculus to come out and take a bow: the quotient rule for derivatives!

$$\frac{dp_i}{dt} = \left(\frac{dP_i}{dt} \sum_j P_j - P_i \sum_j \frac{dP_j}{dt} \right) / \left(\sum_j P_j \right)^2$$

Using our earlier version of the replicator equation, this gives:

$$\frac{dp_i}{dt} = \left(f_i(P)P_i \sum_j P_j - P_i \sum_j f_j(P)P_j \right) / \left(\sum_j P_j \right)^2$$

Using the definition of p_i , this simplifies to:

$$\frac{dp_i}{dt} = f_i(P)p_i - \left(\sum_j f_j(P)p_j \right) p_i$$

The stuff in parentheses actually has a nice meaning: it's just the **mean fitness**. In other words, it's the average, or expected, fitness of an organism chosen at random from the whole population. Let's write it like this:

$$\langle f(P) \rangle = \sum_j f_j(P)p_j$$

So, we get the [replicator equation](#) in its classic form:

$$\frac{dp_i}{dt} = \left(f_i(P) - \langle f(P) \rangle \right) p_i$$

This has a nice meaning: for the fraction of organisms of the i th type to increase, their fitness must exceed the mean fitness. If you're trying to increase [market share](#), what matters is not how good you are, but how much *better than average* you are. If everyone else is lousy, you're in luck.

Entropy

Now for something a bit new. Once we've gotten a probability distribution into the game, its entropy is sure to follow:

$$S(p) = - \sum_i p_i \ln(p_i)$$

This says how 'smeared-out' the overall population is among the various different species. Alternatively, it says how much *information* it takes, on average, to say which species a randomly chosen organism belongs to. For example, if there are 2^N species, all with equal populations, the entropy S works out to $N \ln 2$. So in this case, it takes N bits of information to say which species a randomly chosen organism belongs to.

In biology, entropy is one of many ways people measure biodiversity. For a quick intro to some of the issues involved, try:

- Tom Leinster, [Measuring biodiversity](#), *Azimuth*, 7 November 2011.
- Lou Jost, [Entropy and diversity](#), *Oikos* **113** (2006), 363--375.

But we don't need to understand this stuff to see how entropy is connected to the replicator equation. Marc Harper's paper explains this in detail:

- Marc Harper, [The replicator equation as an inference dynamic](#).

and I hope to go through quite a bit of it here. But not today! Today I just want to look at a pathetically simple, yet still interesting, example.

Exponential growth

Suppose the fitness of each species is independent of the populations of all the species. In other words, suppose each fitness $f_i(P)$ is actually a constant, say f_i . Then the replicator equation reduces to

$$\frac{dP_i}{dt} = f_i P_i$$

so it's easy to solve:

$$P_i(t) = e^{f_i t} P_i(0)$$

You don't need a detailed calculation to see what's going to happen to the probabilities

$$p_i(t) = \frac{P_i(t)}{\sum_j P_j(t)}$$

The most fit species present will eventually take over! If one species, say the i th one, has a fitness greater than the rest, then the population of this species will eventually grow faster than all the rest, at least if its population starts out greater than zero. So as $t \rightarrow +\infty$, we'll have

$$p_i(t) \rightarrow 1$$

and

$$p_j(t) \rightarrow 0 \quad \text{for } j \neq i$$

Thus the probability distribution p will become more sharply peaked, and *its entropy will eventually approach zero.*

With a bit more thought you can see that even if more than one species shares the maximum possible fitness, the entropy will eventually decrease, though not approach zero.

In other words, *the biodiversity will eventually drop* as all but the most fit species are overwhelmed. Of course, this is only true in our simple idealization. In reality, biodiversity behaves in more complex ways—partly because species interact, and partly because mutation tends to smear out the probability distribution p_i . We're not looking at these effects yet. They're extremely important... in ways we can only fully understand if we start by looking at what happens when they're not present.

In still other words, *the population will absorb information from its environment*. This should make intuitive sense: the process of natural selection resembles 'learning'. As fitter organisms become more common and less fit ones die out, the environment puts its stamp on the probability distribution p . So, this probability distribution should gain information.

While intuitively clear, this last claim also follows more rigorously from thinking of entropy as negative information. Admittedly, it's always easy to get confused by minus signs when relating entropy and information. A while back I said the entropy

$$S(p) = - \sum_i p_i \ln(p_i)$$

was the average information required to say which species a randomly chosen organism belongs to. If this entropy is going down, isn't the population *losing* information?

No, this is a classic sign error. It's like the concept of 'work' in physics. We can talk about the work some system does on its environment, or the work done by the environment on the system, and these are almost the same... *except one is minus the other!*

When you are very ignorant about some system—say, some rolled dice—your estimated probabilities p_i for its various possible states are very smeared-out, so the entropy $S(p)$ is large. As you gain information, you revise your probabilities and they typically become more sharply peaked, so $S(p)$ goes down. When you know as much as you possibly can, $S(p)$ equals zero.

So, the entropy $S(p)$ is the amount of information you have left to learn: the amount of information you *lack*, not the amount you *have*. As you gain information, this goes down. There's no paradox here.

It works the same way with our population of replicators—at least in the special case where the fitness of each species is independent of its population. The probability distribution p is like a 'hypothesis' assigning to each species i the probability p_i that it's the best at self-replicating. As some replicators die off while others prosper, they gather information their environment, and this hypothesis gets refined. So, the entropy $S(p)$ drops.

Next time

Of course, to make closer contact to reality, we need to go beyond the special case where the fitness of each species is a constant! Marc Harper does this, and I want to talk about his work someday, but first I have a few more remarks to make about the pathetically simple special case I've been focusing on. I'll save these for next time, since I've probably strained your patience already.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2012 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)



June 4, 2012

Information Geometry (Part 10)

John Baez

Last time I began explaining the tight relation between three concepts:

- entropy,
- information—or more precisely, lack of information,

and

- biodiversity.

The idea is to consider n different species of 'replicators'. A replicator is any entity that can reproduce itself, like an organism, a gene, or a meme. A replicator can come in different kinds, and a 'species' is just our name for one of these kinds. If P_i is the population of the i th species, we can interpret the fraction

$$p_i = \frac{P_i}{\sum_j P_j}$$

as a probability: the probability that a randomly chosen replicator belongs to the i th species. This suggests that we define entropy just as we do in statistical mechanics:

$$S = - \sum_i p_i \ln(p_i)$$

In the study of statistical inference, entropy is a measure of uncertainty, or lack of information. But now we can interpret it as a measure of biodiversity: it's zero when just one species is present, and small when a few species have much larger populations than all the rest, but gets big otherwise.

Our goal here is play these viewpoints off against each other. In short, we want to think of natural selection, and even biological evolution, as a process of statistical inference—or in simple terms, *learning*.

To do this, let's think about how entropy changes with time. Last time we introduced a simple model called the replicator equation:

$$\frac{dP_i}{dt} = f_i(P_1, \dots, P_n) P_i$$

where each population grows at a rate proportional to some 'fitness functions' f_i . We can get some intuition by looking at the pathetically simple case where these functions are actually *constants*, so

$$\frac{dP_i}{dt} = f_i P_i$$

The equation then becomes trivial to solve:

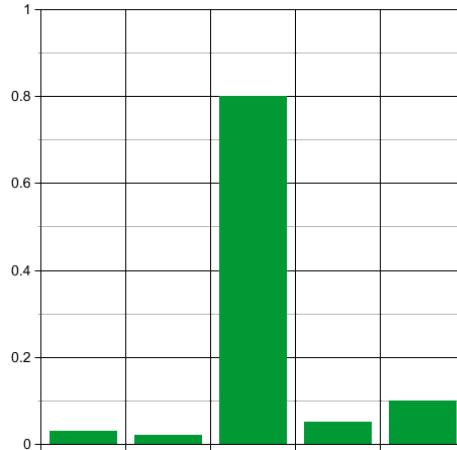
$$P_i(t) = e^{tf_i} P_i(0)$$

Last time I showed that in this case, the entropy will eventually decrease. It will go to zero as $t \rightarrow +\infty$ whenever one species is fitter than all the rest and starts out with a nonzero population—since then this species will eventually take over.

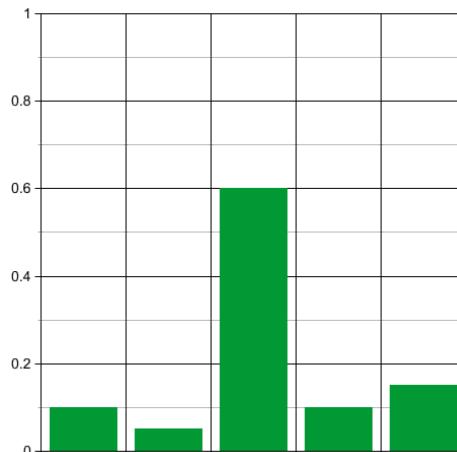
But remember, the entropy of a probability distribution is its *lack* of information. So the decrease in entropy signals an increase in information. And last time I argued that this makes perfect sense. As the fittest species takes over and biodiversity drops, *the population is acquiring information about its environment*.

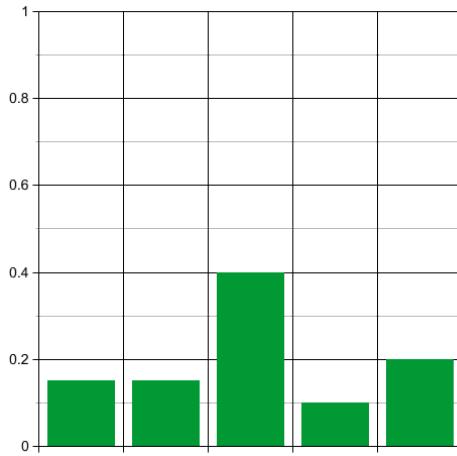
However, I never said the entropy is *always* decreasing, because that's false! Even in this pathetically simple case, entropy can increase.

Suppose we start with many replicators belonging to one very unfit species, and a few belonging to various more fit species. The probability distribution p_i will start out sharply peaked, so the entropy will start out low:



Now think about what happens when time passes. At first the unfit species will rapidly die off, while the population of the other species slowly grows:





So the probability distribution will, for a while, become less sharply peaked. Thus, *for a while*, the entropy will increase!

This seems to conflict with our idea that the population's entropy should decrease as it acquires information about its environment. But in fact this phenomenon is familiar in the study of statistical inference. If you start out with strongly held *false* beliefs about a situation, the first effect of learning more is to become *less* certain about what's going on!

Get it? Say you start out by assigning a high probability to some wrong guess about a situation. The entropy of your probability distribution is low: you're quite certain about what's going on. But you're wrong. When you first start suspecting you're wrong, you become more uncertain about what's going on. Your probability distribution flattens out, and the entropy goes up.

So, sometimes learning involves a decrease in information—*false* information. There's nothing about the mathematical concept of information that says this information is *true*.

Given this, it's good to work out a formula for the rate of change of entropy, which will let us see more clearly when it goes down and when it goes up. To do this, first let's derive a completely general formula for the time derivative of the entropy of a probability distribution. Following Sir Isaac Newton, we'll use a dot to stand for a time derivative:

$$\begin{aligned}\dot{S} &= -\frac{d}{dt} \sum_i p_i \ln(p_i) \\ &= -\sum_i \dot{p}_i \ln(p_i) + \dot{p}_i\end{aligned}$$

In the last term we took the derivative of the logarithm and got a factor of $1/p_i$ which cancelled the factor of p_i . But since

$$\sum_i p_i = 1$$

we know

$$\sum_i \dot{p}_i = 0$$

so this last term vanishes:

$$\dot{S} = - \sum_i \dot{p}_i \ln(p_i)$$

Nice! To go further, we need a formula for \dot{p}_i . For this we might as well return to the general replicator equation, dropping the pathetically special assumption that the fitness functions are actually constants. Then we saw last time that

$$\dot{p}_i = \left(f_i(P) - \langle f(P) \rangle \right) p_i$$

where we used the abbreviation

$$f_i(P) = f_i(P_1, \dots, P_n)$$

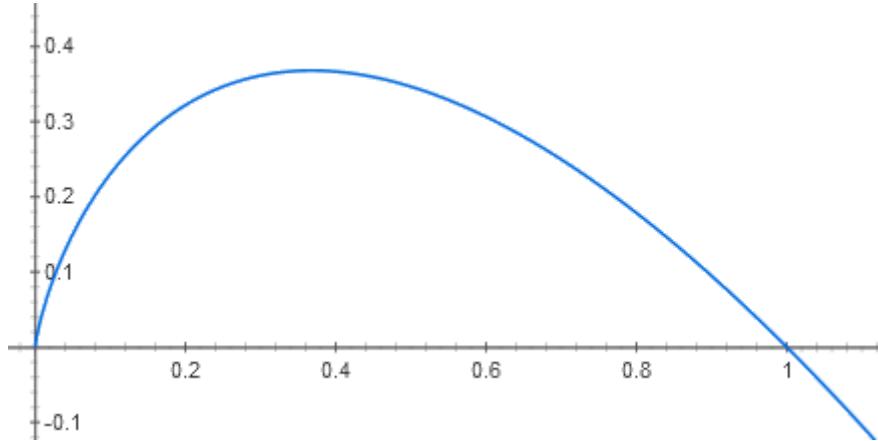
for the fitness of the i th species, and defined the **mean fitness** to be

$$\langle f(P) \rangle = \sum_i f_i(P) p_i$$

Using this cute formula for \dot{p}_i , we get the final result:

$$\dot{S} = - \sum_i \left(f_i(P) - \langle f(P) \rangle \right) p_i \ln(p_i)$$

This is strikingly similar to the formula for entropy itself. But now each term in the sum includes a factor saying how much more fit than average, or less fit, that species is. The quantity $-p_i \ln(p_i)$ is always nonnegative, since the graph of $-x \ln(x)$ looks like this:



So, the i th term contributes positively to the change in entropy if the i th species is fitter than average, but negatively if it's less fit than average.

This may seem counterintuitive!

Puzzle 1. How can we reconcile this fact with our earlier observations about the case when the fitness of each species is population-independent? Namely: a) if initially most of the replicators belong to one very unfit species, the entropy will rise at first, but b) in the long run, when the fittest species present take over, the entropy drops?

If this seems too tricky, look at some examples! The first illustrates observation a); the second illustrates observation b):

Puzzle 2. Suppose we have two species, one with fitness equal to 1 initially constituting 90% of the population, the other with fitness equal to 10 initially constituting just 10% of the population:

$$\begin{aligned} f_1 &= 1, & p_1(0) &= 0.9 \\ f_2 &= 10, & p_2(0) &= 0.1 \end{aligned}$$

At what rate does the entropy change at $t = 0$? Which species is responsible for most of this change?

Puzzle 3. Suppose we have two species, one with fitness equal to 10 initially constituting 90% of the population, and the other with fitness equal to 1 initially constituting just 10% of the population:

$$\begin{aligned} f_1 &= 10, & p_1(0) &= 0.9 \\ f_2 &= 1, & p_2(0) &= 0.1 \end{aligned}$$

At what rate does the entropy change at $t = 0$? Which species is responsible for most of this change?

I had to work through these examples to understand what's going on. Now I do, and it all makes sense.

Next time

Still, it would be nice if there were some quantity that *always goes down* with the passage of time, reflecting our naive idea that the population gains information from its environment, and thus loses entropy, as time goes by.

Often there *is* such a quantity. But it's not the naive entropy: it's the *relative* entropy. I'll talk about that next time. In the meantime, if you want to prepare, please reread [Part 6](#) of this series, where I explained this concept. Back then, I argued that *whenever you're tempted to talk about entropy, you should talk about relative entropy*. So, we should try that here.

There's a big idea lurking here: *information is relative*. How much information a signal gives you depends on your prior assumptions about what that signal is likely to be. If this is true, perhaps biodiversity is relative too.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2012 John Baez

baez@math.removethis.ucr.andthis.edu

[home](#)





June 7, 2012

Information Geometry (Part 11)

John Baez

Last time we saw that given a bunch of different species of self-replicating entities, the entropy of their population distribution can go either up or down as time passes. This is true even in the pathetically simple case where all the replicators have constant fitness—so they don't interact with each other, and don't run into any 'limits to growth'.

This is a bit of a bummer, since it would be nice to use entropy to explain how replicators are always extracting information from their environment, thanks to natural selection.

Luckily, a slight variant of entropy, called 'relative entropy', behaves better. When our replicators have an 'evolutionary stable state', the relative entropy is *guaranteed to always change in the same direction* as time passes!

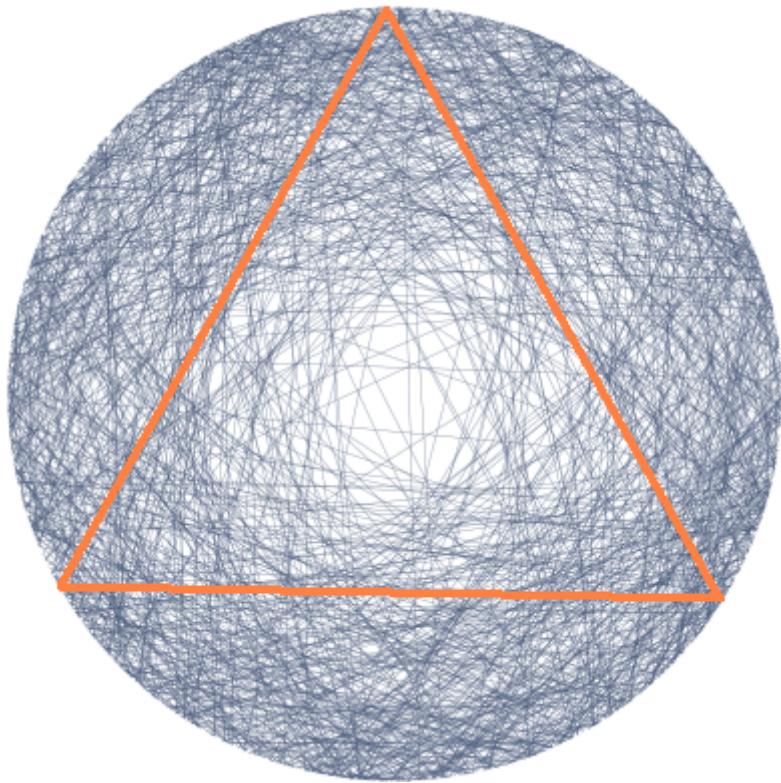
Thanks to Einstein, we've all heard that times and distances are relative. But how is entropy relative?

It's easy to understand if you think of entropy as lack of information. Say I have a coin hidden under my hand. I tell you it's heads-up. How much information did I just give you? Maybe 1 bit? That's true if you know it's a fair coin and I flipped it fairly before covering it up with my hand. But what if you put the coin down there yourself a minute ago, heads up, and I just put my hand over it? Then I've given you no information at all. The difference is the choice of 'prior': that is, what probability distribution you attributed to the coin *before* I gave you my message.

My love affair with relative entropy began in college when my friend Bruce Smith and I read Hugh Everett's thesis, [The Relative State Formulation of Quantum Mechanics](#). This was the origin of what's now often called the 'many-worlds interpretation' of quantum mechanics. But it also has a great introduction to relative entropy. Instead of talking about 'many worlds', I wish people would say that Everett explained some of the mysteries of quantum mechanics using the fact that entropy is relative.

Anyway, it's nice to see relative entropy showing up in biology.

Relative Entropy



Inscribe an equilateral triangle in a circle. Randomly choose a line segment joining two points of this circle. What is the probability that this segment is longer than a side of the triangle?

This puzzle is called [Bertrand's paradox](#), because different ways of solving it give different answers. To crack the paradox, you need to realize that it's meaningless to say you'll "randomly" choose something until you say more about how you're going to do it.

In other words, you can't compute the probability of an event until you pick a recipe for computing probabilities. Such a recipe is called a [probability measure](#).

This applies to computing entropy, too! The formula for entropy clearly involves a [probability distribution](#), even when our set of events is finite:

$$S = - \sum_i p_i \ln(p_i)$$

But this formula conceals a fact that becomes obvious when our set of events is infinite. Now the sum becomes an integral:

$$S = - \int_X p(x) \ln(p(x)) dx$$

And now it's clear that this formula makes no sense until we choose the [measure](#) dx . On a finite set we have a god-given choice of measure, called [counting measure](#). Integrals with respect to this are just sums. But in general we don't have such a god-given choice. And even for finite sets, working with counting measure is a *choice*: we are *choosing* to believe that in the absence of further evidence, all options are equally likely.

Taking this fact into account, it seems like we need two things to compute entropy: a probability distribution $p(x)$, and a measure dx . That's on the right track. But an even better way to think of it is this:

$$S = - \int_X \frac{p(x)dx}{dx} \ln \left(\frac{p(x)dx}{dx} \right) dx$$

Now we see the entropy depends *two* measures: the probability measure $p(x)dx$ we care about, but also the measure dx . Their ratio is important, but that's not enough: we also need one of these measures to do the integral. Above I used the measure dx to do the integral, but we can also use $p(x)dx$ if we write

$$S = - \int_X \ln\left(\frac{p(x)dx}{dx}\right) p(x)dx$$

Either way, we are computing the entropy of one measure *relative to another*. So we might as well admit it, and talk about **relative entropy**.

The entropy of the measure $d\mu$ **relative to** the measure $d\nu$ is defined by:

$$\begin{aligned} S(d\mu, d\nu) &= - \int_X \frac{d\mu(x)}{d\nu(x)} \ln\left(\frac{d\mu(x)}{d\nu(x)}\right) d\nu(x) \\ &= - \int_X \ln\left(\frac{d\mu(x)}{d\nu(x)}\right) d\mu(x) \end{aligned}$$

The second formula is simpler, but the first looks more like summing $-p \ln(p)$, so they're both useful.

Since we're taking entropy to be lack of information, we can also get rid of the minus sign and define **relative information** by

$$\begin{aligned} I(d\mu, d\nu) &= \int_X \frac{d\mu(x)}{d\nu(x)} \ln\left(\frac{d\mu(x)}{d\nu(x)}\right) d\nu(x) \\ &= \int_X \ln\left(\frac{d\mu(x)}{d\nu(x)}\right) d\mu(x) \end{aligned}$$

If you thought something was randomly distributed according to the probability measure $d\nu$, but then you discover it's randomly distributed according to the probability measure $d\mu$, how much information have you gained? The answer is $I(d\mu, d\nu)$.

For more on relative entropy, read [Part 6](#) of this series. I gave some examples illustrating how it works. Those should convince you that it's a useful concept.

Okay: now let's switch back to a more lowbrow approach. In the case of a finite set, we can revert to thinking of our two measures as probability distributions, and write the information gain as

$$I(q, p) = \sum_i \ln\left(\frac{q_i}{p_i}\right) q_i$$

If you want to sound like a Bayesian, call p the [prior probability distribution](#) and q the [posterior probability distribution](#). Whatever you call them, $I(q, p)$ is the amount of information you get if you thought p and someone tells you "no, q !"?

We'll use this idea to think about how a population gains information about its environment as time goes by, thanks to natural selection. The rest of this post will be an exposition of Theorem 1 in this paper:

- Marc Harper, [The replicator equation as an inference dynamic](#).

Harper says versions of this theorem ave previously appeared in work by Ethan Akin, and independently in work by Josef Hofbauer and Karl Sigmund. He also credits others [here](#). An idea this good is rarely noticed by just one person.

The change in relative information

So: consider n different species of replicators. Let P_i be the population of the i th species, and assume these populations change according to the [replicator equation](#):

$$\frac{dP_i}{dt} = f_i(P_1, \dots, P_n) P_i$$

where each function f_i depends smoothly on all the populations. And as usual, we let

$$p_i = \frac{P_i}{\sum_j P_j}$$

be the fraction of replicators in the i th species.

Let's study the relative information $I(q, p)$ where q is some fixed probability distribution. We'll see something great happens when q is a stable equilibrium solution of the replicator equation. In this case, the relative information can never increase! It can only decrease or stay constant.

We'll think about what all this *means* later. First, let's see that it's true! Remember,

$$\begin{aligned} I(q, p) &= \sum_i \ln\left(\frac{q_i}{p_i}\right) q_i \\ &= \sum_i \left(\ln(q_i) - \ln(p_i) \right) q_i \end{aligned}$$

and only p_i depends on time, not q_i , so

$$\begin{aligned} \frac{d}{dt} I(q, p) &= -\frac{d}{dt} \sum_i \ln(p_i) q_i \\ &= -\sum_i \frac{\dot{p}_i}{p_i} q_i \end{aligned}$$

where \dot{p}_i is the rate of change of the probability p_i . We saw a nice formula for this in [Part 9](#):

$$\dot{p}_i = \left(f_i(P) - \langle f(P) \rangle \right) p_i$$

where

$$f_i(P) = f_i(P_1, \dots, P_n)$$

and

$$\langle f(P) \rangle = \sum_i f_i(P) p_i$$

is the **mean fitness** of the species. So, we get

$$\frac{d}{dt} I(q, p) = - \sum_i \left(f_i(P) - \langle f(P) \rangle \right) q_i$$

Nice, but we can fiddle with this expression to get something more enlightening. Remember, the numbers q_i sum to one. So:

$$\begin{aligned} \frac{d}{dt} I(q, p) &= \langle f(P) \rangle - \sum_i f_i(P) q_i \\ &= \sum_i f_i(P) (p_i - q_i) \end{aligned}$$

where in the last step I used the definition of the mean fitness. This result looks even cuter if we treat the numbers $f_i(P)$ as the components of a vector $f(P)$, and similarly for the numbers p_i and q_i . Then we can use the dot product of vectors to say

$$\frac{d}{dt} I(q, p) = f(P) \cdot (p - q)$$

So, the relative information $I(q, p)$ will always decrease if

$$f(P) \cdot (p - q) \leq 0$$

for all choices of the population P .

And now something really nice happens: this is also the condition for q to be an [evolutionarily stable state](#). This concept goes back to [John Maynard Smith](#), the founder of evolutionary game theory. In 1982 he wrote:

A population is said to be in an evolutionarily stable state if its genetic composition is restored by selection after a disturbance, provided the disturbance is not too large.

I will explain the math next time—I need to straighten out some things in my mind first. But the basic idea is compelling: an evolutionarily stable state is like a situation where our replicators 'know all there is to know' about the environment and each other. In any other state, the population has 'something left to learn'—and the amount left to learn is the relative information we've been talking about! But as time goes on, the information still left to learn *decreases!*

Note: in the real world, nature has never found an evolutionarily stable state... except sometimes approximately, on sufficiently short time scales, in sufficiently small regions. So we are still talking about an idealization of reality! But that's okay, as long as we know it.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2012 John Baez

baez@math.removethis.ucr.andthis.edu

[home](#)





June 24, 2012

Information Geometry (Part 12)

John Baez

Last time we saw that if a population evolves toward an 'evolutionarily stable state', then the amount of information our population has 'left to learn' can never increase! It must always decrease or stay the same.

This result sounds wonderful: it's a lot like the second law of thermodynamics, which says entropy must always increase. Of course there are some conditions for this wonderful result to hold. The main condition is that the population evolves according to the [replicator equation](#). But the other is the existence of an evolutionarily stable state. Last time I wrote down the rather odd-looking definition of 'evolutionary stable state' without justifying it. I need to do that soon. But if you've never thought about evolutionary game theory, I think giving you a little background will help. So today let me try that.

Evolutionary game theory

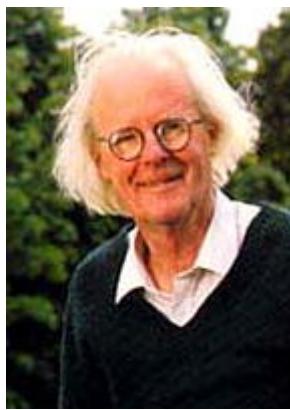
We've been thinking of evolution as similar to *inference* or *learning*. In this analogy, organisms are like 'hypotheses', and the population 'does experiments' to see if these hypotheses make 'correct predictions' (i.e., can reproduce) or not. The successful ones are reinforced while the unsuccessful ones are weeded out. As a result, the population 'learns'. And under the conditions of the theorem we discussed last time, the relative information—the amount 'left to learn'—goes down!

While you might object to various points of this analogy, it's useful—and that's really all you can ask of an analogy. It's useful because it lets us steal chunks of math from the subjects of [Bayesian inference](#) and [machine learning](#) and apply them to the study of biodiversity and evolution! This is what Marc Harper has been doing:

- Marc Harper, [Information geometry and evolutionary game theory](#).
- Marc Harper, [The replicator equation as an inference dynamic](#).

But now let's bring in another analogy, also contained in Harper's work. We can also think of evolution as similar to a *game*. In this analogy, organisms are like 'strategies'—or if you prefer, they have strategies. The winners get to reproduce, while the losers don't. [John Maynard Smith](#) started developing this analogy in 1973, and eventually wrote a whole book on it:

- John Maynard Smith, *Evolution and the Theory of Games*, Cambridge University Press, 1982.



As far as I can tell, evolutionary game theory has brought almost as many chunks of math *to* game theory as it has taken from it. Maybe it's just my ignorance showing, but it seems that game theory becomes considerably deeper when we think about games that many players play again and again, with the winners getting to reproduce, while the losers are eliminated.

According to [William Sandholm](#):

The birth of evolutionary game theory is marked by the publication of a series of papers by mathematical biologist John Maynard Smith. Maynard Smith adapted the methods of traditional game theory, which were created to model the behavior of rational economic agents, to the context of biological natural selection. He proposed his notion of an evolutionarily stable strategy (ESS) as a way of explaining the existence of ritualized animal conflict.

Maynard Smith's equilibrium concept was provided with an explicit dynamic foundation through a differential equation model introduced by Taylor and Jonker. Schuster and Sigmund, following Dawkins, dubbed this model the replicator dynamic, and recognized the close links between this game-theoretic dynamic and dynamics studied much earlier in population ecology and population genetics. By the 1980s, evolutionary game theory was a well-developed and firmly established modeling framework in biology.

Towards the end of this period, economists realized the value of the evolutionary approach to game theory in social science contexts, both as a method of providing foundations for the equilibrium concepts of traditional game theory, and as a tool for selecting among equilibria in games that admit more than one. Especially in its early stages, work by economists in evolutionary game theory hewed closely to the interpretation set out by biologists, with the notion of ESS and the replicator dynamic understood as modeling natural selection in populations of agents genetically programmed to behave in specific ways. But it soon became clear that models of essentially the same form could be used to study the behavior of populations of active decision makers. Indeed, the two approaches sometimes lead to identical models: the replicator dynamic itself can be understood not only as a model of natural selection, but also as one of imitation of successful opponents.

While the majority of work in evolutionary game theory has been undertaken by biologists and economists, closely related models have been applied to questions in a variety of fields, including transportation science, computer science, and sociology. Some paradigms from evolutionary game theory are close relatives of certain models from physics, and so have attracted the attention of workers in this field. All told, evolutionary game theory provides a common ground for workers from a wide range of disciplines.

The Prisoner's Dilemma

In game theory, the most famous example is the [Prisoner's Dilemma](#). In its original form, this 'game' is played just once:

Two men are arrested, but the police don't have enough information to convict them. So they separate the two men, and offer both the same deal: if one testifies against his partner (or **defects**), and the other remains silent (and thus **cooperates** with his partner), the defector goes free and the cooperator goes to jail for 12 months. If both remain silent, both are sentenced to only 1 month in jail for a minor charge. If they both defect, they both receive a 3-month sentence. Each prisoner must choose either to defect or cooperate with his partner in crime; neither gets to hear what the other decides. What will they do?

Traditional game theory emphasizes the so-called 'Nash equilibrium' for this game, in which both prisoners defect. Why don't they both cooperate? They'd both be better off if they both cooperated. However, for them to both cooperate is 'unstable': either one could shorten their sentence by defecting! By definition, a Nash equilibrium has the property that neither player can improve his situation by unilaterally changing his strategy.

In the Prisoner's Dilemma, the Nash equilibrium is not very nice: both parties would be happier if they'd only cooperate. That's why it's called a 'dilemma'. Perhaps the most tragic example today is global warming. Even if all players would be better off if all cooperate to reduce carbon emissions, any *one* will be better off if everybody *except themselves* cooperates while they emit more carbon.

For this and many other reasons, people have been interested in 'solving' the Prisoner's Dilemma: that is, finding reasons why cooperation might be favored over defection.

This book got people really excited in seeing what evolutionary game theory has to say about the Prisoner's Dilemma:

- Robert Axelrod, *The Evolution of Cooperation*, Basic Books, New York, 1984. (A related article with the same title is [available online](#).)

The idea is that under certain circumstances, strategies that are 'nicer' than defection will gradually take over. The most famous of these strategies is 'tit for tat', meaning that you cooperate the first time and after that do whatever your opponent just did. I won't go into this further, because it's a big digression and I'm already digressing too far. I'll just mention that the Prisoner's Dilemma is still full of surprises. Just this week, some fascinating new work has been causing a stir:

- William Press and Freeman Dyson, [Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent](#), *Edge*, 18 June 2012.

I hope I've succeeded in giving you a vague superficial sense of the history of evolutionary game theory and why it's interesting. Next time I'll get serious about the task at hand, which is to understand 'evolutionarily stable strategies'. If you want to peek ahead, try this nice paper:

- William H. Sandholm, [Evolutionary game theory](#), 12 November 2007.

This is where I got the long quote by Sandholm on the history of evolutionary game theory. The original quote contained lots of references; if you're interested in those, go to page 3 of this paper.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!





June 26, 2012

Information Geometry (Part 13)

John Baez

Last time I gave a sketchy overview of evolutionary game theory. Now let's get serious.

I'll start by explaining 'Nash equilibria' for 2-person games. These are situations where neither player can profit by changing what they're doing. Then I'll introduce 'mixed strategies', where the players can choose among several strategies with different probabilities. Then I'll introduce evolutionary game theory, where we think of each strategy as a *species*, and its probability as *the fraction of organisms that belong to that species*.

Back in [Part 9](#), I told you about the 'replicator equation', which says how these fractions change with time thanks to natural selection. Now we'll see how this leads to the idea of an 'evolutionarily stable strategy'. And finally, we'll see that when evolution takes us toward such a stable strategy, the amount of information the organisms have 'left to learn' keeps decreasing!

Nash equilibria

We can describe a certain kind of two-person game using a **payoff matrix**, which is an $n \times n$ matrix A_{ij} of real numbers. We think of A_{ij} as the payoff that either player gets if they choose strategy i and their opponent chooses strategy j .

Note that in this kind of game, there's no significant difference between the 'first player' and the 'second player': *either* player wins an amount A_{ij} if they choose strategy i and their opponent chooses strategy j . So, this kind of game is called **symmetric** even though the matrix A_{ij} may not be symmetric. Indeed, it's common for this matrix to be antisymmetric, meaning $A_{ij} = -A_{ji}$, since in this case what one player wins, the other loses. Games with this extra property are called **zero-sum games**. But we won't limit ourselves to those!

We say a strategy i is a **symmetric Nash equilibrium** if

$$A_{ii} \geq A_{ji}$$

for all j . This means that if both players use strategy i , neither gains anything by switching to another strategy.

For example, suppose our matrix is

$$\begin{pmatrix} -1 & -12 \\ 0 & -3 \end{pmatrix}$$

Then we've got the Prisoner's Dilemma exactly as described last time! Here strategy 1 is **cooperate** and strategy 2 is **defect**. If a player cooperates and so does his opponent, he wins

$$A_{11} = -1$$

meaning he gets one month in jail. We include a minus sign because 'winning a month in jail' is not a good thing. If the player cooperates but his opponent defects, he gets a whole year in jail:

$$A_{12} = -12$$

If he defects but his opponent cooperates, he doesn't go to jail at all:

$$A_{21} = 0$$

And if they both defect, they both get three months in jail:

$$A_{22} = -3$$

You can see that defecting is a Nash equilibrium, since

$$A_{22} \geq A_{12}$$

So, oddly, if our prisoners know game theory and believe Nash equilibria are best, they'll both be worse off than if they cooperate and don't betray each other.



Nash equilibria for mixed strategies

So far we've been assuming that with 100% certainty, each player chooses one strategy $i = 1, 2, 3, \dots, n$. Since we'll be considering more general strategies in a minute, let's call these **pure strategies**.

Now let's throw some probability theory into the stew! Let's allow the players to pick different pure strategies with different probabilities. So, we define a **mixed strategy** to be a probability distribution on the set of pure strategies. In other words, it's a list of n nonnegative numbers

$$p_i \geq 0$$

that sum to one:

$$\sum_{i=1}^n p_i = 1$$

Say I choose the mixed strategy p while you, my opponent, choose the mixed strategy q . Say our choices are made independently. Then the probability that I choose the pure strategy i while you chose j is

$$p_i q_j$$

so the expected value of my winnings is

$$\sum_{i,j=1}^n p_i A_{ij} q_j$$

or using vector notation

$$p \cdot Aq$$

where the dot is the usual dot product on \mathbf{R}^n .

We can easily adapt the concept of Nash equilibrium to mixed strategies. A mixed strategy q is a **symmetric Nash equilibrium** if for any other mixed strategy p ,

$$q \cdot Aq \geq p \cdot Aq$$

This means that if both you and I are playing the mixed strategy q , I can't improve my expected winnings by unilaterally switching to the mixed strategy p . And neither can you, because the game is symmetric!

If this were a course on game theory, I would now do some examples. But it's not, so I'll just send you to page 6 of [Sandholm's paper](#): he looks at some famous games like 'hawks and doves' and 'rock paper scissors'.

Evolutionarily stable strategies

We're finally ready to discuss evolutionarily stable strategies. To do this, let's reinterpret the 'pure strategies' $i = 1, 2, 3, \dots, n$ as **species**. Here I don't necessarily mean species in the classic biological sense: I just mean different kinds of self-replicating entities, or **replicators**. For example, they could be different [alleles](#) of the same gene.

Similarly, we'll reinterpret the 'mixed strategy' p as describing a mixed population of replicators, where the fraction of replicators belonging to the i th species is p_i . These numbers are still probabilities: p_i is the probability that a randomly chosen replicator will belong to the i th species.

We'll reinterpret the payoff matrix A_{ij} as a **fitness matrix**. In our earlier discussion of the replicator equation, we assumed that the population P_i of the i th species grew according to the replicator equation

$$\frac{dP_i}{dt} = f_i(P_1, \dots, P_n)P_i$$

where the **fitness function** f_i is any smooth function of the populations of each kind of replicator.

But in evolutionary game theory it's common to start by looking at a simple special case where

$$f_i(P_1, \dots, P_n) = \sum_{j=1}^n A_{ij}p_j$$

where

$$p_j = \frac{P_j}{\sum_k P_k}$$

is the fraction of replicators who belong to the j th species.

What does this mean? The idea is that we have a well-mixed population of game players—or replicators. Each one has its own pure strategy—or species. Each one randomly roams around and 'plays games' with each other replicator it meets. It gets to reproduce at a rate proportional to its expected winnings.

This is unrealistic in all sorts of ways, but it's mathematically cute, and it's been studied a lot, so it's good to know about. Today I'll explain evolutionarily stable strategies only in this special case. Later I'll go back to the

general case.

Suppose that we select a sample of replicators from the overall population. What is the mean fitness of the replicators in this sample? For this, we need to know the probability that a replicator from this sample belongs to the i th species. Say it's q_j . Then the mean fitness of our sample is

$$\sum_{i,j=1}^n q_i A_{ij} p_j$$

This is just a weighted average of the fitnesses in our earlier formula. But using the magic of vectors, we can write this sum as

$$q \cdot Ap$$

We already saw this type of expression in the last section! It's my expected winnings if I play the mixed strategy q and you play the mixed strategy p .

John Maynard Smith defined q to be **evolutionarily stable strategy** if when we add a small population of 'invaders' distributed according to any other probability distribution p , the original population is more fit than the invaders.

In simple terms: a small 'invading' population will do worse than the population as a whole.

More precisely:

$$q \cdot A((1 - \epsilon)q + \epsilon p) > p \cdot A((1 - \epsilon)q + \epsilon p)$$

for all mixed strategies $p \neq q$ and all sufficiently small $\epsilon > 0$. Here

$$(1 - \epsilon)q + \epsilon p$$

is the population we get by replacing an ϵ -sized portion of our original population by invaders.

Puzzle: Show that q is an evolutionarily stable strategy if and only these two conditions hold for all mixed stategies p :

$$q \cdot Aq \geq p \cdot Aq$$

and also, for all $p \neq q$,

$$q \cdot Aq = p \cdot Aq \Rightarrow q \cdot Ap > p \cdot Ap$$

The first condition says that q is a symmetric Nash equilibrium. In other words, the invaders can't on average be *better* playing against the original population than members of the original population are. The second says that if the invaders are *just as good* at playing against the original population, they must be worse at playing against each other! The combination of these conditions means the invaders won't take over.

Again, I should do some examples... but instead I'll refer you to page 9 of [Sandholm's paper](#), and also these course notes:

- Samuel Alizon and Daniel Cownden, [Evolutionary games and evolutionarily stable strategies](#).
- Samuel Alizon and Daniel Cownden, [Replicator dynamics](#).

The decrease of relative information

Now comes the punchline... but with a slight surprise twist at the end. [Last time](#) we let

$$P = (P_1, \dots, P_n)$$

be a population that evolves with time according to the replicator equation, and we let p be the corresponding probability distribution. We supposed q was some fixed probability distribution. We saw that the relative information

$$I(q, p) = \sum_i \ln\left(\frac{q_i}{p_i}\right) q_i$$

obeys

$$\frac{d}{dt} I(q, p) = (p - q) \cdot f(P)$$

where $f(P)$ is the vector of fitness functions. So, this relative information can never increase if

$$(p - q) \cdot f(P) \leq 0$$

for all P .

We can adapt this to the special case we're looking at now. Remember, right now we're assuming

$$f_i(P_1, \dots, P_n) = \sum_{j=1}^n A_{ij} p_j$$

so

$$f(P) = Ap$$

Thus, the relative information will never increase if

$$(p - q) \cdot Ap \leq 0$$

or in other words,

$$\forall p \quad q \cdot Ap \geq p \cdot Ap \tag{1}$$

Now, this looks very similar to the conditions for an evolutionary stable strategy as stated in the Puzzle above. *But it's not the same!* That's the surprise twist.

Remember, the Puzzle says that q is an evolutionarily stable state if

$$\forall p \quad q \cdot Aq \geq p \cdot Aq \tag{2}$$

and also

$$\forall p \neq q \quad q \cdot Aq = p \cdot Aq \Rightarrow q \cdot Ap > p \cdot Ap \tag{3}$$

Note that condition (1), the one we want, is *neither* condition (2) *nor* condition (3)! This drove me crazy for almost a day.



I kept thinking I'd made a mistake, like mixing up p and q somewhere. You've got to mind your p's and q's in this game!

But the solution turned out to be this. After Maynard Smith came up with his definition of 'evolutionarily stable state', another guy came up with a different definition:

- Bernhard Thomas, On evolutionarily stable sets, *J. Math. Biology* **22** (1985), 105-115.

For him, an **evolutionarily stable strategy** q is one such that

$$\forall p \quad q \cdot Aq \geq p \cdot Aq \tag{2}$$

and also

$$\forall p \quad q \cdot Ap \geq p \cdot Ap \tag{1}$$

Condition (1) is stronger than condition (3), so he renamed Maynard Smith's evolutionarily stable strategies **weakly evolutionarily stable strategies**. And condition (1) guarantees that the relative information $I(q, p)$ can never increase. So, now we're happy.

Except for one thing: why should we switch from Maynard Smith's perfectly sensible concept of evolutionarily stable state to this new stronger one? I don't really know, except that

- it's not much stronger

and

- it lets us prove the theorem we want!

So, it's a small mystery for me to mull over. If you have any good ideas, let me know.

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



Relative Entropy in Biological Systems

 johncarlosbaez.wordpress.com/2015/11/27/relative-entropy-in-biological-systems/

November 27,
2015

Here's a paper for the proceedings of a workshop on Information and Entropy in Biological System this spring:

- John Baez and Blake Pollard, Relative entropy in biological systems, with Blake S. Pollard, *Entropy* **18** (2016), 46.

We'd love any comments or questions you might have. I'm not happy with the title. In the paper we advocate using the term 'relative information' instead of 'relative entropy'—yet the latter is much more widely used, so I feel we need it in the title to let people know what the paper is about!

Here's the basic idea.

Life relies on nonequilibrium thermodynamics, since in thermal equilibrium there are no flows of free energy. Biological systems are also open systems, in the sense that both matter and energy flow in and out of them. Nonetheless, it is important in biology that systems can sometimes be treated as approximately closed, and sometimes approach equilibrium before being disrupted in one way or another. This can occur on a wide range of scales, from large ecosystems to within a single cell or organelle. Examples include:

- A population approaching an evolutionarily stable state.
- Random processes such as mutation, genetic drift, the diffusion of organisms in an environment or the diffusion of molecules in a liquid.
- A chemical reaction approaching equilibrium.

An interesting common feature of these processes is that as they occur, quantities mathematically akin to entropy tend to increase. Closely related quantities such as free energy tend to decrease. In this review, we explain some mathematical results that make this idea precise.

Most of these results involve a quantity that is variously known as 'relative information', 'relative entropy', 'information gain' or the 'Kullback–Leibler divergence'. We'll use the first term. Given two probability distributions p and q on a finite set X , their **relative information**, or more precisely the **information of p relative to q** , is

We use the word ‘information’ instead of ‘entropy’ because one expects entropy to increase with time, and the theorems we present will say that $I(p\|q)$ decreases with time under various conditions. The reason is that the Shannon entropy

contains a minus sign that is missing from the definition of relative information.

$$I(p\|q) = \sum_{i \in X} p_i \ln \left(\frac{p_i}{q_i} \right)$$

$$S(p) = - \sum_{i \in X} p_i \ln p_i$$

Intuitively, $I(p\|q)$ is the amount of information gained when we start with a hypothesis given by some probability distribution q and then learn the ‘true’ probability distribution p . For example, if we start with the hypothesis that a coin is fair and then are told that it landed heads up, the relative information is $\ln 2$, so we have gained 1 bit of information. If however we started with the hypothesis that the coin always lands heads up, we would have gained no information.

We put the word ‘true’ in quotes here, because the notion of a ‘true’ probability distribution, which subjective Bayesians reject, is not required to use relative information. A more cautious description of relative information is that it is a **divergence**: a way of measuring the difference between probability distributions that obeys

$$I(p\|q) \geq 0$$

and

but not necessarily the other axioms for a distance function, symmetry and the triangle inequality, which indeed fail for relative information.

$$I(p\|q) = 0 \iff p = q$$

There are many other divergences besides relative information, some of which we discuss in Section 6. However, relative information can be singled out by a number of characterizations, including one based on ideas from Bayesian inference. The relative information is also close to the expected number of extra bits required to code messages distributed according to the probability measure p using a code optimized for messages distributed according to q .

In this review, we describe various ways in which a population or probability distribution evolves continuously according to some differential equation. For all these differential equations, I describe conditions under which relative information decreases. Briefly, the results are as follows. We hasten to reassure the reader that our paper explains all the jargon involved, and the proofs of the claims are given in full:

- In Section 2, we consider a very general form of the Lotka–Volterra equations, which are a commonly used model of population dynamics. Starting from the population P_i of each type of replicating entity, we can define a probability distribution

$$p_i = \frac{P_i}{\sum_{i \in X} P_i}$$

which evolves according to a nonlinear equation called the replicator equation. We describe a necessary and sufficient condition under which $I(q \| p(t))$ is nonincreasing when $p(t)$ evolves according to the replicator equation while q is held fixed.

- In Section 3, we consider a special case of the replicator equation that is widely studied in evolutionary game theory. In this case we can think of probability distributions as mixed strategies in a two-player game. When q is a dominant strategy, $I(q \| p(t))$ can never increase when $p(t)$ evolves according to the replicator equation. We can think of $I(q \| p(t))$ as the information that the population has left to learn. Thus, evolution is analogous to a learning process—an analogy that in the field of artificial intelligence is exploited by evolutionary algorithms!
- In Section 4 we consider continuous-time, finite-state Markov processes. Here we have probability distributions on a finite set X evolving according to a linear equation called the master equation. In this case $I(p(t) \| q(t))$ can never increase. Thus, if q is a steady state solution of the master equation, both $I(p(t) \| q)$ and $I(q \| p(t))$ are nonincreasing. We can always write q as the Boltzmann distribution for some energy function $E : X \rightarrow \mathbb{R}$, meaning that

where T is temperature and k is Boltzmann's constant. In this case, $I(p(t) \| q)$ is proportional to a difference of free energies:

Thus, the nonincreasing nature of $I(p(t) \| q)$ is a version of the Second Law of Thermodynamics.

- In Section 5, we consider chemical reactions and other processes described by reaction networks. In this context we have populations P_i of entities of various kinds $i \in X$, and these populations evolve according to a nonlinear equation called the rate equation. We can generalize relative information from probability distributions to populations by setting

If Q is a special sort of steady state solution of the rate equation, called a complex balanced equilibrium, $I(P(t) \| Q)$ can never increase when $P(t)$ evolves according to the rate equation.

- Finally, in Section 6, we consider a class of functions called f -divergences which include relative information as a special case. For any convex function f , the **f -divergence** of two probability distributions is given by

$$q_i = \frac{\exp(-E_i/kT)}{\sum_{j \in X} \exp(-E_j/kT)}$$

$$I(p(t) \| q) = \frac{F(p) - F(q)}{T}$$

$$I(P \| Q) = \sum_{i \in X} P_i \ln \left(\frac{P_i}{Q_i} \right) - (P_i - Q_i)$$

$$f : [0, \infty) \rightarrow [0, \infty)$$

Whenever $p(t)$ and $q(t)$ are probability distributions evolving according to the master equation of some Markov process, $I_f(p(t)\|q(t))$ is nonincreasing. The f -divergence is also well-defined for populations, and nonincreasing for two populations that both evolve according to the master equation.

$$I_f(p\|q) = \sum_{i \in X} q_i f\left(\frac{p_i}{q_i}\right)$$



January 11, 2016

Information Geometry (Part 14)

John Baez and [Blake Pollard](#)

It's been a long time since you've seen an installment of the [Information Geometry](#) series on this blog. If you recall, this series turned out to be largely about relative entropy and how it changes in evolutionary games. Some of what we said is summarized and carried further here:

- John Baez and Blake Pollard, [Relative entropy in biological systems](#). (Blog article [here](#).)

But now Blake has a new paper, and I want to talk about that:

- Blake Pollard, [Open Markov processes: a compositional perspective on non-equilibrium steady states in biology](#), *Entropy* **18** (2016), 140.

I'll focus on just one aspect: the principle of minimum entropy production. This is an exciting yet controversial principle in non-equilibrium thermodynamics. Blake examines it in a situation where we can tell exactly what's happening.

Non-equilibrium steady states

Life exists away from equilibrium. Left isolated, systems will tend toward thermodynamic equilibrium. However, biology is about **open systems**: physical systems that exchange matter or energy with their surroundings. Open systems can be maintained away from equilibrium by this exchange. This leads to the idea of a **non-equilibrium steady state** — a state of an open system that doesn't change, but is not in equilibrium.

A simple example is a pan of water sitting on a stove. Heat passes from the flame to the water and then to the air above. If the flame is very low, the water doesn't boil and nothing moves. So, we have a steady state, at least approximately. But this is not an equilibrium, because there is a constant flow of energy through the water.

Of course in reality the water will be slowly evaporating, so we don't really have a steady state. As always, models are approximations. If the water is evaporating slowly enough, it can be useful to approximate the situation with a non-equilibrium steady state.

There is much more to biology than steady states. However, to dip our toe into the chilly waters of non-equilibrium thermodynamics, it is nice to start with steady states. And already here there are puzzles left to solve.

Minimum entropy production

Ilya Prigogine won the Nobel prize for his work on non-equilibrium thermodynamics. One reason is that he had an interesting idea about steady states. He claimed that under certain conditions, a non-equilibrium steady state will *minimize entropy production!*

There has been a lot of work trying to make the '[principle of minimum entropy production](#)' precise and turn it into a theorem. In this book:

- G. Lebon and D. Jou, *Understanding Non-equilibrium Thermodynamics*, Springer, Berlin, 2008.

the authors give an argument for the principle of minimum entropy production based on four conditions:

- **time-independent boundary conditions**: the surroundings of the system don't change with time.
- **linear phenomenological laws**: the laws governing the macroscopic behavior of the system are linear.
- **constant phenomenological coefficients**: the laws governing the macroscopic behavior of the system don't change with time.
- **symmetry of the phenomenological coefficients**: since they are linear, the laws governing the macroscopic behavior of the system can be described by a linear operator, and we demand that in a suitable basis the matrix for this operator is symmetric: $T_{ij} = T_{ji}$.

The last condition is obviously the subtlest one; it's sometimes called [Onsager reciprocity](#), and people have spent a lot of time trying to derive it from other conditions.

However, Blake goes in a different direction. He considers a concrete class of open systems, a very large class called 'open Markov processes'. These systems obey the first three conditions listed above, and the 'detailed balanced' open Markov processes also obey the last one. But Blake shows that minimum entropy production holds only approximately — with the approximation being good for steady states that are *near equilibrium*!

However, he shows that another minimum principle holds exactly, even for steady states that are far from equilibrium. He calls this the 'principle of minimum dissipation'.

We actually discussed the principle of minimum dissipation in an earlier paper:

- John Baez, Brendan Fong and Blake Pollard, [A compositional framework for Markov processes](#). (Blog article [here](#).)

But one advantage of Blake's new paper is that it presents the results with a minimum of category theory. Of course I love category theory, and I think it's the right way to formalize open systems, but it can be intimidating.

Another good thing about Blake's new paper is that it explicitly compares the principle of minimum entropy to the principle of minimum dissipation. He shows they agree in a certain limit — namely, the limit where the system is close to equilibrium.

Let me explain this. I won't include the nice example from biology that Blake discusses: a very simple model of membrane transport. For that, read his paper! I'll just give the general results.

The principle of minimum dissipation

An **open Markov process** consists of a finite set X of **states**, a subset $B \subseteq X$ of **boundary states**, and an **infinitesimal stochastic** operator $H : \mathbf{R}^X \rightarrow \mathbf{R}^X$, meaning a linear operator with

$$H_{ij} \geq 0 \text{ for all } i \neq j$$

and

$$\sum_i H_{ij} = 0 \text{ for all } j$$

I'll explain these two conditions in a minute.

For each $i \in X$ we introduce a **population** $p_i \in [0, \infty)$. We call the resulting function $p : X \rightarrow [0, \infty)$ the **population distribution**. Populations evolve in time according to the **open master equation**:

$$\frac{dp_i}{dt} = \sum_j H_{ij} p_j \text{ for all } i \in X - B$$

$$p_i(t) = b_i(t) \text{ for all } i \in B$$

So, the populations p_i obey a linear differential equation at states i that are not in the boundary, but they are specified 'by the user' to be chosen functions b_i at the boundary states.

The off-diagonal entries H_{ij} , $i \neq j$ are the rates at which population hops from the j th to the i th state. This lets us understand the definition of an infinitesimal stochastic operator. The first condition:

$$H_{ij} \geq 0 \text{ for all } i \neq j$$

says that the rate for population to transition from one state to another is non-negative. The second:

$$\sum_i H_{ij} = 0 \text{ for all } j$$

says that population is conserved, at least if there are no boundary states. Population can flow in or out at boundary states, since the master equation doesn't hold there.

A **steady state** is a solution of the open master equation that does not change with time. A steady state for a closed Markov process is typically called an **equilibrium**. So, an equilibrium obeys the master equation at all states, while for a steady state this may not be true at the boundary states. Again, the reason is that population can flow in or out at the boundary.

We say an equilibrium $q : X \rightarrow [0, \infty)$ of a Markov process is **detailed balanced** if the rate at which population flows from the i th state to the j th state is equal to the rate at which it flows from the j th state to the i th:

$$H_{ji} q_i = H_{ij} q_j \text{ for all } i, j \in X$$

Suppose we've got an open Markov process that has a detailed balanced equilibrium q . Then a non-equilibrium steady state p will minimize a function called the 'dissipation', subject to constraints on its boundary populations. There's a nice formula for the dissipation in terms of p and q .

Definition. Given an open Markov process with detailed balanced equilibrium q we define the **dissipation** for a population distribution p to be

$$D(p) = \frac{1}{2} \sum_{i,j} H_{ij} q_j \left(\frac{p_j}{q_j} - \frac{p_i}{q_i} \right)^2$$

This formula is a bit tricky, but you'll notice it's quadratic in p and it vanishes when $p = q$. So, it's pretty nice.

Using this concept we can formulate a principle of minimum dissipation, and prove that non-equilibrium steady states obey this principle:

Definition. We say a population distribution $p : X \rightarrow \mathbf{R}$ obeys the **principle of minimum dissipation** with boundary population $b : X \rightarrow \mathbf{R}$ if p minimizes $D(p)$ subject to the constraint that

$$p_i = b_i \text{ for all } i \in B$$

Theorem 1. A population distribution p is a steady state with $p_i = b_i$ for all boundary states i if and only if p obeys the principle of minimum dissipation with boundary population b .

Proof. This follows from Theorem 28 in [A compositional framework for Markov processes](#).

Minimum entropy production versus minimum dissipation

How does dissipation compare with entropy production? To answer this, first we must ask: what really is entropy production? And: how does the equilibrium state q show up in the concept of entropy production?

The **relative entropy** of two population distributions p, q is given by

$$I(p, q) = \sum_i p_i \ln \left(\frac{p_i}{q_i} \right)$$

It is well known that for a closed Markov process with q as a detailed balanced equilibrium, the relative entropy is monotonically *decreasing* with time. This is due to an annoying sign convention in the definition of relative entropy: while entropy is typically increasing, relative entropy typically decreases. We could fix this by putting a minus sign in the above formula or giving this quantity $I(p, q)$ some other name. A lot of people call it the **Kullback–Leibler divergence**, but I have taken to calling it **relative information**. For more, see:

- John Baez and Blake Pollard, [Relative entropy in biological systems](#). (Blog article [here](#).)

We say 'relative entropy' in the title, but then we explain why 'relative information' is a better name, and use that. More importantly, we explain why $I(p, q)$ has the physical meaning of *free energy*. Free energy tends to decrease, so everything is okay. For details, see Section 4.

Blake has a nice formula for how fast $I(p, q)$ decreases:

Theorem 2. Consider an open Markov process with X as its set of states and B as the set of boundary states. Suppose $p(t)$ obeys the open master equation and q is a detailed balanced equilibrium. For any boundary state $i \in B$, let

$$\frac{Dp_i}{Dt} = \frac{dp_i}{dt} - \sum_{j \in X} H_{ij} p_j$$

measure how much p_i fails to obey the master equation. Then we have

$$\begin{aligned} \frac{d}{dt} I(p(t), q) &= \sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i} \right) \\ &\quad + \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{Dp_i}{Dt} \end{aligned}$$

Moreover, the first term is less than or equal to zero.

Proof. For a self-contained proof, see [Information geometry \(part 15\)](#), which is coming up soon. It will be a special case of the theorems there. ■

Blake compares this result to previous work by Schnakenberg:

- J. Schnakenberg, Network theory of microscopic and macroscopic behavior of master equation systems, *Rev. Mod. Phys.* **48** (1976), 571–585.

The negative of Blake's first term is this:

$$K(p) = - \sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i} \right)$$

Under certain circumstances, this equals what Schnakenberg calls the **entropy production**. But a better name for this quantity might be **free energy loss**, since for a closed Markov process that's exactly what it is! In this case there are no boundary states, so the theorem above says $K(p)$ is the rate at which relative entropy — or in other words, free energy — decreases.

For an open Markov process, things are more complicated. The theorem above shows that free energy can also flow in or out at the boundary, thanks to the second term in the formula.

Anyway, the sensible thing is to compare a principle of 'minimum free energy loss' to the principle of minimum dissipation. The principle of minimum dissipation is true. How about the principle of minimum free energy loss? It turns out to be approximately true near equilibrium.

For this, consider the situation in which p is near to the equilibrium distribution q in the sense that

$$\frac{p_i}{q_i} = 1 + \epsilon_i$$

for some small numbers ϵ_i . We collect these numbers in a vector called ϵ .

Theorem 3. Consider an open Markov process with X as its set of states and B as the set of boundary states. Suppose q is a detailed balanced equilibrium and let p be arbitrary. Then

$$K(p) = D(p) + O(\epsilon^2)$$

where $K(p)$ is the free energy loss, $D(p)$ is the dissipation, ϵ_i is defined as above, and by $O(\epsilon^2)$ we mean a sum of terms of order ϵ_i^2 .

Proof. First take the free energy loss:

$$K(p) = - \sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i} \right)$$

Expanding the logarithm to first order in ϵ , we get

$$K(p) = - \sum_{i,j \in X} H_{ij} p_j \left(\frac{p_i}{q_i} - 1 - \frac{p_i q_j}{p_j q_i} \right) + O(\epsilon^2)$$

Since H is infinitesimal stochastic, $\sum_i H_{ij} = 0$, so the second term in the sum vanishes, leaving

$$K(p) = - \sum_{i,j \in X} H_{ij} p_j \left(\frac{p_i}{q_i} - \frac{p_i q_j}{p_j q_i} \right) + O(\epsilon^2)$$

or

$$K(p) = - \sum_{i,j \in X} \left(H_{ij} p_j \frac{p_i}{q_i} - H_{ij} q_j \frac{p_i}{q_i} \right) + O(\epsilon^2)$$

Since q is a equilibrium we have $\sum_j H_{ij} q_j = 0$, so now the last term in the sum vanishes, leaving

$$K(p) = - \sum_{i,j \in X} H_{ij} \frac{p_i p_j}{q_i} + O(\epsilon^2)$$

Next, take the dissipation

$$D(p) = \frac{1}{2} \sum_{i,j} H_{ij} q_j \left(\frac{p_j}{q_j} - \frac{p_i}{q_i} \right)^2$$

and expand the square, getting

$$D(p) = \frac{1}{2} \sum_{i,j} H_{ij} q_j \left(\frac{p_j^2}{q_j^2} - 2 \frac{p_i p_j}{q_i q_j} + \frac{p_i^2}{q_i^2} \right)$$

Since H is infinitesimal stochastic, $\sum_i H_{ij} = 0$. The first term is just this times a function of j , summed over j , so it vanishes, leaving

$$D(p) = \frac{1}{2} \sum_{i,j} H_{ij} q_j \left(-2 \frac{p_i p_j}{q_i q_j} + \frac{p_i^2}{q_i^2} \right)$$

Since q is an equilibrium, $\sum_j H_{ij} q_j = 0$. The last term above is this times a function of i , summed over i , so it vanishes, leaving

$$D(p) = - \sum_{i,j} H_{ij} q_j \frac{p_i p_j}{q_i q_j} = - \sum_{i,j} H_{ij} \frac{p_i p_j}{q_i}$$

This matches what we got for $K(p)$, up to terms of order $O(\epsilon^2)$. ■

In short: detailed balanced open Markov processes are governed by the principle of minimum dissipation, not minimum entropy production. *Minimum dissipation agrees with minimum entropy production only near equilibrium.*

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



A Compositional Framework for Markov Processes

 johncarlosbaez.wordpress.com/2015/09/04/a-compositional-framework-for-markov-processes/

September 4,
2015

This summer my students Brendan Fong and Blake Pollard visited me at the Centre for Quantum Technologies, and we figured out how to understand *open* continuous-time Markov chains! I think this is a nice step towards understanding the math of living systems.

Admittedly, it's just a small first step. But I'm excited by this step, since Blake and I have been trying to get this stuff to work for a couple years, and it finally fell into place. And we think we know what to do next.

Here's our paper:

- John C. Baez, Brendan Fong and Blake S. Pollard, [A compositional framework for open Markov processes](#).

And here's the basic idea...

Open detailed balanced Markov processes

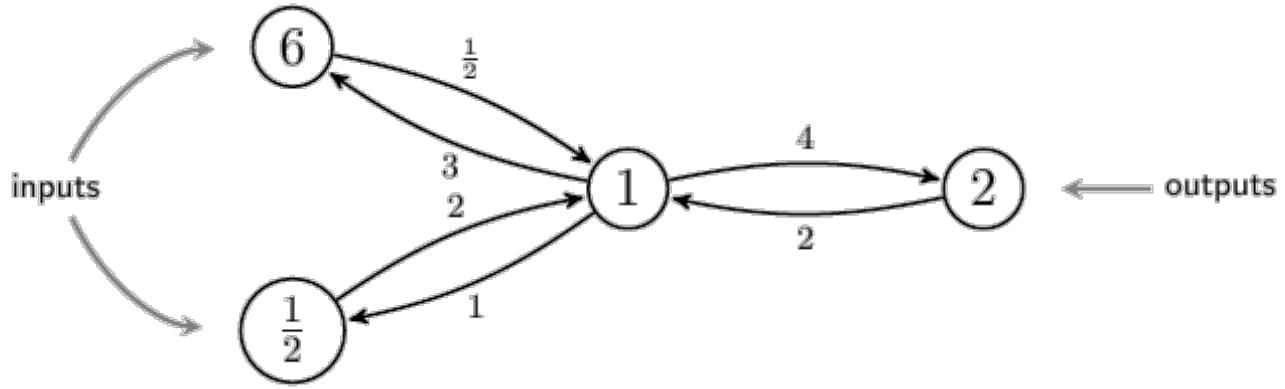
A continuous-time Markov chain is a way to specify the dynamics of a population which is spread across some finite set of states. Population can flow between the states. The larger the population of a state, the more rapidly population flows out of the state. Because of this property, under certain conditions the populations of the states tend toward an equilibrium where at any state the inflow of population is balanced by its outflow.

In applications to statistical mechanics, we are often interested in equilibria such that for any two states connected by an edge, say i and j , the flow from i to j equals the flow from j to i . A continuous-time Markov chain with a chosen equilibrium having this property is called 'detailed balanced'.

I'm getting tired of saying 'continuous-time Markov chain', so from now on I'll just say 'Markov process', just because it's shorter. Okay? That will let me say the next sentence without running out of breath:

Our paper is about *open* detailed balanced Markov processes.

Here's an example:



The detailed balanced Markov process itself consists of a finite set of states together with a finite set of edges between them, with each state i labelled by an equilibrium population $q_i > 0$, and each edge e labelled by a rate constant $r_e > 0$.

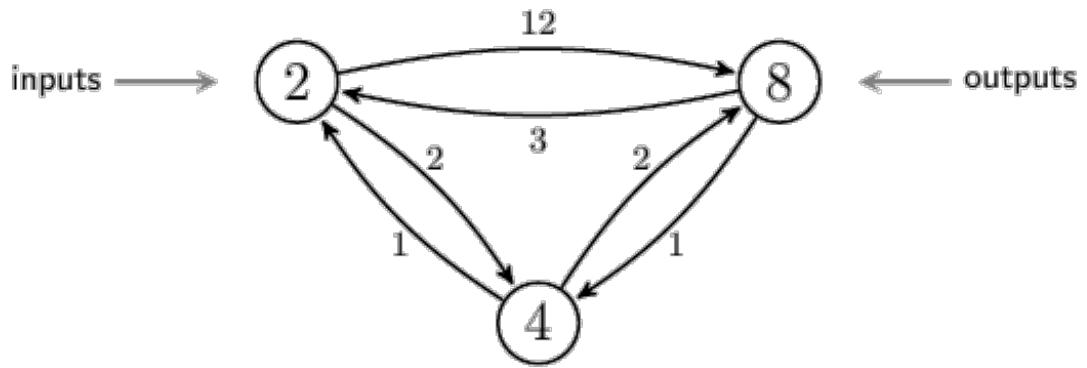
These populations and rate constants are required to obey an equation called the ‘detailed balance condition’. This equation means that in equilibrium, the flow from i to j equal the flow from j to i . Do you see how it works in this example?

To get an ‘open’ detailed balanced Markov process, some states are designated as inputs or outputs. In general each state may be specified as both an input and an output, or as inputs and outputs multiple times. See how that’s happening in this example? It may seem weird, but it makes things work better.

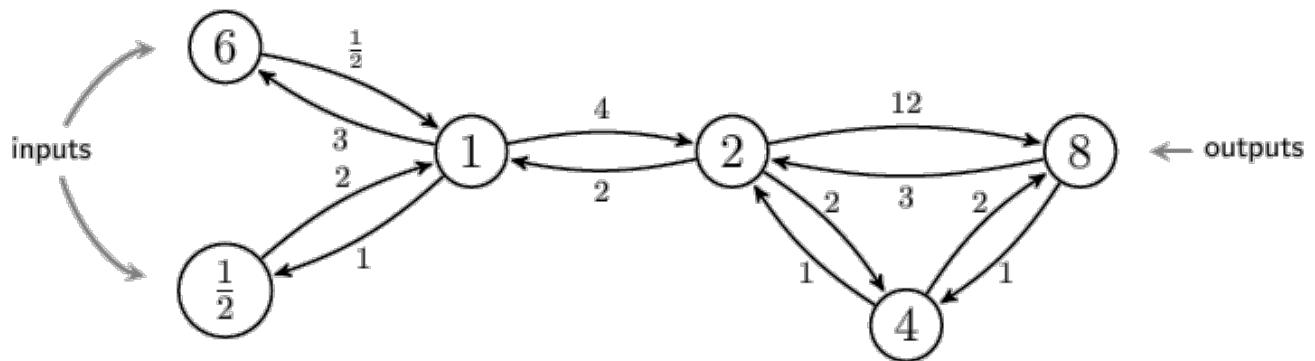
People usually say Markov processes are all about how *probabilities* flow from one state to another. But we work with un-normalized probabilities, which we call ‘populations’, rather than probabilities that must sum to 1. The reason is that in an *open* Markov process, probability is not conserved: it can flow in or out at the inputs and outputs. We allow it to flow both in and out at both the input states and the output states.

Our most fundamental result is that there’s a category DetBalMark where a morphism is an open detailed balanced Markov process. We think of it as a morphism from its inputs to its outputs.

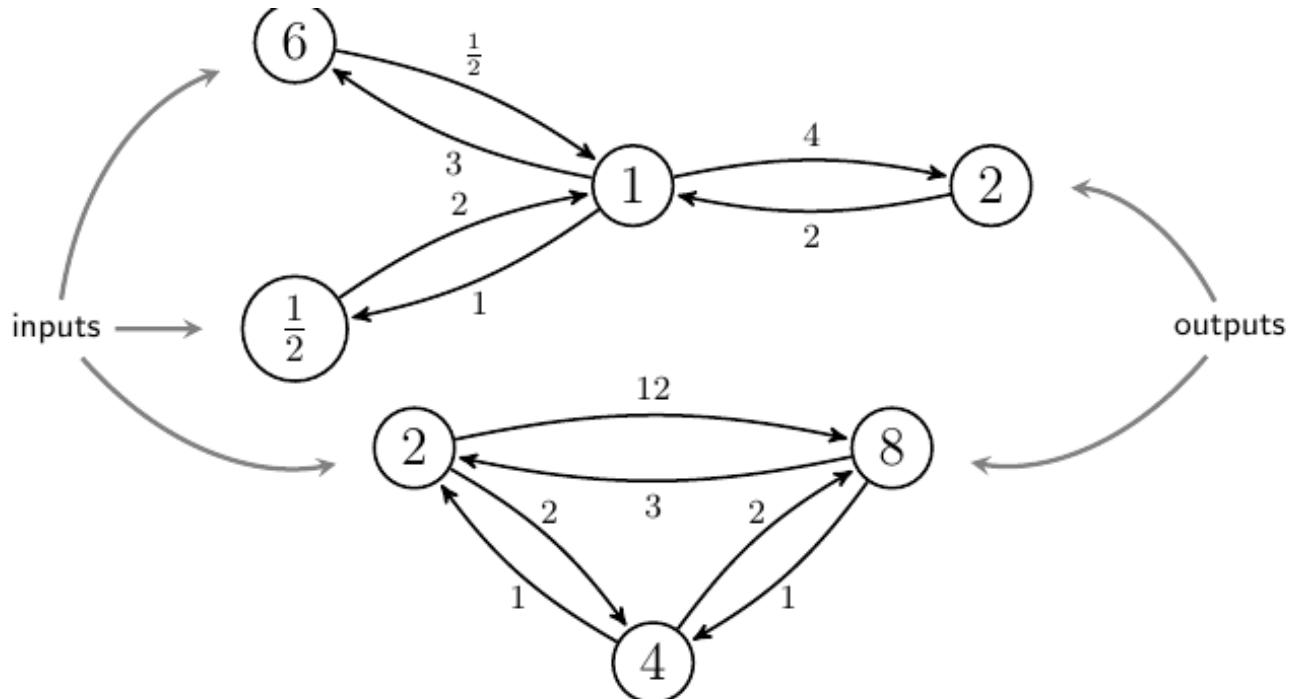
We compose morphisms in DetBalMark by identifying the output states of one open detailed balanced Markov process with the input states of another. The populations of identified states must match. For example, we may compose this morphism N :



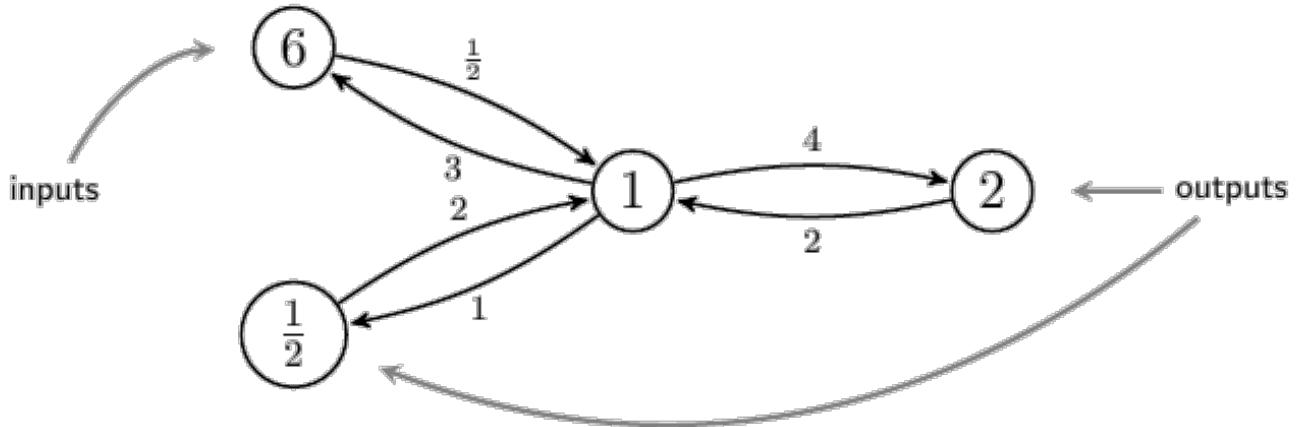
with the previously shown morphism M to get this morphism $M \circ N$:



And here's our second most fundamental result: the category `DetBalMark` is actually a dagger compact category. This lets us do other stuff with open Markov processes. An important one is ‘tensoring’, which lets us take two open Markov processes like M and N above and set them side by side, giving $M \otimes N$:



The so-called compactness is also important. This means we can take some inputs of an open Markov process and turn them into outputs, or vice versa. For example, using the compactness of `DetBalMark` we can get this open Markov process from `M`:



In fact all the categories in our paper are dagger compact categories, and all our functors preserve this structure. Dagger compact categories are a well-known framework for describing systems with inputs and outputs, so this is good.

The analogy to electrical circuits

In a detailed balanced Markov process, population can flow along edges. In the detailed balanced equilibrium, without any flow of population from outside, the flow along from state i to state j will be matched by the flow back from j to i . The populations need to take specific values for this to occur.

In an electrical circuit made of linear resistors, charge can flow along wires. In equilibrium, without any driving voltage from outside, the current along each wire will be zero. The potentials will be equal at every node.

This sets up an analogy between detailed balanced continuous-time Markov chains and electrical circuits made of linear resistors! I love analogy charts, so this makes me very happy:

Circuits	Detailed balanced Markov processes
potential	population
current	flow
conductance	rate constant

power	dissipation
-------	-------------

This analogy is already well known. Schnakenberg used it in his book *Thermodynamic Network Analysis of Biological Systems*. So, our main goal is to formalize and exploit it. This analogy extends from systems in equilibrium to the more interesting case of nonequilibrium steady states, which are the main topic of our paper.

Earlier, Brendan and I introduced a way to ‘black box’ a circuit and define the relation it determines between potential-current pairs at the input and output terminals. This relation describes the circuit’s external behavior as seen by an observer who can only perform measurements at the terminals.

An important fact is that black boxing is ‘compositional’: if one builds a circuit from smaller pieces, the external behavior of the whole circuit can be determined from the external behaviors of the pieces. For category theorists, this means that black boxing is a functor!

Our new paper with Blake develops a similar ‘black box functor’ for detailed balanced Markov processes, and relates it to the earlier one for circuits.

When you black box a detailed balanced Markov process, you get the relation between population-flow pairs at the terminals. (By the ‘flow at a terminal’, we more precisely mean the net population outflow.) This relation holds not only in equilibrium, but also in any nonequilibrium steady state. Thus, *black boxing an open detailed balanced Markov process gives its steady state dynamics as seen by an observer who can only measure populations and flows at the terminals*.

The principle of minimum dissipation

At least since the work of Prigogine, it’s been widely accepted that a large class of systems minimize entropy production in a nonequilibrium steady state. But people still fight about the precise boundary of this class of systems, and even the meaning of this ‘principle of minimum entropy production’.

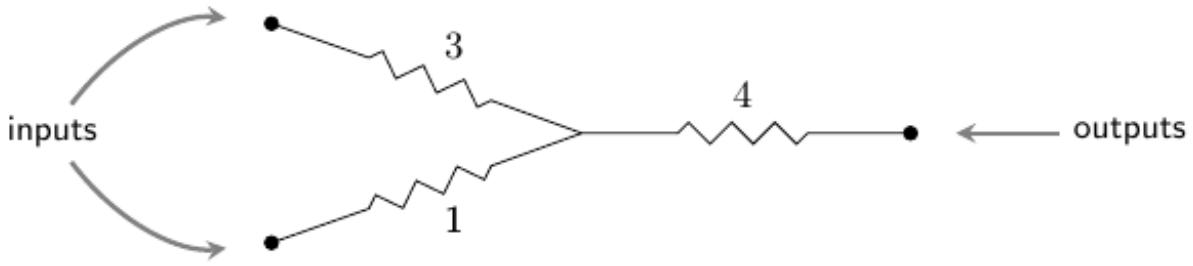
For detailed balanced open Markov processes, we show that a quantity we call the ‘dissipation’ is minimized in any steady state. This is a quadratic function of the populations and flows, analogous to the power dissipation of a circuit made of resistors. We make no claim that this quadratic function actually deserves to be called ‘entropy production’. Indeed, Schnakenberg has convincingly argued that they are only approximately equal.

But still, the ‘dissipation’ function is very natural and useful—and Prigogine’s so-called ‘entropy production’ is also a quadratic function.

Black boxing

I've already mentioned the category DetBalMark , where a morphism is an open detailed balanced Markov process. But our paper needs two more categories to tell its story! There's the category of circuits, and the category of linear relations.

A morphism in the category Circ is an open electrical circuit made of resistors: that is, a graph with each edge labelled by a 'conductance' $c_e > 0$, together with specified input and output nodes:



A morphism in the category LinRel is a linear relation $L : U \rightsquigarrow V$ between finite-dimensional real vector spaces U and V . This is nothing but a linear subspace $L \subseteq U \oplus V$. Just as relations generalize functions, linear relations generalize linear functions!

In our [previous paper](#), Brendan and I introduced these two categories and a functor between them, the 'black box functor':

The idea is that any circuit determines a linear relation between the potentials and net current flows at the inputs and outputs. This relation describes the behavior of a circuit of resistors as seen from outside.

$$\blacksquare : \text{Circ} \rightarrow \text{LinRel}$$

Our new paper introduces a black box functor for detailed balanced Markov processes:

We draw this functor as a white box merely to distinguish it from the other black box functor. The functor \square maps any detailed balanced Markov process to the linear relation obeyed by populations and flows at the inputs and outputs in a steady state. In short, it describes the steady state behavior of the Markov process 'as seen from outside'.

$$\square : \text{DetBalMark} \rightarrow \text{LinRel}$$

How do we manage to black box detailed balanced Markov processes? We do it using the analogy with circuits!

The analogy becomes a functor

Every analogy wants to be a functor. So, we make the analogy between detailed balanced Markov processes and circuits precise by turning it into a functor:

This functor converts any open detailed balanced Markov process into an open electrical circuit made of resistors. This circuit is carefully chosen to reflect the steady-state behavior of the Markov process. Its underlying graph is the same as that of the Markov process. So, the ‘states’ of the Markov process are the same as the ‘nodes’ of the circuit.

$$K : \text{DetBalMark} \rightarrow \text{Circ}$$

Both the equilibrium populations at states of the Markov process and the rate constants labelling edges of the Markov process are used to compute the conductances of edges of this circuit. In the simple case where the Markov process has exactly one edge from any state i to any state j , the rule is this:

$$C_{ij} = H_{ij}q_j$$

where:

- q_j is the equilibrium population of the j th state of the Markov process,
- H_{ij} is the rate constant for the edge from the j th state to the i th state of the Markov process, and
- C_{ij} is the conductance (that is, the reciprocal of the resistance) of the wire from the j th node to the i th node of the resulting circuit.

The detailed balance condition for Markov processes says precisely that the matrix C_{ij} is symmetric! This is just right for an electrical circuit made of resistors, since it means that the resistance of the wire from node i to node j equals the resistance of the same wire in the reverse direction, from node j to node i .

A triangle of functors

If you paid careful attention, you’ll have noticed that I’ve described a triangle of functors:

And if you know anything about how category theorists think, you'll be wondering if this diagram commutes.

In fact, this triangle of functors does not commute! However, a general lesson of category theory is that we should only expect diagrams of functors to commute *up to natural isomorphism*, and this is what happens here:

The natural transformation α 'corrects' the black box functor for resistors to give the one for detailed balanced Markov processes.

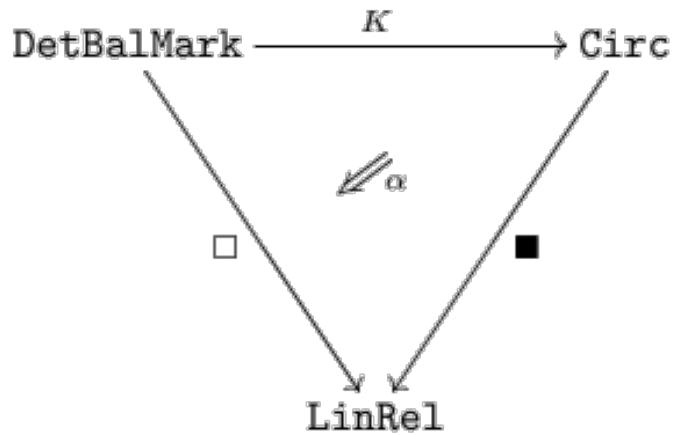
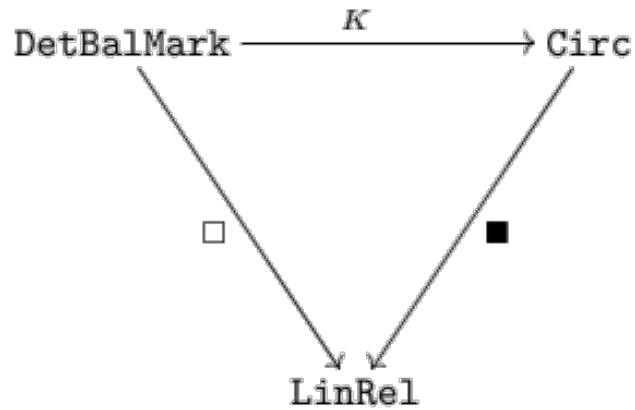
The functors \square and $\blacksquare \circ K$ are actually equal on objects. An object in DetBalMark is a finite set X with each element $i \in X$ labelled a positive populations q_i . Both functors map this object to the vector space $\mathbb{R}^X \oplus \mathbb{R}^X$. For the functor \square , we think of this as a space of population-flow pairs. For the functor $\blacksquare \circ K$, we think of it as a space of potential-current pairs. The natural transformation α then gives a linear relation

in fact an isomorphism of vector spaces, which converts potential-current pairs into population-flow pairs in a manner that depends on the q_i . I'll skip the formula; it's in the paper.

$$\alpha_{X,q} : \mathbb{R}^X \oplus \mathbb{R}^X \rightsquigarrow \mathbb{R}^X \oplus \mathbb{R}^X$$

But here's the key point. The naturality of α actually allows us to reduce the problem of computing the functor \square to the problem of computing \blacksquare . Suppose

is any morphism in DetBalMark . The object (X, q) is some finite set X labelled by populations q , and (Y, r) is some finite set Y labelled by populations r . Then the naturality of α means that this square commutes:



$$M : (X, q) \rightarrow (Y, r)$$

Since $\alpha_{X,q}$ and $\alpha_{Y,r}$ are isomorphisms, we can solve for the functor \square as follows:

This equation has a clear intuitive meaning! It says that to compute the behavior of a detailed balanced Markov process, namely $\square(f)$, we convert it into a circuit made of resistors and compute the behavior of that, namely $\blacksquare K(f)$. This is not *equal* to the behavior of the Markov process, but we can compute that behavior by converting the input populations and flows into potentials and currents, feeding them into our circuit, and then converting the outputs back into populations and flows.

$$\begin{array}{ccc}
 \mathbb{R}^X \oplus \mathbb{R}^X & \xrightarrow{\blacksquare K(M)} & \mathbb{R}^Y \oplus \mathbb{R}^Y \\
 \downarrow \alpha_{X,q} & & \downarrow \alpha_{Y,r} \\
 \mathbb{R}^X \oplus \mathbb{R}^X & \xrightarrow{\square(M)} & \mathbb{R}^Y \oplus \mathbb{R}^Y
 \end{array}$$

$\square(M) = \alpha_Y \circ \blacksquare K(M) \circ \alpha_X^{-1}$

What we really do

So that's a sketch of what we do, and I hope you ask questions if it's not clear. But I also hope you read our paper! Here's what we actually do in there. After an introduction and summary of results:

- Section 3 defines open Markov processes and the open master equation.
- Section 4 introduces detailed balance for open Markov processes.
- Section 5 recalls the principle of minimum power for open circuits made of linear resistors, and explains how to black box them.
- Section 6 introduces the principle of minimum dissipation for open detailed balanced Markov processes, and describes how to black box these.
- Section 7 states the analogy between circuits and detailed balanced Markov processes in a formal way.
- Section 8 describes how to compose open Markov processes, making them into the morphisms of a category.
- Section 9 does the same for detailed balanced Markov processes.
- Section 10 describes the 'black box functor' that sends any open detailed balanced Markov process to the linear relation describing its external behavior, and recalls the similar functor for circuits.

- Section 11 makes the analogy between open detailed balanced Markov processes and open circuits even more formal, by making it into a functor. We prove that together with the two black box functors, this forms a triangle that commutes up to natural isomorphism.
- Section 12 is about geometric aspects of this theory. We show that the linear relations in the image of these black box functors are Lagrangian relations between symplectic vector spaces. We also show that the master equation can be seen as a gradient flow equation.
- Section 13 is a summary of what we have learned.

Finally, Appendix A is a quick tutorial on decorated cospans. This is a key mathematical tool in our work, developed by Brendan in [an earlier paper](#).

A Second Law for Open Markov Processes

 johncarlosbaez.wordpress.com/2014/11/15/a-second-law-for-open-markov-processes/

November 15,
2014

*guest post by **Blake Pollard***

What comes to mind when you hear the term ‘random process’? Do you think of Brownian motion? Do you think of particles hopping around? Do you think of a drunkard staggering home?

Today I’m going to tell you about a version of the drunkard’s walk with a few modifications. Firstly, we don’t have just one drunkard: we can have any positive real number of drunkards. Secondly, our drunkards have no memory; where they go next doesn’t depend on where they’ve been. Thirdly, there are special places, such as entrances to bars, where drunkards magically appear and disappear.

The second condition says that our drunkards satisfy the Markov property, making their random walk into a **Markov process**. The third condition is really what I want to tell you about, because it makes our Markov process into a more general ‘open Markov process’.

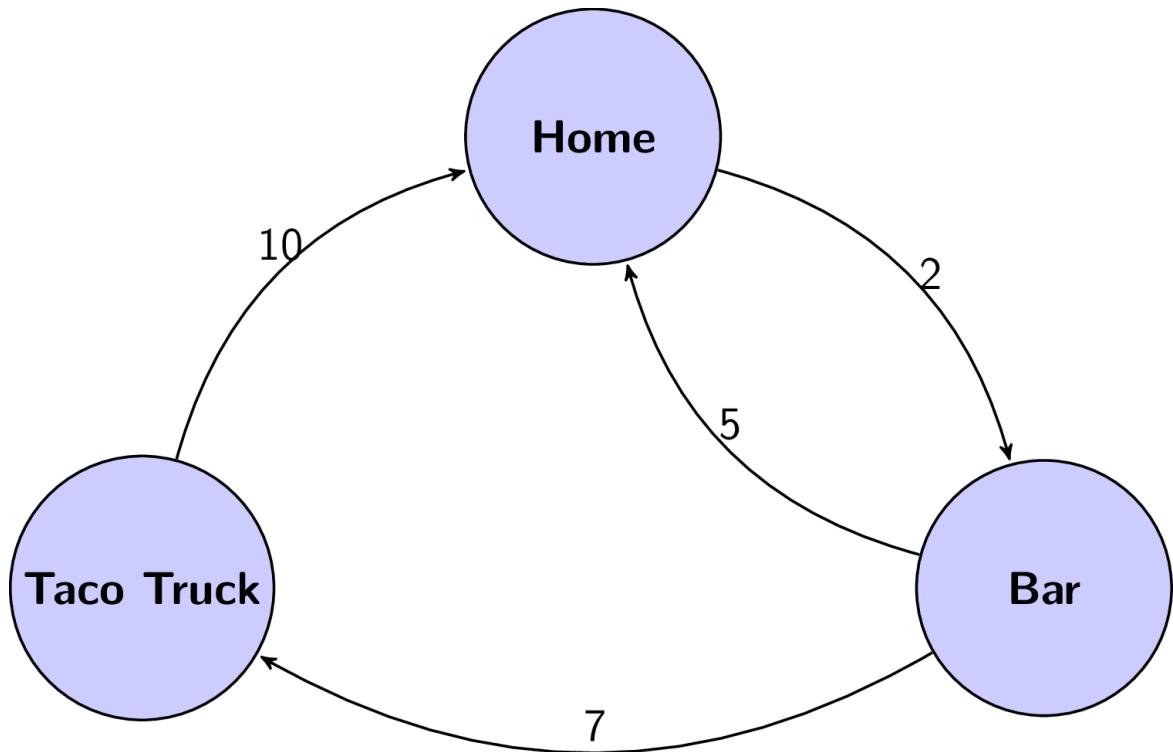
There are a collection of places the drunkards can be, for example:

We call this set V the set of **states**. There are certain probabilities associated with traveling between these places. We call

these **transition rates**. For example it is more likely for a drunkard to go from the bar to the taco truck than to go from the bar to home so the transition rate between the bar and the taco truck should be greater than the transition rate from the bar to home. Sometimes you can’t get from one place to another without passing through intermediate places. In reality the drunkard can’t go directly from the bar to the taco truck: he or she has to go from the bar to sidewalk to the taco truck.

$$V = \{\text{bar, sidewalk, street, taco truck, home}\}$$

This information can all be summarized by drawing a directed graph where the positive numbers labelling the edges are the transition rates:



For simplicity we draw only three states: home, bar, taco truck. Drunkards go from home to the bar and back, but they never go straight from home to the taco truck.

We can keep track of where all of our drunkards are using a vector with 3 entries:

We call this our **population distribution**. The first entry p_h is the number of drunkards that are at home, the second p_b is how many are at the bar, and the third p_{tt} is how many are at the taco truck.

$$p(t) = \begin{pmatrix} p_h(t) \\ p_b(t) \\ p_{tt}(t) \end{pmatrix} \in \mathbb{R}^3$$

There is a set of coupled, linear, first-order differential equations we can write down using the information in our graph that tells us how the number of drunkards in each place change with time. This is called the **master equation**:

$$\frac{dp}{dt} = Hp$$

where H is a 3×3 matrix which we call the **Hamiltonian**. The off-diagonal entries are nonnegative:

$$H_{ij} \geq 0, i \neq j$$

and the columns sum to zero:

$$\sum_i H_{ij} = 0$$

We call a matrix satisfying these conditions **infinitesimal stochastic**. **Stochastic** matrices have columns that sum to one. If we take the exponential of an infinitesimal stochastic matrix we get one whose columns sum to one, hence the label ‘infinitesimal’.

The Hamiltonian for the graph above is

John has written a lot about Markov processes and infinitesimal stochastic Hamiltonians in previous posts.

$$H = \begin{pmatrix} -2 & 5 & 10 \\ 2 & -12 & 0 \\ 0 & 7 & -10 \end{pmatrix}$$

Given two vectors $p, q \in \mathbb{R}^3$ describing the populations of drunkards which obey the same master equation, we can calculate the **relative entropy** of p relative to q :

This is an example of a ‘divergence’. In statistics, a divergence is a way of measuring the distance between probability distributions, which may not be symmetrical and may even not obey the triangle inequality.

$$S(p, q) = \sum_{i \in V} p_i \ln \left(\frac{p_i}{q_i} \right)$$

The relative entropy is important because it decreases monotonically with time, making it a Lyapunov function for Markov processes. Indeed, it is a well known fact that

This is true for any two population distributions which evolve according to the same master equation, though you have to allow infinity as a possible value for the relative entropy and negative infinity for its time derivative.

$$\frac{dS(p(t), q(t))}{dt} \leq 0$$

Why is entropy *decreasing*? Doesn’t the Second Law of Thermodynamics say entropy *increases*?

Don’t worry: the reason is that I have not put a minus sign in my definition of relative entropy. Put one in if you like, and then it will increase. Sometimes without the minus sign it’s called the Kullback–Leibler divergence. This *decreases* with the passage of time, saying that any two population distributions $p(t)$ and $q(t)$ get ‘closer together’ as they get randomized with the passage of time.

That itself is a nice result, but I want to tell you what happens when you allow drunkards to appear and disappear at certain states. Drunkards appear at the bar once they’ve had enough to drink and once they are home for long enough they can disappear. The set of places where drunkards can appear or disappear B is called the set of **boundary states**. So for the above process

is the set of boundary states. This changes the way in which the population of drunkards changes with time!

$$B = \{\text{home, bar}\}$$

The drunkards at the taco truck obey the master equation. For them,

still holds. But because the populations can appear or disappear at the boundary states the master equation no longer holds at those states! Instead it is useful to define the flow of drunkards into the i^{th} state by

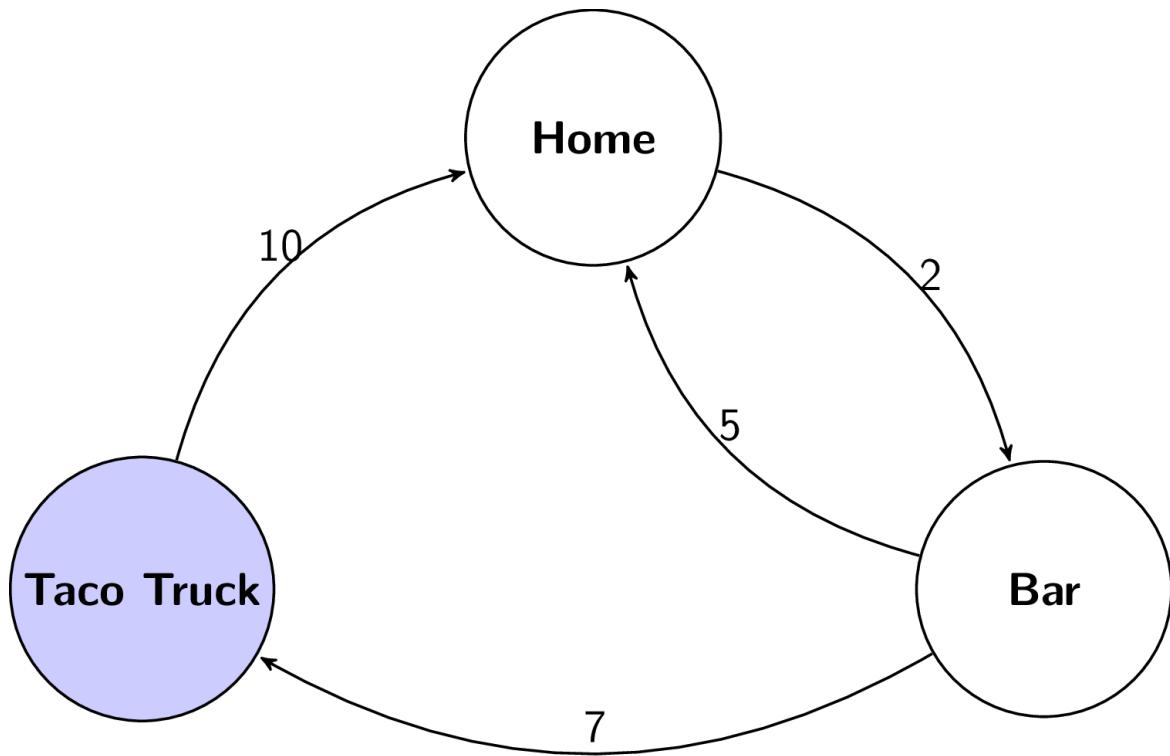
$$\frac{dp_{tt}}{dt} = 7p_b - 10p_{tt}$$

This quantity describes by how much the rate of change of the populations at the boundary states differ from that given by the master equation.

$$\frac{Dp_i}{Dt} = \frac{dp_i}{dt} - \sum_j H_{ij}p_j$$

The reason why we are interested in open Markov processes is because you can take two open Markov processes and glue them together along some subset of their boundary states to get a new open Markov process! This allows us to build up or break down complicated Markov processes using open Markov processes as the building blocks.

For example we can draw the graph corresponding to the drunkards' walk again, only now we will distinguish boundary states from internal states by coloring internal states blue and having boundary states be white:



Consider another open Markov process with states

where

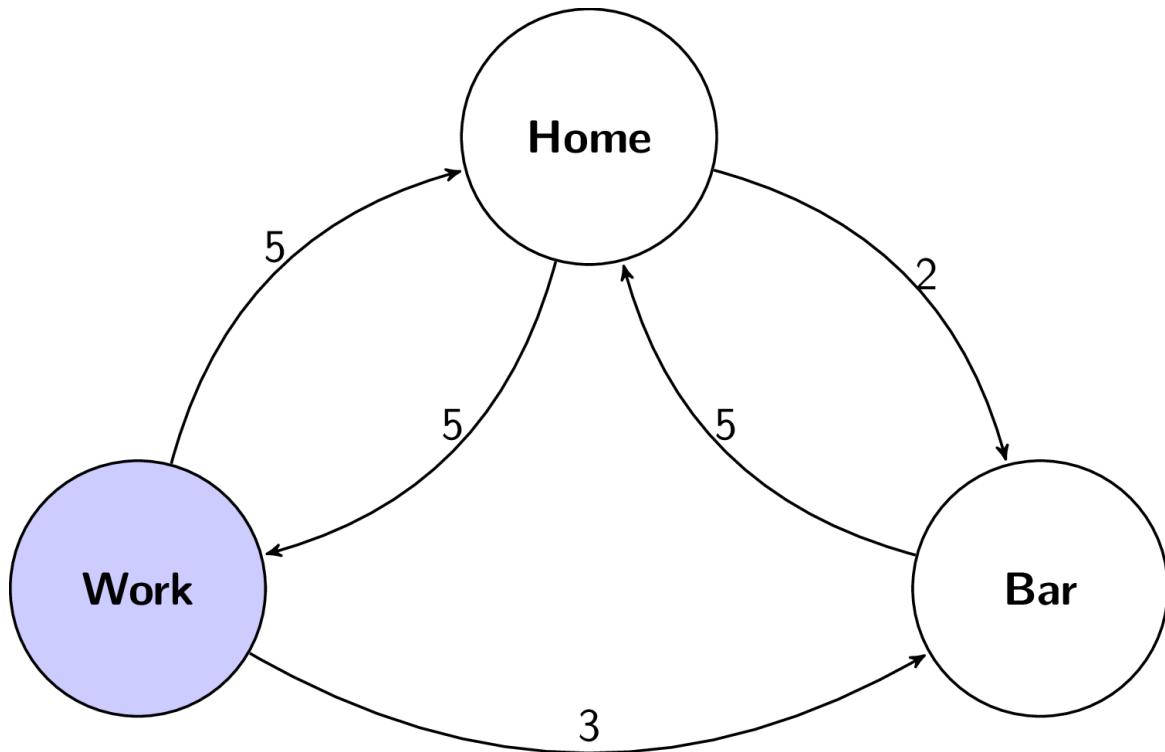
$$V = \{\text{home, work, bar}\}$$

are the boundary states, leaving

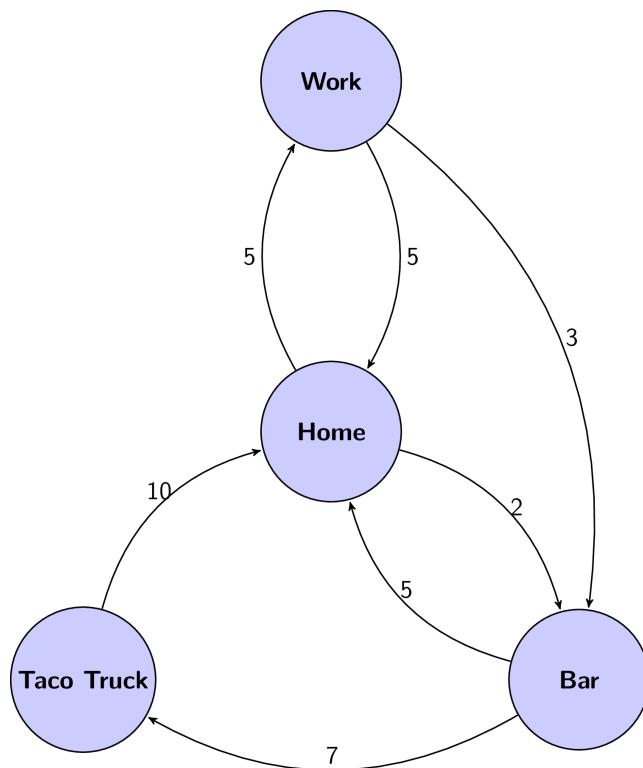
$$I = \{\text{work}\}$$

$$B = \{\text{home, bar}\}$$

as an internal state:



Since the boundary states of this process overlap with the boundary states of the first process we can compose the two to form a new Markov process:



Notice the boundary states are now internal states. I hope any Markov process that could approximately model your behavior has more interesting nodes! There is a nice way to figure out the Hamiltonian of the composite from the Hamiltonians of the pieces, but we will leave that for another time.

We can ask ourselves, how does relative entropy change with time in open Markov processes? You can read my paper for the details, but here is the punchline:

This is a version of the Second Law of Thermodynamics for open Markov processes.

$$\frac{dS(p(t), q(t))}{dt} \leq \sum_{i \in B} \frac{Dp_i}{Dt} \frac{\partial S}{\partial p_i} + \frac{Dq_i}{Dt} \frac{\partial S}{\partial q_i}$$

It is important to notice that the sum is only over the boundary states! This inequality tells us that relative entropy still decreases inside our process, but depending on the flow of populations through the boundary states the relative entropy of the whole process could either increase or decrease! This inequality will be important when we study how the relative entropy changes in different parts of a bigger more complicated process.

That is all for now, but I leave it as an exercise for you to imagine a Markov process that describes your life. How many states does it have? What are the relative transition rates? Are there states you would like to spend more or less time in? Are there states somewhere you would like to visit?

Here is my paper, which proves the above inequality:

- Blake Pollard, [A Second Law for open Markov processes](#), *Open Systems and Information Dynamics* **23** (2016), 1650006.

If you have comments or corrections, let me know!



January 14, 2016

Information Geometry (Part 15)

John Baez and [Blake Pollard](#)

Lately we've been thinking about open Markov processes. These are random processes where something can hop randomly from one state to another (that's the 'Markov process' part) but also enter or leave the system (that's the 'open' part).

The ultimate goal is to understand the nonequilibrium thermodynamics of open systems — systems where energy and maybe matter flows in and out — well enough to understand *in detail* how life works. That's a difficult job! But one has to start somewhere, and this is one place to start.

We have a few papers on this subject:

- Blake Pollard, [A Second Law for open Markov processes](#). (Blog article [here](#).)
- John Baez, Brendan Fong and Blake Pollard, [A compositional framework for Markov processes](#). (Blog article [here](#).)
- Blake Pollard, [Open Markov processes: A compositional perspective on non-equilibrium steady states in biology](#). (Blog article [NOT YET](#).)

However, right now we just want to show you three closely connected results about how relative entropy changes in open Markov processes.

Definitions

An **open Markov process**, is a triple (X, B, H) where X is a finite set of **state**, $B \subseteq X$ is the subset of **boundary states**, and $H : \mathbf{R}^X \rightarrow \mathbf{R}^X$ is an **infinitesimal stochastic** operator, meaning a linear operator with

$$H_{ij} \geq 0, \quad i \neq j$$

and

$$\sum_i H_{ij} = 0$$

For each $i \in X$ we introduce **population** $p_i \in [0, \infty)$. We call the resulting function $p : X \rightarrow [0, \infty)$ the **population distribution**. Populations evolve in time according to the **open master equation**:

$$\frac{dp_i}{dt} = \sum_j H_{ij} p_j, \quad i \in X - B$$

$$p_i(t) = b_i(t), \quad i \in B$$

So, the populations p_i obey a linear differential equation at states i that are not in the boundary, but they are specified 'by the user' to be chosen functions b_i at the boundary states.

The off-diagonal entries H_{ij} , $i \neq j$ are the rates at which population transitions from the j th to the i th state. A **steady state** distribution is a population distribution which is constant in time:

$$\frac{dp_i}{dt} = 0 \quad \text{for all } i \in X$$

A **closed Markov process**, or continuous-time discrete-state Markov chain, is an open Markov process whose boundary is empty. For a closed Markov process, the open master equation becomes the usual **master equation**:

$$\frac{dp}{dt} = Hp$$

In a closed Markov process the total population is conserved:

$$\frac{d}{dt} \sum_{i \in X} p_i = \sum_{i,j} H_{ij} p_j = 0$$

This lets us normalize the initial total population to 1 and have it stay equal to 1. If we do this, we can talk about *probabilities* instead of populations. In an open Markov process, population can flow in and out at the boundary states.

A steady-state distribution in a closed Markov process is typically called an **equilibrium**. We say an equilibrium $q \in [0, \infty)^X$ of a Markov process is **detailed balanced** if

$$H_{ij} q_j = H_{ji} q_i \quad \text{for all } i, j \in X.$$

Given two population distributions

$$p, q : X \rightarrow [0, \infty)$$

we can define the **relative entropy**

$$I(p, q) = \sum_i p_i \ln \left(\frac{p_i}{q_i} \right)$$

There are some nice results about how this changes with time. When q is a detailed balanced equilibrium solution of the master equation, the relative entropy can be seen as the 'free energy' of p . For a precise statement, see Section 4 of [Relative entropy in biological systems](#).

The Second Law of Thermodynamics implies that the free energy of a closed system tends to decrease with time, so for *closed* Markov processes we expect $I(p, q)$ to be nonincreasing. And this is true! But for *open* Markov processes, free energy can flow in from outside.

Results

Theorem 1. Consider an open Markov process with X as its set of states and B as the set of boundary states. Suppose $p(t)$ and $q(t)$ obey the open master equation, and let the quantities

$$\frac{Dp_i}{Dt} = \frac{dp_i}{dt} - \sum_{j \in X} H_{ij} p_j$$

$$\frac{Dq_i}{Dt} = \frac{dq_i}{dt} - \sum_{j \in X} H_{ij} q_j$$

measure how much the time derivatives of p_i and q_i fail to obey the master equation. Then we have

$$\begin{aligned} \frac{d}{dt} I(p(t), q(t)) &= \sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i} \right) \\ &\quad + \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{Dp_i}{Dt} + \frac{\partial I}{\partial q_i} \frac{Dq_i}{Dt} \end{aligned}$$

This result separates the change in relative entropy change into two parts: an 'internal' part and a 'boundary' part.

It turns out the 'internal' part is always less than or equal to zero. So, from Theorem 1 we can deduce a version of the Second Law of Thermodynamics for open Markov processes:

Theorem 2. Given the conditions of Theorem 1, we have

$$\frac{d}{dt} I(p(t), q(t)) \leq \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{Dp_i}{Dt} + \frac{\partial I}{\partial q_i} \frac{Dq_i}{Dt}$$

Intuitively, this says that free energy can only increase if it comes in from the boundary!

There is another nice result that holds when q is an equilibrium solution of the master equation. This idea seems to go back to Schnakenberg:

Theorem 3. Given the conditions of Theorem 1, suppose also that q is an equilibrium solution of the master equation. Then we have

$$\frac{d}{dt} I(p(t), q) = \frac{1}{2} \sum_{i,j \in X} J_{ij} A_{ij} + \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{Dp_i}{Dt}$$

where

$$J_{ij} = H_{ij} p_j - H_{ji} p_i$$

is the **flux** from j to i , while

$$A_{ij} = \ln\left(\frac{p_i q_j}{p_j q_i}\right)$$

is the conjugate **thermodynamic force**.

The flux J_{ij} has a nice meaning: it's the net flow of population from j to i . The thermodynamic force is a bit subtler, but this theorem reveals its meaning: its how much free energy increase is caused by a flow from j to i . We should probably include a minus sign, since free energy wants to *decrease*.

Proofs

Proof of Theorem 1. We begin by taking the time derivative of the relative information:

$$\frac{d}{dt} I(p(t), q(t)) = \sum_{i \in X} \frac{\partial I}{\partial p_i} \frac{dp_i}{dt} + \frac{\partial I}{\partial q_i} \frac{dq_i}{dt}$$

We can separate this into a sum over states $i \in X - B$, for which the time derivatives of p_i and q_i are given by the master equation, and boundary states $i \in B$, for which they are not:

$$\begin{aligned} \frac{d}{dt} I(p(t), q(t)) &= \sum_{i \in X-B, j \in X} \frac{\partial I}{\partial p_i} H_{ij} p_j + \frac{\partial I}{\partial q_i} H_{ij} q_j \\ &\quad + \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{dp_i}{dt} + \frac{\partial I}{\partial q_i} \frac{dq_i}{dt} \end{aligned}$$

For boundary states we have

$$\frac{dp_i}{dt} = \frac{Dp_i}{Dt} + \sum_{j \in X} H_{ij} p_j$$

and similarly for the time derivative of q_i . We thus obtain

$$\begin{aligned} \frac{d}{dt} I(p(t), q(t)) &= \sum_{i,j \in X} \frac{\partial I}{\partial p_i} H_{ij} p_j + \frac{\partial I}{\partial q_i} H_{ij} q_j \\ &\quad + \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{Dp_i}{Dt} + \frac{\partial I}{\partial q_i} \frac{Dq_i}{Dt} \end{aligned}$$

To evaluate the first sum, recall that

$$I(p, q) = \sum_{i \in X} p_i \ln\left(\frac{p_i}{q_i}\right)$$

so

$$\frac{\partial I}{\partial p_i} = 1 + \ln\left(\frac{p_i}{q_i}\right), \quad \frac{\partial I}{\partial q_i} = -\frac{p_i}{q_i}$$

Thus, we have

$$\sum_{i,j \in X} \frac{\partial I}{\partial p_i} H_{ij} p_j + \frac{\partial I}{\partial q_i} H_{ij} q_j = \sum_{i,j \in X} \left(1 + \ln\left(\frac{p_i}{q_i}\right)\right) H_{ij} p_j - \frac{p_i}{q_i} H_{ij} q_j$$

We can rewrite this as

$$\sum_{i,j \in X} H_{ij} p_j \left(1 + \ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i}\right)$$

Since H_{ij} is infinitesimal stochastic we have $\sum_i H_{ij} = 0$, so the first term drops out, and we are left with

$$\sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i}\right)$$

as desired. ■

Proof of Theorem 2. Thanks to Theorem 1, to prove

$$\frac{d}{dt} I(p(t), q(t)) \leq \sum_{i \in B} \frac{\partial I}{\partial p_i} \frac{Dp_i}{Dt} + \frac{\partial I}{\partial q_i} \frac{Dq_i}{Dt}$$

it suffices to show that

$$\sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i} \right) \leq 0$$

or equivalently (recalling the proof of Theorem 1):

$$\sum_{i,j} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) + 1 - \frac{p_i q_j}{p_j q_i} \right) \leq 0$$

The last two terms on the left hand side cancel when $i = j$. Thus, if we break the sum into an $i \neq j$ part and an $i = j$ part, the left side becomes

$$\sum_{i \neq j} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) + 1 - \frac{p_i q_j}{p_j q_i} \right) + \sum_j H_{jj} p_j \ln\left(\frac{p_j}{q_j}\right)$$

Next we can use the infinitesimal stochastic property of H to write H_{jj} as the sum of $-H_{ij}$ over i not equal to j , obtaining

$$\sum_{i \neq j} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) + 1 - \frac{p_i q_j}{p_j q_i} \right) - \sum_{i \neq j} H_{ij} p_j \ln\left(\frac{p_j}{q_j}\right) = \sum_{i \neq j} H_{ij} p_j \left(\ln\left(\frac{p_i q_j}{p_j q_i}\right) + 1 - \frac{p_i q_j}{p_j q_i} \right)$$

Since $H_{ij} \geq 0$ when $i \neq j$ and $\ln(s) + 1 - s \leq 0$ for all $s > 0$, we conclude that this quantity is ≤ 0 . ■

Proof of Theorem 3. Now suppose also that q is an equilibrium solution of the master equation. Then $Dq_i/Dt = dq_i/dt = 0$ for all states i , so by Theorem 1 we need to show

$$\sum_{i,j \in X} H_{ij} p_j \left(\ln\left(\frac{p_i}{q_i}\right) - \frac{p_i q_j}{p_j q_i} \right) = \frac{1}{2} \sum_{i,j \in X} J_{ij} A_{ij}$$

We also have $\sum_{j \in X} H_{ij} q_j = 0$, so the second term in the sum at left vanishes, and it suffices to show

$$\sum_{i,j \in X} H_{ij} p_j \ln\left(\frac{p_i}{q_i}\right) = \frac{1}{2} \sum_{i,j \in X} J_{ij} A_{ij}$$

By definition we have

$$\frac{1}{2} \sum_{i,j} J_{ij} A_{ij} = \frac{1}{2} \sum_{i,j} (H_{ij} p_j - H_{ji} p_i) \ln\left(\frac{q_j p_i}{q_i p_j}\right)$$

This in turn equals

$$\frac{1}{2} \sum_{i,j} H_{ij} p_j \ln \left(\frac{q_j p_i}{q_i p_j} \right) - \frac{1}{2} \sum_{i,j} H_{ji} p_i \ln \left(\frac{q_j p_i}{q_i p_j} \right)$$

and we can switch the dummy indices i, j in the second sum, obtaining

$$\frac{1}{2} \sum_{i,j} H_{ij} p_j \ln \left(\frac{q_j p_i}{q_i p_j} \right) - \frac{1}{2} \sum_{i,j} H_{ij} p_j \ln \left(\frac{q_i p_j}{q_j p_i} \right)$$

or simply

$$\sum_{i,j} H_{ij} p_j \ln \left(\frac{q_j p_i}{q_i p_j} \right)$$

But this is

$$\sum_{i,j} H_{ij} p_j \left(\ln\left(\frac{q_j}{p_j}\right) + \ln\left(\frac{p_i}{q_i}\right) \right)$$

and the first term vanishes because H is infinitesimal stochastic: $\sum_i H_{ij} = 0$. We thus have

$$\frac{1}{2} \sum_{i,j} J_{ij} A_{ij} = \sum_{i,j} H_{ij} p_j \ln\left(\frac{p_i}{q_i}\right)$$

as desired. ■

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2016 John Baez

baez@math.removethis.ucr.andthis.edu

[home](#)





February 1, 2017

Information Geometry (Part 16)

John Baez

This week I'm giving a talk on biology and information:

- John Baez, [Biology as information dynamics](#), talk for [Biological Complexity: Can it be Quantified?](#), a workshop at the [Beyond Center](#), 2 February 2017.

While preparing this talk, I discovered a cool fact. I doubt it's new, but I haven't exactly seen it elsewhere. I came up with it while trying to give a precise and general statement of 'Fisher's fundamental theorem of natural selection'. I *won't* start by explaining that theorem, since my version looks rather different than Fisher's, and I came up with mine precisely because I had trouble understanding his. I'll say a bit more about this at the end.

Here's my version:

The square of the rate at which a population learns information is the variance of its fitness.

This is a nice advertisement for the virtues of diversity: more variance means faster learning. But it requires some explanation!

The setup

Let's start by assuming we have n different kinds of self-replicating entities with populations P_1, \dots, P_n . As usual, these could be all sorts of things:

- molecules of different chemicals
- organisms belonging to different species
- genes of different alleles
- restaurants belonging to different chains
- people with different beliefs
- game-players with different strategies
- etc.

I'll call them **replicators** of different **species**.

Let's suppose each population P_i is a function of time that grows at a rate equal to this population times its 'fitness'. I explained the resulting equation back in [Part 9](#), but it's pretty simple:

$$\frac{d}{dt}P_i(t) = f_i(P_1(t), \dots, P_n(t)) P_i(t)$$

Here f_i is a completely arbitrary smooth function of all the populations! We call it the **fitness** of the i th species.

This equation is important, so we want a short way to write it. I'll often write $f_i(P_1(t), \dots, P_n(t))$ simply as f_i , and $P_i(t)$ simply as P_i . With these abbreviations, which any red-blooded physicist would take for granted, our equation becomes simply this:

$$\frac{dP_i}{dt} = f_i P_i$$

Next, let $p_i(t)$ be the probability that a randomly chosen organism is of the i th species:

$$p_i(t) = \frac{P_i(t)}{\sum_j P_j(t)}$$

Starting from our equation describing how the populations evolve, we can figure out how these probabilities evolve. The answer is called the [replicator equation](#):

$$\frac{dp_i}{dt} = (f_i - \langle f \rangle) p_i(t)$$

Here $\langle f \rangle$ is the average fitness of all the replicators, or **mean fitness**:

$$\langle f \rangle = \sum_j f_j(P_1(t), \dots, P_n(t)) p_j(t)$$

In what follows I'll abbreviate the replicator equation as follows:

$$\frac{dp_i}{dt} = (f_i - \langle f \rangle) p_i$$

The result

Okay, now let's figure out how fast the probability distribution

$$p(t) = (p_1(t), \dots, p_n(t))$$

changes with time. For this we need to choose a way to measure the length of the vector

$$\frac{dp}{dt} = \left(\frac{d}{dt} p_1(t), \dots, \frac{d}{dt} p_n(t) \right)$$

And here information geometry comes to the rescue! We can use the [Fisher information metric](#), which is a Riemannian metric on the space of probability distributions.

I've talked about the Fisher information metric in many ways in this series. The most important fact is that as a probability distribution $p(t)$ changes with time, its speed

$$\left\| \frac{dp}{dt} \right\|$$

as measured using the Fisher information metric can be seen as the *rate at which information is learned*. I'll explain that later. Right now I just want a simple *formula* for the Fisher information metric. Suppose v and w are two tangent vectors to the point p in the space of probability distributions. Then the **Fisher information metric** is given as follows:

$$\langle v, w \rangle = \sum_i \frac{1}{p_i} v_i w_i$$

Using this we can calculate the speed at which $p(t)$ moves when it obeys the replicator equation. Actually the square of the speed is simpler:

$$\begin{aligned}
 \left\| \frac{dp}{dt} \right\|^2 &= \sum_i \frac{1}{p_i} \left(\frac{dp_i}{dt} \right)^2 \\
 &= \sum_i \frac{1}{p_i} ((f_i - \langle f \rangle) p_i)^2 \\
 &= \sum_i (f_i - \langle f \rangle)^2 p_i
 \end{aligned}$$

The answer has a nice meaning, too! It's just the [variance](#) of the fitness: that is, the square of its [standard deviation](#).

So, if you're willing to buy my claim that the speed $\|dp/dt\|$ is the rate at which our population learns new information, then we've seen that *the square of the rate at which a population learns information is the variance of its fitness!*

Fisher's fundamental theorem

Now, how is this related to Fisher's fundamental theorem of natural selection? First of all, what *is* Fisher's fundamental theorem? Here's what [Wikipedia says](#) about it:

It uses some mathematical notation but is not a theorem in the mathematical sense. It states:

"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time."

Or in more modern terminology:

"The rate of increase in the mean fitness of any organism at any time ascribable to natural selection acting through changes in gene frequencies is exactly equal to its genetic variance in fitness at that time".

Largely as a result of Fisher's feud with the American geneticist Sewall Wright about adaptive landscapes, the theorem was widely misunderstood to mean that the average fitness of a population would always increase, even though models showed this not to be the case. In 1972, George R. Price showed that Fisher's theorem was indeed correct (and that Fisher's proof was also correct, given a typo or two), but did not find it to be of great significance. The sophistication that Price pointed out, and that had made understanding difficult, is that the theorem gives a formula for part of the change in gene frequency, and not for all of it. This is a part that can be said to be due to natural selection

Price's paper is here:

- George R. Price, [Fisher's 'fundamental theorem' made clear](#), *Annals of Human Genetics* **36** (1972), 129.140.

I don't find it very clear, perhaps because I didn't spend enough time on it. But I think I get the idea.

My result *is* a theorem in the mathematical sense, though quite an easy one. I assume a population distribution evolves according to the replicator equation and derive an equation whose right-hand side matches that of Fisher's original equation: the variance of the fitness.

But my left-hand side is different: it's the square of the speed of the corresponding probability distribution, where speed is measured using the 'Fisher information metric'. This metric was discovered by the same guy, Ronald

Fisher, but I don't think he used it in *his* work on the fundamental theorem!

Something a bit similar to my statement appears as Theorem 2 of this paper:

- Marc Harper, [Information geometry and evolutionary game theory](#).

and for that theorem he cites:

- Josef Hofbauer and Karl Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, 1998.

However, his Theorem 2 really concerns the rate of increase of fitness, like Fisher's fundamental theorem. Moreover, he assumes that the probability distribution $p(t)$ flows along the gradient of a function, and I'm not assuming that. Indeed, my version applies to situations where the probability distribution moves round and round in periodic orbits!

Relative information and the Fisher information metric

The key to generalizing Fisher's fundamental theorem is thus to focus on the speed at which $p(t)$ moves, rather than the increase in fitness. Why do I call this speed the 'rate at which the population learns information'? It's because we're measuring this speed using the Fisher information metric, which is closely connected to [relative information](#), also known as relative entropy or the Kullback–Leibler divergence.

I explained this back in [Part 7](#), but that explanation seems hopelessly technical to me now, so here's a faster one, which I created while preparing my talk.

The information of a probability distribution q **relative to** a probability distribution p is

$$I(q, p) = \sum_{i=1}^n q_i \ln \left(\frac{q_i}{p_i} \right)$$

It says how much information you learn if you start with a hypothesis p saying that the probability of the i th situation was p_i , and then update this to a new hypothesis q .

Now suppose you have a hypothesis that's changing with time in a smooth way, given by a time-dependent probability $p(t)$. Then a calculation shows that

$$\frac{d}{dt} I(p(t), p(t_0)) \Big|_{t=t_0} = 0$$

for all times t_0 . This seems paradoxical at first. I like to jokingly put it this way:

To first order, you're never learning anything.

However, as long as the velocity $\frac{dp}{dt}(t_0)$ is nonzero, we have

$$\frac{d^2}{dt^2} I(p(t), p(t_0)) \Big|_{t=t_0} > 0$$

so we can say

To second order, you're always learning something... unless your opinions are fixed.

This lets us define a 'rate of learning'---that is, a 'speed' at which the probability distribution $p(t)$ moves. *And this is precisely the speed given by the Fisher information metric!*

In other words:

$$\left\| \frac{dp}{dt}(t_0) \right\|^2 = \left. \frac{d^2}{dt^2} I(p(t), p(t_0)) \right|_{t=t_0}$$

where the length is given by Fisher information metric. Indeed, this formula can be used to *define* the Fisher information metric. From this definition we can easily work out the concrete formula I gave earlier.

In summary: as a probability distribution moves around, the relative information between the new probability distribution and the original one grows approximately as the *square* of time, not linearly. So, to talk about a 'rate at which information is learned', we need to use the above formula, involving a second time derivative. This rate is just the speed at which the probability distribution moves, measured using the Fisher information metric. And when we have a probability distribution describing how many replicators are of different species, and it's evolving according to the replicator equation, this speed is also just the variance of the fitness!

You can read a discussion of this article [on Azimuth](#), and make your own comments or ask questions there!



© 2017 John Baez
baez@math.removethis.ucr.andthis.edu
[home](#)