The Internal Model Principle

John Baez

12 months ago

"Every good key must be a model of the lock it opens."

That sentence states an obvious fact, but perhaps also a profound insight if we interpret it generally enough.

That sentence is also the title of a paper:

• Daniel L. Scholten, Every good key must be a model of the lock it opens (the Conant & Ashby Theorem revisited), 2010.

Scholten gives a lot of examples, including these:

• A key is a model of a lock's keyhole.

• A city street map is a model of the actual city streets

• A restaurant menu is a model of the food the restaurant prepares and sells.

• Honey bees use a kind of dance to model the location of a source of nectar.

• An understanding of some phenomenon (for example a physicist's understanding of lightning) is a mental model of the actual phenomenon.

This line of thought has an interesting application to control theory. It suggests that *to do the best job of regulating some system, a control apparatus should include a model of that system.*

Indeed, much earlier, Conant and Ashby tried to turn this idea into a theorem, the 'good regulator theorem':

• Roger C. Conant and W. Ross Ashby, Every good regulator of a system must be a model of that system), *International Journal of Systems Science* **1** (1970), 89–97.

Scholten's paper is heavily based on this earlier paper. He summarizes it as follows:

> What all of this means, more or less, is that the pursuit of a goal by some dynamic agent (Regulator) in the face of a source of obstacles (System) places at least one particular and unavoidable demand on that agent, which is that the agent's behaviors must be executed in such a reliable and predictable way that they can serve as a representation (Model) of that source of obstacles.

It's not clear that this is true, but it's an appealing thought.

A particularly self-referential example arises when the regulator is some organism and the System is the world it lives in, *including itself*. In this case, it seems the regulator should include a model of *itself!* This would lead, ultimately, to self-awareness.

It all sounds great. But Scholten raises an obvious question: if Conant and Ashby's theorem is so great, why isn't more well-known? Scholten puts it quite vividly:

> Given the preponderance of control-models that are used by humans (the evidence for this preponderance will be surveyed in the latter part of the paper), and especially given the obvious need to regulate that system, one might guess that the C&A theorem would be at least as famous as, say, the Pythagorean Theorem (
>
>), the Einstein mass-energy equivalence (
>
> which can be seen on T-shirts and bumper stickers), or the DNA double helix (which actually shows up in TV crime dramas and movies about super heroes). And yet, it would appear that relatively few lay-persons have ever even heard of C&A's important prerequisite to successful regulation.

There could be various explanations. But here's mine: when I tried to read Conant and Ashby's paper, I got stuck. They use some very basic mathematical notation in nonstandard ways, and they don't clearly state the hypotheses and conclusion of their theorem.

Luckily, the paper is short, and the argument, while mysterious, seems simple. So, I immediately felt I should be able to *dream up* the hypotheses, conclusion, and

argument based on the hints given.

Scholten's paper didn't help much, since he says:

> Throughout the following discussion I will assume that the reader has studied Conant & Ashby's original paper, possesses the level of technical competence required to understand their proof, and is familiar with the components of the basic model that they used to prove their theorem [....]

However, I have a guess about the essential core of Conant and Ashby's theorem. So, I'll state that, and then say more about their setup.

Needless to say, I looked around to see if someone else had already done the work of figuring out what Conant and Ashby were saying. The best thing I found was this:

• B. A. Francis and W. M. Wonham, The internal model principle of control theory, *Automatica* **12** (1976) 457–465.

This paper works in a more specialized context: linear control theory. They've got a linear system or 'plant' responding to some input, a regulator or 'compensator' that is trying to make the plant behave in a desired way, and a 'disturbance' that affects the plant in some unwanted way. They prove that to perfectly correct for the disturbance, the compensator must contain an 'internal model' of the disturbance.

I'm probably stating this a bit incorrectly. This paper is much more technical, but it seems to be more careful in stating assumptions and conclusions. In particular, they seem to give a precise definition of an 'internal model'. And I read elsewhere that the 'internal model principle' proved here has become a classic result in control theory!

This paper says that Conant and Ashby's paper provided "plausibility arguments in favor of the internal model idea". So, perhaps Conant and Ashby inspired Francis and Wonham, and were then largely forgotten.

My guess

My guess is that Conant and Ashby's theorem boils down to this:

Theorem. Let

and

be finite sets, and fix a probability distribution

on

. Suppose

is any probability distribution on

such that

Let

be the Shannon entropy of

and let

be the Shannon entropy of

Then

and equality is achieved if there is a function

such that

∎

Note that this is not an 'if and only if'.

The proof of this is pretty easy to anyone who knows a bit about probability theory and entropy. I can restate it using a bit of standard jargon, which may make it more obvious to experts. We've got an

-valued random variable, say

We want to extend it to an

-valued random variable

whose entropy is small as possible. Then we can achieve this by choosing a function

and letting

Here's the point: if we make

be a function of

we aren't adding any extra randomness, so the entropy doesn't go up.

What in the world does this have to do with a good regulator containing a model of the system it's regulating?

Well, I can't explain that as well as I'd like—sorry. But the rough idea seems to be this. Suppose that

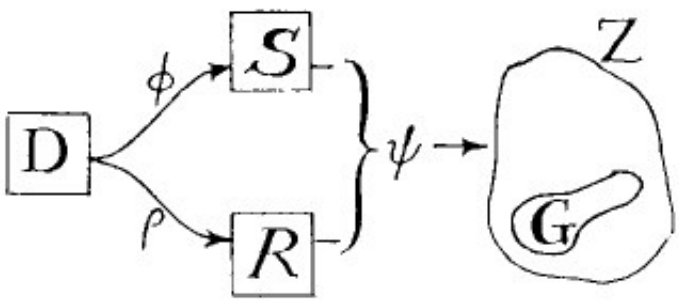
is a **system** with a given random behavior, and

is another system, the **regulator**. If we want the combination of the system and regulator to behave as 'nonrandomly' as possible, we can let the state of the regulator be a function of the state of the system.

This theorem is actually a 'lemma' in Conant and Ashby's paper. Let's look at their setup, and the 'good regulator theorem' as they actually state it.

Their setup

Conant and Ashby consider five sets and three functions. In a picture:



The sets are these:

• A set

$$Z$$

of possible **outcomes**.

• A **goal**: some subset

$$G \subseteq Z$$

of **good** outcomes

• A set

$$D$$

of **disturbances**, which I might prefer to call 'inputs'.

• A set

$$S$$

of states of some **system** that is affected by the disturbances.

• A set

$$R$$

of states of some **regulator** that is also affected by the disturbances.

The functions are these:

• A function

$$\phi : D \to S$$

saying how a disturbance determines a state of the system.

• A function

$$\rho : D \to R$$

saying how a disturbance determines a state of the regulator.

• A function

$$\psi : S \times R \to Z$$

saying how a state of the system and a state of the regulator determines an outcome.

Of course we want some conditions on these maps. What we want, *I guess*, is for the outcome to be good regardless of the disturbance. I might say that as follows: for every

$$d \in D$$

we have

$$\psi(\phi(d), \rho(d)) \in G$$

Unfortunately Conant and Ashby say they want this:

$$\rho \subset [\psi^{-1}(G)]\phi$$

I can't parse this: they're using math notation in ways I don't recognize. Can you figure out what they mean, and whether it matches my guess above?

Then, after a lot of examples and stuff, they state their theorem:

Theorem. The simplest optimal regulator R of a reguland S produces events R which are related to events S by a mapping $h : S \to R$.

Clearly I've skipped over too much! This barely makes any sense at all.

Unfortunately, looking at the text before the theorem, I don't see these terms being explained. Furthermore, their 'proof' introduces extra assumptions that were not mentioned in the statement of the theorem. It begins:

> The sets $R, S,$ and Z and the mapping $\psi : R \times S \to Z$ are presumed given. We will assume that over the set S there exists a probability distribution $p(S)$ which gives the relative frequencies of the events in $S.$ We will further assume that the behaviour of any particular regulator R is specified by a conditional distribution $p(R|S)$ giving, for each event in $S,$ a distribution on the regulatory events in $R.$

Get it? Now they're saying the state of the regulator R depends on the state of the system

S

via a conditional probability distribution

$$p(r|s)$$

where

$$r \in R$$

and

$$s \in S.$$

It's odd that they didn't mention this earlier! Their picture made it look like the state of the regulator is determined by the 'disturbance' via the function

$$\rho : D \to R.$$

But okay.

They're also assuming there's a probability distribution on

$$S.$$

They use this and the above conditional probability distribution to get a probability distribution on

$$R.$$

In fact, the set

$$D$$

and the functions out of this set seem to play no role in their proof!

It's unclear to me exactly what we're given, what we get to choose, and what we're trying to optimize. They do try to explain this. Here's what they say:

> Now
>
> $$p(S)$$
>
> and
>
> $$p(R|S)$$
>
> jointly determine
>
> $$p(R,S)$$
>
> and hence
>
> $$p(Z)$$
>
> and
>
> $$H(Z),$$
>
> the entropy in the set of outcomes:
>
> $$H(Z) = -\sum_{z \in Z} p(z) \log(p(z))$$
>
> With
>
> $$p(S)$$
>
> fixed, the class of optimal regulators therefore corresponds to the class of optimal distributions
>
> $$p(R|S)$$

> for which $H(Z)$ is minimal. We will call this class of optimal distributions π.

I could write a little essay on why this makes me unhappy, but never mind. I'm used to the habit of using the same letter p to stand for probability distributions on lots of different sets: folks let the argument of p say which set they have in mind at any moment. So, they're starting with a probability distribution on S and a conditional probability distribution on $r \in R$ given $s \in S.$ They're using these to determine probability distribution on $R \times S.$ Then, presumably using the map

$$\psi : S \times R \to Z,$$

they get a probability distribution on $Z.$ $H(Z)$ is the entropy of the probability distribution on $Z,$ and for some reason they are trying to minimize this.

(Where did the subset $G \subseteq Z$ of 'good' outcomes go? Shouldn't that play a role? Oh well.)

I believe the claim is that when this entropy is minimized, there's a function $h : S \to R$ such that

$$p(r|s) = 1 \text{ if } r = h(s)$$

This says that the state of the regulator should be completely determined by the the state of the system. And this, I believe, is what they mean by

> Every good regulator of a system must be a model of that system.

I hope you understand: I'm not worrying about whether the setup is a good one, e.g. sufficiently general for real-world applications. I'm just trying to figure out what the setup actually *is*, what Conant and Ashby's theorem actually *says*, and whether it's *true*.

I think I've just made a lot of progress. Surely this was no fun to read. But it I found it useful to write it.

Categories: mathematics, networks

Leave a Comment