"Inductive Bias"

21 Eliezer_Yudkowsky 08 April 2007 07:52PM

(Part two in a series on "statistical bias", "inductive bias", and "cognitive bias".)

Suppose that you see a swan for the first time, and it is white. It does not follow *logically* that the next swan you see must be white, but white seems like a better guess than any other color. A machine learning algorithm of the more rigid sort, if it sees a single white swan, may thereafter predict that any swan seen will be white. But this, of course, does not follow logically - though AIs of this sort are often misnamed "logical". For a purely logical reasoner to label the next swan white as a *deductive* conclusion, it would need an additional assumption: "All swans are the same color." This is a wonderful assumption to make if all swans are, *in reality,* the same color; otherwise, not so good. Tom Mitchell's *Machine Learning* defines the inductive bias of a machine learning algorithm as the assumptions that must be added to the observed data to transform the algorithm's outputs into logical deductions.

A more general view of inductive bias would identify it with a Bayesian's prior over sequences of observations...

Consider the case of an urn filled with red and white balls, from which we are to sample without replacement. I might have prior information that the urn contains 5 red balls and 5 white balls. Or, I might have prior information that a random number was selected from a uniform distribution between 0 and 1, and this number was then used as a fixed probability to independently generate a series of 10 balls. In either case, I will estimate a 50% probability that the first ball is red, a 50% probability that the second ball is red, etc., which you might foolishly think indicated the same prior belief. But, while the marginal probabilities on each round are equivalent, the probabilities over sequences are different. In the first case, if I see 3 red balls initially, I will estimate a probability of 2/7 that the next ball will be red. In the second case, if I see 3 red balls initially, I will estimate a 4/5 chance that the next ball will be red (by Laplace's Law of Succession, thus named because it was proved by Thomas Bayes). In both cases we refine our future guesses based on past data, but in opposite directions, which demonstrates the importance of prior information.

Suppose that your prior information about the urn is that a monkey tosses balls into the urn, selecting red balls with 1/4 probability and white balls with 3/4 probability, each ball selected independently. The urn contains 10 balls, and we sample without replacement. (E. T. Jaynes called this the "binomial monkey prior".) Now suppose that on the first three rounds, you see three red balls. What is the probability of seeing a red ball on the fourth round?

First, we calculate the prior probability that the monkey tossed 0 red balls and 10 white balls into the urn; then the prior probability that the monkey tossed 1 red ball and 9 white balls into the urn; and so on. Then we take our evidence (three red balls, sampled without replacement) and calculate the likelihood of seeing that evidence, conditioned on each of the possible urn contents. Then we update and normalize the posterior probability of the possible remaining urn contents. Then we average over the probability of drawing a red ball from each possible urn, weighted by that urn's posterior probability. And the answer is... *(scribbles frantically for quite some time)*... 1/4!

Of course it's 1/4. We specified that each ball was independently tossed into the urn, with a known 1/4 probability of being red. Imagine that the monkey is tossing the balls to you, one by one; if it tosses you a red ball on one round, that doesn't change the probability that it tosses you a red ball on the next round. When we withdraw one ball from the urn, it doesn't tell us anything about the other balls in the urn.

If you start out with a maximum-entropy prior, then you never learn anything, ever, no matter how much evidence you observe. You do not even learn anything wrong - you always remain as ignorant as you began.

The more inductive bias you have, the faster you learn to predict the future, but only if your inductive bias does in fact concentrate more probability into sequences of observations that actually occur. If your inductive bias concentrates probability into sequences that don't occur, this diverts probability mass from sequences that *do* occur, and you will learn more slowly, or not learn at all, or even - if you are unlucky enough - learn in the wrong direction.

Inductive biases can be probabilistically correct or probabilistically incorrect, and if they are correct, it is good to have as much of them as possible, and if they are incorrect, you are left worse off than if you had no inductive bias at all. Which is to say that inductive biases are like any other kind of belief; the true ones are good for you, the bad ones are worse than nothing. In contrast, statistical bias is always bad, period - you can trade it off against other ills, but it's never a good thing for itself. Statistical bias is a systematic direction in *errors*; inductive bias is a systematic direction in *belief revisions.*

As the example of maximum entropy demonstrates, without a direction to your belief revisions, you end up not revising your beliefs at all. No future prediction based on past experience follows as a matter of strict logical deduction. Which is to say: *All learning is induction, and all induction takes place through inductive bias.*

Why is inductive bias called "bias"? Because it has systematic qualities, like a statistical bias? Because it is a form of pre-evidential judgment, which resembles the word "prejudice", which resembles the political concept of bias? Damned

if I know, really - I'm not the one who decided to call it that. Words are only words; that's why humanity invented mathematics.

Article Navigation

Comments (24) **Tags:** bayesian statistics

Comments (24)

Sort By: **Old**

Barkley_Rosser *08 April 2007 08:37:49PM* 0 points [-]

Well, it is not every day that I can cite something that occurred at a conference that both Robin Hanson and I attended. But, we were at a conference honoring the work of David Grether, a giant of the field of Bayesian decision theory and econometrics, which was held on the George Mason campus on Friday, 4/6.

Anyway, a theme of several papers was that people are slow to update their priors in reality in many situations, although details are important. It is not clear what the source of this "inertia" is.

Eliezer_Yudkowsky *08 April 2007 09:21:30PM* 1 point [-]

Priors don't update. That's why they're called "priors".

Marginal posterior probabilities update; this is learning. Inductive priors over sequences don't update; they are what does the updating, they define your capability to learn. Even if you are a self-modifying AI and can rewrite your own source code, from a Bayesian perspective this is simply folded into an inductive prior over sequences of observations. I previously tried to write a post on this topic, but it got way too long and is now in my backlog of essays to finish someday.

This is exactly what I was trying to get at by distinguishing between the statement, "The marginal probability of drawing a red ball on the third round is 50%", which is true in all three scenarios above; versus the prior distributions over sequences of observations, which are different.

The inductive prior defines your responses to sequences of observations. This does not change over time; it is outside time. Learning how to learn is simply folded into the joint probability distribution.

xamdam *29 June 2010 05:23:42PM* 3 points [-]

> Priors don't update. That's why they're called "priors".

- John shows up on time for meetings 30%.
- John has been reprimanded.
- I think there is 95% chance he will be on time for meetings from now on.

You could just say that 95% is my prior for P(OnTime|Reprimanded), but I am not sure people think this way; "prior has been updated" seems more appropriate (when the condition is history).

DSimon *29 March 2011 03:02:41PM* 0 points [-]

Just call it your "current belief".

Kaj_Sotala *09 April 2007 12:05:00AM* 2 points [-]

(Apologies in advance to the sort-of-off-topic nature of this comment. As you'll see shortly, I had little choice.)

I was wondering, is there an avenue for us non-contributor readers to raise questions we think would be interesting to discuss? As far as I know, there are no public overcoming bias forums or mailing lists where everybody can post. One could ask questions in the comment sections in this blog, but that would be hijacking the commentaries to subjects other than what was actually said in the post - and I believe I've already seen at least one admonishment for a commenter to stick to the topic. Is it best to just post a question in the comments anyway, and trust for one of the regular contributors to make a real post about it if it's deemed interesting enough?

(As for the specific question I had in mind - I was wondering how careful one should be to avoid generalization from fictional evidence [described as a fallacy here, but I'd interprete it as a bias as well - which raises another potentially interesting question, how much overlap is there between fallacies and bias?]. When writing about artificial intelligence, for instance, would it be acceptable to mention Metamorphosis of Prime Intellect as a fictional example of an AI whose "morality programming" breaks down when conditions shift to ones its designer had not thought about? Or would it be better to avoid fictional examples entirely and stick purely to the facts?)

Eliezer_Yudkowsky *09 April 2007 12:17:49AM* 1 point [-]

Excellent suggestion, Kaj. I'm checking with Robin and Nick about putting up a post whose comments could be used for topic suggestions. (No further discussion in this thread though, please.)

On the importance of Less Wrong, or another single conversational locus
by AnnaSalamon | 82v (357c)

MIRI's 2016 Fundraiser
by So8res | 20v (13c)

Attention! Financial scam targeting Less Wrong users
by Viliam_Bur | 24v (93c)

LATEST RATIONALITY QUOTE

> Disasters and miracles follow
by **27chaos** on Rationality Quotes January - March 2017 | 0 points

RECENT WIKI EDITS

RECENT ON RATIONALITY BLOGS

TOP CONTRIBUTORS, 30 DAYS

Lumifer (**165**)
Viliam (**120**)
Zack_M_Davis (**111**)
Elo (**91**)
gjm (**90**)
Alicorn (**83**)
username2 (**59**)
lifelonglearner (**52**)
MrMind (**48**)
freyley (**47**)
Dagon (**43**)
Bound_up (**38**)
gwern (**37**)
ChristianKl (**37**)
Benquo (**33**)

RECENT KARMA AWARDS

HalFinney *09 April 2007 12:30:28AM* 2 points [-]

In practice you don't usually know exactly how the balls got into the urn. In that case you have a set of models for what might have happened, with a prior probability distribution over them. As you observe the sequences, you update the probabilities for these models. How does that fit into this inductive bias framework?

James_Annan *09 April 2007 06:33:29AM* 1 point [-]

> If you start out with a maximum-entropy prior, then you never learn anything, ever, no matter how much evidence you observe. You do not even learn anything wrong - you always remain as ignorant as you began.

Can you clarify what you mean here? Are you referring specifically to the monkey example or making a more general point?

Eliezer_Yudkowsky *09 April 2007 06:49:28AM* 1 point [-]

Finney, if you consider probability distributions over *sequences*, then - for example - a mixture of 33% first distribution, 33% second distribution, and 33% third distribution, produces a new and coherent probability distribution over sequences. This would create an inductive prior that could learn any of the three sequences, given only slightly more evidence to determine which one was most likely.

Annan, I'm making a more general point. (Obviously not so general as to encompass 'maximum-entropy methods' of machine learning, which find the distribution that maximizes entropy subject to constraints; they are not literally *maximum* entropy.) Think of physical matter in a state of very high thermodynamic entropy, such as a black hole or radiation bath. A heat bath doesn't learn from observation, right? There's not enough order present to carry out operations of observing, or learning. Only highly ordered matter, like brains, can extract information from the environment. A probability distribution in a state of maximum entropy likewise lacks structure and does not update in any systematic direction. The marginal posteriors will resemble the marginal priors. It can't learn from experience; it doesn't do induction.

Barkley__Rosser *09 April 2007 08:36:31PM* 0 points [-]

Eliezer,

Yes, thank you for correcting my sloppy wording.

So, it is the marginal posterior probabilities that exhibit inertia, or slow updating through learning, not the eternally unvarying "priors."

simon2 *10 April 2007 04:04:06AM* 0 points [-]

Why do you refer to the difference between a prior and the uniform prior as a bias, rather than the difference from the optimal prior? This doesn't agree with how you previously defined a bias.

Eliezer_Yudkowsky *10 April 2007 04:11:36AM* 0 points [-]

Simon, I don't understand your question. The optimal prior is the one that assigns probability 1 to the exact sequence that will be observed. Also, cognitive biases are not like inductive biases, despite the names, that's kinda the point.

simon2 *10 April 2007 05:02:15AM* 0 points [-]

Well then, what's the point of discussing it on the blog, if the similarity is only due to the names?

As for the optimal prior, if the universe is non-deterministic, or if there are "many worlds", or multiple universes in general, or other ways in which a given observer can have multiple different futures, then the optimal prior is a distribution over all those futures.

simon2 *10 April 2007 05:10:48AM* 0 points [-]

I shouldn't have included non-deterministic, since that only leads to one actual outcome.

Eliezer_Yudkowsky *10 April 2007 05:15:30AM* 0 points [-]

Simon, the point of discussing it on the blog is to help people who were confused by the similarity of names (not a hypothetical scenario, it did happen). And yes, if you are in a many-worlds situation of any type then the optimal prior is a distribution, albeit one that you will never realistically be able to compute.

simon2 *10 April 2007 05:35:54AM* 0 points [-]

OK, that clears it up then.

The point about the optimal prior was that, to the extent that a prior can be considered biased (in the sense I understood the word "bias", not inductive bias), the optimal prior is the unbiased prior it should be compared to. I didn't mean to imply that finding the optimal prior is realistic.

Barkley_Rosser *10 April 2007 06:34:00AM* 0 points [-]

Eliezer,

So, an "optimal prior" is either a subjectively guessed probability or, more optimally, probability distribution that coincides with an objective probability or probability distribution. That is it would equal the posterior distribution one would arrive at after the asymptotic working out of Bayes' Theorem, assuming the conditions for Bayes' Theorem hold.

But, what if those conditions do not hold? Will the "optimal prior" be equal to the "objective truth" or to the distribution that one arrives at after the infinite working out of the posterior adjustment learning process, even assuming that we do not have the sort of inertial slow learning that seems to exist in much of reality?

To give an example of such a non-convergence, consider the sort of example posed by Diaconis and Freeman, with an infinite dimensional space and a disconnected basis, one can end up in a cycle rather than on the mean.

Eliezer_Yudkowsky *10 April 2007 07:00:23AM* 1 point [-]

Barkley, priors aren't meant to be detailed objective models of the world - that's why they're called "priors". :)

A good prior learns from evidence, and the more probability mass it concentrates into sequences of the sort that are actually likely to occur, the faster it will learn. In a certain sense, the "optimal prior" is the one that learns so fast that it doesn't need any evidence at all - but that's not really what a "prior" is *for*. Even with an excellent prior, nearly all of the information will come from the environment.

Sense data is light, the prior is a camera. Most of the information is in the light, but you need a camera to develop it; a rock won't do. A good camera needs less light to develop an accurate picture, but the detailed picture is still carried by the light's message, not factory-preprinted inside the camera.

As for the Diaconis and Freedman paper, I haven't read it, but kindly remember that I am an infinite set atheist. In any case it is easy for poor priors to not learn, or anti-learn. Every prior that assigns more mass than maxent to "plausible" sequences, does so by draining mass from "implausible" sequences. If reality falls into one of the "implausible" sequences, we will do worse than maximum entropy, anti-learn from experience, and not pass on our genes to a whole lot of offspring.

Barkley__Rosser *10 April 2007 06:14:47PM* 0 points [-]

Eliezer,

Ah, so you are a constructivist, perhaps even an intuitionist? Even so, the point of such theorems is that they can happen in a long transient within finite constraints, with the biggie here being the non-connectedness of the support. One can get stuck in a cycle going nowhere for a long time, just as in such phenomena as transient chaos. With a suitably large, but finite, dimensionality and a disconnected support, one can wander in a wilderness with not much serious convergence for a very long time.

I find the idea of a "prior learning" to be a bit weird. It is an agent who learns, although the prior the agent walks in with will certainly play a role in the ability of the agent to learn. But the problem of inertia that I raised has more to do with the nature of agents than with their priors.

Getting to the raison d'etre of this blog, the question here is does bias arise from the nature of the prior an agent brings to a decision or analytical process, or is it something about the open-mindedness or willing to adjust posteriors in the face of evidence that is more important? Presumably both are playing at least some role.

Joe3 *19 April 2007 10:39:58PM* 0 points [-]

Why would anyone use a prior so strong that when presented with data, they would be unable to learn from it. In that case, if your prior is that strong, did you really have any intention of attempting to learn from new data?

Barkley,

I think that the concept of a prior deserves more attention as the strength of your current beliefs in the face of new evidence.

Presumably, if you have a subjective prior, you brought some "prior" experience or knowledge to the problem.... so philosphically, where does the original prior come from, and if it comes from your experience, is it really a prior, or have you actually reasoned your way to a posterior without even realizing it? Perhaps more time should actually be spent justifying your prior if you are going to bring a subjective prior to the problem. If you have good reasons and a lot of quality evidence, then the prior should receive a lot of weight.... deciding how much weight and how strongly you believe in your prior is a tough question.

I think that any time you create a prior without objective evidence able to support it, you have the potential to bias your results. But then again, if you truly believe in your subjective prior, do you really care about the potential to "bias" your results?

Linus_Quigley *21 May 2008 10:35:08AM* 0 points [-]

Thanks for this magnificent post. My only concern is that the point seems slightly overstated when you write: "All learning is induction, and all induction takes place through inductive bias." I wish this had been phrased slightly differently. The definition of learning seems a bit narrow. Is there no such thing as deductive learning? But even considering only the realm of inductive learning (based on observation), let's assume I see a swan for the first time, and the swan is white. Wouldn't it be correct to say that I've learned that at least one swan is white? (This may be slow learning, given the context, but wouldn't it still be learning?) And isn't the "inductive bias" in this case so minimal that it's not really properly called "bias" at all, since the assumption cannot be false?

Nick_Tarleton *12 November 2008 10:56:38PM* 2 points [-]

> Why is inductive bias called "bias"?

Because it represents a divergence from an imagined mind of pure emptiness that can learn equally well in any environment.

Mauve *13 November 2008 12:39:35AM* 1 point [-]

It is a *bias* because it is a prior *assumption* rather than something that is learned in the course of training. Mitchell's *Machine Learning* has a very clear explanation of inductive bias and why it is necessary for learning to occur. There are some examples of inductive bias at Wikipedia: http://en.wikipedia.org/wiki/Inductive_bias

Peacewise *13 November 2011 01:26:19AM* -1 points [-]

Seems to me that the educational psychology term "overextension" has some relevance to the white swan scenario mentioned above. "overextension - inappropriate use of a word for a class of things rather than for one particular thing." Definition provided by Krause, K., Bochner, S., Duchesne, S., & McMaugh, A. (2010). Educational psychology for learning & teaching (3rd ed.). South Melbourne: Cengage Learning Australia. Strictly going from seeing one white swan to labelling therefore, all swans are white is inappropriate, hence why I think overextension is relevant, it mainly occurs within very young children. I imagine that if AI are overextending then they may be displaying characteristics of 2/3 year old children, this may or may not be useful. Some parts of the below discussion mention prior's in the same way that a psychologist would use the term heuristic. "heuristic - a thinking strategy that enables quick, efficient judgements." Social Psychology 10th Edition by David Myers. It may well be useful to go from seeing one white swan to all swans are white, in that it may be a thinking strategy that enables quick efficient recognition of a swan. Perhaps this may be a first look scenario, a person (or ai) glimpses the whiteness and rough shape of a swan and provides a quick working label of "swan", then if necessary firms up that label with a refresh to gather more specific information, or simply holds the swan label if it's not necessarily needed.