Doing what has worked well in the past leads to evidential decision theory

Caspar Oesterheld, Foundational Research Institute

January 9, 2018

Because counterfactuals are untestable, decision theories may be viewed as untestable as well (Yudkowsky, 2010, ch. 13). However, that does not stop one from using a simple learning procedure, often called the law of effect (Sutton and Barto, 1998, sect. 1.6), for a series of Newcomb-like problems: when faced with a Newcomb-like problem, do what has worked well – i.e. what has been succeeded by high rewards or utilities – in past problems of a similar structure.

In this short note, we show that adopting such a learning procedure results in evidential decision theory (EDT) (Ahmed, 2014; Almond, 2010; Price, 1986; Horgan, 1981). While the result is trivial to prove, it fulfills two purposes in the context of Newcomb-like problems. First, the result contributes to understanding EDT – in my experience, people perceive the characterization of EDT as doing what has worked well in the past as surprising or counter-intuitive. Second, the result contributes to our understanding of how simple decision-making policies such as those used in artificial intelligence behave in Newcomb-like problems and how to implement specific decision theories within standard artificial intelligence frameworks.

A decision problem $DP = (\mathbb{A}, \Omega, U, P)$ consists of

- a finite set of possible actions $\mathbb{A}$,
- a finite set of possible outcomes Ω,
- a utility function $U : \Omega \to \mathbb{R}$, and
- a conditional probability distribution $P(\cdot \mid \cdot)$ mapping pairs of an outcome o and an action a onto the probability of outcome o occurring given that the agent takes action a. The probability distribution may or may not be known to the agent.

Of course, we can associate other information with a decision problem. For example, an agent may have some causal model of the decision problem. However, we will see that the probabilistic model P is sufficient for predicting the agent's behavior.

Note that the elements of $\mathbb{A}$ may be meta-actions such as decision theories or randomized strategies.

We now imagine some probabilistic learning policy $\pi : (A \times O)^* \rightsquigarrow A$. We will let that policy interact iteratively with independent instances of a decision problem DP.[1] This interaction yields histories H_n, which are random variables over $(A \times O)^n$.

[1] For example, the agent may face Newcomb's problem every day. Note, however, that the predictor has to make a new prediction every day so that the outcomes can be independent of each other given the actions chosen by the agent.

Smoking lesion-type problems can be serialized in a similar way. Again, it has to be ensured that the instantiations are independent. So, for example, for each day n there could exist a different disease that shows the harmless symptom in the form of an action in the morning of day n and a strong symptom in the form of an outcome on the evening of day n. That said, there is reason to be skeptical of the smoking lesion as a Newcomb-like problem in general, see, e.g., Ahmed (2014, ch. 4).

Note that this setup is formally equivalent to a multi-armed bandit problem (see, e.g., Sutton and Barto, 1998, ch. 2). The only difference is that we will interpret DP as potentially Newcomb-like.

For π to qualify as implementing what has worked well in the past, we require two properties. First, we will require that in the limit it employs all actions infinitely often, i.e. that

$$\lim_{n\to\infty} P(\mathit{count}(a, H_n) > K) = 1 \tag{1}$$

for arbitrarily large $K \in \mathbb{R}$ and all $a \in A$, where $\mathit{count}(a, H_n)$ denotes the number of times a has been taken in H_n. We will write this as

$$\mathit{count}(a, H_n) \to \infty \text{ for } n \to \infty. \tag{2}$$

Secondly, π should eventually converge on the empirically optimal solution, i.e.

$$\lim_{n\to\infty} P(\pi(H_n) \in \arg\max_{a\in A} \mathit{performance}(a, H_n)) = 1, \tag{3}$$

where $\mathit{performance}(a, H_n)$ is the average over the utilities of the outcomes succeeding a.

An example of a policy fulfilling both desiderata is one that, in time step n, picks a random action with probability $\frac{1}{n}$ and a random element from $\arg\max_{a\in\mathbb{A}} \mathit{performance}(a, H_n)$ with probability $\frac{n-1}{n}$.

Equation 2 means that as n approaches infinity, so do the sample sizes $\mathit{count}(a, H_n)$ for each action $a \in \mathbb{A}$. The law of large numbers implies that as the sample size approaches infinity, the mean of the sample utilities approaches the expected value, i.e. that

$$\mathit{performance}(a, H_n) \to \mathbb{E}\left[u(O) \mid a\right] \text{ for } n \to \infty \tag{4}$$

for all $a \in \mathbb{A}$. As a corollary with equation 3, π converges on actions that are optimal according to EDT.

Note again how trivial and unremarkable this result is in the context of multi-armed bandit problems. It only becomes interesting in the context of Newcomb-like problems.

Also note that the problem setup does not expose the full spectrum of Newcomb-like problems. For example, it excludes issues of "updatelessness" (Meacham, 2010; Soares and Fallenstein, 2014, sect. 3; Yudkowsky, 2010, sect. 2) and anthropics (Bostrom, 2010; Armstrong, 2011).

As for further work, I suspect that model-free learning algorithms (see Kaelbling, Littman, and Moore, 1996, sect. 4) more generally converge to EDT (or an updateless version of). Model-based algorithms may be harder to predict. Because classic algorithms for learning models presuppose Cartesianism, they cannot build correct models for a naturalized setting (Soares, 2015). Their decisions may thus depend heavily on the details of the learning algorithm.

Related Work

Gardner (1973, p. 108) remarks that in a series of Newcomb problems, "acting pragmatically, on the basis of past experience" results in one-boxing.

Acknowledgements

I'm indebted to Brian Tomasik and Max Daniel for comments. I also benefited from discussions at AISFP 2017.

References

Ahmed, Arif (2014). *Evidence, Decision and Causality*. Cambridge University Press.

Almond, Paul (2010). *On Causation and Correlation Part 1: Evidential decision theory is correct*. URL: https://casparoesterheld.files.wordpress.com/2016/12/almond_edt_1.pdf.

Armstrong, Stuart (2011). *Anthropic Decision Theory*. Future of Humanity Institute. URL: https://arxiv.org/abs/1110.6437.

Bostrom, Nick (2010). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Ed. by Robert Nozick. Studies in Philosophy. Routledge.

Gardner, Martin (1973). "Free will revisited, with a mind-bending prediction paradox by William Newcomb". In: *Scientific American* 229.1, pp. 104–109.

Horgan, Terence (1981). "Counterfactuals and Newcomb's Problem". In: *The Journal of Philosophy* 78.6, pp. 331–356.

Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. Moore (1996). "Reinforcement Learning: A Survey". In: *Journal of Artificial Intelligence Research* 4, pp. 237–285. URL: http://www.jair.org/media/301/live-301-1562-jair.pdf.

Meacham, Christopher J. G. (2010). "Binding and its consequences". In: *Philosophical Studies* 149.1, pp. 49–71. DOI: 10.1007/s11098-010-9539-7.

Price, Huw (1986). "Against Causal Decision Theory". In: *Synthese* 67, pp. 195–212.

Soares, Nate (2015). *Formalizing Two Problems of Realistic World-Models*. Tech. rep. 2015-3. Machine Intelligence Research Institute. URL: https://intelligence.org/files/RealisticWorldModels.pdf.

Soares, Nate and Benja Fallenstein (2014). *Toward Idealized Decision Theory*. Tech. rep. 2014-7. Machine Intelligence Research Institute. URL: https://arxiv.org/abs/1507.01986.

Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Yudkowsky, Eliezer (2010). *Timeless Decision Theory*. The Singularity Institute. URL: http://intelligence.org/files/TDT.pdf.