A Note Regarding "Active Entropy"

In section 3.2 of the paper, "The Search for a Search: Measuring the Information Cost of Higher Level Search", we read:

> Because the active entropy is strictly negative, any uninformed assisted search (φ) will on average perform worse than the baseline search.

But if active entropy is an indicator of relative search[1] performance, then the fact that active entropy is strictly negative implies that *every search is worse than every other search*, an obvious absurdity.

The paper continues:

> Moreover, the degree to which it performs worse will track the degree to which the assisted search singles out and confers disproportionately high probability on only a few targets in the partition. This suggests that success of an assisted search depends on its attending to a few select targets at the expense of neglecting most of the remaining targets.

Again, this cannot be correct. If performance is averaged over all possible targets[2], then it makes no difference how the search chooses its query – the average performance is the same in every case. This is a simple application of Wolpert and Macready's NFL principle.

These problems are rooted in the definition of active entropy. To put this definition in context, we'll look first at the definition of *active information* at the end of section 3.1:

$$I_+(\varphi|\psi) := \log_2 \frac{\varphi(T)}{\psi(T)} = \log_2 \varphi(T) - \log_2 \psi(T) \qquad (10)$$

> for φ and ψ arbitrary probability measures over the compact metric space Ω (with metric *D*), and *T* an arbitrary Borel set of Ω such that $\psi(T) > 0$.

Active entropy is then defined as follows, at the beginning of section 3.2:

> Let φ and ψ be arbitrary probability measures and let $\tilde{T} = \{T_i\}_{i=1}^N$ be an exhaustive partition of Ω all of whose partition elements have positive probability with respect to ψ . Define *active entropy* as the average active information that φ contributes to ψ with respect to the partition $\tilde{T}$ as

$$H_+^{\tilde{T}}(\varphi|\psi) := \sum_{i=1}^{N} \psi(T_i) I_+^{T_i}(\varphi|\psi) = \sum_{i=1}^{N} \psi(T_i) \log_2 \frac{\varphi(T_i)}{\psi(T_i)} \qquad (11)$$

And from this definition follows the Horizontal No Free Lunch Theorem and proof:

> Then $H_+^{\tilde{T}}(\varphi|\psi) \le 0$ with equality if and only if $\varphi(T_i) = \psi(T_i)$ for $1 \le i \le N$.
>
> *Proof:* The expression in Eq. (11) is immediately recognized as the negative of the *Kullback-Leibler distance*. Since the Kullback-Leibler distance is always nonnegative, the

1 Like the paper, this discussion is restricted to single-query searches. The ramifications of the single-query restriction have been discussed elsewhere by Dietmar Eben.

2 Or at least over a set of target functions that is closed under permutation. The set of all target functions $f : \tilde{T} \to \{\text{target, non-target}\}$ that contain exactly one target is closed under permutation, so the NFL conditions are met in the present case.

expression in Eq. (11) does not exceed zero. Zero is achieved if for each i, $\varphi(T_i)=\psi(T_i)$.

Note that these definitions and theorem make no mention of searches or targets. Stated thus, the HNFLT is unobjectionable, as it simply defines *active entropy* as negative relative entropy and concludes that active entropy is always non-positive. It's when we try to use active entropy as a measure of relative search performance that we run into trouble.

In the context of this and other papers by the same authors, it makes no sense to speak of active information $I_+^T(\varphi|\psi)$ where T is not a target. We therefore assume that T_i in (11) refers to a target.[3] Since not every T_i in $\tilde{T}$ can be a target simultaneously, the summation must be over various search problems, each with a different target.[4] We can therefore interpret active entropy as follows:

> Given a set $\tilde{T}$ of potential targets, and probability distributions φ and ψ over $\tilde{T}$, *active entropy* is the active information $I_+^{T_i}$ averaged over all allowed targets, weighted by $\psi(T_i)$.

The first problem is with the weight. $\psi(T_i)$ is the probability of the search query falling in T_i, but to give a meaningful average, the weight should be the probability of T_i being the target. The latter probability would have to be provided in the problem description in order to remedy this problem, or we can change (11) to an unweighted average as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\log_2\frac{\varphi(T_i)}{\psi(T_i)}$$

The result of this expression is not guaranteed to be negative, so we might think that the sign of the result will tell us whether φ or ψ is the better performer. But if “performance” refers to the probability of the search succeeding, the above expression still doesn't tell us which search is better. To find the probability of success, we need to average φ and ψ, not their respective logs, over the possible targets. So the correct expression is:

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(T_i) - \frac{1}{N}\sum_{i=1}^{N}\psi(T_i) = \frac{1}{N} - \frac{1}{N} = 0$$

This agrees with the NFLT – all searches perform the same when averaged over the possible targets.

It's not clear why the author thought that (11) would work as a relative performance measure, but he may have been attempting to apply a principle from coding theory to searches. In coding theory, incorrectly guessing a distribution results in an inefficient coding scheme. Relative entropy is a measure of this inefficiency, so it's tempting to think that relative entropy would likewise measure the relative inefficiency resulting from a poorly chosen query distribution. But searching is not encoding, and relative entropy does not apply to searches in the way that the paper claims, as shown in this note.

---

3 This interpretation, which I think reflects the intent of the paper, hinges on the assumption that T in (10) and T_i in (11) always refer to a target. If we dispense with this assumption, we can interpret active information such that it makes sense to speak of $I_+^{T_i}$ where T_i is not a target. To avoid confusion, we'll refer to this new concept as *non-target active information*. From this arises an alternate interpretation of (11), which we'll call *non-target active entropy*:

> Given a set of partition elements, one of them designated as a target, *non-target active entropy* is the non-target active information averaged over all partition elements in which the search's query may fall, weighted by $\psi(T_i)$.

In contrast to the previous interpretation, the weight in this summation is appropriate. But non-target active entropy is clearly not a performance measure, since all but one of the summands deal with partition elements that are not the target.

4 It's not clear why the set of allowed targets would be a partition instead of the full powerset of Ω.