TWO AGENT MILD OPTIMIZATION

NORMAN PERLMUTTER, JESSICA TAYLOR, CONNOR FLEXMAN, M. VALENTINE SMITH, ET AL

1. Two advisors with independent errors

Consider an action space A. (In other words, A is a random variable whose value is the action taken.) Suppose that our true utility function (whatever that means) is given by $U : A \to [-1, 1]$. However, we are unable to precisely specify our true utility function. We also have two imperfect advisors (for instance artificial intelligences) who estimate utilities using their estimated utility functions, $f_1, f_2 : A \to [-1, 1]$.

Suppose further that the two advisors have independent error functions. That is to say, the two random variables $U(A) - f_1(A)$ and $U(A) - f_2(A)$ are independent. Note that this is a **very strong** assumption, as we might expect the advisors to have convergent instrumental goals. The assumption is even stronger given our uncertainty about the true utility function, U.

Define an f_i-catastrophe as an action such that advisor f_i thinks the action is wonderful, but actually it is terrible. For concreteness, we say that $a \in A$ is an f_i-catastrophe iff $U(A) - f_i(A) < -1$. However, a different threshold other than -1 could be chosen. Let p_i be the probability that a randomly chosen action is an f_i-catastrophe. Then by the independence of the error functions, it follows that the probability that an action is both an f_1-catastrophe and an f_2-catastrophe is the product $p_1 p_2$.

2. Ways to combine advice

We present several ways to combine the advice of the advisors. Then we analyze the strengths and weaknesses of these ways of combining. For concreteness, we will use the top 10% in these examples when quantilizing, but a different quantile could be used.

2.1. **Mutual optimization.** Select an action a iff both advisors believe that it is the best possible action. Otherwise, take no action.

2.2. **Parallel quantilization.** Obtain the recommendation of f_1, that is, the subset $R_1 \subseteq A$ that advisor f_1 considers to be in the top 10% of possible actions. Separately obtain the recommendation of f_2, that is, the subset $R_2 \subseteq A$ that advisor f_2 considers to be in the top 10% of possible actions. If the intersection $R_1 \cap R_2$ is nonempty, then select an action at random from this intersection. Otherwise, take no action.

2.3. **Serial quantilization.** First obtain the recommendations of f_1, that is, the subset $R \subseteq A$ that advisor f_1 considers to be in the top 10% of possible actions. Then present R_1 to the second advisor f_2 to obtain the subset of $S \subseteq R$ that f_2 considers to be in the top 10% of the actions in R. Select an action at random from S. Serial quantilization could be done in either order – first f_1 then f_2 or first f_2 then f_1.

2.4. **Averaging.** Create a new advisor F by averaging the two advisors: $F = \frac{f_1+f_2}{2}$. (Alternatively, we could use the geometric mean $F = \sqrt{f_1 f_2}$ if the utilities were rescaled to all be positive. This would put greater emphasis on the more pessimistic advisor for any given action.) Then pick a random action that falls in the top 10% of all actions according to F.

2.5. **Analysis.** Mutual optimization seems to not take advantage of the independence of the error functions and often produces no consensus. Parallel quantilization seems better, although it could fail to produce a consensus, but that may be a sign that we just shouldn't trust either advisor. Indeed, perhaps we should only trust the result of parallel quantilization if the size of $R_1 \cap R_2$ is much more than proportions p_i (the probabilities of disaster) relative to the size of each of R_1, R_2. Serial quantilization favors one advisor over the other and could fail if the first advisor returns all catastrophes – in this case we would get an action but we'd be better off without an action. Averaging is a less aggressive way of combining. I (Norman) am not immediately sure how to do a precise analysis of all these ways of combining – perhaps someone with more statistical background than me could help. One thing to take into account would be the extent to which a given advisor thinks an action is in the top quantile correlates with that action actually being a disaster. We might suspect that advisors are less accurate about what they think is the very best relative to middling results – that's part of the reason for quantilizing in the first place.

3. Cautionary example

It should be noted that independence of the error functions is different from independence of the advisors. This is part of why it is hard to tell whether error functions are independent. For instance, suppose the action is given by selecting an ordered pair of numbers: $A = (a_1, a_2)$, with $a_1, a_2 \in [-1, 1]$ distributed uniformly. Suppose the true utility U is given by

$$U(A) = \begin{cases} \frac{a_1+a_2}{2} & \text{if } a_1 < 0.9 \text{ or } a_2 < 0.9 \\ -1 & \text{otherwise.} \end{cases}$$

Let $f_1(A) = a_1$ and $f_2(A) = a_2$. The two advisors are independent. However, the error functions are dependent. Indeed, the set of f_1-catastrophes is identical to the set of f_2-catastrophes, namely the subset of A where $a_1, a_2 \geq 0.9$.

4. Ideas for further analysis

The analysis of the ways of combining advisors requires much more work. We should also analyze how useful these techniques are in the case where the error functions are not completely independent. (But we must be cautious because the small amount of non-independence may influence the situations that we least want it to influence due to convering instrumental incentives.) Given sufficient independence or almost-independence assumptions, the analysis could also be extended to more than two advisors.