On the Possible Pitfalls in the Evaluation of Brain Computer Interface Mice

Riccardo Poli and Mathew Salvaris

Brain-Computer Interfaces Lab, School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK

E-mail: {rpoli,mssalv}@essex.ac.uk

Abstract. In a recent article by Bradberry, Gentili and Contreras-Vidal published in this journal, a interesting method for the control of a 2–D mouse cursor was proposed which apparently attained excellent control and good speed with relatively simple techniques. We believe some of the results in the paper have been misinterpreted due to a failure in appreciating the self-fulfilling nature of the success criteria adopted. In this short article we explain the nature of the problem and attempt to assess its influence on the results reported in the aforementioned article.

Submitted to: *J. Neural Eng.*

1. Introduction

In a recent article published in this journal [1], Bradberry, Gentili and Contreras-Vidal proposed a method for the control of a 2–D mouse cursor which adopted particularly simple techniques, where the control of the velocity of the mouse cursor was a simple linear combination of the temporal derivatives of the voltages recorded in 34 EEG channels. Voltages were low-pass filtered with a cut-off frequency of 1 Hz before differentiation.

The display used in that work is reported in Figure 1. Results with 5 subjects were very positive, indicating that they could hit a target within 15 seconds in 73% of trials, with a median hit time (for the successful trials) of 5.4 seconds. However, we believe that these encouraging results were partly to be attributed to the evaluation criteria adopted rather than the subjects attaining a reasonable level of control of the interface. As we will explain in the following section there are at least two reasons for this.

2. Issues When Evaluating BCI Mice

As shown in Figure 1 the targets used in [1] represent approximately 40% of the length of the edges of the display. In each trial only one of the targets appeared. The cursor started at the centre of the screen and the subject was tasked with directing it to hit the designated target. As soon as a target was hit the trial ended. The cursor trajectories obtained in successful trials were then rescaled so as to make them the same length and averaged to produce plots of mean cursor paths. The success criterion adopted and the averaging technique used led us to ask two questions. We will look at these questions in the following two subsections.

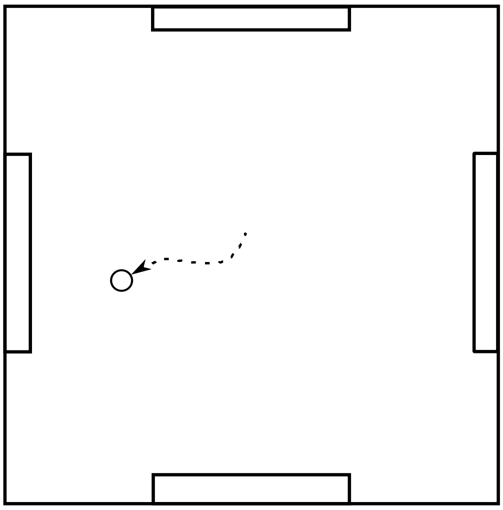

Figure 1: Display used in [1]. The rectangles at the margins of the display represent 4 potential targets. In each trial only one of the targets appeared. The cursor started at the centre of the screen and the subject was tasked to direct it to hit the designated target.

2.1. Is Subjects' Performance Above Chance?

As it is standard practice in psychology, when evaluating performance of subjects in a situation such as this where success could simply be obtained as a result of the cursor randomly drifting within the display, one needs to prove that subjects performance is above chance. This wasn't done in [1], probably because performance appeared sufficiently good to dispel any such concerns. However, we believe this was a potential mistake.

The question to be addressed is how likely is it that a cursor driven by some sort of random process would hit the target. The training procedure adopted in [1] was sufficiently complex that we cannot reasonably attempt to exactly replicate their experimentation. However, we can idealise the system and attempt to get ballpark figures for success rate and expected hit times. As we will see these are comparable to those reported in [1] suggesting that perhaps subjects' real performance was actually inferior to what was reported in the article.

Our idealisation of the system is as follows. We created a simulation where the display is identical to that in Figure 1 except that targets are simple line segments (with no thickness) occupying 40% of the edge of the display. Also, we idealised the cursor making it a point rather than the circle with a non-zero diameter used in [1]. To keep the simulation as similar as possible to that in [1], in each trial we gave a maximum of 1,500 time steps (15 seconds at 100 Hz) to the cursor to hit the target. Also, if the cursor hit a boundary of the display we zeroed the component of the velocity orthogonal to the boundary, thereby simulating an inelastic collision, as was done in [1].

The cursor movement in our simulation was controlled by acting on the two components of the cursor velocity. These were computed as follows:

$$v_x = v_x \times \alpha + U() \times (1-\alpha) \quad \text{and} \quad v_y = v_y \times \alpha + U() \times (1-\alpha)$$

where α is a constant and U is a random number generator which returns values uniformly distributed in the range $[-0.5, +0.5]$. The cursor's velocity was initialised to 0 at the beginning of each trial. We set the value of α to 0.99, so the sequence of random numbers was significantly smoothed, thereby producing relatively slow random drifts in the vicinity of the zero axis. This is compatible with the low pass filtering at 1 Hz of EEG signals adopted in [1]. As an illustration, we show the type of random drift produced by this process in Figure 2.

In the runs with our simulator we obtained a success rate of 71.81% with a median number of iterations for the successful trials of 551 samples, which correspond to 5.51 seconds at the sampling rate of 100 Hz as adopted in [1].‡ These results match worryingly closely the results reported in [1] and summarised in Section 1, suggesting

‡ We performed 10,000 runs to ensure these figures were reasonably unaffected by sampling errors.

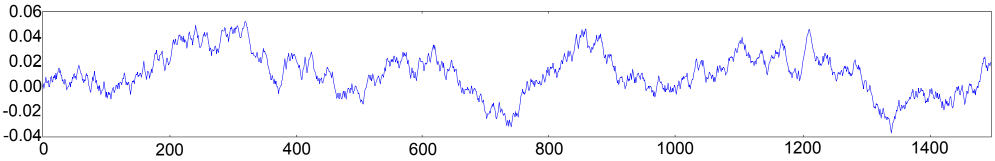

Figure 2: Sample random drift process used to drive our simulated mouse.

that, while the subjects in [1] may have had some degree of control, that control was likely overestimated.

2.2. Does Averaging of Trajectories Lead to Biased Averages?

The second issue we need to look at is the averaging of trajectories for the purpose visualisation. The procedure used in [1] involves taking the trajectories recorded in successful trials, normalising them so that they all are of the same length, averaging them and, finally, plotting the resulting averages. We believe that this procedure leads to biased results.

Firstly, even if all trajectories had the same length (thereby making it easier to average them), it is clear that by selecting only the successful trajectories one already biases the mean. The mean trajectory thus obtained wouldn't represent the real mean, but the *conditional mean* (i.e., the mean subject to the event that a trajectory was successful at hitting the target). Because of this, it is not surprising to see that in [1] all such trajectories hit the target and they do so by hitting it approximately in the middle. Indeed the mean trajectories obtained by averaging 30 trajectories obtained in our pseudo-random simulation show exactly the same behaviour, as illustrated in Figure 3 (bottom). It actually does not matter which process one uses to obtain the original trajectories: by definition they all end up on the target and so must their conditional means. Also, because single trajectories are initially uncorrelated but later become correlated by the fact that they all hit the target, averages appear initially convoluted but they become more directed and smooth as they approach the target, which is exactly what happened in the plots in [1, see their Figure 4].

Secondly, there is the question of whether length-normalising trajectories before averaging has an impact on the veracity of the results. We believe it has. The normalisation process leads to the longer trajectories, which are the more convoluted ones, to be sub-sampled more with respect to the shorter trajectories. This leads to them appearing smoother than they were originally. The shorter trajectories will be smoother because they did approach the target more directly. So, averaging length-normalised trajectories leads to the impression that such trajectories were on average much straighter that they actually were. This is clearly illustrated in Figure 3 (top) which reports the single successful trajectories which led to the means in Figure 3 (bottom). As one can see the mean trajectories are not really representative of the successful trials, which effectively occupy the whole display. It would be very hard for anyone to infer which was the target in these trials if we hadn't drawn it in the plots.

3. Conclusion

What have we demonstrated with the simulations reported in this paper? Certainly, we have not proven that the results reported in [1] are bogus, nor did we conduct these experiments with that purpose in mind. We acknowledge the fact that we *did* adjust the parameter α so as to achieve a close match between performance criteria and mean trajectories. However, the simulations shown in the paper and our discussion about the biases implicit in the success criteria and averaging methods adopted in [1] indicate that care is needed when evaluating BCI mice and that, perhaps, additional investigation is required to fully evaluate the impact of the results presented in [1].

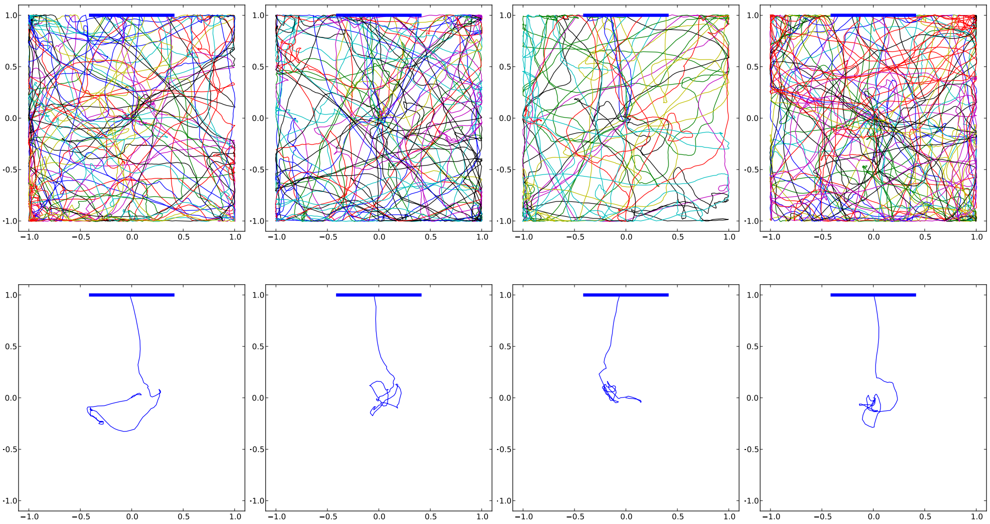

Figure 3: Successful trajectories (top) and corresponding length-normalised averages (bottom) obtained in 4 simulations of 30 trials each. All trajectories had the “up” target (the target is drawn with a thicker line for display clarity). Because of symmetries, similar results can be obtained for the other three target positions.

Acknowledgements

We would like to thank the UK Engineering and Physical Sciences Research Council (grant EP/F033818/1) for financial support. The authors would also like to thank Luca Citi for useful comments.

References

[1] Bradberry T J, Gentili R J and Contreras-Vidal J L 2011 Fast attainment of computer cursor control with noninvasively acquired brain signals *Journal of Neural Engineering* **8**