On the Impact of EEG Re-referencing on Classifier Performance

Joseph Rudoler

August 2018

1 Introduction

The overall goal of my summer project is to determine the optimal voltage reference scheme for electroencephalographic data. For the purposes of the project, the "optimal" scheme is the method of referencing voltage signals that maximizes classifier performance. Classifier performance is quantified by the AUC metric. In this report I will discuss only intracranial data, and include my analysis of four reference schemes: bipolar reference, average reference, region-of-interest reference, and weighted average reference. I will review the data processing steps that I took to assess these reference schemes and summarize and compare the classifier results for each scheme.

2 Methods

2.1 Subject Selection

My analysis dealt with the following 42 participants:

'R1060M', 'R1061T', 'R1065J', 'R1066P', 'R1067P', 'R1068J', 'R1077T', 'R1083J', 'R1094T', 'R1111M', 'R1112M', 'R1113T', 'R1121M', 'R1122E', 'R1123C', 'R1125T', 'R1134T', 'R1135E', 'R1137E', 'R1146E', 'R1147P', 'R1151E', 'R1153T', 'R1154D', 'R1156D', 'R1158T', 'R1161E', 'R1166D', 'R1168T', 'R1172E', 'R1189M', 'R1191J', 'R1193T', 'R1195E', 'R1200T', 'R1215M', 'R1217T', 'R1222M', 'R1223E', 'R1230J', 'R1236J', 'R1243T'

These participants are a subset of the intracranial patients included in the RAM project who participated in the FR1 experiment. In order to have the flexibility to re-reference voltage in multiple ways, it is essential to use only participants for whom monopolar EEG data can be obtained. Since this is

generally not the case for participants under the ENS recording system, I excluded all ENS participants (participants with a subject ID number greater than 275). I assessed classifier performance using a leave-one-session-out cross-validation approach; this made it necessary to restrict my analysis to participants who completed at least two sessions. 96 participants fit these criteria. I encountered a variety of technical issues with 16 of these participants, some related to cluster permission errors and others due to flawed, inconsistent, or nonexistent events or voltage data. The need for proper lobe-level brain region labeling caused me to drop another 9 subjects for whom this data was either non-existent or inconsistent. To reduce computation time I further restricted my analysis to subjects who had completed at least three sessions.

2.2 Measuring Voltage

Voltage is a difference in electric potential between two points: for any measurement of electric potential at some point a in space, it is necessary to subtract the potential at a "reference" point b from that value.

$$\Delta V = V_a - V_b$$

In physical terms, voltage represents the energy or work required to move a point charge between two points. Moving a point charge requires work because it is necessary to exert force in order to move through a gradient in the electric field. As such, the voltage between two points is a reflection of how much the electric field changes over the distance separating those points. This means that selecting a nearby reference will yield information about local changes in the electric field (signal components with a high spatial frequency), while selecting a distant reference will yield information about global changes in the electric field (signal components with a low spatial frequency). Re-referencing is thus a spatial filter which can optimize the quality and resolution of the EEG data for the purposes of classification.
In electrophysiological measurements, voltage is often initially recorded separately at each implanted electrode channel, and referenced to a *common reference*. The common reference might be one of the many electrode channels implanted in the brain or some point chosen on or outside of the scalp. This data is *monopolar* because it contains a single voltage reading for each individual electrode. Before detailing the rest of the data processing pipeline, I will explain how each scheme re-references monopolar data.

2.2.1 Average Referencing

The signals are re-referenced to the average of all electrodes channels. This is done by subtracting the mean voltage across channels from the individual voltage at each channel. This effectively cancels out the original common reference.

$$(\Delta V_{1..n})_{avg} = \frac{\sum_{k=1}^{n} V_k - V_{common}}{n} = V_{avg} - V_{common}$$

$$\Delta V'_k = (V_k - V_{common}) - (V_{avg} - V_{common})$$

$$\Delta V'_k = V_k - V_{avg}$$

2.2.2 Bipolar Referencing

The electrodes channels are grouped in pairs (thus the name *bipolar*) and referenced to one another. There are many possible combinations of channels, though in a bipolar scheme channels are paired with one of their closest neighbors. A single electrode channel can be a member of more than one bipolar pair. The distribution of electrode pairs is often called a *montage*. As was the case with average referencing, re-referencing the data removes the original common reference.

$$\Delta V_a = V_a - V_{common}$$

$$\Delta V_b = V_b - V_{common}$$

$$\Delta V_{ab} = \Delta V_a - \Delta V_b = V_a - V_b$$

2.2.3 Region-of-interest (ROI) Referencing

The electrodes are grouped according to the region of the brain in which they are implanted, and re-referenced to the average of all electrodes in that region. For this referencing scheme I grouped electrodes within the frontal lobe, parietal lobe, temporal lobe, limbic lobe, and occipital lobe in each hemisphere of the brain. Electrodes that did not fall strictly within the bounds of these regions were grouped with the closest neighboring electrode that had a clearly defined region.

If a region was populated by less than three electrodes channels, those channels were grouped along with the nearest well-populated region in order to ensure that their data was actually informative of the electrical activity in that region. Under this referencing scheme, a lone electrode channel is useless because the average of channels in its region is equal to the signal of

that single channel; thus, the re-referencing process completely eliminates that signal and yields zero voltage at every data point. This data would completely lack informative features and would not be useful for spectral analysis or classification. While averaging over a region with two electrode channels is possible, the re-referenced data would still not be particularly useful. Since electrodes are grouped linearly or in a grid, it is likely that one or two electrodes channels that are alone in a region are actually close together and at the edge of that region, with most of the channels that share their strip, grid, or depth electrode residing nearby in the closest neighboring region. Therefore, it is better to group these lone electrodes with their nearest neighbors rather than creating faulty data for a region that is not well-populated.

2.2.4 Weighted Average (Spatial Laplacian)

This method is a measure of the second spatial derivative, also called the spatial Laplacian. It accounts for not simply the gradient or change in the electric field, but also the distribution of that change in space (i.e. the spatial derivative/rate of change of the gradient). The idea is that from each individual channel you subtract a *weighted* average of the activity at all other channels. The weighting is based on the Euclidean distance between channels. Consequently, activity at nearby electrodes has a greater effect on the re-referenced signal than activity at distant electrodes. So, for each electrode i among n total electrodes:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

$$w_{ij} = e^{-(\frac{d_{ij}}{\sigma})^2}$$

$$W_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n-1} w_{ij}}$$

$$\Delta V_i' = \Delta V_i - \sum_{j=1}^{n-1} V_j \cdot W_{ij}$$

The parameter σ represents the standard deviation of the normal distribution represented by the Gaussian function e^{-x^2}, or the width of the well-known "bell-shaped curve". Adjusting this parameter changes the degree to which this scheme is localized. When the σ value is small, the weighting function only assigns significant weights to electrodes to the nearest neighbors of the electrode in question and is therefore similar to the bipolar reference scheme. When the σ value is very large, the weighting function assigns virtually equal weights to all electrodes and is therefore similar to an average

reference scheme. For three subjects ('R1060M', 'R1061T', 'R1112M'), I calculated AUC (methods described in the section below) for 30 logarithmically spaced values of σ between 1 and 1000. Since AUC values plateaued above values of 100, I was able to determine that the scale of interest for σ is between 10 and 100. I then calculated AUC for 10 linearly spaced values of σ between 10 and 100 for 16 subjects (I excluded the rest of the subjects from this optimization in order to avoid over-fitting the data). I averaged the values of σ that maximized AUC for these subjects and obtained an optimal parameter of $\sigma = 50.625$, which I used for my analysis across all 42 subjects.

2.3 Data Processing and Computing Power

In my analysis, I studied encoding events from the FR1 experiment. I looked at an encoding period from 0.0 seconds to 1.366 seconds after onset of the study word, with an additional 1.365 second buffer period at both ends. These parameters, and all other parameters described in this section, are standardized for all lab analyses on FR experiment participants. They are stored as an object in *ramutils.parameters*.

All data processing methods described in this section are identical for both referencing schemes. After loading the voltage values for each channel, I used a Butterworth Filter to remove 60 Hz line noise. Next, I used a Morlet wavelet transform (with a wavenumber of 5) to compute power at 8 different frequencies: 6.0 Hz, 9.75368156 Hz, 15.85571732 Hz, 25.77526961 Hz, 41.90062864 Hz, 68.11423148 Hz, 110.72742057 Hz, and 180.0 Hz. After removing the buffer period, I log-transformed the power values. I averaged the power over the time dimension and then z-transformed the power values across all encoding events within each session. These processes generated normalized power values for every combination of channels and frequencies for all encoding events.

2.4 Classification

Using the normalized power values and the outcomes of all encoding events (recalled vs. not recalled), I trained and tested a logistic regression classifier with a leave-one-session-out cross-validation approach. The classifier used a penalty parameter of $C = 0.00072$. In each session, I employed a weighting scheme based on *ramutils* code to account for the proportions of recalled and not recalled words within that session. Then, I generated ROC curves and computed their AUC to assess the performance of the classifier for each referencing scheme.

To determine if the difference in classifier performance was significant, I found the difference in AUC between referencing schemes for each subject and conducted a single-sample t-test comparing those values to zero. P-values smaller than $p = 0.05$ were considered statistically significant.

3 Figures

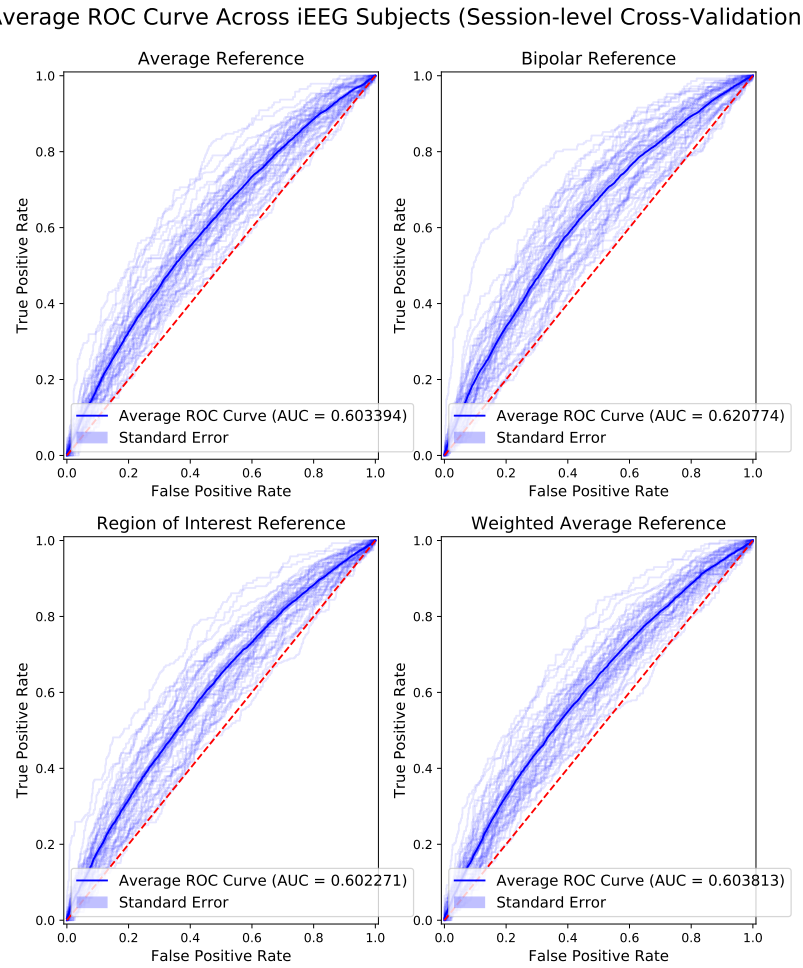

Figure 1: ROC curves for all 42 participants under each referencing scheme, along with the mean ROC across participants and error bars indicating the standard error of the mean.

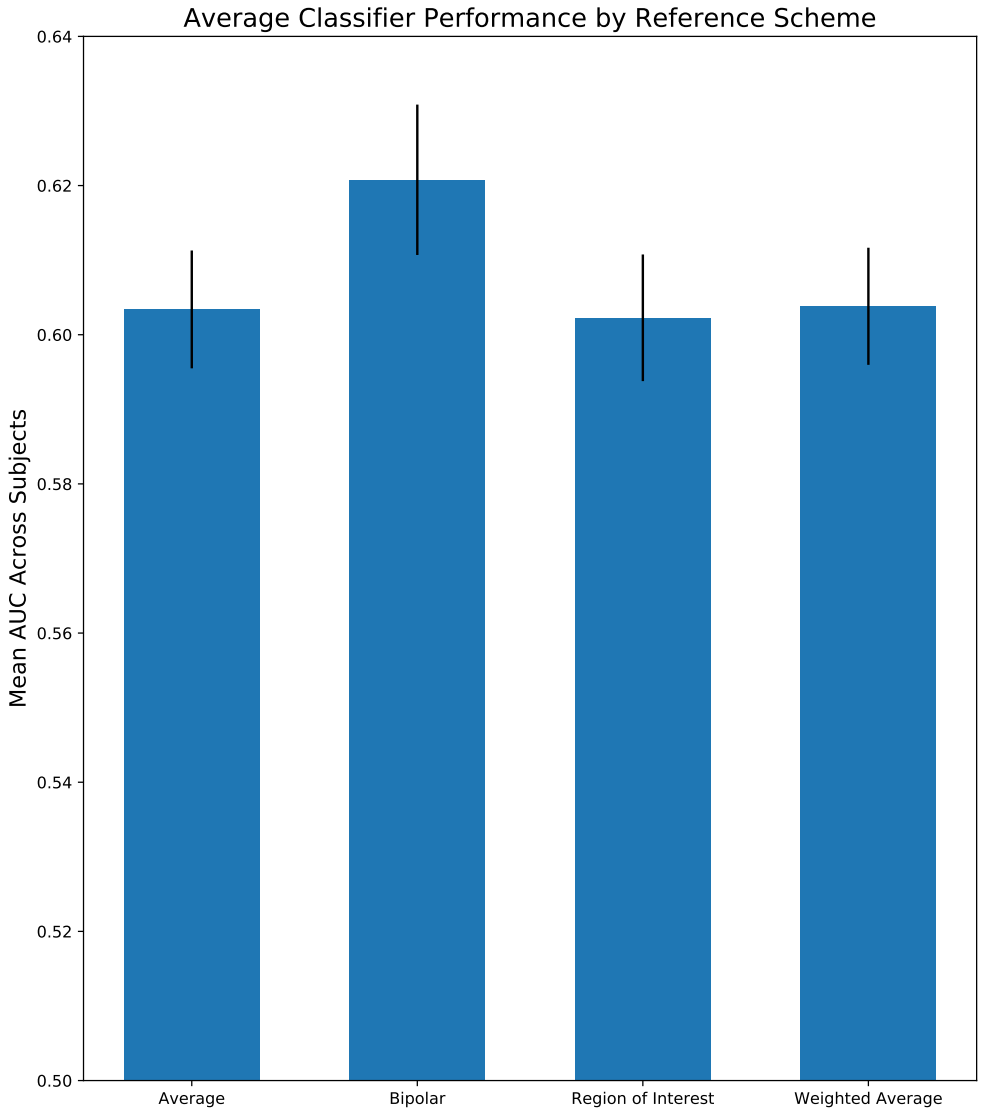

Figure 2: Mean AUC across participants for each referencing scheme. Error bars indicate standard error of the mean. The bipolar scheme yields the greatest mean AUC value, but it is important to note the small scale of this graph. The difference between schemes is marginal.

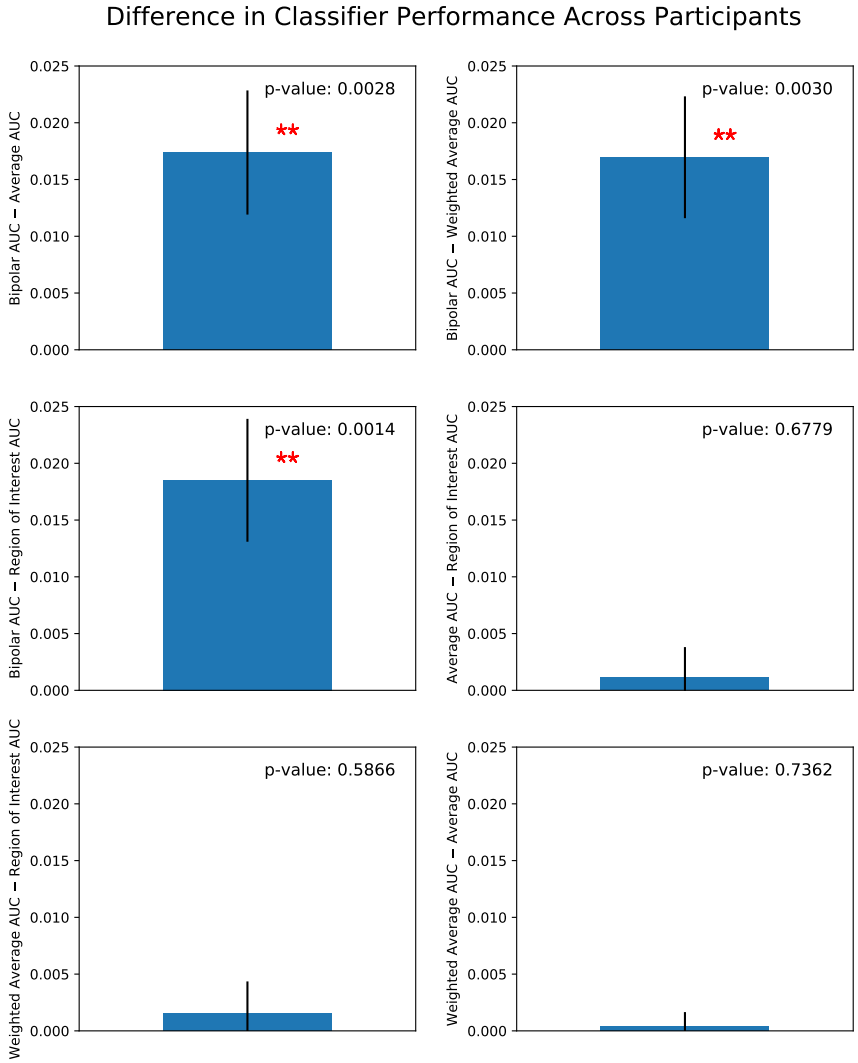

Figure 3: Mean ΔAUC between various reference schemes. Error bars indicate standard error of the mean. Double asterisks indicate statistical significance ($p < 0.05$), which was calculated by performing a one-sample t-test comparing ΔAUC values to zero across subjects. While the bipolar reference is significantly different from all three other schemes, none of those three schemes are significantly difference from one another.

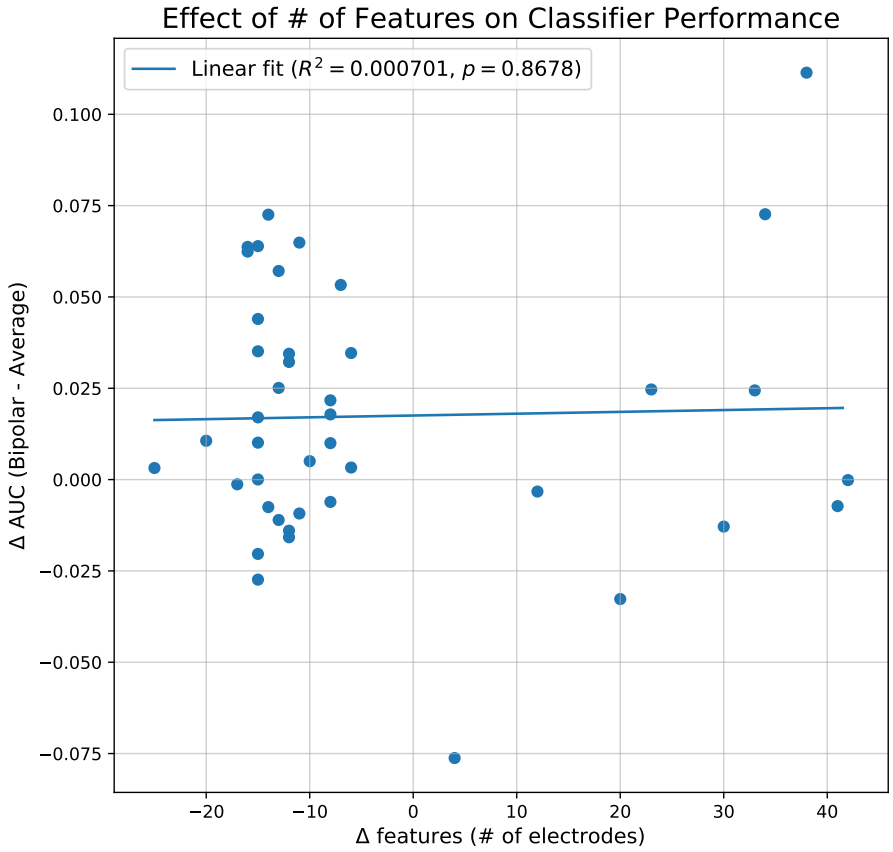

Figure 4: Plot showing the relationship between a discrepancy in the number of features between bipolar and average reference scheme and the difference in performance of those schemes. Since the bipolar scheme often has a different number of electrodes than an average reference scheme, the number of features interpreted by the classifier could potentially be a serious confounding factor that might cause a difference in AUC between schemes. This graph, however, shows that there is no relationship between the difference in number of features and the difference in AUC.

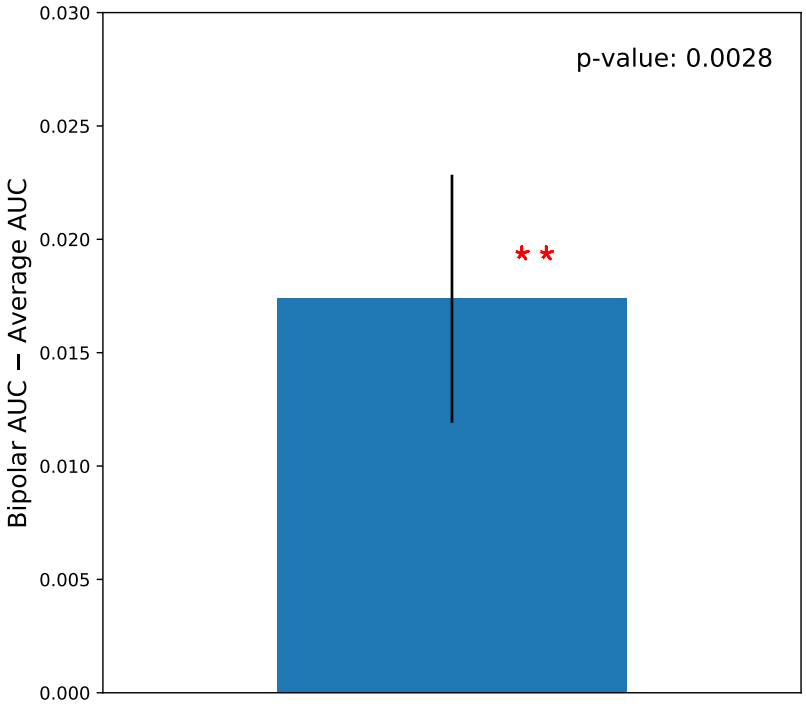

Figure 5: Difference in AUC between the bipolar and average reference scheme among subjects for whom the average reference had more features than the bipolar reference. Even among these subjects, a bipolar reference scheme performs significantly better than an average reference scheme.

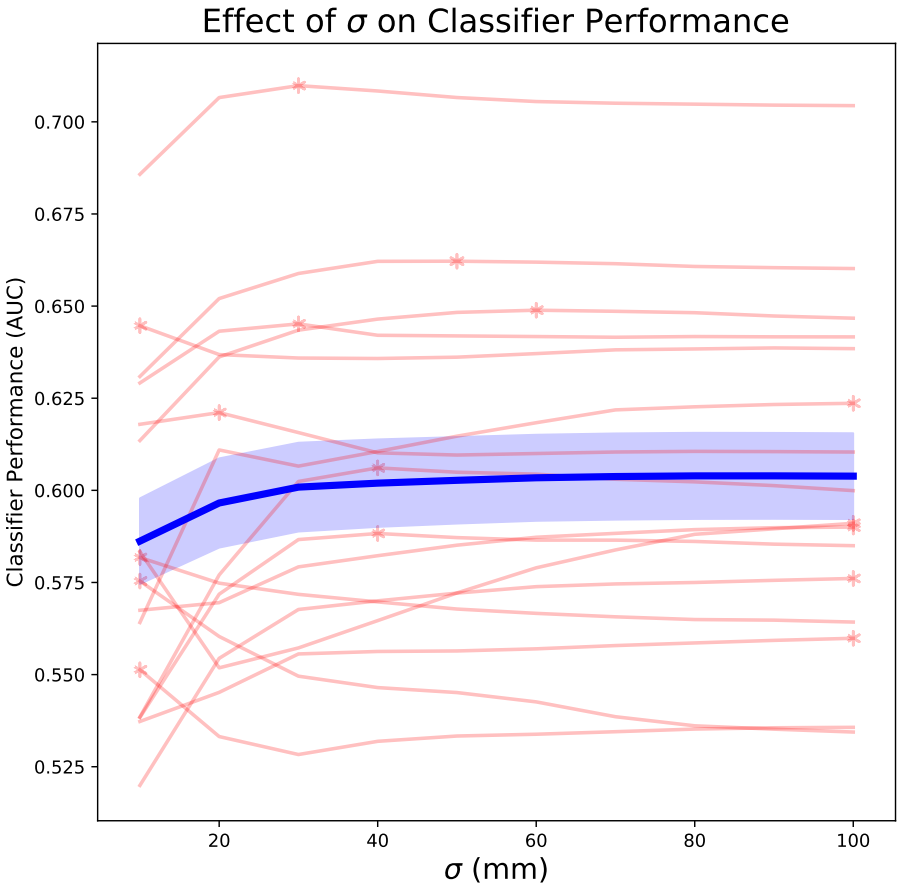

Figure 6: Classifier performance at the 10 assessed values of σ for 16 subjects. Asterisks indicates maximum AUC for a given subject. Blue line indicates mean classifier performance across all 16 subjects, with standard error of the mean indicated by the blue error bars. On average, AUC increases steadily with σ for the first 30-50 millimeters.

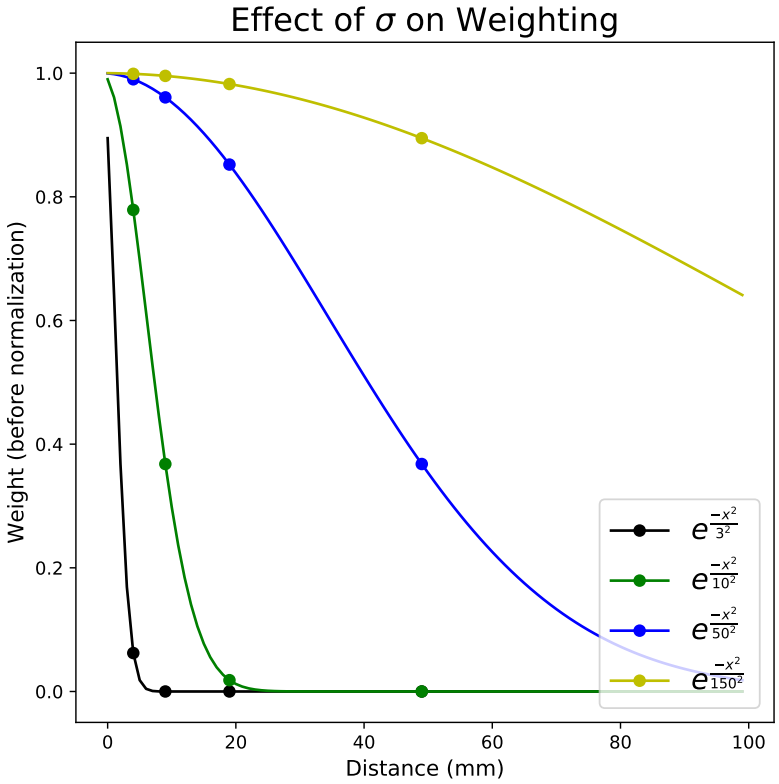

Figure 7: Electrode weights, before normalization, as a function of distance. Each line represents a weighting function with a different parameter σ. The circular markers represent some possible locations of electrodes. Importantly, electrodes past the point of inflection have significantly diminished weights. If the electrodes are all on the same side of the point of inflection, their weights are not so different from one another. In order to isolate the activity of the nearest neighboring electrode (thus approximating the bipolar reference), the point of inflection (which equals σ) must lie in between the nearest electrode and the distant electrodes that should be filtered out.

4 Conclusion

The data support the conclusion that a bipolar reference scheme is more optimal for the purposes of classification than an average reference, a reference based on regions of interest, or a weighted average reference. While this result is statistically significant, the difference in classifier performance is marginal ($\Delta AUC < 0.02$).

This result suggests that successful classification depends upon high spatial frequency components of the EEG signal rather than low spatial frequency components. In other words, the informative features which are predictive of successful recall are local changes in the electric field rather than global changes. It is surprising that there is no significant difference in the performance of the average, ROI, and weighted average schemes. Though none are as sensitive to local changes as the bipolar reference, the ROI and weighted average schemes are still significantly more localized than the average reference.

It is possible that the informative features of the electrophysiological data are localized within regions significantly smaller than lobes. Thus, a more spatially proximate reference, as used in the bipolar scheme, is necessary to optimize classifier performance.

It seems confusing at first that the optimal parameter $\sigma = 50.625$ (mm) is so much larger than the small distances which typically separate the electrodes in a bipolar pair. On average across subjects, classifier performance clearly improves steadily over the first 30-50 millimeters. This counter-intuitive trend is caused by inconsistency in electrode distances across subjects. If σ is smaller than the smallest distance between any two channels (the distance separating bipolar pairs), then the nearby informative channels will not be assigned a large weight. This means it will not be sufficiently highlighted by the spatial filter, and will approximate an average reference much more closely than a bipolar reference. Figure 7 is helpful in visualizing this problem.

The weighted average scheme might also yield better performance if the parameter σ were optimized for each subject through a more thorough and computationally intensive method such as a nested cross validation.