An algorithm for automated spike detection in large-scale extra-cellular electrophysiological recordings with electrical stimulation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Simultaneous electrical stimulation and recording of the neural tissue in micro-electrode arrays (MEAs) has been proposed as a prominent technology for achieving new advances in neuroscience. However, to fully exploit its capabilities in large-scale contexts some technical problems have to be overcome. One of such problems has to do with the corruptions in recordings induced by the stimulus (artifacts), which limits heavily the identifiability of neural signals. Here we provide a principled solution to this problem: by modeling the stimulation artifact as a Gaussian Process we are able to faithfully represent its structure, and to come up with an algorithm that quickly estimates and subtract this artifact. The effectiveness of our method is demonstrated in both real and simulated large-scale data (512 electrodes).

1 Introduction

There is a consensus that many new developments in systems neuroscience and neural engineering (e.g. prosthetic devices) will rely in our ability to control neural activity through the exogenous stimulation of the neural tissue [4]. In this line, the feedback control of the neural tissue has been proposed as a particularly fruitful research avenue [5, 1]. However, such closed-loop interactions require a proper read-out of elicited neural activity in response to stimuli, calling for a methodological shift, as many current methods and technologies for this read-out were conceived for the passive observation of neurons, but break down in regimes dominated by stimulation.

We focus on the simultaneous extra-cellular electrical stimulation and recording in micro-electrode arrays, where a proper computational infrastructure for the analysis of neural signals in the large-scale case is still lacking, as spike sorting —the method that allows identification of neurons from their electrophysiological spatio-temporal fingerprints— cannot handle the corruptions induced by electrical stimuli, since these corruptions —the stimulation artifacts— can be of much greater magnitude than and overlap with the actual neural activity.

Although many approaches have been proposed to tackle this problem [2, 6, 3], they all fall short in important aspects: they are often based on very restrictive assumptions and don't exploit the inherently spatial setup imposed by the MEAs. In consequence, often there is no other solution than throwing away an important proportion of recordings, which leads to biased research results that not necessarily have to do with the regimes where the most interesting neuronal dynamics do occur.

In this work we develop a modern large-scale principled framework for the analysis of neural data in this regime corrupted by artifacts, based on a comprehensive account of the variability of the artifact in the spatio-temporal and stimulus dimensions. Specifically, we characterize this highly structured artifact and model it in the Gaussian Process framework, where we leverage recent advances in machine learning to enable a fast and scalable implementation. Then, we come up with a spike sorting algorithm and demonstrate its effectiveness by comparison to human-curated ground truth from the primate retina. Although some features of our method are context-dependent, we discuss extensions to other scenarios, stressing the generality of our approach.

2 Method

In this section we develop a method for identifying neural activity in response to electrical stimulation. At the highest level, it is based on the following generative model

$$Y = A + s + \epsilon, \tag{1}$$

where Y represent the observed traces, A is the stimulation artifact, s is the neural activity and ϵ is a noise term. We divide the exposition of our method in four parts. First, we describe the nature of the data that is available for analysis (2.1). Second, we describe the structure of the stimulation artifacts (2.2). Third, we propose a Gaussian Process model for these stimulation artifacts (2.3). Finally, we come up with an algorithm that produces an estimate of A and s (2.4).

2.1 Setup

Data is made up by: (1) recordings in response to stimulation, including stimulation covariates and (2) electrical images (EI); the spatio-temporal fingerprints of targeted neurons.

Regarding (1), we assume voltage traces are available over a set of E electrodes (the array) and a discrete time window of length T (representing a multiple of the sampling rate) in response to J different stimuli on a single stimulating electrode (fixed). Although stimuli can depend on many features (e.g. specific structure of the current pulses that are passed through through stimulating electrode) we assume strength or amplitude of stimulus can be finally summarized in terms of the quantities a_j, assumed increasing. Also, for each stimulus, we assume a number of n_j repetitions are available.

Regarding (2), we assume EIs are available for all of the N neurons under study: each of these templates is represented as a $E \times T'$ matrix containing an estimate of the voltage deflections produced by a spike over the array in a length T' time window, and aligning the onset of a spike to an arbitrary value. EIs can be obtained in a separate experiment in the absence of electrical stimulation, using standard large-scale spike sorting methods. Figure 1a contains an example of many EIs obtained during a visual stimulation experiment.

2.2 Stimulation Artifacts

Electrical stimulation experiments where neural responses are ablated (e.g. using the neurotoxin TTX) provide qualitative insights about the structure of the stimulation artifact $A(e, t, j, i)$; that is, how it varies as a function of all the relevant covariates: space (represented by electrode, e), time (t), amplitude of stimulus a_j and stimulus repetition i. The very first distinction to make is between stimulating and non-stimulating electrodes: magnitude of artifacts are much stronger in the former than in the latter, which is specially problematic for the recording of neural activity in response to somatic stimulation. However, because the modeling is more context-specific for this electrode we defer its treatment for the appendix and focus here on the non-stimulating electrodes, where the artifact has the follow properties: i) in the time component, its magnitude peaks following the onset of stimulus, and then stabilizes smoothly ii) in the spatial component, artifact decays smoothly with distance to the stimulating electrode iii) in the stimulus component, magnitude of artifact increases smoothly with increasing stimulus strength iv) repetitions of stimulation lead to the same artifact up to random fluctuations. Then, omitting such trial-dependent random fluctuations, we can understand artifact as a stack of movies $A(e, t, j)$, one for each amplitude of stimulation a_j. Figures 1b, 1c and 1d depicts the main features of this artifact

We finish our description of the artifact with a useful metaphor that will help us think in the structure of the artifact and the problem in more concrete terms: one can think of the artifact as the wave that results from throwing a food pellet to a pond, where bigger pellets (stronger stimulus) leads to waves of larger magnitude. Our aim is to observe the fish (neurons) that come to eat the food, but this is difficult because the perturbations make the water blurry.

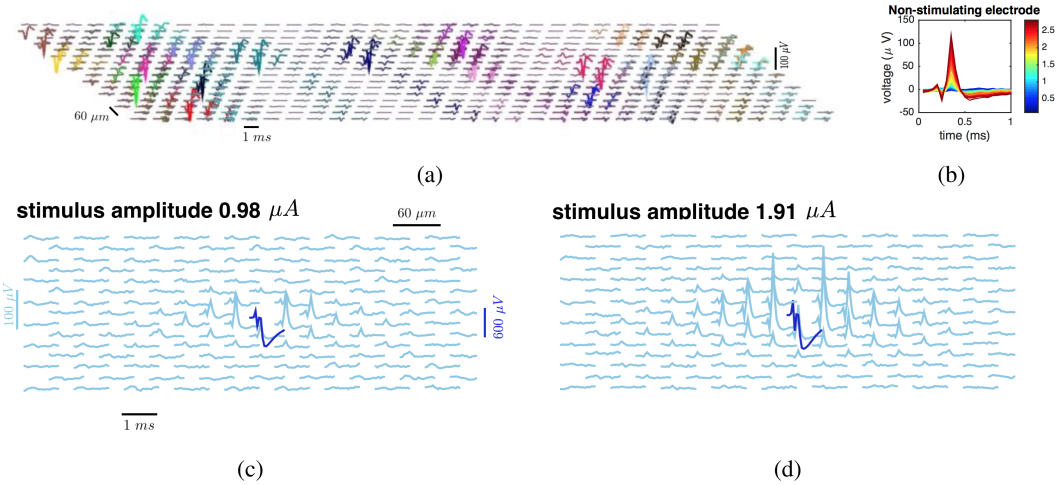


Figure 1: (a). Overlapping EIs of 24 neurons (different colors) over a MEA, aligned to onset of spiking at $t = 0.5ms$. Each trace represent the time course of voltage at a certain electrode (for each neuron, traces are plotted only in the electrodes that contribute the most to the EI). (b) Artifact, as a function of amplitude of stimulus (different colors, in μA) for one of the nearest neighbors of the stimulating electrode. (c) and (d): artifacts on a neighborhood of the stimulating electrode. Each trace represent the time course of voltage at a certain electrode. Notice the stimulating electrode (blue) and non-stimulating electrodes (light blue) are plotted in different scales.

2.3 A scalable GP model for stimulation artifact

Here we propose a generative model for the artifact; namely, we model observed artifacts as draws from a GP, $A \sim GP(\mu, K)$. We choose the GP framework since it has shown to provide a parsimonious framework for the estimation of non-linear functions in the context of uncertainty, and because it leads to a straightforward implementation of all the usual operations that are relevant for our purposes (e.g. extrapolation) in terms of some conditional gaussian distributions. Now we address how to make a sensible choice of the parameters (μ, K), that is, which parametric families are to be chosen and how to perform statistical inference in there.

Regarding μ, we follow the standard in the applied statistics community: μ is a centering parameter and all the non-random aspects of data should be captured by it. In our case this component is given by what we call the switching artifact, a waveform $A_0 = A_0(e,t)$ that is present regardless of the amplitude of stimulation. As a result, we can estimate $\hat{\mu}$ by taking the mean of recordings at the lowest amplitude of stimulation (see appendix for details) leading to a zero-mean Gaussian process.

The choice of K is more involved: it is well-known that the bottleneck in kernel methods lies in the inversion of K, which has complexity $O(d^3)$. In our context, $d = T \times E \times J \sim 10^6$, which renders the above inversion intractable unless further kernel structure is exploited. Here we implement a recent insight in the machine learning community: imposing a Kronecker product structure in K leads to tractable and scalable inferences. In detail, we assume the following decomposition

$$K = \rho K_t \otimes K_e \otimes K_s + \sigma^2 I_d. \qquad (2)$$

where K_t, K_e and K_s are the kernels that account for variations in the time, space and stimulus dimensions of data, respectively. The dimensionless quantity ρ is used to control de overall magnitude of variability and the term $\sigma^2 I_d$ is included in acknowledgement that what is finally observed is a noise-corrupted version of the actual artifact.

Now this structured Kernel has been stated it remains to specify parametric families for the elementary kernels K_t, K_e, K_s. As shown below, we construct them from the Matérn family, using extra parameters to account for the non-stationarities described in 2.3.

2.3.1 A non-stationary family of kernels

Consider the Matérn(3/2) Kernel, the continuous version of an AR(2) process. It's (stationary) Covariance is given by

$$K_\lambda(x_1, x_2) = K_\lambda(\delta = |x_1 - x_2|) = \left(1 + \sqrt{3}\delta\lambda\right) \exp\left(-\sqrt{3}\delta\lambda\right). \tag{3}$$

Here $\lambda > 0$ represents the (inverse) lengthscale and determines how fast correlations decay with distance. We induce non-stationarities by introducing the family of unnormalized gamma densities $d_{\alpha,\beta}(\cdot)$:

$$d_{\alpha,\beta}(x) = \exp(-x\beta)x^\alpha. \tag{4}$$

By an appropriate choice of the pair $(\alpha, \beta) > 0$ we aim to expressively represent non-stationary 'bumps' in variability. Then, to actually come up with a non-stationarity family we consider the process $Z \equiv Z(x) = d_{\alpha,\beta}(x)Y(x)$ where $Y \sim GP(0, K_\lambda)$. Then, Z is a *bona fide* GP [1] with the following covariance matrix ($D_{\alpha,\beta}$ is a diagonal matrix with entries $d_{\alpha,\beta}(\cdot)$):

$$K(\lambda, \alpha, \beta) = D_{\alpha,\beta} K_\lambda D_{\alpha,\beta}. \tag{5}$$

We choose all the kernels K_t, K_e, K_s as in (5): for the time and stimulus Kernel we use time and stimuli as covariates (δ in equation (3) and x in (4)), respectively. However, the case of the spatial Kernel is different: while δ represents distance between recording electrodes, now x represents distance between stimulating and recording electrodes, as it is this distance the one that covaries with artifact magnitude.

2.3.2 Kernel Learning

From equations (2), (3) and (5) our GP for the artifact is completely specified in terms of the parameters $\theta = (\rho, \alpha, \lambda, \beta)$ (one triplet (α, λ, β) for each of the space, time and stimulus kernels) and σ^2. In other word, we encode all the peculiarities — lengthscales, orders of magnitudes, size of non-stationarities — of observed artifacts in different experiments through these parameters. Let's then call $K^\theta = \rho K_t \otimes K_e \otimes K_s$ and $K^{(\theta,\sigma^2)} = K^\theta + \sigma^2 I_d$. Our parameter learning problem is simply the maximization of the following function:

$$\max_\theta \log p(\tilde{A}|\theta, \sigma^2) = -\frac{1}{2}\tilde{A}^t \left(K^{(\theta,\sigma^2)}\right)^{-1} \tilde{A} - \frac{1}{2}\log\left|K^{(\theta,\sigma^2)}\right|. \tag{6}$$

Notice first that to avoid numerical instabilities we not include σ^2 in the optimization; alternatively one can come up with simple estimates of the background noise by using stimulus-free data. Also, here $\tilde{A}$ represents some guess of the artifact: naturally the artifact itself is not revealed (that would solve the problem) but as we are only concerned with parameters that inform about the global and overall properties of the artifact, using a guess instead would not lead to big changes in θ. In practice, we use data corrupted by neural activity, that is $\tilde{A} \equiv Y$. This is a sensible decision as neural activity is sparse and small compared to artifact, and does indeed lead to results that are very close to guesses $\tilde{A}$ from TTX experiments (where available). Figure 2 shows an example of learned Kernels K_t, K_e and K_s using this approach.

2.4 Algorithm

We start by fully specifying the different aspects in equation (1), where we aim to produce estimates $\hat{A}$ and $\hat{s}$. We adopt the convention that whenever indexes are specified we refer to the subvector of the huge vector (either Y,s,A) [2] that created by fixing that component (e.g. $A(t, e)$ is the artifact at time t and electrode e moving freely along different stimulus amplitudes). With these convention, to clearly state trial-wise dependencies, we re-express equation (1) as:

$$Y(t, e, j, i) = A(t, e, j) + s(t, e, j, i) + \epsilon(t, e, j, i) \quad t \leq T, e \leq E, j \leq J, i \leq n_j. \tag{7}$$

[1] Indeed, the above covariance can be thought as the Hadamard or pointwise product between K_λ and the degenerate Kernel dd^t with d being the diagonal of $D_{\alpha,\beta}$.

[2] The same convention also applies for submatrices of K^θ or $K^{(\theta,\sigma^2)}$.

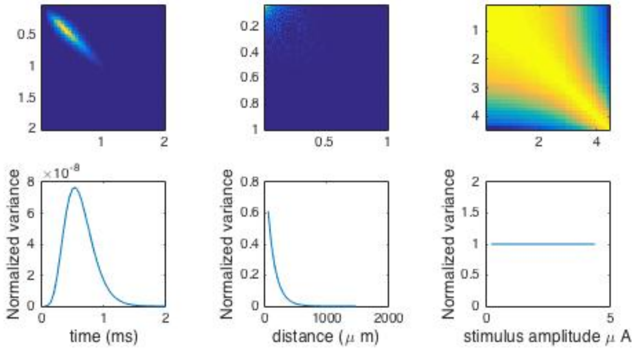


Figure 2: Example of kernels K_t, K_e and K_s learned from real data with $\tilde{A} = Y$

Noise is assumed uncorrelated gaussian with variance σ^2, that is, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{d'})$ with $d' = T \times E \times J \times \sum_{i=1}^{J} n_j$. This assumption is rather restrictive and we assume it here for computational ease, but defer the reader to the supplementary material for a sketch of a formulation that takes into account correlated noise.

In agreement with 2.3 we take $A \sim GP(0, K^\theta)$. This means we conceptualize A as a noise-free artifact and then $A + \epsilon$ can be thought as its noise corrupted version, with covariance $K^{(\theta,\sigma^2)}$.

We represent neural activity s as the superimposed activity of the N neurons involved, each represented by an appropriate convolution of the EI. In detail, we first consider binary vectors b^n that indicate our constraint on spike occurrence and timing: for each trial and neuron there can be at most one spike, which can only occur on a pre-defined discrete set of times (a strict subset of the entire recording window to prevent aliasing). This is the case in reality, as the recording window is comparable with the refractory period of neurons. Then, the actual activities s^n are obtained through products with the Toeplitz matrices M^n that contain copies of the EIs with spike onset aligned at different times. All of this is summarized in equation (8): [3]

$$s = \sum_{n=1}^{N} s^n, s^n = M^n b^n. \tag{8}$$

Now we turn to the parameter estimation problem, based on posterior inference for $p(A, s|Y, \theta, \sigma^2) \propto p(Y|s, A, \sigma^2)p(A|\theta)$. The non-convexity of the set over which the binary vectors b_n are defined makes this problem difficult as many local optima do exist in practice and, as a result, for global optimization there may not be a better alternative than to look at a huge number of possible cases. We circumvent this cumbersome global optimization by taking a principled greedy approach, with two main characteristics i) joint optimization over A and s is addressed with alternate ascent and ii) data is divided in batches corresponding to the same stimulus amplitude, and the analysis for the $j+1$-th batch starts only after definite estimates $\hat{s}_{1,\ldots,j}$ and $\hat{A}_{[j]}$ [4] have already been produced. Moreover, this latter estimate of the artifact is used to produce an initial estimate for A_{j+1}. Intuitively, we borrow strength from lower amplitude of stimulation conditions to counteract the worsened effect of artifacts and increased responses of neurons due to high amplitudes of stimulation.

Now we describe in detail these features, which finally lead to the algorithm that is sketched at the end of this section.

2.4.1 Coordinate Ascent

Given the batch Y_j and an initial artifact estimate A_j^0 (see 2.4.2) we alternate between neural activity estimation given a current artifact estimate, and artifact estimation $\hat{A}_j$ given the current estimate

[3] Note the slight abuse of notation: while the above definitions make sense trial-wise, when considering the entire dataset, actually one deals with many matrices M^n concatenated vertically.

[4] $[j]$ denotes the set $\{1, \ldots, j\}$

of neural activity. This alternate optimization stops when changes in $\hat{s}_j$ are sufficiently small, or nonexistent.

Matching pursuit for spike maximization: given the current artifact estimate $\hat{A}_j$ we maximize the conditional for neural activity, which turns out to be the following sparse regression problem (the set S embodies our constraints on spike occurrence and timing):

$$\min_{b_j^n \in S, n=1,\dots,N} ||(Y_j - \hat{A}_j) - \sum_{n=1}^{n} M^n b_j^n||^2. \tag{9}$$

Intuitively, we seek to place spikes so that they match with the residuals $(Y_j - \hat{A}_j)$ the best; that is, they lead to the smallest sum of squares. For computational ease, we proceed greedily: by concatenating all the M^n and b^n we can find the neuron and spike time that decreases the above residuals the most, and after extracting the corresponding (b^n, M^n) from (b, M) we can continue iteratively until no spike placement of the remaining neurons leads to further decrease in the sum of squares.

Bayesian regression for spike maximization:gGiven the current estimate of neural activity $\hat{s}_j$ we maximize the conditional of the artifact, that is, $\max_{A_j} p(A_j|Y_j, \hat{s}_j, \theta, \sigma^2)$, which — by well known properties of the gaussian distribution — leads to the posterior mean estimator (the overline indicates mean across the n_j stimulation repetitions):

$$\hat{A}_j = E(A_j|Y_j, \hat{s}_j, \theta, \sigma^2) = K^{\theta}_{j,j}\left(K^{(\theta,\sigma^2)}_{j,j}\right)^{-1}(\bar{Y}_j - \bar{\hat{s}}_j). \tag{10}$$

2.4.2 Iteration over batches and artifact extrapolation

The procedure described in 2.4.1 is repeated in a for loop that iterates through the batches corresponding to different stimulus strengths, from the lowest to the highest. Also, when doing $j \to j+1$ an initial estimate for the artifact A^0_{j+1} is generated by extrapolating from the current, faithful, estimate of the artifact up to the j-th batch. This extrapolation is easily implemented as the mean of the noise-free posterior distribution in this GP setup, that is:

$$A^0_{j+1} = E(A_{j+1}|\hat{A}_{[j]}\theta, \sigma^2) = K^{\theta}_{(j+1,[j])}\left(K^{(\theta,\sigma^2)}_{([j],[j])}\right)^{-1}\hat{A}_{[j]}. \tag{11}$$

Importantly, in practice this initial estimate ends up being extremely useful: without a proper estimate of the artifact, coordinate ascent will often lead to poor optima, but the accurate estimates provided by this extrapolation prevent that from being the case.

```
Algorithm: Spike sorting with electrical stimulation artifacts.
    Input: traces $Y = (Y_j)_{j=1,\dots,J}$, in response to $J$ stimulus. EIs of $N$ neurons.
    Output: estimates of artifact and neural activity (spikes) for each neuron.

Initialization: estimate $\sigma^2$ and $\theta$ as in (6). (Can re-use these parameter for data from
                different stimulating electrodes)
Then: repeat for j=1,..., J
                Construct $A^0_j$ from [j-1] using equation (11) ($A^0_1 \equiv 0$)
                repeat until little or no changes in binary vectors $b^n_j$.
                        • Estimate $\hat{s}_j$ by greedily finding spikes as in (9)
                        until no spike of no neuron leads to increase in likelihood.
                        • Estimate $\hat{A}_j$ using (10)
```

3 Results

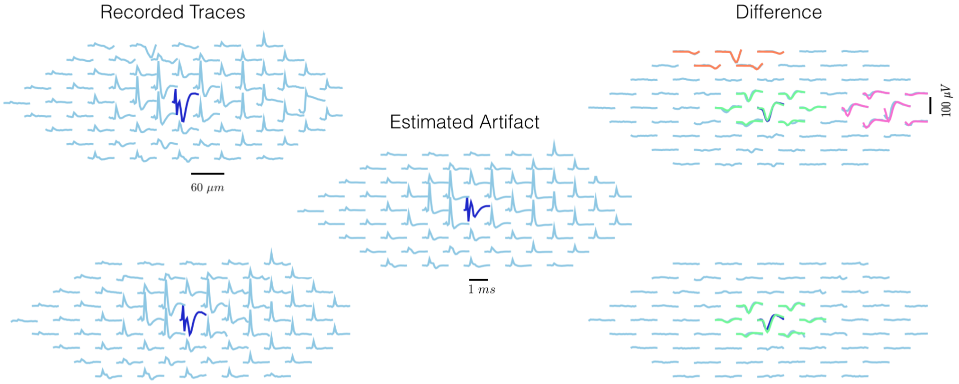
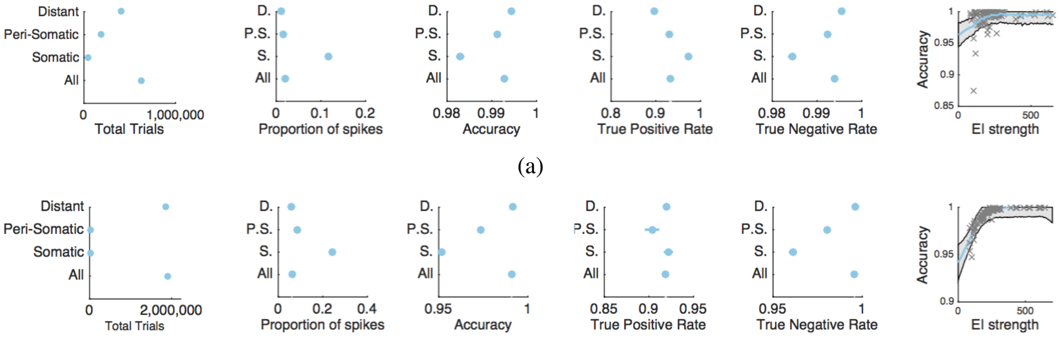


Figure 3: An example of neural activity and artifact inference in a neighborhood of the stimulating electrode. *Left:* Two recordings in response to a 2.01 μA stimulus. *Center:* estimated artifact (as the stimulus doesn't change, it is the same for both trials). *Right:* Difference between raw traces and estimated artifact, with inferred spikes in color. In one case (above) three spiking neurons were detected, while in the other (below) there was only one.



(a)

(b)

Figure 4: Population-level results for both real (a) and simulated (b) data. In each case, the first five plots from left to right show different measures — total number of available trials, observed frequency of spikes, accuracy, and true positive and true negative rates — as a function of distance between targeted neuron and stimulating electrode (somatic, peri-somatic, distant and any distance). 95% Confidence intervals are usually small enough so they cannot be seen. Finally, the last plot shows the accuracy (GP fit with confidence bands) as a function of the l_2 norm of the EI over the 10 more relevant electrodes of a neuron, a proxy for the strength of the EI.

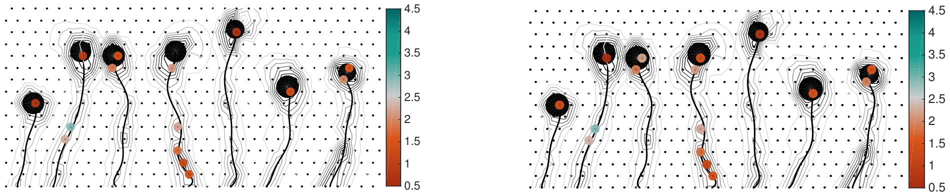
References

[1] Felix Franke, Rodrigo Quian Quiroga, Andreas Hierlemann, and Klaus Obermayer. Bayes optimal template matching for spike sorting–combining fisher discriminant analysis with optimal filtering. *Journal of computational neuroscience*, 38(3):439–459, 2015.

[2] Takao Hashimoto, Christopher M. Elder, and Jerrold L. Vitek. A template subtraction method for stimulus artifact removal in high-frequency deep brain stimulation. *Journal of Neuroscience Methods*, 113:181–186, 2002.

[3] Leon F. Heffer and James B. Fallon. A novel stimulus artifact removal technique for high-rate electrical stimulation. *Journal of Neuroscience Methods*, 170:277–284, 2008.

[4] Lyric A Jorgenson, William T Newsome, David J Anderson, Cornelia I Bargmann, Emery N Brown, Karl Deisseroth, John P Donoghue, Kathy L Hudson, Geoffrey SF Ling, Peter R MacLeish, et al. The brain initiative: developing technology to catalyse neuroscience discovery. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1668):20140164, 2015.



(a)

(b)

Figure 5: Comparison of spatial sensitivity maps, human (*left*) v.s. algorithm (*right*). Individual neurons are shown in black (black circle representing inferred soma and black line representing a polynomial fit to the axon). Colored dots are electrodes that activated a neuron, and the with color scale the stimulus strength (in μA) required to achieve a 50% probability of activation

[5] Steve M Potter, Ahmed El Hady, and Eberhard E Fetz. Closed-loop neuroscience and neuroengineering. *Closing the Loop Around Neural Systems*, page 7, 2014.

[6] Daniel a. Wagenaar and Steve M. Potter. Real-time multi-channel stimulus artifact suppression by local curve fitting. *Journal of Neuroscience Methods*, 120:113–120, 2002.

4 Appendix

4.1 Stimulating Electrode

Second, regarding the stimulating electrode: i) generally speaking, magnitude of artifact is much greater than in the non-stimulating electrodes ii) variations in time also peak right after the onset of stimulation but oscillations are wilder and iii) artifact magnitude increases with stimulus strength smoothness is lost: due to stimulation hardware idiosyncrasies, there might be some breakpoint stimulus strengths such that artifact shape and magnitude will ostensibly change after such strengths are surpassed.