A Computable Expression for the Spatial Distribution of Extracellular Spike Waveforms Enables *in Silico* Localization of Neurons

Alexander Olsson
alexander.olsson@bme.lth.se

I. INTRODUCTION

With implanted electrodes, it is possible to record the aggregate activity of demarcated populations of neurons within the central nervous system for long periods of time [1]. In a process typically referred to as *spike sorting*, recordings obtained in this way can (assuming sufficient signal quality) blindly be decomposed into contributions originating from individual neurons by exploiting inter-neuronal variation in extracellular spike waveforms arising from the spatial dependency of waveform morphology [2]. Such techniques can be invaluable for researchers that wish to study neuronal behaviours both at the individual and network level.

As most assays that involve invasive neural recordings are aimed at uncovering neuronal firing behaviour, processing pipelines for spike sorting typically do not produce any explicitly spatial information on the geometry of the latent neuronal signal sources that are detected. Notably, extracting this information is not trivial, as it requires solving an *inverse problem*. Challenging as it may be, the ability to infer the relative positions of individual neurons by processing extracelluar recordings could enable some interesting applications, clinical and otherwise—conjectured examples include electrode drift detection [3] and high-resolution monitoring of neuronal density [4].

Armed with explicit knowledge of the process that generates extracellular recordings, it is possible to treat the inverse problem of localizing individual neurons as an optimization problem. More formally, if one as access to a closed-form (computable) expression for the so-called *forward mapping* $f : G \longrightarrow R$ that transforms a given geometry $g \in G$ to a corresponding recording $r \in R$, it is possible, assuming that f is locally invertible, to formulate the inverse problem as the problem of finding the *candidate* geometry $\tilde{g}$ that minimizes $D(f(\tilde{g}), \hat{r})$, where $\hat{r}$ is a recording obtained from the true geometry and $D(\cdot, \cdot)$ is some appropriate distance metric. In this paper, I present and evaluate a rudimentary method based on this reasoning for solving the inverse problem of neuron localization by using a computationally efficient model of the forward mapping introduced in [5].

II. METHODS

In this section, I give an overview of (1) the proposed algorithmic framework for neuron localization and (2) a selection of *in silico* experiments performed for the purpose of evaluating the framework. The entirety of the current study was performed using custom Python 3.5 code[1].

If not otherwise specified, the sampling rate is assumed to be $f_s = 25000$ samples/s and the durations of recordings is assumed to be to be $D = f_s \cdot 60s = 1500000$ samples.

A. Neuronal Model

As the aim of this work was to simulate and subsequently locate neurons, a necessary first step was to define a suitable model of individual neurons. For the purposes of this study, a neuron was conceptualized as having 3 free parameters:

1) A spatial *position* $\mathbf{P}_n \in \mathbb{R}^3$, to be interpreted as the location of the soma in some 3D coordinate system. Succinctly, the aim of this study was to estimate this value from observations of extracelluar recordings.
2) A constant *firing rate* $F \in \mathbb{R}^+$, to be interpreted as the expected number of action potentials that the neuron fires per unit of time. While a stationary firing rate is not necessarily a biologically reasonable assumption, it is arguably a justifiable simplification in sufficiently short recordings.
3) A categorical *type* T. This auxiliary variable is intended to encode all inter-neuronal morphological and bioelectrical variation.

To use (populations of) this neuronal model to simulate extracellular voltage recordings, a computable expression for the *extracellular spike waveform*, dependent on the relative positions of neurons and electrodes, is needed. For this purpose, let $\mathcal{W}_T(\mathbf{P}_e, \mathbf{P}_n) \in \mathbb{R}^L$ denote the L voltages sampled isochronously by a hypothetical electrode located at $\mathbf{P}_e = (x_e, y_e, z_e)$ over a time of $\frac{L}{f_s}$ seconds following the onset an action potential fired by a neuron of type T at position $\mathbf{P_n} = (x_n, y_n, z_n)$. Whereas the localization framework presented in this paper is agnostic to the choice of mathematical formulation for $\mathcal{W}_T$, the model introduced in [5] was used in all parts of the current study for practical reasons. This specific formulation, wherein $T \in [1, 4]$, has several desirable properties for my aims here, most saliently its computational simplicity and full translation-invariance, i.e. $\mathcal{W}_T(\mathbf{P}_e, \mathbf{P}_n) = \mathcal{W}_T(\mathbf{P}_e - \mathbf{P}_n)$. To be able to make

[1]All code written for this paper is available at www.github.com/aeolsson/neuron_localization.

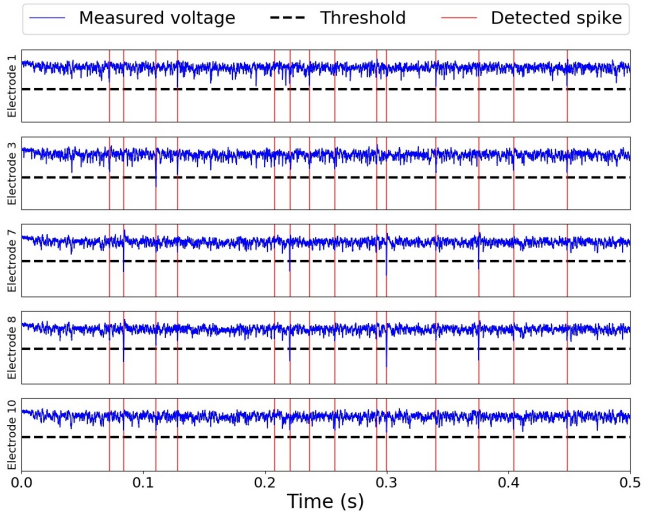

Fig. 1: A simulated recording, comprised of 5 randomly selected electrode channels, with appertaining spike detections.

straightforward use of preimplemented *.mat* files from [5], the duration of the waveform was in this study always set as $L = 100$ samples.

To quantify the aggregate effect which a single neuron has on an electrode over the course of an entire recording, knowing the spatial distribution of waveforms associated with a neuronal firing is not sufficient; it is also necessary to simulate at what times the neuron actually fires. For this, a gamma distribution [6] with shape parameter $k = 5$ and scale parameter $\theta = \frac{1}{kF}$ was used to sample neuron-specific inter-spike intervals (ISIs) that could subsequently used to generate a neuron-specific spike train $\mathbf{S}$, where $\mathbf{S}_t = 1$ if the neuron fires at time t and $\mathbf{S}_t = 0$ otherwise. The signal $\mathbf{X} \in \mathbb{R}^D$ originating from a neuron of type T located at $\mathbf{P}_n$ observed by an electrode located at $\mathbf{P}_e$ can then be obtained by convolving $\mathbf{S}$ with $\mathcal{W}_T(\mathbf{P}_e, \mathbf{P}_n)$. For a geometry with N neurons with types $T_1, ..., T_N$ at positions $\mathbf{P}_n^1, ..., \mathbf{P}_n^N$, each with an associated simulated spike train, the voltage signal observed by an electrode can be obtained by summation over the contributions of all neurons. I here assume linearity (i.e. adding an additional electrode does not affect the recordings of any other electrode)—the signal observed by each electrode in a multi-electrode system can be thus computed independently by following the aforementioned procedure electrode-wise.

B. Data Synthesis

To simulate extracellular neural recordings, a measurement geometry (i.e. a set of neurons and a set of electrode location) was first created. The neurons of the geometry consisted of (1) a population of *target neurons* and (2) a population of *noise neurons*. As has already been alluded to, the aim of the experiments of the current study was to locate, i.e. infer the positions of, all target neurons. Conversely, the purpose of the noise neurons was to realistically represent background noise that an electrode would be subject to.

For all experiments in this study, a population of noise neurons was constructed by defining a hollowed-out cylinder centered on the origin with outer radius $r_o = 250\mu m$, inner radius $r_i = 100\mu m$, height $h = 500\mu m$, and neuronal density $\rho = 9.5 \cdot 10^6$ neurons/μm^3. With this, a homogeneous population of noise neurons was generated as follows:

1) The expected number of noise neurons $\rho\pi(r_o^2 - r_i^2)(h) \approx 717.8$ was computed and rounded to the nearest integer $N_{noise} = 718$.
2) For all $i \in [1, N_{noise}]$:
   a) ξ_1 was sampled from the uniform distribution $\mathcal{U}(\frac{r_i^2}{r_o^2}, 1)$, ξ_2 and ξ_2 from the uniform distribution $\mathcal{U}(0, 1)$.
   b) Set $r = r_0 \cdot \sqrt{\xi_1}$, $\theta = 2\pi \cdot \xi_2$, and $z = \xi_3 \cdot h - \frac{h}{2}$.
   c) The position of the i:th noise neuron was set to $\mathbf{P}_n = (r\cos\theta, r\sin\theta, z)$.

Each noise neuron was randomly assigned a type between 1 and 4, and a firing rate between 1 Hz and 50 Hz.

The population of target neurons and the numbers and positions of electrodes differed between experiments; experiment-specific details are discussed in sections II-G, II-H, and II-I. In this paper, $\mathbf{P}_n^i$ and T_i refer to the true position and type, respectively, of the ith target neuron.

Once a geometry had been constructed, a recording matrix $\mathbf{X} \in \mathbb{R}^{E \times D}$, where E is the number of electrodes, was simulated by superimposing the contributions of all neurons as described in section II-A. As preprocessing, the electrode-wise sample mean was subtracted from the recording.

C. Spike Detection

For the ith electrode of the recording, the *spike detection threshold* was set to $-8 \cdot MAD_i$, where MAD_i is median absolute deviation of $\mathbf{X}_{i,*}$ (the recording of the ith electrode). A firing event (not yet assigned to any target neuron) was assumed to occur at any sample where the recorded voltage fell below this threshold. A first-order difference filter was used to ensure that only the first sample of a continuous sequence of samples below the threshold was counted as a detection. Spike times obtained in this way were lastly aligned by shifting them forward in time to the peak value reached by the signal within 10 samples.

To obtain a electrode-independent estimate of firing times, the individual times detected from all electrodes separately were simply pooled. A refractory period of 25 samples was assumed, meaning that spike events detected within $25/f_s = 1$ ms after a previously detected spike were discarded. A global list of firing times was thusly obtained, each representing the firing time of any target neuron. Whereas this procedure can underestimate the firing rates of neurons (spikes will not be detected if they occur during the refractory period of another neuron), this is of minimal consequence for the purpose of localization. An example of a recording with firing times detected can be viewed in Figure 1.

From each of the S detected firing times t_s, $s \in [1, S]$, a spike waveform matrix $\mathbf{W}^s \in \mathbb{R}^{E \times L}$ was constructed by extracting the $L/4$ preceding samples and succeeding $3L/4$ samples from all electrodes of the recording.

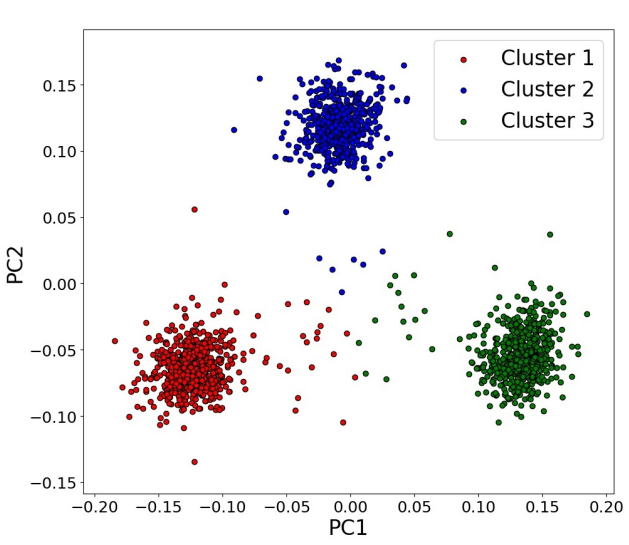

Fig. 2: An example of results from spike sorting in a case with $N = 3$ target neurons. Each plotted data point represents a single waveform extracted from a detected spike and is colored according to what neuron it was attributed to by KMeans clustering. Note that the full $E \cdot L$-dimensional waveform feature space was used as the basis for clustering; principal component analysis was used only for dimensionality reduction to allow for visualization.

D. Spike sorting

Using the sequence of spike times $t_1, ..., t_S$, each with an appertaining waveform matrix $\mathbf{W}^1, ..., \mathbf{W}^S$, a spike sorting procedure was used to determine from which of the neurons each spike originated. Specifically, KMeans clustering (SKLearn [7] implementation) applied on the full $E \cdot L$-dimensional waveform feature space was used to assign a neuronal membership $M_s \in [1, N]$ (N is the number of target neurons) to each waveform and spike time. The number of clusters was set to the number of target neurons in the geometry, and as such the method currently requires prior knowledge of the true number of target neurons. An example of the result of this clustering is shown in Figure 2.

E. Estimation of Waveform Morphology

Even when their neuronal affiliations are known, the individually extracted waveform matrices $\mathbf{W}^1, ..., \mathbf{W}^S$ make for poor proxies for the true extracellular spike waveforms (too much interference from other neurons). Instead, for the ith detected target neuron, an *archetypal waveform matrix* $\bar{\mathbf{W}}^i \in \mathbb{R}^{E \times L}$ was computed by averaging over the all waveform matrices assigned as originating from that neuron. Incidentally, $\bar{\mathbf{W}}^i$ is also the centroid of the ith cluster found during spike sorting. Under the assumption that all electrode-channels are zero-mean (enforced as per section II-B), $\bar{\mathbf{W}}^i$ is an unbiased estimate of the true, recording-generating extracellular waveform matrix $[\mathcal{W}_{T_i}(\mathbf{P}_e^1, \mathbf{P}_n^i), ..., \mathcal{W}_{T_i}(\mathbf{P}_e^E, \mathbf{P}_n^i)]^T$. An example of archetypal waveform morhologies estimated by this approach is shown in Figure 3.

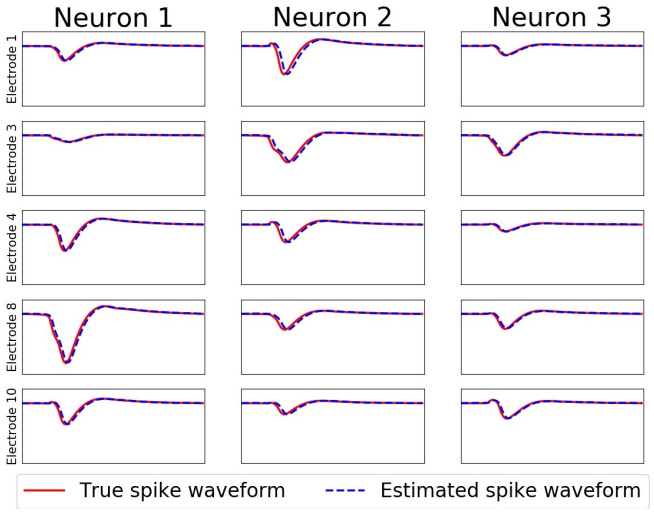

Fig. 3: An example of true (from geometry) and estimated (by spike-triggered averaging) waveforms. To generate this image, pairings between target neurons and detected clusters were made manually on the basis of waveform similarity.

F. Source Localization

Recall from section II-A that $\mathcal{W}_T(\mathbf{P}_e, \mathbf{P}_n) \in \mathbb{R}^L$ is the waveform that an electrode at position $\mathbf{P}_e = (x_e, y_e, z_e)$ would observe if a neuron at position $\mathbf{P}_n = (x_n, y_n, z_n)$ with type T fired (approximated using the model from [5]). Now, let $\tilde{\mathbf{P}} = (\tilde{x}, \tilde{y}, \tilde{z})$ denote a candidate neuron position and $\tilde{T}$ denote a candidate neuron type. I define the *loss function* $\mathcal{L}_i : \mathbb{R}^3 \times \mathbb{Z}^+ \longrightarrow \mathbb{R}^+$ associated with the ith target neuron as:

$$\mathcal{L}_i(\tilde{\mathbf{P}}, \tilde{T}) = \frac{1}{E} \sum_{j=1}^{E} \frac{\|\mathcal{W}_{\tilde{T}}(\mathbf{P}_e^j, \tilde{\mathbf{P}}) - \bar{\mathbf{W}}_{j,*}^i\|_2^2}{\|\bar{\mathbf{W}}_{j,*}^i\|_2^2}$$

where $\|\cdot\|_2$ is the Euclidean norm, $\mathbf{P}_e^j$ is the position of the jth electrode, and $\bar{\mathbf{W}}_{j,*}^i$ is the jth row of the archetypal waveform matrix pertaining to the ith detected neuron. In words, the value of $\mathcal{L}_i(\tilde{\mathbf{P}}, \tilde{T})$ represents how dissimilar the archetypal waveform matrix is from the waveform matrix that would be observed by the electrodes if the ith target neuron had been located at $\tilde{\mathbf{P}}$ and had type $\tilde{T}$.

Importantly, $\mathcal{L}_i(\mathbf{P}_n^i, T_i) = 0$, i.e. the loss is zero if the candidate position and candidate type are the true position and true type, respectively, of the ith target neuron[2]. As $\mathcal{L}_i(\tilde{\mathbf{P}}, \tilde{T}) \geq 0$, it follows that this is a global minimum of the loss function. Thus, if this minimum is unique (not true for all geometries!), it follows that the position and type of the ith neuron are the solution to the minimization problem:

$$\mathbf{P}_n^i, T_i = \underset{\tilde{\mathbf{P}}, \tilde{T}}{\operatorname{argmin}} \, \mathcal{L}_i(\tilde{\mathbf{P}}, \tilde{T})$$

For geometries where $N > 1$, this minimization problem can be considered independently for each target neuron.

[2]Assuming that the waveform morphology estimation has no errors, i.e. that $\bar{\mathbf{W}}_{j,*}^i = \mathcal{W}_{T_i}(\mathbf{P}_e^j, \mathbf{P}_n^i)$. If the waveform morphology estimation result is slightly noisy, we instead expect $\mathcal{L}_i(\mathbf{P}_n^i, T_i)$ to be small but nonzero.

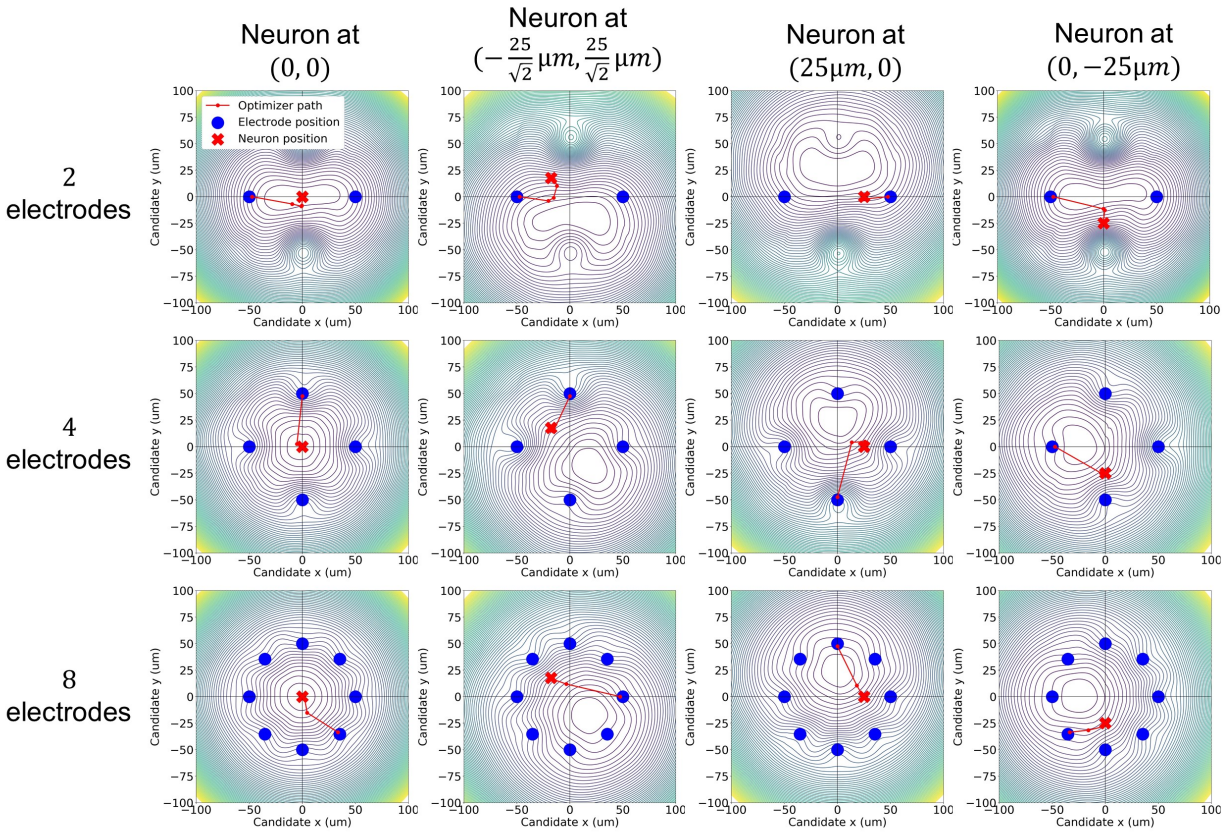

Fig. 4: Level curves of the loss function for the localization problem for different electrode setups and neuron locations. For illustration purposes, the loss function is here set as $\mathcal{L}^*_{2D}(\tilde{x}, \tilde{y}) = \mathcal{L}^*(\tilde{x}, \tilde{y}, 0)$ and it is thus not necessarily representative of the full 3-dimensional problem formulation. As can be seen from the optimizer path, conjugate gradient optimization starting from a random electrode position successfully converges on the true neuron position in all depicted cases.

For the remainder of this paper, I assume that all target neurons are of type $T = 1$. This assumption entails a significant loss of generality, but was practically necessary to keep the project within time limits. With this assumption, the localization problem for the ith neuron reduces to finding the value of $\tilde{\mathbf{P}}$ that minimizes $\mathcal{L}^*_i(\tilde{\mathbf{P}}) = \mathcal{L}_i(\tilde{\mathbf{P}}, 1)$.

Any number of optimization algorithms could be applied to approximate the candidate position that minimizes $\mathcal{L}^*_i$. In this study, the conjugate gradient method [8] (SciPy implementation [9]) was selected, as (1) $\mathcal{L}^*_i$ is differentiable w.r.t. $\tilde{\mathbf{P}}$ and (2) the success or failure this method would give an indication of whether or not the minimization problem is convex. The gradient $\mathbf{g}_i = \nabla\mathcal{L}_i \in \mathbb{R}^3$ and Hessian $\mathbf{H}_i = \nabla\mathbf{g}_i \in \mathbb{R}^{3\times3}$ were approximated[3] by the central finite difference method with $\epsilon = 10^{-8}$. The algorithm was assumed to have converged once $\|\mathbf{g}_i\|_2 < 10^{-9}$. Examples of this localization method running in two dimensions with different geometries are shown in Figure 4.

[3]As I have access to a closed-form expression of $\mathcal{W}_T$ from [5], I could compute the gradient analytically. However, by instead assuming that I have no knowledge of the gradient, the localization method can straightforwardly be used with any extracellular waveform model, even those that are more difficult to differentiate analytically.

G. Experiment 1: Cylindrical Electrode Array

The geometry of the first experiment is shown in Figure 5. 3 target neurons, all of type $T = 1$ and with firing rate $F = 10$ Hz, were placed at coordinates $\mathbf{P}^1_n = (25\mu m, 0, 0)$, $\mathbf{P}^2_n = (0, -25\mu m, 25\mu m)$, and $\mathbf{P}^3_n = (\frac{-25}{\sqrt{2}}\mu m, \frac{25}{\sqrt{2}}\mu m, -25\mu m)$. A population of 718 noise neurons was initialized as per the description in section II-B. A cylinder-shaped 3×5 electrode array centred on the origin with height $100\mu m$ and radius $50\mu m$ (5 electrodes placed equidistantly along the peripheries of 3 circles, all with radius $50\mu m$, at each of the 3 heights $z = -50\mu m$, $z = 0$, and $z = 50\mu m$) was used to simulate recordings. Spike detection, sorting, and waveform morphology estimation was carried out in accordance with sections II-C, II-D, and II-E, respectively. Once archetypal waveform matrices had been established for each neuron in this way, the iterative localization algorithm from section II-F was initialized at the coordinates of a randomly selected electrode (independently for each of the $N = 3$ neuron localization tasks) and allowed to run until convergence or until 1000 iterations had passed, whichever came first. To be able to evaluate the accuracy of estimated neuron positions, each estimated neuron position was compared to the closest true neuron position in the geometry.

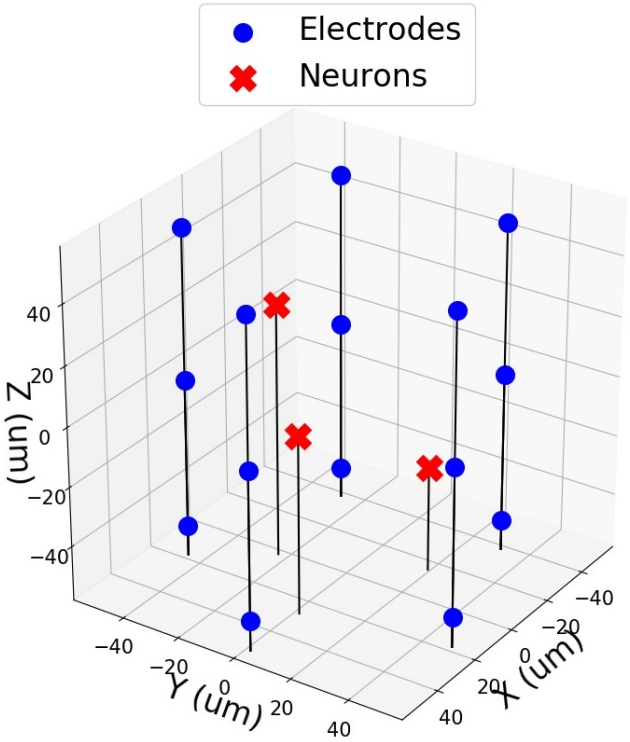

Fig. 5: Measurement geometry of the first experiment (cylindrical electrode array).

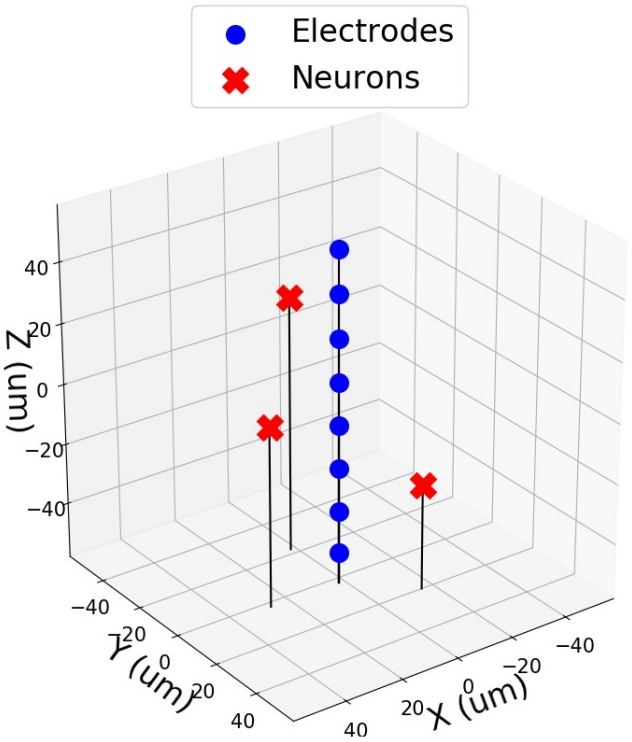

Fig. 6: Measurement geometry of the second experiment (linear electrode array).

H. Experiment 2: Linear Electrode Array

The geometry of the second experiment is shown in Figure 6. Target neurons and noise neurons were initialized identically to experiment 1. A linear electrode array of length 8 electrode, equidistantly placed along the line connecting the end points $(0, 0, -50\mu m)$ and $(0, 0, 50\mu m)$, was used to simulate recordings. Spike detection, sorting, and waveform morphology estimation was carried out in accordance with sections II-C, II-D, and II-E, respectively. Once archetypal waveform matrices had been established for each neuron, the iterative localization algorithm from section II-F was initialized at the origin $(0, 0, 0)$ and allowed to run until convergence or until 1000 iterations had passed, whichever came first. To be able to evaluate the accuracy of estimated neuron positions, each estimated neuron position was compared to the closest true neuron position in the geometry.

Due to the radial symmetry of this geometry, I hypothesized that it would be impossible to exactly locate neurons, i.e. the solution to the minimization problem of section II-F would not be unique. Specifically, in cylindrical coordinates, spike waveforms pertaining to a neuron located at (r, θ_1, z) would be highly similar[4] to spike waveforms pertaining to a neuron located at (r, θ_2, z), even if $\theta_1 \neq \theta_2$. As such, estimating neuronal θ via waveform dissimilarity minimization should be infeasible with a linear electrode array, even though r and z should be sufficiently constrained.

[4]To get identical waveforms, isotropic neurons are required. This is not true for the waveform model $\mathcal{W}_1$ of the current study. The conjecture is thus that the anisotropy is insufficient to overcome the radial symmetry.

I. Experiment 3: Sensitivity Analysis

In the work presented so far, the same waveform model $\mathcal{W}_T = \mathcal{W}_1$ (from [5]) has been used both to (1) simulate recordings, and to (2) compute the loss function that is used to evaluate the quality of candidate neuron positions. Arguably, this 'double dipping' casts doubt on the practical efficacy of the localization method: Neural recording acquired *in vivo* would likely not decompose into spike waveforms *exactly* equal to those predicted by any (tractable) computable model. Thus, an important question to ask is this: would the localization framework outlined in section II-F perform adequately if there was some small but nonzero bias separating empirically estimated archetypal waveforms from waveforms predicted by the expression for the spatial distribution of waveforms used to compute the loss function?

To investigate the robustness of the localization framework to such systematic model-reality discrepancies, a small sensitivity analysis was carried out. The same electrode setup as the one from in experiment 1 (cylindrical) was used, but with only a single target neuron located at the origin (with F=10 Hz and $T = 1$). Once archetypal waveforms had been estimated, they were corrupted with Brownian noise (intended to represent a discrepancy between modelled waveforms and empirically estimated waveforms). This procedure was repeated 11 times with noise standard deviation ranging linearly between 0 and $5\mu V$. The neuron localization problem via loss minimization was solved 100 times at each such noise level and the error of the result recorded.

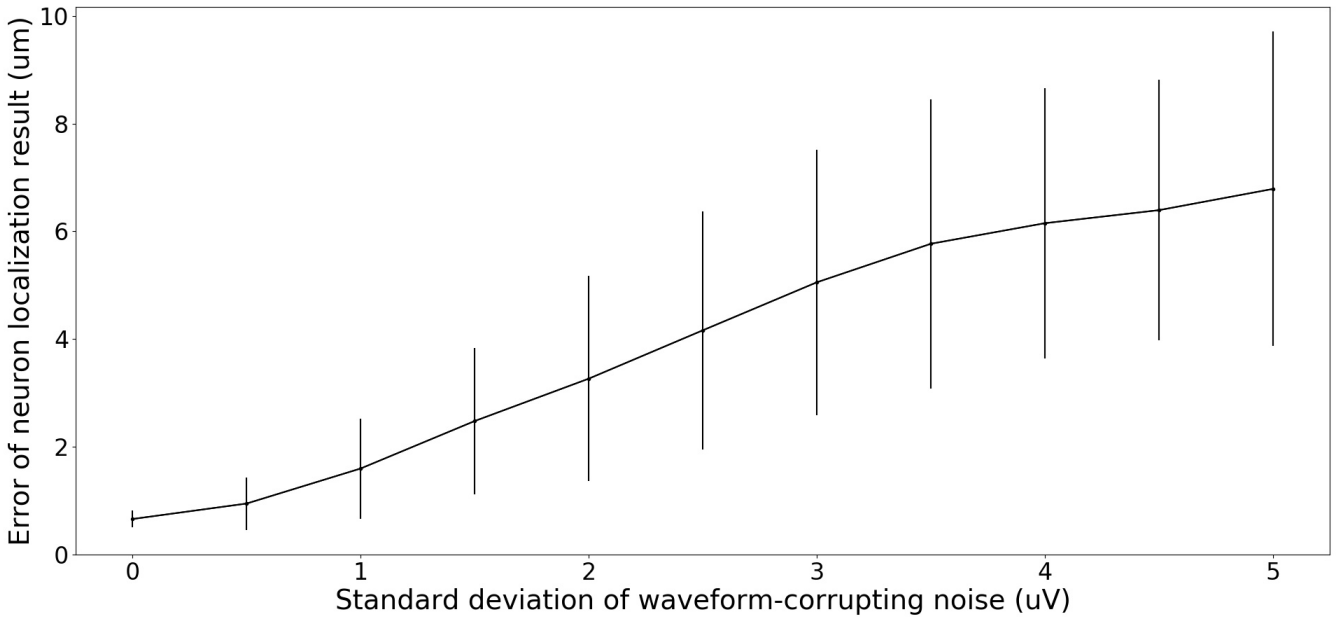

Fig. 7: Mean and standard deviation of localization error for different levels of Brownian noise added to the archetypal waveform matrices $\bar{\mathbf{W}}^1, ..., \bar{\mathbf{W}}^N$ before starting the iterative localization algorithm.

III. RESULTS

A. Experiment 1: Cylindrical Electrode Array

Estimated neuron positions from the geometry with a cylindrical electrode array (experiment 1) paired with actual positions of target neurons are presented in Table I. These results correspond to Euclidean distances between true and estimated neurons of $4.41\mu m$, $2.20\mu m$, and $2.29\mu m$. As these distances are all smaller than the diameter of a typical soma, this result arguably constitutes accurate localization.

B. Experiment 2: Linear Electrode Array

Estimated neuron positions from the geometry with a linear electrode array (experiment 2) paired with true positions of target neurons are presented in Table II. These results correspond to Euclidean distances between true and estimated neurons of $29.00\mu m$, $7.18\mu m$, and $8.98\mu m$. With errors of this size, the localization must be viewed as a failure for this geometry, as was hypothesized. However, the estimated z-coordinates and to some extend the estimated radial distances were similar to the true values of the target neurons, indicating that in this geometry, the method can estimate radial distances from the array and distances along the array, but not angular positions.

C. Experiment 3: Sensitivity Analysis

A plot if the relationship between Brownian noise variance and localization error is shown in Figure 7. Interestingly, the mean error of the localization method increases quite slowly with waveform corruption, indicating that localization can be reasonably accurate even in cases when the spatial distribution model $\mathcal{W}_T$ employed by the loss function does not exactly correspond to the latent waveform model that generated the recording.

IV. DISCUSSION

In this study, I have assumed that all target neuron are of the same type $T = 1$. This is a major limitation, as the spatial distribution of spike waveforms surrounding a neuron can differ radically based on neuron morphology, such as the direction of the axon. The most pressing next step in developing this framework would thus be to incorporate some search over the space of candidate neuron types in order to find the best fit before starting the gradient-based localization step. Furthermore, the number of target neurons N to be located should ideally not have to be provided *ex-ante*, but instead be estimated by processing the recording. Fortunately, methods to estimate the number of latent target neurons from a recording already exist [3].

On a similar note, I have assumed exact knowledge of all electrode positions. This is clearly not a realistic assumption for many types of recordings. An interesting research question could thus be to expand the framework introduced here to simultaneously estimate electrode and neuron locations. Given that the positions of at least some electrodes and/or neurons can be provided to the algorithm (in order to counteract complete rotational and translational invariance), I conjecture that expanding the framework introduced in this paper could be as simple as expanding the search space for candidate positions from $\mathbb{R}^3$ to $\mathbb{R}^{3I}$ in the case of I unknown electrode and/or neuron positions.

Whereas the results of the sensitivity analysis are at least encouraging, simulating model error as Brownian noise is clearly insufficient for establishing validity. To continue this line of inquiry, one would need more theoretically sound way to perturb archetypal waveforms in order to convincingly model the discrepancy between real extracellular spike waveforms and those generated by the simplified model.

TABLE I: Localization results for experiment 1 (cylindrical electrode array).

Target neuron #	True position $[\mu m]$	Estimated position $[\mu m]$	Euclidean error distance $[\mu m]$
1	$(25, 0, 0)$	$(24.20, -0.87, -4.24)$	4.41
2	$(0, -25, 25)$	$(0.70, -23.54, 26.50)$	2.20
3	$(\frac{-25}{\sqrt{2}}, \frac{25}{\sqrt{2}}, -25)$ $\approx (-17.68, 17.68, -25)$	$(-17.00, 15.96, -26.35)$	2.29

TABLE II: Localization results for experiment 2 (linear electrode array).

Target Neuron #	True position $[\mu m]$	Estimated position $[\mu m]$	Euclidean error distance $[\mu m]$
1	$(25, 0, 0)$	$(13.42, -26.13, -4.75)$	29.00
2	$(0, -25, 25)$	$(5.59, -26.85, 29.11)$	7.18
3	$(\frac{-25}{\sqrt{2}}, \frac{25}{\sqrt{2}}, -25)$ $\approx (-17.68, 17.68, -25)$	$(-12.14, 24.56, -26.63)$	8.98

Unavoidably, the localization framework investigated in the current paper would need to be validated on real (i.e. not simulated) neural recordings with known geometry in order to be of any practical value whatsoever.

An interesting question (and one towards which no effort at all was directed in this study) is for what configurations of waveform models and recording geometries localization is theoretically possible. For instance, localization would clearly not be possible, irrespective of electrode geometry, if the spatial distribution of waveforms is uniform, i.e. if the observed waveform is independent of the relative position of the neuron and the electrode. On the other hand, for a waveform distribution were the waveform amplitude is inversely proportional to distance separating the electrode form the neuron, $3-4$ electrodes would typically be sufficient as simple triangulation could be used.

REFERENCES

[1] G. Buzsáki, "Large-scale recording of neuronal ensembles.," *Nat. Neurosci.* 2004, vol. 7, no. 5, pp. 446-51.

[2] M. S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials.," *Network* 1998, vol. 9, no. 4, pp. 53-78.

[3] P. T. Thorbergsson, M. Garwicz, J. Schouenborg, and A. J. Johansson, "Spike-Feature Based Estimation of Electrode Position in Extracellular Neural Recordings," in *Int. Conf. IEEE EMBC* 2012.

[4] T. Kubo, N. Katayama, A. Karashima, and M. Nakao. (2008). "The 3D position estimation of neurons in the hippocampus based on the multi-site multi-unit recordings with silicon tetrodes," in *Int. Conf. IEEE EMBC* 2008.

[5] P. T. Thorbergsson, M. Garwicz, J. Schouenborg, and A. J. Johansson, "Computationally efficient simulation of extracellular recordings with multielectrode arrays," *J. Neurosci. Methods*, 2012, vol 211, issue 1, pp. 133-44.

[6] D. Heeger, "Poisson model for spike generation," 2000, available from cns.nyu.edu/ david/handouts/poisson.pdf

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al.*, "Scikit-learn: machine learning in Python," *J. Mach. Learn. Res.*, 2011, vol. 12, pp. 2825–30.

[8] J Nocedal and S. Wright, "Numerical Optimization," 1999, pp. 120-2.

[9] Documentation available at docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin_cg.html

V. REFLECTION ON ETHICS

In this project I have attempted, with some success, to locate individual neurons by processing simulated extracellular neural recordings. All parts of the study were conducted entirely *in silico* and did not involve neither moral agents nor moral subjects—at least not in a contemporary understanding of these terms. Thus, if one ignores questions of opportunity cost (counterfactually, how I could have spent my time and resources if I had refrained from doing this project?), I find it difficult to see how any of the standard normative theories (consequentialism, deontology, virtue ethics, care ethics, etc.) could ascribe any moral significance at all to my current project in and of itself. To find anything of interest for a discussion on ethics related to the project, I will move to consider ethical concerns related to continued empirical work on neuron localization.

To validate methods for neuron localization (such as the algorithmic framework presented in this paper) *in vivo*, living subjects (nonhuman animals or people) would doubtlessly have to be involved. As with all experiments involving agents capable of suffering, issues related to the 4 principles prescribed by the Georgetown mantra—autonomy, justice, beneficence, and non-maleficence—would arise. As such experiments would concern the central nervous system—the proverbial seat of the soul—these issues would arguably become significantly more poignant. All risks, such as those of violating autonomy or causing bodily or psychological harm, would have to be weighted against any and all potential short- and long-term benefits—medical, scientific, or related to the flourishing of moral subjects more generally—of the technology. If one was to opt for validating methods on non-human animals, a related point could certainly be made that, in accordance with the three Rs of animal research, care should be taken to (1) conduct experiment in ways that minimize stress and pain, and (2) maximize the information gained per unit of unavoidable suffering.

In conclusion, I see no *unique* ethical risks associated with continued work on neuron localization; all risk seem to be held in common with every other endeavor that requires invasive neural engineering. However, of arguably equal importance is the fact that, for the foreseeable future, there are no obvious ethical gain to be had from continued

work, i.e. no immediate alleviation of suffering would be enabled by having better neuron localization capabilities. The bottom line of any near-term cost-benefit analysis would thus depend on the intended application, with some caution being warranted due to the relatively limited ethical upside. Most of both the possible benefits and the risks of the technology instead seem to be concentrated in the relatively distant future, when such knowledge (in conjunction with other technologies) could enable precise influencing of the central nervous system and, by extension, of the mind. It is difficult to ascertain whether the ability to exogenously induce arbitrary qualia would constitute a net ethical benefit for humanity.