2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Automatic Channel Selection in Neural Microprobes: A Combinatorial Multi-Armed Bandit Approach

Camilo Gordillo Barbara Frank Istvan Ulbert Oliver Paul Patrick Ruther Wolfram Burgard

***Abstract*— State-of-the-art neural microprobes contain hundreds of electrodes within a single shaft. Due to hardware and wiring restrictions, it is usually only possible to measure a small subset of the available electrodes simultaneously. The selection of the best channels is typically performed offline either manually or automatically. However, having a fixed selection for long-term observation does not allow the system to react to changes in the neural activity, and may therefore lead to the loss of important information. In this paper, we formulate the process of autonomously selecting the best subset of electrodes as a combinatorial multi-armed bandit problem with non-stationary rewards, thus allowing the probe to adapt its selection policies online. In order to minimize exploratory actions of the probe, we furthermore take advantage of the existing dependencies between neighboring channels. Our approach is an adaptation of the discounted upper confidence bounds (D-UCB) algorithm, and identifies the electrodes providing the largest amount of non-redundant information. To the best of our knowledge, this is the first online approach for the problem of electrode selection. In extensive experiments, we demonstrate that our solution is not only able to converge towards an average optimal selection policy, but it is also able to react to changes in the neural activity or to damages of the recording electrodes.**

I. Introduction

There is a growing interest in the development of high-resolution neural microprobes capable of recording the activity of single neurons. This capacity allows for further investigation of the complex interactions between brain regions, and may lead to a better understanding of the neural processes occurring in the brain. In the future, intelligent microprobes are expected to become a key component in the field of brain-computer interfaces (BCI), thus inheriting a wide spectrum of applications. Envisaged applications include monitoring and reacting to diseases such as epilepsy or depression, for instance, by detecting pathological patterns in brain activity and by providing appropriate electrical stimulation to specific brain regions. Such a desired smart behavior requires the probe to interpret the measured signals and to select actions without the intervention of human experts.

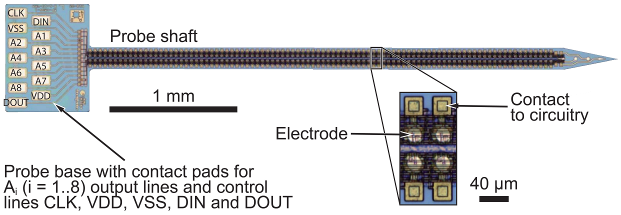
CLK
DIN
VSS
A1
A2
A3
A4
A5
A6
A7
A8
VDD
DOUT
Probe shaft
1 mm
Probe base with contact pads for A_i (i = 1..8) output lines and control lines CLK, VDD, VSS, DIN and DOUT
Electrode
Contact to circuitry
40 µm

Fig. 1: Neural microprobe with 188 electrodes distributed along the shaft.

With state-of-the-art integrated circuits technology, it is possible to integrate hundreds of electrodes along a single probe and to electronically switch the available output channels to read out a subset of electrodes. Figure 1 shows the optical micrograph of such a probe [1]. The contact pads for reading out measurements on the probe base are used to interface the probe in order to record electrophysiological signals from selected electrodes on the probe shaft and to control the integrated switch matrix. Neves *et al.* [2] originally proposed this concept of an electronic depth control (EDC) probe. Typically, an experimenter scans the shaft for neural activity at the individual recording sites and selects the most informative ones for long-term recording. As the number of available electrodes increases, however, the task of manually selecting an optimal subset of electrodes becomes impracticable and inefficient. Therefore, we addressed the problem of autonomously selecting informative electrodes in our previous work [4]. While this approach is able to provide near-optimal selections with a greedy approach, it relies on information collected in a pre-processing step.

In this paper, we present a novel approach to autonomously learn and update electrode selection policies during long-term observation, which is desirable in a complex and dynamic environment such as the human brain. We model this problem within the multi-armed bandit (MAB) framework. In this way, our approach is able to explore the environment and learn optimal selection policies online. Furthermore, we extend the selection strategy by taking into account dependencies between neighboring electrodes. Thereby, we avoid recording of redundant information and increase exploitation of acquired knowledge. By monitoring changes in the reward distribution, we are able to react to changes in the environment and to adapt the selection policies. We validate our approach in extensive experiments on real neural data.

II. Related work

There are just a few studies in the area of automatic channel selection for neural microprobes. Seidl *et al.* [5]

This work was supported by BrainLinks-BrainTools, Cluster of Excellence funded by the German Research Foundation (DFG, grant number EXC 1086). The probe development was supported by the Information Society Technologies Integrated Project NeuroProbes of the 6th Framework Program (FP6) of the European Commission under Project IST-027017.

All authors except I. Ulbert are with the University of Freiburg, Faculty of Engineering, 79110 Freiburg, DE.

I. Ulbert is with the Inst. of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Science, Budapest, Hungary and the Faculty of Information Technology and Bionics, Péter Pázmány Catholic University, Budapest, HU.

978-1-5090-3762-9/16/$31.00 ©2016 IEEE

propose a semi-supervised approach which computes the signal-to-noise ratio (SNR) of the recorded electrodes in order to assist the experimenter in the selection process. Similarly, Van Dijck *et al.* [6] present an automatic approach which employs the SNR as a quality measure for each electrode, and additionally penalizes each channel according to its similarity with respect to the already selected ones. Van Dijck *et al.* [6] also evaluate different strategies for measuring the similarity between the recorded signals based on the spike trains detected on each channel. Applying the penalization avoids the recording of redundant information and leads to a better distribution of the recording channels in a simulated neuronal model. In our previous work [4], we proposed a selection strategy based on nonparametric sparse Gaussian processes for predicting neural activity across neighboring channels. In this approach, we use the signal prediction capabilities of each recorded channel in order to find the subset of electrodes that minimizes the overall prediction error and thus, maximizes the amount of collected information. In contrast to these selection strategies, which find a subset of optimal electrodes for long-term recordings offline, the approach introduced in this paper operates online and updates its selection policies according to the recorded information. In this way, it is able to react to changes in the environment.

Our approach is based on the widely studied multi-armed bandit (MAB) framework [7], in which the received rewards may change over time corresponding to the non-stationary nature of the recorded signals. The non-stationary MAB problem has received considerable attention and different solutions have been proposed in the last decades. Thierens [8] proposes a modification of the standard pursuit approach called Adaptive Pursuit. This method outperforms probability matching strategies when dealing with non-stationary environments. Exploiting the success and mathematical background of the upper confidence bounds (UCB) policies [9], Kocsis and Szepesvári [10] and Garivier and Moulines [11] have implemented some adaptations to the standard algorithm that solve the non-stationary problem: Discounted UCB (D-UCB) and sliding window UCB (SW-UCB), respectively. The D-UCB strategy relies on a discount factor which assigns higher weights to recent rewards, while SW-UCB makes use of a fixed-size horizon, thus forgetting old rewards. Hartland *et al.* [12] present an extension of the UCB policies named *Adapt-EvE*, which features an adaptive change point detection test devised for controlling the ratio between exploration and exploitation in abruptly changing environments.

III. Learning channel selection policies

The task of our smart neural probe system is to select a subset of electrodes for recording neural activity: from a set of K available electrodes, it has to choose a subset $m < K$ in order to maximize the amount of recorded information. We frame this as a multi-armed bandit problem [7] in which the agent repeatedly chooses between different possible actions and obtains a reward after each trial. In the stationary case, these rewards come from an initially unknown but constant probability distribution. The goal of the agent is to maximize the expected cumulative reward over time by identifying optimal selection policies. Thus, it needs to explore the utility of the possible actions while also exploiting the already acquired information.

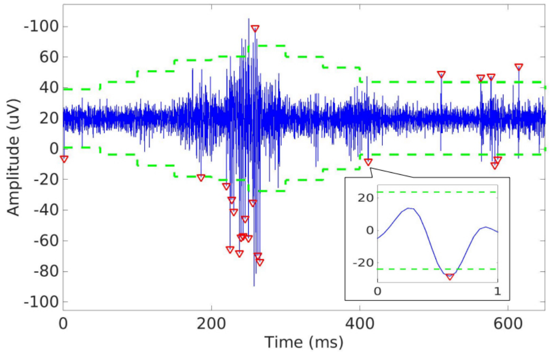

Fig. 2: Neural signal measured by our probe: we consider spikes which are at least 5 times larger than the local noise variance (dashed line) during the computation of the reward. An extracted spike is shown at the bottom.

In our channel selection problem, the agent can choose more than one action, that is multiple electrodes at a time. Following the terminology proposed by Chen *et al.* [13], we define a *super arm* as a group of electrodes which are selected and recorded simultaneously in each trial. Furthermore, the reward distribution might change over time, that is the signals measured on individual channels may change due to varying neural activity, tissue adaptation or relative probe movements. This requires the algorithm to adapt its selection policies. In the following, we detail the computation of the rewards, the identification of correlations and dependencies between electrodes and the detection of changes in the environment.

A. Computing the reward of an electrode

The reward obtained after recording a set of electrodes depends on the quality of the recorded signals. Because of their practical importance in the field of neuroscience [4,6,14,15], we define quality in terms of the number and the amplitude of the neural spikes detected on each recorded electrode. Fig. 2 shows a segment of one of the recorded signals along with the spikes detected with an automatic amplitude threshold [5]. For each detected spike, we extract a 1 ms time window around its peak. Thereby, we compute the reward for each individual recorded electrode i as

$$R_i = \sum_{j=1}^{J} \mathrm{RMS}(\mathrm{Spike}_j), \quad (1)$$

where J is the number of spikes extracted from electrode i, and RMS computes the root mean square value of each spike. The total reward received at a given play would be equivalent to the sum of the individual rewards R_i.

B. The D-UCB selection strategy

In order to learn and adapt electrode selection policies, we build on the D-UCB algorithm proposed by Kocsis and Szepesvári [10]. It is a variant of the UCB policies designed to deal with non-stationary environments by adapting its selection policies according to the most recent observations. In each round or trial T, the algorithm selects a set of electrodes i which maximize their upper confidence bound $I_T(i)$

$$I_T(i) = \overline{\mu}_T(i) + c_T(i). \tag{2}$$

The upper confidence bound considers an estimate of the expected reward $\overline{\mu}_T(i)$ and a measure of the uncertainty of our observations given by a so called padding function $c_T(i)$. Considering the uncertainty ensures exploration of unobserved or presumably less rewarding regions. Due to the combinatorial nature of our problem [13], the algorithm not only selects one channel but a subset *Sel* composed of the top m electrodes with the highest expected rewards.

For each of the K available electrodes i, D-UCB maintains an estimate of these quantities: it updates the discounted average $\overline{\mu}_T(i)$ at the end of each play T according to

$$\overline{\mu}_T(i) = \frac{\sum_{t=1}^{T} w_t R_i^t}{\sum_{t=1}^{T} w_t} \qquad \forall i \in Sel, \tag{3}$$

where the weights w_t assigned to each reward are equal to

$$w_t = \frac{1}{\gamma^{t-1}}, \tag{4}$$

and the discount factor $\gamma < 1$ ensures that recent observations receive higher weights. Decreasing γ makes the algorithm ”forget” previous observations faster, but may also prevent convergence to strong and valid estimates.

The padding function $c_T(i)$ increases the likelihood of selecting electrodes which have not been recorded for a long period of time

$$c_T(i) = \sqrt{\frac{\varepsilon \ln n_T}{N_T(i)}} \qquad \forall i \in K, \tag{5}$$

with ε controlling the trade-off between *exploration* and *exploitation*: high values favoring exploration and small values encouraging exploitation. The value of $N_T(i)$ is updated as

$$N_T(i) = \begin{cases} \gamma(N_{T-1}(i)+1), & \text{if } i \in Sel \\ \gamma N_{T-1}(i), & \text{otherwise} \end{cases} \qquad \forall i \in K. \tag{6}$$

When electrode i is selected, that is when $i \in Sel$, the discount factor γ limits the increment of $N_T(i)$. On the contrary, when electrode i is not selected, scaling down $N_T(i)$ increases the uncertainty of the corresponding electrode. The value of n_T is usually defined as the total number of plays so far, and we compute it as

$$n_T = \sum_{i=1}^{K} N_T(i). \tag{7}$$

This selection strategy is easy to implement and, as our experiments show, ensures a rapid exploration and convergence to reasonable reward estimates. Furthermore, it is able to quickly adapt to changes. In environments with small infrequent changes such as the human brain, however, exploitation is suboptimal due to the required constant exploration. Therefore, we describe in the following how to exploit dependencies between electrodes in order to reduce exploration time and improve exploitation of information.

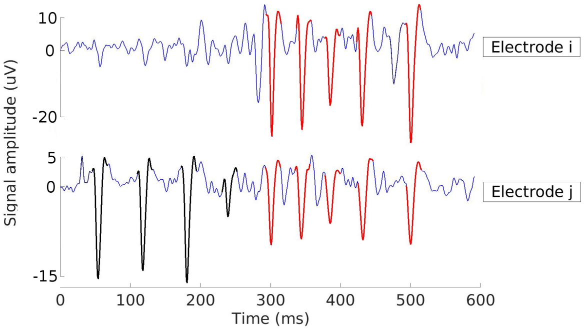

Fig. 3: Dependencies between neighboring electrodes: the spikes observed in both electrodes are shown in red and the spikes detected only in one of the electrodes are shown in black.

IV. Modeling dependencies between electrodes

Since the electrodes are closely spaced along the probe [1], neighboring electrodes may record similar neural activity. As an example, Fig. 3 shows a portion of the neural activity recorded simultaneously from two close-by electrodes. In this case, the first electrode measures redundant information and it would be sufficient to record the second electrode. Following the definition of the reward stated in Eq. (1), we define the dependency of electrode i from electrode j as the percentage of the reward from i which we simultaneously obtain from electrode j.

The following table shows an estimate of the dependency matrix Σ for the electrodes presented in Fig. 3.

	i	j
i	100 %	100 %
j	55 %	100 %

When recording i, we simultaneously obtain around 55 % of the reward expected from j. When we record j, in contrast, we simultaneously obtain 100 % of the information expected from i. By learning and keeping track of these dependencies between electrodes, our aim is to reduce exploration time and to encourage exploitation of highly informative electrodes.

A. Initializing the dependency matrix

At the beginning of the experiment, we sequentially scan the microprobe to obtain an initial estimate of the expected rewards. Similar to updating the estimates about the expected rewards in Eq. (3), we continuously update the dependency matrix by means of a weighted average. Here, we also assign higher weights to recent observations. We compute both the mean Σ_μ and the standard deviation Σ_σ of the observed dependencies.

The length of the initialization period depends on the number of available electrodes and the selection size. As the

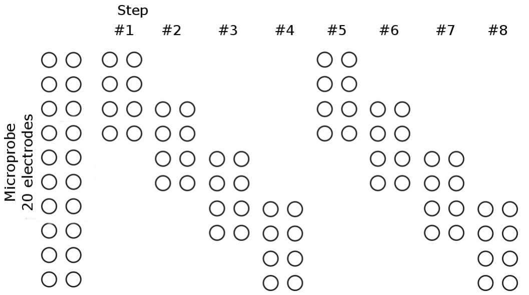

Fig. 4: Initialization sequence: we simultaneously record neighboring electrodes to learn their dependencies. In this example, we can select 8 out of 20 electrodes. Thus, scanning the entire probe completes after 4 plays.

distance between the electrodes increases, we expect their dependency to decrease. Potentials of single neurons, for instance, can only be measured at sufficiently high SNR for distances below 50 µm [16]. For this reason, it is not necessary to initialize and update the entire matrix. Instead, we only consider the dependencies between neighboring electrodes. In general, we allow for four to five complete scans of the microprobe. Fig. 4 shows the scanning sequence for initializing the algorithm.

B. Detecting changes

Knowing the dependencies between electrodes reduces the time needed for exploration of the probe. Similar to the estimates of the expected rewards, however, the dependencies between the electrodes may change over time due to different neural activity or due to undesired displacements of the microprobe. Thus, we need to ensure that the matrix Σ is up-to-date. In principle, we could achieve this by scheduling periodic re-initializing scans, as presented in Fig. 4. Without knowing the rate of change of the environment, however, it is suboptimal to periodically stop exploiting in order to update the estimates. In contrast, it is more efficient to scan specific regions of the probe only when we believe that a change has occurred. To detect such changes in our rewards distribution, we use the change-point-detection algorithm of Hartland *et al.* [12]. The key idea is to monitor, using the Page-Hinkley (PH) statistical test [17], whether or not the series of rewards gathered during the last steps belong to a single statistical distribution (null hypothesis) or not (change-point-detection). In detail, the PH test considers a cumulative variable m_T^i defined as the difference between the observed reward values $\overline{\mu}_1(i), \dots, \overline{\mu}_T(i)$ from Eq. (3) and their mean $\overline{X}_i$ at electrode i up to the current moment T

$$m_{T_i} = \sum_{t=1}^{T} (\overline{\mu}_t(i) - \overline{X}_i - \delta_i), \qquad (8)$$

where δ_i corresponds to the magnitude of change that should not raise an alarm. We furthermore compute the minimum value of m_T: $M_T = \min(m_t)$ for $t = 1, \dots, T$, and monitor the difference between M_T and m_T. When the difference is greater than a given threshold λ, we reject the null hypothesis and trigger a change detection alarm. Table I summarizes the procedure. The PH test involves two parameters: we set δ equal to two times the standard deviation of our individual estimates, thus only considering changes which lie outside a 95 % confidence level around our current estimate; the second parameter λ depends on the desired false detection rate. While Hartland *et al.* [12] propose an adaptive control for autonomously adjusting λ, its implementation lies outside the scope of this work and we set it manually in this paper.

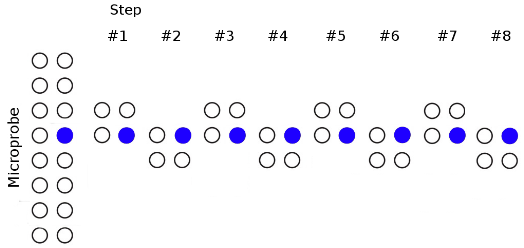

Fig. 5: Scanning sequence after detecting a change in the recorded neural activity in the highlighted electrode: we use a portion of the available channels to scan the region of interest and to update the dependency matrix.

Whenever we observe a change in a specific area of the microprobe, we update our estimates about rewards and dependencies in that region. To achieve this, we use a portion of the available channels to scan the electrodes around the region of interest in which the change occurred, as illustrated in Fig. 5. Before scanning, we reset the estimates regarding dependency and expected reward.

V. Exploiting dependencies for electrode selection

So far, we described how we model and keep track of dependencies between electrodes by means of the dependency matrix Σ. In the following, we use this information to adapt the D-UCB selection strategy (Sec. III-B) in order to improve the selection of informative electrodes.

A. Avoiding redundant information

In each play, our algorithm selects electrodes in a greedy and sequential way. It first sorts the electrodes according to their upper confidence bound I_t (Eq. (2)), and iteratively selects the one with the highest estimate. To avoid the recording of redundant information, we discourage the simultaneous selection of strongly dependent electrodes. To this end, we penalize, after each selection, all remaining reward estimates

TABLE I: The PH statistical test returns true if it detects a change on one of the available electrodes i.

$\overline{X_i} = \frac{1}{T}\sum_{t=1}^{T} \overline{\mu}_t(i)$
$m_{T_i} = \sum_{t=1}^{T}(\overline{\mu}_t(i) - \overline{X_i} - \delta_i)$
$M_{T_i} = \max(m_{t_i}), \quad t = 1, \dots, T$
$PHT_i = M_{T_i} - m_{T_i}$
$\lambda_i = 0.4\overline{X_i}$
Return$(PHT_i > \lambda_i)$

in proportion to their maximum dependency with any of the already selected electrodes

$$\overline{\mu}_T(i) = \overline{\mu}_T(i) \left(1 - \max_{s \in Sel} \; (\Sigma_\mu(i,s) - \Sigma_\sigma(i,s))\right), \quad (9)$$

Due to the negative effect that wrong penalties could have on the selection process, we aim for a conservative penalization by subtracting the standard deviation Σ_σ from our dependency estimates Σ_μ. We repeat this process until *m* electrodes have been selected.

B. Avoiding unnecessary exploration

When we record an electrode and do not observe significant changes in our estimates, we have no reason to believe that any of the non-recorded and strongly dependent neighbors have changed. Thus, in order to avoid exploration of such likely non-informative regions, we avoid increasing the uncertainty of those non-recorded electrodes.

As explained in the beginning of Section III-B, the uncertainty of each electrode i can be controlled by means of the padding function $c_T(i)$, which in turn is a function of $N_T(i)$. Decreasing $N_T(i)$ in Eq. (6) increases the uncertainty of those electrodes which were not selected. This step in the standard D-UCB algorithm ensures permanent exploration independently of the rewards of the non-optimal electrodes. In contrast, we dynamically adapt the discount factor by modifying Eq. (6) as follows

$$N_T(i) = \begin{cases} \gamma(N_{T-1}(i)+1), & \text{if } i \in Sel \\ \eta(i) N_{T-1}(i), & \text{otherwise} \end{cases} \qquad \forall i \in K, \quad (10)$$

where our new discount factor $\eta(i)$ is a function of the estimated dependencies to the selected electrodes

$$\eta(i) = \gamma + \max_{s \in Sel} \; (\Sigma_\mu(i,s) - \Sigma_\sigma(i,s)) * \left(\frac{1-\gamma}{100}\right). \quad (11)$$

In other words, $\eta(i) \in [\gamma, 1]$ is a linear function of the dependency between electrode i and the current selection set *Sel* which takes a maximum value of 1 when the electrodes are 100% dependent. Thus, the uncertainty of electrode i increases as the dependency of the non-observed electrode i to the current selection decreases. When electrode i is independent of our current selection, its uncertainty increases as in the D-UCB algorithm, which increases its likelihood for selection in the future.

As we will show in the experimental evaluation, exploiting the dependencies across electrodes in this way improves the long-run performance of our selection strategy. Furthermore, since we detect changes and update our estimates, the strategy is still capable of reacting to changes in the environment.

VI. Data acquisition

In order to evaluate the performance of the described algorithm, we used a dataset of neural activity recorded in vivo from the neocortex of Wistar rats. The data was collected at the Institute for Psychology of the Hungarian Academy of Sciences following the respective animal care regulations [18]. Four different microprobes with 188 electrodes each were placed in different brain regions and measurements were recorded in highly active regions of interest on each probe. The recorded extracellular potentials were sampled at 20 kHz and bandpass filtered in the range of 300 Hz to 3 kHz. Since only eight electrodes can be recorded simultaneously, the region of interest consisting of 20 electrodes was observed in sequential scans of overlapping regions similar to the protocol shown in Fig. 4. This scanning protocol was designed to collect data for our offline selection procedure [4] in order to maximize information about neighboring electrodes.

VII. Results

With the datasets described above, we simulate the problem of online electrode selection and build an artificial dataset by combining the recordings of 72 observed channels. In our experiments, we evaluate the performance of the algorithm to learn and adapt selection policies for the simultaneous observation of eight electrodes. We compare our approach that takes into account dependencies between electrodes to a standard D-UCB selection strategy and to an offline greedy selection strategy that keeps the selection fixed during the entire recording session.

The performance in the MAB framework is usually defined in terms of the *regret*. The regret is computed as the difference between the received reward and the reward of an optimal policy. In this paper, we define the optimal policy as the subset of electrodes which, on average, would receive the highest *cumulative reward* while avoiding the recording of redundant information. In the following experiments, we experimentally set the duration of a play to a value of 0.5 s. Increasing this value makes the algorithm converge too slowly, while decreasing it does not allow the algorithm to collect enough information during the recording period. Additionally, we set the values of γ and ε to 0.99 and 0.05, respectively. These values were chosen by tuning the performance of the standard D-UCB algorithm to the point where the system quickly converges to a near-optimal selection policy while still reacting effectively to changes in the environment. It is important to note that if we compute the reward as stated in Eq. (1), the standard D-UCB algorithm has no means of avoiding redundant information. To address this issue, we apply a penalization such that only the largest of any simultaneous spikes contributes to the overall received reward in a play. Since our approach learns the dependencies between the electrodes, we apply this penalization only to the reward estimates of the D-UCB algorithm. When comparing the performance of both algorithms, however, we compute the regrets after penalizing the rewards.

The main advantage of our online selection strategy is the capability to adapt its selection policies to recent neural activity. For this reason, we evaluate how our approach deals with simulated changes in the environment. We consider two types of changes: first, a displacement of the neural probe, for instance, due to tissue relaxation, and second, the activation or deactivation of specific neural areas, and thus changes in the signals of individual electrodes. Whenever we

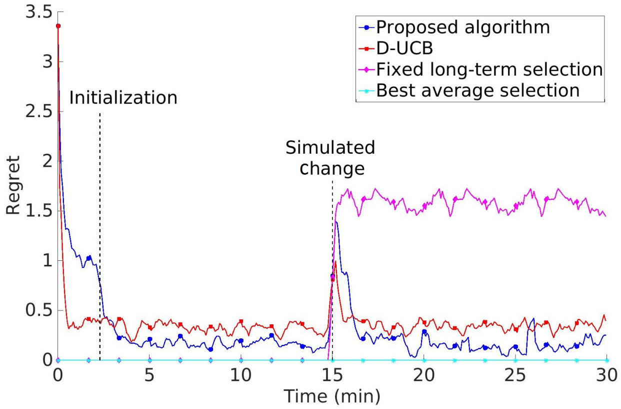

Fig. 6: Regret of different selection policies when simulating a three-row downward displacement of the neural microprobe after 15 minutes. The data was smoothed by averaging 50 consecutive plays.

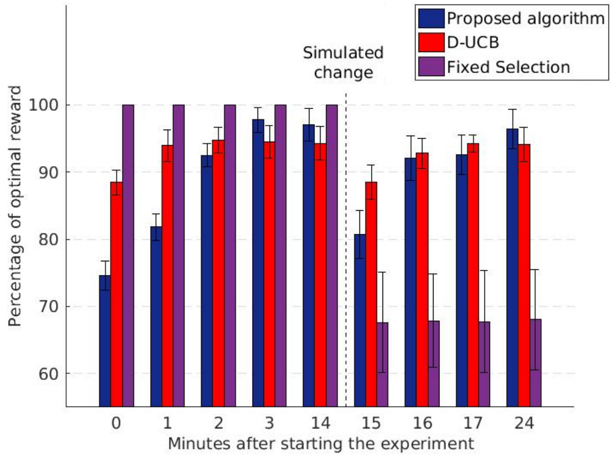

Fig. 7: Comparison of different selection policies in terms of the average percentage of the optimal reward over time for ten different datasets. In this experiment, we simulated an eight-row downward shift after 15 minutes.

artificially introduce these types of changes, we determine a new optimal policy.

In the first experiment, we simulate a three-row downward displacement of the neural probe by shifting the signals in the available dataset after 15 minutes. Fig. 6 shows the performance of both algorithms and a fixed long-term selection for this experiment. In the first part of the experiment, before the change, both algorithms converge towards the average optimal selection. Due to permanent exploration, neither of them achieves optimal performance. By taking advantage of the strong dependencies between electrodes, however, our approach achieves higher rewards than the D-UCB strategy. Since the fixed selection has no means of adapting its selection, its performance is only optimal as long as the neural activity does not change and decreases after the simulated change. The standard D-UCB algorithm, in contrast, adapts its selection policies according to the most recent recorded activity, thus reacting quickly to the simulated change. Similarly, our algorithm is able to detect

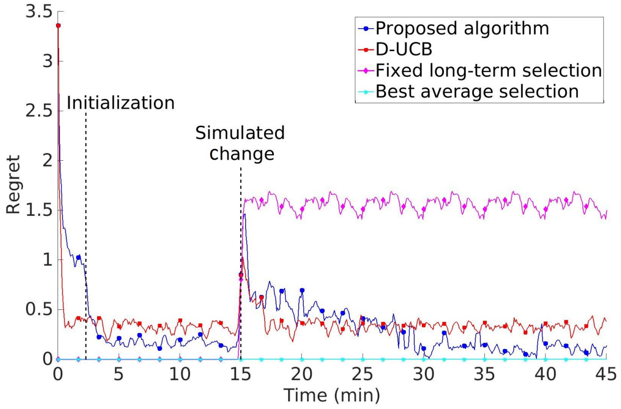

Fig. 8: Regret of the different selection policies when simulating a change in half of the optimal electrodes. The data was smoothed by averaging 50 consecutive plays.

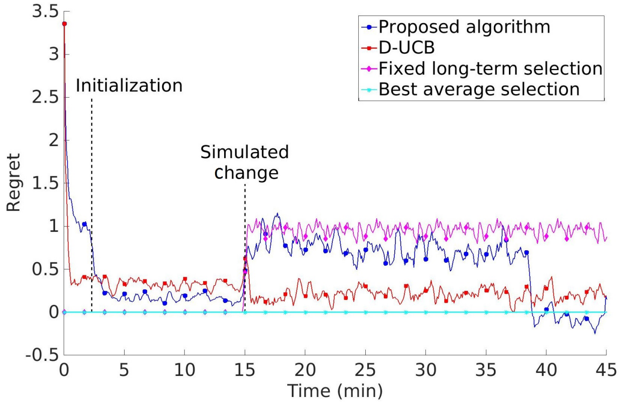

Fig. 9: Simulating a significant change in two non-optimal electrodes. Although our approach eventually detects the introduced change, the outdated dependency matrix misleads the exploration process and avoids the frequent recording of the new optimal electrodes. The D-UCB algorithm, in contrast, quickly reacts to the simulated change.

the change in the distribution and updates both its reward estimates and the dependency matrix. It requires a longer time to adapt to the change but it outperforms the standard D-UCB algorithm after convergence of the reward estimation. We also carried out a statistical evaluation for a simulated 8-row downward displacement of the probe in ten different data sets. Fig. 7 summarizes the results in terms of the fraction of the optimal reward achieved at different points in time. Again, our approach requires more time to converge to an accurate reward estimate but achieves a higher reward after convergence compared to the standard D-UCB strategy.

Another kind of change may occur due to the activation or deactivation of specific neurons. Fig. 8 shows the performance of the different policies when we manipulate the signals from half of the optimal electrodes in order to appear in different locations. Non-optimal recordings are then mapped to the previously optimal channels. The task of the algorithm is to identify that a change has occurred and to adapt its selection policies to the new optimal electrodes. In this experiment, although our approach quickly identifies the change, thus outperforming the fixed selection policy, it requires almost 15 minutes to identify the new optimal electrodes. In contrast, the D-UCB policy only requires two minutes to adapt. Once the reward estimates are updated, however, our algorithm outperforms the D-UCB policy.

One of the worst-case scenarios in a non-stationary MAB problem is a change in one or more of the non-optimal arms such that they suddenly become optimal. Since these arms are selected less frequently, their estimates get updated much slower. Fig. 9 shows the performance of the proposed algorithm when we manipulate only two of the non-optimal electrodes such that their rewards suddenly increase. Our proposed approach requires almost 25 minutes to detect the change and adjust the selection policy. This lack of adaptability is due to an outdated dependency matrix. Although the algorithm eventually explores each of the available electrodes, the dependency estimates acquired before the change mislead the exploration process. It is important to note, however, that in a real experiment and due to the proximity between electrodes, not only one but several neighboring channels would provide evidence of any significant change in neural activity. Thus, our approach should be able to identify changes much faster and to adjust its selection policies accordingly.

VIII. Conclusion

In this paper, we presented an approach to autonomous channel selection for neural microprobes using the multi-armed bandit framework. By adapting a discounted upper confidence selection strategy, we are able to deal with the non-stationary nature of neural signals and to learn electrode selection policies in an online manner. Since we exploit dependencies between close-by electrodes that potentially measure redundant information, we furthermore reduce the time required for exploration of the probe and are able to achieve higher rewards compared to a standard D-UCB strategy. In experiments with real neural data, we demonstrated that our approach is able to identify near-optimal electrode selection policies and to adapt to changes in the environment. In this way, it maximizes the amount of recorded information and outperforms fixed long-term selections in changing environments.

Although we considered a computationally simple way of estimating the quality of the recorded signals, our approach is flexible in this respect. Future research could explore different ways of computing the reward of the recorded signals according to the type of experiment. It could, for instance, include spike sorting capabilities, in order to identify and monitor specific neuronal activity and to improve the evaluation of the electrode's dependencies. In this study, we evaluated our approach using offline recorded datasets and simulated changes. It would be interesting to analyze the behavior of neural activity and to evaluate the performance of our approach over time on a larger set of recording channels as the amount of available electrodes reaches the capabilities of state-of-the-art microprobes.

References

[1] K. Seidl, S. Herwik, T. Torfs, H. P. Neves, O. Paul, and P. Ruther, "CMOS-Based High-Density Silicon Microprobe Arrays for Electronic Depth Control in Intracortical Neural Recording," *Journal of Microelectromechanical Systems*, vol. 20, no. 6, pp. 1439–1448, 2011.

[2] H. Neves, T. Torfs, R. Yazicioglu, J. Aslam, A. Aarts, P. Merken, P. Ruther, and C. van Hoof, "The NeuroProbes Project: A Concept for Electronic Depth Control," in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008.

[3] A. S. Herbawi, F. Larramendy, T. Galchev, T. Holzhammer, B. Mildenberger, O. Paul, and P. Ruther, "CMOS-Based Neural Probe with Enhanced Electronic Depth Control," in *18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*, 2015.

[4] O. Vysotska, B. Frank, I. Ulbert, O. Paul, P. Ruther, C. Stachniss, and W. Burgard, "Automatic Channel Selection and Neural Signal Estimation across Channels of Neural Probes," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[5] K. Seidl, T. Torfs, P. D. Maziere, G. van Dijck, R. Cseresa, B. Dombovari, Y. Nurcahyo, H. Ramirez, M. van Hulle, G. Orban, O. Paul, I. Ulbert, H. Neves, and P. Ruther, "Control and Data Acquisition Software for High-Density CMOS-Based Microprobe Arrays Implementing Electronic Depth Control," *Biomedizinische Technik*, vol. 55, pp. 183–191, 2010.

[6] G. van Dijck, K. Seidl, O. Paul, P. Ruther, M. van Hulle, and R. Maex, "Enhancing the Yield of High-Density Electrode Arrays Through Automated Electrode Selection," *International Journal of Neural Systems*, vol. 22, no. 01, pp. 1–19, 2012.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[8] D. Thierens, "An Adaptive Pursuit Strategy for Allocating Operator Probabilities," in *Proceedings of the Genetic and Evlutionary Computing Conference (GECCO)*, 2005.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Nonstochastic Multiarmed Bandit Problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[10] L. Kocsis and C. Szepesvari, "Discounted UCB," in *2nd PASCAL Challenges Workshop*, April 2006.

[11] A. Garivier and E. Moulines, "On Upper-Confidence Bound Policies for Switching Bandit Problems," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6925, pp. 174–188.

[12] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, "Multi-Armed Bandit, Dynamic Environments and Meta-Bandits," *Hal-00113668*, 2006.

[13] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial Multi-Armed Bandit: General Framework, Results and Applications," in *Proceedings of the 30 th International Conference on Machine Learning*, 2013.

[14] Z. Nenadic and J. Burdick, "A Control Algorithm for Autonomous Optimization of Extracellular Recordings," *IEEE Transactions on Biomedical Engineering*, vol. 53, 2006.

[15] G. van Dijck, A. Jezzini, S. Herwik, S. Kisban, K. Seidl, O. Paul, P. Ruther, F. U. Serventi, L. Fogassi, M. M. van Hulle, and M. A. Umiltà, "Toward Automated Electrode Selection in the Electronic Depth Control Strategy for Multi-unit Recordings," in *Neural Information Processing. Models and Applications*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6444, pp. 17–25.

[16] G. Buzsaki, "Large-Scale Recording of Neuronal Ensembles," *Nature Neuroscience*, no. 7, pp. 446–451, 2004.

[17] E. S. Page, "Continuous Inspection Schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[18] K. Seidl, M. Schwaerzle, I. Ulbert, H. Neves, O. Paul, and P. Ruther, "CMOS-Based High-Density Silicon Microprobe Arrays for Electronic Depth Control in Intracortial Neural Recording - Characterization and Application," *Journal of Microelectromechanical Systems*, vol. 21, pp. 1426–1435, 2012.