Deep Site-Invariant Neural Decoding from Local Field Potentials

Marko Angjelichinoski*, Bijan Pesaran†, and Vahid Tarokh*
*Department of Electrical and Computer Engineering, Duke University
†Center for Neural Science, New York University

***Abstract*—The non-stationary nature of neural activity prevents Brain-Computer Interfaces (BCIs) from leveraging data sets collected at different recording sites, such as cortical depths. As a result, invasive BCIs are continuously retrained on up-to-date site-specific data which is inefficient. In this paper we study the important intra-subject problem where neural activity signals collected at different cortical sites within a given subject are *jointly* used to train a decoder of motor intentions that generalizes well across the entire range of sites. We use directed graphical model with hidden latent variables that are censored via adversarial training that produces site-invariant low-dimensional representations of the neural activity. We evaluate the performance of our method on an experiment in which two macaque monkeys perform memory-guided visual saccades to one of eight target locations. The data sets are collected across range of cortical sites, from superficial to deep cortical sites. The results demonstrate that the site-invariant latent representations result in improved neural decoding by a peak margin of $\approx 50\%$, relative to the baseline approach where the neural decoder is trained on site-dependent features. The findings reported in this paper are an important step towards the development of efficient intra-subject BCIs that generalize well across range of cortical sites.**

I. Introduction

A. Motivation

The main objective of *Brain-Computer Interfaces (BCIs)* is to translate neural activity signals into control commands with the aim of restoring, supplementing and enhancing neurological functions [1], [2]. Some of the most important applications of BCIs include clinical practices for treating neurological disorders such as epilepsy, Parkinson's disease, Alzheimer's disease and other debilitating conditions, neural prosthetics for restoring lost or impaired motor functions, public safety, and the tactical domain [3], [4].

Invasive BCIs that are implanted chronically in the brain tissue and measure neural activity modalities that offer high spatial and temporal resolution such as action potentials and local field potentials (LFPs), offer strong candidate solutions for emerging domain applications. Nevertheless, their utility depends on satisfying a number of criteria with respect to (w.r.t.) the reliability, efficiency and safety summarized in the following three requirements. First and foremost, invasive BCIs should generalize well across different recording sites within the same same subject (also known as the intra-subject problem), including different cortical depths and cortical regions. Second, invasive BCIs should generalize well across unseen representatives of the same population (i.e., the inter- or cross-subject problem). Last but not least, invasive BCIs should generalize well even with limited training data. This requirement is of particular importance since acquiring training data is an expensive and time-consuming process; it is also a motivating factor for addressing the inter- and intra-subject problems mentioned earlier as this could potentially alleviate data scarcity by allowing BCI algorithms to be trained on data collected across different recording sites and/or subjects.

Despite the tremendous amount of progress over the last several decades, the implementation of reliable and efficient invasive BCIs remains a challenging problem. An important factor contributing to the difficulty arises from the *non-stationary* nature of the neural activity signals, whose statistical properties vary dramatically even under slight changes of the recording conditions [1], [5]. As a result, BCI algorithms trained and optimized on data collected from a given recording site (for instance, superficial cortical depth), fail to perform reliably when directly applied to data collected from different recording site (e.g., deep cortical site) even when both the training and the testing data are collected from the same subject at the same time. The issues also arise in the time-domain, i.e., when a BCI algorithm trained on data collected from given recording site in a given time interval, is applied to decode data collected from the same site at different time, even if the subject performs an identical task. Finally, the problem becomes notoriously challenging in inter-subject setups where, unless the subject-dependent variability of the data is not addressed properly, otherwise highly reliable BCI algorithms are rendered ineffective when applied to different subjects.

The highly variable statistical nature of the neural activity remains one of the most consequential challenges in neural engineering that can push BCIs into inefficient designs, where the algorithms are continuously retrained using only up-to-date site-specific data. It is of utmost importance to enable BCIs to leverage data acquired across different recording sites. We conclude that the non-stationary nature of the neural activity signals is one of the most consequential aspects of emerging invasive BCIs. Motivated by this, in this paper we focus on the intra-subject problem and show how to efficiently leverage data collected across range of cortical depths in macaque cortex to train a reliable neural decoder of motor movements that generalizes well across the entire range of depths.

In our previous work [6]–[9], we dealt with this issue to a certain extent. For instance, in order to increase the size of the data set at given cortical depth and driven by the assumption that similar recording sites (i.e., similar electrode depths in the

same subject) yield neural activity signals for similar properties, we leveraged data collected from neighboring regions and recording sites at different depths. In this way, we were able to form larger data sets at a given cortical depth and train more reliable neural decoders. However, the approach remains limited by the non-stationarity of the neural signals and does not efficiently leverage data collected over wide range of recording sites and depths.

Recent advances in deep learning and transfer learning, particularly the area of domain adaptation have had notable success in learning domain-invariant representations of data coming from variety of sources [10]. Some of these methods have been already adapted to BCI setups with varying degrees of success [5]. A notable example is the work in [11] where the authors use adversarial censoring to generate subject-invariant representations of EEG signals.

Motivated by the success of these methods, we propose to use a Variational Autoencoder (VAE) to obtain an invariant hidden representation of the neural activity that does not depend on the specific recording sites. In the proposed method, the invariance is enforced through an information-theoretic modification of the evidence lower bound (ELBO) which leads to an adversarial objective function where a separate adversary network aims to recover the corresponding descriptor variable that uniquely characterizes the recording depth, while, at the same time, the encoder/decoder pair of the VAE minimize a proxy for the mutual information between the depth descriptor and the latent code, making it increasingly difficult for the adversary to deduce any depth information from the latent code; we refer to the architecture as Adversarial VAE (A-VAE). Moreover, motivated by the robustness of the non-parametric methods for feature extraction from LFPs, we seek site-invariant latent codes over Pinsker's feature space [6]; this helps alleviate the data scarcity and allows the deep A-VAE model to be trained reliably. We apply the approach for decoding intended eye movement directions from LFP data collected from the prefrontal cortex (PFC) of two macaque monkeys performing memory-guided visual saccades to one of eight target locations on a screen. The data is collected across range of cortical depths, ranging from superficial sites, near the surface of PFC to deeper cortical sites, approaching white matter. The results show that the low-dimensional site-invariant latent representation of Pinsker's features, produced by the A-VAE result in improved classification performance of the neural decoder by a margin of $\approx 50\%$ in one of the subjects, relative to the baseline approach where the neural decoder is trained on site-specific features.

II. Methods

A neural decoding system consists of three main building blocks: 1) *data acquisition* in which the neural signals are conditioned, amplified and digitized, 2) *feature extraction*, where the acquired neural signals are further processed and the neural activity is represented in a (lower-dimensional) feature space, and 3) *neural decoder*, where the motor intentions are inferred from the feature space representation representations using a suitable classifier. Using this as a guideline, we organize this section into two parts. In Section II-A we describe the experiment and the acquired data, whereas in Section II-B we present the approach we use to obtain site-invariant feature representations of LFP signals.

A. Experimental Setup

We begin with an overview of the experiment and the acquired data. All procedures were approved by the NYU University Animal Welfare Committee (UAWC) and in accordance with the NIH . For additional details regarding the experimental setup, we refer the interested reader to [12], in which the dataset was originally reported.

1) Protocol: We study a classic experimental behavioral task involving memory-guided saccades to a target location [12]. Adult macaque monkeys (*M. mulatta*) are trained to perform memory-guided saccades to one of eight target locations on a screen, see Fig. 1 (top row). Individual trials are initiated by instructing the subject to fixate a central visually-presented target (event A). Once the subject maintains ocular fixation for a baseline period, one of the eight peripheral visual targets (drawn uniformly at random from the corners and edge midpoints of a square centered on the central target) is illuminated for 300 ms (event B); on each trial, the target light is chosen *independently* from previous trials. The extinguishing of the peripheral target (event C) marks the beginning of the *memory period* during which the subject must maintain fixation on the central target until it is extinguished (event D); this event instructs the subject to saccade to the remembered location of the peripheral target. The trial is completed successfully if the subject accurately maintains the gaze on the remembered target location event (E). Regardless of the outcome, the target light is reilluminated at the end of the trial (event F). We use only segments of neural activity recorded during the memory periods of successful trials; this epoch is especially interesting in memory-guided behaviors as the epoch presents information that determines the dynamics of the decision-making process and the subsequent motor response [6], [12].

2) Data Acquisition: The animals were instrumented with a head restraint prosthesis that enables head position fixation and video-based eye movement tracking. The recording chambers were surgically implanted over the lateral prefrontal cortex (PFC) and a microelectrode array consisting of $N = 32$ individually movable electrodes, i.e., channels was semichronically implanted in the chamber. Recent advances have suggested that *local field potential (LFP)* signals present a viable alternative to action potentials for designing invasive BCIs where the neural activity is recorded directly from brain tissue, via chronically implanted arrays of micro-electrodes (see [13] and references therein for a comprehensive overview of main advantages of LFP-based invasive BCI designs). LFPs refer to the potential of the extracellular currents surrounding individual neurons and, unlike the spiking activity of individual neurons, the LFP modality is more resilient to signal degradation [13]. In our experiment, LFP activity was sampled at $\nu_S = 1$ kHz.

3) Data Description: As the experiment progressed, the positions of individual electrodes were gradually advanced

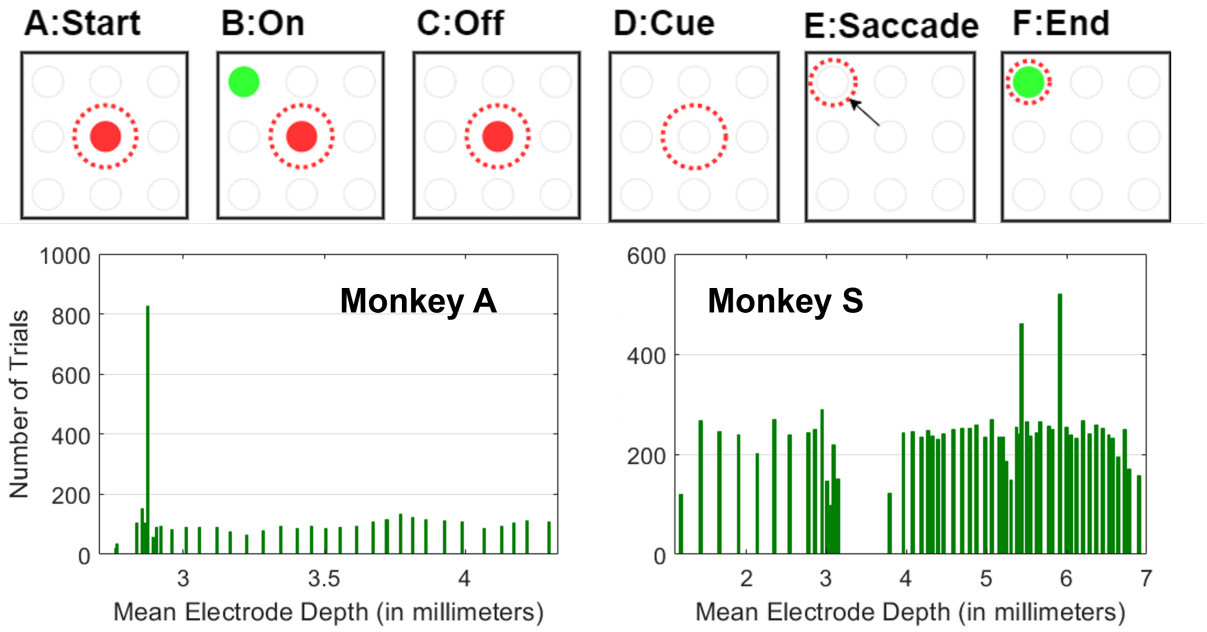

Fig. 1. Top row: experimental protocol and timeline of individual trials. Bottom row: number of trials per Electrode Depth Configuration (EDC). Left: Monkey A. Right: Monkey S.

deeper into the PFC. A fixed configuration of electrode positions over which multiple trials are performed is referred to as *electrode depth configuration (EDC)*, see also [6]. Each EDC is described by a 32-dimensional real vector; each entry in the vector contains the depth of each individual electrode with respect to its initial position. The experiment was performed over a total of 34 and 55 EDCs for Monkey A and Monkey S, respectively. The distribution of successful trials across EDCs is visually depicted in Fig. 1 (bottom row) where the horizontal axes denote the mean electrode depth, computed as a simple average of the entries of the EDC vector. The total number of trials across all EDCs is 3922 for Monkey A and 13064 for Monkey S. The respective averages are ≈ 90 and ≈ 250 trials per EDC for Monkey A and Monkey S, respectively; the only exception is EDC-6 with mean electrode depth ≈ 2.9 mm in Monkey A for which a total of 827 trials were collected across 10 recording sessions. It should be noted that the experiment for Monkey S began before action potential were detected, i.e., before the electrodes penetrated the surface of the PFC, and the recordings for the first 14 EDCs (that is, EDCs with mean electrode depths lower than 3.5 mm) were taken while some (or all) of the electrodes were still outside the PFC [12]. Given the size of the data sets for each EDC and considering that the dimension of the feature space exceeds 100 (see [6], [12]), we conclude that the number of trials for each individual EDC is insufficient to train a reliable decoder.

B. Site-invariant Features

The theory of non-parametric regression has proven to be useful for extracting meaningful features that enable deep models to be trained reliably on limited data [6], [8]. Due to the suitability for LFP-based neural decoding, here we rely on this recently proposed method as an intermediate step that helps obtain compact, low-dimensional representations of the raw LFP signals in addition to alleviating issues related with limited data, and we seek to find the site-invariant representations of the features provided by Pinsker's method described below.

1) Nonparametric Regression Framework: Let $\tilde{x}_t, t = 1, \ldots, T$ to denote the t-th LFP sample collected from some electrode during a given trial. We assume that the LFP signal consists of at least two additive components: (1) useful, information-carrying signal that guides behavior, represented by an unknown function x, and (2) noise-like component σw representing the remaining part of the LFP which does not contribute to behavior. Furthermore, we assume that the process w can be modelled as independent and identically distributed (i.i.d.) Gaussian noise. Hence, we have the following model:

$$\tilde{x}_t = x_t + \sigma w_t, \quad w_t \sim \mathcal{N}(0,1), \quad t = 0, \ldots, T-1, \quad (1)$$

where $x_t = x(t\nu_S)$ and $w_t = w(t\nu_S)$ are the corresponding discrete versions of x and w respectively, and σ denotes the standard deviation of the noise-like component of the signal. We do not assume parametric model for x; we only assume that the model lives in a space of smooth functions [14]. Furthermore, each behavior the subject performs yields different representation in the function space. In addition, the signal x will be also vary across repeated trials due to various neurological reasons [6], [12]. Hence, it is accurate to say that each specific task forms a class of functions in the function space. In such case the neural decoder reduces to conventional, multiple-class composite hypothesis testing; namely, we aim to find a plug-in discriminant function that maps the estimate of x into one of the behavioral tasks in the action set.

A desirable property of the neural decoder is to be consistent which can be guaranteed by taking the worst-case miss-classification probability to zero [14], [15]. This motivates

the use of *minimax-optimal* function estimators [14]. The theory of Gaussian sequence models provides a framework for designing finite-dimensional representations of the minimax-optimal function estimators. We first project the LFP model (1) onto an orthonormal set of functions, such as the Fourier basis functions to obtain the following sequence space representation:

$$\tilde{X}_l = X_l + \frac{\sigma}{\sqrt{T}} W_l, \quad W_l \sim \mathcal{N}(0,1), \quad l = 1, 2, \ldots, \quad (2)$$

where, $\tilde{X}_l$, X_l and W_l are the projections of the vectors $(\tilde{x}_0, \ldots, \tilde{x}_{T-1})$, $(x_0, \ldots, x_{T-1})$ and $(w_0, \ldots, w_{T-1})$ onto the l-th Fourier basis function. Now, instead of estimating x in the function space (1), we alternatively estimate the sequence of Fourier coefficients $\{X_l\}$ in the sequence space using (2). Pinsker's theorem gives an (asymptotically) minimax-optimal estimator for the Gaussian sequence model provided that the Fourier coefficients satisfy some predefined criteria. Let the Fourier coefficients X_l live in an ellipsoid such that $\sum_l a_l^2 X_l^2 \leq C$ where $a_1 = 0$, $a_{2m} = a_{2m+1} = (2m)^\alpha$ with $\alpha > 0$ denoting the smoothness parameter. The minimax-optimal estimator of X_l is given by [14]

$$X_l \approx \left(1 - \frac{a_l}{\mu}\right)_+ \tilde{X}_l = c_l \tilde{X}_l, \quad \mu > 0, \quad l = 1, 2, \ldots. \quad (3)$$

The function $(\cdot)_+$ operates as a rectified linear unit (ReLU), i.e., $(\cdot)_+ = \max\{\cdot, 0\}$. We see that Pinsker's estimator shrinks the observations $\tilde{X}_l$ by an amount $c_l = 1 - a_l/\mu$ if $a_l < \mu$; otherwise, it attenuates them to zero. Thus, the optimal estimator (3) yields only a *finite* number of L (complex) Fourier coefficients that correspond to the lowest L frequencies (including the DC), where L is the largest integer such that $a_L < \mu$ and $a_{L+1} \geq \mu$.

An important special case of Pinsker's estimator (3) is the truncation estimator where with $c_l = 1$ for $l \leq L$ and $c_l = 0$ for $l > L$, corresponding to a Sobolev class of infinitely-differentiable functions, where c_l is given in eq. (3); in other words, the finite-dimensional representation of the estimate of x is obtained by simply retaining the first L Fourier coefficients, corresponding to the L dominant frequency components of the complex spectrum of the LFP signal (including the DC component). The main difference between the general Pinsker's estimator (3) and the truncation estimator is the rate of convergence: Pinsker's estimator converges fastest to the true LFP waveform as $T \to \infty$ among all minimax-optimal estimators [14]. In practice, when the number of LFP samples T is limited, as in the current experiment, this is a rather subtle difference and one should not expect significant deviation in the decoding performance between (3) and its simplified truncation based variant, consistent with our previous work [6]. The truncation estimator is also simpler to implement than Pinsker's estimator since it introduces only a single free parameter, namely the number of retained complex Fourier coefficients L.

Finally, one last comment regarding Pinsker's estimator is in order. Through closer inspection of equation (3), we see that Pinsker's estimator essentially acts as a single non-linear (ReLU) layer of neural network, with weights that are fixed. Thus, Pinsker's method can be seen as an efficient and simple alternative of common adaptive feature extractors based on convolutional layers such as the EEGNet [16], that significantly reduces the implementation complexity and is suitable for limited data setups [9].

2) Adversarial Variational Autoencoder: Pinsker's feature extraction produces low-dimensional representations of the LFP signals that incorporate the relevant information stored in the amplitude and the phase of the signal, which in turn, allows for reliable training of deep models (such as deep neural network for classification [17]). This makes Pinsker's feature space attractive for further processing. The procedure is, however, applied on a per trial, per channel basis; as a result, Pinsker's features are not independent from the cortical sites of each electrode. To address this issue and find site-invariant representations of Pinsker's features, we adopt a common approach from transfer learning, based on directed graphical modelling with adversarial censoring of the latent variables, also known as Adversarial Variational Autoencoder (A-VAE) [11], see Fig. 2.

Let $X \in \mathbb{R}^D$ denote the vector comprising Pinsker's features from all channels, obtained via (3) (or its truncation variant) with $D = N \cdot (2L - 1)$ where N and L denote the number of channels and the number of retained complex Fourier coefficients, respectively. We introduce two additional variables. Let $Z \in \mathbb{R}^M, M < D$ denote the latent variable, and let S be a nuisance variable that incorporates information about the cortical site the signal X was collected from. A straightforward way to model S is via categorical random variable that indicates the EDCs at which the recording of the trials occurred. For instance, $S \in \{1, 2, \ldots, 34\}$ for Monkey A and $S \in \{1, 2, \ldots, 55\}$ for Monkey S. Note that both X and S are observable, whereas Z is hidden. The objective is to find hidden codes Z of X that are independent of S.

The joint pdf of X, Z and S can be factorized as $p_\theta(X, Z, S) = p(S)p(Z)p_\theta(X|Z, S)$ where we have assumed that $p(Z|S) = p(Z)$ to make Z and S independent explicitly. Evidently, learning this model amounts to maximizing the log-likelihood of the training data computed with respect to the conditional distribution $p_\theta(X|S) = \mathbb{E}_{Z \sim p(Z)}[p_\theta(X|Z, S)]$. Note that $p(S)$, even though easily estimated empirically from the training data, is not needed directly during training. Learning $p_\theta(X|S)$ is difficult due to the intractable posterior $p_\theta(Z|X, S)$. However, using a tractable variational posterior $q_\phi(Z|X, S)$ as an approximation of $p_\theta(Z|X, S)$, we can use the *evidence lower bound (ELBO)* as a surrogate for $\ln p_\theta(X|S)$ which can be written as:

$$\begin{aligned} \ln p_\theta(X|S) \geq \mathcal{L}_{\theta,\phi}(X, S) = \\ - \mathrm{KL}(q_\phi(Z|X, S)||p(Z)) \\ + \mathbb{E}_{Z \sim q_\phi(Z|X,S)}[p_\theta(X|Z, S)]. \end{aligned} \quad (4)$$

The encoder and decoder of the VAE, $q_\phi(Z|X, S)$ and $p_\theta(X|Z, S)$ are also known as recognizer and generator, respectively, terminology we shall use in the following. Note that independence is a particularly strong assumption and simply fixing the latent prior to be independent from the nuisance variable is usually insufficient to yield latent codes Z that are

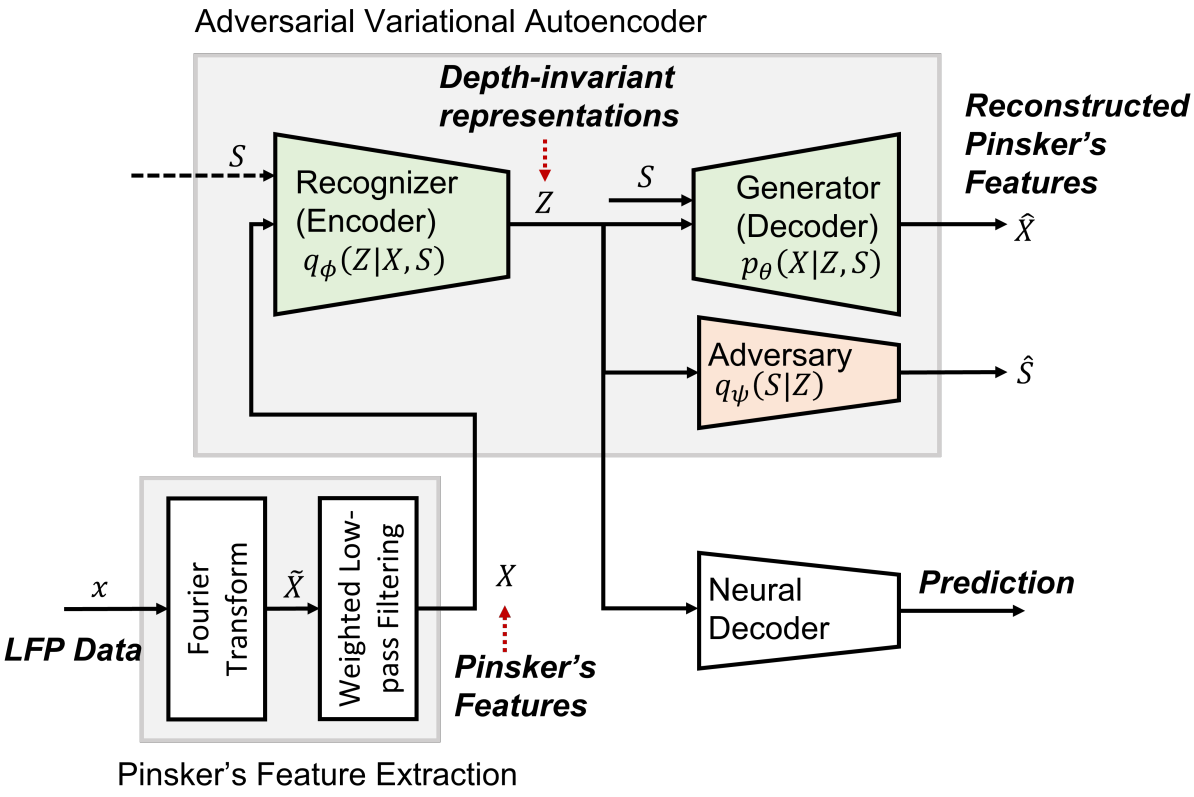

Fig. 2. Block diagram of a site-invariant neural decoding system with Adversarial Variational Autoencoder. In our earlier work [6], [8], [9], the Pinsker's features were directly fed to the neural decoder.

entirely independent from S in practice; this is again in part due to the inherently limited training data. Further invariance can be enforced by minimizing the mutual information $I(S;Z)$ where $Z \sim q_\phi(Z|X,S)$ is the latent code generated by the encoder; this criterion can be easily incorporated in the inequality from equation (4) as follows:

$$\ln p_\theta(X|S) \geq \mathcal{L}_{\theta,\phi}(X,S) - \lambda I(S;Z), \tag{5}$$

where the number $\lambda > 0$ is a free hyperparameter. The equality holds when $q_\phi(Z|X,S) = p_\theta(Z|X,S)$ and $\lambda I(S;Z) = 0$. By keeping λ strictly positive, the second term on the right-hand side of (5) can be zero only if the mutual information $I(S;Z)$ is zero which is achieved when the variables Z and S are independent. The regularization of the ELBO through the subtraction of the mutual information does not change the inequality and (at least in principle), both the right-hand and the left-hand side of (5) will have their optimums in the same θ. Since the mutual information is hard to compute empirically, we incorporate the variational lower bound into (5) which finally leads to the objective function $l_{\theta,\phi,\psi}(X,S)$ given as follows:

$$\begin{aligned}\ln p_\theta(X|S) &\geq l_{\theta,\phi,\psi}(X,S) \\ &= \mathcal{L}_{\theta,\phi}(X,S) - \lambda \mathbb{E}_{Z\sim q_\phi(Z|X,S)}[q_\psi(S|Z)],\end{aligned} \tag{6}$$

where $q_\psi(S|Z)$ is the variational posterior approximating the intractable true posterior $p(S|Z)$. The objective in (6) is adversarial in nature, as shown in Fig. 2. This can be most easily seen for categorical nuisance variable S. In this case, the variational posterior $q_\phi(Z|X,S)$ can be realized as a neural network for classification which is referred to as adversary and is trained to minimize the cross-entropy loss $-\mathbb{E}_{Z\sim q_\phi(Z|X,S)}[q_\psi(S|Z)]$. Concurrently, the VAE maximizes $l_{\theta,\phi,\psi}(X,S)$ which maximizes the cross-entropy loss $-\mathbb{E}_{Z\sim q_\phi(Z|X,S)}[q_\psi(S|Z)]$ producing latent codes Z that challenge the adversary's ability to infer S. Formally, this is equivalent to the following optimization problem:

$$\max_{\theta,\phi} \min_{\psi} l_{\theta,\phi,\psi}(X,S). \tag{7}$$

In other words, while the adversary is becoming better in inferring S from Z, the recognizer and the generator of the VAE try their best to fool the adversary.

III. Results

In this section we evaluate the performance of neural decoders of motor intentions from dept-invariant representations of LFPs, obtained using the A-VAE described in Section II-B2. First, in Section III-A we discuss the formation of the training and the test data sets from the experimental data, the technical elements of the implemented architecture of the A-VAE and the neural decoders and summarize the values of the hyper-parameters used in the evaluations. Section III-B presents the results along the main conclusions.

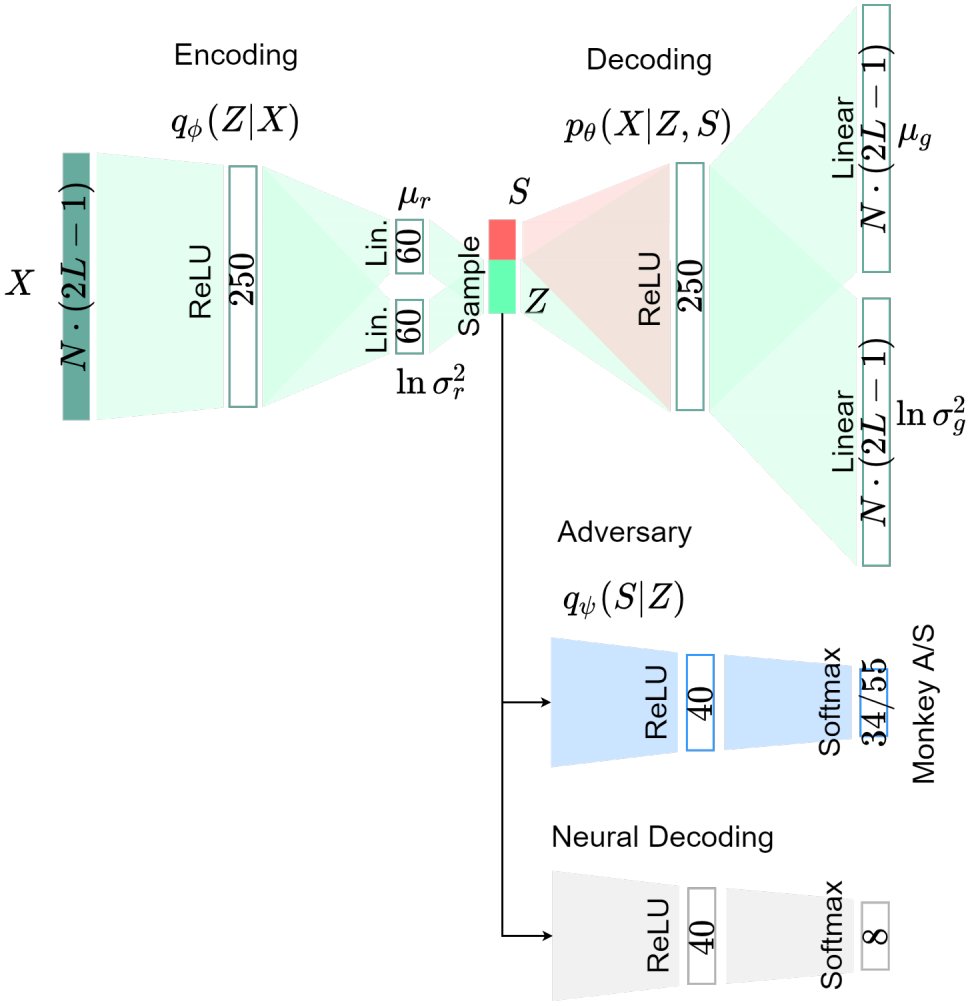

Fig. 3. Block diagram of a site-invariant neural decoding system.

A. Preliminaries

1) Training/Testing Data Sets: To understand the performance of the site-invariant codes for neural decoding, we need to devote special attention to the allocation of the training and testing sets. The raw data sets are discussed in Section II-A3, see also Fig. 1. The training and testing data sets are selected randomly for each subject separately: for Monkey A, we retain 400 trials for testing, and for Monkey S we allocate 2000 trials for testing, whereas the remaining trials in each case are used for training. Recall that in the case of Subject S, the EDCs with mean electrode depth smaller than 3.5 mm were collected while the electrodes were partially (or completely) outside the PFC; in addition, it has been shown that this data is particularly unreliable for training neural decoder and for the most part the classification performance here is close to random choice [6], [9]. Therefore, for Subject S we exclude the first 14 EDCs from the analysis and focus on the remaining 41 EDCs, with mean electrode depth larger than 3.5 millimeters. Finally, To avoid over-fitting and obtain reliable conclusions, we apply statistical averaging where the performance metric, namely the *average decoding accuracy*, is computed as an empirical average over 100 randomly selected testing data sets for each subject.

2) Architecture and Hyperparameters: We adopt disentangled Gaussian distributions for the prior, the encoder and the decoder, i.e., $p(Z) = \prod_{m=1}^{M} \mathcal{N}(0,1)$, $q_\phi(Z|X) = \mathcal{N}(\mu_r, \text{diag}(\sigma_r^2))$ and $p_\theta(X|Z,S) = \mathcal{N}(\mu_g, \text{diag}(\sigma_g^2))$, respectively. Note that we made the encoder to be independent from S explicitly, i.e., $q_\phi(Z|X,S) = q_\phi(Z|X)$, so that the system would not require knowledge of S during testing. Using these models, the ELBO, which is the first term on the right-hand side in (6), can be written in the following form:

$$\mathcal{L}_{\theta,\phi}(X,S) \approx \sum_{m=1}^{M} \left(1 + \ln \sigma_{r,m}^2 - \sigma_{r,m}^2 + \mu_{r,m}^2\right) - \sum_{d=1}^{D} \left(\ln \sigma_{g,d}^2 + \frac{(X_d - \mu_{g,d})^2}{\sigma_{g,d}^2}\right). \quad (8)$$

In the notation above, we used a single sample $Z \sim q_\phi(Z|X)$ to approximate the expectation appearing in the second term in (4) as this has shown to be sufficient in our evaluations. We use Multilayer Perceptron (MLP) networks to parametrize the encoder and the decoder, see also Fig. 3. Hence, μ_g and σ_g are

TABLE I
THE HYPER-PARAMETERS OF THE MODEL AND ITS NETWORKS.

Hyper-parameter	Value
Pinsker's Feature Extractor (with truncation)	
Number of LFP samples	$T = 650$ (first 0.65 seconds of memory period) [6]
Number of retained (complex) Fourier coefficients	$L = 9$
Optimization parameters valid for all networks (learning rates indicated separately)	
Optimization method	Adadelta
Minibatch size	75
Number of training epochs	300
Recognizer (VAE Encoder) and Generator (VAE Decoder)	
Topology	fully-connected
Number of hidden layers	1
Number of neurons in layers in recog./gen.	$288/60$ (input) - $250/250$ (hidden), $120/576$ (output)
Hidden neuron activation functions	ReLU
Learning rates	0.1
Adversary and Classifier (Neural Decoder)	
Topology	fully-connected
Number of hidden layers	1
Number of neurons in layers in adv./clf.	$60/288$ (input) - $40/40$ (hidden) - 34(M.A), 55(M.S)$/8$ (output)
Hidden neuron activation functions	ReLU
Learning rates	0.1(clf.)$/0.05$(adv.)
λ	1

the outputs of the decoder MLP excited by $Z = \mu_r + \sigma_r \odot \epsilon$, in addition to S, with $\epsilon \sim p(Z)$ generated by the standard multivariate Gaussian prior and μ_r and σ_r represented by the outputs of the encoder MLP excited by X. As the output features in the above model are assumed to be independent, the training examples are first decorrelated before training the A-VAE. In similar fashion as the encoding/decoding networks, the adversary is parametrized by an MLP classification network that for each X takes the latent code Z generated by the encoder and uses the softmax activation function at the output layer to compute the probabilities across the support of S (i.e., the set of unique EDCs) and infer S. Beside the adversary, we also connect an additional MLP classifier over the latent space; this network is the neural decoder, trained to infer the motor intention of the subject using the site-invariant representations, as shown in Fig. 3. It should be noted that the neural decoder was trained using the same training data used in training the A-VAE.

The specific values of the hyperparameters of the system are summarized, in part in Fig. 3 (hidden layers, hidden units) and, in part in Table I. In this work, we primarily focus on optimizing the hyperparameters pertaining to the architecture of the A-VAE and its corresponding networks, including the neural decoder. In other words, we did not fine tune the performance over the parameters of Pinsker's feature extraction method; rather, we fixed these parameters to values which have been shown previously to yield reliable performance [6]. For instance, we fixed T to 650 samples, corresponding roughly to the first half of the memory period, as suggested in [6]. Furthermore, we retain only $L = 5$ complex Fourier coefficient per channel, corresponding to the 5 lowest frequencies (that includes the DC component). Regarding the selection of the architecture of the A-VAE networks and the neural decoder as well as the values of related hyperparameters (such as optimization method, learning rate, minibatch size and so on), our aim was to find a configuration that generalizes well for both representative subjects. We therefore used exploratory cross-validations on randomly selected training data sets and recorded the performance. We selected the configuration that led to most reasonable performance in both subjects; it should be noted however that further, subject-specific fine tuning might produce even more reliable results, but this is outside the scope of the present work.

B. Evaluations

In all evaluations in this section, we use the average classification accuracy, averaged over all targets, as a performance measure. Our investigations as well some of our earlier works [6], where we investigated the performance with respect to the confusion matrices, suggest that the average accuracy is an adequate performance measure since the data sets are well-balanced across target classes and the confusion matrices are diagonally dominated. As a result, we find that additional performance metrics such as recall and score, do not contribute with new insights.

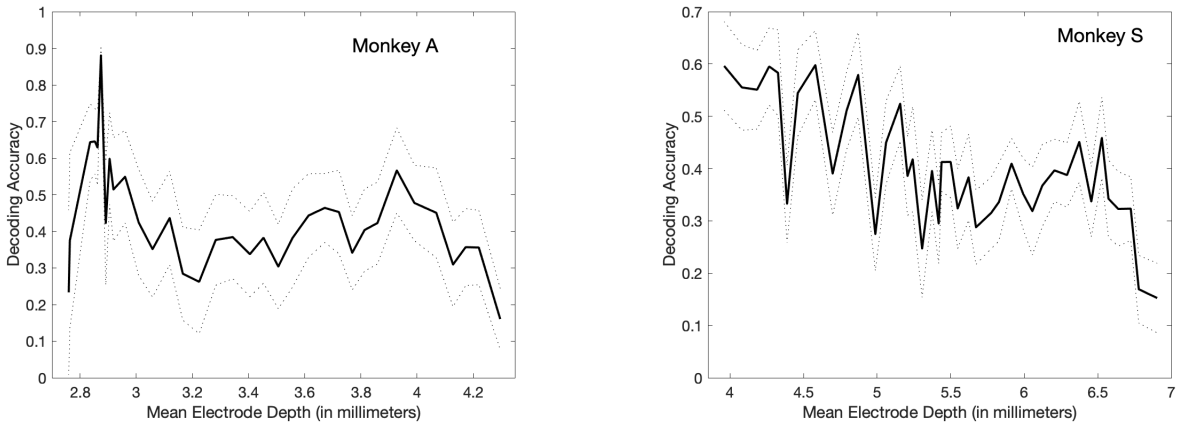

Fig. 4. Average site-specific decoding accuracy (solid line) $\pm$ one standard deviation (dotted lines). 15% of trials per depth were retained for testing.

Before discussing the performance of the the site-invariant neural decoder, it is insightful to see the performance of the site-specific decoder; that is a decoder trained on the data of individual depth. To be consistent in our analysis and to keep the complexity comparable, we fix the configuration of the depth-specific neural decoder to be the same configuration of the neural decoder used in the depth-invariant architecture with the only difference being in the size of the input which in the depth-specific case is $N \cdot (2L-1) = 288$; the remaining parameters, including the optimizer are given in Table I.

The average decoding accuracy of the depth-specific neural decoder is shown in Fig. 4 for both Monkey A and Monkey S across all recording depths. The relatively poor performance of the depth-specific neural decoder can be attributed to the size of the data sets at each depth; recall from Fig. 1 that the number of collected trials per EDC is rather limited. Conversely, we observe that for EDCs with relatively large number of trials as EDC-6 in Monkey A, the performance tends to improve. One way to mitigate the impact of limited data sets is to use *data bundling* as in [6], i.e., create larger depth-specific data sets by bundling several data smaller data sets from similar cortical depths (e.g., data sets whose EDC vectors are close to each other in Euclidean distance sense); nevertheless, this approach is susceptible to the non-stationarity of the data across depths which ultimately limits its application.

Next, we turn our attention to the depth-invariant neural decoding approach described in Section II-B2. We consider two baseline methods that we use as benchmark against which we evaluate the performance of the depth-invariant neural decoder. The first baseline method relies on straightforward application of the neural decoder over the depth-dependent Pinsker's features. We do so by using a separate MLP classifier network with similar topology as the neural decoder from Table I, with the only difference in the input layer which in this case has $N \cdot (2L-1) = 288$ neurons; in addition, we use the same optimization parameters as in Table I to keep the complexity comparable and the performance analysis streamlined. The other baseline relies on the vanilla VAE method without the adversary network. Recall that the prior distribution of the latent code was made explicitly independent from the nuisance variable representing the depth; therefore the purpose of this analysis is to check to what extent this design assumption alone is sufficient to yield depth-invariant latent representations of Pinsker's features. We implement the VAE model by using the same configuration of the recognizer and generator from Table I while removing the adversary; keeping the same configuration again helps with keeping the complexity comparable with the corresponding constituent blocks of the A-VAE architecture.

TABLE II
AVERAGE CLASSIFICATION ACCURACY OF THE NEURAL DECODER $\pm$ ONE STANDARD DEVIATION (IN $\%$) COMPUTED FOR 100 TRAINING/TESTING DATA SPLITS: 400 AND 2000 TESTING TRIALS USED FOR MONKEY A AND MONKEY S, RESPECTIVELY..

	Subjects	
	Monkey A	Monkey S
Direct decoding over Pinsker's features	56.38±1.60	36.33±0.85
Vanilla Variational Autoencoder	52.91±1.59	34.56±0.70
Adversarial Variational Autoencoder	**71.63±1.41**	**52.12±0.81**
Relative Gain w.r.t. direct decoding	27.2±4.70	50.84±4.12

The results are given in Table II. In both subjects, the neural decoder trained over the latent codes generated via adversarial censoring yields improved performance with respect to both baseline methods. While the relative improvement is larger for Monkey S, the performance peaks for Monkey A. This is a common pattern with this data set, originally observed [12] and through subsequent investigations [6]; namely, the Monkey A data set consistently yields more reliable neural decoders.

Interestingly, the neural decoder trained over latent codes generated by vanilla VAE does not show any noticeable improvement over the other baseline, namely the direct application of neural decoder over the Pinsker's features. This suggest that the assumption does little in terms of enforcing invariance with respect to the nuisance variable and making the prior independent from the depths is not enough to yield depth-invariant latent code. In other words, the vanilla VAE

TABLE III
AVERAGE CLASSIFICATION ACCURACY OF THE ADVERSARY ± ONE STANDARD DEVIATION (IN %) COMPUTED FOR 100 TRAINING/TESTING DATA SPLITS. 400 AND 2000 TESTING TRIALS USED FOR MONKEY A AND MONKEY S, RESPECTIVELY.

	Subjects	
	Monkey A	Monkey S
Depth inference from Pinsker's features	75.12±2.10	58.00±1.80
Vanilla Variational Autoencoder	51.93±2.26	37.53±1.2
Adversarial Variational Autoencoder	12.63±2.18	6.89±0.88

merely produces low-dimensional latent codes which are still depth-dependent. This can be further seen in Table III where we show the classification accuracy of the adversary. Note that in the case of the vanilla VAE, the second term in (6) is removed, causing the optimization problem in (7) to decouple. We notice that in this case the adversary can still recover the depth indicator from the latent representation relatively well, especially when compared with the random choice baseline ($\approx 3\%$ for Monkey A and $\approx 2.3\%$ for Monkey S). Comparing these results with the classification accuracy when inferring the depth indicator directly from Pinsker's features, we see that some invariance has still been enforced in the vanilla VAE case, which can be attributed to the independent prior. However, the invariance is much stronger in the case of the A-VAE and in this case the adversary is able to recover significantly less site information from the latent codes.

IV. DISCUSSION

In this paper we addressed the issues of the non-stationarity of neural activity across cortical depths by finding depth-invariant feature representations of the neural activity. To this end, we proposed a solution based on directed graphical model with adversarial, i.e., Adversarial Variational Autoencoder (A-VAE). We verified the viability of the method in the context of neural decoding of motor intentions from LFPs using an experiment in which two macaque monkeys perform memory-guided visual saccades to one of eight target locations on a screen, and where the data was collected across range of cortical depths in both subjects. The results demonstrate that a neural decoder trained over depth-invariant low-dimensional representations of the neural activity outperforms a neural decoder trained directly over depth-dependent features, by a relative margin of up to 50%.

The findings we report have potentially far-reaching practical implications for the development of invasive BCIs in the domains of healthcare and public safety. Apart from civilian applications, BCIs are also foreseen as powerful emerging technological tool in the tactical domain. Even though our approach is application-agnostic, we note that the technological components of the proposed system are also applicable in scenarios of potential importance to national security. We note that further investigations are required to generalize our results across different subjects and across a range of motor tasks. Additional preclinical studies in animal models as well as clinical studies in patients are needed; this is a non-trivial, time-consuming and expensive endeavour.

ACKNOWLEDGMENT

This work was supported by the Army Research Office MURI Contract Number W911NF-16-1-0368 and DURIP Contract W911NF-21-1-0080.

REFERENCES

[1] R. P. N. Rao, *Brain-Computer Interfacing: An Introduction*. Cambridge University Press, 2013.

[2] N. Tiwari, D. R. Edla, S. Dodia, and A. Bablani, "Brain computer interface: A comprehensive survey," *Biologically Inspired Cognitive Architectures*, vol. 26, pp. 118 – 129, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212683X18301142

[3] S. J. Bensmaia and L. E. Miller, "Restoring sensorimotor function through intracortical interfaces: progress and looming challenges." *Nature Neuroscience*, vol. 15, no. 8, pp. 313–325, May 2014.

[4] K. A. Moxon and G. Foffani, "Brain-machine interfaces beyond neuroprosthetics," *Neuron*, vol. 86, no. 1, pp. 55 – 67, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627315002603

[5] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, Feb 2016.

[6] M. Angjelichinoski, T. Banerjee, J. Choi, B. Pesaran, and V. Tarokh, "Minimax-optimal decoding of movement goals from local field potentials using complex spectral features," *Journal of Neural Engineering*, vol. 16, no. 4, May 2019.

[7] M. Angjelichinoski, J. Choi, T. Banerjee, B. Pesaran, and V. Tarokh, "Cross-subject decoding of eye movement goals from local field potentials," *Journal of Neural Engineering*, vol. 17, no. 1, p. 016067, feb 2020.

[8] M. Angjelichinoski, M. Soltani, J. Choi, B. Pesaran, and V. Tarokh, "Deep james-stein neural networks for brain-computer interfaces," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1339–1343.

[9] ——, "Deep pinsker and james-stein neural networks for decoding motor intentions from limited data," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1058–1067, 2021.

[10] T. Phung, T. Le, L. Vuong, T. Tran, A. Tran, H. Bui, and D. Phung, "On learning domain-invariant representations for transfer learning with multiple sources," *CoRR*, vol. abs/2111.13822, 2021. [Online]. Available: https://arxiv.org/abs/2111.13822

[11] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Learning invariant representations from eeg via adversarial inference," *IEEE Access*, vol. 8, pp. 27 074–27 085, 2020.

[12] D. A. Markowitz, Y. T. Wong, C. M. Gray, and B. Pesaran, "Optimizing the decoding of movement goals from local field potentials in macaque cortex," *Journal of Neuroscience*, vol. 31, no. 50, pp. 18 412–18 422, 2011.

[13] B. Pesaran, M. Vinck, G. T. Einevoli, A. Sirota, P. Fries, M. Siegel, W. Truccolo, C. E. Schroeder, and R. Srinivasan, "Investigating large-scale brain dynamics using field potential recording: analysis and interpretation." *Nature Neuroscience*, vol. 21, no. 8, pp. 903–919, Jul. 2018.

[14] I. M. Johnstone, *Gaussian estimation : Sequence and wavelet models*, 2017.

[15] T. Banerjee, J. Choi, B. Pesaran, D. Ba, and V. Tarokh, "Classification of local field potentials using gaussian sequence model," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*, June 2018, pp. 683–687.

[16] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, jul 2018. [Online]. Available: https://doi.org/10.1088/1741-2552/aace8c

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.