Developing machine-learning methods for the analysis of electromagnetic brain activity

Ivan Zubarev
Developing machine-learning methods for the analysis of electromagnetic brain activity

Ivan Zubarev
Abstract

Traditionally, analysis of electromagnetic brain activity focuses on modeling the data-generating process and identifying which components in the measured signal are associated with experimental manipulations.

At the beginning of XXI century, machine-learning based approaches aiming to infer brain states from the measurements started to gain increasing popularity. These methods rely on extracting complex multivariate patterns allowing to predict experimental conditions from the measurements. This thesis summarizes how such prediction-based methods can be applied to measurements of electromagnetic brain activity in a way that allows to advance our understanding of the underlying neural processes.

Because these techniques belong to a class of inverse probability problems and do not model the data-generating process directly, interpreting the learning outcomes in terms of the underlying neurophysiological processes is not straightforward.

Instead, predictive models allow testing the generalization properties of brain activity across e.g. experimental tasks (Publication I) and individuals (Publication II), as well as employ model comparison techniques to gain additional insights about the statistical properties of the data-generating process indirectly i.e. by comparing models with different structural constraints (Publication II). Moreover, projecting relevant model parameters learned from the data back into the input space can provide additional insights into the data-generating process and thus complement traditional approaches. These approaches are implemented in an open-source academic software described in Publication III.
Preface

I have a difficult time seeing scientific results, especially in neurobiology, as anything but provisional approximations, to be enjoyed for a while and discarded as soon as better accounts become available. But skepticism about the current reach of science ... does not imply diminished enthusiasm for the attempt to improve provisional approximations

António Damásio
Descartes’ Error, 1994

I always believed that the end goal of studying the human brain is to bring us closer to understanding ourselves and other human beings. I also believe that these problems worth spending life on. Modern advances in neuroimaging technologies offer us a chance to observe neural phenomena with ever-increasing richness and complexity. I hope that methods developed here would allow researchers to harness this complexity and to gain further insights into who we are.

The work comprising this Thesis was carried out at the Department of Neuroscience and Biomedical Engineering (NBE) of Aalto University. It took me some 7 years to complete it, which is a pretty long time. Luckily, I had the luxury to work on the topic that interested me profoundly and has never failed to provide enough challenges and opportunities for growth. I also had a unique opportunity to pursue my interests independently, formulate the problems and decide on the best course of action to approach them. For all this, I am grateful to my supervisor, prof. Lauri Parkkonen.

During this time, I have had the pleasure of working with many outstanding researchers from a wide variety of backgrounds. Most of all I want to thank my key collaborators Dr. Rasmus Zetter and L.Sci Hanna-Leena Halme for sharing the joys, excitement, and also some pains of doing research together. I would also like to thank Dr. Elina Nagaeva, Dr. Sergey Shishkin, Dr. Anastasia Ovchinnikova for working on projects that although not included in this thesis have contributed to the ideas developed here greatly. Most of all, I would like to thank prof. Alexei
Ossadtchi for his mentorship, inspiration, and some of the most fruitful scientific discussions.

NBE is an excellent place to observe scientists at work. Working in such a vibrant environment exposed me to a variety of perspectives about how one can do research, think about science, and approach life in general. I am grateful to all my friends and colleagues at the NBE, especially Dovile Kurmanaviciute, Dr. Victor Souza, Dr. Niko Mäkelä, Dr. Tuomas Mutanen, Dr. Johanna Metsomaa, Dr. Joonas Iivanainen, Dr. Andrey Zhdanov, and many others who always managed to provide me with rich food for thought on these matters.

Finalizing this work would not be possible without the support from my closest friends and family and, especially, Liliia Andriichuk who was there during the toughest days of this journey.

I dedicate this thesis to the memory of my father, prof. Pyotr Nikolaevich Zubarev, who has always been an example for me both professionally and personally.

Espoo, November 2020,

Ivan Zubarev
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>1</td>
</tr>
<tr>
<td>Contents</td>
<td>3</td>
</tr>
<tr>
<td>List of Publications</td>
<td>5</td>
</tr>
<tr>
<td>Author's Contribution</td>
<td>7</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>9</td>
</tr>
<tr>
<td>Symbols</td>
<td>11</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>13</td>
</tr>
<tr>
<td>2. Aims of the study</td>
<td>15</td>
</tr>
<tr>
<td>3. Background</td>
<td>17</td>
</tr>
<tr>
<td>3.1 Measurements of electromagnetic brain activity</td>
<td>17</td>
</tr>
<tr>
<td>3.1.1 Nervous system overview</td>
<td>17</td>
</tr>
<tr>
<td>3.1.2 Electromagnetic brain activity</td>
<td>18</td>
</tr>
<tr>
<td>3.1.3 Magnetoencephalography</td>
<td>19</td>
</tr>
<tr>
<td>3.1.4 Generative model of the event-related EEG-MEG signals</td>
<td>20</td>
</tr>
<tr>
<td>3.1.5 Analysis of the MEG data</td>
<td>22</td>
</tr>
<tr>
<td>3.2 Statistical Inference</td>
<td>23</td>
</tr>
<tr>
<td>3.2.1 Uncertainty and probability</td>
<td>24</td>
</tr>
<tr>
<td>3.2.2 Modelling assumptions and hypothesis space</td>
<td>24</td>
</tr>
<tr>
<td>3.2.3 Maximum Likelihood Principle</td>
<td>25</td>
</tr>
<tr>
<td>3.2.4 Generative and Discriminative learning</td>
<td>26</td>
</tr>
<tr>
<td>3.3 Machine learning</td>
<td>27</td>
</tr>
<tr>
<td>3.3.1 Objective function</td>
<td>27</td>
</tr>
<tr>
<td>3.3.2 Computational graph</td>
<td>28</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Neural Networks and deep learning</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Convolutional Neural Networks</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Optimization</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Probabilistic interpretation of learning outcomes</td>
</tr>
<tr>
<td>3.4</td>
<td>MEG data as a specific application domain of machine learning methods</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Statistical properties of the MEG data</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Encoding and decoding approaches</td>
</tr>
<tr>
<td>3.5</td>
<td>Applications of machine learning methods in neuroimaging</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Knowledge discovery</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Brain-Computer Interfaces</td>
</tr>
<tr>
<td>4.1</td>
<td>Transfer learning</td>
</tr>
<tr>
<td>4.2</td>
<td>Generalization tests</td>
</tr>
<tr>
<td>4.3</td>
<td>Convolutional neural networks for MEG data</td>
</tr>
<tr>
<td>5.1</td>
<td>Publication I: Evidence for a general performance monitoring system in the human brain</td>
</tr>
<tr>
<td>5.2</td>
<td>Publication II: Adaptive neural network classifier for MEG signals</td>
</tr>
<tr>
<td>5.3</td>
<td>Publication III: MNEflow neural networks for MEG-EEG analysis</td>
</tr>
<tr>
<td>6.</td>
<td>Discussion</td>
</tr>
</tbody>
</table>

Bibliography

Publications
This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.


Author’s Contribution

Publication I: “Evidence for a general performance-monitoring system in the human brain”

The candidate conceived the experiment, collected the data, performed the analysis and wrote the initial manuscript. Together with the second author the candidate designed the experiment and finalized the manuscript.

Publication II: “Adaptive neural network classifier for decoding MEG signals”

The candidate developed and implemented the method. Together with the second author the candidate designed the and conducted the experiments, designed and implemented interpretation heuristics and wrote the initial version of the manuscript. Third author collected the data for experiment 2. All authors contributed to finalizing the manuscript.

Publication III: “MNEflow: Neural networks for EEG/MEG decoding and interpretation”

The candidate designed and implemented the software and wrote the initial version of the manuscript. The second author implemented unit testing. The last author contributed to finalizing the manuscript.
**Abbreviations**

**MEG** Magnetoencephalography  
**EEG** Electroencephalography  
**ECoG** Electrocorticography  
**fMRI** Functional magnetic resonance imaging  
**fNIRS** Functional near-infrared spectroscopy  
**MNE** Minimum norm estimate  
**MLE** Maximum likelihood estimate  
**MAP** Maximum a posteriori estimate  
**ICA** Independent component analysis  
**MAE** Mean absolute error  
**MSE** Mean squared error  
**BCI** Brain–computer interface  
**ERF** Event-related field  
**ERS** Event-related synchronisation  
**ERD** Event-related desynchronisation  
**EPSP** Excitatory post-synaptic potential  
**IPSP** Inhibitory post-synaptic potential  
**SNR** Signal-to-noise ratio  
**CNN** Convolutional Neural Network  
**ROC AUC** Receiver operating characteristic: area under curve  
**iid** Independent and identically distributed
Symbols

$X$ data, inputs or features

$y$ target variable

$D$ dataset comprising $X$ and $y$

$\theta$ trainable model parameters

$b$ bias coefficients

$w$ weight coefficients

$A$ mixing matrix

$W$ de-mixing matrix

$H$ modelling assumptions, hypothesis, hyperparameters

$\alpha$ learning rate

$\lambda$ regularization coefficient

$J$ objective function, cost function

$f(x)$ any function of $x$

$p(x)$ probability density function of $x$

$h^i_j$ output of $j$-th node of $i$-th layer of a neural network

$p(x|y)$ conditional probability density function of $x$ given $y$
1. Introduction

Advances in brain imaging technologies and computational methods in recent decades made it possible to accumulate tremendous amounts of data about the structure and the function of the human brain. Depending on the measurement technique, these data elucidate various aspects of the electromagnetic, metabolic, and other processes occurring in the brain across different spatial and temporal scales. While no single imaging method holds promise to "solve" the brain entirely, their variety provides us with an opportunity to integrate these observations using our own mental machinery. As instrumentation develops further, however, the complexity of these data grows proportionally. Because our cognitive capacity is limited, amassing such a multitude of data does not immediately guarantee a better understanding. It seems that the bottleneck becomes our ability to extract knowledge or insights from these data.

Machine learning algorithms hold great promise to alleviate this problem by making the extraction of statistical regularities (or patterns) from these data automatic. However, there is a growing concern that complex machine learning algorithms themselves can be too complicated for human interpretation and understanding (Chiang, 2000).

**Understanding** refers to a cognitive process, ultimately allowing an individual to deal adequately with an object or phenomenon. Merriam-Webster dictionary defines understanding as "a capacity to apprehend general relations of particulars" and "the power to make experience intelligible," i.e., allowing others to interact with it constructively (Merriam-Webster).

Algorithmic information theory describes understanding as a compression of information. In his "Discourse on Metaphysics" Leibniz ((1686) as cited in Chaitin Gregory, 2006) posited that a theory has to be simpler than the data it explains; otherwise, it is worthless. Without this constraint, even completely random data can be described by an arbitrarily complex theory. At the extreme, the data can always be described by the theory of the same size as the data itself. Such description, however, will be trivial, and a model describing it will be not useful and, most likely,
Modeling natural phenomena involves identifying valid, novel, useful, and comprehensible regularities in the data (Fayyad et al., 1996) allowing us to, e.g., make useful predictions or describe the data-generating process. Depending on the modeling task, multiple useful models of the same phenomena can exist at the same time. A model that gives accurate predictions is valuable because it can provide novel means to solve practical problems, while models providing informative interpretations are "enabling others to fruitfully think about an idea" (Buja, 2000).

Parts of the data that are not described by a given model are often treated as random noise (Shawe-Taylor and Cristianini, 2004). Intuitively, the noise can arise from, e.g., measurement process or be inherent to the underlying phenomena. In any case, it introduces uncertainty into the inference process.

Uncertainty can arise from two different sources. It can be associated with randomness intrinsic to the studied phenomena (i.e., aleatory uncertainty), or we can be uncertain about something simply because we lack knowledge (i.e., epistemic uncertainty). The latter epistemic uncertainty can be reduced by further investigation (O'Hagan, 2004).

This thesis is organized as follows. Section 3.1 reviews the origins of the electromagnetic brain activity and neuroimaging approaches used to study it non-invasively. Sections 3.2 and 3.3 introduce relevant concepts in statistics and general machine-learning used throughout the thesis and in publications. Section 3.4 focuses on describing neuroimaging as an application domain of machine-learning methods. Section 3.5 covers practical applications. Computational approaches used and developed in Publications are described in detail in Section 4.
The goal of this thesis to summarize my observations on how machine-learning methods can be applied to the measurements of the electromagnetic brain activity in a way that would lead to solving practical real-world problems and advance our understanding of the human brain.

- Publication I Proposes to use time-resolved classification in combination with across-condition generalisation tests to probe similarity between brain responses elicited in different tasks.

- Publication II Introduces compact convolutional neural networks, designed to capture known regularities in the MEG signals efficiently while being robust to inter-individual variability.

- Publication III Introduces an open source software implementing methods developed in Publication II
3. Background

3.1 Measurements of electromagnetic brain activity

3.1.1 Nervous system overview

*Neurons*

Neurons, basic cells of the nervous system, are electrically active cells. At rest, their cellular membranes maintain a stable potential difference between extracellular and intracellular space. Transmembrane ion currents, result in changes in the transmembrane potential. Whenever the transmembrane potential reaches an activation threshold, neurons produce action potentials (AP) - abrupt changes in their potentials due to the opening of specific voltage-gated ion channels. AP propagates through the axon resulting in the release of neurotransmitters - signaling molecules released from the axon terminals into the synaptic cleft. Once released, neurotransmitters reach specific binding sites in the postsynaptic cell producing changes in its transmembrane potential. These changes can, e.g., move the postsynaptic cell closer to the activation threshold resulting in a depolarizing or excitatory postsynaptic potential (EPSP). On the contrary, (hyperpolarizing) inhibitory postsynaptic potential (IPSP) moves the cell further away from the activation threshold. In a postsynaptic cell, excitatory and inhibitory inputs are integrated as they propagate along the dendritic tree towards the soma. Depending on whether the input was received on distal or proximal parts of the dendrite, the soma, or the axon hillock, such inputs can affect the postsynaptic transmembrane potential differently. Generally speaking, more proximal inputs have greater effects. Dendritic integration allows the post-synaptic cell to carry out complex computations subserving the processing of information. Needless to say that it is an intricate and highly complex process affected
Neuroplasticity is one of the core principles allowing the nervous system to learn these representations from the observed environment. In brief, neuroplasticity can be described by Hebb’s law (Hebb, 1949), according to which the efficiency of synaptic transmission between two neurons increases whenever activation of a given pre-synaptic cell is followed by activation of the postsynaptic cell. This process occurs on multiple levels of the nervous system, including synaptic, morphological, circuit-wide, and system-level mechanisms (Zenke and Gerstner, 2017). In computational terms, the weight of a particular synaptic input increases whenever activation of a pre-synaptic cell leads to an activation of a postsynaptic cell and vice versa (Markram et al., 1997; Caporale and Dan, 2008).

These two mechanisms combined allow the central nervous system to learn and maintain complex representations of the surrounding world, subserving perception, action, and higher cognitive functions (Keysers and Gazzola, 2014).

3.1.2 Electromagnetic brain activity

The adult human brain is estimated to consist of $86.1 \pm 8.1$ billion neurons, with some 19% of which located in the cerebral cortex (Azevedo et al., 2009). In the cortex, pyramidal cells account for the majority of the neuronal population and are considered its primary excitation units.

In this thesis, we focus on electromagnetic signals arising from the coordinated activity of large populations of the pyramidal cells in the human neocortex (Murakami and Okada, 2006). This activity can occur spontaneously or in response to external stimuli.

Several properties of the pyramidal cells allow studying their coordinated activity non-invasively. Most notably, their long apical dendrites are aligned in orientation towards the cortical surface. Whenever these dendrites receive EPSPs or IPSPs, the electrical potential difference between the distal and proximal parts of apical dendrites produces a net primary current orthogonal to the cortical surface (Murakami et al., 2003). Whenever a large number of the pyramidal cells receive similar inputs, the summation of these spatially-aligned net currents produces an electric field which can be detected outside of the head (Murakami and Okada, 2006; Baillet, 2017). Moreover, as IPSPs and EPSPs develop more slowly over time, compared to APs, net currents produced by the postsynaptic events in this neuronal population are also summed over time. This summation gives rise to an electromagnetic field sufficiently strong to be detected outside of the head non-invasively using techniques such as Electroencephalography (EEG) and Magnetoencephalography (MEG).
3.1.3 Magnetoencephalography

Magnetoencephalography is a non-invasive technique for measuring the magnetic fields produced primarily by the coordinated electrical activity of large populations of neocortical pyramidal cells (Hämäläinen et al., 1993). Neuronal currents generate the magnetic induction that can be measured, e.g., with a pick-up coil. MEG signal is complementary to that of electroencephalography (EEG), which measures the potential distribution caused by the neural activity by placing electrodes on the scalp. MEG is considered to have higher spatial resolution than EEG because the EEG signal is distorted by the heterogeneous conductivity profile of head tissues to a greater extent than the MEG signal. The magnetic permittivity, however, is homogeneous across all tissues, resulting in a lesser spatial distortion of the MEG signal (see, e.g. Baillet, 2017).

Neural currents corresponding to simultaneous activation of between 10 000 and 50 000 pyramidal cells are strong enough to be detectable non-invasively. The resulting magnetic induction measured extracranially is on a scale of femtoteslas ($10^{-15}$ T). The sensitivity of MEG to a given neural source varies considerably across the brain, depending on the location and orientation of the pick-up coils with respect to the source (Baillet, 2017).

MEG data typically include 1) evoked responses (event-related fields; ERFs) that are phase-locked to specific sensory, cognitive, or motor events, 2) induced modulations of ongoing oscillations time-locked but not phase-locked to external events, and 3) spontaneous (i.e., not related to any event) brain activity during sleep or resting wakefulness (Figure 3.1) (Hari and Salmelin, 2012).

![Figure 3.1. Typical brain event-related responses measured by MEG and spontaneous activity in time-domain](image_url)

As methods developed in this thesis aim to infer brain states occurring
in response to the presented stimuli, we primarily focus on evoked and induced responses in the time domain. Other event-related brain responses can involve, e.g., transient changes in the various estimates of functional connectivity between several brain regions (He et al., 2019) or phase-amplitude coupling between several frequency bands (Canolty et al., 2006; Florin and Baillet, 2015).

**Evoked responses** are stereotypical patterns of electromagnetic brain activity emerging within a fixed latency after a reference event (e.g., presentation of sensory stimuli) (Kappenman and Luck, 2012). The amplitude of such event-related responses is relatively small compared to the ongoing background brain activity. Considering such background activity as noise, single-trial evoked responses typically considered to have a signal-to-noise ratio (SNR) $\sim 1$. In MEG, such event-related fields (ERFs) are often challenging to detect and identify in an un-processed measurement. The traditional analysis relies on repeating measurements of the ERFs to each type of stimuli many times within the experiment and obtaining the average ERF waveform. Averaging away the unrelated background activity increases the SNR but leads to considerable loss of information contained in, e.g., trial-to-trial variations in the ERF amplitude or latency (see e.g. David et al., 2006).

**Induced oscillatory responses** are modulations of ongoing oscillatory activity occurring consistently after a reference event. Such modulations typically occur within a narrow frequency band and are not-necessarily phase-locked to the stimulus onset. This activity can manifest as consistent increases or decreases of oscillatory power in particular brain regions and frequency bands and are referred to as event-related synchronization and desynchronization, respectively.

In the time domain, these event-related MEG signals share several statistical properties related to the underlying neurophysiological activity and the measurement process (see Section 3.4.1). The following sections briefly introduce the known statistical properties of the data-generating process and common approaches to analyzing MEG data.

### 3.1.4 Generative model of the event-related EEG-MEG signals

An event-related MEG measurement comprises a simultaneous reading of a large number (typically 200-300) of MEG sensors sampled over short periods of time after the reference event (e.g., a presentation of visual stimuli) at a high temporal resolution (typically up to $\sim 1000$ Hz). Mathematically, these data can be represented by an $n \times t$ data matrix $X \in \mathbb{R}^{n \times t}$ containing measurements from $n$ sensors at $t$ time points.

These data are generated by an unknown number of loosely-defined latent sources (Daunizeau and Friston, 2007). Depending on the modeling
approach, these sources can be thought of as focal populations of synchronously activated neurons (i.e., "neurophysiological source") or widely-distributed networks of brain regions that display similar statistical properties (i.e., "statistical source").

Importantly, activation of a single focal source is picked up by multiple MEG sensors. Similarly, a single MEG sensor can pick-up a superposition of the signals generated by multiple sources that are active at the same time (which is often the case). Thus, the resulting MEG measurement $X \in \mathbb{R}^{n \times t}$ typically contains a mixture of an unknown number $k$ of simultaneously active sources $s$ such that at each time point $t$.

$$x_t = As_t + \epsilon$$

(3.1)

Where $A$ is referred to as a mixing matrix, $s$ is an instantaneous amplitude of each of the $k$ sources at time $t$, and $\epsilon$ is additive observation noise.

Each of the latent sources projects into the sensor space differently, resulting in a characteristic spatial pattern, often represented as topographies in the sensor array in columns of $A$ (Figure 3.2). Due to the linear properties of such spatial mixing and dense spatial and temporal sampling of MEG, it is often possible to separate signals originating from different sources by utilizing local spatiotemporal correlations (Cardoso, 1998).

**Figure 3.2.** a. Directed acyclic graph representation of the generative model for MEG signal. b. Spatial topography and time course representation of the latent sources.

Time courses of each of $k$ sources represented in the rows of $S$ (Figure 3.2). These $k$ sources evolve in time in a structured way and, generally speaking, may or may not be statistically dependent.

If the number of sources $k$ is greater than the number of measuring channels $n$ or the true rank of $X$, Eq. 3.1 becomes an ill-posed problem and requires constraining the solution space (see Section 3.1.5) to estimate the underlying sources.

Generally speaking, existing modeling approaches can be classified based on the assumptions that they impose on $A, S, k$, or $\epsilon$. For example, if $k$ is smaller than the number of channels, assuming rows of $S$ orthogo-
nal leads to principle-component analysis (PCA) based (Hotelling, 1933; Uusitalo and Ilmoniemi, 1997) approaches, while a weaker statistical independence assumption on rows of $S$ results in an ICA solution (Onton et al., 2006; Hyvärinen, 2013). Similarly, if columns of $A$ are derived from an anatomically-informed biophysical model and each column of $S$ is constrained to have a minimum norm at each time instance $t$ results in Minimum-Norm Estimate (MNE) family of inverse solutions (Hämäläinen and Ilmoniemi, 1994; Lin et al., 2006; Hauk et al., 2011). We review these methods in detail in the following section.

3.1.5 Analysis of the MEG data

The generative model of MEG and EEG signals suggests specific spatial (e.g., topography in the sensor space) and temporal (e.g., frequency) structure of these observations. Because neural sources of M/EEG cannot be observed directly, they have to be inferred statistically through model inversion techniques. Such inversion is fundamentally an ill-posed problem since the number of sources, their location, amplitude, and other properties are unknown. Various assumptions and structural constraints have been introduced, resulting in various families of inverse solutions.

Traditionally, biophysical source modeling (Mosher et al., 1999) is used to introduce anatomical constraints onto the inversion methods with assumptions including a fixed number of electrical-current dipoles (Mosher et al., 1992), or constraints to the overall current distribution over a pre-specified dipole grid (Hämäläinen and Ilmoniemi, 1994; Van Veen et al., 1997). Anatomical constraints are introduced using a well-defined forward model of propagation of a given source to the sensor space (Hämäläinen and Ilmoniemi, 1994; Mosher and Leahy, 1998). These techniques primarily focus on estimating neural sources generating the observed activity.

In contrast to the above, blind-source separation (BSS) based methods treat latent sources as statistical objects and introduce constraints on their statistical properties (e.g., independence, orthogonality, etc.) (Hyvärinen and Oja, 2000; Makeig et al., 1995). These methods model the measurement data as a spatial mixture of some number of latent statistical sources or components. Each of these sources projects into the sensor space by a so-called source mixing matrix. Such decomposition methods can often be regarded as simplified generative models of the signal without explicit anatomical constraints. In practice, these methods can be combined with biophysical modeling, and statistical sources are interpreted in terms of the underlying neural sources.

Applying machine learning techniques to infer brain states from the EEG and MEG signals has recently become an increasingly popular approach (Subasi and Gursoy, 2010; Blankertz et al., 2011; Dahne et al., 2015; Kaplan et al., 2015; Woo et al., 2017; Sabbagh et al., 2020). These
methods treat experimental data as a supervised discriminative learning problem (see Section 3.4.2), where the goal is to predict an experimental condition encoded by a target variable $y$ from the measurement data $X$ either by modeling conditional probability density $p(y|X)$ or approximating discriminant or regression functions directly.

Most of the work in this field focuses on linear methods that allow relatively straightforward interpretation (Haufe et al., 2014). Often such linear discrimination is preceded with various feature extraction techniques, such as (time-)frequency transforms, univariate feature selection, scaling, connectivity estimates, etc. (see e.g. Herman et al., 2008; Omedes et al., 2013; Zhang et al., 2019). Of particular note is the use of supervised spatial filters that follow the logic of the generative model (Blankertz et al., 2008; Nikulin et al., 2011; Dähne et al., 2014; Halme and Parkkonen, 2016).

Finally, deep learning methods have recently gained popularity in the neuroimaging field (Schirrmeister et al., 2017; Lawhern et al., 2018; Roy et al., 2019). One of the most prominent advantages of deep learning approaches is combining feature extraction and discriminative predictive modeling within a single computational framework, allowing the model to learn optimal feature representations. Moreover, some deep learning models (e.g., convolutional neural networks, CNNs) can capture local statistics and are robust to certain invariances in the data (Bengio, 1997). As we discuss in Sections 3.4.2 and 4.3 due to their high complexity, unambiguous interpretation of the deep learning models is generally hard and is an area of active ongoing research.

In the following section, we focus on describing relevant concepts in machine learning and discussing these methods’ application to the MEG data in Section 3.5.

### 3.2 Statistical Inference

Statistical inference can be described as a process of deducing the properties of some unobserved data-generating process by observing samples of data drawn from it (Mackay, 1995). The end goal of statistical inference is often to estimate properties of a broader population by assuming that the observed data set is sampled from the same data-generating process as the general population (McCullagh, 2002).

The data-generating process is represented in the form of a statistical model comprising a set of assumptions describing how the sample data is generated as well as the relationships between different variables in the data.

In practice, it is not often possible (or even useful) to fully describe the natural phenomena with an elaborate statistical model. Instead, the goal of the analysis is often to learn about the regularities in the data-
generating process as well as the associated variability. In this sense, according to an aphorism attributed to George Box, "all models are wrong (but some are useful)" and any inference process includes a degree of uncertainty (Mackay, 1995). Such uncertainty can arise because observations are finite and often limited. Such uncertainty "due to the lack of knowledge" is referred to as epistemic (O’Hagan, 2004). Alternatively, uncertainty can arise from, e.g., a noisy measurement process or be inherent to the data-generating process itself; in this case, it is referred to as aleatory uncertainty. One can argue that the process of scientific discovery is concerned with reducing the epistemic uncertainty while adequately estimating the aleatory uncertainty.

3.2.1 Uncertainty and probability

Probability theory provides a consistent framework to deal with the uncertainty in a systematic way (Barber, 2011; Bishop, 2014). Probabilities can be used to describe, e.g., frequencies of outcomes in random experiments or, more broadly, to describe 'degrees of belief' in propositions that do not necessarily involve randomness. The latter view is known as the Bayesian approach (Bishop, 2014).

In this framework, the observations \(X\) are treated as random variables. Whenever dealing with different sets of variables, e.g., \(X\) and \(y\), the relationships between them are treated as non-deterministic in terms of probability distributions.

In a typical inference problem, we are interested in specifying a probability distribution generating the data \(D\). We can specify our hypothesis \(H\) as a family of parametric probability distributions with some unknown parameters \(\theta\). Thus, the model will have the form

\[
p(D, \theta | H) = p(D | \theta, H)p(\theta | H)
\]

(3.2)

, where \(p(\theta | H)\) referred to as the prior distribution represents our initial guess on the distribution of \(\theta\) and \(p(D | \theta, H)\) is the likelihood function, describing how likely it is to observe \(D\) for each given value of \(\theta\).

3.2.2 Modelling assumptions and hypothesis space

The process of inference is conditional on the modeling assumptions. As discussed above, given a set of modeling assumptions \(H\), the inference problem is to estimate \(\theta\) that are likely to produce \(D\). One common assumption in the context of a machine learning task is that all the samples are drawn from the general population (i.e., the same probability distribution) and are statistically independent of each other. This \(iid\) (independent and identically distributed) assumption makes it possible to describe the data-generating process with a probability distribution over a single exam-
ple. The iid assumption is crucial for training and testing out-of-sample generalization of the model (Goodfellow et al., 2016). Other assumptions are introduced to constrain the nature of interactions between the variables (e.g., conditional independence assumptions) or constrain the set of functions that the model can represent (e.g., linearity). These assumptions are often based on the structural knowledge of the domain, such as known causal relationships between the variables or other physical properties of the studied phenomena (Rudin, 2018). Knowing these properties often allows us to constrain the set of possible solutions and make the inference problem more tractable (Barber, 2011).

The Bayesian approach to probability usually requires to formulate assumptions explicitly in the form of the prior probability distribution on the model parameters. After the model observed the data, we can use Bayes theorem to estimate the posterior distribution $p(\theta|D, H)$.

$$p(\theta|D, H) = \frac{p(D|\theta, H)p(\theta|H)}{p(D|H)}$$ (3.3)

Where $p(D|\theta, H)$ is referred to as the data likelihood function evaluated for the observed data set $D$, estimating how likely the observed data set is to arise from different values of $\theta$. The term in the denominator $P(D|H)$ is called a marginal likelihood or model evidence. Collectively $H$ and $\theta$ comprise a parametric statistical model of the process generating $D$.

### 3.2.3 Maximum Likelihood Principle

Training machine learning algorithms involves adjusting $\theta$ in a way that maximizes the data likelihood function. Assuming that $m$ examples used to train the model are drawn from the true but unknown data-generating distribution $p(D)$, the maximum likelihood estimator for $\theta$ is defined as

$$\theta_{ML} = \arg \max_\theta p(D|\theta)$$ (3.4)

Maximizing the data likelihood is equivalent to minimizing the Kullback-Leibler divergence (KL-divergence), or the cross-entropy between the empirical distribution $p(D)$, defined by the training set, and the model distribution $p(D|\theta, H)$. Machine learning algorithms are often trained to minimize negative log-likelihood, referred to as the loss function. While the relation between the loss function and the underlying probability distribution associated with the likelihood function often remains implicit, there is often a computable mapping from one to the other (Saerens et al., 2002). For example, mean squared error is the cross-entropy between the empirical distribution and a Gaussian model. Similarly, binary cross-entropy is related to Bernoulli distribution. (Goodfellow et al., 2016).
In case of the supervised learning the maximum likelihood estimator is usually generalized to estimate a conditional probability $P(y|X, \theta)$. Thus,

$$\theta_{ML} = \arg\max_{\theta} p(y|X; \theta).$$ (3.5)

An iid assumption allows to decompose this problem into a sum of likelihoods for each sample

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p(y_i|X_i; \theta).$$ (3.6)

If the model has sufficient capacity to approximate the actual empirical distribution $p(D)$ and such a distribution corresponds to a single value of $\theta$, the maximum likelihood estimate is shown to converge to this true value as the number of training samples $m$ approaches infinity (Goodfellow et al., 2016).

### 3.2.4 Generative and Discriminative learning

In the most general case, a supervised machine learning problem can be expressed by a deterministic mapping $y = f(x)$, where the underlying function $f$ is unknown and learned from the data. Alternatively, we can formulate the same problem in terms of probability distributions. Consider a neuroimaging dataset $D$ comprising the measurement data $X$ and target variables $y$ is sampled from an unknown joint probability distribution $p(X,y)$. The model observes several realizations of a random vector $x$ and an associated target variable $y$, and then attempt to predict new values of $y$ from new samples $x$, usually by estimating the conditional mean of $p(y|x)$ (Bishop, 2014).

Two general approaches to solving these problems can be identified.

**Generative approaches**

Generative approaches rely on a specifying fully-structured joint probability distribution over the samples and the corresponding target variables. Joint probability density function $p(x, y)$ is defined by building a statistical model of the data-generating process and deriving conditional probability densities of interest (e.g. $p(y|x)$) from it (Bishop, 2014).

A key feature of generative models is that they allow modeling distribution of inputs ($x$) (explicitly or implicitly) as well as outputs ($y$). Because of this property, it is often possible to encode prior knowledge about the structure of the problem or to generate synthetic data points in the input space. Consequently, generative methods are relatively easy to interpret in terms of the data-generating process (e.g., parameters of the measured brain activity) (Bishop and Lasserre, 2007).
**Discriminative methods**

Discriminative methods attempt to infer the posterior conditional density $p(y|x)$ and estimating the conditional mean according to the learning objective.

In practice, however, $p(y|x)$ is rarely inferred explicitly, and an algorithm approximates only conditional mean by mapping inputs $x$ deterministically onto corresponding the $y$ from the training data.

### 3.3 Machine learning

Machine learning is a discipline that studies how computer programs optimize their performance in a given task without being explicitly programmed i.e. relying on regularities occurring in the data and statistical inference (Goodfellow et al., 2016).

Unlike classical statistics that focuses on describing the data-generating process in terms of the underlying probability distribution, machine learning often is primarily concerned with obtaining accurate out-of-sample predictions, i.e., how well does the algorithm perform on new, previously unseen inputs. Such generalization performance can be estimated by computing the expected value of the error on a new input, that is separate from the data used for training.

Implementing a machine learning method to solve a practical problem can be split into several major parts. First of all, one needs to define the problem and the goal that an algorithm needs to achieve by specifying the objective function and metrics used to evaluate model performance. Having defined the problem, one needs to specify the model architecture, i.e., define the computational graph, i.e., the type and the sequence of computations that the model can apply to the inputs to obtain the mapping onto the desired outputs. Finally, one needs to specify the optimization strategy by choosing the optimization approach, performance evaluation procedure, and regularization. In this section, we will briefly discuss key concepts in machine learning following this basic structure.

#### 3.3.1 Objective function

An objective function is a single, overall measure of loss incurred in taking any of the available decisions. In a supervised learning context, such loss represents the distance between the model prediction and the ground truth target variable. During model training, partial derivatives of the objective function with respect to all the adaptive parameters are used to adjust the parameter values and thus minimize the total expected loss (Goodfellow et al., 2016).

Selecting an appropriate objective function is one of the crucial choices
Background

Figure 3.3. Effect of the objective function on the prediction of the finger trajectory in the Berlin BCI competition dataset. MAE (mean absolute error) results in a few large errors. MSE (mean squared error) results in many smaller errors. Combining the two cost functions results in a more accurate prediction when designing a machine learning model. Depending on the application domain objective function can represent our assumptions regarding the statistical properties of noise in the data or the distribution of the target variables. For example, consider a regression problem with a mean-squared error objective function. Such a model will penalize the system more if it frequently makes medium-sized mistakes. Alternatively, mean-absolute error objective will weight rare large mistakes more (3.3). These kinds of design choices depend on the application and the desired behavior of the system.

3.3.2 Computational graph

A computational graph (see, e.g., Figure 3.4) is a convenient way to represent complex operations occurring, e.g., in an artificial neural network. In such a representation, each node is a mathematical operation performed on inputs. Nodes transform their inputs and feed their outputs into other nodes connected via edges.

Depending on the architecture and the types of operations specified in the computational graph, a machine learning model can be tuned to learn specific input features.

Representational capacity

The representational capacity of a model characterizes the complexity of functions, which it can approximate. In general, machine learning algorithms perform best when their capacity is appropriate for the true
complexity of the task (Goodfellow et al., 2016).

The correspondence between the model capacity and the complexity of the learning problem affects the performance of the model on the training and test sets differently. Models with insufficient capacity tend to have a higher error rate on the training set. This situation is known as underfitting. Alternatively, if the capacity of the model exceeds that of the true data distribution, the model performs well on the training set but fails to generalize to new, unseen samples. Such high generalization error on the test set indicates that the representational capacity of the model is so large that it not only fits both the patterns and the noise. This situation is referred to as overfitting.

Statistical learning theory posits that the discrepancy between training error and generalization error is bounded from above by a quantity that grows as the model capacity grows but shrinks as the number of training examples increases (Vapnik, 2000). Besides the trade-off between the input dimensionality and the training set size, or, samples-to-variables ratio, expected performance of a given model also depends on the following factors (Bzdok et al., 2020):

- Proportion of the informative variables or dimensions in the data
- Known redundancies and sources of variability
- Signal-to-noise ratio
- Model violations

Thus, model capacity is affected not only by a number of trainable parameters but also by the complexity of functions it can fit and the capabilities of the optimization algorithm. Thus, even if the capacity of the model is adequate to the complexity of the data, the model can still overfit if, e.g., it converges to a local minimum, as is often the case with maximum likelihood estimation for complex distributions (Goodfellow et al., 2016).

### 3.3.3 Neural Networks and deep learning

Artificial neural networks are a particular type of machine learning methods that use basis functions so that each basis function is itself a nonlinear function of a linear combination of the inputs. Eq. (3.7) describes a typical computation performed at a single node of the network.

\[
h_i = f(w_i x_i + b) \tag{3.7}
\]

where, \( w \) and \( b \) are adaptive parameters and \( f \) is a non-linear activation of \( i \)-th node. Nodes are organized in layers as shown in Figure 3.4 such that
outputs of the nodes in \( j \)-th layer serve as inputs for layer \( j + 1 \) (Bishop, 2014). Layers between the input and the output layers are referred to as hidden layers. Neural networks with more than one hidden layer are known as deep neural networks (DNNs).

\[
h_i^2 = f(w_{ij}^1 x_j + b_{ij}^1) \quad y_i = f(w_{ij}^2 h_i^2 + b_{ij}^2)
\]

\[
h^2 = f(w^T x + b) \quad y = f(w^T h + b)
\]

**Figure 3.4.** Feedforward neural network architecture

When a neural network is initialized, adaptive coefficients \( w \), are initialized at random by drawing from, e.g., a uniform distribution to provide the initial state of the networks and ensure efficient training. Bias coefficients \( b \) are typically initialized by setting them to some constant non-zero values. Initial predictions \( \hat{y} \) are made by computing Eq. 3.7 for each layer successively, thus, propagating the input information forward through the network. The result of this forward propagation step is then matched to the ground truth target variable \( y \) using an objective function \( J \). After that, the obtained error term is propagated backward through the network to adjust the weights so that the error function is minimized (see Section 3.3.5).

### 3.3.4 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a specific type of feedforward neural networks that use convolution in at least one of their layers (LeCun et al., 2010). A convolution is an operation on two functions of a real-valued argument: the input and the convolution kernel. Discrete convolution can be viewed as a multiplication of the input by a sparse matrix, which has
several entries constrained to be equal to other entries. For example, for univariate discrete convolution, each row of the matrix is constrained to be equal to the row above shifted by one element.

Because kernels are often regarded as feature extraction filters, the output of this operation is referred to as a feature map. The kernel is usually a multidimensional array of parameters that are adapted by the optimization algorithm.

A typical layer of a convolutional network consists of three stages. In the first stage, a layer performs convolutions to produce a set of feature maps. In the second stage, each feature map is run through a nonlinear activation function, such as, e.g., the rectified linear activation function (ReLU) (Glorot et al., 2011). In the third stage, the pooling function is applied to downsample the output of the layer and introduce invariance to local translations in the input (Figure 3.5).

CNNs possess three important properties making them a particularly suitable approach for application to neuroimaging data: sparsity, parameter sharing and invariance to local translations.

![Convolutional neural network architecture](image)

**Figure 3.5.** Convolutional neural network architecture

**Parameter sharing**

Unlike traditional neural networks, which use separate parameters for each feature of the input, CNNs apply the same convolution kernel to all features of the input by repeatedly translating the kernel along one or more dimensions. Thus, parameter sharing allows detecting stereotypical patterns that may occur at different locations in the input space. CNN's were initially developed in the image processing domain where the
same patterns can occur at any location in the image and have been very successful on other data with an underlying grid-like structure.

Besides extracting such local features, the convolutional layer reduces the number of adaptive parameters in the network compared to traditional feedforward neural networks without sacrificing the representative capacity.

**Translation equivariance and invariance**

The use of shared convolution kernels results in a so-called translation equivariance of the representation, meaning that translation of the input produces a similar translation of the output. When applied to MEG data, this means that a trained convolution kernel can detect, for example, evoked response even if they occur at varying latency after the reference event. Thus, convolution can be viewed as a pattern-detection operation by template matching.

Combining convolution with a pooling operation can introduce invariance to local translation within the pooling step and also downsample the output. A pooling function replaces the output of the preceding layer with a summary statistic of the nearby outputs. For example, the max-pooling (Zhou and Chellappa, 1988) operation reports the maximum output within a rectangular neighborhood. Other popular pooling functions include the average, $l_2$ norm, or a weighted average based on the distance from the central pixel (Boureau et al., 2010).

### 3.3.5 Optimization

Most machine learning algorithms are trained via an iterative procedure minimizing the objective function. This is typically done in two stages. First, the derivatives of the objective function with respect to the adjustable parameters are computed. Because objective function essentially computes an error term between the prediction and the ground truth, the error is said to propagate backward through the computational graph. Thus, the evaluation of derivatives of the objective function is referred to as backpropagation. In the second stage, these derivatives are used to adjust the adaptive parameters in a way that minimizes the objective function (Eq. 3.8). The simplest such technique involves gradient descent (Rumelhart et al., 1986).

$$\theta_l := \theta_l - \alpha \frac{\partial J}{\partial \theta_l}$$  \hspace{1cm} (3.8)

where $\alpha$ is a constant referred to as the learning rate.

If successful, such an optimization scheme is known to converge towards the parameter values corresponding to their respective maximum likelihood estimates.

In practice, because optimizing the objective function is often a non-convex problem, the basic scheme described above might converge to the
local minima, which is a known property of the maximum likelihood estimation on complex distributions. In the context of neural networks, several optimization schemes were shown to perform considerably better than the basic gradient descent when dealing with local minima (Hinton et al., 2006; Bottou, 2010; Kingma and Ba, 2015).

Moreover, regularization techniques and empirical methods discussed in the next Section also address the local minima problem.

**Regularisation**

Inferring model parameters is often an underdetermined problem. As discussed above, this problem becomes more severe as the ratio between training set size and the model capacity decreases, or optimization involves a non-convex error function.

Regularization is a term describing a set of techniques of solving ill-posed problems by adding constraints, assumptions, or other information to the solution. Regularization is any modification to a learning algorithm intended to reduce its generalization error but not its training error (Goodfellow et al., 2016).

**Weight decay techniques**

One popular regularization approach is known as weight decay. It introduces the penalty term proportional to the total sum of model weights which is added to a loss function to favor solutions involving less non-zero parameters during training.

\[ J(\theta) = L(\theta) + \lambda \|\theta\|_2, \]  

(3.9)

where \( L \) is a negative log-likelihood of the data, \( \theta \) is a vector of all model parameters and \( \lambda \) is regularization coefficient. Thus, weight decay approaches introduce a trade-off between model complexity and the goodness of fit, making simpler solutions more preferable unless complex solutions fit the data significantly better. This trade-off is controlled with a regularization coefficient of \( \lambda \), which is often defined empirically.

The way in which the weight decay term is computed can have varying effects on the solution. Similarly to using different loss functions in Figure 3.3, using an l2-penalty (corresponding to the squared sum of the model weights) will force the model to prioritize solutions without very large weight coefficients. Alternatively, l1-penalty (corresponding to the sum of the absolute model weights) will prioritize sparse solutions with fewer but possibly larger non-zero weights.

From a Bayesian perspective, introducing a weight decay term to a loss function corresponds to imposing different prior distributions on the model parameters leading to a maximum aposteriori (MAP) solution. Thus, l2-norm corresponds to a Gaussian prior on \( \theta \) with zero mean and variance of \( \lambda^{-2} \), while l1-norm corresponds to a sparse Laplacian prior.
Dropout

The dropout regularization approach was designed to prevent complex co-adaptation of nodes of the artificial neural networks during their training. Each training iteration, a random portion of the model parameters is set to zero or "dropped out", allowing the model to use all its capacity and learn redundant representations of the patterns in the data. Such redundancy makes the inference process robust to the noise by combining multiple representations (Srivastava et al., 2014).

3.3.6 Probabilistic interpretation of learning outcomes

Neural networks are typically regarded as function approximation methods because the relations between the variables are deterministic. However, it is also useful to interpret the network outputs and in terms of the underlying (often implicit) probability model. In doing so, it is important to consider network architecture, non-linearity of the output layer, and the cost function. As the objective function is related to the likelihood function, maximum likelihood estimation is used during model training to determine the network parameters. Regularized maximum likelihood can be interpreted as a MAP (maximum a-posteriori) estimate in which the regularizer can be viewed as the logarithm of a prior parameter distribution. Saerens et al. (2002), demonstrate that "any reasonable cost function" has a computable transformation into MAP estimate. In the case of supervised learning, both MLE and MAP are point estimates of the conditional density \( p(y|X) \) and the functional form and the scale parameter of which are defined by the loss function.

Probabilistic interpretation of the output of a trained machine learning algorithm can be obtained if we consider it in conjunction with an objective function.

It is well known, for instance, that artificial neural networks, trained to minimize the mean square error produce an approximation of the expected value of the desired output conditional on the explanatory input variables if two conditions are met (Goodfellow et al., 2016):

- a global minimum of the objective function is achieved, and
- model’s capacity is sufficient to approximate the optimal estimator.

Similar results were obtained for classification problems, e.g., minimizing categorical cross-entropy objective leads to approximating the conditional expectation of the desired output.

Finally, Saerens et al. (2002) provided formal proof that given a "perfectly trained" model and "reasonable cost function", there always exists a computable transformation mapping the model output to the corresponding
3.4 MEG data as a specific application domain of machine learning methods

This section discusses the specific properties of the MEG data that differentiate it from traditional application domains of machine learning, such as, e.g., computer vision or natural language processing. These properties can be roughly divided into those related to 1) statistical properties of the data itself, 2) our epistemic uncertainty, or the lack of knowledge, about the underlying brain activity, and 3) the fact that these data were acquired as a result of a controlled experiment.

3.4.1 Statistical properties of the MEG data

As discussed in Section 3.1.3 event-related MEG data is acquired from an array of sensors (magnetometers or gradiometers) at a high temporal resolution (e.g., 1000Hz). These sensors sample the magnetic field produced by the coordinated electrical activity of cortical neurons. In a typical decoding study, short (typically up to 2 seconds) epochs of the MEG measurement centered around the event of interest combined with the corresponding label of this event are used as one training example. As already discussed in the context of the generative model of the MEG signals (Section 3.1.4), an event-related MEG epoch can be viewed as a multivariate time-series containing a mixture of spatial projections of an unknown number of simultaneously active neural sources. It is often assumed that only a small portion of these sources is relevant for the studied experimental task, while the rest is often regarded as background brain activity or, simply, noise. Because relevant and irrelevant components of the MEG signal can, in principle, share the same spatial and temporal statistical properties, the key task for a successful learning algorithm is separating these components efficiently. For the same reason, a typical assumption when analyzing single-trial MEG data is that the signal-to-noise ratio (SNR) is close to 1. Apart from the SNR and high dimensionality, MEG signals also display high inter-individual variability arising from differences in cortical anatomy as well as the current functional state of the brain across subjects within the same study (Olivetti et al., 2014; Varoquaux et al., 2017).

*Dimensionality*

As measurement technologies continue to evolve, larger and denser MEG sensor arrays become increasingly available, allowing to sample whole-head brain activity at finer spatiotemporal resolution. Inevitably the information content of these measurements increases. At the same time,
the higher resolution of the measurements does not translate into better insights directly. An increasing number of channels in MEG systems make it prohibitively time-consuming to conduct explorative analysis or use classical univariate techniques of analysis due to a massive multiple comparisons problem. Consequently, advances in measurement techniques stimulate a trend towards more powerful multivariate analysis methods.

While using more sensors provides a natural approach to increase the SNR (e.g., via spatial averaging), it also increases the dimensionality of the feature space, which in the simplest case, requires more complex models. Given the fact that the data collection remains relatively expensive (compared to, e.g., image processing domain) and is often limited by the duration of the experiment, such an increase in model complexity increases the risk of overfitting.

On the other hand, the effective representative capacity (Section 3.3.2) required to capture both the signal-of-interest and the "brain noise" is relatively low compared to, e.g., image processing where extremely complex models are required to extract high-level visual features. In EEG and MEG, such high complexity is often not required due to the limited sensitivity of the non-invasive measurement techniques. This calls for adequate dimensionality reduction techniques that can exploit the redundancies and local correlations efficiently, but increasing model complexity rarely leads to significant increases in performance.

In other words, effective dimensionality reduction (Lemm et al., 2005; Blankertz et al., 2008; Nikulin et al., 2011; Halme and Parkkonen, 2016) and source separation approaches (Taulu and Kajola, 2005; Taulu and Simola, 2006) may hold greater potential to advance the field of decoding brain signals non-invasively than extracting more complex features. At the same time, invasive measurements of the brain activity like multi-unit activity or electrocorticogram (ECoG) have richer information content and thus can benefit from applying more complex models. Regardless of the measurement technique, however, the principle remains the same - optimal decoding methods should be informed by an adequate generative model if they are to advance our knowledge.

**Signal-to-noise ratio**

Low SNR in EEG and MEG measurements is naturally related to the fact that the major source of the "noise" originates from the brain activity irrelevant to the decoding task. While sensor noise, physiological artifacts, and external electromagnetic interference also contribute to the problem, there are efficient approaches for separating those components from the brain signals based on their statistical properties (Uusitalo and Ilmoniemi, 1997; Taulu and Kajola, 2005; Taulu and Simola, 2006; Taulu and Hari, 2009; Zeman et al., 2007; Haumann et al., 2016). One complication arises from the fact the background brain activity can, in principle, have the same
Background

spatial (e.g., smoothness, locality, etc.) or temporal (e.g., frequency content) statistical properties as the signal-of-interest, making it extremely difficult to tell them apart (Huang et al., 2014; Puce and Hämäläinen, 2017). There is generally no reason to believe that the statistical properties of such "brain noise" are significantly different from those of the "signal of interest." As experimenters, we are often interested in stereotypical patterns of brain activity (e.g., evoked responses, see Section 3.1.3) arising consistently under different experimental conditions. Based on decades of previous brain research and signal analysis, we know that the signal of interest can be consistent either in latency, spatial topography, or confined to a certain frequency band. Thus, to extract the signal of interest efficiently, it seems reasonable to utilize this consistency, combined with prior information about the spatiotemporal properties of the signal of interest, as well as the known sources of variability in the signal.

Epistemic uncertainty about brain activity

Besides the fact that both the signal and the "brain noise" may share spatiotemporal properties, the ground truth model of these properties is often only available in the form of the experimental hypothesis and cannot be immediately observed. These hypothetical properties are often of primary interest to the researcher. This leads to a situation where it is difficult to give an unambiguous interpretation of the negative decoding results. Are we unable to achieve sufficient decoding performance because our algorithm is inadequate to the problem, or our experimental design fails to produce the expected brain responses reliably? Similar ambiguity can also be present when interpreting positive findings: is our algorithm performing well because our hypothesis is correct, or are these effects due to confounding factors that we failed to control with our experimental design? Either way, it is evident that the use of machine learning methods requires imposing specific constraints into designing the experiments in order to be able to interpret the findings in a way that extends our knowledge about the neural processes under study.

Experimental design and data collection procedure

Neuroimaging datasets are typically generated in a well-controlled experimental, designed to test a specific hypothesis about the brain function. Although other data sets exist and are becoming increasingly popular, here we focus specifically on the experiments, where the subjects' brain activity is measured while they are presented with a set of predefined stimuli or are performing a specific task. These stimuli or task conditions are typically encoded in a target variable ($y$). Such a target variable can be continuous (e.g., stimulus intensity), leading to a regression problem or discrete classes of stimuli or movements (e.g., movements performed with right or left hand) resulting in a classification problem. In more complex
Background experiments, \( y \) can also be a multidimensional vector combining different types of variables.

Naturally, this leads to supervised learning problems where we want to predict the type of stimuli or response from the brain data. Such tasks are formulated as classification problems and particularly popular in the context of brain-computer interfaces (Section 3.5.2). Alternatively, a researcher can be interested in predicting a continuous variable (e.g., the force of muscle contraction) from the brain measurements leading to regression problems (Sabbagh et al., 2020).

Since \( y \) is a product of the experimental design, we can note several interesting properties. First, \( p(y) \) is known and describes the complete set of studied conditions within a given experiment. Background brain activity is either assumed to be the same across the studied conditions or modeled explicitly as the "rest" or "null" condition. Second, because \( y \) is a product of the experimental hypothesis, the ground truth relations between \( X \) and \( y \) are often not known, and the extent to which information about \( y \) is represented in \( X \) is of primary interest.

The goal of an experimental design aims to establish validity (correspondence to the real-world phenomena), reliability (e.g., precision and consistency), and replicability (e.g., likelihood of obtaining the same outcome under similar conditions) of the findings. When applied to neuroimaging data sets, machine learning methods probe a hypothesis that the information encoded in the target variable can be decoded from the measurements. Consequently, the predictive performance of the algorithm on the held-out test set is often considered as a combined measure of all three criteria. Whether the predictive performance of the algorithm is significantly better than guessing at random can be identified, e.g., using a permutation test (Ojala and Garriga, 2009). In this procedure, an association between the data \( X \) and the target variables \( y \) is randomized to obtain an empirical distribution of the prediction results under the null hypothesis. In practice, however, having a "statistically significant" predictive performance can only be interpreted that the information about \( y \) is present in the signal. Further insights can be gained by comparing the performance of several algorithms on the same dataset (Roy et al., 2019). If any given model outperforms other methods significantly (e.g., based on a statistical test comparing the performance of different models), it might serve as evidence that its representational capacity and modeling assumptions are the closest to the true capacity of the data generating process. One problem interpreting these findings is how to attribute high predictive performance to the brain activity and not to other signal components, e.g., oculomotor artifacts that can appear consistently in some experimental settings and thus inform the algorithm. The risk here is that such correlated activity can yield high predictive performance but be completely unrelated to the studied neural phenomena (Hagemann and Naumann, 2001; Whitham
et al., 2008; Ma et al., 2012; Carl et al., 2012). Thus, designing a study applying machine learning techniques requires additional care in order to control such possible confounds.

### 3.4.2 Encoding and decoding approaches

As already discussed above, applying machine learning methods to the neuroimaging data is most typically defined as a supervised discriminative learning problem, as defined in Section 3.2.4. Given a set of measurements $X$ and the associated target variables $y$, encoding the experimental manipulations, we seek to estimate whether information about $y$ can be extracted from $X$. This is achieved by modeling an expectation of posterior conditional density $p(y|X)$. Alternatively, we can be interested in components of $X$ that vary in association with changes in $y$ by estimating $p(X|y)$. In the neuroimaging literature, these approaches are known as decoding and encoding methods, respectively.

**Encoding approach**

A typical way to study differences in brain responses to different types of stimuli is to produce an experimental design $y$ based on the hypothesis $H$ summarised by $p(y|H)$. Ideally, such design would elucidate meaningful differences in the conditional distribution of the data $p(X|y, H)$ that could be estimated with various hypothesis testing techniques (Friston et al., 1994; Delorme and Makeig, 2004). This is known as encoding or forward modeling approach (Haufe et al., 2014; Weichwald et al., 2015; Kriegeskorte and Douglas, 2019).

Encoding models seek to explain how the measurements $X$ are generated by describing the relations between the observed variables as functions of some latent (i.e., not directly observable) factors $y$. An encoding model predicts the response at each measurement channel at a given time point given the experimental condition by modeling the conditional density $p(X|y, H)$. Consequently, encoding models allow identifying features that display significant association with the target variable.

Due to the fact that neuroimaging data is usually produced in a controlled experiment, the prior distribution of the target variable $p(y|H)$ is known by design. In fact, successful experiments are designed in a way to simplify the inference process. For example, when designing an evoked response experiment, it is often desirable to have an equal number of repetitions of each type of stimuli to avoid dealing with problems arising from class imbalance during the analysis. Thus, because $p(y|H)$ is known by design, factorising $p(X, y|H) = p(X|y, H)p(y|H)$ is tractable, making encoding models equivalent to generative models in practice (Haufe et al., 2014). In principle, encoding models can also be used to estimate conditional density $p(y|X)$; however, they are often outperformed by discriminative models.
(Bishop and Lasserre, 2007). Because generative models primarily focus on estimating \( p(X, y|H) \) and are sensitive to violations of modelling assumptions \( H \). Thus, there is generally no reason to believe that the "encoding" factorisation \( p(X, y) = p(X|y, H)p(y|H) \) above can result in optimally separated conditional densities \( p(X|y) \). In other words, modelling the joint density \( p(X, y) \) generally does not guarantee an informative factorization in class-conditional densities \( p(X|y) \) (Jordan et al., 1999).

Encoding approaches employ null-hypothesis significance testing to quantify the likelihood of the data under the (null-)hypothesis of no association between \( y \) and \( X \). This can easily lead to a massive multiple comparisons problem, as the change of false discovery increases proportionally to the dimensionality of \( X \) (Nichols and Holmes, 2002; Maris and Oostenveld, 2007; Helwig, 2019; Alberton et al., 2020). A subset of features in \( X \), for which the null-hypothesis is rejected, are considered relevant for the encoding model in this experimental paradigm (Weichwald et al., 2015; Kriegeskorte and Douglas, 2019).

**Decoding approach**

By contrast, a decoding model seeks to obtain a conditional probability of the target variables \( y \), given the observations \( X \). Decoding models invert the data-generating process by extracting patterns associated with the variations in the target variable \( y \). Thus, decoding models focus on estimating \( p(y|X) \), without modelling the \( p(X) \) or \( P(X, y) \) explicitly.

Decoding methods seek to obtain a representation optimizing the discrimination between the experimental conditions, or, more broadly, focus on estimating the conditional mean of \( p(y|X) \). As such, they are constrained to supervised (discriminative) learning problems (Bishop, 2014; Kriegeskorte and Douglas, 2019).

The performance of decoding models is estimated based on their ability to accurately predict target variables \( y \) from previously unseen data \( X \). Such focus on empirical tests of out-of-sample generalization is a key advantage of decoding approaches since it allows testing several models on the same data and employ model selection techniques to make inferences regarding the data-generating process.

Decoding methods employ multivariate techniques to extract latent factors associated with the target variable from the observed data (Haufe et al., 2014). This is achieved by amplifying a signal of interest and suppressing all "signals of no interest." Because of that, features relevant for a decoding model can correspond to sources of signal and the noise alike, making the interpretation of decoding features less straightforward because these models can attribute large weights to the activity of which can be statistically unrelated to the process under study. Thus, interpreting the parameters of decoding models requires additional steps to obtain an approximation of the corresponding forward model (Haufe et al., 2014).
Interpretation of encoding and decoding methods

Since the goal of applying machine learning methods to the neuroimaging data is often to advance our understanding of the underlying neural processes, we are often interested in interpreting the features relevant for each approach.

Encoding and decoding methods provide different views of the data, and their usefulness interpretation depends on the goals of the analysis.

Because encoding models model the data-generating process, their parameters can be interpreted in terms of the data-generating (in our case neurophysiological) process in a straightforward way. Thus, encoding approaches are preferable whenever the goal is to obtain an insight into specific observed variables are affected by variations in target variable (Bzdok and Ioannidis, 2019). The relevance of features in an encoding model can be translated into conditional independence (or, more formally, d-separation) statements that provide insights into the causal structure of the data-generating process (Weichwald et al., 2015). By contrast, decoding models integrate all available data into a single performance metric allowing to test if the information about the target variable is encoded in the observed data. Thus, decoding models are preferred whenever the goal of the analysis is capturing patterns that allow future predictions (Bzdok and Ioannidis, 2019). Yet, features relevant for such predictions are not necessarily associated with the data-generating process (Haufe et al., 2014). To interpret the patterns learned by a decoding model, their parameters need to be mapped back onto the input space to obtain corresponding activation patterns (Haufe et al., 2014; Kindermans et al., 2018). Patterns obtained by these procedures can be used as an approximation of the underlying generative process but cannot substitute encoding analysis fully. Nevertheless, patterns informing a successful decoding model can still provide useful insights for a quick explorative analysis aimed at guiding and motivating further more refined experiments. The relevance of features identified by a decoding model to the data generating process can be estimated by removing each feature recursively and monitoring how it affects the predictive performance (Breiman, 2001).

To summarize, decoding analysis allows answering the following questions:

- is the information about the target variable encoded in any combination of patterns observed in the data;
- when comparing several decoding approaches: which model allows capturing these regularities more efficiently;
- how well does the model generalize across subjects or different populations;
As follows from Figure 3.6 only encoding models in a stimulus-based setting support unambiguous causal statements. However, using both encoding and decoding approaches combined can provide further insights into the causal structure of the data. Particularly combining the two approaches can be useful to identify the features relevant in only one type of model. Figure 3.7 summarizes the findings of Weichwald et al. (2015), demonstrating that, e.g., features only relevant in encoding can not be regarded to have a direct causal relationship with target variables while features only relevant for decoding can only provide “brain state context” w.r.t. to the target (Weichwald et al., 2015).

Bzdok et al. (2020) provides an example of how applying both approaches to the same data can yield novel insights. Features identified as significant (p < 0.05) by a mass-univariate test did not always yield high predictive
performance in the out-of-sample. By contrast, features relevant for prediction typically coincided with significant p-values in almost all cases (Bzdok et al., 2020).

3.5 Applications of machine learning methods in neuroimaging

3.5.1 Knowledge discovery

When working with neural data, we often seek not only to obtain an accurate predictive performance but also to discover regularities that allow the model to make these predictions, i.e., discover the relevant parameters of the data generating process.

Fayyad et al. (1996) define knowledge discovery as a "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al., 1996). Defined in this way, knowledge discovery in the neuroimaging domain can occur in a research context, leading to new insights into the functioning of the human brain as well as in a clinical setting, leading to the identification of new biomarkers associated with neurological or psychiatric conditions (Woo et al., 2017).

Validity refers to a degree of certainty that discovered patterns would also be present in similar data that was not used to train the model and thus guarantee accurate predictions. Measures of out-of-sample performance (such as validation accuracy, area under the ROC-curve, etc.) are typically used as an estimate of validity.

Usefulness and understandability are related to our ability to interpret the patterns discovered by the model in a way that leads to human insight and informs further useful actions (Rudin, 2018). In the context of neuroscience, interpretability means being able to translate patterns that the model extracts from the data into neurophysiological insights allowing to, e.g., adjust our assumptions about the neural activity under study. In practice, it could mean adjusting assumptions about the statistical properties of the signal of interest or noise, scientific hypothesis, experimental design, or measurement procedure. Finally, novelty means that patterns identified by the inference process were not known beforehand.

Recent advances in machine learning and especially deep learning make it very tempting to envision feeding the neuroimaging data into a complex algorithm and get back accurate predictions and useful insights. Yet, to date, the improvements in the predictive performance of complex deep neural networks over classical methods have been quite limited (Schirrmeister et al., 2017; Kawahara et al., 2017; Vieira et al., 2017; Lawhern et al., 2018; He et al., 2018; Zubarev et al., 2019) and our ability to interpret the learning outcomes of these methods decreases proportionally to increases
in model complexity.

Indeed, deep learning methods achieved remarkable performance by making feature extraction an integral part of the learning process. As follows from the no-free-lunch theorem, however, no single machine learning model is capable of achieving superior performance in any conceivable task (Wolpert and Macready, 1997). Here, we recall that in order to perform optimally, the representational capacity of the learning algorithm should be proportional to that of the informative patterns in the data. It seems reasonable that for any given application domain, a model should be designed to capture regularities inherent to this domain.

**Figure 3.8.** Knowledge discovery cycle necessarily involves interpretation of the findings. Future studies are informed by the previously obtained findings.

Because different algorithms capture these regularities in different ways, it is often the case that several models can produce reasonably accurate predictions on the same data. At the same time, patterns that such close-to-optimal models can learn from the data may vary considerably. Such a set of models, known as Rashomon set, could thus contain functions that can be approximated well by simpler functions, which are also a part of this set. Rudin (2018) expresses this argument as follows "uncertainty arising from the [finite amount of] data leads to a Rashomon set, a larger Rashomon set probably contains interpretable models, thus interpretable accurate models often exist" (Rudin, 2018).

Whenever non-proprietary algorithms are concerned, interpretability mainly depends on their complexity. Introducing structural constraints based on domain knowledge is an efficient way to reduce such complexity, allowing the data to vary in a predefined and meaningful way. This seems in line with the algorithmic information theory perspective (Chaitin-Gregory, 2006), arguing that understanding requires a degree of information compression. Thus, it seems reasonable that future decoding approaches should rely on keeping the model complexity limited. From a practical
point of view, this approach also has a number of advantages as simpler models require less training data and computational resources to perform efficiently, etc. High costs of acquiring neuroimaging data have always been one of the major limiting factors for applications of deep learning models. Today, large-scale normative data sets of functional brain data are becoming increasingly available, alleviating this problem. The goal of these massive projects is to identify biomarkers of various neurological conditions starting from healthy aging to potential early markers of neurological disorders (Niso et al., 2016; Taylor et al., 2017). Although these data sets create a unique opportunity to apply machine learning methods very efficiently, it also emphasizes the need for interpreting such findings and turning them into actionable information informing the public health policies. Yet, regardless of the dataset sizes, it is might still be reasonable to minimize the complexity of the models applied to these data from a pure knowledge discovery perspective.

3.5.2 Brain-Computer Interfaces

Brain-Computer interfaces (BCIs) are an emerging class of technologies allowing an individual to interact with the environment without any overt motor output (Wang et al., 2008; Lebedev et al., 2011; Nicolas-Alonso and Gomez-Gil, 2012; Sreedharan et al., 2013; Chaudhary et al., 2016). BCIs appeared holding promise for patients with severe motor disabilities to communicate or control external devices. Over the years, the potential domain of application of BCIs has expanded to rehabilitation, entertainment, and other domains (see Figure 3.10). It became apparent, however, that

Figure 3.9. Basic elements of Brain-computer interface (BCI) system. BCI design focuses on developing new applications and control strategies (stimuli, experimental protocols, etc.). Sensing technology research focuses on developing hardware allowing more precise and accurate measurements of the brain activity. Decoding algorithms research focuses on developing computational tools allowing to extract accurate information about the brain states from the measurements in real time.

a number of technological challenges need to be overcome to allow these non-invasive systems to be used outside of research laboratories.
We define a BCI is non-invasive if it is based on the measurements of the brain activity that do not require to disturb the integrity of the operator’s body with, e.g., surgical procedures. The examples of non-invasive BCI systems can be those based on EEG (Wolpaw et al., 1991; Lotte et al., 2007, 2018), MEG (Mellinger et al., 2007; Halme and Parkkonen, 2016), fMRI (Heunis et al., 2020) and fNIRS (Naseer and Hong, 2015). BCIs based on EEG are the most popular and widely used due to their relatively low cost and high temporal resolution. In this thesis, we focus on MEG-based systems for two major reasons: MEG possesses a higher spatial resolution and thus presents the cutting edge of the non-invasive systems in terms of the number of control signals that can be extracted from the measurements. Second, MEG sensor technology is still actively developing, holding promise to even further developments in signal quality and decreasing the costs. It remains to be seen, however, if such gains in signal quality will justify the increased equipment costs for MEG-based BCI systems to be widely adopted to clinical practice. Apart from the complications arising from the nature of the data obtained from the non-invasive measurements discussed in Section 3.4.1, it became apparent that non-invasive BCI systems have very few control signals that can be used reliably in real-world applications. At present most BCI systems are based on the following brain signals and control strategies:

- **Motor imagery**: Somatomotor \( \mu \)-rhythm modulation (Pfurtscheller et al., 2006; Halme and Parkkonen, 2018)

- **Steady-state evoked responses** to rhythmic stimulation. These systems are based on focusing an individual's attention on the concurrently presented streams of auditory, visual, or somatosensory stimuli.

![Figure 3.10. An estimate of current and prospective requirements in various applications of the BCI technology. Advances in measurement hardware and decoding methods expand the possible applications of the BCIs.](image-url)
- P300-based system. These BCIs focus on detecting an attention-related component of an evoked response.

- Slow cortical wave, also known as the readiness potential

Thus, the number of control signals used to control the non-invasive Brain–Computer Interface systems is rather small, limiting their practical applications to domains where the control can be limited to a relatively small number of commands and high misclassification rate, which are still quite likely to occur, will not have catastrophic consequences (Figure 3.10).

As brain-sensing technologies continue to evolve, the bandwidth of the BCI systems is likely to improve if the hardware improvements will be matched by progress in decoding methods.
4. Methods

This section introduces the machine-learning approaches used in publications comprising this Thesis. This section starts by describing how transfer learning techniques and generalization tests can be used to probe the similarity between brain responses observed across different experimental conditions. Next, we discuss how structural constraints from the generative model can be used to develop a compact and efficient convolutional neural network architecture that can perform decoding brain states from MEG signals efficiently, providing insights into patterns of the activity contributing to its performance. Finally, we discuss model interpretation techniques used in Publication I and Publication II and implemented in software described in Publication III.

4.1 Transfer learning

Transfer learning is defined as applying information learned in one task to a different yet related task. In the context of machine learning, and especially in the field of deep learning, transfer learning is often used for saving computational time and resources. In image processing, for example, one can use several layers from a network trained on a different image set for extracting low-level visual features and focus the training only on high-level abstract representations. From a modeling perspective, this means that we introduce a modeling assumption that local low-level features are sufficiently similar across the data sets. If the model is able to perform well, we could then conclude that such an assumption is useful and thus justified. Intuitively, this approach seems loosely-related to empirical generalization tests where we test the iid assumption by evaluating the performance of our model trained on the training set on the held-out test set. In transfer learning, however, this analogy holds only if the transferred parameters are not adjusted to the new data.
Temporal generalization method
A similar idea was used by King and Dehaene (2014) where the performance of a classifier trained on one time point of the MEG signal was evaluated on all other time-points of the event-related MEG epoch to obtain a temporal generalization map, allowing to track the emergence and maintaining of the same patterns across the time-resolved signals like EEG and MEG regardless of the exact timing of these events with regard to the reference event (e.g., the presentation of the stimulus).

Transfer across conditions
More generally, this approach can be extended beyond using different time samples from the same MEG dataset. Similarly, a classifier can be trained to, e.g., discriminate between two experimental conditions in one dataset and applied to identify the presence of similar patterns in another dataset. Such cross-condition generalization can identify the invariant patterns of the brain activity consistent across, e.g., imagery and perception or performance monitoring in motor and non-motor domains in Publication I.

Transfer across subjects
Empirical tests of generalization required for evaluating the predictive power of the decoding models can offer other valuable perspectives into the analysis of the neuroimaging data. In an across-subject design, the logic of transfer learning can be used to estimate the generalization of the decoding models across subjects. Traditional group analysis typically involves averaging brain responses for each test subject and computing group-level statistics. By contrast, techniques like leave-one-subject-out cross-validation (LOSO-CV) train a decoding model is trained on pooled unaveraged data from all participants except one, and then evaluated on the data of this held-out subject. This procedure is then repeated for each subject in the study, allowing not only to estimate the generalization with regard to each individual subject but also to use the information about within-subject variability in the estimation of the decision boundaries, potentially leading to more accurate estimates.

Figure 4.1. LOSO cross-validation allows assessing the across-subject generalization directly.
4.2 Generalization tests

As discussed in Section 3.2.2, machine learning methods typically rely on iid assumptions to approximate the data-generating process in terms of the probability distribution over a single sample. When evaluating the performance of a machine learning model on held-out data, we assume that these previously unseen data originate from the same distribution as the training set. Thus, high performance on the validation or test set indicates two things 1) that the model generalizes to the new data from drawn from the same underlying probability distribution, and 2) that the data is indeed drawn from the same probability distribution, i.e., the iid assumption holds.

Alternatively, if a machine learning algorithm trained on one dataset performs equally well on the data from a different dataset, this can indicate that the data-generating processes producing these data sets are similar to the degree that allows such generalization. Thus, we can formulate a null-hypothesis of no generalization between the two tasks that would correspond to the predictive performance for the model on the new data within the empirical chance level (Combrisson and Jerbi, 2015). Finally, permutation tests can be used to obtain the empirical distribution of the data under the null hypothesis and estimating how likely the observed performance is to arise, given no similarity between the tasks (Ojala and Garriga, 2009; Maris and Oostenveld, 2007).

In line with this reasoning, we use time-resolved across-condition generalization method (King and Dehaene, 2014) in Publication I to probe whether evoked responses following errors in the motor task have shared neural signatures observed after the presentation of negative feedback in learning tasks or errors committed when operating a brain-computer interface. In Publication I, a separate logistic regression classifier was trained to discriminate between correct vs. erroneous outcomes at each time-point of 500ms window following the reference event in condition A (e.g., speeded motor task). Each of these classifiers was then applied to each time point of condition B to discriminate correct vs. erroneous outcomes (e.g., correct vs. erroneous BCI control) to obtain a time-resolved generalization map 5.1. Clusters of time-point pairs where across-condition generalization scores were greater than the chance level with $p<.01$ were tested for significance using permutation tests. The Cluster p-values were computed as a probability to observe a cluster of larger positive mass over 10,000 random permutations.

Thus, in Publication I, combining generalization tests with standard hypothesis testing approaches allowed us to demonstrate quantitative evidence of similarity of neural responses arising under different experimental conditions.

In Publication II, we compare the across-subject generalization perfor-
mance of several machine learning models to probe how introducing different structural constraints allows extracting features to make decoding robust to inter-individual differences. Such design is based on the assumption that for each dataset, brain responses of different subjects share some statistical properties while others vary in a certain way. Thus, applying models with different structural constraints to the same dataset allows us to understand which of these constraints reflect the true underlying data-generating process better than others.

4.3 Convolutional neural networks for MEG data

In the previous section, we showed how training a separate classifier at each time-point addresses allows us to estimate generalization between the responses in two different experimental tasks even if these responses occur at different latency. In Publication II, we addressed the same problem using convolutional neural networks.

In section 3.3.4, we discussed how convolutional neural networks use parameter sharing to introduce invariance to local translations of the input. In the context of decoding MEG signals, such invariances to local translation in the time domain can be extremely useful because event-related brain responses may not always occur within the exact same latency with regard to the triggering event as, e.g., is the case with event-related modulations of ongoing oscillatory activity.

In Publication II, we develop a convolutional neural network whose architecture follows the known properties of the generative model of the MEG signal (see Section 3.1.4). Specifically, it combines linear spatial filtering with a convolutional layer operating across the time dimension. Introducing these structural constraints allowed us to reduce the number of trainable parameters and computational load, making this model suitable for experiments with relatively small sample sizes typical for most neuroimaging data sets and usable in real-time brain-computer interface experiments.

The spatial de-mixing layer in Figure 5.2 learns a set of spatial filters that use spatial correlations across sensors to extract time courses of latent statistical sources. This operation can be viewed as a dimensionality reduction via de-correlation across the spatial domain. Unlike unsupervised de-mixing techniques used for the same purpose, training these spatial filters is not based on additional constraints onto the time-courses of the resulting latent components. Instead, these filters are trained by backpropagation to optimize the global learning objective. In the context of classification tasks, this would mean that these filters optimize the extraction of spatial features allowing the discrimination between the classes.
The temporal convolution layer receives the time-courses of latent components extracted by the spatial de-mixing layer. By using a set of 1-dimensional convolution kernels, this layer learns to detect temporal patterns that may occur at different time-points across the time axis. Convolution is followed by a pooling operation that introduces invariance to local translations and reduces the feature dimensionality further. Apart from introducing invariance and dimensionality reduction, this layer can be seen as a temporal filter that learns to enhance the signal of interest in a specific frequency band. Similarly to standard FIR filters, the coefficients of the convolution kernels can be used to obtain their frequency response, which can provide some insight into spectral properties of the signal of interest. The temporal convolution layer uses rectified linear units to introduce an adaptive threshold to the output.

Spatial and temporal layers combined can be viewed as adaptive feature extraction operations optimizing the global training objective. The final fully-connected layer uses the reduced set of thus de-correlated features for sparse classification or regression by imposing the $l_1$ penalty onto its weights. This sparsity constraint forces the network to learn a minimum possible set of informative features for each class, which are more likely to correspond to the underlying sources. This, however, does not guarantee the straightforward interpretation of the extracted components in neurophysiological terms for the following reasons:

- The primary training objective is discriminative, and thus, the extracted features can equally correspond to either the underlying data-generating process or the noise;

- Because there are no constraints on statistical independence of the latent sources and the fact that the network is forced to use it’s representational capacity fully by dropout regularization the extracted features are most likely redundant and thus are non-unique;

- Due to this non-uniqueness, no single latent component can generally be interpreted as an informative feature (see Section 3.4.2);

These factors, however, do not render exploring these components completely useless. First, it can be useful for, e.g., quick exploratory analysis or to verify that the model performance is not driven by components that originate from, for example, oculomotor artifacts. Second, as discussed in Section 3.4.2 exploring features relevant for discrimination can provide additional insights when used in combination with encoding approaches. Finally, comparing the performance of several decoding models (e.g., using different activation functions or temporal invariances) can lead to useful, although indirect, insights into the statistical properties of the signal of
interest.
5. Summary of Publications

5.1 Publication I: Evidence for a general performance monitoring system in the human brain

In this paper, we use generalization tests (see Section 4.2) to probe the similarity of brain responses across different tasks and experimental conditions. Specifically, we addressed a long-lasting theoretical debate whether evoked responses to motor errors (Holroyd and Coles, 2002) are produced by the same neural system as the visual feedback communicating negative or unexpected outcomes in a learning task, or errors committed without an explicit motor output (e.g., when controlling a brain-computer interface).

We designed a learning task where the subjects used BCI to learn the values of the four abstract stimuli. Each trial subjects chose one of the two stimuli presented on the screen and observed feedback communicating whether they earned a point in this trial. Each stimulus had a predefined probability of generating a point. The subjects were instructed to maximize the number of points by identifying and choosing the most valuable stimuli over 360 learning trials. Once the learning was complete, the subjects were tested in a separate motor task involving the same stimuli. The analysis was focused on the generalization of logistic regression classifiers between 4 pairs of experimental conditions: 1) correct vs. erroneous BCI control, 2) positive vs. negative visual feedback, 3) expected vs. unexpected visual feedback, and 4) correct vs. erroneous motor output. A separate logistic regression classifier was trained on each time-point in one task (e.g., motor task) and applied to each time-point of another task (e.g., BCI control task) (King and Dehaene, 2014). Generalization performance across each two discrimination tasks was summarized by temporal generalization maps. Cluster-based permutation tests were used to identify time points where the across condition generalization was significantly above chance. The results indicated that a classifier trained to discriminate correct vs. erroneous motor responses generalizes successfully to discriminate evoked
5.2 Publication II: Adaptive neural network classifier for MEG signals

In this publication, we suggest a convolutional neural network (Section 4.3) that follows the general logic of the generative model of EEG-MEG signals and incorporates similar independence assumptions.

Specifically, we assume (1) neural sources informing the classification can be separated linearly in the sensor space, (2) the time-courses of these sources can be approximated reasonably well by an adaptive finite-impulse response (FIR) filter implemented as a 1-dimensional convolutional layer across time domain, and (3) the combination of spatial and temporal filtering results in a relatively small set of informative features justifying the sparse l1-penalty on the weights of the final output layer.

This leads to a compact network architecture comprising a spatial linear de-mixing layer, temporal convolution layer, and a single classification layer shown in Figure 5.2. This design results in a relatively small number of trainable parameters, which are grouped into the "latent components," allowing relatively simple interpretation in terms of classical analysis of...
MEG signals. These methods were implemented in MNEflow software and are discussed in the summary of Publication III.

We demonstrate that these models can outperform traditional machine-learning methods and more complex neural networks in terms of across-subject generalization on three different data sets: (1) 5-class dataset of multimodal sensory evoked responses, (2) 3-class motor imagery dataset, and (3) subset of 250 subjects taken from a Cam-CAN dataset. Moreover, in a separate experiment, we demonstrated that the model trained on the pooled data from 18 subjects could be used to control a BCI in real-time without a dedicated calibration session.

From a more general perspective, neural networks provide a flexible framework to probe the properties of the generative model by exploring how modifications to the network architecture affect their predictive performance. For example, in our experiments, we observed that, e.g., increasing the number of layers or adding a non-linearity into a spatial de-mixing layer did not improve the classification performance. These results suggest that the relatively low representational capacity of LF-CNN could be close to that of the data-generating process, at least for these data sets.

### 5.3 Publication III: MNEflow neural networks for MEG-EEG analysis

This paper introduces an open-source academic software designed to provide neuroimaging researchers with a set of tools to conveniently apply neural networks for classification, regression, and sequence prediction tasks involving EEG/MEG data. The software provides an implementation of several popular CNNs (Schirrmeister et al., 2017; Lawhern et al., 2018), including those developed in Publication II, as well as several domain-specific utilities for pre-processing the data, validating and interpreting the results.
In the present version of the software, interpretation of the model weights is only implemented for LF-CNN and includes several heuristic approaches to identify spatio-temporal properties of latent components associated with the variations in target variables based on three following heuristic approaches.

- **Recursive elimination** - ranks the components based on the effect that their removal has on the value of the loss function (Breiman, 2001).

- **Correlation with the target variable**, computed for each feature individually. For regression problems, feature relevance is ranked by their (absolute) Spearman correlation coefficients with the target variable. For classification problems, categorical cross-entropy is used as a distance metric.

- **Weight-based contributions**. Finally, feature relevance can be ranked by the value of the corresponding model weights.

By streamlining the workflows of classical MEG analysis and machine learning, the software aims at providing a convenient and time-efficient tool for applying neural networks to neuroimaging data. Moreover, MNEflow provides methods for inspecting and interpreting patterns that the networks learn from the data, which can be of particular interest to the neuroscience community.
6. Discussion

As neuroimaging technologies continue to evolve, the need for advanced multivariate approaches to analyze the functional neuroimaging data increases. Predictive machine learning methods hold great promise for practical applications of neuroimaging, particularly in identifying functional biomarkers of various neurological and psychiatric conditions and developing brain-computer interfaces. As data-sharing and standardization initiatives gain popularity in the field, large open data sets allow applying complex machine learning algorithms to these data. Indeed, machine learning techniques allow extracting complex patterns of the brain activity associated with studied neural phenomena from the measurements. Nevertheless, their application to EEG/MEG data compared to traditional methods has been limited to date. One reason for this could be that the increasing complexity of these methods, which is particularly true in deep neural networks, often complicates the interpretability of the learning outcomes (Rudin, 2018). In a short-term perspective, knowledge discovery is likely to remain the primary application area of machine learning methods to functional brain imaging data.

This thesis summarizes how these methods can be applied to advance our knowledge about the underlying neural processes. As discussed in Section 3.4.2 discriminative learning approaches, or decoding methods, are generally not designed to describe the data-generating process. Instead, they attempt to model a set of conditional probability densities \( p(y|x) \) by learning the best possible representation of the informative patterns in data within their hypothesis space and representational capacity. As pointed out in Haufe et al. (2014), there is generally no guarantee that these conditional densities can be used to describe the data-generating process fully. Thus, conclusions that one can draw from applying machine learning techniques to neuroimaging data can be used to characterize the neural processes only indirectly. Such indirect evidence can, however, complement traditional analysis methods in a powerful way allowing to address research questions that cannot be addressed otherwise. Interpretation of decoding outcomes can generally be classified into three main...
Probing generalization properties of predictive models  Decoding models focus on out-of-sample prediction. A carefully designed experiment can use these properties to probe whether the patterns learned in one dataset generalize across subjects, experimental conditions, or populations. Combined with transfer learning techniques, decoding models provide efficient means to probe whether similar patterns of neural activity emerge under different experimental conditions, as well as across different individuals, allowing to address intra- and inter-individual variability in brain responses at new levels of depth.

Applying machine learning techniques to neuroimaging data requires particular attention when designing experiments and data collection procedures. Specifically, an efficient experimental design should include a sufficient number of training examples as well as ensure that the decoding performance can be unambiguously attributed to the hypothesized brain activity and not to other correlated phenomena such as movement artifacts, eye blinks, etc.

Model comparison  Insights about the generative process can be obtained by applying several approaches that incorporate different modelling assumptions and comparing their predictive performances. For example, in Publication II, we demonstrate that interaction between latent components leads to an increase in prediction accuracy while introducing non-linearity functions into spatial mixing does not.

When comparing the performance of several alternative discriminative models, it is rarely sufficient to show that model A outperforms model B without estimating the external validity of the best-performing model. Indeed, both models – A and B can, in principle, be very poor at predicting the target variable to be useful in practice. Finally, as discussed in the context of representational capacity and interpretability, given equal predictive performance, a simpler model should be always be preferred.

Feature interpretation  Due to the discriminative nature of decoding approaches, the interpretation of parameters that such models extract from the data is not straightforward. Interpreting features that discriminative models use to produce their predictions is always dependent on the quality of these predictions and produces useful information only if such performance is close to optimal. Taking these nuances into account, various model inversion techniques are required to link discriminative learning outcomes to the generative model. This area remains a topic of active research (Haufe et al., 2014; Kindermans et al., 2018).

Combined with encoding approaches, decoding methods can provide additional insights into causal relationships between the observed brain activity and the experimental conditions. This combination can be partic-
uliarly valuable for features relevant for, e.g., an encoding model, but are irrelevant for a decoding model and vice versa.

**Limitations and future directions**  This thesis focuses primarily on applying machine learning techniques to the time-domain event-related MEG data in supervised classification or regression problems. Analysis of ongoing spontaneous brain activity, functional connectivity, or complex phase-amplitude perturbations in the time-frequency domain can require a more refined approach to fit the complexity of these more advanced problems.

Similarly, methods developed and applied in the publications belong to the class of discriminative methods. Generative approaches such as, e.g., hierarchical Bayesian networks can be used in the future to combine the power of the multivariate machine learning methods with high interpretability inherent to generative approaches. One example of applying simple generative models to explain the richness of the neurophysiological data can be provided in the recent publication by Nagaeva et al. (2020) in which we used Gaussian Mixture Model combined with Bayesian model selection to identify distinct subtypes of somatostatin-expressing interneurons in the ventral tegmental area in mice based on single-cell firing patterns. This work is not included in the present thesis.

Finally, despite the considerable advances in the field, I believe that this work is far from being over. I can only hope that the observations presented in this thesis contribute to this ongoing process of discovery.


D. Bzdok and J. P. Ioannidis. Exploration, Inference, and Prediction in Neuroscience and Biomedicine, 4 2019. ISSN 1878108X.


N. Caporale and Y. Dan. Spike timing-dependent plasticity: A Hebbian learning rule, 2008. ISSN 0147006X.


N. E. Helwig. Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging, 3 2019. ISSN 19390068.


N. Kriegeskorte and P. K. Douglas. Interpreting encoding and decoding models, 4 2019. ISSN 18736882. URL /pmc/articles/PMC6705607/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6705607/.


N. Naseer and K. S. Hong. fNIRS-based brain-computer interfaces: A review, 1 2015. ISSN 16625161. URL /pmc/articles/PMC4309034/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4309034/.


Bibliography


A. Puce and M. S. Hämäläinen. A review of issues related to data acquisition and analysis in EEG/MEG studies, 6 2017. ISSN 20763425. URL /pmc/articles/PMC5483631/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5483631/.


S. Vieira, W. H. Pinaya, and A. Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications, 3 2017. ISSN 18737528.


