Reverse Engineering Neural Networks From Many Partial Recordings

Elahe Arani*, Sofia Triantafillou†, and Konrad P. Kording†

*Donders Institute for Brain, Cognition, and Behaviour; Biophysics Department; Radboud University; Nijmegen, The Netherlands
†Department of Bioengineering; University of Pennsylvania; Philadelphia, PA, USA

Abstract

Much of neuroscience aims at reconstructing brain function, but we only record a small number of neurons at a time. We do not currently know if simultaneous recording of most neurons is required for successful reconstruction, or if multiple recordings from smaller subsets suffice. This is made even more important as novel techniques allow recording from selected subsets of neurons. To get at this question, we analyze a neural network, trained on the MNIST dataset, using only partial recordings and characterize the dependency of the quality of our reverse engineering on the number of simultaneously recorded "neurons". We find that prediction in the nonlinear neural network is meaningfully possible if a sufficiently large number of neurons is simultaneously recorded but that this number can be considerably smaller than the number of neurons. Moreover, recording many times from small random subsets of neurons yields surprisingly good performance. This type of analysis we perform here can be used to calibrate approaches that can dramatically scale up the size of recorded data sets in neuroscience.

Keywords: Systems Identification, Experiment Planning, ANNs, Partial Recordings.

Introduction

Reconstructing input/output (I/O) function in the brain plays an important role in neuroscience. To do so, neuroscientists record the activity of neurons and try to model their relationship using machine learning (ML) and statistical methods (e.g. see Pillow et al. (2008); OLeary et al. (2015)). However, current technology allows us to record only a small subset of the neurons that participate in the solution of even simple tasks.

Even though it is not possible to simultaneously record all neurons involved in a task, it is possible to make multiple partial recordings potentially with different observed neurons in each recording. For example, multi-electrode measurements (Ballini et al., 2013), 2-photon calcium imaging (Kerr & Denk, 2008) and optogenetics (Deisseroth, 2011), allow us to simultaneously observe a subsets of neurons. The optical methods even allow us to select which exact set of neurons we want to record. Therefore, it is natural to question if integrating multiple partial recordings can compensate for partial observation.

Based on these technological advances, a few studies have recently focused on dealing with sub-sampled observations of neural activity. Pillow & Latham (2007) extend the linear-nonlinear Poisson (LNP) framework to include the activity of unmeasured (hidden) neurons to estimate connectivity patterns among observed and unobserved neurons. However, the method is limited to small numbers of unobserved neurons. Wohrer et al. (2010) propose recovering the full noise correlation matrix from partial electrophysiology recordings, based on fully observed signal correlations. Turaga et al. (2013) use a latent dynamical system model to combine two non-simultaneously recorded but strongly interacting populations of neurons into one model, and show that combining partial observations improves the prediction of neural activity. These studies show that there is a growing interest in combining partial observations to improve our understanding of the brain.

Quantifying the added benefit of combining partial information to elucidate neural function is hard, particularly in the absence of known ground truth. Recently, Jonas & Kording (2017) advocate using man-made systems, where the ground truth is fully accessible, to evaluate methods used for studying biological systems and identify their strengths and caveats. Based on this idea, in this paper, we explore how well we can reverse engineer the neural activation in an artificial neural network (ANN) based on partial observations of its hidden neurons. We see this as a step forward in answering the following questions: Is it meaningful to try and reconstruct a complex neural circuit when we only partially observe it? What is the best way of dealing with the missing information in partial recordings? Which of the ML methods is more successful for the task? And finally, what is the best strategy for combining partial observations, given that the sampling capacity is technically limited?

In addition, in light of the recent successes of ANNs in a wide range of ML problems, reconstruction an ANN has become an interesting problem in its own right. In particular, reconstructing part of a networks' I/O functions based on partial observations is related to the redundancy of the specific network. Izui & Pentland (1990) show that redundant neural networks are more accurate, faster and more stable. Denil et al. (2013) show that there is significant redundancy in the parameterization of several DL models and that based on a few parameters of the model it is possible to accurately predict the remaining values. Cheng et al. (2015) exploit the redundancy in ANNs to reduce memory footprint. Techniques based on knowledge distillation (Hinton et al., 2015) compress the knowledge of a network into a more compact model, which is trained to predict the soft outputs of the larger model. Being able to understand neural networks based on "recordings" is important both in biology and in machine learning.

To approach this question, we train ANNs on the MNIST dataset. We use these ANNs as the ground-truth networks, and use state-of-the-art ML methods to estimate their intermediate I/O functions using partial observations of the hidden layers as input. We then examine how estimation accuracy

behaves for settings of partial observation. This resembles prediction in neuroscience, where subsets of neurons in different parts of the brain are measured, and ML methods are used to estimate the I/O relationship between them.

Results

In neuroscience, we often try to reconstruct the functionality or I/O function of areas in the brain. However, brain measurements are noisy and limited by the recording technologies, and the ground truth is not known. Here, we use standard ANNs trained on a classical hand-written digit dataset (MNIST) as the ground truth systems. We use 3000 samples from the MNIST training data to train our ANNs, while the rest is used to produce partial recordings data, described below. Assuming a network of $L+1$ layers (with $l=0$ being the input layer and $l=L$ the output layer), with N_l neurons on each layer, the activity of the neurons in the output layer L is a function of the activity of any previous layer l:

$$\{O_i^L\}_{i=1}^{N_L} = f(\{O_j^l\}_{j=1}^{N_l}), \tag{1}$$

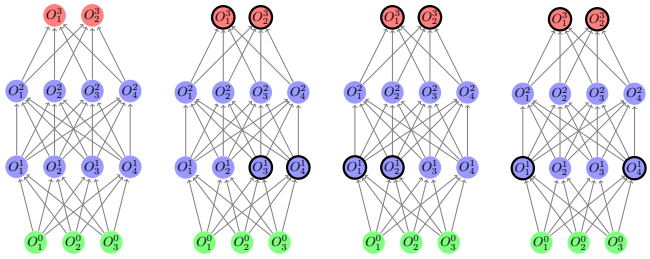
where O_n^l is used to denote the output of neuron n in layer l. In this work, we train two different ANNs on the MNIST data (i: NN with 4 hidden layers of 128 neurons in each layer and 99.62 and 97.36 accuracy on training and test data; ii: DNN with 7 hidden layers of [512, 256, 256, 128, 128, 128, 128] neurons and 99.55 and 97.78 accuracy on training and test data). These serve as our ground truth system whose intermediate I/O function we try to approximate based on multiple partial recordings. We assume that at any given point of time, we can only record a subset of neurons, and will thus have considerable missing information. We want to quantify how well reconstruction is possible based on an ML technique, the approach to deal with the missing data (unrecorded neurons), the number of recorded neurons, and the noise level.

To address these issues, we simulate partial recordings from the trained ANNs and use ML to reconstruct the hidden-unit to output-unit mapping. We assume that there is a sequence of settings k. For each setting, only a random subset of neurons is observable. We can think of each k, as one setting of the electrical or optical recording apparatus. For each setting, we thus observe the activity of a different subset of the neurons $\{O_j^l\}_{j\in Obs_k}$, while the rest are unobserved. An example of partial data is shown in Fig.1. This way, we produce simulated training sets that approximate what we may record partially from a real brain.

Based on the data collected from the (partial) recordings, we want to estimate the activity of the output neurons as a function of all the observed neurons in layer l:

$$\{\hat{O}_i^L\}_{i=1}^{N_L} = \hat{f}(\{O_j^l\}_{j\in\cup_k Obs_k}). \tag{2}$$

This problem is not trivial because the partial recordings produce many missing values corresponding to unobserved neurons in each recording. The missing data have a specific structure: in each recording, only a subset of variables (neurons) are observed. The final model in Eq.2 is defined over the union of all variables that have been observed in at least one of the recordings. The ratio of missing over observed data is higher for recordings with fewer observed neurons per recording.



(a) Trained net (b) Subset 1 (c) Subset 2 (d) Subset 3

Figure 1: An example of three recording subsets. (a) The ground-truth neural network. (b,c,d): An example of three ($K=3$) settings. In each setting, the all output neurons and two of the neurons of layer 1 are observed. Observed neurons are outlined in black.

Missing data is a common problem in statistics, and the way it is dealt with can affect the results, depending on the missingness mechanism. In the simplest case, where data are missing independently of observed and unobserved data, the data are said to be Missing Completely At Random (MCAR) Rubin (1976). This is the case for the partial data, since the inclusion of a neuron in a recording is based on the sampling design, and does not depend on the activity of the neurons, observed or unobserved. When data are MCAR, excluding the samples with missing values do not bias the results. However, discarding every sample that includes a missing value is not possible here, since every recording has some missing values. Thus, we must select a strategy for dealing with the missing data.

The most common strategy for missing values is mean value imputation (MVI), where every missing value for a variable is replaced with its sample mean. More complex approaches in the area of multiple imputation (Rubin, 1996) fill in the missing values by assuming a model for the missing values that can be estimated from observed values. The estimated model is then used to impute the missing values. For example, Soft-Impute (SI) iteratively uses a soft-threshold singular value decomposition (SVD) to replace the missing values (Mazumder et al., 2010). In this work, we use MVI and SI to complete the missing values in our partial recordings data.

An alternative approach to using all the data in the same model is to train a separate estimator for each recorded subset and combine the predictions. In this case, the estimator in Eq.2 can be computed as the average of K marginal estimators:

$$\hat{f}(\{O_j^l\}_{j\in\cup_k Obs_k}) = \frac{1}{K}\sum_k \hat{f}_k(\{O_j^l\}_{j\in Obs_k}) \tag{3}$$

The advantage of marginal estimators is that each $\hat{f}_k$ can be identified with any appropriate ML method without preprocessing, in parallel. However, each marginal estimator is then estimated based on fewer samples.

We now use simulated experiments to examine how the quality of reconstruction based on partial recordings depends on: (a) data imputation method, (b) noise level, (c) size of the ground truth network, and (d) number of recorded subsets.

We use NN and DNN networks as our ground truth networks, and simulate observations by sampling with replacement from the second half of the MNIST training data that were not used for training the ground truth networks. We typically simulate $K = 10$ partial recorded subsets. This means that we consider K subsets of neurons to record from, each chosen randomly (and thus often overlapping), which corresponds to switching the subset of neurons recorded $K - 1$ times. In each subset, all output neurons and N neurons from a given hidden layer are observed.

In real neuroscientific experiments, there is a fixed information throughput that limits the overall data set size: You can either opt to record more neurons (larger subset sizes) for fewer samples, or fewer neurons for more samples. Thus, to make more meaningful comparisons, we only compare situations where the same total amount of information is acquired, i.e. each training set will have the same product of the number of observed neurons and the number of observations.

We then train an ML method to predict the values of the output neurons based on the partial data of the given hidden layer. We used XGBoost and ANNs coupled with imputation techniques (MVI and SI), as well as with combining separate marginal models (denoted by MP for mean prediction). In addition, we used XGBoost with the built-in approach for handling missing data without explicitly imputing the missing values, by adding a default direction to each tree node (denoted by XBG-G). For each partial data set, we trained an independent model with XGBoost for every output neuron, and a single ANN model with 10 output neurons. XGBoost was used with the following parameters: $\eta = 0.5$ and $max\_depth = 6$. We used an ANN with three hidden layers (16 neurons in each, ReLU activation function) and applied early-stopping based on loss value. We repeated this procedure for 5 iterations of random subsamples.

To compare the performance of different methods on partial recordings, for each iteration, we calculate the average (over all output neurons) Root Mean Squared Error (denoted by $\overline{\text{RMSE}}$). Fig.2 shows the reconstruction accuracy for different methods as a function of the number of observed neurons in each recording. Both layers have similar performance. The error decreases as the number of observed neurons increases, although the total number of non-missing data points remains nearly the same. For both layers, ANNs with MVI perform better for almost all settings.

The recordings of neural networks include several sources of noise. The effective noise in the recordings has a component that comes from measurement, e.g. Johnson noise in the electrode, and a component that comes from the brain itself (e.g. Poisson noise in spiking). To study the effect of noise, we add signal-dependent noise to the partial recordings. For each sample, we add zero-mean Gaussian noise with variance that is proportional to the activity of the neuron in the given sample. We set the noise variance to $20, 50$ or 100 percentage of the neurons' activity value. Here, 100% corresponds to the level expected for a Poisson process. We add the noise after simulating the partial recordings and use ANNs to estimate the I/O function. To check which of the imputation techniques can handle the noise better, we consider both MVI and SI imputation techniques. MVI still outperforms SI, but both methods are robust to all levels of additive signal-dependent noise (Fig.2).

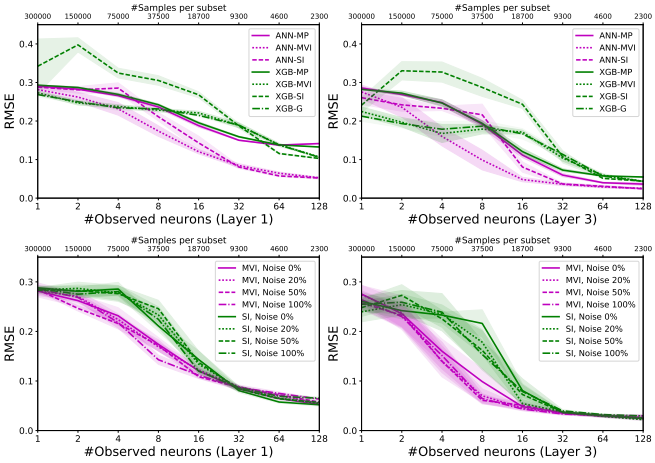


Figure 2: **Neural networks with mean value imputation work best at reverse engineering an ANN, even in the presence of noise.** $\overline{\text{RMSE}}$ as a function of the subset size from the first (left) and third (right) layer of the ground truth NN for various top: reconstruction techniques, and bottom: additive noise levels.

We also simulated data from a deeper NN (DNN) as a ground truth network, to compare estimation accuracy when the ground truth network is bigger and has more redundancy. Results in Fig.3 show similar trends in both NN and DNN data: increasing the number of observed neurons improves performance. Moreover, for wider layers, a smaller ratio of neurons is required for adequate performance. In all layers, by observing roughly twenty percent of the neurons, we can achieve nearly the same error rate as with a fully observed layer.

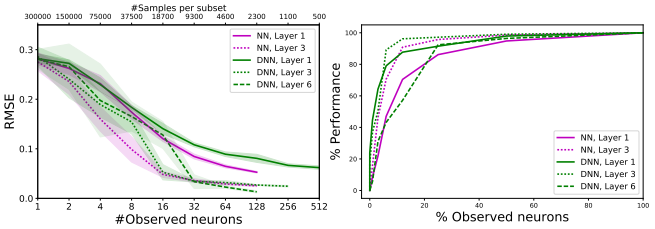


Figure 3: **Simultaneously recording more neurons helps reverse engineering for a fixed number of samples, just as well for shallow and deep networks.** Left: $\overline{\text{RMSE}}$ for reconstruction as a function of observed subset size using ANN-MVI in different layers of the NN and DNN. Right: Performance (as a percentage of maximum decrease in $\overline{\text{RMSE}}$) vs percentage of observed neurons for each layer.

Results so far show that increasing the number of the observed neurons at the expense of presenting fewer stimuli improves reconstruction accuracy for a fixed number of selections of subsets of neurons. However, it is possible that increasing the number of selected subsets can compensate for few simultaneously recorded neurons. Choosing many different subsets is often experimentally feasible, e.g. by focusing a laser on different subsets of neurons on a different plane. To test the trade-off between the number of subsets and number of recorded neurons in each, we simulated partial data for

varying K, and a different number of observed neurons per subset. While no meaningful infenences are possible with a single selection of neurons (Fig.4), increasing the number of selected subsets improves the prediction accuracy; this improvement is more noticeable when fewer neurons are simultaneously recorded. Changes in performance are not always monotonic; in general, however, combining more subsets can compensate for the lack of full observations and achieve similar performance to combining fewer recordings of larger subsets.

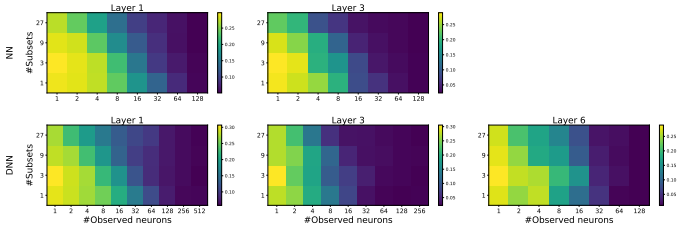


Figure 4: **Sequentially recording from many subsets of neurons can compensate for recording only small subsets at a time.** $\overline{\text{RMSE}}$ for reconstruction using ANN-MVI for a different number of sequentially recorded subsets K, keeping the total number observed values approximately constant.

Discussion

Modern techniques in neuroscience allow simultaneous recordings from many neurons. Motivated by this, we investigated how well we can predict neural activity in ANNs from multiple partial observations. Surprisingly, MVI -the simplest and easiest imputation technique- achieves a lower error rate even compared to a more complex and computationally expensive method, like SI. Also, we found that increasing the number of simultaneously observed neurons increases the quality of our prediction. However, many recordings from small subsets of neurons can well approximate what we get if we simultaneously record from all neurons. Thus, combining multiple partial recordings may improve our understanding I/O functions in the brain.

Obviously, an artificial neural network is not a brain. Even though ANNs are inspired by neuroscience, brains are biophysically more complex and probabilistic. Biological neurons can perform on many different time scales and are more heterogeneous in many ways. In the brain, the communication takes place through spikes, whereas it happens via abstract rates in neural networks. Moreover, in our specific work here, we dealt with small ANNs, whose architectures are obviously idiosyncratic,as they are simple feed-forward neural networks with no lateral or recurrent connections. In addition, ground truth networks were only trained on the MNIST dataset. Arguably, the manifold of hand-written digits has a far simpler structure than typical manifolds of things in the real world. Thus, our analysis might give categorically different answers if we could apply it to the real brain.

Nevertheless, this work presents a test case where artificial systems are used to evaluate and guide computational neuroscience. Using ANNs as a stand-in for an actual neural circuit, we found that reverse engineering neural function based on a small fixed subset of recorded neurons is not possible, even for a low dimensional task. This may suggest that the number of neurons that need to be simultaneously recorded may be larger for more complicated tasks. We, therefore, believe that a similar scaling analysis should be standard for reverse engineering in neuroscience applications.

Modern neuroscientific techniques allow running similar analyses to the one we presented here on real brains. Using Ca2+ imaging along with modern optical targeting techniques allows trading of the number of recorded neurons with the noise level. It also allows simultaneous recording from the input parts of a system and the outputs. As such, the neural network that we used here, could be readily replaced with a real sample of brain tissue. Doing scaling analyses on reverse engineering approaches is the only way of knowing how recording techniques need to be optimized.

References

Ballini, M., Mueller, J., Livi, P., Chen, Y., Frey, U., Shadmani, A., ... others (2013). A 1024-channel cmos microelectrode-array system with 26'400 electrodes for recording and stimulation of electroactive cells in-vitro. In *Vlsi circuits, 2013 symposium* (pp. C54–C55).

Cheng, Y., Yu, F. X., Feris, R. S., Kumar, S., Choudhary, A., & Chang, S.-F. (2015, December). An exploration of parameter redundancy in deep networks with circulant projections. In *ICCV*.

Deisseroth, K. (2011). Optogenetics. *Nature methods*, *8*(1), 26–29.

Denil, M., Shakibi, B., Dinh, L., de Freitas, N., et al. (2013). Predicting parameters in deep learning. In *NIPS* (pp. 2148–2156).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *In deep learning and representation learning workshop, NIPS*.

Izui, Y., & Pentland, A. (1990, April). Analysis of neural networks with redundancy. *Neural Comput.*, *2*(2), 226–238. Retrieved from `http://dx.doi.org/10.1162/neco.1990.2.2.226` doi: 10.1162/neco.1990.2.2.226

Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLOS Computational Biology*, *13*(1), e1005268.

Kerr, J. N., & Denk, W. (2008). Imaging in vivo: watching the brain in action. *Nature Reviews Neuroscience*, *9*(3), 195–205.

Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, *11*(Aug), 2287–2322.

OLeary, T., Sutton, A. C., & Marder, E. (2015). Computational models in the age of large datasets. *Current opinion in neurobiology*, *32*, 87–94.

Pillow, J. W., & Latham, P. E. (2007). Neural characterization in partially observed populations of spiking neurons. In *NIPS* (pp. 1161–1168).

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, *91*(434), 473–489.

Turaga, S., Buesing, L., Packer, A. M., Dalgleish, H., Pettit, N., Hausser, M., & Macke, J. (2013). Inferring neural population dynamics from multiple partial recordings of the same neural circuit. In *NIPS* (pp. 539–547).

Wohrer, A., Romo, R., & Machens, C. K. (2010). Linear readout from a neural population with partial correlation data. In *NIPS* (pp. 2469–2477).