Inferring causal connectivity from pairwise recordings enabled by optogenetics

Mikkel Elle Lepperød, Konrad Paul Kording

May 1, 2018

Abstract

Neurons interact through spikes and a central objective of neuroscience is measuring how neurons causally affect one another. To probe such interactions scientists often use optogenetics, which typically leads to stimulation effects on multiple cells. This then produces a so-called confounding problem - we cannot know which of the stimulated neurons affected the activity of a given target neuron. Here we show how the resulting biases can be large and how causal inference techniques, in particular instrumental variables from econometrics, can ameliorate this confounding problem. Our approach utilizes the fact that neurons have absolute refractory periods where the stimulation has no influence. This missing stimulation response can then allow disentangling causality. If we can not randomly stimulate individual neurons, then causal inference techniques are needed to identify mechanisms.

1 Introduction

We want to understand the mechanisms or causal chains that give rise to activity in the brain, to perception, action, and cognition. For such an understanding it is not sufficient to know the correlations between variables or even be able to predict them. After all, there can be many ways how the same activities come about by distinct causal chains [Drton et al., 2011, Peters et al., 2017]. Identifiability of causal networks is complicated when it comes to brain data. Complex systems such as the brain are hard to understand because of the numerous ways they interact internally [Jonas and Kording, 2017] and therefore will almost never satisfy the criteria needed for identifiability from observational data [Pearl, 2009]. While observing correlations within the system is usually relatively easy, transitioning from observed correlations to a causal or mechanistic understanding is hard or,

maybe more typically, impossible. Getting at such an understanding is particularly hard because the brain consists of countless neurons, each of which influences many other neurons. Only under certain assumptions about nonlinearity or noise sources does a fully observed system become identifiable [Daniusis et al., 2012, Shimizu et al., 2006]. Even if we could record all neurons at the same time, estimating causality and producing a mechanistic understanding would be hard.

We generally only record from a small subset of all neurons. The data we obtain from typical recordings, e.g. from electrophysiology or calcium imaging, is very low dimensional relative to the dimensionality of the brain. Moreover, it is observational, which means that it does not result from randomized perturbations. In such cases, we can never know to which level the observed activity was caused by other observed activity, or by the activity of the unobserved neurons. Such unobserved activity is then called confounders. If we, in the presence of confounders, estimate mechanisms from observational data we will generally make large errors and draw incorrect conclusions[Angrist and Pischke, 2008]. Unobserved neural activity confounds estimates of causal interactions and makes it hard to estimate real mechanisms.

Whether we use simple regression techniques or advanced functional connectivity techniques[Stevenson et al., 2008, Honey et al., 2009], confounding is the big threat to causal validity [**?**]. One popular way to estimate the output of single neurons is to perform a multiple regression [Pillow et al., 2008], modeling each neuron with a generalized linear model (GLM) . Multiple regression may be perceived a solution to confounding problems as well as it supports the concept of "explaining away". However, in this paper we will exclude approaches that look at multiple neurons at the same time. In practice, "explaining away" can only be a good strategy if we record from most neurons and we are not close to that in most modern experimental setting, especially in the mammalian brain. As such, we focus on cases where we only know stimulus, pre-synaptic and post-synaptic neuron.

Completely recording all neurons is not yet possible in mammalian brains, and we therefore focus on identification strategies from neuron pairs. Let us say we want to estimate causal connectivity between two neurons, A and C (Fig. 2(a)). But let us say that two neurons A and B are driven by a common input D. Because B and C are connected they are strongly correlated. Consequently, A and C are also correlated and a regression $C = \beta A$ will, when causally interpreted, misleadingly conclude that there is a direct interaction. In this case we say the regressor A is endogenous and the regression coefficient β estimates the magnitude of association rather than the

magnitude of causation. Naïve regressions in partially observed systems will generally not reveal causality.

To estimate causal relationships between neurons, stimulation is the gold standard. In fact, a common definition of causality is in terms of what would happen if one would change the value of one variable in the system, independently of changing other variables – an intervention [Pearl, 2009]. If we stimulate single neurons, the ability to estimate causal relationships by regression is within grasp. However, this is experimentally challenging and yield very low cell count because it either requires intracellular electrodes or two-photon stimulation [Lerman et al., 2017, Nikolenko et al., 2007, Emiliani et al., 2015]. Because gold-standard perturbations are so hard, it would be highly desirable if we could back out causality from regular optogenetic stimulation [Boyden et al., 2005, Zemelman et al., 2002].

Interpreting the results of optogenetic stimulation in terms of causal interactions is difficult. Regular optogenetic stimulation affects many neurons as it is generally impossible to direct photons to only one neuron. Hence, the stimulus will produce a distributed pattern of activity. For example, if we focus a stimulation beam on one neuron, there will be a cone ahead and a cone behind that neuron. This distributed pattern of stimulation produces down-stream activity which then percolates through the network of neurons. So any down-stream activity induced by stimulation could in principle come from any of the stimulated neurons introducing problematic confounders.

The inference of causality from observational data is studied largely in the fields of statistics [Pearl, 2009], machine learning [Peters et al., 2017] and econometrics [Angrist and Pischke, 2008]. Within these fields, the problem of endogenous regressors is commonly considered. We may thus look towards these fields for insights into how we may resolve the confounding problem induced by optogenetic stimulation.

One particularly popular approach towards causal inference in economics are instrumental variables. Let's say that we want to estimate the return β from education x to yearly income y with the regression $y = \beta x + u$. Here u is the factors other than schooling that contributes to yearly income. One of the factors in u is a person's ability. However, a person's ability may also affect schooling and thus the regressor x is correlated with the error term u. This then implies that the regression estimate β will not estimate the magnitude of causation from schooling on wages, but rather, the association. In this case one may use the proximity to a college or university as an instrumental variable (IV) z [Card, 1993]. After all, we expect that intelligent children are raised in a somewhat homogeneous way. Then, in order to attribute the causal effect of schooling on wages one may

calculate the ratio of covariances $\beta = cov(z, y)/cov(z, x)$. This then corrects for the confounding.

We might wonder under which circumstances the IV technique gives meaningful results. The IV technique has been used extensively in econometrics and is provable causal given three assumptions [Angrist and Pischke, 2008]. First, the instrument must be uncorrelated with the error term. Second, the instrument must be correlated with the regressor. Third, there must be no direct influence of the instrument on the outcome variable but only an influence through the regressor variable. The validity of these assumptions is central when using the IV.

For an instrument to be good, it needs to be unaffected by other variables. However, in the brain there may be nothing that is truly unaffected by the network state. However, there are certain variables that are more or less affected. For example, the overall activity of the network is through slow dynamics very strongly nonrandom. But the refractory period of an individual neuron may be, locally, rather random. First, if neurons are spiking according to conditional Poisson distributions, their exact timing will, conditioned on the network state, be random. Moreover, after a strong, long stimulation the phases of integrate and fire neurons will effectively be random. While refractoriness may not be perfectly random, the exact times of spiking are notoriously hard to predict [Stevenson et al., 2008] making it likely that they are quite random.

Here we show that the instrumental variable (IV) technique can be employed if one seeks to estimate the causal connectivity between neuron pairs. We begin by showing how confounding is produced by regular optogenetic stimulations. We then simulate this confounding effect in a simple network of three leaky integrate and fire (LIF) neurons. With this simple simulation we show that using the refractory period as an IV we are able to distinguish between connected and unconnected neuron pairs. We compare these estimates with a naive, but widely used cross correlation histogram (CCH) method that fails. Furthermore we simulate a recurrent randomly connected network of excitatory and inhibitory LIF neurons with distributed synaptic weights. With this data at hand we calculate the mean squared errors of the IV method and show that it is robust to different simulated network states. Finally, we compare the amount and size of false positives/negatives and goodness of fit on synaptic weights with pairwise assessments using CCH and logistic regression.

2 Results

2.1 Optogenetics is not local

Optogenetics is generally seen as a perturbation method that by and large affects local neurons. However, we may wonder if this is really the right way of conceptualizing the spatial effect of stimulation. The stimulation effects depend on multiple factors. It depends on light intensity and opsin density, where more light and more molecules will produces a stronger effect on each cell. It also depends on the number of potentially stimulated neurons, and if light is received by more neurons it will have more impact on the overall population activity. Lastly it can depend on physiological properties of the cells, e.g. light may only have a strong effect on spiking when neurons are sufficiently close to their threshold. The induced effect of optogenetics as a function of distance should be the product of four functions. The light intensity, the opsin concentration, a function characterizing the spiking properties of neurons, and lastly the density of neurons.

To estimate the light intensity we calculated the spatial extent of laser light delivered by fiber-optics under experimental plausible conditions according to [Aravanis et al., 2007]; see Section 4.5. This modeling of light intensity yield an approximately $1/r^2$ reduction with distance r from the light source as seen in Fig. 1 (cyan line). This is intuitive as the surface of the 3d shell grows is $4\pi r^2$ and photons will be roughly isotropic beyond the scattering length Fig. 1 (inset). Intuitively this makes sense, after all, the same number of photons have to cross each of the spheres around the stimulation location unless they are absorbed. The density of photons decreases rapidly with distance.

To estimate the density of neurons at a given distance we observe that there are more neurons the further away they are from the stimulation site when neurons are uniformly distributed in brain tissue as shown in Fig. 1 (black line). In fact, in good approximation this density will have an increase of approximately r^2 with distance. This again derives from the same surface scaling that we have in the 3d shell. The number of neurons that can be activated increases rapidly with distance.

To estimate the effect of stimulation we also need to consider the non-linearity for spiking which can largely be characterized by the distribution of membrane potentials of neurons. Surprisingly, that distribution has been observed to be largely flat [Paré et al., 1998, Rudolph and Destexhe, 2006]. This suggest that the response in membrane potential to a perturbation is roughly proportional to the light intensity. Moreover, assuming that opsins

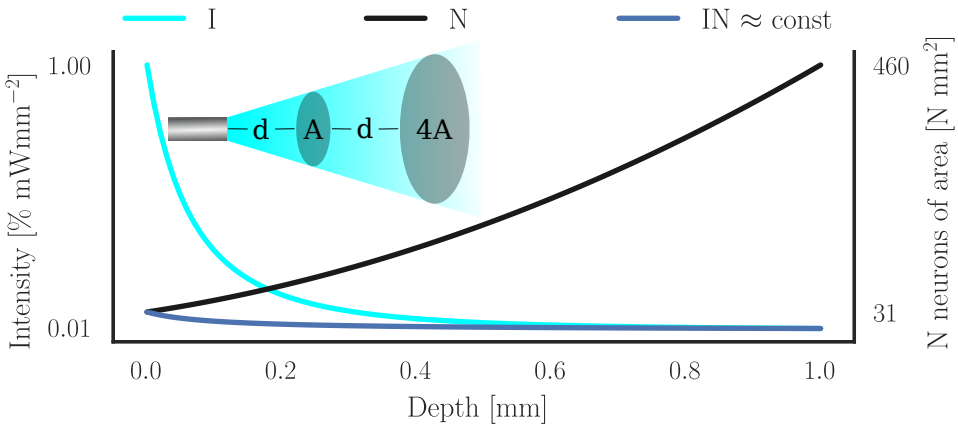


Figure 1: **Spatial extent of optogenetic stimulus**. Due to scattering and geometric loss the light intensity (I, cyan line) plotted as percentage of intensity exiting the optogenetic fiber follows approximately an inverse square law r^{-2} where r is the distance from the fiber. If neurons are uniformly distributed the number of affected neurons increase by r^2 (N, black line) rendering the probability of activating a neuron approximately constant (IN, blue line).

are not localized by means of specialized viruses, the opsin density will be relatively flat. We can thus calculate the overall stimulation effect to be the product of neuron density and light intensity which is approximately flat in distance (up to the distance where absorption becomes important) Fig. 1 (blue line). Thus, single photon optogenetics does not actually produce a localized effect.

2.2 Confounding as a problem for the estimation of causal effects

When we stimulate many neurons at the same time, and observe a post-synaptic neuron to be active after our stimulation, it is hard to know which of the stimulated neurons produced the activity. To illustrate such confounding effects we simulated a network comprised of three neurons (A,B,and C). These neurons were simulated to implement a Poisson spiking mechanism. In addition we gave them additive Gaussian white noise to the membrane potential. We also gave them interactions, where spikes of neuron B increase the probability of firing for neuron C but there are no other interactions

Fig. 2(a). Finally, we allowed simulated optogenetic stimulation to affect neurons A and B (but not C). We thus have a simple system to think through causal questions.

After running the simulation, the peristimulus time histogram of the stimulated neurons show the result of both the stimulation itself (suppressed for visibility) and the neuron's refractory period Fig. 2(b) (AA, BB). Since the stimulation affects A and B simultaneously it induces a strong correlation between A and B Fig. 2(c) (AB). This further generates a strong correlation between A and C, confounding the system by rendering the cross correlation histograms (CCHs) between BC and AC both statistically significant ($p_{fast} < 0.001, p_{diff} < 0.001$; see Section 4.2) shown in Fig. 2(c). Even though the correlation peak between B and C is larger than between A and C due to correlated spikes outside of the stimulation times one may imagine a situation where only A and C is measured, giving rise to a false prediction that they are connected. If stimulation affects multiple neurons simultaneously, there is a real confounding problem.

2.3 Instrumental variables to resolve confounding

If we want to discover the actual influence of stimulation of a neuron on post-synaptic neurons we need something that can distinguish the influence from one stimulated neuron from the influence of another stimulated neuron. We would thus need something that affects the stimulation effect on only one neuron. Arguably, refractoriness is such a variable. If a neuron is in its absolute refractory period, then no amount of stimulation will make any difference. This allows us an interesting way of getting at causality, by comparing the network state between a time when a neuron was able to spike and a time where the neuron was unable to spike.

Instrumental variables requires the existence of a random (or random enough) variable that affects a variable of interest. This independent influence then allows quantifying the influence of the variable of interest on the rest of the network. In our case, a neuron being refractory is in good approximation independent (see Discussion for caveats). It affects the influence of stimulation on the potentially refractory neuron Fig. 3(a). And the spikes that are thus missing from the refractory neuron which otherwise would have been induced by the stimulation can then be used to identify causality.

We can now test if, for our simple three neuron system, an instrumental variable estimator would do better than simply analyzing the correlations by looking at the cross correlation histogram (CCH). We thus use the IV

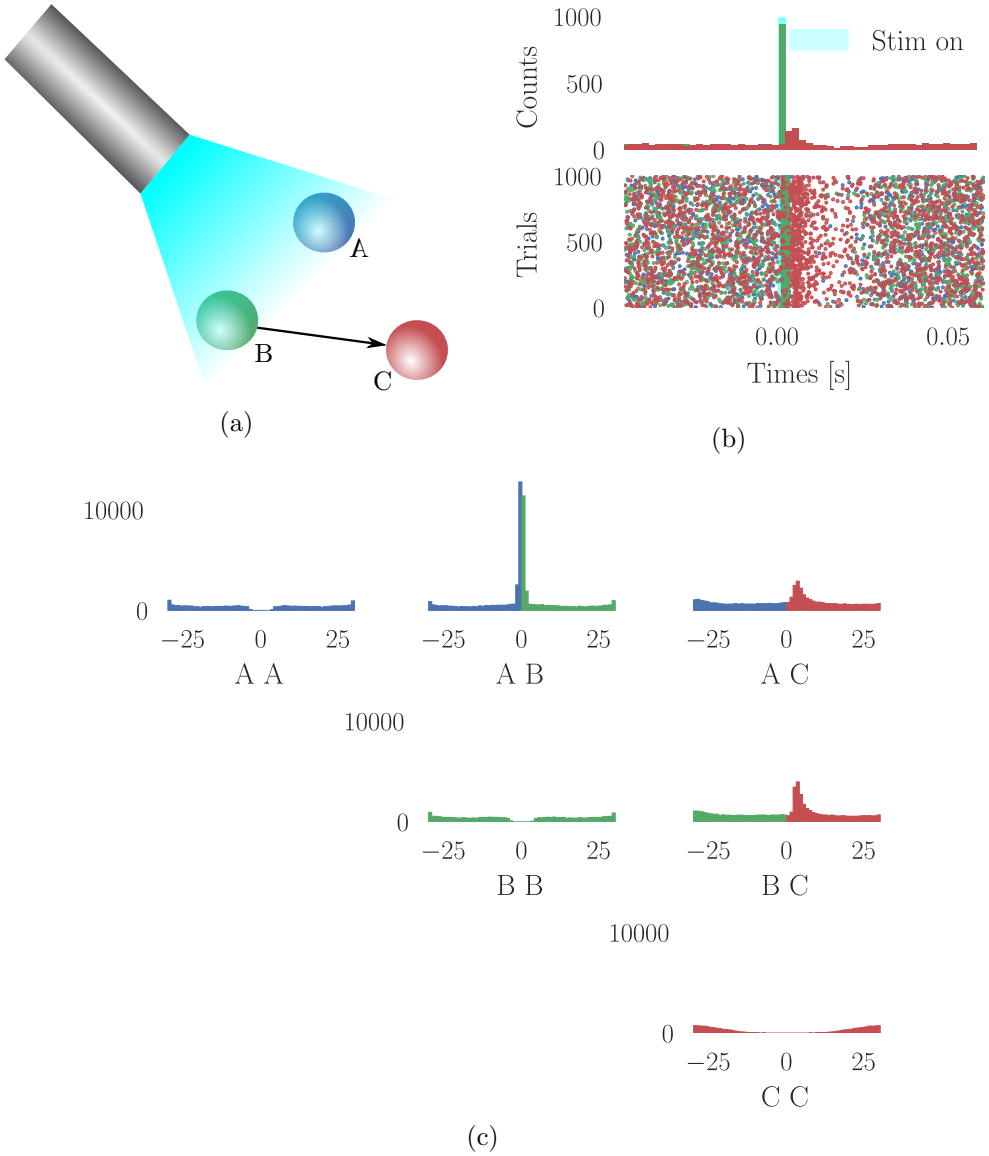


Figure 2: **A,B,C network**. A sketch of the simple network containing three neurons shows stimulation configuration with blue laser light and the connections with arrows (a). The neurons A and B are stimulated 1000 trials and the corresponding peristimulus time histogram are shown in (b) upper panel with a raster plot in the lower panel. Cross correlation histograms (CCHs) are shown in (c) where horizontal and vertical axes represents time lag in ms and counts of coincident spikes in bins of 1 ms.

estimator Eq. (4) on the three neuron system Fig. 3(b) and (c). It converges to the correct causal conclusions that the weights $w_{BC} > 0$ and $w_{AC} = 0$. For such a simple system, it produces meaningful estimates of the causal interactions between neurons.

2.4 Larger simulated networks

The interacting neurons in a real network exhibit inhibition and interact in many kinds of ways. To evaluate the IV method in a more meaningful setting we thus simulated a recurrent neural network consisting of 1250 randomly connected linear integrate and fire neurons where 250 had inhibitory synapses. The network was tuned to be in an asynchronous regime; see Fig. 6(a) with distributed synaptic weights according to patch clamp experiments [Sayer et al., 1990, Mason et al., 1991]; see Fig. 6(c) and Table 1 for parameters. To evaluate how well the IV method estimates the weights as a function of number of trials we calculated the mean squared error of the weight connecting 100 neuron pairs Fig. 4. As seen here, the IV estimator's precision decreases similarly in three different settings with varying amounts of relative inhibition g.

We want to compare the IV estimator which exploits the refractory period with the CCH method given by Eq. (6) which ignores network confounding. We calculated the amount of false positives as the percentage of estimated synapses larger than 0.05 where the true weight was 0, finding 13.3% for CCH and 0.2% for the IV estimator; see Fig. 5(a). In addition we compared the size of the estimations at false positive instances and found that the CCH give significantly higher false positive weight-estimates than IV ($p = 0.03$ $\Delta = 0.049$, calculated by permutation re-sampling [Wassermann, 2006]; see Fig. 5(a). The IV approach, while not being perfect, thus outperforms the simple CCH approach.

It might be that modeling refractory periods in the context of a naive regression improves results. We thus performed a logistic regression as seen in Fig. 5(a) denoted logit. As seen here, logit performs even worse than CCH showing that it really helps to use the refractory period as an instrumental variable. To further evaluate the methods we calculated false negatives as instances where the true weight is $w > 0$ but estimated to be equal to zero in Fig. 5(b) showing that the CCH and IV estimators performs equally well. Finally we wanted to evaluate the estimated weights as a function of true weights shown in Fig. 5(c) and (d) after 30000 trials. The IV estimator yields a good prediction $r^2 = 0.55$, while the CCH estimator performs surprisingly bad with $r^2 = 0.002$. Utilizing refractory periods as

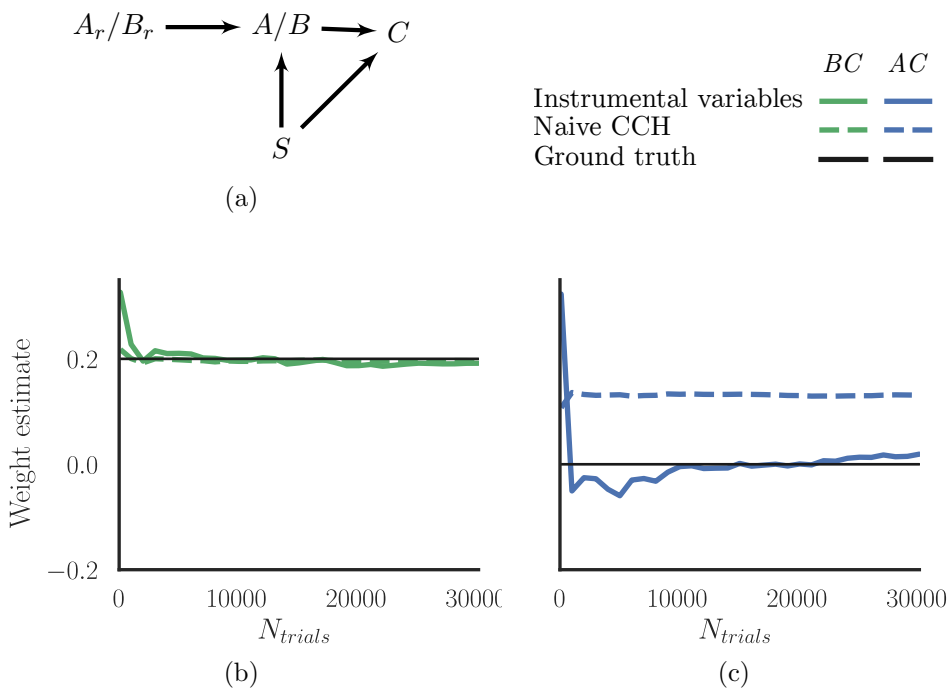
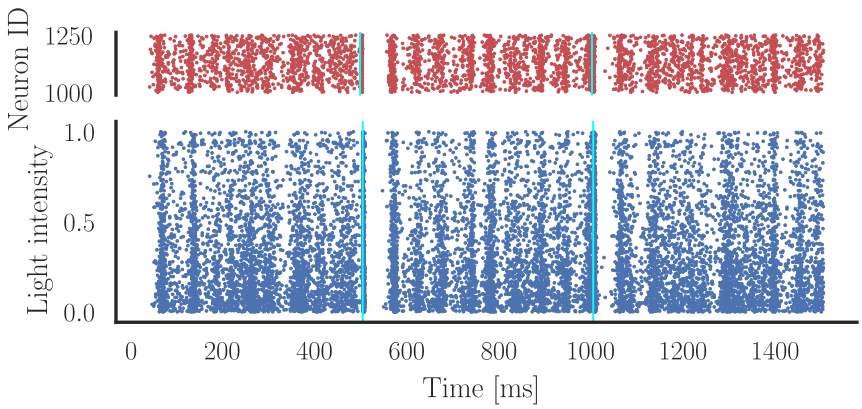
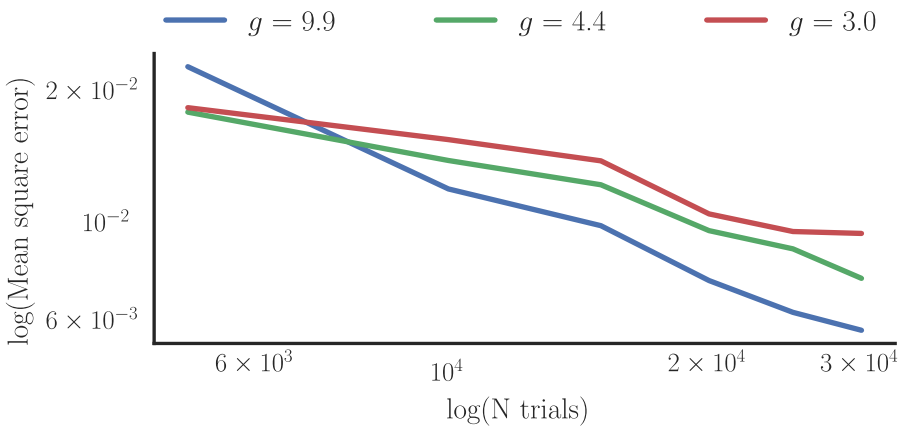


Figure 3: **Instrumental variable estimation (IV) of connectivity**. (a) In an instrumental variable estimation procedure we use a variable that is assumed to be random (here refractoriness) which influences a variable of interest (here spiking) and to use this influence to infer the causal interaction of that variable on other variables (here spiking of A or B onto C). (b) A popular estimation approach for IVs, the Wald technique correctly estimates causal connectivity in the A,B,C neural network using the refractory period. Subfigure (a) shows a path diagram [Wright, 1921] between the upstream neuron A or B, the stimulation S, the post-synaptic neuron C and the IV as A_r or B_r. Arrows represent associations where S is associated with A, B and potentially also with C both directly and through A, B. The IV estimator calculated by Eq. (4) converges to $\hat{\beta}_{BC} \approx 0.2, \hat{\beta}_{AC} \approx 0$ after approximately 5000 trials as seen in (b) and (c) respectively. Insets represent high resolution zoom of cross correlation histograms of BC and AC where horizontal and vertical axes represents time lag and counts of coincident spikes in bins of 0.1 ms respectively.



(a)



(b)

Figure 4: **Mean square error (MSE) of IV estimator in a large network.** (a) Raster plot showing stimulated excitatory population in blue and inhibitory population in red, the stimulation onset is colored in cyan t = 500 and 1000 ms. (b) The IV estimator is evaluated in the asynchronous recurrent neural network at three different amounts of relative inhibition g (decreasing with model number). The MSE as a function of number of trials is shown on a logarithmic scale where the slopes was found to be $-0.77, -0.48, -0.40$ for model 1,2,3 respectively.

an (imperfect) instrumental variable considerably improves estimates.

3 Discussion

Here we have asked if the refractory period of neurons can be used as an instrumental variable to reverse engineer the causal flow of activity in a network of simulated neurons. We have found that this approach performs considerably better than the naive method. We have found that neither naive linear nor naive logit models produce good estimates of weights. Our system effectively reverse engineers causality by looking at the response that is missing because of refractoriness which effectively allows better estimates of causal effects.

One popular way at estimating causal effects is fitting generalized linear models (GLMs) to simultaneously recorded neurons[Pillow et al., 2008]. The GLMs are basically multiple regressions and require multiple neurons in order to perform well. In fact, if one recorded all neurons GLMs could be sufficient to estimate causal connections. However, this is not the case for research in the mammalian brain, especially not for primates, where we only record a very small subset of the neurons that do the actual computation. The GLM field is very strong at modeling latency distributions and sequences of spikes in individual neurons. These ideas should, arguably, be merged with IV approaches. The main strength of the IV estimator presented here compared with GLM methods is that we only require one pair because we can utilize the randomness in refractoriness.

The main problem with optogenetics when using it to infer connectivity is it's non-local property. This is due to the long reach of light, high density of neurons and the observation that distributions of membrane potentials over neurons are generally flat [Paré et al., 1998, Rudolph and Destexhe, 2006]. One could however imagine situations where optogenetics were more local. If membrane potential distributions were in general skewed with the mode far from threshold the optogenetic perturbation would be more local. This is since each neuron would require a very strong stimulus in order to elicit spikes at all. But there could be other ways of making optogenetics more local. For example, if one could engineer animals to have brains that are far more absorbent one could stimulate locally. How to engineer more localized stimulation is an important problem if one wants to causally interrogate systems.

Very weak laser pulses in noisy networks might only affect very few neurons each trial [English et al., 2017]. However, the stimulus will still

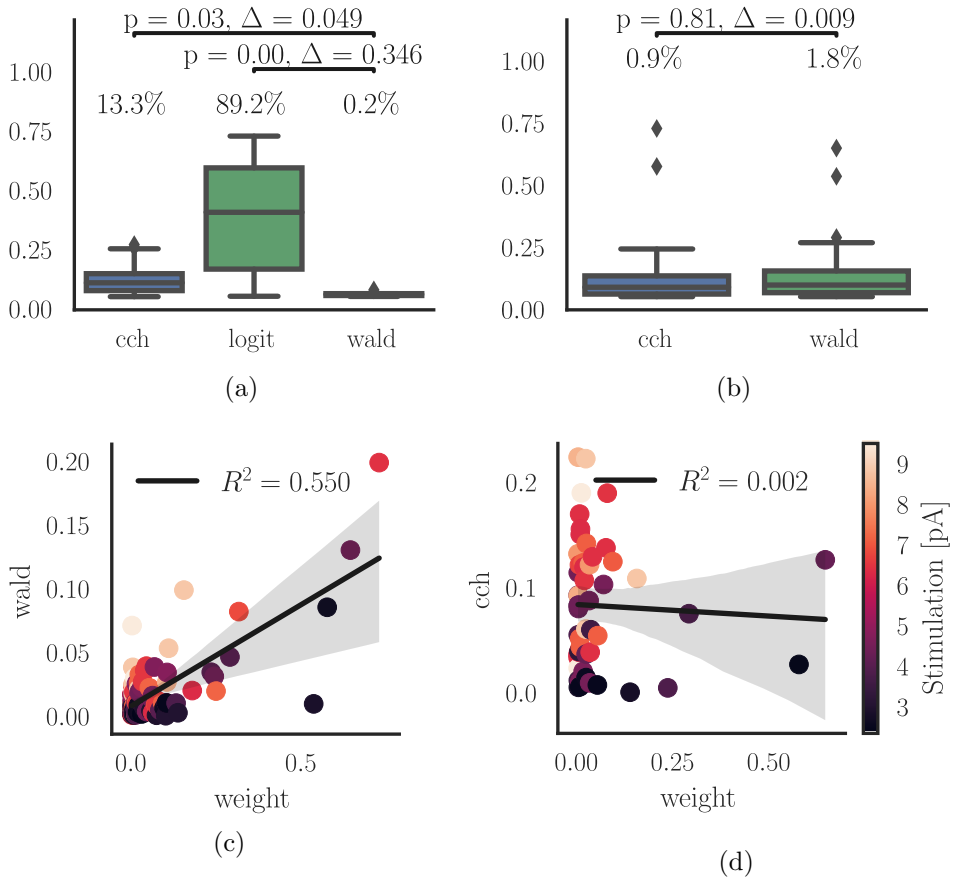


Figure 5: **False estimates and goodness of fit**. False positives are shown in (a) for the cross correlation histogram (cch) method, logistic regression (logit) and the IV estimator (wald). False negatives for cch and wald is shown in (b). Positive estimates of weight as a function of true weight is scattered for the wald estimator in (c) and cch in (d) color coded by the size of perturbation intensity.

affect the many far away neurons by a tiny bit. Therefore, weak stimulation does not remove the logical problem of the CCH estimator. Moreover, the network still acts as a confounder and, if anything, the weak stimulation will dramatically reduce the statistical power of the approach. Lowering stimulation amplitudes does not appear to be a way of obtaining meaningful causal estimates.

For the refractory period to be a good instrument, it is necessary that it is not overly affected by the network activity. This clearly is going to be problematic in many cases, after all the network activity affects refractoriness. However, there are multiple scenarios where refractorinness will be a good instrument. For example, if we have balanced excitation and inhibition we may expect largely random refractoriness of individual neurons. If a neuron biophysically implements something like conditional Poisson spiking its refractoriness will be pretty random. Even if neurons refractoriness is strongly correlated during normal network operation there may be ways of randomizing it. Giving one burst of stimulation which is strong enough to elicit multiple spikes from each neuron may effectively randomize the phase of each neuron. Importantly, we may expect the phase of a neuron to be far more random than the activity of the network as a whole.

We found some negative values of the IV estimator which were suppressed as we knew that we only stimulated excitatory neurons. This happens because neurons have correlated refractory times which is likely when looking at the distribution of CC seen in Fig. 6(a). Furthermore, the neural network simulated here introduces much response overlap due to synapses having equal synaptic time constants and transfer delays. This makes inference quite hard since multiple neurons are affecting the same cell at the same time for each stimulation. However, this is less important in the brain, where the variability of connections and synaptic weights would most likely work to our advantage and where firing patterns are sparser. Independence of refractoriness would be further improved, if in addition a clever stimulation routine was implemented such that the distribution of stimulation strength varies spatially from trial to trial.

The independence of refractory times is the one factor which makes or breaks our approach. Therefore we may think of ways of making the refractoriness distribution more random. First, it would help to use a task and situation where neurons are as uncorrelated as possible. Second, we may use a set of conditioning pulses of stimulation to increase independence of refractory times. Third, we may utilize chemical, behavioral, or molecular perturbations to randomize refractoriness. There has not, to the authors knowledge, been any attempts in neuroscience to randomize refractory pe-

riods, so there may be a lot of possibilities to improve.

Generalizing this idea, we may ask if there are ways of constructing good instrumental variables. One may assume a way of building molecular oscillators or otherwise pattern generators into neurons which affect their firing rate. Such modulatory activity would be observable in the extracellular activity of the neuron. Any neuron that then correlates to this modulation must be post-synaptic of the recorded neuron. Any kind of a signal that is local to a cell and not affected by network activity could produce a meaningful instrumental variable and it might be reasonably doable to link membrane channels to intracellular signal generation. Even if perfect IVs do not exist in brains, we may be able to make them.

There are many techniques for causal inference and most of them are largely unknown in neuroscience and are based on approximating randomness in a world that does not have it. In many cases, one could use regression discontinuity designs if one has a spiking system [**?**, **?**]. One could use a difference in difference approach [**?**]. One can use matching approaches [**?**] where one compares similar network states and their evolution over time. In general, neuroscience is in a quest for causality, we should be able to benefit considerably by utilizing techniques that are popular in the field of causal inference.

4 Methods

4.1 Instrumental variable estimation

A simple approximation of the connectivity strength between a pre-synaptic neuron x and a post-synaptic neuron y can be to ignore external excitation and simply calculate the relation between the spike times in x and y with a regression model given by

$$y = \beta x + u. \tag{1}$$

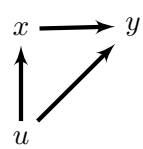
Here y is the dependent variable, x is the explanatory variable, β is the slope of the x, y curve and u is an unknown error term. This system follows the path diagram

$$\begin{tikzcd} x \arrow[r] & y \\ u \arrow[ur] & \end{tikzcd}$$

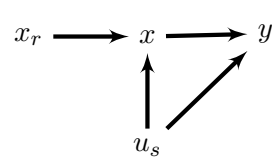
Assuming that changes in spike times y are described by βx i.e. $\frac{\mathrm{d}y}{\mathrm{d}x} = \beta$ for spike times x. One problem with this idea is that in a confounded system, perfectly correlated neurons will give statistically indistinguishable β. In the extreme case where two neurons are both made to fire every time they are stimulated, they will have the same weights according to Eq. (1), after all, during stimulation $y = 1$ for both, even if only one of them drives the post-synaptic neuron. Another problem is if the network state affects both the probability of a neuron to fire and also the probability of post-synaptic neurons to fire. In this case, the network state can induce a correlation which will make the estimation highly biased. Arguably, the network state will, in all realistic models, have a dramatic influence on all neurons and the regression model is rather described by

$$y = \beta x + u(x). \tag{2}$$

Corresponding to the following path diagram [Wright, 1921]



Here we have the relation $\frac{\mathrm{d}y}{\mathrm{d}x} = \beta + \frac{\mathrm{d}u}{\mathrm{d}x}$. To get at causality we thus require some stimulation that only highlight the activity in y caused by x, disassociating x from u. The optogenetic stimulation is not specific to x and will activate parts of the network activity u leaving only a subset of u correlated with x, namely u_s. To disassociate x from u_s we need something that can distinguish between different neurons that are stimulated. We thus require some instrument x_r which is (1) uncorrelated with the network u and (2) is correlated with the regressor x[Angrist and Pischke, 2008]. We assume that the neurons are independent at small time scales and that stimulation additionally randomize membrane potential individually in neurons. We may thus use the fact that a neuron that has fired just before the stimulation will be in absolute refractory period and hence have $x_r = 0$ independently of u_s. This introduces times where the spike from one of the stimulated neurons are missing. Thus we may use the refractory period as an instrumental variable illustrated with the following path diagram



Here x_r represent times where the pre-synaptic neuron is refractory during stimulation. This is then an estimator that compares the post-synaptic activity when a given neuron is non-refractory with the post-synaptic activity when it is, thus removing the confounding. The true β is given by

$$\beta_{IV} = \frac{\mathrm{d}y}{\mathrm{d}s_r} / \frac{\mathrm{d}x}{\mathrm{d}s_r} \tag{3}$$

Since our instrument s_r is binary we may use the IV (or more precisely Wald) estimator [Wald, 1940, Cameron and Trivedi, 2005] to estimate β_{IV} by

$$\hat{\beta}_{IV} = \frac{\bar{y}_s - \bar{y}_{s_r}}{\bar{x}_s - \bar{x}_{s_r}} = \bar{y}_s - \bar{y}_{s_r} \tag{4}$$

Here $\bar{y}_s$ is the average number of trials where stimulating x resulted in a response in y and $\bar{y}_{s_r}$ is the average number of trials where an unsuccessful stimulation resulted in a response in y.

To utilize the refractory period as an IV we first picket out one window of 4 ms for each of the pre-synaptic and post-synaptic neuron with a latency of 0 and $\tau_{syn} + D$ ms (see Eq. (10)) respectively. By binary classifying each window for each trial weather it contained a spike we obtained the two binary arrays y_s and y_{sr}.

4.2 Cross correlation histogram

The statistical tests giving the probabilities p_{diff} and p_{fast} were done according to [Stark and Abeles, 2009, English et al., 2017]. Briefly, to test if the cross correlation histogram (CCH) peak was significant we employed two tests. By using the Poisson distribution with a continuity correction [Stark and Abeles, 2009] given by Eq. (5) we calculated p_{diff} by comparing the peak in positive time lag with the maximum peak in negative time lag [English et al., 2017]. The probability p_{fast} represents the difference between CCH and it's convolution with a hollow Gaussian kernel [Stark and Abeles, 2009].

$$p(N|\lambda(m)) = 1 - \sum_{k=0}^{N-1} \frac{e^{-\lambda(m)}\lambda(m)^k}{k!} - \frac{e^{-\lambda(m)}\lambda(m)^N}{2N!} \tag{5}$$

Here λ represents the counts at bin m and N is the number of bins considered. To estimate the connection weight between pairs we used the spike transmission probability defined in [English et al., 2017] as

$$P_{trans} = \frac{1}{n} \sum_{m=4ms}^{8ms} CCH(m) - \lambda_{Gauss}(m), \tag{6}$$

where n is the number of spikes detected in the presynaptic neuron.

4.3 Logistic regression

To utilize the refractory period without using it as an IV we estimated synaptic weights using a logistic regression. To do this we first picket out one window of 4 ms for each of the pre-synaptic and post-synaptic neuron with a latency of 0 and $\tau_{syn}+D$ ms (see Eq. (10)) respectively. By binary classifying each window for each trial weather it contained a spike we obtained two binary arrays, the regressor x and the dependent variable y where we want to estimate the probability $P(y = 1|x)$ by fitting the parameters $\boldsymbol{\beta}$ such that

$$y = \begin{cases} 1 & \text{if } \beta_0 + \beta_1 x + u > 0 \\ 0 & \text{else} \end{cases} \tag{7}$$

where u is an error term. Further, we used the logit link function such that the the probability giving the proxy for synaptic weight is given by

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}} \tag{8}$$

The model was fitted using the python package scikit-learn [Pedregosa et al., 2011]

4.4 Simulated network

To simulate a recurrent network of excitatory and inhibitory neurons we used the leaky integrate and fire (LIF) model given by

$$\frac{\mathrm{d}V_m^i}{\mathrm{d}t} = -\frac{(V_m^i - E_L)}{\tau_m} + \frac{I_{syn}^i(t)}{C_m}. \tag{9}$$

When the membrane potential V_m^i of neuron i reaches a threshold V_{th} an action potential is emitted and V_m^i reset to the leak potential E_L followed by an absolute refractory period τ_{ref}. The membrane time constant is represented by τ_m and $I_{syn}^i(t)$ denotes the post synaptic current (PSC) for neuron i modeled as a sum of alpha functions given by

$$I_{syn}^i(t) = \sum_{j=1}^{C} J_j \alpha(t - t_j - D), \tag{10}$$

where t_j denotes an incoming spike through synapse j at delay D and C is the number of incoming synapses on neuron i. The PSC amplitude is given by J_j and the alpha function is given by

$$\tau_{syn}\alpha(t) = te^{-\frac{t}{\tau_{syn}}}H(t). \tag{11}$$

Here τ_{syn} denotes the synaptic integration time constant and H is the Heaviside step function. All neurons were driven by an external Poisson process with rate $rate_p$.

Synaptic weights were log-normally distributed such that the increase in membrane potential V_m^i due to one spike were restricted to lie between $V_{syn} = 0.05mV$ and $V_{syn} = 2.05mV$ based on experimental findings [Sayer et al., 1990, Mason et al., 1991]. The synaptic distribution is shown in Fig. 6(c) where the inhibitory PSC amplitude is given by $J_{in} = gJ_{ex}$ where J_{ex} denotes the excitatory synaptic weight.

To find suitable parameters yielding asynchronous activity we measured the population correlation coefficient given by

$$\langle CC \rangle_{pop} = \left\langle \left\langle \frac{h_i - \langle h_i \rangle}{std(h_i)} \frac{h_j - \langle h_j \rangle}{std(h_j)} \right\rangle \right\rangle_{pop}, \tag{12}$$

where h is the spike time histogram with binsize at $5ms$ for neuron i, j and $\langle \cdot \rangle$ is the mean operator. The distribution of CC is shown in Fig. 3 which were found by performing several parameter sweeps picking three parameter sets which mainly differed in firing rate (data not shown).

To further evaluate the network state we calculated the coefficient of variance (CV) of the population given by

$$\langle CV \rangle = \left\langle \frac{std(ISI_i)}{\langle ISI_i \rangle} \right\rangle_{pop}, \tag{13}$$

where ISI denotes the inter-spike interval of neuron i. Due to the finite time synaptic integration time constant $\tau_{syn} = 1ms$ we were unable to have the network showing an irregular state; see Fig. 6(b). To verify that indeed this was due to τ_{syn} we performed several simulations with lower τ_{syn} obtaining $\langle CV \rangle_{pop} > 1$ (data not shown). It would likely be easier to achieve irregular network state if synapses were conductance based [Kumar et al., 2008]. However, we settled with current based synapses as we were mainly interested in achieving an asynchronized state ($\langle CC \rangle_{pop} < 0.01$).

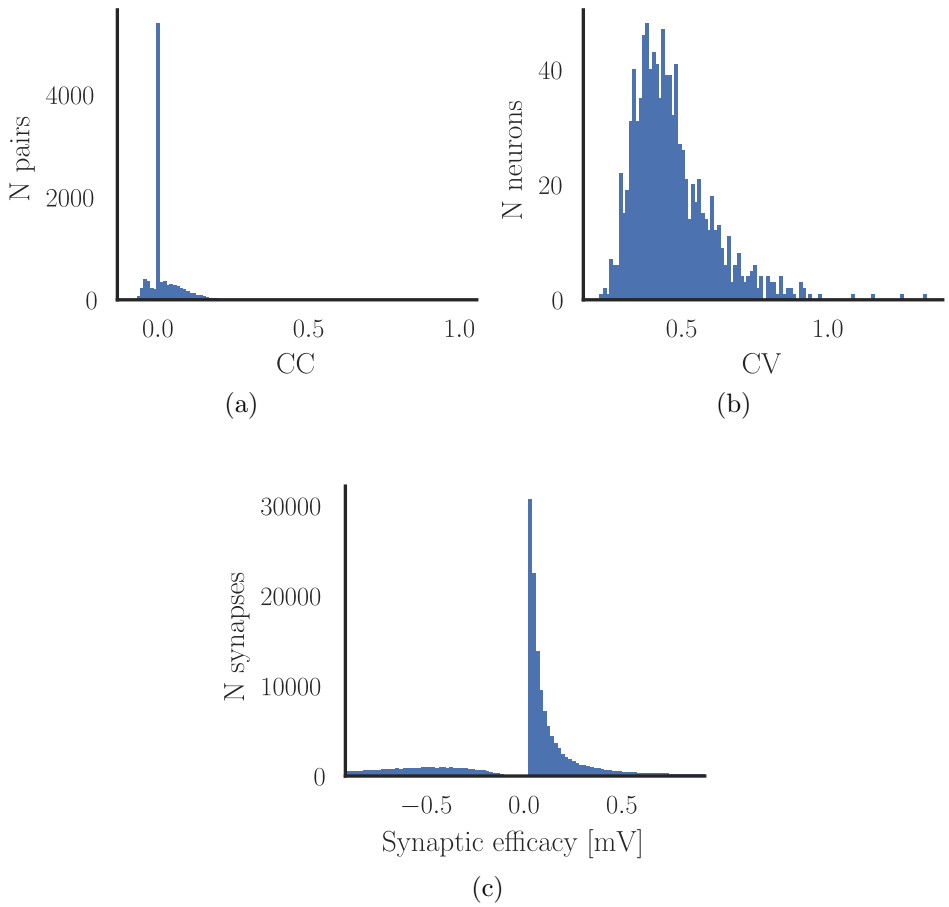


Figure 6: Network state

	model 1	model 2	model 3	units
$N_{neurons}$	1250			
Δt	0.1			
N_{ex}	1000			
N_{in}	250			
eta	0.9			
$rate_p$	3694.26			Hz
V_{reset}	0			mV
V_m	0			mV
E_L	0			mV
t_{ref}	2			ms
τ_m	20			ms
V_{th}	20			mV
C_m	1			pF
V_{syn}	0.2			mV
g	9.9	4.4	3	
V_{syn}^{high}	2.05			mV
V_{syn}^{low}	0.05			mV
var_{syn}	0.5			mV^2
τ_{syn}^{in}	1			ms
τ_{syn}^{ex}	1			ms
$delay$	1.5			ms
eps	0.1			
C_{ex}	100			
C_{in}	25			
J_{in}	0.88727	0.394342	0.26887	pA
J_{ex}	0.0896232			pA
J_{high}^{ex}	0.918638			pA
J_{low}^{ex}	0.0224058			pA
J_{high}^{in}	0.918638			pA
J_{low}^{in}	0.0224058			pA
$stim_N^{in}$	0			
$stim_N^{ex}$	800			
$stim_{amp}^{in}$	0			pA
$stim_{amp}^{ex}$	10			pA
$stim_{duration}$	2			ms
$stim_{period}$	100			ms
$stim_{max}^{period}$	150			ms
$rate_{in}$	7.17024	9.66335	11.754	Hz
$rate_{ex}$	9.05793	10.9176	13.1032	Hz
$density$	100000			Nmm^{-3}
S	10.3			mm^{-1}
NA	0.37			
r	0.1			μm
n	1.36			

Table 1: Simulation parameters of three different models.

4.5 Perturbation intensity

In order to replicate an optogenetic experiment we modeled transmission of light through brain tissue with the Kubelka-Munk model for diffuse scattering in planar, homogeneous, ideal diffusing media given by

$$T = \frac{1}{Sz + 1}. \tag{14}$$

Here T denotes a transmisison fraction, S is the scattering coefficient for mice [Aravanis et al., 2007] and z is the distance from a light source [Ho et al., 2017]. Further we combined diffusion with geometric loss assuming that absorption is negligible as in [Aravanis et al., 2007] and computed the intensity as presented in Fig. 1 by

$$\frac{I(r)}{I(r=0)} = \frac{\rho^2}{(Sr+1)(r+\rho)^2} \tag{15}$$

where r is the distance from the optical fiber and

$$\rho = \frac{d}{2}\sqrt{\left(\frac{n}{NA}\right)^2 - 1}. \tag{16}$$

Here d is the diameter of the optical fiber, NA is the numerical aperture of the optical fiber and n is the refraction index for gray matter [Ho et al., 2017]; see numerical values for parameters in Table 1.

To estimate the distribution of light intensity on affected neurons we assumed a neuron density of $10^4 Nmm^{-3}$ and found the volume of a cut cone that could contain the number of stimulated excitatory neurons which were found to yield the depth $r_{max} = 0.175mm$. We then selected N_stim neurons that were given a random position in the range $[0, r_{max}]$ and were assigned a stimulation strength as the maximum stimulation strength multiplied by Eq. (15). Then we selected 50 of the excitatory neurons that were not stimulated as the "target" population which together with the inhibitory neurons were not perturbed directly by the light stimulus.

In order to keep the stimulus model as simple as possible we let set the maximum stimulation strength to $10pA$ which was found suitable by investigating the percentage of successful stimulations to be around 50%.

References

[Angrist and Pischke, 2008] Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

[Aravanis et al., 2007] Aravanis, A. M., Wang, L. P., Zhang, F., Meltzer, L. A., Mogri, M. Z., Schneider, M. B., and Deisseroth, K. (2007). An optical neural interface: in vivo control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.*, 4(3):S143–S156.

[Boyden et al., 2005] Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., and Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience*, 8(9):1263.

[Cameron and Trivedi, 2005] Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

[Card, 1993] Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research.

[Daniusis et al., 2012] Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. (2012). Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*.

[Drton et al., 2011] Drton, M., Foygel, R., and Sullivant, S. (2011). Global identifiability of linear structural equation models. *The Annals of Statistics*, pages 865–886.

[Emiliani et al., 2015] Emiliani, V., Cohen, A. E., Deisseroth, K., and Häusser, M. (2015). All-optical interrogation of neural circuits. *Journal of Neuroscience*, 35(41):13917–13926.

[English et al., 2017] English, D. F., McKenzie, S., Evans, T., Kim, K., Yoon, E., and Buzsáki, G. (2017). Pyramidal Cell-Interneuron Circuit Architecture and Dynamics in Hippocampal Networks. *Neuron*, 96(2):505–520.

[Ho et al., 2017] Ho, A. H. P., Kim, D., and Somekh, M. G. (2017). *Handbook of photonics for biomedical engineering*.

[Honey et al., 2009] Honey, C., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., and Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040.

[Jonas and Kording, 2017] Jonas, E. and Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Comput. Biol.*, 13(1):1–24.

[Kumar et al., 2008] Kumar, A., Schrader, S., Aertsen, A., and Rotter, S. (2008). The high-conductance state of cortical networks. *Neural Comput.*, 20(1):1–43.

[Lerman et al., 2017] Lerman, G. M., Gill, J. V., Rinberg, D., and Shoham, S. (2017). Two photon holographic stimulation system for cellular-resolution interrogation of olfactory coding. In *Optics and the Brain*, pages BrM3B–5. Optical Society of America.

[Mason et al., 1991] Mason, a., Nicoll, A., and Stratford, K. (1991). Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro. *J. Neurosci.*, 11(January):72–84.

[Nikolenko et al., 2007] Nikolenko, V., Poskanzer, K. E., and Yuste, R. (2007). Two-photon photostimulation and imaging of neural circuits. *Nature methods*, 4(11):943.

[Paré et al., 1998] Paré, D., Shink, E., Gaudreau, H., Destexhe, A., and Lang, E. J. (1998). Impact of spontaneous synaptic activity on the resting properties of cat neocortical pyramidal neurons in vivo. *Journal of neurophysiology*, 79(3):1450–1460.

[Pearl, 2009] Pearl, J. (2009). *Causality*. Cambridge university press.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Peters et al., 2017] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT Press.

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

[Rudolph and Destexhe, 2006] Rudolph, M. and Destexhe, A. (2006). On the use of analytical expressions for the voltage distribution to analyze intracellular recordings. *Neural computation*, 18(12):2917–2922.

[Sayer et al., 1990] Sayer, R. J., Friedlander, M. J., and Redman, S. J. (1990). The time course and amplitude of EPSPs evoked at synapses between pairs of CA3/CA1 neurons in the hippocampal slice. *J. Neurosci.*, 10(3):826–836.

[Shimizu et al., 2006] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

[Stark and Abeles, 2009] Stark, E. and Abeles, M. (2009). Unbiased estimation of precise temporal correlations between spike trains. *J. Neurosci. Methods*, 179:90–100.

[Stevenson et al., 2008] Stevenson, I. H., Rebesco, J. M., Miller, L. E., and Körding, K. P. (2008). Inferring functional connections between neurons. *Current opinion in neurobiology*, 18(6):582–588.

[Wald, 1940] Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

[Wassermann, 2006] Wassermann, L. (2006). All of nonparametric statistics. *New York*.

[Wright, 1921] Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557–585.

[Zemelman et al., 2002] Zemelman, B. V., Lee, G. A., Ng, M., and Miesenböck, G. (2002). Selective photostimulation of genetically charged neurons. *Neuron*, 33(1):15–22.