Transmission matrix inference via pseudolikelihood decimation

Daniele Ancora¹,∗ and Luca Leuzzi¹,²

¹ Dipartimento di Fisica, Università di Roma ‘Sapienza’, Piazzale A. Moro 2, 00185, Roma, Italy
² Institute of Nanotechnology, Consiglio Nazionale delle Ricerche,
CNR-NANOTEC-Soft and Living Matter Laboratory, Piazzale A. Moro 2, 00185, Roma, Italy

E-mail: daniele.ancora@uniroma1.it

Received 27 January 2022, revised 27 July 2022
Accepted for publication 23 August 2022
Published 7 September 2022

Abstract
Recently, significant efforts in medical imaging are towards the exploitation of disordered media as optics tools. Among several approaches, the transmission matrix description is promising for characterizing complex structures and, currently, has enabled imaging and focusing through disorder. In the present work, we report a statistical mechanics description of the transmission problem. We convert a linear input–output transmission recovery into the statistical inference of an effective interaction matrix. We do this by relying on a pseudolikelihood maximization process based on random intensity observations. Our aim is to bridge results from spin-glass theory to the field of disordered photonics, uncovering insights from the scattering problem and encouraging the development of novel imaging techniques for better medical investigations.

Keywords: pseudolikelihood, transmission matrix, statistical inference, disordered optics

(Some figures may appear in colour only in the online journal)

1. Introduction

Linear problems are largely used in many scientific fields, among which modern topics range from general machine learning aspects [1, 2], to spectra deconvolution [3], stock market forecast [4], till the generalization of signal transmission through disordered systems [5]. Despite its generality, our study finds its principal motivation in the latter application. The non-trivial problem of the transmission matrix recovery of an opaque device would enable its usage
as an ordinary optical tool, capable of focussing or transmitting an image through disorder [6, 7], turning the scattering into a beneficial feature exploitable to enhance resolution. Turbidity is a feature of the random media that depends upon the microscopic arrangement of its constituents and, to investigate imaging in such conditions, several approaches have been studied [8]. The light traveling through a non-uniform environment experiences several changes in its original trajectory due to the complex distribution of the refractive index, losing the memory of its initial direction. Although the microscopic process is complicated, the overall effect can be efficiently described by the transmission matrix approach [5]. Such matrix describes how a disordered medium linearly transports a given input to the recorded output, formalizing its transmission rule [5, 9, 10]. Defining a robust way to measure such a matrix is a challenging problem in the field of disordered optics [11]. It may allow using new tools for biomedical inspections, such as multimode-fibers [12] or scattering structures designed ad-hoc [13]. It is not the interest of this paper to describe the benefits for which the community pursues toward this goal; the interested reader could follow modern (and more specific) bibliography [11, 14, 15]. The state-of-art for the calculation of the transmission matrix (from now on $T$) is the method introduced by Popoff et al [5], which uses the Hadamard basis as input and reconstruct the $T$ based on output observations. Although the protocol has proven excellent performance, one would prefer a less strict sampling approach, ideally not relying on a pre-defined input basis. Moreover, the method works in the forward direction (input to output) and permits the reconstruction of the direct transmission matrix. Operatively, to control the turbid device, it is required to invert such $T$. In the presence of noise, matrix inversion is a delicate process that renders the system very sensitive to perturbation, where it is often preferable to use other approaches [16]. The need for a procedure that is robust and efficient also for random sampling datasets is the reason why we approach statistics.

With this work, we want to offer another perspective for the estimation of the transmission matrix via a random sampling statistical approach. The randomization of the input and the relative complex response of the system renders the problem similar to those studied in the field of disordered systems. In fact, the motivation behind our study is to borrow tools and procedures from statistical physics into the description of transmission channels. We will link concepts from Ising spin models, interaction Hamiltonian formulations, and information criteria to estimate the number of couplings. We aim at switching the paradigm of the transmission matrix reconstruction to the field of disordered systems. The model we propose in the following is directly derived from the linear propagation of light in arbitrary media [17], and it is interpretable as a spin-glass-like model with input/output variables (I/O) rather than spins. This translation permits us to borrow statistical tools to tackle the inference of the $T$-parameters. In particular, we rearrange the formalism to define an effective interaction matrix of the system that contains more information than the sole transmission rule. To recover this, we make use of the pseudolikelihood maximization approach [18, 19] coupled with a decimation strategy [20]. This solution allows us to estimate and select the most representative parameters of the I/O model considered. In this way, we demonstrate that the interaction matrix of such a system converges to a block matrix that contains the estimations for the transmission, information about the channel noise, and an input self-coupling rule. Using the pseudolikelihood formalism allows the separation of the problem, making it linearly scalable (with the number of parameters) and easy to distribute in parallel on GPU hardware. Indeed, our approach can be seen as a thermodynamical description of multiple and multivariate linear regression problems, leaving it applicable to any situation of this kind.
Figure 1. Schematics of the numerical experiment considered. We want to transmit an input signal through a channel characterized by an unknown linear response. The response is fully encoded in the transmission matrix of the system, that gives access to the correct exploitation of the communication channel in both directions.

We introduce, in the following, the statistical mechanical problem implicated by reconstructing the intensity transmitted through a disordered channel. To start, we create a dataset of random inputs, collecting the noisy measurements at the output after the propagation through an unknown-yet-fixed-channel. In the specific case study, we consider a sparse $\mathcal{T}$, and we incorporate a decimation procedure to perform feature selection during the optimization. We devise a thorough study concerning this aspect, analyzing several information criteria used to stop the decimation. Finally, we use our formalism to recover both the direct and the inverse transmission, concluding our study with the reproduction of a focusing and imaging experiment (figure 1).

2. Signal transmission through a disordered device

Let us start by considering a linear problem with multiple input–output transmission channels. To keep the analogy with disordered photonics, we could interpret it as an I/O system in which a random medium (such as a multimode fiber) acts as a linear light scrambler:

$$E_{\text{out}} = \mathcal{T}E_{\text{in}} + \varepsilon,$$

where $E_{\text{in}}$ is the controlled input, $E_{\text{out}}$ is the recorded output, and $\varepsilon$ is a complex random noise. In this case, $\mathcal{T}$ is the electromagnetic transmission matrix [5, 21]. Unfortunately, due to technological limitation, we are always bound to learn $\mathcal{T}$ from measures of output intensities

$$I_{\text{out}} = \|E_{\text{out}}\|^2$$

even when we have complete control over the input field [22, 23]. Thus, to infer the transmission parameters, an intensity model is the starting point:

$$I_{\text{out}} = \|\mathcal{T}E_{\text{in}}\|^2 + \varepsilon.$$

In this preliminary disordered system description, we begin our study by considering an intensity-only model that would be exact in the case of imaging of incoherent sources (such as fluorescence, phosphorescence, black body radiation, incandescent lamps, etc). Any input vector of intensity distribution $I^n = \{I^n_\eta\}, \eta = 1, \ldots, N_I$, is connected to the output pattern $I^\text{out} = \{I^\text{out}_\gamma\}, \gamma = 1, \ldots, N_O$, via an intensity transmission matrix $\mathcal{T} = \{T^{n}_{\gamma\eta}\}$ by the rule:

$$I^\text{out} = \mathcal{T}I^n + \varepsilon, \quad I^\text{out}_\gamma = \sum_{\eta=1}^{N_I} T^{n}_{\gamma\eta} I^n_\eta + \varepsilon_\gamma.$$
For simplicity, we assume the number of input modes to be equal to the number of modes recorded at the output, thus \( N_0 = N_I = N/2 \) with \( N = N_I + N_0 \) and \( \gamma, \eta = 1, \ldots, N/2 \). In the last term of equation (2), the noise terms \( \epsilon_\gamma \) take into account possible statistical perturbations in the measurement, due, e.g., to stochastic temperature changes or to bending or vibrations of the fiber. Such \( \epsilon \) term is assumed to be a random vector with zero-mean entries and finite mean square displacement \( \sigma_\gamma \). Initially, we will impose \( \sigma_\gamma = 0, \forall \gamma \). We want to stress that equation (2) describes intensity interactions only and would work rigorously in case of incoherent monochromatic sources. Similar assumptions, however, were used for image reconstruction through fiber bundles even in broadband illumination schemes [24]. Thus, in the present representation, we will consider \( \mathbb{T} \) as a matrix connecting the intensities between the fiber’s ends, not the electromagnetic transmission matrix of the disordered device. Without losing generality, from now on, we refer to it as effective-\( \mathbb{T} \) or, simply, transmission matrix.

3. Multivariate least-squares fitting

The transmission problem described in the previous chapter involves the input of an array of pixels, of which we have complete control. Therefore, we are able to multivariate the input. During the transmission, the \( I^i \) is transported by the disordered device in a random and, a priori, unpredictable fashion. Because of this, the probed transmission channels output an image that is the result of the linear light-field scrambling. Within each pixel, the linear combination of all the inputs intensity concurs in the formation of the intensity output. That said, each output pixel underlines a multivariate least-square fitting:

\[
\mathbb{T}_{\text{inf}} = \arg\min_{\mathbb{T}} \sum_{\gamma} \left( I_{\text{out}}^{\gamma} - \sum_{\eta=1}^{N/2} T_{\gamma\eta} I_{\text{in}}^{\eta} \right)^2.
\] (3)

Since there are many pixels \( \gamma \) in the output image, the problem in its entirety is called a multiple multivariate linear regression and can be solved independently from one to another. In general, the solution can be approached by convex optimization or, if the number of output measurements is sufficient, by direct inversion [25]. In fact, when the number of measures \( M \) oversamples the problem \( (M > N) \), the solution can be explicitly found by multiplying \( I^\text{out} = \mathbb{T} I^\text{in} \) by the transpose of the measurement matrix:

\[
I^\text{out} I^\text{in} = \mathbb{T}^\dagger I^\text{in},
\] (4)

where, with the \( \dagger \) we indicate the transposition operation. In the expression above, \( (I^\text{in} I^\text{in})^\dagger \) is the transpose of the Gramian matrix (of the measurement \( I^\text{in} \)) that is invertible. This permits us to write the solution to the problem explicitly as:

\[
\mathbb{T}_{\text{inf}} = I^\text{out} I^\text{in} (I^\text{in} I^\text{in})^{-1}.
\] (5)

Please note that we have defined, here, the transmission matrix as an operator acting on the left side\(^3\). Once the matrix \( \mathbb{T}_{\text{inf}} \) has been inferred, the variance of the noise can be estimated a posteriori by computing the residuals per each output channel. This procedure is exact when the noise follows Gaussian statistics. In the following, we will derive an analogous expression

\(^3\)It is easy to find often alternative definitions, in which the operator to be inferred acts on the right side, as in \( y = x^\mathbb{T} \). In the latter case, the solution is \( \mathbb{T} = (x^\dagger x)^{-1} x^\dagger y \) where, between brackets, we have the Gramian matrix.
from principles of statistical mechanics, importing tools normally used in this discipline to put at the service of transmission matrix recovery.

4. Inverse statistical mechanics approach

Let us now build our inference model based on disordered systems’ theory. Satisfying the system equation (2), at zero noise, means that:

$$\int \frac{N/2}{\prod_{\gamma} I_{\text{out}}^\gamma \prod_{\eta} I_{\text{in}}^\eta} \delta \left( I_{\text{out}}^\gamma - \sum_{\eta=1}^{N/2} T_{\gamma\eta} I_{\text{in}}^\eta \right) = 1$$  \hspace{1cm} (6)

for any given entry $I_{\text{in}}^\gamma$. This is exact in the case of zero noise. To see what happens in presence of noise, we first rewrite the integral (6) expressing each $\delta$-function with a Gaussian in the limit of its standard deviation $\sigma_{\gamma}$ going to zero:

$$\lim_{\sigma_{\gamma} \to 0} \int \frac{N/2}{\prod_{\gamma} I_{\text{out}}^\gamma} \prod_{\eta} \exp \left\{ -\frac{1}{2\sigma_{\gamma}^2} \left( I_{\text{out}}^\gamma - \sum_{\eta=1}^{N/2} T_{\gamma\eta} I_{\text{in}}^\eta \right)^2 \right\} = 1.$$  \hspace{1cm} (7)

Keeping $\sigma_{\gamma}$‘s finite amounts to allow Gaussian noise, recovering the noise statistics introduced in equation (2) for a perturbative term that follows a Gaussian distribution. For the sake of generality, we consider different noise for each output channel $\gamma$. If $\sigma_{\gamma} = \sigma, \forall \gamma$, the exponent in equation (7) can be expressed as a bilinear cost function:

$$H = N/2 \sum_{\gamma=1}^{N/2} \left( I_{\text{out}}^\gamma - \sum_{\eta=1}^{N/2} T_{\gamma\eta} I_{\text{in}}^\eta \right)^2$$

$$= \sum_{\gamma=1}^{N/2} \left( I_{\text{out}}^\gamma \right)^2 - 2 \sum_{\gamma,\eta}^{N/2} T_{\gamma\eta} I_{\text{in}}^\eta I_{\text{in}}^\eta + \sum_{\gamma,\eta,\eta'}^{N/2} T_{\gamma\eta} T_{\gamma\eta'} I_{\text{in}}^\eta I_{\text{in}}^{\eta'}$$

$$\equiv -\sum_{i,j} I_{i} J_{ij} I_{j} = -\mathbf{I} \mathbf{J} \mathbf{I},$$  \hspace{1cm} (8)

where we define the concatenated input–output intensity vector $\mathbf{I} = \{ I_{\text{in}}^\gamma, I_{\text{out}}^\gamma \}$ and new indices are defined such that $i, j = \{ \gamma, \alpha + N/2 \} = \{ 1, \ldots, N \}$. In this description, $\mathbf{I}$ is a vector in which the second half (the output) depends on the first (input). The interaction matrix, $\mathbf{J} = \{ J_{ij} \}$, has the form of a four-block tensor containing the transmission matrix $\mathbb{T}$ in the out-of-diagonal blocks:

$$\mathbf{J} = \left( \begin{array}{cc} -\mathbf{U} & \mathbb{T} \\ \mathbb{T}^\dagger & -\mathbf{I} \end{array} \right) \hspace{1cm} (9)$$

Here, we can also recognize the input self-coupling matrix $\mathbf{U} = \mathbb{T}^\dagger \mathbb{T}$ as the Gramian matrix, and the identity matrix $\mathbf{I}$. The latter form of equation (8) is similar to the spin-glass Hamiltonian.

In the statistical mechanics’ formulation, $\mathcal{H}$ is the Hamiltonian of the system of which the interaction matrix $\mathbf{J}$ (structured as in equation (9)) is the subject of our inference procedure.
Writing $\beta = 1/2 \sigma^2$, the stationary probability to have an input–output configuration $\mathbf{I}$ with a certain energy given the set of interaction $\{J_{ij}\}$ is the Boltzmann distribution:

$$P(\mathbf{I} | \mathbb{J}) = \frac{1}{Z(\mathbb{J}, \mathbf{I})} e^{-\beta \sum_{i}^{N} J_{ij} I_{ij}}. \quad (10)$$

In this context, the normalization $Z$ is the canonical partition function that normalizes the probability distribution:

$$Z(\mathbb{J}, \mathbf{I}) = \left( \frac{1}{2\pi \sigma^2} \right)^{N/2} \int \prod_{j} dI_{j} e^{-\beta H(\mathbf{J}, \mathbf{I})}. \quad (11)$$

By having defined the probability, it is straightforward to write the log-likelihood $L$ of the system. The function $L$ describes how the set of interaction parameters $\{J_{ij}\}$ are likely to describe the observed I/O intensity configuration $\mathbf{I}$:

$$L(\mathbf{I} | \mathbb{J}) = \ln \left[ P(\mathbf{I} | \mathbb{J}) \right]. \quad (12)$$

The maximum likelihood estimation (MLE) principle applied to our case states that the interaction matrix $\mathbb{J}_{\text{inf}}$ that maximize the function $L$ is the one that most likely describes the system interactions:

$$\mathbb{J}_{\text{inf}} = \arg \max \ L(\mathbf{I} | \mathbb{J}). \quad (13)$$

In the following, we rely on the MLE principle to find the transmission matrix by inferring the whole interaction matrix $\mathbb{J}$. As described in section 3, the maximization of $L$ is, in principle, possible for the case of intensities and corresponds to equation (3). With a view to further extensions to the case of more complicated systems, in which, e.g., next to the linear transmission local constraints are imposed on the input or output pixels (thus, inducing effective couplings between light modes), we introduce the pseudolikelihood description of the problem.

### 4.1. Pseudolikelihood of the I/O model

The calculation of $L$ can be separated by introducing the concept of the local-likelihood function $L_{i}$ of the $i$th element. We decided to follow a common approach tested in spin-glass inference [17, 27], looking at the probability of the $i$th element conditional to the realization of the others \(\setminus i\) (not $i$):

$$P(I_{i} | I_{\setminus i}) = \frac{P(I_{i}, I_{\setminus i})}{P(I_{\setminus i})}, \quad (14)$$

where $I_{\setminus i}$ is the set of all intensities but $i$. Before proceeding with the calculation of the probability $P(I_{i} | I_{\setminus i})$, it is useful to explicitly separate the Hamiltonian contributions $\mathcal{H}_{i}$ that directly involves the variable $i$.

Starting from equation (7), for a given output pixel variable $\gamma = 1, \ldots, N/2$, we can write its contribution to the (temperature rescaled) Hamiltonian as:

$$H_{\gamma} \equiv \beta_{\gamma} \mathcal{H}_{\gamma}(T_{\gamma}^{\text{out}}(I_{\gamma}^{\text{out}}, I_{\gamma}^{\text{in}})) = \beta_{\gamma} (T_{\gamma}^{\text{out}})^{2} - 2\beta_{\gamma} T_{\gamma}^{\text{out}} \sum_{\eta=1}^{N/2} T_{\gamma\eta} I_{\gamma}^{\text{in}} P_{\eta}^{\text{in}} + \beta_{\gamma} \sum_{\eta\eta'} T_{\gamma\eta}^{\dagger} T_{\gamma\eta'} I_{\gamma}^{\text{in}} P_{\eta}^{\text{in}}, \quad (15)$$

...
where we have explicitly considered a channel-dependent ‘inverse’ noise:

\[ \beta_\gamma \equiv \frac{1}{2\sigma_\gamma^2}. \] (16)

We also define the related partial canonical partition function for the single variable Hamiltonian, given all the other variable values as quenched, as:

\[ Z_\gamma = \int dI_{\text{out}}^\text{out} e^{-H_\gamma(I_{\text{out}}^\text{out}|I_{\text{out}}^\text{out},I_{\text{in}})}. \] (17)

Analogously, the (noise-rescaled) input variable Hamiltonian contributions relevant to the local-likelihood maximization will have the form:

\[ H_\eta(I_{\text{in}}^\text{in}|I_{\text{in}}^\text{in},I_{\text{out}}) \equiv -2I_{\text{in}}^\text{in} \frac{\beta_\eta}{N} \sum_{\gamma=1}^{N/2} \beta_\gamma T_\gamma^\eta T_\gamma^\eta + 2I_{\text{in}}^\text{in} \sum_{\xi<\eta} V_{\eta\xi} I_{\text{in}}^\text{in} + (I_{\text{in}}^\text{in})^2 V_{\eta\eta} \] (18)

\[ Z_\eta = \int dI_{\text{in}}^\text{in} e^{-H_\eta(I_{\text{in}}^\text{in}|I_{\text{in}}^\text{in},I_{\text{out}})}, \] (19)

where we have defined:

\[ V_{\eta\xi} \equiv \sum_{\gamma} \beta_\gamma T_\gamma^\eta T_\gamma^\xi. \] (20)

At this point, we can write the local-probability distribution of the generic \( i \)th variable (input or output) given the parameters \( V, T \) and \( \{ \beta_\gamma \} \) and the intensities \( I \) as:

\[ P(I_i|I_\text{\neg}i) = e^{-H_i}/Z_i, \quad i = 1, \ldots, N \] (21)

and define the corresponding logarithm of the local-likelihood per each element \( i \) as:

\[ \mathcal{L}_i = \ln P(I_i|I_\text{\neg}i) = -H_i - \ln Z_i. \] (22)

In general, the maximization of each individual local-likelihood function \( \mathcal{L}_i \) is carried out by searching in the entire parameter’s space. In this way, from the parallel maximization of each \( \mathcal{L}_i \), the same parameters will be inferred \( N \) times. Since the values after independent maximization procedures are usually different, its average value is taken as the final estimate.

Alternatively, one can exploit the knowledge of the symmetry properties of the system and maximize the sum of all local-likelihoods:

\[ \mathcal{PL} = N \log \prod_{i=1}^{N} P(I_i|I_\text{\neg}i) \] (23)

that we refer to as the logarithm of the pseudolikelihood function. From now on, we drop the ‘log’ prefix, and we refer directly to pseudolikelihood (\( \mathcal{PL} \)) functions without losing specificity.

To put together equations (15) and (18) in equation (23), it is be convenient to introduce \( N/2 \times N/2 \) diagonal matrix \( \beta \), whose elements are the ‘inverse temperatures’ \( \{ \beta_\gamma \} \), and to give the following generic expression in terms of \( I = (I_{\text{\text{in}}}, I_{\text{\text{out}}}) \):
\[ H_i = I_i \sum_{j \neq i} M_{ij} I_j + I_i^2 M_{ii} = I_i B_i - I_i^2 A_i, \quad i = 1, \ldots, N \]  

(24)

with:

\[
M \equiv \begin{pmatrix}
-2\mathcal{V} + \hat{\mathcal{V}} & 2\mathcal{T}^\dagger \beta \\
2\beta \mathcal{T} & -\beta
\end{pmatrix}
\]  

(25)

\[ A_i \equiv -M_{ii}, \quad B_i \equiv \sum_{j \neq i} M_{ij} I_j. \]  

(26)

Here, with the hat \(\hat{\mathcal{V}}\), we denote a matrix that contains only the diagonal part of \(\mathcal{V}\). The local partition function is, then, written as:

\[ Z_i = \int dI_i e^{-H_i}. \]  

(27)

By definition the diagonal part of \(M\) is always negative, cf equations (20) and (25): \(A_i = \sum_{j=1}^{N/2} \beta_j T_{ji}^2 > 0\), for \(i = 1, \ldots, N/2\) and \(A_i = \beta_{i-N/2} > 0\), for \(i = N/2 + 1, \ldots, N\). In this way, the partition function converges without worrying about the intensity saturation cut-off, that is, the dominion of \(I_i\) in integral equation (27). We stress that, since \(M\) is symmetric, only the elements \(M_{i \leq j}\) will be inferred since there are at most \(N(N+1)/2\) independent terms. When considering the pseudolikelihood maximization, we stress that we infer the pseudo-interaction matrix \(M\) rather than the \(J\) matrix of equation (9). The form of equation (25) is the one of a purely transmission problem. Specifically the lower diagonal block is a diagonal matrix and \(\mathcal{V}\) is the convolution of the transmission matrix elements, equation (20).

4.2. Which partition function?

\(L_i\) depends upon the choice of the integration range in the partition function. This choice is bound to the values each \(I_i\) can take. First of all, it is possible to analytically solve the indefinite integral of equation (27), obtaining:

\[ Z_i(I_i) = \sqrt{\frac{\pi}{4A_i}} e^{\frac{B_i^2}{4A_i}} \text{erf}\left(\frac{2A_i I_i - B_i}{\sqrt{4A_i}}\right) = \sqrt{\frac{\pi}{4A_i}} e^{\frac{B_i^2}{4A_i}} F(I_i). \]  

(28)
The only term dependent on \( I_i \) is the argument of \( F \), which changes upon different choice of the integral limits \([I_{\min}, I_{\max}]\) as, for instance:

\[
F_i = \begin{cases} 
2 & \text{for } F_i|_{-\infty}^\infty \\
1 + \text{erf} \left( \frac{B_i}{\sqrt{4A_i}} \right) & \text{for } F_i|_{0}^\infty \\
\text{erf} \left( \frac{2A_i - B_i}{\sqrt{4A_i}} \right) + \text{erf} \left( \frac{B_i}{\sqrt{4A_i}} \right) & \text{for } F_i|_{0}^1 \\
\text{erf} \left( \frac{2A_i - B_i}{\sqrt{4A_i}} \right) + \text{erf} \left( \frac{2A_i + B_i}{\sqrt{4A_i}} \right) & \text{for } F_i|_{-1}^1 
\end{cases}
\] (29)

In case of pure transmission, that we analyze here, the choice of \( F_i \) does not alter the inference process since, in standard linear least-square problems, we usually do not make any assumption on the partition function when solving equation (3). However, an optimal selection of the integral limits may be useful to account for all the possible intensities allowed in the system. Thus, in our statistical framework, the \( L_i \) can be explicitly written as:

\[
L_i = I_i B_i - I_i^2 A_i - \frac{1}{2} \ln \left( \frac{\pi}{4A_i} \right) - \frac{B_i^2}{4A_i} - \ln \frac{F_i|_{I_{\min}}^{I_{\max}}}{I_{\min}}. 
\] (30)

To have a graphical view on the various functions introduced in equation (30), we plot the generic pseudolikelihood as a function of the parameters \( \exp(L_i) = P(A_i, B_i | I_i) \) given \( I_i \), for different choices of the integration extremes in figure 2. We can qualitatively appreciate that the maxima of the functions follow similar directions along the plane \((A_i, B_i)\). As \( I_i \) varies, the maximum of \( P(A_i, B_i | I_i) \) varies\(^4\). We reported four cases, though the simplest choice of \( F_i|_{-\infty}^{-\infty} = 2 \) turns out to be efficient for all the intensities tested.

4.3. Pseudolikelihood derivatives

The formulation of the pseudolikelihood (30) defines a concave function that admits (under low enough noise) a unique maximum. To ease the maximization process is, thus, reasonable to calculate its first derivatives. Let us first compute the derivatives of the parameters \( A_i' \) and \( B_i' \), defined in equation (26), to simplify the notation:

\[
A_i' = -\delta_{im} B_i', \quad B_i' = \delta_{im} A_i', \quad \delta_{ij} = \delta_{m} \delta_{jm}.
\] (31)

With these, it is possible to calculate the derivatives of \( L_i \) with respect to the parameters of the interaction matrix \( M \):

\[
\frac{\partial L_i}{\partial M_{nm}} = I_i B_i' - \frac{1}{F_i} \frac{\partial F_i}{\partial M_{nm}}, 
\] (32)

\[
\frac{\partial L_i}{\partial M_{nn}} = -I_i^2 A_i' + \frac{A_i'}{2A_i} - \frac{1}{F_i} \frac{\partial F_i}{\partial M_{nm}}.
\] (33)

Nonetheless, the derivatives of \( F_i \) depend upon the choice of the integration limits. Thus, for the cases analyzed in section 4.2, we have that the out-of-diagonal terms of \( M \) can be expressed

\(^4\)We do not have a unique maximum, since it will be given by the choice of the \( \{I_{ij}\} \) rather than the two parameters of equation (26).
Figure 2. Pseudolikelihood for the $i$th intensity elements given all the others. The function is plotted against the parameters $A_i$ and $B_i$, using $I_i = 0.25$ as a reference value for the intensity. In the first plot we draw the contours lines for a different case $I_i = 0.10$. Although the plots differ in terms of absolute values, it is possible to notice how the maximum follow the same direction in all the cases.

\[
\frac{\partial F_i}{\partial M_{nm}} = B_i \left( \frac{e^{-\frac{n^2}{\pi A_i}}}{\sqrt{\pi A_i}} \right) \begin{cases} 
0 & \text{for } F_i|\infty, \\
1 & \text{for } F_i|0, \\
1 - e^{-A_i + B_i} & \text{for } F_i|1, \\
e^{-A_i - B_i} - e^{-A_i + B_i} & \text{for } F_i|1. 
\end{cases}
\]

and the diagonal part gives:

\[
\frac{\partial F_i}{\partial M_{nn}} = \frac{A_i'}{2A_i} \left( \frac{e^{-\frac{n^2}{\pi A_i}}}{\sqrt{\pi A_i}} \right) \begin{cases} 
0 & \text{for } F_i|\infty, \\
B_i & \text{for } F_i|0, \\
B_i - e^{-A_i + B_i}(B_i + 2A_i) & \text{for } F_i|1, \\
e^{-A_i - B_i}(B_i - 2A_i) - e^{-A_i + B_i}(B_i + 2A_i) & \text{for } F_i|1. 
\end{cases}
\]
These derivatives are required to use quasi-Newtonian methods for finding the maximum of a function without explicitly calculating the Hessian. Second derivatives also can be explicitly calculated, even if the algorithm should pose dedicated care to avoid out-of-memory issues.

4.4. Measurement sampling

Having set a statistical framework, we have to consider the quantities introduced above as averaged over the repetition of $M$ measurements:

$$
\mathcal{P} \mathcal{L} = \frac{1}{M} \sum_{i=1}^{N} \sum_{\mu=1}^{M} \mathcal{L}_{i,\mu},
$$

that depends on the $N \times N\{M_{ij}\}$ coupling parameters. The $N(N+1)/2$ derivatives are, then,

$$
\frac{\partial \mathcal{P} \mathcal{L}}{\partial M_{nm}} = \frac{1}{M} \sum_{i=1}^{N} \sum_{\mu=1}^{M} \frac{\partial \mathcal{L}_{i,\mu}}{\partial M_{nm}},
$$

where $M$ is the number of coupled I/O observations of a system with a fixed $T$. The sampling rate of a given system will impact on the inference procedure: a larger statistical sampling will correspond to a more accurate learning process, i.e., to a better estimate of the model parameters. Thus, we define the sampling rate $\xi$ as:

$$
\xi(M, K_{\text{tot}}) = \frac{M}{K_{\text{tot}}},
$$

where $\xi$ depends on the number of independent parameters $K_{\text{tot}}$ to be inferred in the model. Three different situations can occur in this case, depending on the sampling rate achieved:

- $\xi < 1$, undersampling regime. The number of measurements is lower than the number of parameters, worst inference scenario.
- $\xi \gtrsim 1$, sufficient sampling. $M$ approximately matches $K$, thus, the sampling rate is appropriate for a stable learning procedure.
- $\xi \gg 1$, oversampling regime. Best-case scenario, in which the number of measurements is high enough to reasonably rule out pre-asymptotic finite measurement effects in the convergence of the pseudolikelihood to the actual likelihood function.

Ideally, we would like to be able to correctly infer the parameters in all the possible sampling rates, particularly in the undersampled regime. In the following, we will try to analyze the behavior of our learning framework under different measurement scenarios.

5. Numerical implementation

The algorithm is written in MATLAB, making use of the unconstrained low-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) minimization routines\(^5\) within the minFunc package [28]. The algorithm uses a low-rank Hessian approximation that reduces memory usage. Of course, we are not bound to use L-BFGS algorithms for the optimization, but we found it a good compromise between accuracy, speed, and memory requirement. We also have tested the

\(^5\) Normally, all the optimization routines available minimize cost functions. In our case, the maximization of $\mathcal{P} \mathcal{L}$ corresponds to the minimization of $-\mathcal{P} \mathcal{L}$.
use of constrained optimization since we have full control over the parameter range in the system. However, to be more general and for assessing the robustness of the method, we decided to report unconstrained results.

5.1. Dataset creation and its statistics

To test the inference model, we created a numerical set of intensity measurements that simulates a laboratory experiment. First of all, any real camera detection returns positive values, limited in magnitude by the pixel sensitivity: values above the maximal sensitivity $I_{\text{max}}$ will saturate, below the minimal sensitivity $I_{\text{min}}$ will not be recorded. For the sake of generality, we scale these values, and we assume that our simulated camera works in the range $I_i \in [0, 1]$ for both the input and the output. For the input, we choose to generate $M$ squared patterns of pixels having size $w$. Each of these contains $N_i = N_o = N/2 = w^2$ random values normally distributed, accordingly to a truncated Gaussian probability distribution with $\sigma = 0.1$, $\forall \gamma$ and $\mu = 0.5$. We point out that $\sigma$ is not related to the noise but defines the distribution of intensity values sampled in the input. These patterns are scrambled into the corresponding outputs via the $T$ matrix, which must preserve the positiveness and the intensity range.

In order to build our $T$, we first decide that a fraction $s$ of all possible entries will be active (non-zero), i.e., $s = S_{\text{active}}/w^4$. Given the sparsity, we randomly activate $S_{\text{active}}$ couplings in a temporary (binary) $T'$-matrix, setting their value to 1. We want also the output to be in the same $[0, 1]$ range, thus, we need to normalize the $T'$ row-wise, so that the matrix eventually be a stochastic matrix:

$$T_{\gamma \alpha} = \frac{T'_{\gamma \alpha}}{\sum_{\alpha} T'_{\gamma \alpha}}, \quad \sum_{\alpha = 1}^{N_o} T_{\gamma \alpha} = 1.$$  

With this choice, in every experiment performed, we are guaranteed to have also the output in the same intensity range of the input, like in-camera acquisitions. In general, we found that

Figure 3. On the left, an example of a transmission matrix for a $4 \times 4$ input–output system. On the right the probability distribution chosen for all the pixel in the input (blue histogram) and the corresponding distribution after passing through the transmission channel (orange histogram). Both have Gaussian statistics with mean value centered at 0.5.
the histogram of the intensities in the output followed a narrower Gaussian distribution, as it can be seen in figure 3. Depending on which integral (29) is chosen for the inference, when the region of integration is taken symmetric around zero (explicitly, when we use $F_i|_{-\infty}^{\infty}$ or $F_i|_{-1/2}^{1/2}$), it is convenient to translate the intensities in each channel by their mean $I_i - \mu_i$. The reason for this will be discussed in section results and section 5.3.1.

According to (4), we will study the inference method under noisy measurements. We perturb the output with a noise term, leaving the input unaltered. The noise follows Gaussian statistics with zero mean and variance $\sigma$. To probe different noise regimes, we repeated our analysis changing $\sigma \in [0, 0.5]$, in steps of 0.02, including possible saturation effects of the data. Per each $\sigma$, we will run an optimization and consequently a complete decimation of the parameters, studying the best model selection via the usage of several information criteria. The effect introduced by the noise perturbation is represented in figure 4; where it is clear that when the noise grows up to 20%, the channel is so perturbed that saturation effects become relevant.

5.2. \( P_L \)-maximization with all parameters

We decided to start the optimization process using weak prior information. Recalling the tensor form of $M$, we initialize each block with the following criteria:

(a) $I$ block. We start with a diagonal unity matrix;
(b) $T$ blocks. We set all the parameters to $1/w^2$;
(c) $V$ block. We calculate it as $T'\beta T$ using the previously initialized matrix, with $\beta_\gamma = \beta = 1/2/\sigma$, $\forall \gamma$.
We call the initial coupling matrix as $M^{(0)}$ that we use to start the minimization routine, obtaining the estimated couplings $\{M_{ij}\}$ at the end of the process. We point out that any initialization scheme should lead to similar convergence for the parameters. With the described choice, we avoid potentially wrong starting directions in the optimization process that can delay the convergence. The resulting optimized interaction matrix is referred to as $M_{\text{inf}}$.

5.3. Decimation strategy and model selection

After the initial $\mathcal{PL}$-maximization, we proceed by repeatedly decimating small couplings inferred within the interaction matrix $\hat{M}$. By doing so, we aim to recover a certain degree of sparsity that the transmission matrix may have (or may not) in the attempt of reducing overfitting issues. To do this, we proceed iteratively, sorting the values of the inferred $T$ and discarding a fraction (equal to $1/128$) of the smallest elements. Then, we repeat the optimization starting with the previously inferred matrix. The decimation strategy is applied directly on the $T$-region via the definition of a binary parameter mask $T^{\text{bin}}$, which is $T_{ij}^{\text{bin}} = 1$ when the coupling is active and 0 otherwise. The decimation is transported into the $V$-region by calculating another mask $V^{\text{bin}} = (T^{\text{bin}}/\beta T^{\text{bin}})$. To avoid divergence of the partition function, we restrict the decimation procedure only in the out-of-diagonal part of $\hat{M}$, leaving the diagonal parameters $\{M_{ii}\}$ acting as regularizers at advanced stages of decimation. Per each decimation step, we call $K$ the total number of parameters active in $\hat{M}$.

At first, the decimation strategy was introduced in [20] to recover couplings in an Ising model along with a stopping criterion, which allows estimating an appropriate number of parameters without proceeding further. Intuitively, the maximized value of $\mathcal{PL}$ should not change when decimating irrelevant couplings, whereas we expect a substantial decrease when discarding essential parameters. By doing so, we are implicitly testing a number of different linear transmissions in which weakly transmitting channels get progressively deactivated. Although we are not changing the transmission model itself (that we assume to be known and linear), we are effectively changing how we infer the interactions. All these linear models are not independent since the succession of decimated channels cannot be predicted $a$ $priori$; however, they form an ensemble of inferred transmission rules among which we have to select the most representative one [29]. Ideally, we would like to carry out model selection simultaneously with the decimation to be able to stop the procedure as soon as we find the optimal representation. The tools that we look for, then, are measures that are fast to compute and quantitative in assessing which model to prefer. Bayesian and cross-validation analysis can be orders of magnitude slower than approaches based on information theory [29] and, in this context, we prefer to investigate these latter. To this aim, we compare several criteria borrowed from information theory and statistical mechanics, which we use to stop the decimation when an acceptable $\hat{M}$ estimate is achieved.

- **Tilted-pseudolikelihood function (tPLF).** Numerically, the PLF does not remain flat when decimating; it can be hard to evaluate the precise number of relevant couplings by observing the point when it drops. To overcome this genuinely technical-issue, a tilted version of the PL was proposed to search for a maximum rather than for the end of a plateau. Following the procedure introduced by Decelle and Ricci-Tersenghi in [20], we begin by defining the tilted-$\mathcal{PL}$ function as:

$$t\mathcal{PLF} = \mathcal{PL} - k\mathcal{PL}_{\text{max}} - (1 - k)\mathcal{PL}_{\text{min}}.$$  \hspace{1cm} (39)

There, $k = K/K_{\text{tot}}$ weights the fraction of the active couplings among the total number of parameters $K_{\text{tot}}$, whereas $(1 - k)$ counts the fraction of those decimated. In this context, $\mathcal{PL}_{\text{max}}$ is the value achieved after fitting the fully connected model, while $\mathcal{PL}_{\text{min}}$
is the pseudolikelihood of the non-interacting one. This choice in equation (39) sets the boundaries of the tPLF at zero, both when the model has the maximum and the minimum number of parameters allowed. In a typical operational scenario, the tPLF increases when starting to decimate. Then, it reaches a maximum and, after that, it decreases again to zero [17, 20]. Ideally, we want to stop the decimation when \( k^* = s \), the true fraction of non-zero elements in \( T \), as defined in section 5.1. We locate this point at \( k_{tPLF} \) where the tPLF is maximum and, thus, we call \( M^{tPLF} \) the optimized decimated interaction matrix. Tilting the PL is effective when the true model is unknown and when it is possible to calculate the PL of the empty model a priori. It is only in this situation that we can assess the location of the maximum during the decimation procedure. Otherwise (and this will not be our case), the tPLF can be calculated only as a posteriori of the decimation procedure.

- **Akaike information criterion (AIC).** In an ideal case study, one of the most widely used tools to estimate the reconstruction quality is to compute the Kullback–Leibler divergence between the true and fitted models. The less this distance is, the better gets the estimated model. Unfortunately, in many cases, we do not have access to the true model for comparison, and we need to rely on information theory to make a qualitative assessment of which model performs better. One of the most important results in this field was provided by Akaike’s work [30]. Given a certain number of model parameters \( K \) and a set of \( M \) measurements, the quantity:

\[
AIC = 2K - 2M(\mathcal{P}\mathcal{L}) \tag{40}
\]

permits to compare between different models fitted on the same data. Of course, here we considered our log-pseudolikelihood quantity in place of the original \( \log(\mathcal{L}) \). The value \( K^{AIC} \) at which the AIC is minimized corresponds to the model matrix \( M^{AIC} \) that minimizes the information loss in terms of K–L distance and, thus, at the best parameter number representing the system. Although AIC permits comparing structurally different models [29], here, we make use of it to pick between models belonging to the same class (we assume knowing that the transmission is linear), being a decreasing number of interaction parameters the only difference between them.

- **Corrected AIC (AICc).** When the sampling \( M \) is small (though larger than \( K \)), there is a higher chance that AIC will overfit, selecting models with a higher number of parameters. To take into account this, a modified version of the AIC, namely AICc, was introduced considering a second order approximation of the bias term [31]:

\[
AICc = AIC + \frac{2K(K+1)}{M-K-1}. \tag{41}
\]

The above formulation takes into account the sampling number into a penalty factor, that loses importance when \( M \to \infty \), converging to AIC value. The value \( K^{AICc} \) at which the AICc is minimized corresponds to a second-order estimate of the optimal model. \( M^{AICc} \) is the corresponding decimated matrix.

- **Bayesian information criterion (BIC).** Firstly introduced by Schwarz in [32], it can be written as a function of the estimable parameters \( K \) and the number of sampling \( M \) as:

\[
BIC = K \log(M) - 2M(\mathcal{P}\mathcal{L}). \tag{42}
\]

The number of parameters \( K^{BIC} \) at which BIC is minimized is the best inference point according to the BIC. Here, the interaction matrix selected is called \( M^{BIC} \). This method
resulted to be efficient in the activation method [33] for the inference of Ising couplings starting from empty matrices.

It is interesting to recall that each ICs considered is related to the Kullback–Leibler divergence and in each formulation directly appears the log-likelihood term (the pseudolikelihood in our case), eventually multiplied by a factor proportional to the number of measures. The most relevant difference among these ICs is the remaining factor, called the ‘bias’ term, which gives a different penalty to the number of parameters during the model selection. It is the bias that, in the ultimate analysis, will determine which model selection performs better in a given scenario. These information criteria will help us estimate the best model fitting the data, providing an indication of the sparsity and the corresponding number of parameters needed to describe it correctly.

5.3.1. PL_{\text{min}} of the fully disconnected model. According to the strategy proposed in section 5.3, only the values on the diagonal of the interaction matrix \{M_{ij}\} are active when the model is completely decimated. This implies that, for each \(i\)th local likelihood \(L_i\), the parameter \(B_i = 0\) vanishes. In practice, we are left with:

\[
L_i \xrightarrow{\text{decimation}} \ln \left( \frac{e^{M_{ii}I_i^2}}{dI_i e^{M_{ii}I_i^2}} \right) = \ln \left( \frac{e^{-A_iI_i^2}}{dI_i e^{-A_iI_i^2}} \right). \tag{43}
\]

The argument of the logarithm is a Gaussian distribution of variance \(2/A_i\), with, cf. equations (16) and (20),

\[
-A_i = A_i = \begin{cases} \frac{N/2}{2\sigma_i^2}T_i^2 & \text{for the input, if } i = 1, \ldots, N/2, \\ \frac{1}{2\sigma_i^2} & \text{for the output, if } i = N/2 + 1, \ldots, N. \end{cases} \tag{44}
\]

When the model is completely decimated, the maximization of each \(L_i\) corresponds to fit a Gaussian to the distribution of the intensities per channel \(i\). We are not interested in the mean channel values, so we pre-compute the shift \(I - \mu\), where \(\mu = \{\mu_i\}\) is the vector mean value calculated over \(M\) samplings. However, for the case described in section 5.1, all the channels means are identical \(\mu_i = 0.5\forall i\), as can be seen in the output distribution plotted in figure 3. Thus, in our case study, we shift all the intensities by \(I - 0.5\).

Let us recall, that in this case we need to use either \(L_i\) or \(L_i^{1/2}\), to accept negative values of the integration variables. This shift will not change the definition of \(T\). Now, we can calculate the minimum value that \(\mathcal{P}L\) can reach as:

\[
\mathcal{P}L_{\text{min}}^{1/2} \bigg|_{-1/2}^{1/2} = -\sum_{i=1}^{N} \left[ \frac{I_i^2}{2\sigma_i^2} + \ln \left( \frac{\pi\sigma_i^2}{2} \right) + \ln \left| \frac{I_i}{\sigma_i} \right|^{1/2} \right] \\
= -\frac{1}{2} \sum_{i=1}^{N} \frac{I_i^2}{\sigma_i^2} - N \ln \frac{\pi}{2} - 2\sum_{i=1}^{N} \ln \sigma_i - \sum_{i=1}^{N} \left[ 2\text{erf} \left( \frac{1}{\sqrt{2}\sigma_i} \right) \right], \tag{45}\]

or

\[
\mathcal{P}L_{\text{min}}^{\infty} \bigg|_{-\infty}^{\infty} = -\frac{1}{2} \sum_{i=1}^{N} \frac{I_i^2}{\sigma_i^2} - N \ln 2\pi - \sum_{i=1}^{N} \ln \sigma_i. \tag{46}\]


Figure 5. Pseudolikelihood boundaries as a function of the noise plotted in a log-scale. The blue values are the $\mathcal{P}L$ minimized with all the parameters, that progressively decimate to the final orange values. The continuous yellow and dashed violet line represents the values expected from the theory, respectively from equations (46) and (47).

In the scenario in which the variance is equiparted between $N/2$ inputs $\sigma_I$ and $N/2$ outputs $\sigma_O$, the formula above reduces to:

$$\mathcal{P}L_{\text{min}} = -\frac{1}{2} \left( \frac{1}{\sigma_I^2} \sum_{i=1}^{N/2} I_i^2 + \frac{1}{\sigma_O^2} \sum_{i=N/2+1}^{N} I_i^2 \right) - \frac{N}{2} \ln(2\pi \sigma_I \sigma_O). \quad (47)$$

In the $4 \times 4$ case of figure 3, for example, we have $\sigma_I = \sigma^{\text{in}}$ and $\sigma_O = 0.05$. In figure 5 we plot the numerically optimized minimum and maximum value for the $\mathcal{P}L$, against the theoretical results. The plot confirms that equations (46) and (47) correctly predict the lower boundary for the pseudolikelihood.

6. Results

In the previous chapters, we have described extensively the learning framework proposed for an I/O spin-glass model. We present, in the following, the results obtained under different operative conditions to define the modus operandi. At the same time, we will always try to link our model description to the spin-glass theory to interpret the results from a different perspective.

6.1. Inferring $M$ with all the parameters

We start the analysis of the results by focusing on algorithm convergence. Initially, we keep all the parameters free to adjust to maximize the pseudolikelihood without decimation. Here, we explicitly consider the inference results of a $4 \times 4$ I/O system, which underlines an $M$ matrix having $32 \times 32$ parameters. We initially study the inference of a transmission matrix $T$
Figure 6. Schematics of the initial inference results: the matrix $M$ obtained after the PL maximization on independent active parameters. The black regions are the values close to zero, while with green and red we can distinguish between three (respectively negative and positive) different matrix sub-regions. The fact that the interaction matrix $M$ converged to a block matrix confirms that the link between the I/O pixels is well described by a transmission matrix.

with positive entries characterized by sparsity $s = 0.2$ in a weak noise $\sigma = 0.02$ regime. After maximizing the pseudolikelihood with respect to $M_{ij}$, we obtain the result shown in figure 6. We call it $M^\text{inf}$. It is evident that the interaction matrix converged into the expected tensorial form predicted by (25), detecting that we are in a purely transmission case. The color code chosen aids the visualization of negative values (in green) against positive ones (in red). We label three parts in the matrix, as shown in the figure:

- Part 1. The self-output coupling matrix converges into a diagonal form, $-\beta$, whose elements are $\{\beta_{i-N/2}\} = \{\beta_1, \beta_2, \ldots, \beta_{N/2}\}$, for $i = N/2 + 1, \ldots, N$.
- Part 2. The input–output coupling matrix closely follows the transmission matrix $2\beta T$ used for the dataset creation.
- Part 3. The self-input coupling matrix reported is $-V$, with, cf equation (20),

$$-V_{\eta\xi} = -\sum_{\alpha=1}^{N/2} T_{\eta\alpha}^\dagger \beta_\alpha T_{\alpha\xi} \quad \forall \, \eta, \xi = 1, \ldots, N/2, \quad i \leq j,$$

$$-V = -T^\dagger \beta T$$

To better understand how the noise propagates within this matrix, it is convenient to separate the formal analysis per each of the parts labeled in the interaction matrix $M^\text{inf}$, separating the noise mask $\beta$ from the matrix (25).
6.1.1. The noise channel as statistical temperature. First of all, it is worth noticing that the diagonal term (part 1 of $M_{\text{inf}}$, figure 6) gives us an estimation of the noise per each channel of the system. According to the tensor form (25), in the diagonal part, we have that $M_{ii} = \beta_i$ for $i > N/2$. Thus, the diagonal elements of $M_{\text{inf}}$ explicitly recovers the noise $\beta\gamma = \gamma - N/2$ per each I/O channel as in equation (24). It is interesting noticing that noise statistics can be recovered by looking directly at the diagonal part. We studied a range of noise $\sigma$, that we can relate to the factor $\beta\gamma$ for each output channel $\gamma$, via: $\beta\gamma = (2\sigma^2)^{-1}$, $\gamma = 1, \ldots, N/2$, i.e., in matrix notation:

$$\beta = (2\sigma^2)^{-1},$$

where the diagonal matrix $\sigma$ is introduced in equation (2). Giving a thermodynamical interpretation, the average temperature of the measurement is:

$$\theta = 2 \frac{N/2}{N} \sum_{\gamma=1}^{N/2} \frac{1}{\beta\gamma},$$

where, theoretically, the temperature can be expressed in terms of the variance of the noise introduced in equation (7) or (16), $\theta_{\text{th}} = 2\sigma^2$. In figure 7, we plot the average temperatures estimated from the inference of the full model for the systems studied as a function of the noise inserted in the dataset. We can appreciate a good agreement with the theoretical prediction, which tells that the system temperature scales quadratically with the inserted noise. The estimation departs from the parabolic trend after a certain point around $\sigma = 0.20$. This effect is due to the statistics of the output intensity after strong saturation effects start to take place, see figure 4.

We can conclude that the diagonal matrix of part (1) is effectively giving an estimation for the noise experienced within each transmission channel. This can be visualized in part (b) plot.
of figure 7, where per each channel we calculate the inserted noise standard deviation $\sigma^c_i$, and we divide it for the estimated one $\sigma_i$, as a function of the variance of the noise $\epsilon$ introduce in equation (2). It is possible to notice that this ratio stays around 1 (the gray plane), indicating that we are effectively estimating the channel noise up to a perturbation of 20%. Beyond this value, the noise estimation departs from the real one due to saturation effects.

6.1.2. The transmission matrix. The diagonal matrix in block (1) gives us the possibility to calculate the transmission matrix elements $T^i_{in}$ from the inferred $M^i_{in}$ block (2). We divide row wise $M^i_{in}$ for $i \in [1, N/2]$ and $j \in [N/2 + 1, N]$ by the elements $\beta_i$, so that the estimation of the system transmission matrix is given by:

$$T^i_{in} = \frac{M^i_{in}}{2\beta_i}.$$  

Part (a) of figure 8 shows the sorted values against the matrix sparsity $s$ reconstructed at $\sigma = 0.05$. We can see how the calculated $T^i_{in}$ values closely match the expected blue curve, abruptly falling at the exact matrix sparsity. However, at this stage, also the other parameters are active, and we are interested in cutting out all those which are irrelevant for the description of the $T$ matrix.

6.1.3. Self-input coupling. The block (3) of $M^i_{in}$ is closely connected with the noise estimation, cf equation (48). We formally expect $M_{ij} = M_{ji}$, and, as we already said, we implement the symmetry in the minimization procedure. For this part, we compare $V^i_{in}$ with $T^{i\dagger} \beta^i T^i$, where the $T$ is the true transmission matrix used to produce the dataset. Part (b) of figure 8 reveals the appropriateness of the values recovered for the $V^i_{in}$ against the theoretically expected trend, with the inferred set of output channel noises $\beta^i$. 

![Figure 8](image-url)
6.2. Decimation and information criterion comparison

The results discussed so far are referred to the optimal model, learned with all the parameters active in $M$. However, this is not always the best choice due to possible overfitting problems. Using all the parameters may fit well the training dataset but, it may poorly perform on the estimation of the transmission matrix representing the system, eventually failing the fit procedure on the validation dataset. To avoid this, we are interested in estimating the exact number of parameters active in the channel. We, therefore, proceed as proposed in the decimation procedure. That is, we recursively eliminate small couplings in successive $\mathcal{P}\mathcal{L}$-maximization processes. While doing this, we expect an abrupt change in the $\mathcal{P}\mathcal{L}$ when the model starts to under-fit the connectivity. This point indicates the best number of parameters that reproduce the dataset.

The quality of the inference procedure depends on the sampling rate, equation (38), that weight the number of measurements against the number of parameters to be estimated in the model. Thus, we study the behaviour of our inference method under a variety of numerical conditions. As a function of the linear input size $w$, the sampling rate is:

$$\xi(w) = \frac{M}{K_{\text{tot}}(w)}.$$  

By construction a symmetric matrix $M$ has $K_{\text{tot}}(w) = w^2(2w^2 + 1)$ independent entries. However, taking into account that we are considering a problem of pure noisy transmission all elements of the $M$ matrix are built from the $w^4$ elements of the transmission matrix plus the $w^2$ channel noise variances. In total, $K_{\text{tot}}(w) = w^4 + w^2$. In the following, considering $M = 10000$ of I/O couples and varying $w$, we find the sampling ratio values reported in table 6.2 (table 1).

At this stage, we run independent optimizations as a function of the noise $\sigma$ and study the behavior of the $\mathcal{P}\mathcal{L}$ progressively eliminating a fraction of $1/128$ couplings at each decimation step. To pick the optimal model among the decimation procedure, we evaluate the information criteria described in section 5.3, and we compare them on the $\mathcal{P}\mathcal{L}$-surface plots in figure 9. In each plot, the exact number of parameters to be inferred is represented with a blue line with white dots. Per each sampling rate, we encounter different behaviors, according to the regimes sketched in section 4.4:

- **I/O $4 \times 4$ ($\xi \gg 1$).** This is the best case for Bayesian inference; all the IC correctly guess the model number at zero noise $\sigma = 0$. The situation changes as soon as the noise is inserted. AIC and AICc follow the same trend (this is expected since AICc converges to AIC in case of a high sampling rate). Both are prudent, always preferring models with a higher number of parameters. On the other hand, tPLF excessively decimates in every noise regime. BIC outperforms all the information criteria in this scenario, correctly guessing the parameter number up to $\sigma = 0.36$.

- **I/O $8 \times 8$ ($\xi > 1$).** The situation changes with respect the previous one. AIC is always very prudent and stable against noise, while AICc pushes more toward the correct result. However, at high noise, AICc excessively decimates. tPLF, instead, oscillates around the
Figure 9. $\mathcal{P}\mathcal{L}$-surface plots of four I/O systems tested in this study, differing in the input size $w$. The surface represent the pseudolikelihood optimized value as a function of the noise and the parameter number (thus, the decimation).

correct value, over-decimating at low noise and under-decimating at higher $\sigma$ values. BIC correctly guesses the parameter number up to moderate noise but, after that, BIC transits towards excessive decimation. BIC is really not effective for $\sigma \geq 0.16$. At low noise, noticeably, tPLF and BIC closely guess the number of parameters.

- **I/O $12 \times 12$ ($\xi \lesssim 1$).** The previous behaviour changes in this sampling scenario. AIC is always the most prudent while AICc quickly over-decimates. tPLF is very robust at this sampling rate, being very close to the correct solution and never over-decimate. BIC is good up to 10% noise but, after that, it fails, picking a practically empty model as its best guess.

- **I/O $16 \times 16$ ($\xi \ll 1$).** In this regime, BIC fails at lower noises, recovering exact solutions up to a perturbation of 4%. tPLF performs very well, staying prudent at higher noise and converging to the same results as AIC. Remarkably, AICc tends to overfit, favoring results with a higher number of parameters active.
Figure 10. Reconstruction error $Q$ of the direct transmission matrix $T_{\text{inf}}$. The plots report the quantity $Q$ under different inference scenarios. We can appreciate that all the curves start from $Q \approx 0$, growing as the noise and the size increase (with fixed number of measurements).

We have analyzed how ICs help find the most accurate number of parameters among the decimated models quantitatively by defining the inference reconstruction error of the $T_{\text{inf}}$ against the original one (used to produce the datasets):

$$Q(T, T_{\text{inf}}) = \sqrt{\frac{\|T - T_{\text{inf}}\|}{\|T\|}}. \quad (53)$$

Here, $Q = 0$ corresponds to the exact recovery of $T_{\text{inf}}$ and in general low $Q$ represents good inference results. Results are presented in figure 10 for all the ICs and model sizes. In general, at low noise, we are able to get an excellent estimation of $T$ and at all the sampling rates investigated. Although some IC perform better than others in picking the closest number of parameters, all the models are similar in terms of operative results. There are exceptions: tPLF poorly performs on oversampled systems (i.e., $4 \times 4$) at any noise and BIC progressively fails at lower noises as $\xi$ decreases. These facts are consistent with the plots of figure 9, giving a further indication in favor of an operative equivalence among the different selection criteria tested. In particular, no decimation always resulted worse than the others (thus appropriate decimation worth better results), and knowing the exact number of effective couplings decreases the $Q$-reference curve.
6.3. The inverse transmission matrix

The same protocol can be used to infer the inverse transmission matrix $T^{-1}$, rather than inferring $T$ and then inverting it. We will use this procedure for the reconstruction of an image pattern that travels from the I-edge to the O-edge. Let us analyze the reliability on the inference of the inverse, recalling that each dataset was generated with the direct $T$ matrix, and the noise is added to the output. In general, the $T^{-1}$ may not have the same sparsity as the corresponding direct matrix, yet it is a stochastic matrix, as well. As for the previous case, we proceed with the inference of inverse matrices at increasing noise levels and using the decimation. To be consistent, we present the results in analogy with section 6.2, using the same datasets and color-code.

- **I/O 4 × 4 ($\xi \gg 1$).** This is the only size at which the inverse matrix has some null entries. Thus, there is a small degree of diluteness to be recovered. In the green $\mathcal{PL}$-surface AIC and AICc are prudent with the decimation, BIC gets closer to the actual value of parameters to be estimated slightly, preferring setting to zero smaller couplings. Interestingly, tPLF strongly decimates and has a less predictable behavior, with large fluctuations at high noise.

- **I/O 8 × 8 ($\xi \gtrsim 1$).** At this size, the $T^{-1}$ is not sparse and, accordingly, AIC, AICc, and BIC always determine a weak or no decimation. Like in the previous case, tPLF tends to over-decimate the coupling matrix.

- **I/O 12 × 12 ($\xi < 1$).** In this case, the situation is pretty similar to the 8 × 8 case, tPLF being the worse information criterion among the others.

- **I/O 16 × 16 ($\xi \ll 1$).** With the largest system considered BIC tends to excessively decimate compared with AIC and AICc, which are the most accurate selection criteria. tPLF fails like the previous cases, always returning a quite sparse matrix.

Interestingly, for the $T^{-1}$ inference, tPLF always produced wrong estimations, strongly decimating coupling while the matrix was not sparse. This is more a test on tPLF (and other IC, in general) than on the inference quality itself. Being $T^{-1}$ a dense matrix, we know in advance that any decimation trial should be irrelevant. In principle, there are only non zero elements, though many of them can be small, because of the stochasticity property

$$\sum_j T_{ij} = 1.$$  

So, decimating setting the smallest ones to zero does not relevantly change the reconstruction outcome. Yet, in our case, the inverse matrix is substantially different from the direct one. First, it is not possible to control directly the ground truth $T^{-1}$, and to estimate it we have to rely on numerical inversion techniques. Moreover, $T^{-1}$ has also access to negative values that vary along each line, and the condition $\sum_j T_{ij} = 1$ does not constrain the magnitude of the entries that can become large if proper cancellations occur. In these terms, the inference of direct and inverse $T$ represent two opposite scenarios.

To control the inference quality in this regime, we compute the reconstruction error $Q(T^{-1}, (T^{-1})_{\text{ml}})$ like in section 6.2. Contrarily to the previous situation, no reconstruction approaches $Q \approx 0$ as can be seen in figure 12. Moreover, it seems that there is an optimal noise around $\sigma \approx 0.1$ at which the inference is more accurate (minimum in the $Q$ curve). All the IC selections follow the same quality trend being almost indistinguishable at the exception of tPLF, which departs from reasonable estimations at $\sigma = 0.04$. From figures 11 and 12, tPLF seems not reliable while inferring $T^{-1}$. In fact, it grounds its estimation based on sensitive features of the pseudolikelihood curve: a net drop of the $\mathcal{PL}$ is easily detected by tPLF.

\[6\text{We recall that } T \text{ is a stochastic matrix, with random entries } T_{ij} \in [0, 1] \text{ and } \sum_i T_{ij} = 1.\]
whereas noisy curves could misplace this maximum. Lastly, if the matrix is densely populated, we point out that tPLF cannot be used by definition. In fact, the definition of tPLF sets its values at full or empty matrix connectivity to be zero, and the maximum cannot be located at the edges.

6.4. Consistency of the inferred matrices

Consistency is important, in particular, when evaluating $\mathbf{T}$ and its inverse. For each of the two matrices, the inference procedures are carried out separately, and, as a consequence, there is no guarantee that consistency is preserved after the matrix recovery. In this section, we try to study if reciprocal properties between direct and inverse matrices are preserved after the inference procedure.

6.4.1. Pseudo-unitary product. To test the consistency of the direct and inverse matrices, it is worth studying the product $\mathbf{T}^{-1}\mathbf{T} = \mathbf{T}\mathbf{T}^{-1} = \mathbf{I}$. Moreover, we used different IC selections that might perform better on direct or inverse inference but not on both. We tested the consistency

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure11.png}
\caption{$\mathcal{P}L$-surface plots of several I/O systems tested in this study differing in the input size $w$. The surface represent the pseudolikelihood optimized value as a function of the noise and the parameter number (thus, the decimation).}
\end{figure}
of the method by picking the best-inferred matrices (direct and inverse) multiplying them to estimate the identity matrix. Per each $w$, we use the information criterion that best matched the correct number of couplings of the direct matrix without any consideration for the inverse. In the plot of figure 13, we plot the sorted values of the pseudo-unity matrix $(T^{-1})_{\inf}$ calculated in the $w = 4$ case, choosing AIC as the best selection criterion for both direct and inverse.

We use a progressive color map to differentiate between reconstructions at different noises. In low noise transmission (bluish curves), the diagonal strongly dominates over the out-of-diagonal region, exhibiting a visible step that tends to vanish as the rumor increases. However, we found that a dominant diagonal feature emerged at all the noises considered. For all the other sizes explored, we found similar behavior, with the general trend that the higher the number of measurements $M$, the better is the reconstruction of the diagonal part of the pseudo-unity matrix. As already discussed in section 6.3, tPLF cannot be used to pick inverse matrices because it is not devised for dense matrices.

6.4.2. Stochasticity of inferred direct and inverse matrices. Since both $T$ and $T^{-1}$ are stochastic, it is interesting to observe if this property is recovered by the inference process, in which stochasticity is not imposed. Per each AIC reconstruction, we calculate the average row sum of $T_{\inf}$ and $(T^{-1})_{\inf}$ that we plot against the noise level in figure 14. For low noise conditions, the matrices recovered at any size appear stochastic. However, above the saturation threshold,
Figure 13. Sorted values of the matrix $\left( T^{-1}\right)_{\text{inf}} T_{\text{inf}}$ (top) and $T_{\text{inf}} \left( T^{-1}\right)_{\text{inf}}$ (bottom) obtained with AIC selection with a $4 \times 4$ system. The baseline of the identity matrix is drawn with a black solid line, while the sorted values at increasing noise are plotted following a progressive colorcode. The bluish lines (low $\sigma$) are closely following the expected behaviour, having a strong diagonal component dominating an almost null out of diagonal region. The situation progressively worsens at increased noise, though diagonal elements remain larger than the off-diagonal ones.

their stochasticity starts to fluctuate while their mean value decreases considerably. Remarkably, the trend followed at every system size and noise seems to coincide, regardless of their relative sampling ratio.

6.4.3. Correlation analysis

For an additional analysis, let us examine the correlation coefficient of the $C(T, T_{\text{inf}})$ and $C(T^{-1}, T_{\text{inf}}^{-1})$. The results are presented in figure 15 as a function of the noise and the IC used for selecting among decimated models. In panel (a) the results of the correlation of the inferred transmission matrix against the true-matrix are presented. The green curves denote better results when sampling $\xi \gg 1$, returning high correlation along the noise range explored. The situation progressively worsens as soon as $\xi$ decreases below 1, i.e., the number of parameters becomes larger than the measurements used for the learning. On the other hand, we could look at the correlation trend of the $T_{\text{inf}}^{-1}$, reported in part (b). It is standing out a different situation. Only the $\xi \gg 1$ case behaved like the previous inference, where the others start from a lower correlation at low noise. Afterward, the correlation reaches a maximum then decreases again. The same behavior can be observed for any size, though for the two larger sizes of
Figure 14. Measure of the stochasticity of the inferred matrix. We can notice that, at zero noise, it is fully recovered in both cases while, by increasing the noise, the stochasticity decreases with huge fluctuations along lines.

$12 \times 12$ and $16 \times 16$, the correlation with the original inverse matrix is so low that we may suspect poor reconstruction performances.

6.5. Testing image transmission and focusing

The motivation that driven this study is to create a statistical framework able to recover the transmission matrix of an optical disordered system via a random sampling approach, opening the possibility to use turbid media as optically opaque lenses. Ideally, one wants a reliable system, able to correctly infer the properties of the channel on both directions even in the presence of strong noise fluctuation and saturation effects. To test the robustness of the inference procedure, we decided to accomplish two tasks defined in this way:

- **Focusing**. Using the inferred $T_{\text{inf}}$ matrix to deliver energy to a given spatial pattern at the output. From the knowledge of $T_{\text{inf}}$, we are able to construct a given output pattern (like focusing on a given spot) changing the input.

- **Imaging**. Using the inferred $(T^{-1})_{\text{inf}}$ matrix to reconstruct the image of an object at the input starting from the noisy and random pattern in output.

In a perfect scenario, it may be reasonable that both situations are equivalent, reversing the input–output order in the vector containing the pixel intensities $I$, due to the exchange of role between $T$ and $T^{-1}$. To test the inferred matrices, we take another validation dataset $I_{\text{val}}$ composed of $M = 1000$ new patterns that we transmit through the disordered channel. This test acts as a validation for the training procedure, given that none of $I_{\text{val}}$ vectors were used in the learning dataset. By using $T_{\text{inf}}$ and $[(T^{-1})_{\text{inf}}]^{-1}$, we want to reconstruct the intensity pattern focused at the output, $I_{\text{rec}}$, given an input. On the other hand, by using $(T^{-1})_{\text{inf}}$ and $(T_{\text{inf}})^{-1}$, we want to reconstruct the image at the input, $I_{\text{rec}}$, given a pattern at the output. To
Figure 15. Part (a), correlation of the $T_{\text{inf}}$ with the original one used to create both training and validation dataset as a function of the noise and for the various IC used. We can notice how every curve starts with $C \approx 1$ decaying faster as the system size increases. Part (b), $T_{\text{inf}}^{-1}$ correlation against the expected one. Here we notice a particularly low correlation, especially in the bigger (and undersampled) systems.

To test the goodness of the reconstructions, we calculate the connected correlation $C$ between the true $I^x$ and the reconstructed object $I_{\text{rec}}^x$:

$$C(I^x, I_{\text{rec}}^x) = \frac{\sum_l (I^x - \bar{I}^x)(I_{\text{rec}}^x - \bar{I}_{\text{rec}}^x)}{\sqrt{\sum_l (I^x - \bar{I}^x)^2 \sum_l (I_{\text{rec}}^x - \bar{I}_{\text{rec}}^x)^2}},$$

(54)

where $x = \{\text{in}, \text{out}\}$, $l = \{\alpha, \gamma\}$ and $\bar{I}^x$ is the mean value of $I^x$. $C = 1$ means exact reconstruction whereas a small $C$ means failed reconstruction.$^7$

6.5.1. Focusing procedure. In the top row of figure 16, we report the analysis of the correlation behavior for focusing versus noise at various sampling rates. Here, the transmission matrices used were inferred by pseudolikelihood-maximization and AIC decimation.

(a) Focusing with $T_{\text{inf}}$. It works extremely well at zero noise for any system linear size $w$. The correlation drops as soon as we increase the noise, decreasing the faster, the smaller is the sampling rate. For $\xi < 1$, the focusing is substantially lost as soon as $\sigma > 0$, whereas for $\xi \gg 1$, the focusing with the real and the inferred matrices yield very correlated results also with a large noise $\sigma = 0.5$.

$^7$In principle $C$ could also access negative values, being $C = -1$ a perfectly reconstructed object with opposite sign.
Figure 16. Correlation of the reconstructed patterns using the inferred $T$ and $T^{-1}$ and their corresponding inverted quantities.

(b) Focusing with $\left[\left(T^{-1}\right)_{\text{inf}}\right]^{-1}$. The inversion of the $T_{\text{inf}}^{-1}$ appears to be robust against inversion at high sampling rate for any noise, quickly dropping as $\xi \geq 1$. In these cases, the correlation behavior is not monotonous with the noise but there seems to be an optimal noise value ($\sigma \approx 0.1$) at which the inversion returns the best result.

6.5.2. Imaging procedure. In the bottom row of figure 16, we report the correlation analysis for the imaging procedure. As for the focusing, the inverse matrices were inferred by pseudolikelihood maximization coupled with AIC.

(c) Imaging with $(T_{\text{inf}})^{-1}$. The numerical inversion of $T_{\text{inf}}$ is highly unstable already at moderate noise for practically any sampling rate $\xi$. At $\sigma = 0$, however, numerical inversion was correctly outperformed at all investigated sizes, even for $\xi \ll 1$.

(d) Imaging with $\left(T^{-1}\right)_{\text{inf}}$. Good results are found for the inference of the inverse. The correlation with the true pattern remains quite stable in the whole noise range explored, moreover being robust at any sampling scenario.

Although the correlation plot of figure 15 seems to condemn the inference of $T^{-1}$ to low quality reconstructions (in particular in low sampling regimes), the correlation analysis in the validation set in figure 16 indicates that imaging seems feasible. Let us, then, have a look at images reconstructed using $(T_{\text{inf}})^{-1}$ and using $(T_{\text{inf}})^{-1}$ (also termed $\text{inv}(T_{\text{inf}})$ in the figures). As a test image, we used a $16 \times 16$ draw of a bomb that we reconstruct after the propagation from the output back to the input edge. The reconstructions are presented in figure 17 using $(T_{\text{inf}})^{-1}$ (set on top) and $(T^{-1})_{\text{inf}}$ (bottom). In the first case, the reconstruction is perfect at zero noise,
Figure 17. Upper part (a), image reconstructions using \((T_{\text{inf}})^{-1}\) and bottom part (b), using \((T^{-1})_{\text{inf}}\). The insets reproduce the correlation between the image reconstructed with the original and with the inferred transmission matrices, cf the 16 × 16 curves of panel (c) (top) and (d) (bottom) of figure 16. It is possible to notice how the performances are better in the inverse inference at any noise, where the direct gives excellent reconstruction only at zero noise.

but the quality immediately drops, and the image is not recognizable at any noise other than \(\sigma = 0\). The latter, instead, returns good reconstructions at finite noise, even if the correlation of \(T^{-1}\) was poor and the reconstruction error \(Q\) large. At noises around the saturation threshold \(\sigma \neq 0\), the reconstructions start to fade into a noisy background.

As expected, matrix inversion is something that we need to avoid, especially in the presence of noise. Both the matrix inversions resulted in less (or not-at-all) robust reconstruction; thus, it is always better to directly infer the matrix needed for the given application: focusing or imaging.

7. Conclusions and perspectives

In this last section, we want to examine briefly some aspects of our study. First, this work was inspired by recent development in statistical mechanics of random optical systems [34–36], from either thermodynamics and inference perspectives [17, 27]. As extensively discussed, the main idea behind our manuscript was to explore the analogy between the transmission matrix problem and the statistical mechanics description of a system of interacting variables. From this field, we borrowed several tools and ideas [17, 20] but more can be done to expand the results to other cases other than noisy transmission. Considering a random input and measuring a randomized output, in fact, reminds the situation in which unknown couplings describe
the interactions taking place in continuous random spins. To do so, we have rearranged the classical multivariate least-square approach into a representation of an interaction matrix $\mathbf{M}$ that acts on an I/O representation. We have set a statistical mechanics framework, where the noise variance is proportional to the heat-bath temperature of a canonical ensemble. By maximizing the pseudolikelihood of such a system, thus, we are looking at the thermodynamics of the I/O pixels variables in which their mutual interaction is expressed in terms of the elements of the $\mathbf{T}$. Our model is designed to be reversible and accepts the swap of I/O, giving access also to a straightforward inference of the $\mathbf{T}^{-1}$. We found that, when the input/output are linked via a transmission matrix, the generalized interaction matrix converges into a block representation containing the transmission rule, the noise estimation and the input self-couplings. All these results are consistent with the literature in multivariate least-square problems but our model allows for a generalized description. This validation permits us to step further with the following studies towards the definition of a phase-problem for the inference of a complex-valued $\mathbf{T}$, or a problem with local feedback or local filters and conditional constraints. Our work can be generalized in the presence of coherent sources and data available for input and output complex amplitudes (i.e., measuring intensities and phases). Instead, reconstructing a complex transmission based on intensity-only measurement needs dedicated care.

Our model does not recover only the sole transmission matrix: more information is encoded in the interaction matrix $\mathbf{M}$. On one hand, the inference provides us with an estimation of the noise variance under Gaussian statistics that assess the presence of noisy channels. Second, the matrix $\mathbf{M}$ contains the self-input coupling term. In the case of pure transmission, this is proportional to the a Gramian matrix [37], but the approach can be generalized to more involved optical systems. On top of yielding a tool for focusing and image reconstruction, the whole interaction matrix, then, may permit the study of properties of the materials composing the medium. At the moment, our method could find direct application in the case of linear incoherent imaging [24]. However, the aim is to achieve a statistical mechanics description of the complete processes of focusing and imaging through disorder [7, 38]. Possibly, this will enable the comprehension of structured focusing [13, 39], to study Anderson localization in disordered fibers [40] or signal transmission through hyperuniform channels [41]. The bidirectional approach in calculating direct and inverse transmission could be suitable, additionally, for the study of truncated channels. When characterizing a fiber, we can look at the whole input/output facets; but this is not always possible in generic disordered media, in which only a portion of the output could be collected by the camera [38]. Open transmission channels potentially reduce the chance of direct $\mathbf{T}$-matrix inversion, thus the interest in inferring its inverse directly. Finally, our model is linearly scalable in complexity. Each local likelihood term $\mathcal{L}_i$ is an independent entity and can be optimized in parallel. This independence permitted us to write our model exploiting GPU hardware resources to speed up the optimization process and leave room for further algorithm optimization.

Acknowledgments

The authors thank A Marruzzo, G Gradenigo, F Antenucci, J Rocchi, F Ricci-Tersenghi, and L Biferale for useful discussions.
Funding information

We acknowledge the support from the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Program, Project LoTGlasSy (Grant Agreement No. 694925) and Prof. Giorgio Parisi. We also acknowledge the support of LazioInnova-Regione Lazio under the program Gruppi di ricerca 2020-POR FESR Lazio 2014–2020, Project NanoProbe (Application code A0375-2020-36761), and the Italian Ministry of Education, University and Research under the PRIN2015 program, Grant code 2015K7KK8L-005.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Daniele Ancora  https://orcid.org/0000-0002-1232-4112
Luca Leuzzi  https://orcid.org/0000-0002-5971-7191

References

[1] Robert C 2014 Machine Learning: A Probabilistic Perspective (London: Taylor and Francis)
[2] Friedman J, Hastie T and Tibshirani R 2001 The Elements of Statistical Learning (Springer Series in Statistics vol 1) (New York: Springer)
[3] Peckner R, Myers S A, Jacome A S V, Egertson J D, Abelin J G, MacCoss M J, Carr S A and Jaffe J D 2018 Nat. Methods 15 371
[4] Altay E and Satman M H 2005 J. Financ. Manag. Anal. 18 18
[5] Popoff S M, Lerosey G, Fink M, Boccara A C and Gigan S 2011 New J. Phys. 13 123021
[6] Bertolotti J, van Putten E G, Blum C, Lagendijk A, Vos W L and Mosk A P 2012 Nature 491 232
[7] Vellekoop I M, Lagendijk A and Mosk A P 2010 Nat. Photon. 4 320
[8] Sebbah P 2001 Waves and Imaging Through Complex Media (Berlin: Springer)
[9] Yoon J, Lee K, Park J and Park Y 2015 Opt. Express 23 10158–67
[10] Carpenter J, Eggleton B J and Schröder J 2014 Opt. Express 22 96
[11] Wiersma D S 2013 Nat. Photon. 7 188
[12] Boonzajer Flaes D E, Stopka J, Turtava S, De Boer J F, Tyc T and Čižmár T 2018 Phys. Rev. Lett. 120 233901
[13] Di Battista D, Ancora D, Zhang H, Lemonaki K, Marakis E, Liapis E, Tzortzakis S and Zacharakis G 2016 Optica 3 1237–40
[14] Chaigne T, Katz O, Boccara A C, Fink M, Bossy E and Gigan S 2014 Nat. Photon. 8 58
[15] Frostig H, Small E, Daniel A, Oulevey P, Derevyanko S and Silberberg Y 2017 Optica 4 1073–9
[16] Penrose R 1956 On best approximate solutions of linear matrix equations Math. Proc. Camb. Phil. Soc. 52 17–9
[17] Tyagi P, Maruzzo A, Pagnani A, Antenucci F and Leuzzi L 2016 Phys. Rev. B 94 024203
[18] Ravikumar P, Wainwright M J and Lafferty J D 2010 Ann. Stat. 38 1287–319
[19] Ekeberg M, Lövkvist C, Lan Y, Weigt M and Aurell E 2013 Phys. Rev. E 87 012707
[20] Decelle A and Ricci-Tersenghi F 2014 Phys. Rev. Lett. 112 070603
[21] Bianchi S and Di Leonardo R 2012 Lab Chip 12 635
[22] Kim D, Moon J, Kim M, Yang T D, Kim J, Chung E and Choi W 2014 Opt. Lett. 39 1921–4
[23] Ancora D, Dominici L, Gianfrate A, Cazzato P, De Giorgi M, Ballarini D, Sanvitto D and Leuzzi L 2021 Photonics Res. 10
[24] Sharma M K, Metzler C A, Nagesh S, Baraniuk R G, Cossairt O and Veeraraghavan A 2019 IEEE Trans. Comput. Imaging 6 95–108
[25] Calisesi G, Ghezzi A, Ancora D, D’Andrea C, Valentini G, Farina A and Bassi A 2021 Prog. Biophys. Mol. Biol. 168 66–80
[26] Barber D 2012 Bayesian Reasoning and Machine Learning (Cambridge: Cambridge University Press)
[27] Marruzzo A, Tyagi P, Antenucci F, Pagnani A and Leuzzi L 2018 SciPost Phys. 5 002
[28] Schmidt M 2005 minFunc: unconstrained differentiable multivariate optimization in Matlab http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html
[29] Anderson D and Burnham K 2004 Model Selection and Multi-Model Inference 2nd edn vol 63 (New York: Springer)
[30] Akaike H 1974 IEEE Trans. Autom. Control 19 716–23
[31] Hurvich C M and Tsai C-L 1989 Biometrika 76 297–307
[32] Schwarz G 1978 Ann. Stat. 6 461–4
[33] Franz S, Ricci-Tersenghi F and Rocchi J 2019 arXiv:1901.11325
[34] Antenucci F, Crisanti A and Leuzzi L 2015 Phys. Rev. A 91 053816
[35] Antenucci F, Crisanti A, Ibáñez-Berganza M, Marruzzo A and Leuzzi L 2016 Phil. Mag. 96 704
[36] Antenucci F 2016 Statistical Physics of Wave Interactions (Berlin: Springer)
[37] Sreeram V and Agathoklis P 1994 IEEE Trans. Circuits Syst. 1 41 234–7
[38] Popoff S, Lerosey G, Fink M, Boccara A C and Gigan S 2010 Nat. Commun. 1 81
[39] Di Battista D, Ancora D, Leonetti M and Zacharakis G 2016 Appl. Phys. Lett. 109 121110
[40] Leonetti M, Karbasi S, Mafi A and Conti C 2014 Phys. Rev. Lett. 112 193902
[41] Di Battista D, Ancora D, Zacharakis G, Ruocco G and Leonetti M 2018 Opt. Express 26 15594–608