# Physically optimizing inference

Audrey Huang,[1] Benjamin Sheldan,[2] David A. Sivak,[2] and Matt Thomson[3]

[1] *Department of Computer Science, California Institute of Technology, Pasadena, California, 91125, USA*
[2] *Department of Physics, Simon Fraser University, Burnaby, BC, V5A1S6, Canada*
[3] *Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125, USA*

Data is scaling exponentially in fields ranging from genomics to neuroscience to economics. A central question is whether modern machine learning methods can be applied to construct predictive models based on large data sets drawn from complex, natural systems like cells and brains. In machine learning, the predictive power or generalizability of a model is determined by the statistics of training data. In this paper, we ask how predictive inference is impacted when training data is generated by the statistical behavior of a physical system. We develop an information-theoretic analysis of a canonical problem, spin network inference. Our analysis reveals the essential role that thermal fluctuations play in determining the efficiency of predictive inference. Thermal noise drives a system to explore a range of configurations providing 'raw' information for a learning algorithm to construct a predictive model. Conversely, thermal energy degrades information by blurring energetic differences between network states. In general, spin networks have an intrinsic optimal temperature at which inference becomes maximally efficient. Simple active learning protocols allow optimization of network temperature, without prior knowledge, to dramatically increase the efficiency of inference. Our results reveal a fundamental link between physics and information and show how the physical environment can be tuned to optimize the efficiency of machine learning.

The emergence of 'Big Data' is a central theme in contemporary science [1–3]. An important challenge is utilizing large data sets to model and understand complex interacting systems found in fields such as biology and economics [4, 5]. Recently, machine learning has demonstrated the ability to extract and model patterns from large data sets across domains as disparate as object recognition, speech interpretation, and game playing [6–10]. Thus, modern machine learning methods provide an attractive route for automating the construction of scientific models from data.

In science and engineering, the key objective of modeling is prediction. For example, an important goal in modern biology is to engineer biological networks to achieve practical goals like chemical synthesis and disease treatment [11]. This engenders a drive to learn predictive models of biological networks that generalize to conditions outside the natural range of operation.

In machine learning, the statistics of training data control the accuracy, generalizability, and efficiency of learning [12–14]. A challenge in scientific model construction is that the statistical properties of the training data are determined by the physical behavior of the system under investigation: training data can only be indirectly controlled by human intervention or experimentation. A fundamental research question, then, is to understand how a system's natural statistics affect the learning of predictive models. In this paper, we ask how the physical behavior of a network determines the feasibility of learning predictive models from observations: when is the natural behavior of a system sufficient for predictive inference?

We develop an information-theoretic analysis of a classic physical learning problem, the inference of a spin network from observations. Spin networks have been applied to model a range of disordered systems with non-uniform interactions between elements, including neural networks, bird flocks, and economic systems [7, 15–18]. Our analysis shows that the physical environment of a spin network, specifically its temperature, determines when a unique model can be constructed from a finite set of observations. Further, we link the efficiency of inference—the number of observations required to achieve a specific error tolerance—to the scale of thermal spin fluctuations in the system.

Learning is optimally efficient at a specific temperature that maximizes the information content of each observation. Through active learning, an observer can optimize the scale of a system's fluctuations, thereby bringing the system to its optimal inference temperature and maximizing the efficiency of inference. Thus, our results elucidate a fundamental connection between physics and machine learning by revealing that the physical environment (specifically temperature) of a network constrains the amount of information that is available for inference.

## I. SPIN-NETWORK INFERENCE

We consider the problem of inferring a predictive model of a spin network from network state observations. We specifically ask how the physical properties of the network (its physical environment) impact the efficiency of inference (Fig. 1).

A spin network consists of $m$ binary elements or 'spins.' A network microstate is specified by each spin $i$ being either up ($\sigma_i = 1$) or down ($\sigma_i = -1$). (As an alternative to their original physical interpretation, spins can represent Boolean variables that are 'true' or 'false,' neurons that are 'firing' or 'silent,' genes that are 'on' or 'off,' and so on.) Spins $i$ and $j$ interact in a pair-wise fashion parameterized by a coupling constant $J_{i,j} \in \mathbb{R}$. Interac-

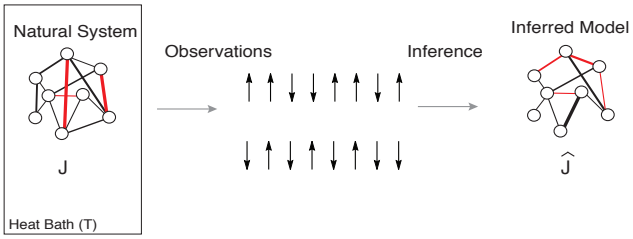


FIG. 1. The spin-network inference problem. A spin network is coupled to a heat bath at temperature $T$. Negative couplings are shown as red links and positive couplings as black links. An observer infers a predictive model of the network based on observations of network configurations drawn from the equilibrium distribution.

tions are symmetric, $J_{i,j} = J_{j,i}$, with no self-interactions, $J_{i,i} = 0$. These symmetry and diagonal constraints mean that the matrix $\mathbf{J}$ has $(m^2 - m)/2$ unknowns.

A given spin network can exist in one of $2^m$ different microstates or configurations, $s = (\sigma_1, \sigma_2, \ldots, \sigma_m)$. Each network configuration has an associated energy due to spin-spin interactions:

$$E(s|\mathbf{J}) = -\sum_{ij} J_{i,j}\sigma_i\sigma_j - \sum_i h_i\sigma_i \ . \tag{1}$$

A magnetic field $h_i$ biases spin $i$ to either point up or down. For simplicity, in our analysis we set $\mathbf{h} = 0$, but the extension to non-zero fields is conceptually straightforward.

The network exchanges energy with a heat bath at temperature $T$. At equilibrium, the network fluctuates between configurations, and the probability of observing a specific spin configuration $s_k$ is determined by the Boltzmann distribution,

$$P(s_k|\mathbf{J}) = \frac{\exp[-E(s_k|\mathbf{J})]/T}{Z} \ , \tag{2}$$

for network partition function $Z \equiv \sum_k \exp[-E(s_k|\mathbf{J})/T]$ (throughout we set $k_\mathrm{B} = 1$). In more general situations (*e.g.*, a neural network), $T$ would quantify the effective temperature, corresponding to the magnitude of stochastic fluctuations in network state.

We observe a given network at equilibrium, sampling $N$ states from the equilibrium distribution. We focus our analysis on estimation of the coupling matrix $\mathbf{J}$, because this allows estimation of $E(s|\mathbf{J})$ at any temperature.

To infer $\mathbf{J}$ from observations, we construct a likelihood function for the coupling matrix $\mathbf{J}'$ given a set of $N$ independent identically distributed observations $\{s_k\}$ of network state:

$$\mathcal{L}(\mathbf{J}'|\{s_k\}) = P(\{s_k\}|\mathbf{J}') \tag{3}$$

$$\log \mathcal{L}(\mathbf{J}'|\{s_k\}) = -\sum_k E(s_k|\mathbf{J}')/T - N \log Z(\mathbf{J}') \ . \tag{4}$$

For specific systems, we solve the spin-network inference problem by estimating the network couplings as

$$\widehat{\mathbf{J}} = \mathrm{argmax}_{\mathbf{J}'}\mathcal{L}(\mathbf{J}'|\{s_k\}) \ , \tag{5}$$

the parameter values $\widehat{\mathbf{J}}$ that maximize the likelihood $\mathcal{L}$ given a set of network state observations $\{s_k\}$.

## II. INFORMATION AND PHYSICS IN MODEL INFERENCE

When is accurate inference possible in a spin network? We seek to convert observations drawn from the true equilibrium distribution $P(s_k|\mathbf{J})$ into information about the couplings (spin interactions) $\mathbf{J}$ in the underlying network. How much information does $P(s_k|\mathbf{J})$ carry about $\mathbf{J}$?

Fundamentally, inference is impacted by the environment of a network. For example, $P(s_k|\mathbf{J})$ is strongly dependent on temperature. As $T \to 0$, $P(s_k|\mathbf{J})$ becomes dominated by ground states that minimize $E(s_k|\mathbf{J})$, and many coupling matrices $\mathbf{J}$ have the same ground states. Similarly, at high temperature, all networks achieve the same state distribution $P(s_k|\mathbf{J}) \sim 1/2^m$: samples are uniformly distributed across all $2^m$ possible network states, independent of $E(s_k|\mathbf{J})$, and so all networks share the same $P(s_k|\mathbf{J})$. These simple arguments suggest an optimal regime for inference at an intermediate temperature, when training data is not confined to just ground states, yet remains distributed non-uniformly to provide information about the relative energy of different states and hence the underlying couplings $\mathbf{J}$.

### A. The information content of observations

A good regime for learning is characterized by a data distribution that changes dramatically for a small change in underlying system parameters. A standard measure for distinguishability—here between the state distribution $P(s|\mathbf{J})$ generated by the network with couplings $\mathbf{J}$ and the distribution $P(s|\mathbf{J}')$ predicted by the model with inferred couplings $\mathbf{J}'$ [19]—is the relative entropy (Kullback-Leibler divergence) [20],

$$D[P(s|\mathbf{J})||P(s|\mathbf{J}')] \equiv \sum_k P(s_k|\mathbf{J}) \log \tfrac{P(s_k|\mathbf{J})}{P(s_k|\mathbf{J}')} \ . \tag{6}$$

$D[P(s|\mathbf{J})||P(s|\mathbf{J}')] = 0$ if and only if the two distributions are identical. When the two distributions are similar ($\mathbf{J}' \approx \mathbf{J}$), the leading-order contribution to the relative entropy is

$$D[\mathcal{L}(\mathbf{J}|S)\|\mathcal{L}(\mathbf{J}'|S)] \approx \tfrac{1}{2}\mathbf{J}^T\mathcal{I}\mathbf{J} \ , \tag{7}$$

in terms of the Fisher information matrix [20],

$$\mathcal{I}_{ij,\ell m}(\mathbf{J}) \equiv \left\langle \frac{\partial^2 \log \mathcal{L}(\mathbf{J}|s_k)}{\partial J_{i,j}\partial J_{\ell,m}} \right\rangle_{P(s_k|\mathbf{J})} \ , \tag{8}$$

where the $ij, \ell m$-entry gives the mixed partial derivative of the log-likelihood with respect to the coefficient coupling spins $i$ and $j$, and the coefficient coupling spins $k$ and $\ell$. The Fisher information measures the curvature of the relative entropy near the true parameters, and it thus provides a quantitative measure of the expected information content of an observation drawn from the equilibrium distribution $P(s_k|\mathbf{J})$.

For spin networks, diagonal entries have a simple analytic form:

$$\mathcal{I}_{ij,ij}(\mathbf{J}) = \frac{\left\langle \left(\frac{\partial E}{\partial J_{i,j}}\right)^2 \right\rangle - \left\langle \frac{\partial E}{\partial J_{i,j}} \right\rangle^2}{T^2} \tag{9}$$

$$= \frac{\mathrm{Var}[\sigma_i \sigma_j]}{T^2} \ . \tag{10}$$

The Fisher information is the ratio of $\mathrm{Var}[\sigma_i\sigma_j]$, the fluctuations in spin-spin alignment within the network, to the squared temperature. This reveals a trade-off: greater thermal energy increases the magnitude of network fluctuations—populating excited states to permit the inference of their energies and hence spin couplings—while also reducing the difference in equilibrium probabilities of states with different energies. The optimal temperature for learning balances these two effects.

The spin-spin alignment variance is directly related to the spin-spin correlation function $\langle \sigma_i \sigma_j \rangle$ via

$$\mathrm{Var}[\sigma_i \sigma_j] = 1 - \langle \sigma_i \sigma_j \rangle^2 \ . \tag{11}$$

$\langle \sigma_i \sigma_j \rangle$ has been a central object in the study of spin networks [21], providing an operational measure of spin-network ordering [15, 21]. Moreover, $\langle \sigma_i \sigma_j \rangle$ can be computed directly from observations, permitting estimation of the Fisher information.

Figure 2 illustrates the relationship between correlations, information, and network state for a simple model spin network containing two groups of spins, with positive intra-group couplings and a single negative inter-group coupling. As $T \to 0$, the network has intrinsically low Fisher information, and inference is degenerate because the true network occupies a similar set of ground states to a test network with similar topology. As temperature increases, network fluctuations increase, and network states become occupied differently between the true and test networks. Specifically, the negative couplings in network $\mathbf{J}'$ enforce coherence between spins within each group. As $T \to \infty$, inference again becomes degenerate, as for both networks $P(s_k) \to \frac{1}{2^m}$ for all states $s_k$.

### B. Information and sampling complexity

$\mathcal{I}_{ij,ij}$ provides a direct measure of inference efficiency. Specifically, the Cramer-Rao bound relates the Fisher information to a lower bound on the inference error, defined

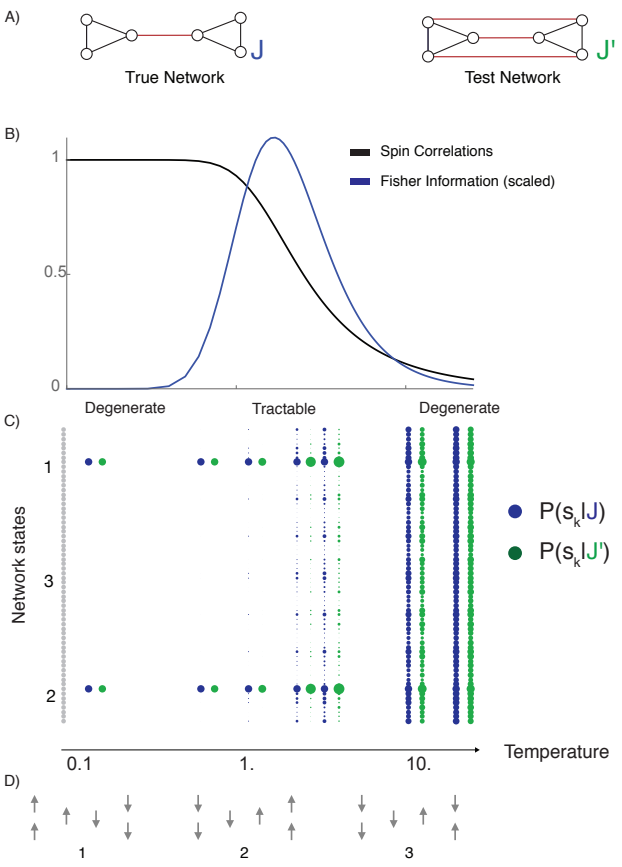


FIG. 2. Fluctuations and information. (A) True and test network architectures, with black links indicating positive couplings and red links negative couplings. (B) The Fisher information $\mathcal{I}_{ij,ij}$ and spin-spin correlation function $\langle \sigma_i \sigma_j \rangle$, averaged over spin pairs in true network $\mathbf{J}$. (C) Equilibrium probability $P(s_k|\mathbf{J})$ of spin-network states, given true coupling $\mathbf{J}$ (blue) or test coupling $\mathbf{J}'$ (green). Size of dot indicates probability. (D) Example spin-network states, with numerical labels corresponding to those in (C).

as the mean-squared error between $\mathbf{J}$ and an unbiased estimator $\widehat{\mathbf{J}}$ [20].

In general, the Cramer-Rao bound is stated for the inverse of the complete Fisher information matrix, however a convenient loose bound is [22]

$$\left\langle \left(\widehat{J}_{i,j}(\{s_k\}) - J_{i,j}\right)^2 \right\rangle_{P(\{s_k\}|\mathbf{J})} \geq \frac{1}{N\mathcal{I}_{ij,ij}(\mathbf{J})} \ , \tag{12}$$

for $N$ observations drawn from $P(s_k|\mathbf{J})$. In this way, $\mathcal{I}$ quantifies the information that observations drawn from $P(s_k|\mathbf{J})$ provide about the value of network couplings $\mathbf{J}$.

For inference error (averaged over unique spin pairs $\{i,j\}$)

$$\epsilon \equiv \frac{\sum_{\{i,j\}} \left(\widehat{J}_{i,j}(\{s_k\}) - J_{i,j}\right)^2}{\frac{m^2-m}{2}} \ , \tag{13}$$

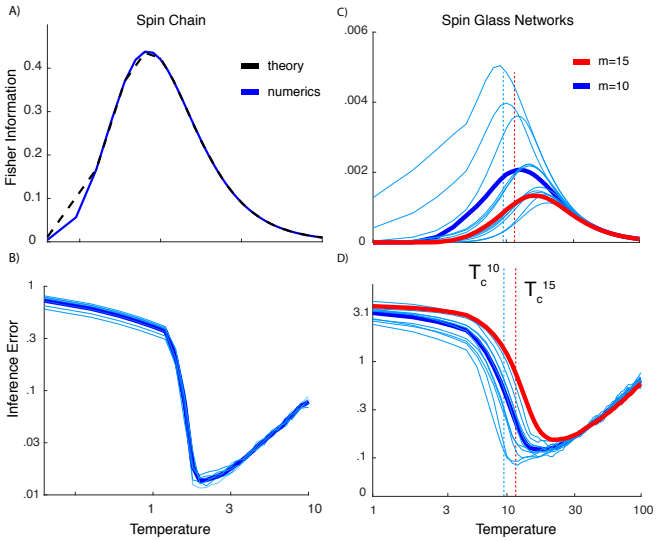


FIG. 3. Inference in example network architectures. Fisher information (top row) and inference error (bottom row) versus temperature for uniform 1D chain composed of $m = 10$ spins (left column) and spin glass network for $m = 10$ or 15 (right column). See Appendix for numerical details. (A) Fisher information theoretical expression (17) (black dashed curve) closely matches numerical estimation of $\min_{ij} \mathcal{I}_{ij,ij}$, the minimum diagonal entry of the Fisher information matrix (blue solid curve). (B) Inference error from maximum-likelihood inference of $m = 10$ chain, using L1 norm, as a function of temperature for 50 replicates (light blue) and mean (dark blue). (C) Minimum diagonal entry of Fisher information matrix, $\min_{ij} \mathcal{I}_{ij,ij}$, for spin glass network with $m = 10$ (blue) or $m = 15$ (red). For $m = 10$, 50 individual runs (light blue), mean (dark blue), and $T_c$ (blue dotted line) are shown. For $m = 15$, mean (dark red) and $T_c$ (red dotted line) are shown. (D) Inference error for corresponding networks.

the Fisher information provides a direct measure of the sampling complexity, defined in machine learning as the number of samples $N_{\epsilon_0}$ required to achieve a desired bound $\epsilon_0$ on the average error:

$$N_{\epsilon_0} \geq \frac{1}{\epsilon_0} \sum_{\{i,j\}} \frac{1}{\mathcal{I}_{ij,ij}(\mathbf{J})} \tag{14}$$

$$\geq \frac{1}{\epsilon_0} \sum_{\{i,j\}} \frac{T^2}{\left(1 - \langle \sigma_i \sigma_j \rangle^2_{P(s_k|\mathbf{J})}\right)} . \tag{15}$$

## III. INFERENCE IN MODEL SPIN-NETWORK ARCHITECTURES

### A. Optimal inference for 1D spin chains

Analytic forms for $\langle \sigma_i \sigma_j \rangle$ (and hence the Fisher information) cannot in general be directly computed, however some simple networks do permit analytic expressions. An interesting special case is the 1-dimensional spin chain, a network with $J_{i,i+1} = 1$ and $J_{i,j} = 0$ otherwise, which has simple correlations and Fisher information [21]:

$$\langle \sigma_i \sigma_j \rangle_{P(s_k|\mathbf{J})} = \left(\tanh \frac{1}{T}\right)^{|i-j|} \tag{16}$$

$$\mathcal{I}_{ij,ij}(\mathbf{J}) = \frac{1 - \left(\tanh \frac{1}{T}\right)^{2|i-j|}}{T^2} . \tag{17}$$

Figure 3A shows the Fisher information increasing as a function of temperature, obtaining a maximum at $T_{\text{opt}}$, and decaying asymptotically to zero as $T \to \infty$. Similarly, maximum-likelihood inference error is minimized at a finite temperature (Fig. 3B).

Further, (17) shows that the average information content of observations increases as a function of distance $|i - j|$ between spins, because $\langle \sigma_i \sigma_j \rangle_{P(s_k|\mathbf{J})}$ decays exponentially with distance. Thus, long-range links are 'easier' to learn because correlations between distant spins are weaker. Nearby links are harder to learn because correlations are strong and less information accrues per observation. In the 1D spin chain, the challenge of inference is dominated by the effort of learning the coupling $J_{i,i+1}$ of the most highly correlated nearest-neighbor spins.

### B. Optimal inference in spin glass networks

For specific random ensembles of $\mathbf{J}$, we can numerically evaluate the Fisher information. An important class of spin networks are Gaussian random spin glasses, where couplings are Gaussian-distributed according to $P(J_{i,j}) \sim \mathcal{N}(0; K^2)$, a normal distribution with mean 0 and variance $K^2$. These networks have both positive and negative couplings and rich physical behaviors, so have been a model system for studying complex interacting systems.

Figure 3 shows $\mathcal{I}_{ij,ij}(\mathbf{J}, T)$ averaged over a set of spin glasses, for network size $m = 10$ or 15. Consistent with our asymptotic arguments in Section II, Fig. 3 reveals a local maximum in $\mathcal{I}_{ij,ij}(\mathbf{J}, T)$ as a function of $T$, across many randomly sampled parameter sets. Moreover, the maximum in $\mathcal{I}_{ij,ij}(\mathbf{J}, T)$ approximately coincides with a minimum in inference error.

This optimal inference temperature $T_{\text{opt}}$ appears to track the critical temperature

$$T_c = \frac{1}{2}\sqrt{\frac{K^2 m}{\ln 2}} \tag{18}$$

of a well-characterized phase transition in a Gaussian spin glass in the absence of a magnetic field $h$ [21, 23]. Below this critical temperature, the network's state distribution is concentrated on a small number of low-energy states, and for finite samples explores only an exponentially small fraction of possible configurations. Near the phase transition, the state distribution $P(s_k)$ transitions abruptly from being 'frozen' (sharply peaked on a small number of ground states) to being dispersed, and there

is a correspondingly abrupt change in the spin-spin correlations, the Fisher information, and the efficiency of inference.

## IV. OPTIMIZING INFERENCE WITH ACTIVE LEARNING

In general, the optimal temperature for inference, like $T_c$ for Gaussian networks, depends on parameters including the network topology and the variance of network couplings. Such network-specific parameters, and thus the optimal temperature for inference, are generally unknown *a priori* by an observer. This motivates the development of active learning paradigms that can determine $T_{\rm opt}$ through interaction with the system.

We design a learning protocol that operates in two stages. In the first stage, we directly manipulate temperature to optimize inference efficiency. We estimate the optimal temperature $T_{\rm opt}$ by empirically maximizing the minimum estimated Fisher information across all coupling coefficients,

$$\widehat{T}_{\rm opt} \equiv \text{argmax}_T \left[ \min_{ij} \widehat{\mathcal{I}}_{ij,ij}(\mathbf{J}) \right] \ . \tag{19}$$

Importantly, the spin-network Fisher information can be estimated on-line via the empirical spin-spin correlation function $\frac{1}{N}\sum_{\{s_k\}} \sigma_i^{(k)}\sigma_j^{(k)}$ sampled during an initial temperature 'sweep:'

$$\widehat{\mathcal{I}}_{ij,ij}(\mathbf{J}) = \frac{1 - \left(\frac{1}{N}\sum_{\{s_k\}} \sigma_i^{(k)}\sigma_j^{(k)}\right)^2}{T^2} \ , \tag{20}$$

In the second stage, we use maximum-likelihood estimation (see Appendix) to infer $\mathbf{J}$ from observations collected at $\widehat{T}_{\rm opt}$, the inferred optimal temperature for inference.

To test an interesting case, we selected a relatively large network (with $m = 60$ nodes) based on a biological reconstruction of a cortical brain circuit [24]. Figure 4 shows the initial estimation of Fisher information as a function of temperature for the true network, which found a global maximum near $T = 11$. Maximum-likelihood inference at $T = 11$ produced an L1 error of less than 2% per coupling. Similar inference at $T = 1$ produced a network with uniformly high error. Figure 4D shows the topology of the networks inferred at $T = 11$ and $T = 1$.

In this way, temperature optimization allows efficient learning of large networks. The example suggests that optimization protocols might be useful for experimentally determining models for natural networks of biological interest. More sophisticated strategies could incorporate additional classes of perturbations.

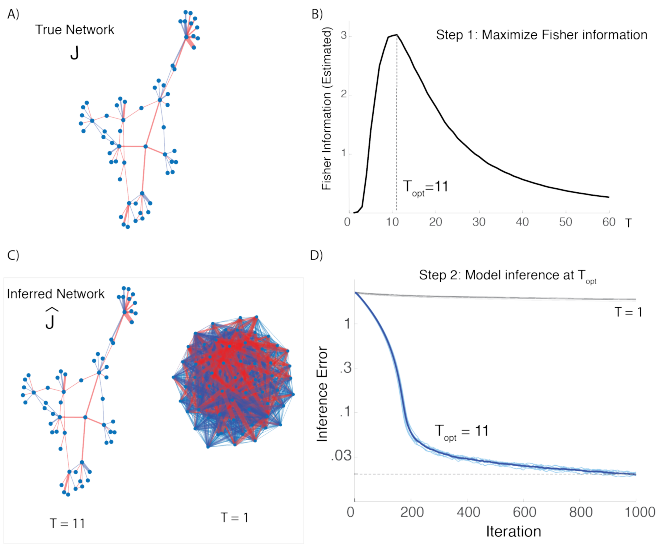


FIG. 4. Inference with active learning. (A) Spin network modeled on mouse cortex reconstruction from [24]. Edges are red for $J_{i,j} < 0$ and blue for $J_{i,j} > 0$. (B) In first step of active learning, a temperature sweep finds that Fisher information is maximized at $T = 11$. (C) Network reconstructed through likelihood maximization (see Appendix) from equilibrium samples collected at $T = 11$ (left) correctly identifies topology of the network and the sign of couplings $J_{i,j}$, whereas the network inferred at $T = 1$ has many erroneous linkages. In both cases, links omitted for $|J_{i,j}| < 0.1$. (D) Inference error as function of iteration, averaged over 10 runs, at $T = 11$ (blue) and $T = 1$ (black). Dotted line shows the 2% error threshold indicated in the text.

## V. CONCLUSION

In this paper, we studied fundamental limits imposed by the physical behavior of a spin network on the learning of predictive models. We showed that the physics of a spin network strongly affects the efficiency of inference. Specifically, fluctuations play a particularly important role in determining when predictive inference is possible.

Intuitively, inference is degenerate when many different networks produce similar distributions over states. This occurs generically both at low temperature, when a network explores a small number of configurations and is focused on ground states, and at high temperature when a network samples all states equally. Thus, below an optimal temperature, learning efficiency is enhanced by increasing the temperature, thereby increasing the strength of the thermal noise. Greater thermal energy allows the network to explore a broader range of states, thereby enabling the inference of sufficiently constrained models. Above the optimal temperature, additional thermal noise becomes detrimental to learning: when random driving is too high, couplings are hidden by the dominance of thermal noise.

Our results reveal a physical dimension of inference

and machine learning. In machine learning, standard paradigms—including PAC learning [13] and Empirical Risk Minimization [12, 14]—relate the complexity of the learning task to the statistical properties of training data. In PAC learning, for example, the probability of obtaining training samples from a class of interest determines the number of samples required to meet a desired error bound. In physical systems, the properties of training data are impacted directly by physical variables, providing a fundamental link between physics and learning.

Here, we argue that for a spin network there are optimal environmental conditions for learning. If this result holds more generally, a system such as a neural network or a gene regulatory network might benefit from experimental strategies that inject noise into the system. Further, there might be 'optimal' noise magnitudes when inference becomes most tractable.

Experimentally, our work motivates feedback protocols for interacting with a system to bring it near conditions of optimal inference. The framework we describe here would require closed-loop monitoring of spins and a 'noise source.' However, we anticipate that broader strategies for optimizing inference might include more direct perturbations to a system; for example, in spin networks one could manipulate magnetic field as well as temperature.

Originally developed in statistical mechanics to model phase transitions in simple magnetic materials, spin-network models have provided useful descriptions of non-equilibrium systems such as bird flocks and neural circuits [7, 15–18]. The broad utility of these models suggests that our results may provide an avenue to develop 'optimal inference protocols' for the construction of predictive models in a wide range of domains beyond the formal scope of equilibrium physics, including problems in chemistry, biology, and ecology.

## ACKNOWLEDGMENTS

This work was supported by grants from the NIH, the CZI, Amgen, and Rosen Center at Caltech (MT); a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, the Canada Research Chairs program, and the Faculty of Science, Simon Fraser University through the President's Research Start-up Grant (DAS); and WestGrid (www.westgrid.ca) and Compute Canada Calcul Canada (www.computecanada.ca). The authors thank Andrew Stuart, Joel Tropp, Rob Phillips, John Doyle, Carl Pabo, Malcolm Kennett, Emma Lathouwers, and Erik Winfree for useful discussions and feedback on the manuscript.

## Appendix A: Numerical Methods

In the numerical examples shown in Figs. 3 and 4, we perform maximum-likelihood estimation of spin network architectures from data. Briefly, we calculate the estimator $\widehat{\mathbf{J}}$ of the coupling matrix by numerically maximizing $\mathcal{L}(\mathbf{J}'|\{s_k\})$ via gradient ascent,given state observations $\{s_k\}$ drawn from $P(s_k|\mathbf{J})$, the equilibrium state distribution of the 'true' network with couplings $\mathbf{J}$.

Explicitly, we average over $N$ observations ($N = 50,000$ in Fig. 3; $N = 100,000$ in Fig. 4) the gradient of the log likelihood, $\log \mathcal{L}$:

$$\frac{\partial}{\partial J'_{ij}} \left[ \frac{1}{N} \sum_{\{s_k\}} \log \mathcal{L}(J'_{ij}|s_k) \right] = \beta \left( \frac{1}{N} \sum_{\{s_k\}} \sigma_i^{(k)} \sigma_j^{(k)} - \langle \sigma_i \sigma_j \rangle_{P(s_k|\mathbf{J}')} \right) . \tag{A1}$$

$\langle \sigma_i \sigma_j \rangle_{P(s_k|\mathbf{J}')}$ is the spin-spin correlation for equilibrium fluctuations under the estimated coupling $\mathbf{J}'$, and $\frac{1}{N} \sum_{\{s_k\}} \sigma_i^{(k)} \sigma_j^{(k)}$ is the empirical spin-spin correlation under the true coupling $\mathbf{J}$.

Each iteration of the learning algorithm applies the simple update rule to the current estimate $\mathbf{J}(n)$:

$$J'_{ij}(n+1) = J'_{ij}(n) + \eta \, \frac{\partial}{\partial J'_{ij}} \left[ \frac{1}{N} \sum_{\{s_k\}} \log \mathcal{L}(J'_{ij}|s_k) \right] . \tag{A2}$$

Here, $\eta$ is a parameter commonly known as the learning rate that modulates the size of the gradient step.

We evaluate error using the L1 norm,

$$\epsilon = \frac{\sum_{\{i,j\}} |J_{i,j} - \widehat{J}_{i,j}|}{\frac{m^2 - m}{2}} , \tag{A3}$$

because in examples the L1 norm provides an intuitive and direct measure of the total absolute magnitude of network coupling difference between the estimated and true networks.

[1] J. W. Lichtman, H. Pfister, and N. Shavit, Nat. Neurosci. **17**, 1448 (2014).
[2] J. Blumenstock, G. Cadamuro, and R. On, Science **350**, 1073 (2015).
[3] S. A. T. V. Svensson, R. Vento-Tormo, "Exponential scaling of single-cell rna-seq in the last decade," ArXiv:1704.01379.
[4] R. M. Shiffrin, Proc Natl Acad Sci USA **113**, 7308 (2016).
[5] K. Cerniol, G. Lahav, G. Suel, and O. Troyanskaya, Cell (2016).
[6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou,

V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Nature **529**, 484 (2016).
[7] G. E. Hinton and R. R. Salakhutdinov, Science **313**, 504 (2006).
[8] M. Schmidt and H. Lipson, Science **324**, 81 (2009).
[9] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, A. Lavin, and D. S. Phoenix, Science **358**, 2612 (2017).
[10] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Science Advances **3**, e1701816 (2017).
[11] J. Bailey, Science **252**, 1668 (1991).
[12] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1999).
[13] L. G. Valiant, in *Proceedings of the sixteenth annual ACM symposium on Theory of computing - STOC '84* (ACM Press, New York, New York, USA, 1984) pp. 436–445.
[14] T. Poggio and S. Smale, in *2005 International Conference on Neural Networks and Brain* (IEEE).
[15] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[16] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).
[17] Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin, Proc. Natl. Acad. Sci. USA **108**, 18720 (2011).
[18] S. Cocco, S. S. Leibler, and R. Monasson, Proc. Natl. Acad. Sci. USA **106**, 14058 (2009).
[19] S. Watanabe, *Algebraic geometry and statistical learning theory*, Vol. 25 (Cambridge University Press, 2009).
[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, 2006).
[21] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).
[22] B.-Z. Bobrovsky, E. Mayer-Wolf, and M. Zakai, The Annals of Statistics , 1421 (1987).
[23] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).
[24] D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, and R. C. Reid, Nature **471**, 177 (2011).