Model parameter learning using Kullback-Leibler divergence

Lin, C.; Marks, T.K.; Pajovic, M.; Watanabe, S.; Tung, C.

Abstract

In this paper, we address the following problem: For a given set spin configurations whose probability distribution is of the Boltzmann type, how do we determine the model coupling parameters? We demonstrate that directly minimizing the Kullback-Leibler divergence is a very efficient method. We test this method against the Ising and XY models on the one-dimensional (1D) and two-dimensional (2D) lattices, and provide two estimators to quantify the model quality. We apply this method to two types of problems. First we apply it to the real-space renormalization group (RG), and find that the obtained RG flow is sufficiently good for determining the phase boundary (within 1% of the exact result) and the critical point, but not accurate enough for critical exponents. The proposed method provides a simple way to numerically estimate amplitudes of the interactions typically truncated in the real-space RG procedure. Second, we apply this method to the dynamical system composed of self-propelled particles, where we extract the parameter of a statistical model (a generalized XY model) from a dynamical system described by the Viscek model. We are able to obtain reasonable coupling values corresponding to different noise strengths of the Viscek model. Our method is thus able to provide quantitative analysis of dynamical systems composed of self-propelled particles.

Physica A

Model parameter learning using Kullback-Leibler divergence

Chungwei Lin[a], Tim K. Marks, Milutin Pajovic, Shinji Watanabe

Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, USA

Chih-kuan Tung[b]

North Carolina Agricultural and Technical State University,
1601 E. Market Street, Greensboro, NC, 27411, USA

(Dated: July 3, 2017)

In this paper, we address the following problem: For a given set spin configurations whose probability distribution is of the Boltzmann type, how do we determine the model coupling parameters? We demonstrate that directly minimizing the Kullback-Leibler divergence is a very efficient method. We test this method against the Ising and XY models on the one-dimensional (1D) and two-dimensional (2D) lattices, and provide two estimators to quantify the model quality. We apply this method to two types of problems. First we apply it to the real-space renormalization group (RG), and find that the obtained RG flow is sufficiently good for determining the phase boundary (within 1% of the exact result) and the critical point, but not accurate enough for critical exponents. The proposed method provides a simple way to numerically estimate amplitudes of the interactions typically truncated in the real-space RG procedure. Second, we apply this method to the dynamical system composed of self-propelled particles, where we extract the parameter of a statistical model (a generalized XY model) from a dynamical system described by the Viscek model. We are able to obtain reasonable coupling values corresponding to different noise strengths of the Viscek model. Our method is thus able to provide quantitative analysis of dynamical systems composed of self-propelled particles.

---

[a] clin@merl.com
[b] ctung@ncat.edu

I. INTRODUCTION

In a general sense, the heart of any machine learning (ML) method involves learning a representation for patterns in a set of data. The representation can have many forms, such as a function, a probability distribution with many parameters, or the weights of a neural network [1, 2]. Combined with density functional theory (DFT) [3, 4], ML techniques have been used to design/select materials with certain properties of interest. Generally, DFT calculations of many materials (or the same material with different configurations) are needed. At this stage, some generic ML algorithms can be used to generate/select the configurations [5]. The next step is to fit the results to some kind of model with undetermined parameters [6]; the fitted model is assumed to be general and can be used to predict material properties. For example, in [5], more than 20,000 Au_{12}, Au_{13}, and Au_{14} nanoclusters are evaluated to fix the parameters of an empirical potential, which can then be used for larger clusters of arbitrary shapes. Similar strategies can be used to predict/screen functional molecules [7, 8] and thermoelectric materials [9, 10]. At a more fundamental level, ML has been used to find the density functional [11] or to find the pseudopotential [12] in DFT. ML has also been applied to solve many-body problems, such as the Anderson impurity model [13] and lattice problem in the framework of dynamical mean field theory [14–16]. Recently, Broecker et al. show that ML techniques (neural networks) can overcome the Fermion sign problem under some circumstances [17].

In this paper, we focus on the following problem: Suppose one has a set of Ising or Heisenberg spin configurations, whose distribution obeys the Boltzmann distribution for a given model energy functional, how do we extract the model parameters from this set of configurations [18–21]? In this case, "learning" means extracting the coupling values from a predefined model. We demonstrate that this model parameter learning can be done efficiently by *directly* minimizing Kullback-Leibler (KL) divergence [22], which is a major objective criterion used to obtain model parameters in ML. KL-based learning is commonly used in ML, which motivates us to apply it to some physical systems. We apply this method to two types of problems. First, we apply the method to the real-space renormalization group (RG), where the RG flow is explicitly calculated and the phase boundaries are determined. Second, we use it to study the system of interacting active or self-propelled particles. Specifically, we extract the parameter of a statistical model (generalized XY model) from the dynamics described by the Viscek model [23]. Overall, we find that if the functional form of the model is known, this method is very efficient. The efficiency originates from that only *model-specific* observables are required. For the same reason, this method easily give an answer even if the model is not correct, and some quantitative estimators of the model quality are needed.

The rest of the paper is organized as follows. In Section II, we formulate the problem, and describe how to use KL divergence to fit the model parameters. We use the two-dimensional (2D) Ising model as an illustrative example and discuss limitations and possible pitfalls of this method. Two estimators of the model quality are proposed. In Section III, we apply this method to the real-space renormalization group (RG) to determine the RG flow and thus the critical point. In Section IV, we apply this method to determine the statistical model parameter from a dynamical system of self-propelled particles. A brief introduction of the problem will be given. A short conclusion is provided in Section V.

II. MAIN ALGORITHM

A. Problem formulation and Kullback-Leibler divergence

We first formulate the problem in terms of probability distributions. Suppose we have a "target" distribution $\bar{P}(\{S\})$, with $\{S\}$ labeling the spin configuration. The goal is to find the best functional $E(\{S\}; J_1, J_2, ..., J_n)$ such that the model probability $P(\{S\}; J_1, J_2, ..., J_n) \propto e^{-E(\{S\}; J_1, J_2, ..., J_n)}$, parametrized by $J_1, J_2, ..., J_n$, is as close to $\bar{P}(\{S\})$ as possible. The probability follows a Boltzmann distribution, a form which can be derived from the maximum entropy principle [24]. In the following derivation, we assume there are only two parameters J_1 and J_2, but the generalization to more parameters is straightforward. The Kullback-Leibler (KL) divergence (from $\bar{P}$ to P) [22] is defined as

$$D_{\text{KL}}(\bar{P}||P(J_1, J_2)) = D_{\text{KL}}(J_1, J_2) \equiv \sum_{\{S\}} \bar{P}(\{S\}) \log \left[\frac{\bar{P}(\{S\})}{P(\{S\}; J_1, J_2)} \right], \tag{1}$$

where the summation is over all configurations $\{S\}$. This is a scalar function of J_1 and J_2 that measures the difference between two probability distributions. When the target distribution $\bar{P}(\{S\})$ and the model distribution $P(\{S\}; J_1, J_2)$ are identical, $D_{\text{KL}} = 0$; when the difference is small, D_{KL} is also small. We consider the general energy functional

$E(\{S\}; J_1, J_2) = -J_1 F_1(\{S\}) - J_2 F_2(\{S\})$, leading to a model probability distribution

$$P(\{S\}; J_1, J_2) = \frac{e^{J_1 F_1(\{S\}) + J_2 F_2(\{S\})}}{\sum_{\{S'\}} e^{J_1 F_1(\{S'\}) + J_2 F_2(\{S'\})}}. \tag{2}$$

Here $F_i(\{S\})$ represents some form of inter-spin interactions, which are not restricted to the pairwise form and will be specified later. In this case, the KL divergence becomes

$$D_{\mathrm{KL}}(J_1, J_2) = \sum_{\{S\}} \bar{P}(\{S\}) \log \bar{P}(\{S\}) - \sum_{\{S\}} \bar{P}(\{S\}) \log P(\{S\}; J_1, J_2) \tag{3}$$

$$= \text{const.} - \sum_{\{S\}} \bar{P}(\{S\}) \left[J_1 F_1(\{S\}) + J_2 F_2(\{S\}) \right] + \sum_{\{S\}} \bar{P}(\{S\}) \log \left[\sum_{\{S'\}} e^{J_1 F_1(\{S'\}) + J_2 F_2(\{S'\})} \right] \tag{4}$$

$$= \text{const.} - \sum_{\{S\}} \bar{P}(\{S\}) \left[J_1 F_1(\{S\}) + J_2 F_2(\{S\}) \right] + \log \left[\sum_{\{S\}} e^{J_1 F_1(\{S\}) + J_2 F_2(\{S\})} \right], \tag{5}$$

where const. represents terms that have no dependence on J_1 or J_2. To get from (4) to (5), we used $\sum_{\{S\}} \bar{P}(\{S\}) = 1$. To shorten the expression, we define $\langle O \rangle_{target} \equiv \sum_{\{S\}} \bar{P}(S) O\{S\}$, and $\langle O \rangle_{model} \equiv \sum_{\{S\}} P(\{S\}; J_1, J_2) O(\{S\})$. The derivatives of $D_{\mathrm{KL}}(J_1, J_2)$ are given:

$$\frac{\partial D_{\mathrm{KL}}}{\partial J_1} = \langle F_1 \rangle_{model} - \langle F_1 \rangle_{target} \equiv \alpha_1(J_1, J_2),$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial J_2} = \langle F_2 \rangle_{model} - \langle F_2 \rangle_{target} \equiv \alpha_2(J_1, J_2), \tag{6}$$

$$\frac{\partial^2 D_{\mathrm{KL}}}{\partial J_i \partial J_j} = \langle F_i F_j \rangle_{model} - \langle F_i \rangle_{model} \langle F_j \rangle_{model} \equiv M_{ij}(J_1, J_2).$$

Note that $M_{ij} = M_{ji}$, and that all second derivatives depend only on the model distribution (not on the target distribution).

With these derivatives determined, we use Newton's method to estimate the values of J_1, J_2 that minimize $D_{\mathrm{KL}}(J_1, J_2)$. To do so, we write the second-order Taylor expansion

$$D_{\mathrm{KL}}(J_1 + \Delta J_1, J_2 + \Delta J_2) \approx \alpha_1 \Delta J_1 + \alpha_2 \Delta J_2 + \frac{1}{2} \sum_{ij=1}^{2} M_{ij} \Delta J_i \Delta J_j, \tag{7}$$

up to a J_i-independent constant which we ignore. For later use, we also write down the third derivative:

$$\frac{\partial^3 D_{\mathrm{KL}}}{\partial J_i^3}(J_1, J_2) = \langle F_i^3 \rangle_{model} - 3 \langle F_i^2 \rangle_{model} \langle F_i \rangle_{model} + 2 \langle F_i \rangle_{model}^3. \tag{8}$$

To derive the optimization step, we set the first derivatives of Eq. (7) to zero, i.e. $\frac{\partial D_{\mathrm{KL}}}{\partial J_1} = \frac{\partial D_{\mathrm{KL}}}{\partial J_2} = 0$, and solve for the increment in the J_i values:

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} \Delta J_1 \\ \Delta J_2 \end{pmatrix} + \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = 0$$

$$\Rightarrow \begin{pmatrix} \Delta J_1 \\ \Delta J_2 \end{pmatrix} = - \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \tag{9}$$

where the α_i and M_{ij} all depend on J_1 and J_2. Newton's method iteratively updates the model parameters J_1 and J_2 using Eq. (9). In practice, we use a relaxed Newton's method in which the step size is a fraction of the one given by Eq. (9): $J_i^{new} = J_i^{old} + \tilde{\eta} \Delta J_i$, where $\tilde{\eta} \leq 1$ is known as the learning rate. The iterations stop when ΔJ_1 and ΔJ_2 are smaller than some predefined value. The same procedure can be generalized to models with more parameters. The most crucial and time-consuming part is to compute $\langle \cdots \rangle_{model}$ at different model parameters. Fortunately, these model expectation values can be computed efficiently using a Markov chain Monte Carlo (MCMC) method, such as Metropolis or "heat bath" algorithms.

In addition to Newton's method, one can also use Levenberg-Marquardt (LM) algorithm to solve the "root finding" problem, i.e. to find J_1, J_2 such that $\alpha_1 = \alpha_2 = 0$ in Eq. (6). The LM method is particularly useful if the second-derivative matrix M_{ij} is ill-conditioned. In all the examples we have tested, M_{ij} is well-conditioned and computing its inverse does not result in any numerical problems. Overall, we find LM and Newton's method are quite comparable in performance. However, when the number of parameters becomes large, LM method appears to be more appropriate (ill-conditioning can happen [25]). An example will be given in the next subsection [Table I].

B. Illustrative examples

(a) 2D Ising model $\tilde{\eta} = 0.8$ **(b) 2D XY model** $\tilde{\eta} = 0.4$
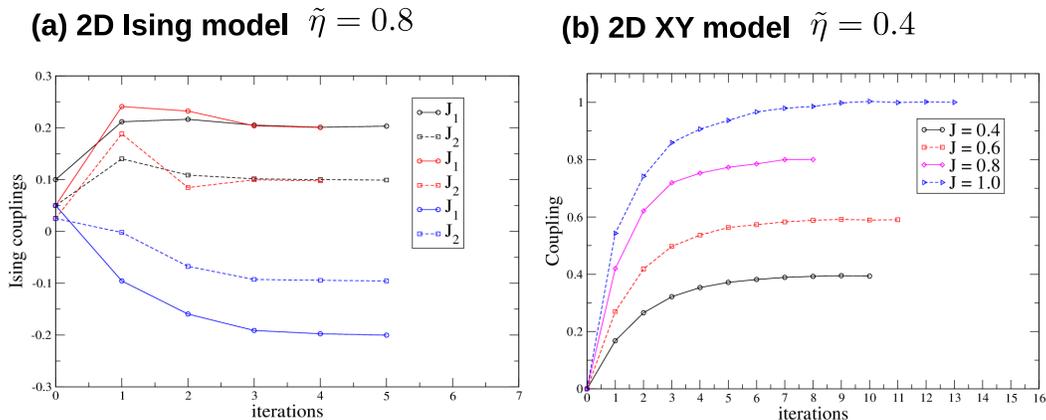
FIG. 1. (a) The black and red curves show convergence trajectories of 2D Ising model with $J_1 = 0.2$, $J_2 = 0.1$, using two different initial conditions. The blue curves show convergence trajectories for $J_1 = -0.2$, $J_2 = -0.1$ for an initial condition with the wrong signs. The learning rate $\tilde{\eta} = 0.8$. (b) The convergence trajectories of 2D XY model with nearest neighbor coupling $J = 0.4$, 0.6, 0.8, and 1. The learning rate $\tilde{\eta} = 0.4$.

To test our method, we first generate a set of configurations for a given fixed value of the model parameters, using the typical "heat bath" algorithm. That set of samples defines the target probability $\bar{P}(\{S\})$. We then use the procedure described in Section II A to estimate the model parameters. As an illustrative example, we consider the Ising model on a periodic 8×8 square lattice, with both nearest-neighbor (NN) and second-nearest-neighbor (next-nearest-neighbor, or NNN) couplings. In this case, we define

$$F_1(\{S\}) = \sum_{NN} S_i S_j, \ \ F_2(\{S\}) = \sum_{NNN} S_i S_j. \tag{10}$$

For the energy defined by $J_1 F_1(\{S\}) + J_2 F_2(\{S\})$, J_1 and J_2 respectively represent the nearest neighbor (NN) and next-nearest-neighbor (NNN) couplings. We choose $(J_1, J_2) = (0.2, 0.1)$ and $(J_1, J_2) = (-0.2, -0.1)$. The convergence trajectories, computed from Eq. (9), for different initial conditions are shown in Fig. 1(a). We find that the convergence is generally reached using the small initial values of couplings and the learning rate $\tilde{\eta} < 1$; it typically takes about 10 to 20 iterations to converge. Fig. 1(b) shows the convergence trajectories of XY model on a periodic 8×8 square lattice with only the nearest neighbor coupling J from 0.4 to 1. We do not see any numerical issue when J crosses the Kosterlitz-Thouless (KT) critical $J \sim 2/\pi \approx 0.64$ of the 2D XY model [26–28].

Using the described procedure, we have tested our method on the 1D and 2D (square and triangular lattices, up to size 40×40) Ising model with NN, second NN, and third NN couplings (the sign of the coupling can be positive or negative). We have also tested it on the 1D and 2D (square lattice, up to size 30×30) XY model with NN coupling (up to the coupling strength of 10). The boundary condition can be periodic or confined (i.e. the spins at the boundary only couple to those inside the boundary). As long as the system is in the disordered state (discussed in next subsection), the method succeeds on all of these models with high accuracy (error $< 1\%$). The method also works in the ferromagnetic phase for the model of continuous spins (XY model), but requires more sampling points. We have applied this method to quantitatively determine the coupling between collective swimming particles [29].

To further test the proposed method, we consider a 1D chain of 40 Ising spins, with 10 inter-spin couplings as $J_1 = 0.2$, $J_n = (-1)^{n+1} J_1 [1 - 0.1(n-1)]$ for $n = 2$ to 10. Fig. 2 shows the convergence trajectories of each J_i,
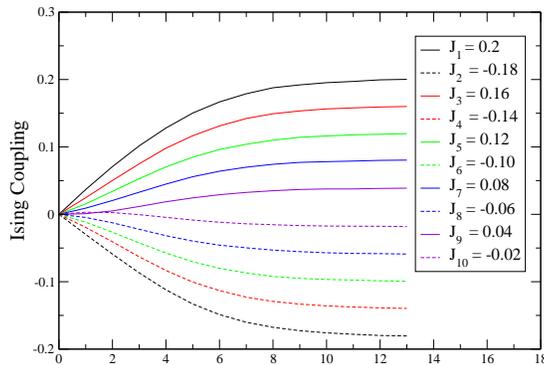
FIG. 2. The convergence trajectories of 1D chain of 40 Ising spins. J_i is the inter-spin coupling between nth and $n + i$th sites. The exact values are given in the box, and the converged values are given in Table I. The learning rate $\tilde{\eta} = 0.4$.

with zero initial couplings. The converged values are compared to the exact in Table I. We see that both Newton's method and LM method give good parameter estimations in this 10-parameter case. We find that Newton's method usually takes less iterations to converge, but the total computational times are quite comparable. We have increased the parameter number up to 14 and does not find any problems.

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
exact	0.200	-0.180	0.160	-0.140	0.120	-0.100	0.080	-0.060	0.040	-0.020
Newton	0.199	-0.178	0.158	-0.137	0.118	-0.097	0.077	-0.056	0.036	-0.020
LM	0.202	-0.181	0.163	-0.140	0.121	-0.100	0.081	-0.059	0.037	-0.018

TABLE I. The inter-spin couplings for the 1D chain of 40 Ising spins. J_i represents the coupling between spins at nth and $n + i$th site. Both Newton's and LM methods give good parameter estimations. Both Newton and LM iterations start from zero couplings. In this particular case, LM takes 20 iterations to converge, whereas Newton's method takes 13. The computed values are rounded to three decimal digits.

C. Limitations of the method and model quality

We now discuss the limitation of the proposed method. First, our method becomes less efficient when the state is more ordered, as the method relies on spin fluctuations to determine the coupling constants. Considering the extreme case in which all spins are aligned ferromagnetically, it is then not possible to determine whether this ordering comes from J_1, J_2, or any combination of them. Second, there is a theoretical limit of the model parameter estimation, given by the Fisher information

$$I(J) = \sum_{\{S\}} \left[\frac{\partial}{\partial J} \log P(\{S\}; J) \right]^2 P(\{S\}; J). \tag{11}$$

We note that Eq. (11) is simply $\langle F_i^2 \rangle_{model}$. The eigenvalues of Fisher information have recently been used to connect the microscopic model and the emergent macroscopic model [25]. The Cramér-Rao bound [30], which gives the *minimum* of parameter variance, is given by

$$Var\{J\} > \frac{1}{I(J)}. \tag{12}$$

It turns out that in all models considered here, the variance of J is smaller than 0.01. Therefore, for the rest of the paper we do not explicitly mention this limit.

To quantitatively characterize the model quality, there are two general strategies: (i) comparing to the values obtained using a model with more parameters; and (ii) comparing more expectation values [for example the second moment $\langle F_i^2 \rangle$ with F_i defined in Eq. (10)] between target and model distributions. For (i), we increase the number of

couplings and observe how the added couplings impact the results. The ratio

$$\frac{\left|J_i^{[N+m]} - J_i^{[N]}\right|}{\left|J_i^{[N]}\right| + 1} \equiv R_{i,[N,m]}, \tag{13}$$

where N is the number of original couplings and m is the number of added couplings, can be used as an estimator of model quality of the N-parameter model. The $+1$ term in the denominator is to account for the case when $\left|J_i^{[N]}\right|$ approaches zero. Certainly, this ratio depends on the form of the added interactions. Some physics principles, e.g. the coupling strength decays with increasing distance, provide a guide on how the additional couplings should be chosen. For (ii), we expand the KL divergence to the third order, from which we can estimate the error of the coupling that was fitted using the second-order expansion. For simplicity (but relevant to the application in Section IV), we consider the case where there is only one coupling. This estimation is based an assumption that the target distribution can be fully described by a model with only the coupling J_0, and an assumption that the fitted J is close to J_0. Using $\Delta = J - J_0$ as the expanding variable, we get a Taylor series around $\Delta = 0$ (or $J = J_0$) as

$$D_{\mathrm{KL}}(\Delta) = \frac{1}{2}\alpha_{target}\Delta^2 + \frac{1}{6}\beta_{target}\Delta^3 + \cdots$$

Here α_{target} and β_{target} are respectively the second and third derivative given in Eq. (6) and Eq. (8), with the "model average" replaced by the "target average" (since the model of coupling J_0 corresponds to the target distribution). The first derivative vanishes as J_0 corresponds to the target distribution. When we compute the second derivative using a model distribution, we get α_{model}, which is different from α_{target}. This discrepancy comes from the difference between J and J_0. Explicitly evaluating the second derivative, we get $\frac{d^2}{d\Delta^2}D_{\mathrm{KL}}(\Delta) = \alpha_{target} + \beta_{target}\Delta$, from which Δ can be estimated via $\alpha_{model} = \alpha_{target} + \beta_{target}\Delta$. Define $d\alpha = \alpha_{target} - \alpha_{model}$, we get $\Delta = \frac{|d\alpha|}{\beta_{target}}$. Note that if $d\alpha = 0$ ($\alpha_{target} = \alpha_{model}$, implying the model and target average of the second moment are identical), then $\Delta = 0$ (the fitted J coincides with the exact J_0). For computational convenience, we replace β_{target} by β_{model}, and define

$$R = \frac{\Delta}{|J| + 1} = \frac{|d\alpha|/\beta_{model}}{|J| + 1} \tag{14}$$

as another estimator. If the model form is not known in advance, at least one of the two estimators should be used to check the model quality. In some situations, it is not obvious how to add couplings (especially when there is no lattice, as discussed in Section IV), and using Eq. (14) is an apparent choice. On the other hand, if the model already has many parameters, an estimator based on (ii) easily becomes complicated (as all third derivatives are needed), so using (i) is perhaps more practical.

D.   Important features

We conclude this section by discussing some features of our proposed method. First, our method is straightforward and simple. Although the use of KL divergence or log-likelihood function is common [18–21], most of algorithms avoid direct minimizing it. We demonstrate that the direct minimization can be done (using just the standard Newton's method) when properly choosing the model-specific observables. Second, this method is non-perturbative in nature, in a sense that all possible interaction terms are treated equally. In particular, without introducing complexities to the code, the proposed method applies to non-pairwise interactions whose strengths may not be easy to obtain by analyzing the spin-spin correlation function. An example will be provided in the next section. Third, we note that the proposed method computes only the *model specific* expectation values ($\langle F_i \rangle$), not some model-independent correlation function. Therefore, the proposed method is very likely to give an answer even if the model is not correct. For this reason one should be careful about the obtained answer, especially if the model form is not known. Two estimators provided in the previous subsections provide a measure of model quality. However, if we are certain about the model, the proposed method is very efficient. We explicitly show that there is no numerical issues for a model of 10 parameters. Finally, we mention that our method works less efficiently when the state is ordered.
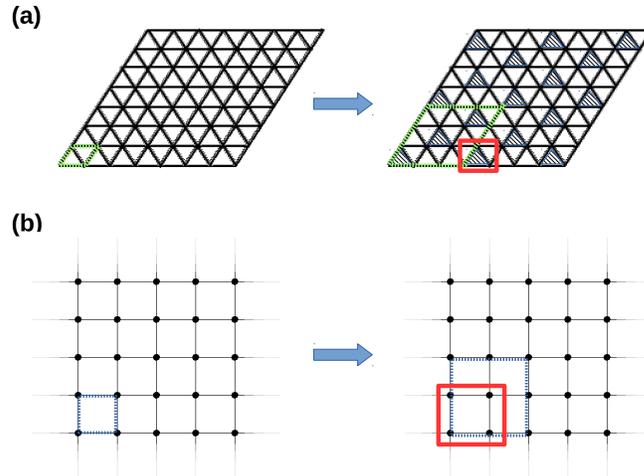
FIG. 3. (a) The triangular lattice. Three Ising spins in the red box in the right panel represent a block spin. (b) The square lattice. Four Ising spins in the red box in the right panel represent a block spin. In both cases, the dashed box in the left panel denotes the original unit cell; the dashed box in the right panel denotes the unit cell of block spins.

III. REAL-SPACE RENORMALIZATION GROUP ON TWO-DIMENSIONAL ISING MODEL

A. Problem formulation

As a first application, we apply the KL divergence to determine the real-space RG flow for 2D Ising model. The triangular and square lattices are considered. We note that the connection between the RG and the deep learning has been recently pointed out and exploited [31, 32]. For the real-space RG calculation, we follow the standard "block spin" procedure [33–35], where a original lattice, with lattice constant a, is coarse grained to the one with larger lattice constant la (so the number of unit cells is reduced by l^2). The spin in the original lattice is S_i, whereas the spin in the coarse grained one is S_I. This "block spin" procedure is illustrated in Fig. 3. For a given spin configuration in the triangular lattice [Fig. 3(a)], the value of block spin follows the simple "majority rule". Namely, for three original spins belonging to the block I, $S_I = 1(-1)$ if two or three original spins (in the same block I) have the value 1 (-1). For a given spin configuration in the square lattice [Fig. 3(b)], there are four original spins belonging to the block I. $S_I = 1(-1)$ if three or four original spins (in the same block I) have the value 1 (-1). If the sum of four original spins happens to be zero, then S_I has 50 % chance to be 1 or -1. The procedure to determine the RG flow is summarized as follows:

- Generate a set of configurations according to the original model $E(\{s_i\})(= -\sum_{ij} J_{ij}^{(0)} s_i s_j)$

- For each original configuration, determine the corresponding configuration of block spins, from which we get $P(\{s_I\})$.

- From $P(\{s_I\})$, minimize KL divergence to find the best $E(\{S_I\})(= -\sum_{IJ} J_{IJ}^{(1)} s_I s_J)$.

- Repeat the procedure, using $J_{IJ}^{(1)}$ in the original lattice.

Following these steps, we obtain how couplings change upon scaling:

$$J^{(0)} \to J^{(1)} \to J^{(2)} \to \dots$$

By initializing $J^{(0)}$ at different values, we can get the whole RG flow in the coupling space.
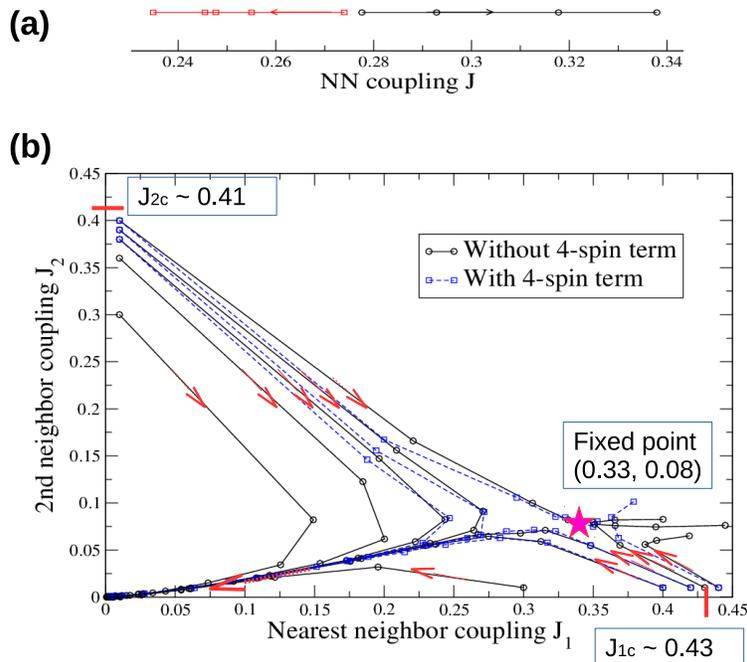
FIG. 4. (a) The RG flow for the Ising model on a triangular lattice, with nearest neighbor coupling only. The original square lattice is 18×18, and we use 2×10^5 runs in the Monte Carlo steps. The critical $J_{1c} \sim 0.277$, as compared to the exact result $J_c = 0.2746$. (b) The RG flow for the Ising model on a square lattice, with first and second neighbor couplings (black solid curves) and with an additional 4-spin interaction (blue dashed curves). The original square lattice is 30×30, and we use 6×10^5 runs in the Monte Carlo steps. The critical fixed point is seen to be around $(J_1, J_2) = (0.33, 0.08)$. The critical point of Ising model of nearest neighbor is $J_{1c} \sim 0.43$; that of second neighbor is $J_{2c} \sim 0.41$. The analytical result is $J_{1c} = J_{2c} \sim 0.44$.

B. Simulation results

Fig. 4 shows our main results. For the triangular lattice, the initial model only has NN coupling, and we only keep on the NN coupling in the subsequent RG procedure. By using an original lattice of size 18×18, we get the critical J of 0.277, which is remarkably close to the exact result $J_c = [\log 3]/4 = 0.2746$ [35]. For the square lattice, the initial model includes both NN (J_1) and 2nd NN (J_2) couplings, and we keep on the NN and 2nd NN coupling in the RG procedure. The RG flow is shown in Fig. 4(b). We have included a four-spin coupling interaction,

$$H_{4\text{-spin}} = -J_3 F_3(\{S\}) = -J_3 \sum_i S_i S_{i+\hat{x}} S_{i+\hat{y}} S_{i+\hat{x}+\hat{y}}. \tag{15}$$

which is usually neglected in the perturbation based calculation. We found that this additional term is indeed much weaker (at most 10 % of J_1 and J_2) than NN and 2NN couplings. As a concrete example, we give one RG trajectory of (J_1, J_2, J_3) round to three decimal places $(0.400, 0.010, 0.000) \rightarrow (0.317, 0.058, -0.005) \rightarrow (0.283, 0.063, -0.007) \rightarrow (0.244, 0.056, -0.006) \rightarrow (0.189, 0.043, -0.002) \rightarrow (0.124, 0.024, -0.001)$, whose projection on (J_1, J_2) plane is shown in Fig. 4(b). Our calculations explicitly show the inclusion of the four-spin term does not change the structure of the RG flow, and a quantitative estimate of its strength is given. We obtain a RG fixed point around $(J_1, J_2) = (0.33, 0.08)$, and the phase boundary for NN only model is around 0.43, and that for 2nd NN only model is around 0.41. These values are again very close to the exact results [34]. Although this method can obtain the phase boundary quite accurately, it fails to obtain a meaningful values for the critical exponents due to the numerical uncertainty around the fixed point. As shown in Fig. 4(a), the flow away from the critical point does not really follow the exponential pattern. Finally, we notice that our method does not give sensible results when $|J_i|$ is large enough to form an ordered state. This is how we cut the RG flow in Fig. 4(b). As the real-space RG is somehow "uncontrollable" (in a sense that some non-zero couplings are usually forced to be zero to make the calculation doable), our method provides a simple way to numerically estimate the amplitudes of these ignored interactions.

IV. DYNAMICAL SYSTEM OF SELF-PROPELLED PARTICLES

A. Problem formulation
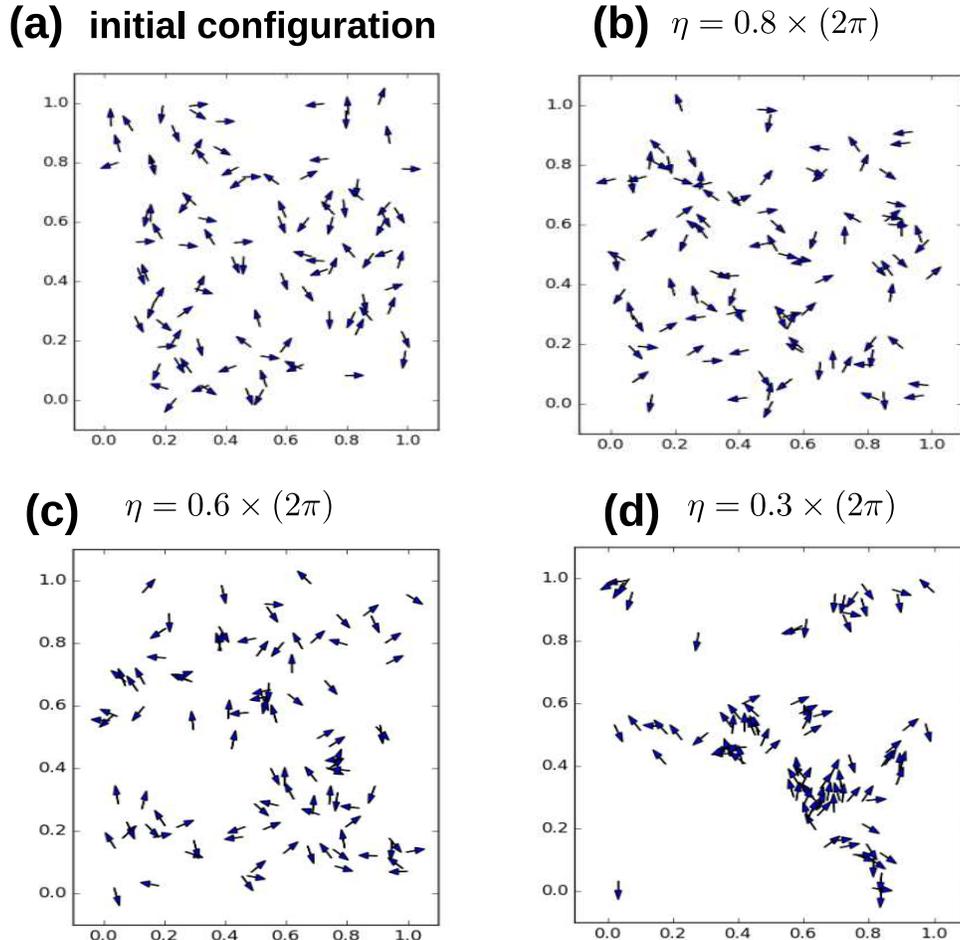
FIG. 5. Representative behaviors, corresponding to different noise strength η, of Viscek model. The arrow represents the moving direction of each particle. (a) A random initial configuration, where the position and velocity direction of each particle are randomly chosen. (b) $\eta = 0.8 \times (2\pi)$. (c) $\eta = 0.6 \times (2\pi)$. In both (b) and (c), the particles still move randomly with some spatial correlation. (d) $\eta = 0.3 \times (2\pi)$. A clear clustering behavior is seen.

Active matter is an emerging field that studies the ensemble behaviors of constituents that constantly consume energy and propel themselves [36]. It is particularly relevant in biological physics, since most cells or animals populations can be considered as an active matter [37, 38]. One of the best studied active matter system is the self-propelled particles which was first proposed by Vicsek et al. [23] (referred to as Vicsek model), which has been experimentally reconstructed [39], and theoretically studied extensively [40–42]. The dynamics of the system is described by a "molecular-dynamics" type of equation of motion with the updating rules:

$$
\begin{aligned}
\mathbf{x}_i(t+1) &= \mathbf{x}_i(t) + \mathbf{v}_i(t)\Delta t \\
\mathbf{v}_i(t+1) &= v_0(\cos\theta_i(t+1), \sin\theta_i(t+1)) \\
\theta_i(t+1) &= \langle\theta(t)\rangle_r + \Delta\theta
\end{aligned}
\tag{16}
$$

Here $\langle\theta(t)\rangle_r$ is the average direction of velocities of particles (including particle i) within a circle of "interaction" radius r surrounding the centered particle i. $\Delta\theta$, a noise term, is a random number uniformly distributed between

$[-\eta/2, \eta/2]$ (the maximum η is 2π, representing a totally random system), mimicking the fluctuations. One crucial aspect of SPP is that its distribution of velocity amplitude is different from that in thermal equilibrium, implying the system of active particles has a lower entropy than the system of normal "Brownian" particles [43]. Maintaining this state consumes energy. This aspect is captured in the Viscek model by assuming a constant velocity amplitude. In our simulation, we consider $N = 100$ particles confined in a 1×1 periodic square cell. We define an average inter-particle separation r_0 to be $N\pi r_0^2 = 1$, and will consider the interaction radius r to be $0.5r_0$, r_0, and $1.5r_0$. Following Ref. [23], we choose $v_0 = 0.001$. We note the cell size and the interaction radius by themselves are not important, it is their ratio that determines the dynamics. A few representative behaviors, corresponding to different noise strength η and $r = r_0$, are given in Fig. 5. We see that as the noise strength is decreased, the system tends to form clusters.

Bialek et al. [44] proposed that the systems of self-propelled particles can be described as a statistical model. In this description, one searches for the "best" energy functional such that the statistical distribution is as close as possible to the time variation obtained by the dynamical equations. To describe the Vicsek model, we use a generalized XY model

$$E(\{\theta_i\}) \equiv -JF_r(\{\theta_i\}) = \begin{cases} -\frac{J}{2}\sum_{i,j}\cos(\theta_i - \theta_j), & \text{if } |\mathbf{x}_i - \mathbf{x}_j| < r \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

In Eq. (17), $\mathbf{x}_i$ and θ_i are respectively the position and the direction of velocity of ith particle. Eq. (17) means that particles i and j have an XY-type of coupling $-J\cos(\theta_i - \theta_j)$ only when their separation is smaller than r. The distribution of a given velocity (or angle since the amplitude is fixed) configuration $P(\{\theta_i\})$ is proportional to $\exp[-E(\{\theta_i\})]$. The coupling J is positive so that the nearby particles tend to align their moving directions with one another.

Generally, the dynamical approach (Vicsek model) is of Langevin type, where the fluctuation is approximated by a random noise in the time domain, whereas the statistical approach is of Fokker-Planck type, where the fluctuation is translated to a probability distribution. Our goal now is to find the best model parameter J in Eq. (17) for the Viscek model of different noise strength η. Using the proposed method, all we need is to compute

$$\langle F_r \rangle_T \equiv \frac{1}{T}\int_0^T dt\, F_r(\{\theta_i(t)\}) \tag{18}$$

Ideally $\langle F_r \rangle_T$ and $\langle F_r^2 \rangle_T$ approach constants at large T, and is independent of initial configurations (ergodic). However this has to be numerically tested, and we will find that it is only approximately true. The uncertainty in $\langle F_r \rangle_T$ (as $T \to \infty$) leads to the uncertainty of fitted J.

B. Simulation results

To obtain the coupling J defined by Eq. (17) from the dynamical equations Eq. (16), we first compute $\langle F_r \rangle_T$ [Eq. (18)] for a given η in the long time limit. A representative behavior of $\langle F_r \rangle_T$ for $\eta = 0.7 \times (2\pi)$ is plotted in Fig. 6(a). The time evolution of four random initial conditions are plotted. We see that while it approaches to a constant in a long-time limit, this limiting value weakly depend on the initial conditions (about 13 %). The variation for $\langle F_r^2 \rangle_T$ in long time limit is larger at about 23 %. We also try a few "unusual" initial configurations, such as all initial positions are within a small region, or all original velocities are identical, and they all yield to a similar long-time behavior (but takes longer, not shown). We also find that the variation on initial conditions decreases when we increase the number of particles. For example, by keeping $\eta = 0.7 \times (2\pi)$, for $N = 300$, the variation of $\langle F_r \rangle_T$ due to initial conditions is about 5 % (between 18 and 20) and that of $\langle F_r^2 \rangle_T$ is about 10 % (between 90 and 100) (not shown).

The determined J as a function of η is given in Fig. 6 (b). The error bar originates from the variation of $\langle F_r \rangle_T$ due to initial conditions. In practice we compute $\langle F_r \rangle_T$ starting from ten random initial conditions to determine this uncertainty. As expected for all interaction radii, a larger η, corresponding to larger fluctuation in the Viscek model, leads to a smaller J, corresponding to a weaker inter-particle coupling in the statistical model. Our results also show that the coupling J is stronger for the smaller interaction radius. This behavior may seem counter-intuitive, as a larger interaction radius favors clustering. However, it can be understood from the velocity averaging procedure of the Viscek model – as the interaction radius increases, more particles contribute to the velocity average, effectively reducing the strength of the pairwise interactions. This perhaps subtle aspect is captured using our method. Our method does not converge when the noise strength $\eta < 0.6 \times (2\pi)$. As discussed in Section II, this originates from that the system being more ordered in the statistical model. From Fig. 6(b), the maximum J we get is around 0.5. It is interesting to contrast this value to the KT critical $J \sim 2/\pi \approx 0.64$ in the 2D XY model [26–28]. As our method does

(a) time average $\eta = 0.7 \times (2\pi)$ **(b) mapping**
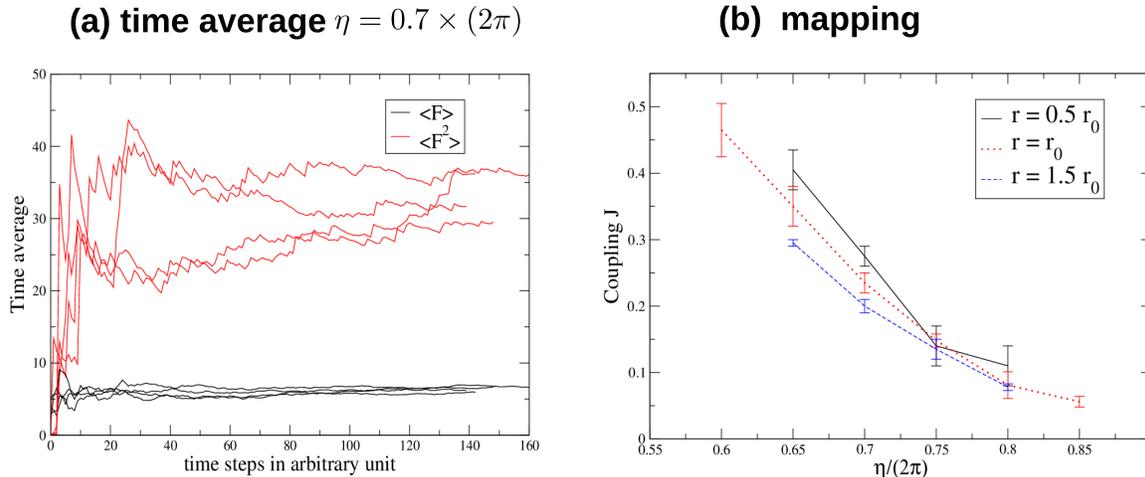
FIG. 6. (a) Time average of $\langle F_r \rangle_T$ and $\langle F_r^2 \rangle_T$ for $\eta = 0.7$ and $r = r_0$, using four different random initial conditions. It is seen that $\langle F_r \rangle_T$ at larger time has about 10% variation, whereas $\langle F_r^2 \rangle_T$ about 20% variation. (b) The fitted J for $r = 0.5r_0$, $r = r_0$, $r = 1.5r_0$ as a function of η . When $\eta < 0.6$, our method does not converge. For all η, the coupling J is stronger for the smaller interaction radius

not pose any numerical issues across the KT transition on a lattice [Fig. 1 (b)], the inability to converge at small η is more related to the tendency of clustering. To estimate of the model quality, we use $|\Delta|$ defined Eq. (14), and find the values are all smaller than 0.15 (mostly smaller than 0.05, especially for larger fitted J). This indicates that the generalized XY model provides a reasonable statistical description for the dynamics described by the Viscek model.

V. CONCLUSION

This work is devoted to solve one type of "inverse" problems: for a given set of configurations whose probability distribution is of Boltzmann type, i.e. $P(\{S\}) \propto \exp[-E(\{S\}; J_1, J_2, ...)]$, how do we determine the model parameters (J_i)? We propose to extract the model parameters by directly minimizing the Kullback-Leibler divergence. The most time consuming part of this procedure is to compute some expectation value with different model parameters, which can be done efficiently using Markov chain Monte Carlo algorithm. This method is non-perturbative in nature (treating all interactions equally) and is relatively easy to implement. We test this method for Ising and XY models on the 1D and 2D lattices. When the form of model is known, this method is very effective. One essential limitation is that this method does not work well when the system is in the ordered state. Two estimators to quantify the model quality are provided.

We then apply our method to two types of problems. First we apply it to the real-space renormalization group, and find the obtained RG flow sufficiently good for determining phase boundary (1% within the exact result), but not accurate enough for critical exponents. In addition, our method provides a simple way to quantitatively estimate amplitudes of the interactions truncated in the usual real-space RG procedure. Second we apply it to the dynamical system of self-propelled particles, where we extract the parameter of a statistical model (a generalized XY model) from a dynamical system described by the Viscek model. We are able to obtain the sensible coupling values as a function the noise strength of the Viscek model. In this respect, we provide an example on how to analyze the data, especially when we have some simple model(s) in mind. We regard this as a first step to quantify the behavior of self-propelled particles.

Acknowledgement

We thank the anonymous reviewer for suggesting the use of Levenberg-Marquardt algorithm.

---

[1] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer, 2006).
[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," Nature **521**, 436–444 (2015).
[3] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133–A1138 (1965).
[4] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864–B871 (1964).
[5] Alper Kinaci, Badri Narayanan, Fatih G. Sen, Michael J. Davis, Stephen K. Gray, Subramanian K. R. S. Sankaranarayanan, and Maria K. Y., "Unraveling the planar-globular transition in gold nanoclusters through evolutionary search," Scientific Reports (2016/11/28/online).
[6] Atsuto Seko, Tomoya Maekawa, Koji Tsuda, and Isao Tanaka, "Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids," Phys. Rev. B **89**, 054303 (2014).
[7] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sanchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alan Aspuru-Guzik, "The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid," The Journal of Physical Chemistry Letters **2**, 2241–2251 (2011).
[8] Roberto Olivares-Amaya, Carlos Amador-Bedolla, Johannes Hachmann, Sule Atahan-Evrenk, Roel S. Sanchez-Carrera, Leslie Vogt, and Alan Aspuru-Guzik, "Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics," Energy Environ. Sci. **4**, 4849–4861 (2011).
[9] Atsuto Seko, Atsushi Togo, Hiroyuki Hayashi, Koji Tsuda, Laurent Chaput, and Isao Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization," Phys. Rev. Lett. **115**, 205901 (2015).
[10] Kevin F. Garrity, "First-principles search for n-type oxide, nitride, and sulfide thermoelectrics," Phys. Rev. B **94**, 045122 (2016).
[11] John C. Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke, "Finding density functionals with machine learning," Phys. Rev. Lett. **108**, 253002 (2012).
[12] Prasanna V. Balachandran, James Theiler, James M. Rondinelli, and Turab Lookman, "Materials prediction via classification learning," Scientific Reports **5**, 13285 (2015/08/25/online).
[13] Louis-Fran çois Arsenault, Alejandro Lopez-Bezanilla, O. Anatole von Lilienfeld, and Andrew J. Millis, "Machine learning for many-body physics: The case of the anderson impurity model," Phys. Rev. B **90**, 155136 (2014).
[14] Antoine Georges, Gabriel Kotliar, Werner Krauth, and Marcelo J. Rozenberg, "Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions," Rev. Mod. Phys. **68**, 13–125 (1996).
[15] G. Kotliar and D. Vollhardt, Physics Today **57** (2004).
[16] L.-F. Arsenault, O. Anatole von Lilienfeld, and A. J. Millis, "Machine learning for many-body physics: efficient solution of dynamical mean-field theory," ArXiv e-prints (2015), arXiv:1506.08858 [cond-mat.str-el].
[17] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, "Machine learning quantum phases of matter beyond the fermion sign problem," ArXiv e-prints (2016), arXiv:1608.07848 [cond-mat.str-el].
[18] Jascha Sohl-Dickstein, Peter B. Battaglino, and Michael R. DeWeese, "New method for parameter estimation in probabilistic models: Minimum probability flow," Phys. Rev. Lett. **107**, 220601 (2011).
[19] Erik Aurell and Magnus Ekeberg, "Inverse ising inference using all the data," Phys. Rev. Lett. **108**, 090201 (2012).
[20] H. Chau Nguyen and Johannes Berg, "Mean-field theory for the inverse ising problem at low temperatures," Phys. Rev. Lett. **109**, 050602 (2012).
[21] Omri Har-Shemesh, Rick Quax, Borja Miñano, Alfons G. Hoekstra, and Peter M. A. Sloot, "Nonparametric estimation of fisher information from real data," Phys. Rev. E **93**, 023301 (2016).
[22] S. Kullback and R. A. Leibler, "On information and sufficiency," Ann. Math. Statist. **22**, 79–86 (1951).
[23] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet, "Novel type of phase transition in a system of self-driven particles," Phys. Rev. Lett. **75**, 1226–1229 (1995).
[24] E. T. Jaynes, "Information theory and statistical mechanics," Phys. Rev. **106**, 620–630 (1957).
[25] Benjamin B. Machta, Ricky Chachra, Mark K. Transtrum, and James P. Sethna, "Parameter space compression underlies emergent theories and predictive models," Science **342**, 604–607 (2013), http://science.sciencemag.org/content/342/6158/604.full.pdf.
[26] J M Kosterlitz and D J Thouless, "Ordering, metastability and phase transitions in two-dimensional systems," Journal of Physics C: Solid State Physics **6**, 1181 (1973).
[27] Jorge V. José, Leo P. Kadanoff, Scott Kirkpatrick, and David R. Nelson, "Renormalization, vortices, and symmetry-breaking perturbations in the two-dimensional planar model," Phys. Rev. B **16**, 1217–1241 (1977).
[28] Paul M. Chaikin and T.C. Lubensky, *Principles of condensed matter physics* (Cambridge University Press, 1995).
[29] Chih-kuan Tung, Chungwei Lin, Benedict Harvey, Alyssa G. Fiore, Florencia Ardon, Mingming Wu, and Susan S. Suarez, "Fluid viscoelasticity promotes collective swimming of sperm," Scientific Reports **7**, 3152 (2017).

[30] Harald Cramér, *Mathematical methods of statistics* (Princeton : Princeton University Press, 1999).

[31] C. Bény, "Deep learning and the renormalization group," ArXiv e-prints (2013), arXiv:1301.3124 [quant-ph].

[32] P. Mehta and D. J. Schwab, "An exact mapping between the Variational Renormalization Group and Deep Learning," ArXiv e-prints (2014), arXiv:1410.3831 [stat.ML].

[33] Leo Kadanoff, "Scaling laws for ising models near tc," Physics **2**, 263 (1966).

[34] R. K. Pathria, *Statistical Mechanics* (Butterworth-Heinemann, 1996).

[35] Nigel Goldenfeld, *Lectures on phase transition and the renoralization group* (Addison-Wesley, 1992).

[36] M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, Madan Rao, and R. Aditi Simha, "Hydrodynamics of soft active matter," Rev. Mod. Phys. **85**, 1143–1189 (2013).

[37] Andrea Cavagna, Alessio Cimarelli, Irene Giardina, Giorgio Parisi, Raffaele Santagati, Fabio Stefanini, and Massimiliano Viale, "Scale-free correlations in starling flocks," Proceedings of the National Academy of Sciences **107**, 11865–11870 (2010), http://www.pnas.org/content/107/26/11865.full.pdf.

[38] H. P. Zhang, Avraham Beer, E.-L. Florin, and Harry L. Swinney, "Collective motion and density fluctuations in bacterial colonies," Proceedings of the National Academy of Sciences **107**, 13626–13630 (2010), http://www.pnas.org/content/107/31/13626.full.pdf.

[39] Jeremie Palacci, Stefano Sacanna, Asher Preska Steinberg, David J. Pine, and Paul M. Chaikin, "Living crystals of light-activated colloidal surfers," Science **339**, 936–940 (2013), http://science.sciencemag.org/content/339/6122/936.full.pdf.

[40] Yaouen Fily and M. Cristina Marchetti, "Athermal phase separation of self-propelled particles with no alignment," Phys. Rev. Lett. **108**, 235702 (2012).

[41] S. C. Takatori, W. Yan, and J. F. Brady, "Swim pressure: Stress generation in active matter," Phys. Rev. Lett. **113**, 028103 (2014).

[42] Alexandre P. Solon, Hugues Chaté, and Julien Tailleur, "From phase to microphase separation in flocking models: The essential role of nonequilibrium fluctuations," Phys. Rev. Lett. **114**, 068101 (2015).

[43] Clemens Bechinger, Roberto Di Leonardo, Hartmut Löwen, Charles Reichhardt, Giorgio Volpe, and Giovanni Volpe, "Active particles in complex and crowded environments," Rev. Mod. Phys. **88**, 045006 (2016).

[44] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak, "Statistical mechanics for natural flocks of birds," Proceedings of the National Academy of Sciences **109**, 4786–4791 (2012), http://www.pnas.org/content/109/13/4786.full.pdf.