Kernel methods for interpretable machine learning of order parameters

Pedro Ponte and Roger G. Melko

Department of Physics and Astronomy, University of Waterloo, Ontario N2L 3G1, Canada
and Perimeter Institute of Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada

Machine learning is capable of discriminating phases of matter, and finding associated phase transitions, directly from large data sets of raw state configurations. In the context of condensed matter physics, most progress in the field of supervised learning has come from employing neural networks as classifiers. Although very powerful, such algorithms suffer from a lack of interpretability, which is usually desired in scientific applications in order to associate learned features with physical phenomena. In this paper, we explore support vector machines (SVMs), which are a class of supervised kernel methods that provide interpretable decision functions. We find that SVMs can learn the mathematical form of physical discriminators, such as order parameters and Hamiltonian constraints, for a set of two-dimensional spin models: the ferromagnetic Ising model, a conserved-order-parameter Ising model, and the Ising gauge theory. The ability of SVMs to provide interpretable classification highlights their potential for automating feature detection in both synthetic and experimental data sets for condensed matter and other many-body systems.

I. INTRODUCTION

Pattern recognition, the task of automatic discovery of features and regularities in data, is a major focus of machine learning, which now powers many different tools in our daily lives [1,2]. From a physics perspective, the problem of searching for patterns in experimental data has long driven theoretical progress, which, by definition, relates the patterns to underlying postulates of our physical theories. Hence, in the modern push to automate the discovery of important and novel physical features in data [3–29], physicists must be mindful of the interpretability of machine learning results if they are truly meant to drive theoretical process.

In condensed matter, physicists face the ultimate big-data challenge. One must search practical measurements, obtained from the exponentially large state space of a system or model, for patterns that relate to underlying theoretical paradigms. The recent success of neural networks in classifying phases of matter [4,6–9,18,29] has been encouraging, in that it demonstrates how relatively standard supervised learning tools can be repurposed for calculations in condensed matter physics. However, contrary to industry applications of machine learning, where performance is the prime metric of success, in physics it is generally desirable to further tie the outcome to some theoretical structure, which can eventually be used, e.g., to make predictions. Neural network behavior can indeed be interpreted on the simplest models of statistical mechanics, such as the demonstration in Ref. [4] that the magnetization order parameter of the two-dimensional (2D) Ising model is learned by the weights of the hidden units. However, in general for more complicated models, similar success in relating network structure to nonlinear or nonlocal order parameters is challenging, especially in the case of deep neural networks. This lack of interpretability presents a challenge for the goal of driving theoretical progress with machine learning. It is therefore crucial that the condensed matter community survey the performance of interpretable machine learning algorithms on data obtained from models of interest to condensed matter physics.

In this paper we study interpretable supervised learning algorithms applied to the discrimination of phases of matter in large synthetic data sets produced by numerical simulation. We focus on a particular class of machine learning algorithms called support vector machines (SVMs). In particular, we ask whether SVMs in the supervised learning setting are able to discover the mathematical structure of physical order parameters. We first introduce the SVM algorithm and describe its general properties, with particular focus on the kernel trick. This trick allows us to perform linear regression on nonlinear features of the data, without explicitly generating the features. In particular, this is relevant when applied to Monte Carlo data from many-body systems, which typically have high-dimensional physical states. Without the kernel trick, it would be unpractical to generate a set of nonlinear features, which could be exponentially large. In Sec. III, we perform phase classification on Monte Carlo configurations produced from the 2D Ising model, a conserved-order-parameter (COP) Ising model [5], and the 2D Ising gauge theory. We find that SVMs with a quadratic polynomial kernel can discriminate the phases of the 2D Ising model by learning the correct physical order parameter (the squared magnetization per spin). In the case of the COP model, we find that SVMs are able to discover the nontrivial order parameter, devised by Wang [5] through visualization of dimensionally reduced data, with less human intervention. Finally, we show that the 2D Ising gauge theory, whose $T = 0$ ground state is defined by local plaquette constraints but no conventional order parameter, can be discriminated by SVMs with polynomial kernels. In this case, no polynomial with order less than 4 can identify the ground state, indicating that the SVM is able to learn the original Hamiltonian in order to evaluate whether the local constraints on four Ising degrees of freedom are satisfied for each plaquette.

II. SUPPORT VECTOR MACHINES

Linear support vector machines [2] were initially designed to perform binary classification. They belong to the class of

supervised learning algorithms that, given an input $\boldsymbol{x} \in \mathbb{R}^p$, predict the class $y \in \{-1,1\}$ in which it is most likely to belong. In the context of this paper, the input $\boldsymbol{x}$ will represent a spin configuration from a statistical mechanical model of interest in condensed matter physics, in which the different components of $\boldsymbol{x}$, called features in the machine learning literature, correspond to the spins at different lattice sites. The output y will label the thermodynamic phase it was sampled from. The main idea behind SVMs is to find the hyperplane, defined by $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, that best separates the two classes. Formally, this is achieved by solving the optimization problem:

$$\underset{\boldsymbol{w}, \xi_i, b}{\arg \min} \left\{ \frac{1}{2} \boldsymbol{w} \cdot \boldsymbol{w} + \frac{1}{\gamma} \sum_{i=1}^{N} \xi_i \right\}$$

such that $y^{(i)}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b) \geqslant 1 - \xi_i, \ i = 1 \ldots N.$ (1)

Here, γ is a constant, ξ are slack variables (described more below) and N is the number of input samples, called the training set. Finding optimal parameters $\boldsymbol{w}$ and b, which can then be used to make predictions on a test set, is the goal of supervised learning with SVMs. Unlike, say, a feed-forward neural network [4], this model does not provide probabilistic predictions. However, the optimization problem Eq. (1) is equivalent to $\arg \min_{\boldsymbol{w}, b} \{ \frac{1}{2} \boldsymbol{w} \cdot \boldsymbol{w} + \frac{1}{\gamma} \sum_{i=1}^{N} \max [0, 1 - y_i d(\boldsymbol{x}^{(i)})] \}$ with $d(\boldsymbol{x}^{(i)}) = \boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b$. Because the hinge function $\max(0, 1 - yd)$ approximates the misclassification error $\Theta(-yd)$ [1], this can be viewed as an approximation to minimizing the misclassification error with a so-called ℓ_2-norm regularization (which penalizes unnecessary coefficients of $\boldsymbol{w}$). This analogy also explicitly shows that γ can be interpreted as a regularization parameter.

When the ξ_i variables are constrained to be zero in Eq. (1), this corresponds to the hard margin case and the optimization algorithm has a solution only for linearly separable classes. In this case, the solution corresponds to the hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ for which the margin, defined as the minimum distance $d_{\min} = 1/||\boldsymbol{w}||$ of the data samples to the hyperplane, is maximum. Thus, this algorithm finds the hyperplane, which maximizes the training set margin and provides the most confident predictions on new inputs. On the other hand, perfectly separable data is not a typical property of data sets and in general the constraints $y^{(i)}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b) \geqslant 1$ are not feasible. Hence, it is necessary to introduce the slack variables ξ, which allow a training set input $\boldsymbol{x}^{(i)}$ to violate the margin at a cost of ξ_i/γ. For a given training sample, the slack variable ξ_i can take different values. $\xi_i = 0$ if the margin is not violated; $0 < \xi_i \leqslant 1$ if it is on the correct side of the hyperplane but violates the margin; and $\xi_i > 1$ if the sample is misclassified. In general it is necessary to use a test set to find the optimal parameter γ. This will provide the best tradeoff between minimizing training errors and the model complexity.

After optimization (i.e., training), the class to which a new input $\boldsymbol{x}$ belongs is predicted as $y = \text{sgn}(d(\boldsymbol{x}))$, hence $d(\boldsymbol{x})$ is referred to as the decision function in the machine learning literature. Even though the previous optimization problem can be solved by quadratic programming [2], it also admits a dual formulation whereby the primal variables $\boldsymbol{w}, \xi, b$ are eliminated and the optimization is performed over N dual variables α_i, which are the Lagrange multipliers associated

with each constraint in Eq. (1). The optimal parameter $\boldsymbol{w}$ is then expressed as $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}^{(i)}$. A crucial feature of this dual formulation is that the optimization algorithm only depends on inner products of the training samples $\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \rangle$. In addition, at prediction time, one only needs to calculate the inner product between the training samples and new samples. Because the algorithm is formulated such that the input vector enters only in the form of a scalar product, this allows us to employ the kernel trick, whereby we replace $\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$ with some other choice of kernel function, $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. Then, at prediction time the decision function has the form

$$d(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}^{(i)}, \boldsymbol{x}) + b, \quad (2)$$

whereby we can learn more complex decision functions depending on the choice of the kernel without explicitly generating more features in our input $\boldsymbol{x}$. For example, the kernel $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = (\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)} + c_0)^d$ corresponds to the mapping to a $\binom{p+d}{d}$ dimensional feature space corresponding to all the monomials of the form $x_{i_1} x_{i_2} \ldots x_{i_k}$ (ignoring permutations) that are up to order d where p is the number of raw features. Importantly, we note that the choice of a kernel (e.g., a polynomial) that results in a highly interpretable decision function also limits the expressive power of the SVM, especially when compared with universal approximator algorithms such as neural networks.

III. RESULTS

In this section, we perform supervised learning with SVMs on data sets generated by sampling spin configurations of classical Hamiltonians, where Ising degrees of freedom $\boldsymbol{\sigma}$ (spins taking binary values) will serve as our input $\boldsymbol{x}$. Finite-size lattices with N spins are considered. In the following, we use the most efficient training algorithm for SVMs by means of the SCIKIT-LEARN library [30]—the sequential minimal optimization algorithm [31]—known to scale as $O(N^2)$, or with a smaller power, for several kernels and types of data. We find, in practice, training on a single core is generally slow for 10^5 samples or more. We explore the behavior of different polynomial kernels and perform grid search to find the optimal regularization parameter γ. In general, the results are averaged over several choices of training and test sets for the same hyperparameters in order to obtain more reliable statistics.

A. 2D Ising model

We first consider the nearest-neighbor Ising model in two dimensions, $H = - \sum_{\langle ab \rangle} \sigma_a \sigma_b$, where $\sigma_a = \pm 1$, and $\boldsymbol{a}$ are the Euclidean coordinates of a given lattice site. Monte Carlo simulations using the Wolff algorithm were performed to collect spin configurations $\boldsymbol{\sigma}^{(i)} = (\ldots, \sigma_a^{(i)}, \ldots)$ where i identifies each configuration collected at different temperatures from $T = 1.6$ to $T = 2.9$. The 2D Ising phase transition occurs at the critical temperature $T_c \approx 2.269$ [32] and separates a ferromagnetic (FM) phase, characterized by a nonzero total magnetization per spin, from a featureless paramagnetic (PM) phase at high temperatures. For a given $L \times L = N$ size lattice,
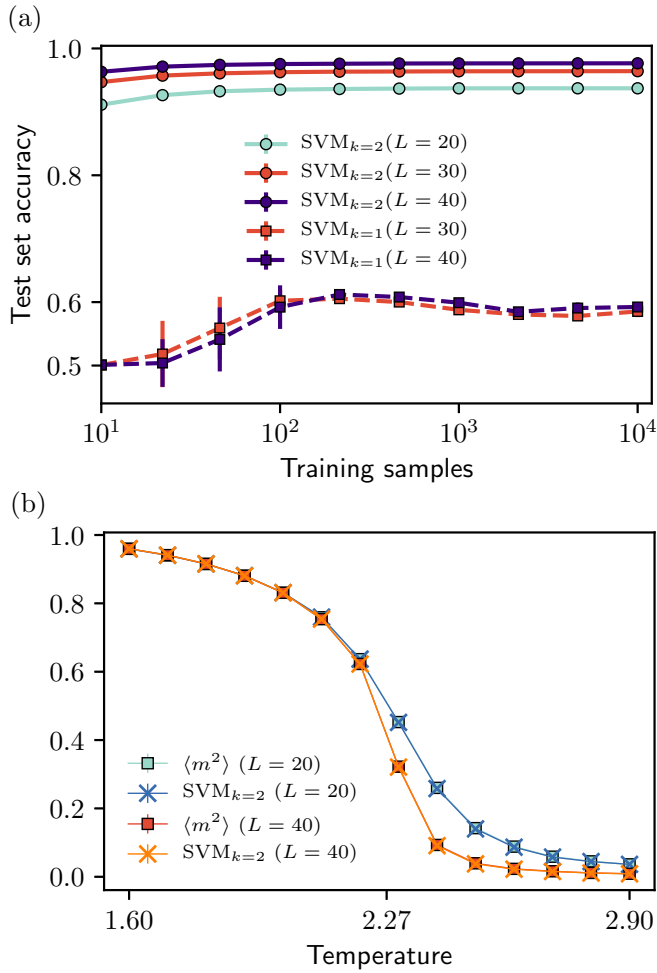
FIG. 1. (a) Average test set accuracy of support vector machines with polynomial kernel $K(\sigma,\sigma') = (\sigma \cdot \sigma')^k$ trained on Monte Carlo sampled configurations from the 2D Ising model. For each number of training samples, the accuracy is averaged over 100 independent training and test sets. (b) The SVM classifies samples according to $\text{sgn}(d(\sigma))$. The decision function $d(\sigma)$ for the SVM with a quadratic polynomial kernel is evaluated by Monte Carlo sampling at different temperatures and compared to the squared magnetization per spin m^2. The arbitrary scale factor and offset in the SVM decision function are fixed by matching the decision function to $\langle m^2 \rangle$ at $T = 1.6$ and $T = 2.9$.

each sample is labeled with its corresponding phase in the binary class $y_i = \pm 1$. We train SVMs to learn to discriminate between the two phases for different numbers of samples in the training set, with a corresponding change in the number of samples at each temperature. For the Ising model, we limit our survey to a linear and a quadratic kernel of general form $K(\sigma,\sigma') = (\sigma \cdot \sigma' + c_0)^k$ with $k = 1,2$ and $c_0 = 0$. Note that in general, it might be necessary for the learning procedure to find the optimal c_0 as well.

In order to quantify the performance of each SVM model, the main metric that we study is the test set accuracy as a function of the number of samples in the training set, for the value of γ, which results in the optimal accuracy of the model, illustrated in Fig. 1(a). For the linear kernel, the exploration of γ is over a log-spaced grid of 11 values from 10^{-5} to 10^5. For

the quadratic kernel the accuracy does not depend significantly on the choice of γ and thus we fixed it at $\gamma = 10^{-5}$. For a given number of training samples and regularization γ, the test set accuracy is additionally averaged over different random selections of training and test sets.

Results for the test set accuracy and for the SVM decision function are shown in Fig. 1. As seen in Fig. 1(a), the quadratic kernel performs extremely well with mean test set accuracy $\approx 97\%$ for $L = 40$. This can be easily interpreted, since we know that this model possesses a quadratic order parameter that linearly discriminates the FM from the PM, i.e., the squared magnetization per spin $m^2 = (\sum_a \sigma_a / N)^2$. We find that the quadratic kernel reaches very significant performance with only a few dozen samples in the training set, which is a result of the simplicity of this model. Moreover, with increasing number of samples the test set accuracy approaches a plateau value, which increases with system size toward 100%. This is the expected behavior, since at the critical point the fluctuations of the order parameter approach zero in the thermodynamic limit and it is thus possible to discriminate perfectly between both phases.

For the linear kernel [Fig. 1(a)], the accuracy shows nonmonotonic behavior with the total number of training set samples and does not improve with increasing system size. This is a consequence of the fact that a linear decision function is unable to discriminate between the FM and PM phases. Namely, in the FM phase configurations have magnetization per spin near ± 1, while for the PM phase most configurations have approximately zero magnetization. Thus, the $k = 1$ kernel is asking a linear decision boundary to separate a data set with three clusters—an impossible task. Close inspection of the decision function learned by the SVM reveals it contains random linear coefficients without any structure, confirming that nothing physically relevant is being learned about the data in this case.

As noted above, the accuracy of the SVM with a quadratic kernel on the test set does not depend significantly on the regularization parameter γ. An advantage of SVMs is that we can visualize the decision function being learned. From Eq. (2), the decision function for an SVM with quadratic polynomial kernel can be expressed as

$$d(\sigma) = \sum_a \sum_x C_x^{(a)} \sigma_a \sigma_{a+x} + b. \qquad (3)$$

In Fig. 2, we display the heat map of $C_x = \langle C_x^{(a)} \rangle_a$, where $\langle \ldots \rangle_a$ denotes averaging with respect to all sites a for $\gamma = 10^6$ and $\gamma = 10^{-6}$ and system size $L = 30$. It is interesting to note that even though the classification performance is very similar, the SVM decision function corresponds to different order parameters depending on the amount of regularization. Clearly, at $\gamma = 10^6$, the SVM is learning m^2 as the order parameter of the model up to finite-size effects. In contrast, at $\gamma = 10^{-6}$, the SVM is learning to calculate the square of the total magnetization within some fixed distance of each spin and summing all these different local contributions. A more detailed Fourier decomposition of $C_x^{(a)}$ shows that the $k = 0$ mode is dominant for any amount of regularization, but at small γ the contributions from other small k modes are larger.
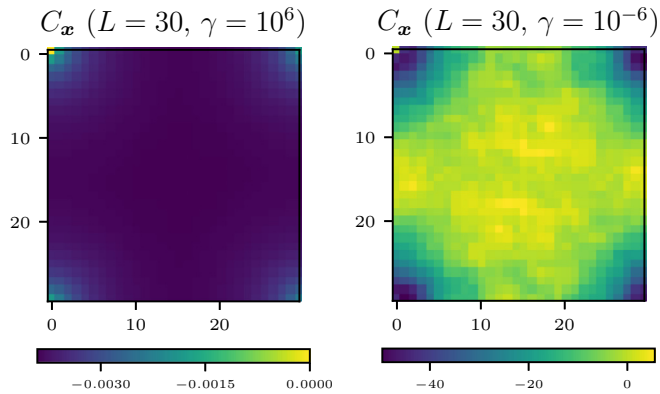
FIG. 2. The decision function for an SVM with quadratic polynomial kernel, Eq. (3). The averaged $C_x = \langle C_x^{(a)} \rangle_a$ is displayed for different values of regularization $\gamma = 10^6$ and $\gamma = 10^{-6}$. Clearly, for large regularization ($\gamma = 10^6$), the decision function is essentially the m^2 order parameter of the 2D Ising model.

In Fig. 1(b), the SVM decision function (for large regularization $\gamma = 10^6$) is averaged over Monte Carlo samples at different temperatures showing essentially perfect agreement with m^2. Of course, the SVM decision function has an arbitrary scale and offset and in order to match with $\langle m^2 \rangle$ a linear transformation is performed, so that they agree at the extreme values of temperature $T = 1.6$ and 2.9.

B. Conserved-order-parameter Ising model

The conserved-order-parameter Ising model [5] is described by the same Hamiltonian as the Ising model but the configuration space is restricted to the subspace where the total magnetization $\sum_a \sigma_a$ is zero. This model describes a half-filled lattice gas of particles with a nearest-neighbor attractive interaction [33]. At low temperatures, domains of up and down spins are separated by either two horizontal or two vertical domain walls for a square lattice geometry with periodic boundary conditions. There is a phase transition to a featureless phase at the same critical temperature as the 2D Ising model. Reference [5] studied this model on the square lattice using principal component analysis (PCA), a dimensional reduction algorithm, and devised the following order parameter:

$$S = \frac{1}{L^4} \sum_{a,b} \sigma_a \sigma_b \left[\cos\left(\frac{2\pi}{L}(a_1 - b_1) \right) + \cos\left(\frac{2\pi}{L}(a_2 - b_2) \right) \right]. \tag{4}$$

S is of the form of Eq. (3) with

$$C_x^{(a)} = \frac{1}{L^4} \left[\cos\left(\frac{2\pi}{L} x_1 \right) + \cos\left(\frac{2\pi}{L} x_2 \right) \right]. \tag{5}$$

In contrast to PCA, which through feature dimension reduction provides the easy visualization required for the determination of this order parameter, SVMs should provide automated order parameter detection in a more systematic way. We explore this ability now.
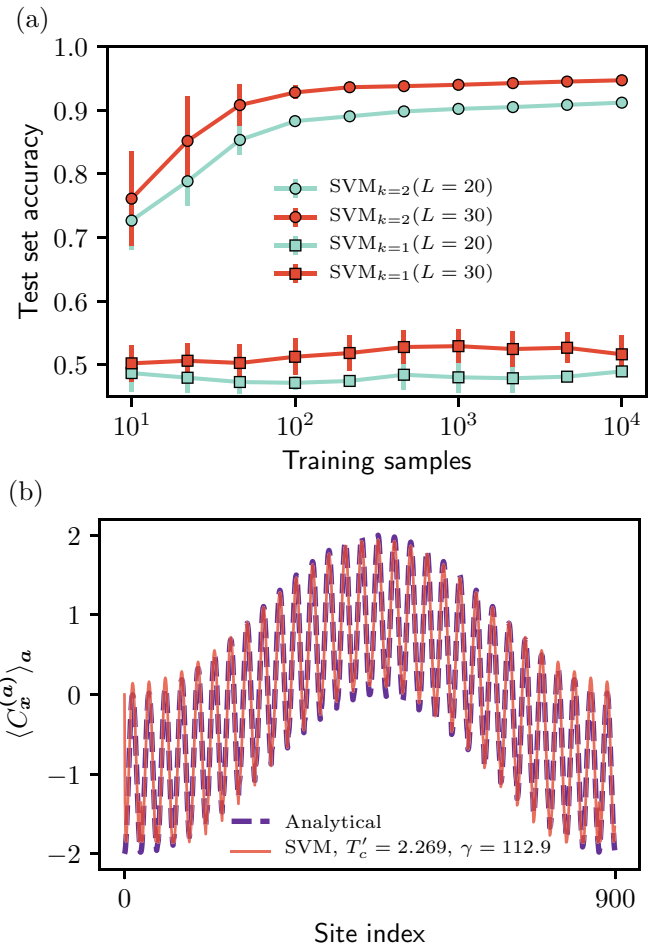

FIG. 3. (a) Average test set accuracy vs number of training samples for support vector machines with polynomial kernel $K(\sigma, \sigma') = (\sigma \cdot \sigma')^k$ trained on Monte Carlo sampled configurations from the conserved-order-parameter Ising model at different temperatures. (b) The spatial dependence of the SVM decision function coefficients $\langle C_x^{(a)} \rangle_a$ learned by an SVM with quadratic kernel and regularization coefficient $\gamma = 112.9$ for system size $L = 30$ shows very good agreement with the analytical form (5) devised in Ref. [5].

As in the case with the Ising model, we collect Monte Carlo samples for a set of temperatures below and above T_c. Figure 3(a) shows the averaged test set accuracy as a function of the number of samples for the optimal value of the regularization parameter γ over a grid of 11 values from 10^{-5} to 10^5. The results are averaged over 400 training and test sets. As in the Ising model case, for a quadratic kernel there is a monotonic improvement of test accuracy with the number of training samples. The limiting value, as the number of training samples becomes very large, also increases with system size. This behavior signals the existence of a quadratic order parameter that discriminates between the two phases. As a check, we compare the explicit spatial dependence $C_x = \langle C_x^{(a)} \rangle_a$ of the SVM decision function for $\gamma = 112.9$ (the selection of this regularization value is discussed below) to the analytical form (5). The scale and offset of the SVM decision function are fixed by requiring the coefficients C_x to have the same mean and standard deviation as the analytical form.
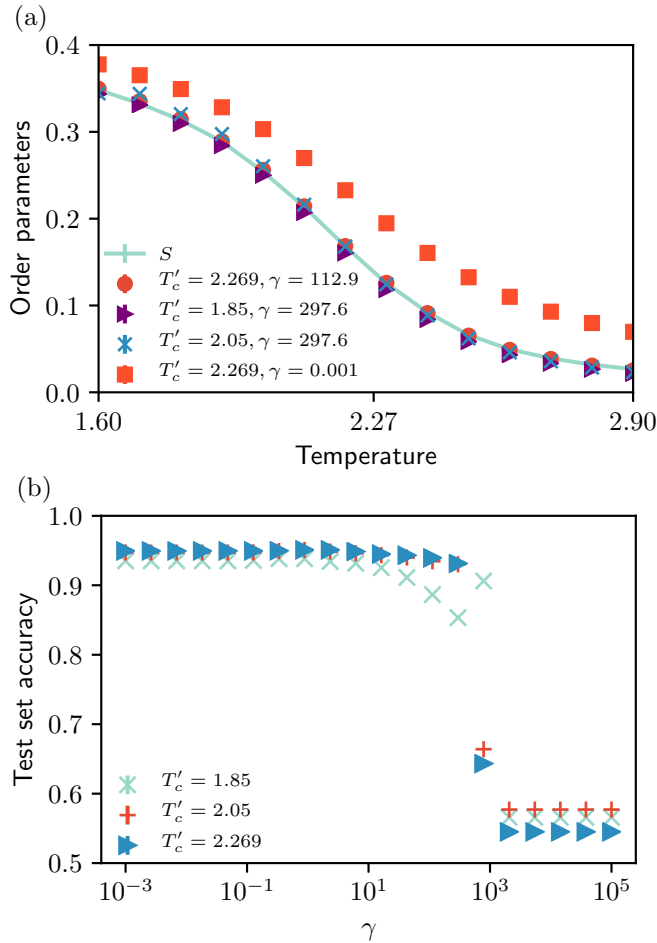
(a)

(b)

FIG. 4. (a) Comparing the decision function averaged over Monte Carlo samples at different temperatures learned by SVMs with quadratic kernel, trained on 30000 samples with $L = 30$, to the analytical order parameter S [Eq. (4)] assuming different critical temperatures T'_c and amount of regularization γ. (b) Dependence on regularization γ of the test set accuracy (averaged over ten sets of 10000 samples) of quadratic SVMs trained on 30000 samples for $L = 30$.

Figure 3(b) shows very good agreement between them. In contrast, the linear kernel does not show a clear improvement of the test set accuracy with increasing number of samples, which indicates there is not an order parameter of that form.

One could ask whether this order parameter could also be learned in the case where the precise value of the critical temperature T'_c is not known. We address this by performing supervised learning with quadratic SVMs assuming different values T'_c for the critical temperature. Finite-size scaling of the learned order parameter (or its moments) can then be used to estimate the value of the critical temperature. Figure 4(a) compares the actual order parameter S [Eq. (4)] with the SVM decision function for different values of $T'_c = 1.85, 2.05, 2.269$. The scale and offset of the SVM decision function are fixed by matching it with S at $T = 0.1$ and $T = 100.0$. We observe that the SVM decision functions learned when assuming $T'_c = 1.85, 2.05$ (for a choice of regularization values discussed below) also agree well with S, suggesting machine learning algorithms are able to learn

important physical information without the precise knowledge of T_c.

We now analyze the role of the regularization parameter γ for learning in this model. Figure 4(b), shows for $L = 30$ and a training set with 30 000 samples, the test set accuracy averaged over ten randomly picked test sets of 10 000 samples versus the regularization parameter γ for $T'_c = 1.85, 2.05, 2.269$. At $\gamma \lesssim 10^3$, the test set accuracy jumps significantly to very high values ($\approx 90\%$) and then reaches a plateau. Interestingly, we find that throughout the plateau region the decision function of the SVM changes continuously. Figure 4(a) also compares the analytical order parameter S with the SVM decision function for different values of γ when $T'_c = T_c$. For $\gamma = 112.9$, right at the beginning of the plateau, there is highly accurate agreement between S and the SVM decision function, while for $\gamma = 0.001$ the two do not match. Similar results were also found for the other values of T'_c. Thus, the physical order parameter can be interpreted as being associated with the least amount of complexity that still allows the model to have good predictive performance. This was also what we found for the Ising model in Sec. III A.

C. 2D Ising gauge theory

Finally, we consider the challenging case of topologically ordered systems, where no conventional local parameter exists. For concreteness we study the 2D Ising gauge theory with Hamiltonian $H = -\sum_p \prod_{i \in p} \sigma_i^z$, where the spins live on the bonds of a 2D square lattice and p represents a plaquette with four spins. The set of ground states is a degenerate manifold with the constraint that for all plaquettes p, $\prod_{i \in p} \sigma_i^z = 1$. In the thermodynamic limit, the constraints are violated at any finite temperature. A conventional order parameter that distinguishes ground states from finite temperature states does not exist. Reference [4] found that the simplest fully connected feed-forward neural networks were unable to classify these two different cases in a supervised learning context. Only with convolutional neural networks, which explicitly exploit locality and translational invariance, did this classification task succeed on raw spin configurations. We note that by engineering predictive features as a preprocessing step, simple feed-forward neural networks may be successful in classification (as demonstrated in Ref. [18]). However, in the present case, we explore the behavior of SVMs for the raw 2D Ising gauge theory data with no preprocessing on the input data.

We perform supervised learning on spin configurations generated for the Ising gauge theory at $T = 0$ and $T = \infty$ (i.e., completely random spin states). We explore kernels of increasing degree, starting from $k = 1$, for different system sizes L. Figure 5(a) shows that for system size $L = 5$, SVMs with polynomial kernel of degree less than 4 fail to discriminate these two phases, exhibiting average test set accuracies $\sim 50\%$, which amounts to random guessing. However, a SVM with polynomial kernel of fourth order is able to perform the task with an accuracy that converges towards 100% as the number of training samples is increased. As apparent in Fig. 5(b), we also note that the number of training samples necessary to learn the fourth-order discriminator increases with system size.
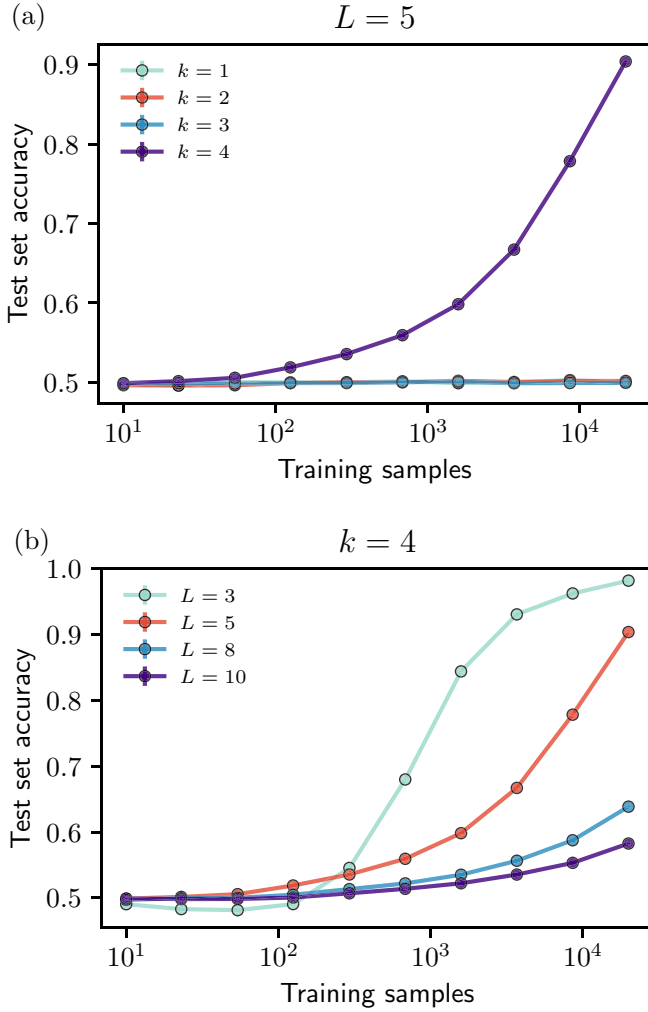
(a)

(b)

FIG. 5. (a) Test set accuracy of SVMs with polynomial kernels of order k in classifying ground states versus infinite temperature states of the 2D Ising gauge theory for system size $L = 5$. Only in the feature space of fourth-order polynomials is the SVM algorithm performance better than a random algorithm. The performance approaches 100% with increasing number of training samples. (b) Test set accuracy for $k = 4$ and different system sizes. For larger system sizes, more training samples are needed to learn the correct decision function.

This illustrates the difficulty that SVMs can have in performing classification for large system sizes.

In order to interpret these results, we further analyze the decision function of the SVM. The smallest degree (fourth-order) polynomial kernel learns a decision function of the form $d(\sigma) = \sum_{abcd} C_{abcd}\sigma_a\sigma_b\sigma_c\sigma_d + b$, which contains the product of four spins. Figure 6 shows the histogram of the coefficients C_{abcd} for an SVM trained on 20 000 samples and $\gamma = 10^{-6}$. From this plot it is clear that there are two sets of coefficients; one near zero, and another set with large negative values. Counting the number of these large (in magnitude) coefficients reveals 600. For this lattice, with 25 square plaquettes, this number corresponds to the number of possible permutations of the four indices of $abcd$, i.e., $25 \times 4!$. Hence, this model provides evidence that SVMs are able to learn complex interpretable decision functions and discover
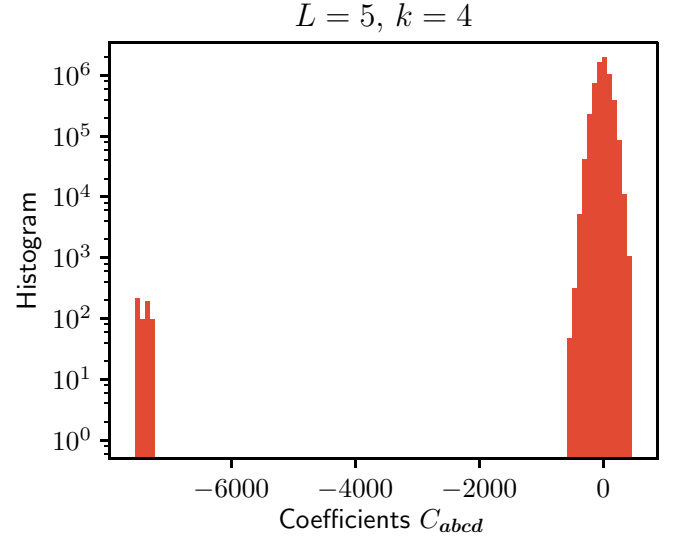

FIG. 6. Histogram of the coefficients C_{abcd} in the decision function learned by a SVM with fourth-order polynomial kernel in classifying ground states versus infinite temperature states of the 2D Ising gauge theory for $L = 5$. There are 600 large coefficients (in absolute value), which correspond to the 25 plaquettes and the permutations of their spins.

the locality of the Hamiltonian directly from raw data on spin configurations.

IV. CONCLUSIONS

We have examined the use of support vector machines (SVMs), one of the most common tools for supervised learning, for the binary classification of phases in several models of interest to condensed matter and statistical physics. SVMs employ a kernel trick to define a decision function, used to discriminate features in a higher-dimensional space. The kernel depends on the inner product of spin vectors, but can otherwise have some freedom in definition. We have shown that in contrast to other methods such as neural networks, the ability to use different kernels for classification tasks gives SVMs significant value in finding interpretable physical discriminators for different models, such as conventional order parameters defined in condensed matter theory.

To allow for some slack to misclassify data, SVMs employ a regularization, which controls the tradeoff between minimizing training errors, while still allowing sufficient model complexity. In this paper we have found that the amount of regularization has a strong impact on the decision function learned. We find that the expected physical order parameter is associated with the least complex model (i.e., the largest amount of regularization) that is still able to obtain near-optimal test set accuracy.

On the other hand, we find that small values of regularization that achieve the same (or better) performance than the physical order parameter learn a nonphysical decision function—one not related to any conventional order parameter. Since such decision functions generalize well to uncorrelated test set samples, we argue that this is not an example of overfitting, even though it may arise due to particular details

contained within the training set. This observation deserves further study, as it may have consequences more generally for the role of regularization in black box algorithms, such as neural networks, when applied to data in condensed matter physics.

While very successful on the Ising-like examples studied in this paper, the SVM algorithm is not without its limitations. As observed for the 2D Ising gauge theory, the number of samples required to learn the physically relevant decision function can grow prohibitively large. The reason that this occurs for the degenerate ground state of the Ising gauge theory, and not the models with conventional order parameters, requires further study. We generally observe that the training of SVMs when the number of samples is larger than 10^5 becomes time consuming. This could also be an issue in using SVMs near phase transitions in models where a large number of samples is necessary to learn the physical discriminator. In such cases, it would be interesting to further explore the interpretability of neural networks, since they can possibly present better scalability to larger system sizes.

It would be interesting to study the utility of SVMs on other classical models of interest in condensed matter physics. We note that for systems with continuous degrees of freedom it might be necessary to consider a radial basis function as a kernel, which maps to an infinite dimensional space, instead of polynomials. For example, it would interesting to see for the 2D XY model, if the spin stiffness can be identified as the physical discriminator for the Kosterlitz-Thouless transition using a suitably modified SVM. Finally, SVMs and other sparse kernel machines may in the future be easily adapted to study quantum phases and phase transitions by using wave functions or density matrices as data. SVMs with linear or quadratic polynomial kernels could be used to determine operators, which discriminate between different quantum phases using density matrices or wave functions as data. Thus, as the condensed matter community increasingly adopts modern machine learning methods into its numerical arsenal, we expect that SVMs will become a standard tool for finding interpretable physical discriminators for generic Hamiltonians in the near future.

[1] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, Secaucus, 2006).

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, 2012).

[3] L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U.-N. Niranjan, P. Duncan, M. Rupp, K.-R. Müller, and K. Burke, Int. J. Quantum Chem. **116**, 819 (2016).

[4] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).

[5] L. Wang, Phys. Rev. B **94**, 195105 (2016).

[6] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[7] G. Torlai and R. G. Melko, Phys. Rev. B **94**, 165134 (2016).

[8] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Sci. Rep. **7**, 8823 (2017).

[9] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Phys. Rev. X **7**, 031038 (2017).

[10] A. Tanaka and A. Tomiya, J. Phys. Soc. Jpn. **86**, 063001 (2017).

[11] T. Ohtsuki and T. Ohtsuki, J. Phys. Soc. Jpn. **85**, 123706 (2016).

[12] L. Li, T. E. Baker, S. R. White, and K. Burke, Phys. Rev. B **94**, 245129 (2016).

[13] D.-L. Deng, X. Li, and S. D. Sarma, arXiv:1609.09060.

[14] L. Huang and L. Wang, Phys. Rev. B **95**, 035105 (2017).

[15] J. Liu, Y. Qi, Z. Y. Meng, and L. Fu, Phys. Rev. B **95**, 041101 (2017).

[16] G. Torlai and R. G. Melko, Phys. Rev. Lett. **119**, 030501 (2017).

[17] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Nat. Phys. **13**, 435 (2017).

[18] Y. Zhang and E.-A. Kim, Phys. Rev. Lett. **118**, 216401 (2017).

[19] N. Portman and I. Tamblyn, J. Comput. Phys. **350**, 871 (2017).

[20] D. Crawford, A. Levit, N. Ghadermarzy, J. S. Oberoi, and P. Ronagh, arXiv:1612.05695.

[21] K.-I. Aoki and T. Kobayashi, Mod. Phys. Lett. B **30**, 1650401 (2016).

[22] D.-L. Deng, X. Li, and S. D. Sarma, Phys. Rev. X **7**, 021021 (2017).

[23] X. Gao and L.-M. Duan, Nat. Commun. **8**, 662 (2017).

[24] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, arXiv:1701.04831.

[25] Y. Huang and J. E. Moore, arXiv:1701.06246.

[26] S. J. Wetzel, Phys. Rev. E **96**, 022140 (2017).

[27] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, arXiv:1703.05334.

[28] W. Hu, R. R. P. Singh, and R. T. Scalettar, Phys. Rev. E **95**, 062122 (2017).

[29] F. Schindler, N. Regnault, and T. Neupert, Phys. Rev. B **95**, 245134 (2017).

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[31] J. Platt, in *Advances in Kernel Methods - Support Vector Learning* (MIT Press, Cambridge, 1998).

[32] L. Onsager, Phys. Rev. **65**, 117 (1944).

[33] M. Newman and G. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, Oxford, 1999).