

# Mapping general inference to physics

J. Shepard Bryan IV and Pressé Lab

Department of Physics, Arizona State University

July 9, 2024

## Abstract

Here, we attempt to create a general inference algorithm by constructing a model in such a way that model parameters evolve with dynamics similar to physical systems when the model becomes maximally predictive. By constructing models in this fashion, we gain a deeper understanding of how they work and how they can become more efficient. A key takeaway from this exercise is that mapping inference to physics seems to imply that a fully connected recurrent neural network can be approximated to second order with a number of parameters that scale linearly with the network size, a significant reduction from the expected quadratic scaling.

**DISCLAIMER: This is a work in progress. Please excuse any typos or errors. Please also excuse the lack of citations.**

## 1 Introduction

It is well known that there is a fundamental connection between physics and learning. For example, recording information requires energy [Landauer, Bechhoeffer, 1], the act of measurement unavoidably alters the measured object [quantumtextbook, 1], and physical objects are known to be capable of learning [1]. Moreover, the whole field of statistical mechanics is based on the concept of “incomplete information” [1], which requires the existence of an observer capable of learning [maxwellsdemon, 1]. Recently, there has been interest in exploiting the connection between physics and learning to better understand deep learning [1] and explore new models [hopfield, 1]. In this work, we expand upon this work by creating a direct mapping from general inference to physics.

The heart of what we would like to explore is as follows: Imagine actively measuring some unlabeled variables over a long series of time steps, with the goal of predicting the value of each variable at the next time step. For example, the measurements could be readouts from an experiment, values of nodes in a neural network, or states (on or off) of neurons in a brain at a time step. How would one best simplify and make sense of this data? What general inference algorithm would allow one to best predict future measurements, regardless of the form of the object being measured?

Here, we create a mapping from this general inference problem to physics. In this paradigm, physical quantities such as mass, charge, space, and spin are defined as statistical quantities of measured variables, and natural constants such as the vacuum permittivity emerge as averaged error. Additionally, machine learning concepts such as gradient descent optimizers are replaced with physics inspired dynamics laws. We show that this paradigm can stably minimize prediction error. Moreover, our mapping has potential advantages over traditional physics based inference models [hopfield, 1] because it allows for long range interactions between variables while minimizing the number of model parameters.

## 2 Methods

Here we lay out the mathematical framework for our work mapping general inference to physics. We start by presenting a graphical model for the general inference problem, starting with the most general case of fully connected binary random variables, then making approximations to simplify the network. We then derive the loss function for a prediction model on the network, decomposing the loss term into two parts: one governing the continuous variables, and another governing the binary variables. We show that when the model is optimized (has minimal error) this loss function effectively acts as a conserved quantity. We then show that careful choices of model parameters lead to model parameter dynamics, i.e. the way in which model parameters are updated after each new measurement, which are identical to electrodynamics and quantum mechanics.

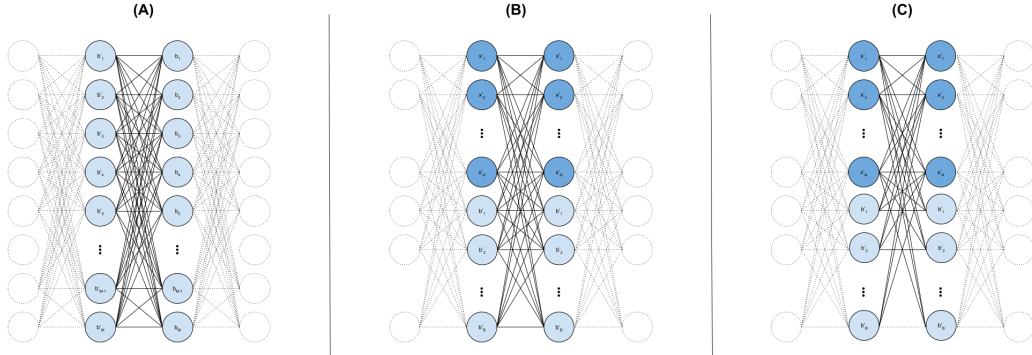


Figure 1: Here, we present a graphical model for the general inference problem under three different cases. (A) A fully connected network of binary random variables. (B) A fully connected mixture network of both binary and continuous variables. (C) A network of fully connected continuous variables, but sparsely connected binary variables.

## 2.1 Graphical model

Our first goal is to rigorously define the problem. Loosely speaking, we want to predict future measurements of a system given a stream of past measurements. Let us start by defining the variables we are measuring.

Some of the variables, we will call “external variables”, in our system will directly correspond to measurements of external observables. For example, think of a sensory neuron in a human body that fires in response to heat [trp, 1]. The firing state of this neuron is directly linked to the temperature of the external environment. Conversely, some of the variables in our system will be “internal variables”. For example, think of an interneuron, which fires in response to other neurons in the body, but not in response to external stimuli [interneuron, 1]. These internal variables are necessary for predicting the external variables, thus they are included in our system. We will make no distinction between internal and external variables in our system.

Let us now assume that all the variables in the system are binary random variables. Let there be  $M$  binary variables, labelled  $b_1, b_2, \dots, b_M$ , and let  $\mathbf{b}'_{1:M}$  be the grouping of all variables. At first, it may seem that by considering only binary random variables, we limit the systems that we can look at, however, when we remember that all measurements will have finite precision we can safely assume that all measurements can be converted into sets of binary random variables. For example, a variable measuring temperature to 32 bit precision, can be decomposed into 32 separate binary variables.

Lastly, regarding the binary variables, we will use a convention where the possible outcomes of the binary variables are  $-1$  and  $1$ , like those we experience in quantum spin mechanics, instead of  $0$  and  $1$  like we use in traditional probability theory. This choice does not lose generality, as we can always convert back to the probability convention via  $b \rightarrow (b + 1)/2$ .

Next, let us introduce memory into our system. Let the prime symbol ( ' ) indicate past measurement. So dividing the measurement process into discrete time steps,  $b_n$  represents the measurement of the  $n$ th variable at some time level and  $b'_n$  represents the measurement of the same variable at the previous time step. We will only consider two time steps at a time, with the hope that the states of internal variables will be able to retain information for longer periods.

We now need to define the network of relationships between variables. In the most general case, a fully connected network, every variable  $\mathbf{b}_{1:M}$  is conditioned on the variables at the previous time step,  $\mathbf{b}'_{1:M}$ . Figure 1 panel A shows a graphical representation of this network. Notice that every variable at the first time step,  $b'_i$ , is connected to every variable at the subsequent time step,  $b_j$ . Notice that this fully connected pattern continues in the future and past, represented by greyed out nodes.

We can write out the probability of the future states,  $\mathbf{b}_{1:M}$ , given the past states,  $\mathbf{b}'_{1:M}$  according to a Boltzmann distribution

$$\mathcal{P}(b_i | \mathbf{b}'_{1:M}) = \frac{1}{Z_i} \exp(-\beta U(b_i, \mathbf{b}'_{1:M})) \quad (1)$$

$$Z_i = \exp(-\beta U(+1, \mathbf{b}'_{1:M})) + \exp(-\beta U(-1, \mathbf{b}'_{1:M})) \quad (2)$$

where  $U$  is some energy of the state. Taylor expanding the energy to first order, we get

$$U(b_i, \mathbf{b}'_{1:M}) \approx b_i(J_{0i} + \sum_j J_{ij}b'_j) \quad (3)$$

$$\mathcal{P}(b_i | \mathbf{b}'_{1:M}) \approx \frac{1}{Z_i} \exp\left(-\beta b_i(J_{0i} + \sum_j J_{ij}b'_j)\right) \quad (4)$$

where  $\beta$  represents an inverse temperature,  $J_{0i}$  represents some base energy for the  $i$ th variable,  $J_{ij}$  represents the interaction energies.

Note that the connection strengths between nodes may change over time, however, we will assume that if the connection strengths change over time, they do so slowly enough that to some certainty, the connection strength between  $b'_i$  and  $b_j$  will be approximately equal to the connection strength between  $b''_i$  and  $b'_j$ .

Equation (4) is effectively a stochastic fully connected binary neural network [1], which has  $M^2 + M$  connections. When  $M$  is large, it is prohibitively costly to model this system, thus our next goal is to simplify the graphical model. We can simplify this graphical model by clustering variables with large coefficients of correlation (see supplementary section S0.1). In other words, we can replace large clusters of correlated our binary variables in our system with variables that only encode the sum of the cluster. For example, if  $\mathbf{b}_{1:K}$  are highly correlated, we may replace them with

$$x = \sum_{k=1}^K b_k. \quad (5)$$

This cluster variable,  $x$ , will be distributed around some mean, with variance related to the error of our clustering process times the number of components in the cluster, which we will discuss in later sections. Note that when the number of binary variables that make up the cluster is large, then we can approximate cluster variables as continuous variables. Moving forward, we will use this approximation and treat  $x$  as continuously distributed.

Let us proceed by replacing every highly correlated subset of  $\mathbf{b}_{1:M}$  with such cluster variables. For concreteness, let us say that there are  $N$  such clusters, labelled  $x_1, x_2, \dots, x_N$ , and  $B$  remaining binary variables, which are not highly correlated with any of the subsets. Let us use similar notation allowing  $\mathbf{x}_{1:N}$  to represent the grouping of the cluster variables and let  $\mathbf{x}'_{1:N}$  represent the cluster variables at the previous time step.

After the substitution process laid out above, our new graphical model will look like the one presented in figure 1 panel B. Here we have  $N$  cluster variables and  $B$  binary variables. The network is fully connected, with  $(N + B)^2 + (N + B)$  parameters. Since  $M > N + B$  this new system has exponentially fewer parameters.

We can simplify this graphical model further by considering that highly correlated binary variables have already been clustered. By definition, this means that the remaining correlations between variables are weak. Thus, we can ignore connections between binary variables. However, we will keep the connections between clusters, since weak correlations may be balanced out by the large variance of each  $x_n$ . This final graphical model is shown in figure 1 panel C, where we see that the clusters are fully connected, but each binary variable,  $b_i$  is connected only to the clusters.

We are now ready to construct a model to predict the system. In the next section, we construct a loss function for the system presented in figure 1 panel C.

## 2.2 Loss function

In the previous section, we built out a graphical model representing a system of fully connected binary variables as a set of clusters and sparsely connected binary variables. In this section, we introduce the concept of a model, which is used to predict the values of the variables in the system. We then construct a loss function for the model.

Consider a single time step. Let us define the ground truth distribution of variables as

$$\mathcal{P}^*(\mathbf{x}_{1:N}, \mathbf{b}_{1:B})$$

where the star indicates that this is the “real” distribution. Next, we define a model that predicts the values of the system variables,

$$\mathcal{P}(\mathbf{x}_{1:N}, \mathbf{b}_{1:B} | \Theta) \mathcal{P}(\Theta)$$

which has two parts: 1) a likelihood,  $\mathcal{P}(\mathbf{x}_{1:N}, \mathbf{b}_{1:B}|\Theta)$ , which gives the probability of the measurements according to some model parameters,  $\Theta$ , and 2) a prior,  $\mathcal{P}(\Theta)$ ,

$$\mathcal{P}(\mathbf{x}_{1:N}, \mathbf{b}_{1:B}|\Theta)$$

and a prior over model parameters which acts as a regularizer over model parameters. The hope is that the model will assign probabilities to measurements that are as close as possible to the ground truth distribution. We can quantify the error of the model using the Kullback-Leibler divergence [1]

$$\text{KLD} = - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log \left( \frac{\mathcal{P}(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}|\Theta) \mathcal{P}(\Theta)}{\mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B})} \right) \quad (6)$$

$$= - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) (\log(\mathcal{P}(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}|\Theta) \mathcal{P}(\Theta)) - \log(\mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}))) \quad (7)$$

$$= - \mathcal{H}_0 - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}|\Theta) \mathcal{P}(\Theta)) \quad (8)$$

$$\mathcal{H}_0 + \text{KLD} = - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}|\Theta) \mathcal{P}(\Theta)) \quad (9)$$

where  $\mathcal{H}_0$  is the intrinsic entropy of the system. Equation (9) states that the average prediction error (defined as the log probability of the model, which is often called the ‘‘surprise’’ [1]) of our model is equal to the intrinsic entropy of the system,  $\mathcal{H}_0$ , plus the error (KL Divergence) between our model and the ground truth. Defining loss as

$$\mathcal{L} = \mathcal{H}_0 + \text{KLD} \quad (10)$$

$$= - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}|\Theta) \mathcal{P}(\Theta)) \quad (11)$$

we see that minimizing the loss minimizes the prediction error.

One key difference between our loss function, equation (11), and traditional loss terms defined elsewhere [1] is that we include the prior term as an inseparable part of the model. In later sections we will show that prior will act as both a regularizer and a momenta term during training, which may be advantageous over other paradigms in which loss functions, regularizers, and optimizers are defined separately.

We can simplify this loss function further, decomposing it into a loss from the clusters and a loss from the binary variables,

$$\mathcal{L} = - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}|\Theta) \mathcal{P}(\Theta)) \quad (12)$$

$$= - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(\mathbf{b}_{1:B}|\mathbf{x}_{1:M}, \Theta) \mathcal{P}(\mathbf{x}_{1:M}|\Theta) \mathcal{P}(\Theta)) \quad (13)$$

$$= - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) (\log(\mathcal{P}(\mathbf{b}_{1:B}|\mathbf{x}_{1:M}, \Theta)) + \log(\mathcal{P}(\mathbf{x}_{1:M}|\Theta)) + \log(\mathcal{P}(\Theta_x))) \quad (14)$$

$$= - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(\mathbf{b}_{1:B}|\mathbf{x}_{1:M}, \Theta)) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \log(\mathcal{P}(\mathbf{x}_{1:M}|\Theta)) - \log(\mathcal{P}(\Theta)) \quad (15)$$

$$= - \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log \left( \prod_{i=1}^B \mathcal{P}(b_i|\mathbf{x}_{1:M}, \Theta) \right) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \log(\mathcal{P}(\mathbf{x}_{1:M}|\Theta)) - \log(\mathcal{P}(\Theta)) \quad (16)$$

$$= - \sum_{i=1}^B \int d\mathbf{x}_{1:N} \sum_{\mathbf{b}_{1:B}} \mathcal{P}^*(\mathbf{x}_{1:M}, \mathbf{b}_{1:B}) \log(\mathcal{P}(b_i|\mathbf{x}_{1:M}, \Theta)) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \log(\mathcal{P}(\mathbf{x}_{1:M}|\Theta)) - \log(\mathcal{P}(\Theta)) \quad (17)$$

$$= - \sum_{i=1}^B \int d\mathbf{x}_{1:N} \sum_{b_i} \mathcal{P}^*(b_i, \mathbf{x}_{1:M}) \log(\mathcal{P}(b_i|\mathbf{x}_{1:M}, \Theta)) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \log(\mathcal{P}(\mathbf{x}_{1:M}|\Theta)) - \log(\mathcal{P}(\Theta)) \quad (18)$$

Where we have used the fact that binary variables are independent in the model to separate the binary variable terms. Finally by splitting  $\Theta$  into parameters which affect only individual binary variables,  $\Theta_{b_i}$ , and parameters which only affect the clusters,  $\Theta_x$ , we get

$$\mathcal{L} = \mathcal{L}_x + \sum_{i=1}^B \mathcal{L}_{b_i} \quad (19)$$

$$\mathcal{L}_x = -\log(\mathcal{P}(\Theta_x)) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \log(\mathcal{P}(\mathbf{x}_{1:M}|\Theta_x)) \quad (20)$$

$$\mathcal{L}_{b_i} = -\log(\mathcal{P}(\Theta_{b_i})) - \sum_{b_i} \int d\mathbf{x}_{1:N} \mathcal{P}^*(b_i, \mathbf{x}_{1:M}) \log(\mathcal{P}(b_i|\mathbf{x}_{1:M}, \Theta_{b_i})) \quad (21)$$

where we have successfully split the loss into two parts, one governing binary variables and one governing cluster variables.

Here we have defined probability distributions over the system variables and our model then constructed a loss function governing prediction error. We have not yet defined the architecture of the model, nor what the parameters of the model represent. In the next sections, we will choose a model architecture that is explicitly designed to minimize this loss. We will start with the loss over cluster variables, then move onto binary variables.

### 2.3 Model over cluster variables

Our goal now is to start designing the architecture of our model. In this section, we will design the model architecture for the distribution over cluster variables, with the goal of minimizing the error defined in equation (20). Because a Gaussian distribution can be thought of as a second order approximation to arbitrary distributions in log space [1], we will use it as the functional form for our model distribution

$$\mathcal{P}(\mathbf{x}_{1:M}|\Theta_x) \approx \mathcal{N}(\mathbf{x}_{1:M}; \mathbf{0}, \mathbf{K}) \quad (22)$$

$$= ((2\pi)^M |\mathbf{K}|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M}\right) \quad (23)$$

where  $\mathbf{K}$  is the covariance of the distribution. Notice that for simplicity we have set the mean of the distribution to zero, which can be interpreted as implicitly shifting the measurements by  $\mathbf{x}_{1:M} \rightarrow \mathbf{x}_{1:M} - \boldsymbol{\mu}_{1:M}$  where  $\boldsymbol{\mu}_{1:M}$  is the mean of the raw model distribution. We will eventually marginalize over measurements, meaning that this shift will have no effect on the final result, but simplifies the derivation.

Plugging in equation (23) into the loss function, equation (20), we get

$$\mathcal{L}_x = -\log(\mathcal{P}(\Theta_x)) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \log\left(\left((2\pi)^M |\mathbf{K}| \right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M}\right)\right) \quad (24)$$

$$= -\log(\mathcal{P}(\Theta_x)) - \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \left(-\frac{1}{2} \log\left((2\pi)^M |\mathbf{K}| \right) - \frac{1}{2} \mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M}\right) \quad (25)$$

$$= -\log(\mathcal{P}(\Theta_x)) + \frac{M}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{K}|) + \frac{1}{2} \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M}. \quad (26)$$

Let us now define exactly what the model parameters,  $\Theta_x$ , are and how they relate to the covariance matrix,  $\mathbf{K}$ . A covariance matrix has two parts: 1) the variance of each variable along the diagonal, and 2) the covariance between variables on the off diagonals [1]. Because our goal is to map inference to physics, we will assign physics inspired labels to each part of the covariance.

Remember that we already know from section 2.1 that each cluster,  $x_i$ , is the sum of binary variables, meaning that each  $x_i$  will be binomial distributed, which has variance proportional to the number of components in the cluster. Therefore, we expect that the variance of  $x_i$  will be proportional to the size of the cluster. Let us define,

$$\text{Var}[x_i] = m_i \quad (27)$$

which we will call the ‘‘mass’’ of the cluster, which has the dual purpose of quantifying both the variance of the variable and the number of binary variables that make up the cluster.

Moving onto the covariance, we notice that covariance encodes two things: the magnitude of correlation and the sign of the correlation. Taking inspiration from physics, we will treat the sign of the covariance as a product of

charges, and the strength of the correlation as an inverse distance

$$\text{Cov}[x_i, x_j] = \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \quad (28)$$

where  $q_i$  and  $q_j$  are the charges of variables and  $\vec{r}_i$  and  $\vec{r}_j$  are the positions of the variables. Through thought experiments like “if A is correlated with B and B is correlated with C then A is correlated with C” and “if A is anti-correlated with B and B is anti-correlated with C then A is correlated with C” we can conclude that charges can be used to keep track of the sign of correlation. As for positions, inverse distance makes sense because typically we expect objects that are close together to be more correlated than objects that are far apart, an assumption already used in many regression methods [1]. Let us start by assuming that the positions of each variable span a space with  $D$  dimensions where  $D$  is large enough that we can effectively tune charges and positions of each variable to match any arbitrary set of covariances. We will later show that we can constrain the number of dimensions to 3 when we define a covariance between cluster variables and binary variables.

All together, plugging our definitions, equations 27 and 28, into the covariance matrix,  $\mathbf{K}$ , we get

$$\mathbf{K} = \begin{bmatrix} m_1 & \frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|} & \frac{q_1 q_3}{|\vec{r}_1 - \vec{r}_3|} & \cdots \\ \frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|} & m_2 & \frac{q_2 q_3}{|\vec{r}_2 - \vec{r}_3|} & \cdots \\ \frac{q_1 q_3}{|\vec{r}_1 - \vec{r}_3|} & \frac{q_2 q_3}{|\vec{r}_2 - \vec{r}_3|} & m_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (29)$$

where the pattern continues along the whole  $N$  by  $N$  matrix. The inverse covariance matrix,  $\mathbf{K}^{-1}$ , can be calculated using a Taylor expansion where we separate the covariance into diagonal and off-diagonal components

$$\mathbf{K} = \mathbf{K}_{\text{diag}} + \mathbf{K}_{\text{off}} \quad (30)$$

$$\mathbf{K}^{-1} = (\mathbf{K}_{\text{diag}} + \mathbf{K}_{\text{off}})^{-1} \quad (31)$$

$$\approx \mathbf{K}_{\text{diag}}^{-1} - \mathbf{K}_{\text{diag}}^{-1} \mathbf{K}_{\text{off}} \mathbf{K}_{\text{diag}}^{-1} \quad (32)$$

$$= \begin{bmatrix} \frac{1}{m_1} & \frac{-q_1 q_2}{m_1 m_2 |\vec{r}_1 - \vec{r}_2|} & \frac{-q_1 q_3}{m_1 m_3 |\vec{r}_1 - \vec{r}_3|} & \cdots \\ \frac{-q_1 q_2}{m_1 m_2 |\vec{r}_1 - \vec{r}_2|} & \frac{1}{m_2} & \frac{-q_2 q_3}{m_2 m_3 |\vec{r}_2 - \vec{r}_3|} & \cdots \\ \frac{-q_1 q_3}{m_1 m_3 |\vec{r}_1 - \vec{r}_3|} & \frac{-q_2 q_3}{m_2 m_3 |\vec{r}_2 - \vec{r}_3|} & \frac{1}{m_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (33)$$

where we have used the fact that magnitude of covariance between clusters is defined to be small in our approximation, in other words, we drop all  $\frac{1}{r^2}$  terms. We can use our inverse covariance matrix to simplify the expression  $\mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M}$  which shows up in the loss function, equation (26),

$$\int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M} = \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \sum_i \sum_j x_i x_j (\mathbf{K}^{-1})_{ij} \quad (34)$$

$$= \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \sum_i \left( x_i^2 (\mathbf{K}^{-1})_{ii} + \sum_{j \neq i} x_i x_j (\mathbf{K}^{-1})_{ij} \right) \quad (35)$$

$$= \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \sum_i \left( \frac{x_i^2}{m_i} - \sum_{j \neq i} \frac{x_i x_j}{m_i m_j} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \right) \quad (36)$$

$$= \sum_i \left( \frac{\mathbb{E}[x_i^2]}{m_i} - \sum_{j \neq i} \frac{\mathbb{E}[x_i x_j]}{m_i m_j} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \right) \quad (37)$$

where  $\mathbb{E}[\dots]$  is the expectation value taken over the ground truth distribution. Finally, we can calculate the

determinant by again separating the covariance matrix into diagonal and off diagonal components,

$$|\mathbf{K}| = |\mathbf{K}_{\text{diag}} + \mathbf{K}_{\text{off}}| \quad (38)$$

$$= |\mathbf{K}_{\text{diag}}| |\mathbf{I} + \mathbf{K}_{\text{diag}}^{-1} \mathbf{K}_{\text{off}}| \quad (39)$$

$$= |\mathbf{K}_{\text{diag}}| \exp(\text{Trace}(\log(\mathbf{I} + \mathbf{K}_{\text{diag}}^{-1} \mathbf{K}_{\text{off}}))) \quad (40)$$

$$\approx |\mathbf{K}_{\text{diag}}| \exp(\text{Trace}(\mathbf{K}_{\text{diag}}^{-1} \mathbf{K}_{\text{off}})) \quad (41)$$

$$= |\mathbf{K}_{\text{diag}}| \exp(0) \quad (42)$$

$$= |\mathbf{K}_{\text{diag}}| \quad (43)$$

$$= \prod_{i=1}^N m_i \quad (44)$$

where we have used the identity  $\log(\mathbf{I} + \mathbf{A}) \approx \mathbf{A}$ .

It is now clear that the parameters of our model,  $\Theta_x$ , are the positions of the variables. Let  $\vec{\mathbf{r}}_{1:M}$  be the grouping of positions defined in equation (28) and let  $\mathbf{r}_{1:M}^{(d)}$  be the grouping of all positions along the  $d$ th dimension. We must choose a prior over these positions. Let us choose

$$\mathcal{P}(\Theta_x) = \mathcal{P}(\vec{\mathbf{r}}_{1:M}) \quad (45)$$

$$= \prod_{d=1}^D \mathcal{N}(\mathbf{r}_{1:M}^{(d)}; \mathbf{r}'_{1:M}{}^{(d)}, \mathbf{K}^{-1}) \quad (46)$$

where we break up the prior into  $D$  separate multivariate Gaussians over positions along each dimension with mean equal to the position at the previous time step,  $\mathbf{r}'_{1:M}$  and covariance equal to the inverse covariance of the likelihood. Let us now simplify the negative log prior which shows up in the loss function, equation (26),

$$-\log(\mathcal{P}(\Theta_x)) = -\log\left(\prod_{d=1}^D \mathcal{N}(\mathbf{r}_{1:M}^{(d)}; \mathbf{r}'_{1:M}{}^{(d)}, \mathbf{K}^{-1})\right) \quad (47)$$

$$= -\sum_{d=1}^D \log\left((2\pi)^M |\mathbf{K}^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{r}_{1:M}^{(d)} - \mathbf{r}'_{1:M}{}^{(d)})^T \mathbf{K} (\mathbf{r}_{1:M}^{(d)} - \mathbf{r}'_{1:M}{}^{(d)})\right)\right) \quad (48)$$

$$= \frac{D}{2} \log((2\pi)^M |\mathbf{K}|^{-1}) + \frac{1}{2} \sum_{d=1}^D \mathbf{v}_{1:M}^{(d)T} \mathbf{K} \mathbf{v}_{1:M}^{(d)} \quad (49)$$

$$= \frac{DM}{2} \log(2\pi) - \frac{D}{2} \log(|\mathbf{K}|) + \frac{1}{2} \sum_{d=1}^D \mathbf{v}_{1:M}^{(d)T} \mathbf{K} \mathbf{v}_{1:M}^{(d)} \quad (50)$$

where we have defined a ‘‘velocity’’,  $\vec{\mathbf{v}}_{1:M} = \vec{\mathbf{r}}_{1:M} - \vec{\mathbf{r}}'_{1:M}$  and have used the identity  $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ . We can

simplify the term on the far right by plugging in our covariance

$$\frac{1}{2} \sum_{d=1}^D \mathbf{v}_{1:M}^{(d)T} \mathbf{K} \mathbf{v}_{1:M}^{(d)} = \frac{1}{2} \sum_{d=1}^D \left( \sum_i \sum_j v_i^{(d)} v_j^{(d)} (\mathbf{K})_{ij} \right) \quad (51)$$

$$= \frac{1}{2} \sum_{d=1}^D \left( \sum_i \left( m_i v_i^{(d)2} + \sum_{j \neq i} v_i^{(d)} v_j^{(d)} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \right) \right) \quad (52)$$

$$= \frac{1}{2} \sum_i \left( m_i v_i^2 + \sum_{j \neq i} \vec{v}_i \cdot \vec{v}_j \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \right) \quad (53)$$

$$= \sum_i \frac{1}{2m_i} \left( (m_i v_i)^2 + q_i m_i \vec{v}_i \cdot \sum_{j \neq i} \frac{q_j \vec{v}_j}{|\vec{r}_i - \vec{r}_j|} \right) \quad (54)$$

$$= \sum_i \frac{1}{2m_i} \left( p_i^2 + q_i \vec{p}_i \cdot \vec{A}_i \right) \quad (55)$$

$$= \sum_i \frac{1}{2m_i} \left( p_i^2 + q_i \vec{p}_i \cdot \vec{A}_i + q_i^2 A_i^2 - q_i^2 A_i^2 \right) \quad (56)$$

$$\approx - \sum_i \frac{1}{2m_i} \left( p_i^2 + q_i \vec{p}_i \cdot \vec{A}_i + q_i^2 A_i^2 \right) \quad (57)$$

$$= \sum_i \frac{1}{2m_i} \left( p_i^2 + q_i \vec{p}_i \cdot \vec{A}_i + q_i^2 A_i^2 - q_i^2 A_i^2 \right) \quad (58)$$

$$\approx \sum_i \frac{1}{2m_i} \left( \vec{p}_i + q_i \vec{A}_i \right)^2 \quad (59)$$

where we have defined a ‘‘momentum’’,  $\vec{p}_i = m_i \vec{v}_i$ , and a ‘‘vector potential’’,  $\vec{A}_i = \sum_{j \neq i} \frac{q_j \vec{v}_j}{|\vec{r}_i - \vec{r}_j|}$ , then dropped all  $\frac{1}{r^2}$  terms. An attentive reader will notice that the final result in equation (59) is equal to the kinetic energy from a system of charged particles in a vacuum, including the momentum terms and vector potential terms. This is, of course, by design.

Plugging this all together into the loss function, equation (26) we get

$$\begin{aligned} \mathcal{L}_x &= -\log(\mathcal{P}(\Theta_x)) + \frac{M}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{K}|) + \frac{1}{2} \int d\mathbf{x}_{1:N} \mathcal{P}^*(\mathbf{x}_{1:M}) \mathbf{x}_{1:M} \mathbf{K}^{-1} \mathbf{x}_{1:M} \\ &\approx \frac{(D+1)M}{2} \log(2\pi) - \frac{(D-1)}{2} \log \left( \prod_i m_i \right) + \sum_i \left( \frac{\mathbb{E}[x_i^2]}{2m_i} - \sum_{j \neq i} \frac{\mathbb{E}[x_i x_j]}{2m_i m_j} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} + \frac{1}{2m_i} \left( \vec{p}_i + q_i \vec{A}_i \right)^2 \right). \end{aligned} \quad (60)$$

Our last simplification will come from an identity

$$\mathbb{E}[x_i x_j] = \text{Cov}[x_i, x_j] + \mathbb{E}[x_i] \mathbb{E}[x_j] \quad (62)$$

where the expectation values are taken over the ground truth distribution. Notice that  $\mathbb{E}[x]$  is the average error between our model (where the expected mean is 0) and the ground truth. We expect that this error is proportional to the size of the cluster,

$$\mathbb{E}[x_i] = \epsilon m_i \quad (63)$$

where  $\epsilon$  is expected error per binary variable in the cluster. Note that  $\text{Cov}[x_i, x_j]$  will be much smaller than  $\mathbb{E}[x_i] \mathbb{E}[x_j]$  which scales with the variance squared, therefore we can approximate

$$\mathbb{E}[x_i x_j] \approx \epsilon^2 m_i m_j. \quad (64)$$



Plugging this in we get

$$\mathcal{L}_x = \frac{(D+1)M}{2} \log(2\pi) - \frac{(D-1)}{2} \log\left(\prod_i m_i\right) + \sum_i \left( \frac{\epsilon^2 m_i}{2} - \sum_{j \neq i} \frac{\epsilon^2 q_i q_j}{2|\vec{r}_i - \vec{r}_j|} + \frac{1}{2m_i} (\vec{p}_i + q_i \vec{A}_i)^2 \right) \quad (65)$$

$$= \frac{(D+1)M}{2} \log(2\pi) - \sum_i \left( \frac{(D-1)}{2} \log(m_i) - \frac{\epsilon^2 m_i^2}{2} \right) - \sum_i \left( \sum_{j \neq i} \frac{\epsilon^2 q_i q_j}{2|\vec{r}_i - \vec{r}_j|} - \frac{1}{2m_i} (\vec{p}_i + q_i \vec{A}_i)^2 \right) \quad (66)$$

$$= \mathcal{C} - \mathcal{M} - \mathcal{H} \quad (67)$$

$$\mathcal{C} = \frac{(D+1)M}{2} \log(2\pi) \quad (68)$$

$$\mathcal{M} = \sum_i \left( \frac{(D-1)}{2} \log(m_i) - \frac{\epsilon^2 m_i^2}{2} \right) \quad (69)$$

$$\mathcal{H} = \sum_i \left( \sum_{j \neq i} \frac{\epsilon^2 q_i q_j}{2|\vec{r}_i - \vec{r}_j|} - \frac{1}{2m_i} (\vec{p}_i + q_i \vec{A}_i)^2 \right) \quad (70)$$

where we have separated our loss into a constant,  $\mathcal{C}$ , a mass term,  $\mathcal{M}$ , and a Hamiltonian,  $\mathcal{H}$ . If we expect that the clusters do not change over time, then when the model reaches minimum prediction error, the model parameters (positions) are allowed to evolve over time only by preserving the Hamiltonian, where the model was carefully chosen such that the Hamiltonian is equivalent to that of electrodynamics [cite].

In this section, we mapped the dynamics model parameters over cluster variables to electrodynamics. In the next section, we apply similar methodology to map dynamics of binary variables to quantum mechanics.

## 2.4 Model over binary variables

In this section, we will design the model architecture for the distribution over binary variables, with the goal of minimizing the error defined in equation (21). We will start by choosing a Hopfield model for our distribution over binary variables, then simplify our loss function until we get terms mimicking kinetic and potential energy. We will show that with a clever transformation, we can compactify the model, allowing model parameters and momenta to span a Hilbert space, such that the loss function becomes identical to the expectation value of a quantum mechanical Hamiltonian.

We will start by considering the probability distribution over a single binary variable,  $b$ . As mentioned, model the distribution over  $b$  as a Hopfield model

$$\mathcal{P}(b|\mathbf{x}_{1:N}) = \frac{1}{Z} \exp\left(b \left( J_0 + \sum_j J_j x_j \right)\right) \quad (71)$$

where  $Z$  is the normalization term,  $J_0$  represents a base energy, and each  $J_j$  represents an interaction energy with the corresponding cluster  $j$ . We are now going to make a substitution,

$$\sum_j J_j x_j = \sqrt{\sum_j J_j^2} \sum_j \frac{J_j}{\sqrt{\sum_j J_j^2}} x_j \quad (72)$$

$$= J_x \tilde{x} \quad (73)$$

$$J_x = \sqrt{\sum_j J_j^2} \quad (74)$$

$$\tilde{x} = \sum_j \frac{J_j}{\sqrt{\sum_j J_j^2}} x_j \quad (75)$$

where  $J_x$  can be considered a ‘‘magnitude of interaction’’ and  $\tilde{x}$  can be considered the dot product of the measurement vector with a unit vector in  $N$  dimensional space. We can use this result to simplify the model,

$$\mathcal{P}(b|\tilde{x}) = \frac{1}{Z} \exp(b(J_0 + J_x \tilde{x})) \quad (76)$$

where we have absorbed all information about cluster measurements into a single variable,  $\tilde{x}$ .

Let us now plug equations (75) and (76) into the loss function defined in equation (21),

$$\mathcal{L}_b = -\log(\mathcal{P}(\Theta_b)) - \sum_b \int d\mathbf{x}_{1:N} \mathcal{P}^*(b, \mathbf{x}_{1:M}) \log(\mathcal{P}(b|\mathbf{x}_{1:M}, \Theta_b)) \quad (77)$$

$$= -\log(\mathcal{P}(J_0, J_x)) - \sum_b \int d\tilde{x} \int d\mathbf{x}_{1:N} \delta\left(\tilde{x} - \sum_j \frac{J_j}{\sqrt{\sum_j J_j^2}} x_j\right) \mathcal{P}^*(b, \mathbf{x}_{1:M}) \log(\mathcal{P}(b|\tilde{x}, J_0, J_x)) \quad (78)$$

$$= -\log(\mathcal{P}(J_0, J_x)) - \sum_b \int d\tilde{x} \log(\mathcal{P}(b|\tilde{x}, J_0, J_x)) \int d\mathbf{x}_{1:N} \delta\left(\tilde{x} - \sum_j \frac{J_j}{\sqrt{\sum_j J_j^2}} x_j\right) \mathcal{P}^*(b, \mathbf{x}_{1:M}) \quad (79)$$

$$= -\log(\mathcal{P}(J_0, J_x)) - \sum_b \int d\tilde{x} \mathcal{P}^*(b, \tilde{x}) \log(\mathcal{P}(b|\tilde{x}, J_0, J_x)) \quad (80)$$

Furthermore, if we choose to condition the prior over model parameters,  $J_0$  and  $J_x$ , on their values at the previous time step, exactly as we did with model parameters in the previous section, we can interpret the log prior over model parameters as a kinetic energy term

$$-\log(\mathcal{P}(J_0, J_x)) = -\log(\mathcal{P}(J_0, J_x | J'_0, J'_x)) \quad (81)$$

$$= -\log(\mathcal{P}(p_0, p_x)) \quad (82)$$

where  $p_0$  is the momentum (change in value at the time step) of  $J_0$  and  $p_x$  is the momentum of  $J_x$ . Interpreting the log prior over model parameters as a kinetic energy our loss becomes

$$\mathcal{L}_b = -\log(\mathcal{P}(p_0, p_x)) - \sum_b \int d\tilde{x} \mathcal{P}^*(b, \tilde{x}) \log(\mathcal{P}(b|\tilde{x}, J_0, J_x)) \quad (83)$$

$$= -\sum_b \int d\tilde{x} \mathcal{P}^*(b, \tilde{x}) \log(\mathcal{P}(p_0, p_x)) - \sum_b \int d\tilde{x} \mathcal{P}^*(b, \tilde{x}) \log(\mathcal{P}(b|\tilde{x}, J_0, J_x)) \quad (84)$$

$$= -\sum_b \int d\tilde{x} \mathcal{P}^*(b, \tilde{x}) (\log(\mathcal{P}(b|\tilde{x}, J_0, J_x)) + \log(\mathcal{P}(p_0, p_x))) \quad (85)$$

$$= -\sum_b \int d\tilde{x} \mathcal{P}^*(b, \tilde{x}) \log(\mathcal{P}(b|\tilde{x}, J_0, J_x) \mathcal{P}(p_0, p_x)) \quad (86)$$

where for simplicity we have merged the kinetic and potential energy terms. The key takeaway from equation (86) is that the model over a binary variable coupled to external clusters can be represented using just two model parameters,  $J_0$  and  $J_x$ , along with their corresponding momenta,  $p_0$  and  $p_x$ . Thus,  $\mathcal{L}_b$  has only 4 degrees of freedom.

We have shown so far that the loss function for a binary variable coupled to external clusters, equation (86), can be represented using just 4 degrees of freedom. We are now going to exploit this insight by reparametrizing the model to span a Hilbert space so that we can map the dynamics of model parameters to quantum mechanics. In particular, quantum mechanical spin states span  $\mathcal{C}^2$ , which also has 4 degrees of freedom. So, let us compactify the four model parameters of our Hopfield model into a single quantum mechanical state vector

$$J_0, J_x, p_0, p_x, \tilde{x} \rightarrow e^{i\hbar\hat{H}(\tilde{x})}|\alpha\rangle \quad (87)$$

where the orientation of  $e^{i\frac{\Delta t}{\hbar}\hat{H}(\tilde{x})}|\alpha\rangle$  encodes  $J_0$ ,  $J_x$ ,  $p_0$ , and  $p_x$ . We will interpret  $|\alpha\rangle$  as the state vector at the previous time point and  $e^{i\frac{\Delta t}{\hbar}\hat{H}(\tilde{x})}$  as the time propagation operator. We will not define a functional form explicitly relating the quantum state representation to the Hopfield representation. However, we will define expressions relating how parts of equation (86) relate to the state vector. First, we define that the probability of a measurement is the squared probability amplitude

$$\left| \langle b | (\vec{\sigma} \cdot \vec{v}) e^{-i\frac{\Delta t}{\hbar}\hat{H}(\tilde{x})} |\alpha\rangle \right|^2 \equiv \mathcal{P}(b|\tilde{x}, J_0, J_x) \quad (88)$$

$$\approx \left| \langle b | (\vec{\sigma} \cdot \vec{v}) \left( \hat{I} - i\frac{\Delta t}{\hbar}\hat{H}(\tilde{x}) \right) |\alpha\rangle \right|^2 \quad (89)$$

where  $|b\rangle$  is a basis vector, the Pauli spin matrices dotted with a vector indicates we are measuring along some orientation  $\vec{v}$ , and we assume that  $\hat{H}(\tilde{x})$  is small enough that we can drop quadratic terms. Next, we will define the log of the model probability as

$$U_b(\tilde{x}) \equiv \log(\mathcal{P}(b|\tilde{x}, J_0, J_x) \mathcal{P}(p_0, p_x)) \quad (90)$$

and we will define the ground truth probability as

$$\left| \langle b | (\vec{\sigma} \cdot \vec{v}) \hat{\lambda}(\tilde{x}) e^{-i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x})} | \alpha \rangle \right|^2 \equiv \mathcal{P}^*(b|\tilde{x}) \quad (91)$$

$$\approx \left| \langle b | (\vec{\sigma} \cdot \vec{v}) \hat{\lambda}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \right|^2 \quad (92)$$

where  $\hat{\lambda}(\tilde{x})$  is some operator which corrects the model state vector to the “true” state vector. Plugging equations (89), (90), and (92) into (86), we get

$$\mathcal{L}_b = - \sum_b \int d\tilde{x} U_b \left| \langle b | (\vec{\sigma} \cdot \vec{v}) \hat{\lambda}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \right|^2 \quad (93)$$

$$= - \sum_b \int d\tilde{x} \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) \hat{\lambda}(\tilde{x})^\dagger (\vec{\sigma} \cdot \vec{v}) | b \rangle U_b \langle b | (\vec{\sigma} \cdot \vec{v}) \hat{\lambda}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \quad (94)$$

$$= - \int d\tilde{x} \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) \sum_b \left( \hat{\lambda}(\tilde{x})^\dagger (\vec{\sigma} \cdot \vec{v}) | b \rangle U_b \langle b | (\vec{\sigma} \cdot \vec{v}) \hat{\lambda}(\tilde{x}) \right) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \quad (95)$$

$$= - \int d\tilde{x} \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) \hat{U}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \quad (96)$$

$$\hat{U}(\tilde{x}) = \sum_b \left( \hat{\lambda}(\tilde{x})^\dagger (\vec{\sigma} \cdot \vec{v}) | b \rangle U_b \langle b | (\vec{\sigma} \cdot \vec{v}) \hat{\lambda}(\tilde{x}) \right) \quad (97)$$

where we see that our loss function becomes equivalent to the expectation value of spin states over some energy function,  $\hat{U}(\tilde{x})$ . Notice that the loss in equation (96) can potentially be complex, unless we require that  $\hat{H}(\tilde{x}) = \hat{U}(\tilde{x})$ . For example,

$$\begin{aligned} \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) \hat{U}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle &= \langle \alpha | \left( \hat{U}(\tilde{x}) + i \frac{\Delta t}{\hbar} [\hat{H}(\tilde{x}), \hat{U}(\tilde{x})] + \frac{\Delta t^2}{\hbar^2} \hat{H}(\tilde{x})^2 \right) | \alpha \rangle \\ &\approx \langle \alpha | \left( \hat{U}(\tilde{x}) + i \frac{\Delta t}{\hbar} [\hat{H}(\tilde{x}), \hat{U}(\tilde{x})] \right) | \alpha \rangle \\ &= \langle \alpha | \hat{U}(\tilde{x}) | \alpha \rangle + i \frac{\Delta t}{\hbar} \langle \alpha | [\hat{H}(\tilde{x}), \hat{U}(\tilde{x})] | \alpha \rangle \end{aligned}$$

is only guaranteed to be real for all values of  $\hat{U}(\tilde{x})$  if  $\hat{H}(\tilde{x}) = \hat{U}(\tilde{x})$  which gives  $[\hat{H}(\tilde{x}), \hat{U}(\tilde{x})] = 0$ . Thus, we interpret that the time propagation operator,  $\hat{H}(\tilde{x})$ , is inseparably linked to the model error. All together, the loss for our binary variable becomes

$$\mathcal{L}_b = - \int d\tilde{x} \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) \hat{H}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \quad (98)$$

$$= - \int d\mathbf{x}_{1:N} \int d\tilde{x} \delta \left( \tilde{x} - \sum_j \frac{J_j}{\sqrt{\sum_j J_j^2}} x_j \right) \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) \hat{H}(\tilde{x}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\tilde{x}) \right) | \alpha \rangle \quad (99)$$

$$= - \int d\mathbf{x}_{1:N} \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) \hat{H}(\mathbf{x}_{1:N}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) | \alpha \rangle \quad (100)$$

where  $\hat{H}(\mathbf{x}_{1:N}) = \hat{H}(\sum_j \frac{J_j}{\sqrt{\sum_j J_j^2}} x_j)$  and have reintroduced our cluster variables.

Now that we have a form for our loss, let us find a functional form for our Hamiltonian,  $\hat{H}(\mathbf{x}_{1:N})$ . Let us start by guessing a form and then checking that it agrees with our definition of covariance. Let us guess that our potential

energy for binary variables takes a similar form to the potential of cluster variables (see equations (33) and (37)), and our kinetic energy of binary variables takes a form similar to the kinetic energy of clusters (equation (59)),

$$\frac{x_i x_j}{m_i m_j} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \rightarrow \frac{(\vec{\sigma} \cdot \vec{v}) x_n}{m_b m_i} \frac{q_b q_i}{|\vec{r}_b - \vec{r}_i|} \quad (101)$$

$$\frac{1}{2m_i} \left( \vec{p}_i + q_i \vec{A}_i \right)^2 \rightarrow \frac{1}{2m_b} \left( \vec{\sigma} \cdot \left( \vec{p} - q_b \vec{A} \right) \right)^2 \quad (102)$$

where the measurement of a binary variable is related to the orientation,  $\vec{v}$ , and the momentum is replaced with a momentum operator,  $vec\mathbf{p}$  dotted with the Pauli spin matrix. All together this becomes

$$\hat{H}_T = \frac{1}{2m_b} \left( \vec{\sigma} \cdot \left( \vec{p} - q_b \vec{A} \right) \right)^2 \quad (103)$$

$$\hat{H}_U(\mathbf{x}_{1:N}) = - \sum_{n=1}^M \frac{(\vec{\sigma} \cdot \vec{v}) x_n}{m_b m_n} \frac{q_b q_n}{r_{bn}} \quad (104)$$

$$\hat{H}(\mathbf{x}_{1:N}) = \hat{H}_T + \hat{H}_U(\mathbf{x}_{1:N}) \quad (105)$$

$$\hat{H}(\mathbf{x}_{1:N}) = \frac{1}{2m_b} \left( \vec{\sigma} \cdot \left( \vec{p} - q_b \vec{A} \right) \right)^2 - \sum_{n=1}^M \frac{(\vec{\sigma} \cdot \vec{v}) x_n}{m_b m_n} \frac{q_b q_n}{r_{bn}} \quad (106)$$

$$= \frac{1}{2m_b} \left( \vec{p} - q_b \vec{A} \right)^2 - \frac{q_b}{2m_b} \vec{\sigma} \cdot \vec{B} - \sum_{n=1}^M \frac{(\vec{\sigma} \cdot \vec{v}) x_n}{m_b m_n} \frac{q_b q_n}{r_{bn}} \quad (107)$$

where  $\vec{B}$  (curl of the vector potential) is the magnetic field acting on the variable,  $m_b$  is the mass of  $b$ ,  $q_b$  is the charge of  $b$ ,  $r_{bn}$  is some distance metric between cluster  $n$  and variable  $b$ , and  $\vec{v}$  is the measurement orientation. Notice that equation (107) is the Hamiltonian defined in the Pauli equation, where the scalar potential is replaced with our measurement term.

We can verify that equation (107) matches our definition of covariance by plugging it into the definition of

covariance,

$$\text{Cov}[b, x_i] = \text{E}[(b - \bar{b})(x - \bar{x}_i)] \quad (108)$$

$$= \text{E}[bx_i] - \bar{b}\bar{x}_i \quad (109)$$

$$= \text{E}[bx_i] \quad (110)$$

$$= \int d\mathbf{x}_{1:N} \sum_b x_i b \mathcal{P}(b|\mathbf{x}_{1:N}) \mathcal{P}(\mathbf{x}_{1:N}) \quad (111)$$

$$= \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) \sum_b x_i b \left| \langle b | (\vec{\sigma} \cdot \vec{\nu}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) | \alpha \rangle \right|^2 \quad (112)$$

$$= \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) \sum_b x_i \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) (\vec{\sigma} \cdot \vec{\nu}) | b \rangle \langle b | (\vec{\sigma} \cdot \vec{\nu}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) | \alpha \rangle \quad (113)$$

$$= \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) x_i \langle \alpha | \left( \hat{I} + i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) (\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}) \left( \hat{I} - i \frac{\Delta t}{\hbar} \hat{H}(\mathbf{x}_{1:N}) \right) | \alpha \rangle \quad (114)$$

$$\approx \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) x_i \langle \alpha | \left( (\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}) - i \frac{\Delta t}{\hbar} [(\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), \hat{H}(\mathbf{x}_{1:N})] \right) | \alpha \rangle \quad (115)$$

$$= -i \frac{\Delta t}{\hbar} \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) x_i \langle \alpha | [(\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), \hat{H}(\mathbf{x}_{1:N})] | \alpha \rangle \quad (116)$$

$$= -i \frac{\Delta t}{\hbar} \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) x_i \langle \alpha | [(\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), \hat{H}_T + \hat{H}_U(\mathbf{x}_{1:N})] | \alpha \rangle \quad (117)$$

$$= -i \frac{\Delta t}{\hbar} \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) x_i \langle \alpha | [(\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), \hat{H}_U(\mathbf{x}_{1:N})] | \alpha \rangle \quad (118)$$

$$= -i \frac{\Delta t}{\hbar} \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) x_i \langle \alpha | \left[ (\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), - \sum_{n=1}^M \frac{(\vec{\sigma} \cdot \vec{\nu}) x_n q_b q_n}{m_b m_n r_{bn}} \right] | \alpha \rangle \quad (119)$$

$$= -i \frac{\Delta t}{\hbar} \langle \alpha | \left[ (\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), - \sum_{n=1}^M \int d\mathbf{x}_{1:N} \mathcal{P}(\mathbf{x}_{1:N}) \frac{(\vec{\sigma} \cdot \vec{\nu}) x_i x_n q_b q_n}{m_b m_n r_{bn}} \right] | \alpha \rangle \quad (120)$$

$$= i \frac{\Delta t}{\hbar} \langle \alpha | \left[ (\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), \frac{(\vec{\sigma} \cdot \vec{\nu}) q_b q_i}{m_b r_{bi}} \right] | \alpha \rangle \quad (121)$$

$$= i \frac{\Delta t}{\hbar} \frac{q_b q_i}{m_b r_{bi}} \langle \alpha | [(\vec{\sigma} \cdot \vec{\nu}) \hat{\sigma}_3 (\vec{\sigma} \cdot \vec{\nu}), (\vec{\sigma} \cdot \vec{\nu})] | \alpha \rangle \quad (122)$$

$$= i \frac{\Delta t}{\hbar} \frac{q_b q_i}{m_b r_{bi}} \langle \alpha | (-2i\nu^2 (\nu_1 \hat{\sigma}_2 - \nu_2 \hat{\sigma}_1)) | \alpha \rangle \quad (123)$$

$$= \frac{q_b q_i}{r_{bi}} \frac{2\Delta t \nu^2}{\hbar m_b} (\langle \hat{\sigma}_1 \rangle \nu_2 - \langle \hat{\sigma}_2 \rangle \nu_1) \quad (124)$$

where we have used that  $\bar{x}_i = \int dx_i x_i \mathcal{P}(x_i) = 0$  and we have dropped  $\frac{1}{r^2}$  terms appearing in  $\hat{H}(\mathbf{x}_{1:N})^2$ . Equation (124) shows that our Hamiltonian gives the proportionally correct covariance if  $\langle \hat{\sigma}_1 \rangle \nu_2 - \langle \hat{\sigma}_2 \rangle \nu_1 = 1$ . Let us define

$$\nu_1 = \frac{-\langle \hat{\sigma}_2 \rangle}{\langle \hat{\sigma}_1 \rangle^2 + \langle \hat{\sigma}_2 \rangle^2} \quad (125)$$

$$\nu_1 = \frac{\langle \hat{\sigma}_1 \rangle}{\langle \hat{\sigma}_1 \rangle^2 + \langle \hat{\sigma}_2 \rangle^2} \quad (126)$$

$$\nu^2 = \frac{\langle \hat{\sigma}_1 \rangle^2 + \langle \hat{\sigma}_2 \rangle^2}{\langle \hat{\sigma}_1 \rangle^2 + \langle \hat{\sigma}_2 \rangle^2} \quad (127)$$

$$= 1 \quad (128)$$

where our measurement angle,  $\vec{\nu}$ , is always orthogonal to the orientation at the previous time step.

$$\text{Cov}[b, x_i] = \frac{q_b q_n}{r_{bi}} \frac{2\Delta t \nu^2}{\hbar m_b} \quad (129)$$

$$\propto \frac{q_b q_n}{r_{bi}} \quad (130)$$

where evidently the proportionality can become an equality with proper choice of units.

### **3 Discussion**

...

### **References**

[1] *Cite me!!!*

# Supplementary information

## S0.1 Clustering correlated variables

Consider two highly correlated variables  $b_i$  and  $b_j$  such that

$$\mathcal{P}(b_i|b_j) \approx \delta_{q_i b_i, q_j b_j} \quad (1)$$

where  $\delta_{b_i, b_j}$  is the Kronecker delta and the variables  $q_i$  and  $q_j$  are the signs of correlation with  $q_i q_j = 1$  if  $b_i$  and  $b_j$  are correlated and  $q_i q_j = -1$  if  $b_i$  and  $b_j$  are anti-correlated.

The sum of interaction terms (see equation (3)) of highly correlated variables is approximately equal to the average of the signed interaction energies times the sum of the variables,

$$\begin{aligned} \mathbb{E}[J_i b_i + J_j b_j] &= \sum_{b_i} \sum_{b_j} (J_i b_i + J_j b_j) \mathcal{P}(b_i, b_j) \\ &= \sum_{b_i} \sum_{b_j} (J_i b_i + J_j b_j) \mathcal{P}(b_i|b_j) \mathcal{P}(b_j) \\ &\approx \sum_{b_i} \sum_{b_j} (J_i b_i + J_j b_j) \delta_{q_i b_i, q_j b_j} \mathcal{P}(b_j) \\ &= \sum_{b_i} \sum_{b_j} (J_i b_i + J_j b_j) \delta_{b_i, q_i q_j b_j} \mathcal{P}(b_j) \\ &= \sum_{b_j} (J_i q_i q_j b_j + J_j b_j) \mathcal{P}(b_j) \\ &= (J_i q_i q_j + J_j) \sum_{b_j} b_j \mathcal{P}(b_j) \\ &= (J_i q_i q_j + J_j) \sum_{b_j} \frac{b_j + b_j}{2} \mathcal{P}(b_j) \\ &= \frac{J_i q_i q_j + J_j}{2} \sum_{b_i} \sum_{b_j} (b_i + b_j) \delta_{b_i, b_j} \mathcal{P}(b_j) \\ &\approx \frac{J_i q_i q_j + J_j}{2} \sum_{b_i} \sum_{b_j} (b_i + b_j) \mathcal{P}(b_i|b_j) \mathcal{P}(b_j) \\ &= \frac{J_i q_i q_j + J_j}{2} \sum_{b_i} \sum_{b_j} (b_i + b_j) \mathcal{P}(b_i, b_j) \\ &= \frac{J_i q_i q_j + J_j}{2} \mathbb{E}[b_i + b_j] \\ &= q_j \frac{J_i q_i + J_j q_j}{2} \mathbb{E}[b_i + b_j] \end{aligned}$$

where in the last step we have used an identity  $q_j = \frac{1}{q_j}$  since it  $q_j$  may only be 1 or  $-1$ . In general, we can cluster any  $K$  highly correlated and anti-correlated variables to get

$$\mathbb{E} \left[ \sum_k J_k b_k \right] \approx q_{rf} \frac{\sum_k q_k J_k b_k}{K} \mathbb{E} \left[ \sum_k b_k \right]. \quad (2)$$

where  $q_{rf}$  is some reference charge. Note that we will not use equation (2) directly in this work. We merely are interested in demonstrating that sets of highly correlated binary variables may be replaced with variables that only consider the sum of the set.