Information Physics

J. Shepard Bryan IV

Department of Physics, Arizona State University

February 14, 2024

Abstract

Here we attempt to solve for dynamical equations governing model parameters during optimal learning. We show that equations similar to electrodynamics and quantum dynamics are solutions to optimal learning.

DISCLAIMER

PLEASE NOTE THAT THIS IS A WORK IN PROGRESS. PLEASE EXCUSE ANY TYPOS IN THE WRITING AND IN THE EQUATIONS. PLEASE ALSO EXCUSE THE LACK OF CITATIONS IN THE WORK SO FAR. WHEN YOU READ THIS WORK, I ASK THAT YOU PLEASE TRY TO FOCUS MORE ON THE BIG PICTURE AND LESS ON THE TYPOS, AS THE TYPOS WILL BE CORRECTED OVER TIME, BUT THE BIG PICTURE WILL LIKELY REMAIN MORE CONSTANT. YOUR FEEDBACK IS APPRECIATED.

1 Introduction

The field of artificial intelligence offers a large selection of models and training protocols to make predictions based on data. However, often times these models are arrived at through vague intuition or trial and error. It would be ideal if we could obtain a model and training protocol through more rigorous means. In this work we attempt to construct a model and derive optimal learning dynamics from first principles. Interestingly, we show that careful definitions of model parameters lead to learning dynamics with equations of motion identical to electrodynamics and quantum mechanics.

The goal of this work is to show that physical equations can be solutions to optimal learning. In order to show this, we frame the learning problem as a relationship between an observer and measurements. In this paradigm, we represent parameters of the observer's model using physical constants such as mass, charge, and distance. While we admit that to a certain extent we choose definitions of physical constants to fit the expected behavior, we also find these definitions can be justified at a more intuitive level as well. This work may hint that there exists a deeper connection between physics and learning than previously realized.

We hope that the long term impact of this work is to advance the field of artificial intelligence by creating a paradigm in which optimal models and training protocols can be selected with better precision.

2 Methods

2.1 Definitions

We start by laying down verbal definitions of terminology used throughout the work. The focus of this section is to introduce terminology and illustrate how the concepts are connected. In the next section we will apply mathematics to these definitions.

Observer An observer is a thing that makes measurements. An observer can be an entity or a computer program. The main goal of the observer is to predict the future measurements.

Measurements Measurements are the data that the observer receives. Measurements are ordered, meaning that the observer can arrange the measurements into a sequence of time steps and will be able to match each measurement to a specific time step. They are distinguishable, meaning that if the observer is making multiple measurements

per time level, they will be able to distinguish which measurement came from which observable. We also will allow redundant measurements, such that the observer can measure the same observable multiple times in a single time step, where redundant measurements are also distinguishable.

Observables Observables are the things that are getting measured. At each time step, the observer receives a measurement from each observable. Observables can be continuous, in which case measurements can take on any value on the real line, or discrete, in which measurements can take on only certain values. From the observer's point of view, observables are effectively random variables, which take on values when measured, but appear non-deterministic in the absence of a perfect model.

Model A model might be best described as "the thing that helps the observer make predictions about the measurements of observables". In our case, the observer's model is a probability distribution over observables. We will work in a Bayesian paradigm, where the model will treat both observables and model parameters as random variables.

Parameters Model parameters are simply the values that go into the observer's model. By tweaking model parameters, the observer is able to tune the predictive distribution to best match the data.

Learning Learning is the act of modifying model parameters in order to best match prior data and predict future data. As each new data point comes in, the observer will adjust the model parameters to minimize prediction error.

Memory The observer has a memory where it can store a record of the values of past measurements as well as the values of model parameters. The observer's memory is finite, meaning that when the memory is full, the observer must forget an old value in order to record a new value. Thus, part of an observer's job is to minimize the amount of values that are written down while maximizing predictive power.

All together, we say, "An observer predicts the measurements of observables using a model that it stores in its memory. The observer learns from measurements by updating the parameters of its model to best match with the data."

2.2 Optimal learning

We are now ready to apply equations to our observer-measurement system. We break this section into three parts. In the first part, we define the observer's prediction error at a single time step as a loss function. In the second section we define model variables over continuous observables and show that modeling continuous observables as charged particles interacting in a vaccum leads to maximum predictive power. In the third section we define model variables over binary observables and show that modeling binary observables as quantum spin states leads to maximum predictive power. In the final section, we construct a learning algorithm where perturbations from inaccurate predictions are modeled as external forces on our system of particles.

2.2.1 Constructing the loss function

Let us imagine an observer-measurement system in which time is discretized into individual time steps. At each time step the observer measures each observable. For simplicity, we will start with an assumption that the observer measures each observable once per time step, but later on we will adjust this assumption. Let there be K observables. Let $\boldsymbol{X}$ be the measurements of the observables at a time level of interest, where X_k is the value of measurement of the kth observable for the time step. For now, let us assume that the observables are continuous.

To the observer, these measurements of observables are effectively random variables drawn from some distribution. Let the "true" distribution over observables be $\mathcal{P}^*(\boldsymbol{X})$, and let the observer's model likelihood be $\mathcal{P}(\boldsymbol{X}|\Theta)$ where Θ are the parameters of the observer's model with a prior $\mathcal{P}(\Theta)$. Notice that the observer's model is not necessarily equal to the true distribution over observables. The observer does not know what the true distribution is, but it is aware that its own model may deviate from the true distribution.

The goal of the observer is to predict measurements of observables with the lowest possible error. One way to quantify the mismatch between the observer's model and the true distribution is with generalization error (also known as risk) [1]

$$\mathcal{L} = -\int d\boldsymbol{X}\mathcal{P}^*(\boldsymbol{X})\log\left(\mathcal{P}(\boldsymbol{X}|\Theta)\mathcal{P}(\Theta)\right) \tag{1}$$

which calculates the negative of the expectation value of the log of the observer's model. We call this generalization error, the "loss". We will assume that the observer wants to choose a model which minimizes this loss term.

We now have an equation that we would like to minimize, but we do not know the form of the observer's model, nor the nature of the model parameters, Θ. To move forward, we will assume that the observer's likelihood follows

a multivariate normal distribution,

$$\mathcal{P}\left(\boldsymbol{X}|\Theta\right) \approx \mathcal{N}\left(\boldsymbol{X};\boldsymbol{\mu}_\Theta,\boldsymbol{\Sigma}_\Theta\right) \tag{2}$$

$$= \left((2\pi)^K|\boldsymbol{\Sigma}_\Theta|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_\Theta)^T\boldsymbol{\Sigma}_\Theta^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_\Theta)\right) \tag{3}$$

where the subscript on $\boldsymbol{\Sigma}_\Theta$ and $\boldsymbol{\mu}_\Theta$ indicates that these are functions of Θ. Since a normal distribution can be thought of as a second order approximation to a distribution in log space [2], we can consider this step in our work as an approximation rather which maintains much generality than an assumption which restricts the work.

Before we move forward, let us make one more simplification by redefining observables such that the mean of the distribution is at zero, $\boldsymbol{\mu}_\Theta = \mathbf{0}$. The observer's new likelihood after this transformation is

$$\mathcal{P}\left(\boldsymbol{X}|\Theta\right) = \mathcal{N}\left(\boldsymbol{X};\mathbf{0},\boldsymbol{\Sigma}\right) \tag{4}$$

$$= \left((2\pi)^K|\boldsymbol{\Sigma}_\Theta|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}_\Theta^{-1}\boldsymbol{X}\right). \tag{5}$$

This can be interpreted as the observer simply redefining observables such that they satisfy this condition. We will ultimately marginalize over this parameter, therefore this transformation will have minimal effect.

We now must choose a form for the prior over model parameters, $\mathcal{P}\left(\Theta\right)$. The first step in this is to specify the hyperparameters that govern the form of the prior. That is, we must specify what additional information the observer must store in its memory that it uses to compute the prior. One clever way to minimize the amount of memory used by the observer is to choose hyperparameters calculated from the model parameters of the previous time step since these values are already stored in the observer's memory. Concretely, we say

$$\mathcal{P}\left(\Theta\right) = \mathcal{P}\left(\Theta|\Theta'\right) \tag{6}$$

where Θ' are the model parameters of the previous time step.

Next we choose a form for the prior over Θ. Let us choose a multivariate Gaussian as the prior over Θ,

$$\mathcal{P}\left(\Theta|\Theta'\right) = \mathcal{N}\left(\boldsymbol{\Theta};\boldsymbol{\alpha}_{\Theta'},\boldsymbol{\Lambda}_{\Theta'}\right) \tag{7}$$

$$= \left((2\pi)^M|\boldsymbol{\Lambda}_{\Theta'}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\Theta}-\boldsymbol{\alpha}_{\Theta'})^T\boldsymbol{\Lambda}_{\Theta'}^{-1}(\boldsymbol{\Theta}-\boldsymbol{\alpha}_{\Theta'})\right) \tag{8}$$

where $\boldsymbol{\Theta}$ is a vector containing each of the elements of Θ, M is the total number of model parameters, and $\boldsymbol{\alpha}_{\Theta'}$ and $\boldsymbol{\Lambda}_{\Theta'}$ are the mean and covariance of this distribution where, again, the subscript indicates that they are functions of the hyperparameters, Θ'. All together the observer's loss becomes

$$\mathcal{L} = -\int d\boldsymbol{X}\mathcal{P}^*\left(\boldsymbol{X}\right)\log\left(\left((2\pi)^{K+M}|\boldsymbol{\Sigma}_\Theta||\boldsymbol{\Lambda}_{\Theta'}|\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}_\Theta^{-1}\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\Theta}-\boldsymbol{\alpha}_{\Theta'})^T\boldsymbol{\Lambda}_{\Theta'}^{-1}(\boldsymbol{\Theta}-\boldsymbol{\alpha}_{\Theta'})\right)\right) \tag{9}$$

We must now choose values for $\boldsymbol{\alpha}_{\Theta'}$ and $\boldsymbol{\Lambda}_{\Theta'}$. One excellent choice for the mean is to set it equal to the previous values of the model parameters,

$$\boldsymbol{\alpha}_{\Theta'} = \boldsymbol{\Theta}' \tag{10}$$

which intuitively indicates that the value of the model parameters at one time level should be related to the model parameters of the previous time level. As for $\boldsymbol{\Lambda}_{\Theta'}$, there is a logical choice, but it is less straightforward.

Working in the Bayesian paradigm, we typically try to model priors in a way that prioritizes simplification of the mathematics of the problem since the prior's influence on the final inference already diminishes as data are collected [3]. In our case we will stick to this philosophy and shamelessly select a prior based entirely on the way it simplifies the problem. The most cumbersome part of our computation at this stage is the normalization terms in our likelihood and prior since they involve log determinants of potentially large matrices. Thus a logic choice for $\boldsymbol{\Lambda}_{\Theta'}$ would be one that allows normalization terms to cancel out. This is achievable if we set $\boldsymbol{\Lambda}_{\Theta'} = \boldsymbol{\Sigma}_\Theta^{-1}$ since

$$\left(|\boldsymbol{\Sigma}_\Theta||\boldsymbol{\Lambda}_{\Theta'}|\right)^{-\frac{1}{2}} = \left(|\boldsymbol{\Sigma}_\Theta||\boldsymbol{\Sigma}_\Theta^{-1}|\right)^{-\frac{1}{2}} \tag{11}$$

$$= \left(|\boldsymbol{\Sigma}_\Theta||\boldsymbol{\Sigma}_\Theta|^{-1}\right)^{-\frac{1}{2}} \tag{12}$$

$$= 1 \tag{13}$$

but there are two caveats to this. First, technically the covariance matrix over Θ cannot be a function of Θ, however if we assume that model parameters evolve slowly enough such that $\Theta \approx \Theta'$, then this becomes a valid approximation. Second, by defining the covariance matrix over model parameters to be $\mathbf{\Sigma}_\Theta^{-1}$, we are implicitly setting the number of model parameters to be K since the side length of the covariance matrix must correspond to the number of model parameters. However, there is a trick to avoid this restriction, which we discuss next.

The covariance matrix for observables, $\mathbf{\Sigma}_\Theta$, has roughly $K^2/2$ elements (one for each pair of observables), but by defining the covariance matrix over model parameters to be $\mathbf{\Sigma}_\Theta^{-1}$ we are setting the number of model parameters to be K, which being a much smaller number, may lead to significant reduction in model capacity. There is a way to avoid this restriction. To do this, we split the model parameters into D disjoint sets, where each set has a prior with our desired covariance

$$\mathcal{P}\left(\Theta|\Theta'\right) = \prod_{d=1}^{D} \mathcal{P}\left(\Theta_d|\Theta'_d\right) \tag{14}$$

$$= \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{\Theta}_d; \mathbf{\Theta}'_d, \mathbf{\Sigma}_\Theta^{-1}\right) \tag{15}$$

$$= \prod_{d=1}^{D} \left((2\pi)^K |\mathbf{\Sigma}_\Theta^{-1}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{\Theta}_d - \mathbf{\Theta}'_d)^T \mathbf{\Sigma}_\Theta (\mathbf{\Theta}_d - \mathbf{\Theta}'_d)\right) \tag{16}$$

$$= \left((2\pi)^{DK} |\mathbf{\Sigma}_\Theta|^{-D}\right)^{-\frac{1}{2}} \exp\left(\sum_{d=1}^{D} -\frac{1}{2}(\mathbf{\Theta}_d - \mathbf{\Theta}'_d)^T \mathbf{\Sigma}_\Theta (\mathbf{\Theta}_d - \mathbf{\Theta}'_d)\right) \tag{17}$$

where the d subscript indicates the dth set of model parameters and where we have substituted $\boldsymbol{\alpha}_{\Theta'} = \mathbf{\Theta}'$ and $\mathbf{\Lambda}_{\Theta'} = \mathbf{\Sigma}_\Theta^{-1}$. Splitting the model parameters into disjoint subsets allows us to increase the number of model parameters, but it also introduces D normalization terms into our loss function. If we still want the likelihood and prior normalization terms to cancel, the solution is to allow duplicate measurements of each observable.

We started this section by assuming that the observer measures each observable once per time step, however, our modeling choices make it much more convenient if the observer makes duplicate measurements of each observable. Let us move forward with this assumption. Let $\boldsymbol{X}_d$ be the dth set of measurements for a time step and let $\boldsymbol{X}_{1:D}$ be the collection of all measurements. Then the new loss function becomes

$$\mathcal{L} = -\int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \log\left(\prod_{d=1}^{D} \mathcal{P}\left(\boldsymbol{X}_d|\Theta\right) \mathcal{P}\left(\Theta_d|\Theta'_d\right)\right) \tag{18}$$

$$= -\int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \log\left(\prod_{d=1}^{D} \left((2\pi)^{2K} |\mathbf{\Sigma}_\Theta||\mathbf{\Sigma}_\Theta^{-1}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{X}_d^T \mathbf{\Sigma}_\Theta^{-1} \boldsymbol{X}_d - \frac{1}{2}(\mathbf{\Theta}_d - \mathbf{\Theta}'_d)^T \mathbf{\Sigma}_\Theta (\mathbf{\Theta}_d - \mathbf{\Theta}'_d)\right)\right) \tag{19}$$

$$= -\int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{d=1}^{D} \left(-K\log\left(2\pi\right) - \frac{1}{2}\boldsymbol{X}_d^T \mathbf{\Sigma}_\Theta^{-1} \boldsymbol{X}_d - \frac{1}{2}(\mathbf{\Theta}_d - \mathbf{\Theta}'_d)^T \mathbf{\Sigma}_\Theta (\mathbf{\Theta}_d - \mathbf{\Theta}'_d)\right). \tag{20}$$

Equation (20) is the loss function for our observer-measurement system. By minimizing this equation, the observer maximizes its predictive power.

Note that while equation (20) relates how the model parameters should evolve with respect to observables, it does not define a relationship between model parameters and the covariance matrix, $\mathbf{\Sigma}_\Theta$. In the next section we will choose a representation for the covariance matrix, $\mathbf{\Sigma}_\Theta$ inspired by electromagnetism. We will show that we can choose a covariance matrix such that model parameters evolve with equations of motion identical to electrodynamics.

2.2.2 Modeling continuous observables as charged particles

In the previous section we set up the loss function for our observer-measurement system, but we did not specifically define how model parameters relate to the observer's model. In this section, we choose a representation for the observer's model inspired by physics.

Let us start by looking closely at the covariance matrix, which can be written as,

$$\boldsymbol{\Sigma}_\Theta = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{1,2} & \sigma_1\sigma_3\rho_{1,3} & \dots \\ \sigma_1\sigma_2\rho_{1,2} & \sigma_2^2 & \sigma_2\sigma_3\rho_{2,3} & \dots \\ \sigma_1\sigma_3\rho_{1,3} & \sigma_2\sigma_3\rho_{2,3} & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{21}$$

where σ_i^2 is the variance of measurements of the ith observable and $\rho_{i,j}$ is the correlation between measurements of the ith and jth observables. The variance, σ_i^2, quantifies the intrinsic uncertainty of measurements of the observables. The correlation, $\rho_{i,j}$, quantifies the strength of the correlation between measurements of observables, where $\rho_{i,j} = 0$ indicates no correlation and $|\rho_{i,j}| = 1$ indicates perfect correlation. The correlation, $\rho_{i,j}$, also has a sign which indicates whether measurements of two observables are correlated or anti-correlated. Eventually we will need to calculate the inverse covariance matrix so let us do that now

$$\boldsymbol{\Sigma}_\Theta = \left(\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{1,2} & \sigma_1\sigma_3\rho_{1,3} & \dots \\ \sigma_1\sigma_2\rho_{1,2} & \sigma_2^2 & \sigma_2\sigma_3\rho_{2,3} & \dots \\ \sigma_1\sigma_3\rho_{1,3} & \sigma_2\sigma_3\rho_{2,3} & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)^{-1} \tag{22}$$

$$= \left(\begin{bmatrix} \sigma_1 & 0 & 0 & \dots \\ 0 & \sigma_2 & 0 & \dots \\ 0 & 0 & \sigma_3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \dots \\ \rho_{1,2} & 1 & \rho_{2,3} & \dots \\ \rho_{1,3} & \rho_{2,3} & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & \dots \\ 0 & \sigma_2 & 0 & \dots \\ 0 & 0 & \sigma_3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)^{-1} \tag{23}$$

$$= (\boldsymbol{\sigma}(\boldsymbol{I} + \boldsymbol{\rho})\boldsymbol{\sigma})^{-1} \tag{24}$$

$$= \boldsymbol{\sigma}^{-1}(\boldsymbol{I} + \boldsymbol{\rho})^{-1}\boldsymbol{\sigma}^{-1} \tag{25}$$

$$\approx \boldsymbol{\sigma}^{-1}(\boldsymbol{I} - \boldsymbol{\rho})\boldsymbol{\sigma}^{-1}. \tag{26}$$

We wish to represent the elements of the covariance matrix with physical labels. We first define variance as mass

$$m_i = \sigma_i^2 \tag{27}$$

where m_i is the mass of observable i. There are many ways to justify variance as mass, and it actually is not an uncommon definition [4], however a discussion on mass as a form of intrinsic uncertainty is beyond the scope of this work. For now we will just justify it on grounds that uncertainty seems to scale with mass since you can know your mass to plus or minus a few kilograms, but you can only know the mass of the sun to plus or minus a few million kilograms!

We next will define covariance as a charge over distance,

$$\sigma_i\sigma_j\rho_{i,j} = \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \tag{28}$$

where q_i is the charge of observable i and $\vec{r}_i$ is the position of observable i. To our knowledge, parametrizing covariance in this way is not common and therefore requires a bit more justification. Focussing first on the signed nature of charge, we notice that if we assign correlated observables with like charges, then their covariance will be positive, and likewise if we assign opposite charges to anti-correlated particles, then their covariance will be negative, therefore it makes sense to use charge as a means of keeping track of the sign of correlation. Focussing next on distance as a representation of correlation magnitude, we can justify this on a principle that two observables that are close together should be strongly correlated and two observables that are far apart should be weakly correlated, much in tune with our day-to-day intuition. This makes sense, but why do we choose to model that correlation dies off like $1/r$ instead of something different like $1/r^2$ or e^{-r}? While there are justifications for this choice (like treating the metric as the first term of a Laurent series, or choosing $1/r$ because it leads to stable orbits in 3 dimensions), the simple and more satisfying answer is that we make this choice because it gives us the answer we want.

Before we move forward, we need to address a couple of points about our choice of metric. First, an astute reader will realize that our definition of covariance requires distances between observables to be large since our $1/r$ term will go to infinity as distances between observables go to zero. As such, moving forward, we will assume that distances are large, and in particular, we will assume that higher order terms such as $1/r^2$ can be ignored.

Second, by representing distances between observables as distances between positions, we are implicitly embedding our covariance metric in a Euclidean space, but we have not yet specified the number of dimensions of this space. Later we will show that when we define a covariance between continuous and discrete observables, there can be at most 3 orthogonal spatial dimensions. For now, let us move forward with an assumption that observables exist in a 3-dimensional space, knowing that we have not proved this yet but will prove this later.

All together, our covariance matrix is

$$\boldsymbol{\Sigma}_\Theta = \begin{bmatrix} m_1 & \frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|} & \frac{q_1 q_3}{|\vec{r}_1 - \vec{r}_3|} & \cdots \\ \frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|} & m_2 & \frac{q_2 q_3}{|\vec{r}_2 - \vec{r}_3|} & \cdots \\ \frac{q_1 q_3}{|\vec{r}_1 - \vec{r}_3|} & \frac{q_2 q_3}{|\vec{r}_2 - \vec{r}_3|} & m_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{29}$$

We can also plug in our definitions into the inverse covariance matrix, equation (26), to obtain

$$\boldsymbol{\Sigma}_\Theta^{-1} = \begin{bmatrix} \frac{1}{m_1} & -\frac{1}{m_1 m_2}\frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|} & -\frac{1}{m_1 m_3}\frac{q_1 q_3}{|\vec{r}_1 - \vec{r}_3|} & \cdots \\ -\frac{1}{m_1 m_2}\frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|} & \frac{1}{m_2} & -\frac{1}{m_2 m_3}\frac{q_2 q_3}{|\vec{r}_2 - \vec{r}_3|} & \cdots \\ -\frac{1}{m_1 m_3}\frac{q_1 q_3}{|\vec{r}_1 - \vec{r}_3|} & -\frac{1}{m_2 m_3}\frac{q_2 q_3}{|\vec{r}_2 - \vec{r}_3|} & \frac{1}{m_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{30}$$

Now that our covariance matrix is defined, we finally know what the model parameters, Θ, represent: the positions of the observables. The D disjoint subsets of model parameters represent the positions along each of the D dimensions. Moving forward, we can substitute

$$\boldsymbol{\Theta}_d = \boldsymbol{r}_d \tag{31}$$

where $\boldsymbol{r}_d$ contains the positions of each observable along the dimension d. We can also interpret that the D measurements of each observable correspond to measurements of the observable along each of the spatial dimensions. Let $\vec{X}_i$ be the vector of measurements for observable i. On a final note regarding model parameters, some may argue that we should also treat the mass and charge of each particle as model parameters. This may be justifiable, however, later we will redefine continuous observables as sums over collections of binary observables, in which case, mass and charge will be directly related to the number of observables in the collection. Thus, we will only consider the positions of observables to be model parameters.

Plugging our definitions, equations (29) and (30), into our loss function, equation (20) we get

$$\mathcal{L} = -\int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{d=1}^{D} \left(-K \log\left(2\pi\right) - \frac{1}{2}\boldsymbol{X}_d^T \boldsymbol{\Sigma}_\Theta^{-1} \boldsymbol{X}_d - \frac{1}{2}(\boldsymbol{r}_d - \boldsymbol{r}'_d)^T \boldsymbol{\Sigma}_\Theta (\boldsymbol{r}_d - \boldsymbol{r}'_d) \right) \tag{32}$$

$$= DK \log\left(2\pi\right) + \int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{d=1}^{D} \left(\frac{1}{2}\boldsymbol{X}_d^T \boldsymbol{\Sigma}_\Theta^{-1} \boldsymbol{X}_d \right) + \sum_{d=1}^{D} \left(\frac{1}{2}(\boldsymbol{r}_d - \boldsymbol{r}'_d)^T \boldsymbol{\Sigma}_\Theta (\boldsymbol{r}_d - \boldsymbol{r}'_d) \right) \tag{33}$$

where we pull terms that do not depend on $\boldsymbol{X}_{1:D}$ out of the integral. There are two main terms to our loss function: a likelihood term which contains an integral over measurements, and a prior term. For simplicity, let us treat each of these terms separately.

Focussing on the likelihood term of equation (33), let us expand the matrix product as a sum over indices

$$\int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{d=1}^{D} \left(\frac{1}{2}\boldsymbol{X}_d^T \boldsymbol{\Sigma}_\Theta^{-1} \boldsymbol{X}_d \right) = \int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{d=1}^{D} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\frac{1}{2} X_{di} \Sigma_{\Theta ij}^{-1} X_{dj} \right) \tag{34}$$

$$= \int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\frac{1}{2} \vec{X}_i \cdot \vec{X}_j \Sigma_{\Theta ij}^{-1} \right) \tag{35}$$

$$= \int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{i=1}^{K} \left(\frac{1}{2} |\vec{X}_i|^2 \Sigma_{\Theta ij}^{-1} + \sum_{j \neq i} \frac{1}{4} \vec{X}_i \cdot \vec{X}_j \Sigma_{\Theta ij}^{-1} \right) \tag{36}$$

$$= \int d\boldsymbol{X}_{1:D} \mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right) \sum_{i=1}^{K} \left(\frac{|\vec{X}_i|^2}{2m_i} - \sum_{j \neq i} \frac{1}{4} \frac{\vec{X}_i \cdot \vec{X}_j}{m_i m_j} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \right) \tag{37}$$

where $\vec{X}_i$ is a vector containing each of the D measurements of observable i, and where $\Sigma^{-1}_{\Theta ij}$ is the element of $\mathbf{\Sigma}^{-1}_{\Theta}$ at the ith row and jth column. To move forward we must treat this integral over measurements, which is not straightforward since the observer does not know the exact form of $\mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right)$ aside from knowing that it is close to the observer's model plus some error. It makes sense then to represent this integral over measurements in terms of the error. To do this let us conjecture that the error should be related to the expectation value over measurements like

$$\int d\boldsymbol{X}_{1:D}\mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right)\vec{X}_i\cdot\vec{X}_j = \epsilon^2 m_i m_j \tag{38}$$

where ϵ^2 is the average expected error. Now, why do we expect that this integral over measurements should be related to the mass? The answer is that if we think of this distribution as a random walk [2] with some small bias towards a direction, then this small bias, which corresponds to ϵ^2, is amplified by the number of steps, which corresponds to the mass of the observable. But, why do we expect that the error, ϵ^2, should be a constant and not dependent on the observables? The answer to this is that while it may be reasonable to expect different errors for different observables, that would require storing K^2 additional variables in the observer's memory, which is memory intensive, therefore for large numbers of observables, an observer may opt to treat error as a global quantity in order to minimize memory requirements. Lastly, why do we not expect a sign in the error? The answer is that during the learning process, we expect the model to overestimate quantities as frequently as it underestimates quantities, such that the time averaged value of error is equal to zero, even while the magnitude of error is nonzero. Now that we have established that (38) is reasonable let us plug it into our expression

$$\int d\boldsymbol{X}_{1:D}\mathcal{P}^*\left(\boldsymbol{X}_{1:D}\right)\sum_{i=1}^{K}\left(\frac{|\vec{X}_i|^2}{2m_i}-\sum_{j\neq i}\frac{1}{4}\frac{\vec{X}_i\cdot\vec{X}_j}{m_i m_j}\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right)=\sum_{i=1}^{K}\left(\frac{\epsilon^2 m_i}{2}-\sum_{j\neq i}\frac{1}{4}\epsilon^2\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right). \tag{39}$$

We are now ready to move onto the next term in our expression.

We now focus on the prior term of equation (33), we again expand the matrix as a sum over indices

$$\sum_{d=1}^{D}\left(\frac{1}{2}(\boldsymbol{r}_d-\boldsymbol{r}'_d)^T\mathbf{\Sigma}_\Theta(\boldsymbol{r}_d-\boldsymbol{r}'_d)\right)=\sum_{d=1}^{D}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\frac{1}{2}(r_{di}-r'_{di})\Sigma_{\Theta ij}(r_{dj}-r'_{dj})\right) \tag{40}$$

$$=\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\frac{1}{2}(\vec{r}_i-\vec{r}'_i)\cdot(\vec{r}_j-\vec{r}'_j)\Sigma_{\Theta ij}\right) \tag{41}$$

$$=\sum_{i=1}^{K}\left(\frac{1}{2}(\vec{r}_i-\vec{r}'_i)^2 m_i+\sum_{j\neq k}\frac{1}{4}(\vec{r}_i-\vec{r}'_i)\cdot(\vec{r}_j-\vec{r}'_j)\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right) \tag{42}$$

$$=\sum_{i=1}^{K}\left(\frac{1}{2}(\vec{r}_i-\vec{r}'_i)^2 m_i+\sum_{j\neq k}\frac{1}{4}(\vec{r}_i-\vec{r}'_i)\cdot(\vec{r}_j-\vec{r}'_j)\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right). \tag{43}$$

Let us first simplify this expression by relating the change in position for the time step as a momentum term $\vec{p}_i = m_i(\vec{r}_i-\vec{r}'_i)$. Let us plug this into our expression

$$\sum_{i=1}^{K}\left(\frac{1}{2}(\vec{r}_i-\vec{r}'_i)^2 m_i+\sum_{j\neq k}\frac{1}{4}(\vec{r}_i-\vec{r}'_i)\cdot(\vec{r}_j-\vec{r}'_j)\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right)=\sum_{i=1}^{K}\left(\frac{p_i^2}{2m_i}+\sum_{j\neq k}\frac{1}{4}\frac{\vec{p}_i\cdot\vec{p}_j}{m_i m_j}\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right). \tag{44}$$

Finally, we simplify further using completing the squares and dropping all $1/r^2$ terms

$$\sum_{i=1}^{K}\left(\frac{p_i^2}{2m_i}+\sum_{j\neq k}\frac{1}{4}\frac{\vec{p}_i\cdot\vec{p}_j}{m_i m_j}\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right)=\sum_{i=1}^{K}\frac{1}{2m_i}\left(p_i^2+2q_i\vec{p}_i\cdot\sum_{j\neq k}\frac{q_j\vec{p}_j}{4m_j|\vec{r}_i-\vec{r}_j|}\right) \tag{45}$$

$$\approx\sum_{i=1}^{K}\frac{1}{2m_i}\left(\vec{p}_i+q_i\sum_{j\neq k}\frac{q_j\vec{p}_j}{4m_j|\vec{r}_i-\vec{r}_j|}\right)^2 \tag{46}$$

$$\approx\sum_{i=1}^{K}\frac{1}{2m_i}\left(\vec{p}_i+q_i\sum_{j\neq k}\vec{A}_{ij}\right)^2 \tag{47}$$

$$\tag{48}$$

where we have defined a vector potential from observable j acting on observable i as $\vec{A}_{ij}=\frac{q_j\vec{p}_j}{4m_j|\vec{r}_i-\vec{r}_j|}$, which, aside from scaling factors, is equivalent to the vector potential we would expect from charged particles in a vacuum.

Putting it all together we get

$$\mathcal{L}=DK\log(2\pi)+\sum_{i=1}^{K}\left(\frac{\epsilon^2 m_i}{2}\right)-\sum_{i=1}^{K}\left(\sum_{j\neq i}\left(\frac{1}{4}\epsilon^2\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right)-\frac{1}{2m_i}\left(\vec{p}_i+q_i\sum_{j\neq k}\vec{A}_{ij}\right)^2\right) \tag{49}$$

$$=DK\log(2\pi)+\mathcal{M}-\mathcal{H} \tag{50}$$

where

$$\mathcal{M}=\sum_{i=1}^{K}\frac{\epsilon^2 m_i}{2} \tag{51}$$

$$\mathcal{H}=\sum_{i=1}^{K}\left(\sum_{j\neq i}\left(\frac{1}{4}\epsilon^2\frac{q_i q_j}{|\vec{r}_i-\vec{r}_j|}\right)-\frac{1}{2m_i}\left(\vec{p}_i+q_i\sum_{j\neq k}\vec{A}_{ij}\right)^2\right). \tag{52}$$

by minimizing this loss function, the observer maximizes its predictive power. Once the observer minimizes this loss function, what we will refer to as "optimal learning", then $\mathcal{L}$ will be constant. The parameter values may still change as the observer learns from more measurements, but they must change while keeping $\mathcal{L}$ constant. In other words, at optimal learning, the loss function, $\mathcal{L}$, will act as a conserved quantity. Therefore, at optimal learning $\mathcal{M}+\mathcal{H}$ must be conserved.

An attentive reader will notice that our condition for optimal learning will lead to a conservation of mass, $\mathcal{M}$, and a conservation of energy for a system with a Hamiltonian, $\mathcal{H}$ that is identical (aside from scaling factors) to the Hamiltonian for charged particles in a vacuum. This is partly by choice as we made a number of modeling assumptions along the way to lead us to this point, but depending on how convincing our arguments for modeling choices are towards the reader, it may hint at a deeper connection between physics and learning.

In this section we have showed that an observer is able to maximize its predictive power by modeling continuous observables as charged particles in a vacuum. In the next section, apply similar logic to binary observables and show that equations of quantum mechanics lead to optimal learning.

2.2.3 Modeling binary observables as quantum spin states

COMING SOON

2.2.4 Modeling learning as an external force

COMING SOON

3 Discussion

In this work we have defined the problem of optimal learning, then showed that laws of physics are solutions to this problem. The advantage of this framework is that it allows us to confidently construct machine learning models and

training schemes from first principles rather than trial-and-error heuristics. We believe that the primary benefit from this work will come from its ability to create smarter AI algorithms. A pleasant secondary benefit is that it may shed light on some connection between physics and learning.

In particular, there has been much interest in relating non-equilibrium dynamics to inference [5]. We hope that our work helps add to this line of thought, since it shows rigorously that dynamical systems are solutions to learning. In a way, though, this result is almost obvious since computers and brains, which are clearly capable of learning, are constructed of protons, neutrons, and electrons interacting with nonlinear dynamics.

References

[1] Jeff Heaton. "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618". In: *Genetic programming and evolvable machines* 19.1-2 (2018), pp. 305–307.

[2] Frederick Reif. *Fundamentals of statistical and thermal physics*. 1998.

[3] Steve Pressé and Ioannis Sgouralis. *Data Modeling for the Sciences: Applications, Basics, Computations*. Cambridge University Press, 2023.

[4] Ariel Caticha. "The entropic dynamics approach to quantum mechanics". In: *Entropy* 21.10 (2019), p. 943.

[5] Agnish Kumar Behera et al. "Enhanced associative memory, classification, and learning with active dynamics". In: *Physical Review X* 13.4 (2023), p. 041043.