Noether's Razor: Learning Conserved Quantities

Tycho F. A. van der Ouderaa
Imperial College London
London, UK

Mark van der Wilk
University of Oxford
Oxford, UK

Pim de Haan
CuspAI
Amsterdam, NL

Abstract

Symmetries have proven useful in machine learning models, improving generalisation and overall performance. At the same time, recent advancements in learning dynamical systems rely on modelling the underlying Hamiltonian to guarantee the conservation of energy. These approaches can be connected via a seminal result in mathematical physics: Noether's theorem, which states that symmetries in a dynamical system correspond to conserved quantities. This work uses Noether's theorem to parameterise symmetries as learnable conserved quantities. We then allow conserved quantities and associated symmetries to be learned directly from train data through approximate Bayesian model selection, jointly with the regular training procedure. As training objective, we derive a variational lower bound to the marginal likelihood. The objective automatically embodies an Occam's Razor effect that avoids collapse of conservation laws to the trivial constant, without the need to manually add and tune additional regularisers. We demonstrate a proof-of-principle on n-harmonic oscillators and n-body systems. We find that our method correctly identifies the correct conserved quantities and U(n) and SE(n) symmetry groups, improving overall performance and predictive accuracy on test data.

1 Introduction

Symmetries provide strong inductive biases, effectively reducing the volume of the hypothesis space. A celebrated example of this is the convolutional layer embedding translation equivariance in neural networks, which can be generalised to other symmetry groups [Cohen and Welling, 2016].

Meanwhile, physics-informed machine learning models [Greydanus et al., 2019, Cranmer et al., 2020], typically relying on neural differential equations [Chen et al., 2018], embed constraints known from classical mechanics into model architectures to improve accuracy on physical dynamical systems.

Rather than strictly constraining a model to certain symmetries, recent works have explored whether invariance and equivariance symmetries in machine learning models can also be automatically learned from data. This often relies on separate validation data [Maile et al., 2022], explicit regularisers [Finzi et al., 2021] or additional outer loops [Cubuk et al., 2018]. Alternatively, we can take a Bayesian approach where we embed symmetries into the prior and empirically learn them through Bayesian model selection [van der Wilk et al., 2018, Immer et al., 2022, van der Ouderaa et al., 2022].

We propose to use *Noether's theorem* [Noether, 1918] to parameterise symmetries in Hamiltonian machine learning models in terms of their conserved quantities. To do so, we propose to symmetrise a learnable Hamiltonian using a set of learnable quadratic conserved quantitites. By choosing the conserved quantities to be quadratic, we can find closed-form transformations that can be used to obtain an unbiased estimate of the symmetrised Hamiltonian.

Secondly, we phase symmetries implied by conserved quantities in the prior over Hamiltonians an leverage the *Occam's razor* effect of Bayesian model selection [Rasmussen and Ghahramani, 2000, van der Wilk et al., 2018] to learn conserved quantities and their implied symmetries directly from train data. We derive a practical lower bound using variational inference [Hoffman et al., 2013]

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-4318

resulting in a single end-to-end training procedure capable of learning the Hamiltonian of a system jointly with its conserved quantities. As far as we know, this is the first case in which Bayesian model selection with variational inference is successfully scaled to deep neural networks, an achievement in its own right, whereas most works so far have relied on Laplace approximations [Immer et al., 2022].

Experimentally, we evaluate our *Noether's razor* method on various dynamical systems, including n-simple harmonic oscillators and n-body systems. Our results suggest that our method is indeed capable of learning the conserved quantities that give rise to correct symmetry groups for the problem at hand. Quantitatively, we find that our method that learns symmetries from data matches the performance of models with the correct symmetries built-in as oracle. We outperform vanilla training, resulting in improved test generalisation and predictions that remain accurate over longer time periods.

2 Background

2.1 Hamiltonian mechanics

Hamiltonian mechanics is a framework that describes dynamical systems in *phase space*, denoted $\mathcal{M} = \mathbb{R}^M$, with M even. Phase space elements $(\boldsymbol{q}, \boldsymbol{p}) \in \mathcal{M}$ follow *Hamiltonian equations of motion*:

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i} \tag{1}$$

where the Hamiltonian $H : \mathcal{M} \to \mathbb{R}$ is an *observable*[1], which are smooth functions on the phase space, that corresponds to the energy of the system. It is often simpler to write $\boldsymbol{x} = (\boldsymbol{q}, \boldsymbol{p})$, so that we have: $\dot{\boldsymbol{x}} = J\nabla H$ and $J = \left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$, where I is the identity matrix, J is called the *symplectic form* and $\nabla H = \nabla_{\boldsymbol{x}} H(\boldsymbol{x})$ is the gradient of phase space coordinates.

Example: n-body problem in 3d. If we consider a d=3 dimensional Euclidean space containing n bodies, our position and velocity spaces are each $\mathbb{R}^{3n}$ making up phase space $\mathcal{M} = \mathbb{R}^{2\cdot 3n}$. Our Hamiltonian $H : \mathbb{R}^{3n} \times \mathbb{R}^{3n} \to \mathbb{R}$, which in this case is a separable function $H(q,p) = K(q) + P(p)$ of kinetic energy $K(q) = \sum_i m_i ||p||^2/2$ and the potential energy $P(p) = \sum_{i \neq j} G m_i m_j / ||q_i - q_j||$ where m_i is the mass of a body i and G is the gravitational constant.

2.2 Learning Hamiltonian mechanics from data

We can model the Hamiltonian from data [Greydanus et al., 2019, Ross and Heinonen, 2023, Tanaka et al., 2022, Zhong et al., 2019]. Concretely, we are interested in a posterior over functions that the Hamiltonian can take $p(H_{\boldsymbol{\theta}} \mid \mathcal{D})$, conditioned on trajectory data $\mathcal{D} = \{(\boldsymbol{x}^n_t, \boldsymbol{x}^n_{t'})\}_{n=1}^N$ sampled from phase space at different time points (t, t'), or time difference Δt=t'−t. Given a new data point $\boldsymbol{x}^*_t$, we would like to make predictions $p(\boldsymbol{x}^*_{t'} | \boldsymbol{x}^*_t, H_{\boldsymbol{\theta}}, \mathcal{D})$ over phase space trajectories into the future t'.

Hamiltonian neural networks Hamiltonian neural networks [Greydanus et al., 2019, Toth et al., 2019, Rezende et al., 2019] model the Hamiltonian H using a learnable Hamiltonian $H_{\boldsymbol{\theta}} : \mathcal{M} \to \mathbb{R}$ parameterised by $\boldsymbol{\theta} \in \mathbb{R}^P$. With a straightforward Gaussian likelihood $p(\boldsymbol{x}_{t'} | \boldsymbol{x}_t, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_{t'} | \boldsymbol{x}_t + J\nabla H_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\Delta t, \sigma^2_{\text{data}}\boldsymbol{I})$ with a small observation noise σ^2_{data}, a maximum likelihood fit can be found by minimising the negative log-likelihood $\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \sum_i \sum_t - \log p(\boldsymbol{x}^i_{t+\Delta t} | \boldsymbol{x}^i_t, \boldsymbol{\theta})$ on minibatches of data using stochastic gradient descent. The mean of this likelihood represents a single Euler integration step (Sec. 2.1 of David and Méhats [2023]), which bounds the possible accuracy of the fit to the true Hamiltonian H. In practice, we may replace this by more accurate differentiable numerical integrators [Kidger, 2022].

2.3 Noether's theorem

The theorem of [Noether, 1918], here presented in the Hamiltonian formalism [Baez, 2020, Arnold, 1989], links the concepts of an observable being conserved, to the Hamiltonian being invariant to the symmetries generated by an observable.

[1] The term *observable* in classical mechanics should not be confused with the statistical notion of a variable being observed or not. In fact, we will model observables as latent variables that are not observed.

https://doi.org/10.52202/079017-4318

Conserved quantity Let $\mathcal{O}$ be the set of observables, which are smooth real-valued functions $\mathcal{M} \to \mathbb{R}$ on the phase space. Given a trajectory $\boldsymbol{x}(t)$ generated by the Hamiltonian H, we can compute the variation of an observable $O \in \mathcal{O}$ in time via the chain rule and Hamilton's equations of motion (Equation (1))

$$\frac{\mathrm{d}O}{\mathrm{d}t} = \sum_i \frac{\partial O}{\partial q_i}\dot{q}_i + \frac{\partial O}{\partial p_i}\dot{p}_i = \sum_i \frac{\partial O}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial O}{\partial p_i}\frac{\partial H}{\partial q_i} = \{O, H\}, \tag{2}$$

where the last equality defines the *Poisson bracket* $\{\cdot,\cdot\} : \mathcal{O} \times \mathcal{O} \to \mathcal{O}$. The Poisson bracket relates to the symplectic form via $\{O, H\}(\boldsymbol{x}) = \nabla O(\boldsymbol{x}) \cdot J\nabla H(\boldsymbol{x})$. An observable that does not change along any trajectory is called a *conserved quantity*. As we can see from Equation (2), an observable O is conserved if and only if $\{O, H\} = 0$.

From two conserved quantities $O, O' \in \mathcal{O}$, we can create a new conserved quantity by linear combination $\alpha O + \beta O' \in \mathcal{O}$ with coefficients for $\alpha, \beta \in \mathbb{R}$, which is conserved because the Poisson bracket is linear in both arguments. Also, we can take the product $OO' \in \mathcal{O}$, with $(OO')(\boldsymbol{x}) = O(\boldsymbol{x})O'(\boldsymbol{x})$, which is conserved because the Poisson bracket satisfies Leibniz's law of differentiation $\{OO', H\} = \{O, H\}O' + O\{O', H\}$. Finally, the Poisson bracket of the conserved quantities $\{O, O'\} \in \mathcal{O}$ is also conserved, because of the Jacobi identity.

Symmetries generated by observables Referring back as to the Hamiltonian equations of motion in Equation (1), note that these equations work not just for the Hamiltonian $H \in \mathcal{O}$ of the system, but for *any* observable $O \in \mathcal{O}$. So given any starting point x_0, we can generate a trajectory $\boldsymbol{x}(\tau)$ satisfying

$$\boldsymbol{x}(0) = \boldsymbol{x}_0 \qquad\qquad \dot{\boldsymbol{x}}(\tau) = J\nabla O(\boldsymbol{x}(\tau)). \tag{3}$$

We have used a different symbol to not conflate the ODE time τ with regular time t of the trajectory generated by the Hamiltonian. Denote the flow associated to this ODE generated by observable O by $\Phi_O^\tau : \mathcal{M} \to \mathcal{M}$, mapping $\boldsymbol{x}_0$ to $\Phi_O^\tau(\boldsymbol{x}_0) = \boldsymbol{x}(\tau)$. Note that any ODE flow satisfies $\Phi_O^0 = \mathrm{id}_\mathcal{M}$ and $\Phi_O^{\tau+\kappa} = \Phi_O^\tau \circ \Phi_O^\kappa$. Hence, the observable O generates a one-dimensional group $\mathcal{G}_O$, parametrized by τ, that is a subgroup of the group $\mathrm{Diff}(\mathcal{M})$ of diffeomorphisms $\mathcal{M} \to \mathcal{M}$.

Theorem 1 (Noether)**.** *The observable $O \in \mathcal{O}$ is a conserved quantity on the trajectories generated by Hamiltonian $H \in \mathcal{O}$ if and only if H is invariant to $\mathcal{G}_O$, meaning that for all $\tau \in \mathbb{R}$, $H \circ \Phi_O^\tau = H$.*

Proof. By reasoning analogous to that in Equation (2), the value of the Hamiltonian changes under the flow generated by observable O as $\frac{\mathrm{d}H}{\mathrm{d}\tau} = \{H, O\}$. Noting that the Poisson bracket is anti-symmetric, we have that: O is a conserved quantity $\iff \{O, H\} = 0 \iff \{H, O\} = 0 \iff H$ is invariant to the flow generated by O. □

2.4 Automatic symmetry discovery

Symmetries play an important role in machine learning models, most notably group invariance and equivariance constraints [Cohen and Welling, 2016]. Instead of having to define symmetries explicitly in advance, recent attempts have been made to learn symmetries automatically from data. Even if learnable symmetries can be differentiably parameterised, learning them can remain difficult as symmetries act as constraints on the functions a model can represent and are, therefore, not encouraged by objectives that solely optimise train data fit. As a result, even if a symmetry would lead to better test generalisation, the training collapses into selecting no symmetry. Common ways to overcome this are designing explicit regularisers that encourage symmetry [Benton et al., 2020, van der Ouderaa et al., 2022], which often require tuning, or use of validation data [Alet et al., 2021, Maile et al., 2022, Zhou et al., 2020]. Learning symmetries for integrable systems was proposed in [Bondesan and Lamacraft, 2019], whereas our framework works more generally also for non-integrable systems, such as the 3-body problem. Recent works have demonstrated effectivity of Bayesian model selection to learn symmetries directly from training data. This works by optimising the marginal likelihood, which embodies an Occam's razor effect that trades off data fit and model complexity. For Gaussian processes, the quantity can often be computed in closed-form [van der Wilk et al., 2018], and can be scaled to neural networks through variational inference [van der Ouderaa and van der Wilk, 2021] and linearised Laplace approximations [Immer et al., 2022].

3 Symmetrising Hamiltonians with Conserved Quantities

Our method introduced in the next section will learn the Hamiltonian of a system together with a set of conserved quantities. First, in this section we discuss how the learned conserved quantities will be parametrised, and how we can make the Hamiltonian invariant to the symmetry generated by conserved quantities.

3.1 Parameterising conserved quantities

In this work, we limit ourselves to modelling up to a fixed maximum number of K conserved quantities $C^1_{\boldsymbol{\eta}}, C^2_{\boldsymbol{\eta}}, \ldots, C^K_{\boldsymbol{\eta}} : \mathcal{M} \to \mathbb{R}$ are observables parameterised by *symmetrisation parameters* $\boldsymbol{\eta}$, to distinguish them from the *model parameters* $\boldsymbol{\theta}$ parameterising the Hamiltonian scalar field.

In this paper, we consider quadratic conserved quantities of the form $C_{\boldsymbol{\eta}}(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}/2 + \boldsymbol{b}^T\boldsymbol{x} + c$. As we use the conserved quantities only through their gradients, the constant is arbitrary and can be ignored. The learnable symmetrisation parameters are thus $\boldsymbol{\eta} = \{\boldsymbol{A}, \boldsymbol{b}\}$, for a symmetric matrix $\boldsymbol{A}$. A quadratic conserved quantity C generates a symmetry transformation whose scalar field $\dot{\boldsymbol{x}} = J\nabla C(\boldsymbol{x}) = J\boldsymbol{A}\boldsymbol{x} + J\boldsymbol{b}$ is affine, or linear on the homogeneous coordinates $(\boldsymbol{x}, 1)$. Its flow can be analytically solved

$$\begin{bmatrix} \Phi^\tau_C(\boldsymbol{x}) \\ 1 \end{bmatrix} = \exp\left(\tau \begin{bmatrix} J\boldsymbol{A} & J\boldsymbol{b} \\ \boldsymbol{0}^T & 0 \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} \tag{4}$$

using the matrix exponential $\exp(\cdot)$ for which efficient numerical algorithms exist [Moler and Van Loan, 2003]. This equation can be verified to have the correct scalar field and boundary condition, and thus forms the unique solution to the ODE in Equation (3).

3.2 Symmetrising observables

Given an observable $C \in \mathcal{O}$, we want to transform an observable f into $\hat{f}$ that is invariant to the transformations generated by C. This means that $\hat{f} \circ \Phi^\tau_C = \hat{f}$ for all symmetry time $\tau \in \mathbb{R}$. Via Noether's theorem, we know that this is equivalent to C being conserved in the trajectories generated by f, and also equivalent to $\{C, \hat{f}\} = 0$. However, this equation does not prescribe how to obtain such $\hat{f}$. Instead, we'll create $\hat{f}$ by symmetrizing over the symmetry group generated by C. This is done by averaging over the orbit of the transformation

$$\hat{f}(\boldsymbol{x}) = \int_{\mathbb{R}} f(\Phi^\tau_C(\boldsymbol{x}))\mu(\tau).$$

with a measure μ over symmetry time τ. This measure μ induces a measure on the 1-dimensional subgroup $\mathcal{G}_C$ of the group of diffeomorphisms $\mathcal{M} \to \mathcal{M}$. If this measure on $\mathcal{G}_C$ is uniform (specifically, a right-invariant measure [Halmos, 1950]), then $\hat{f}$ is indeed invariant.

Instead of a single symmetry generator, we can also have a set $\mathcal{C} = \{C_1, ..., C_K\}$ of observables and we want to make f invariant to all of these. Assume that this set spans a vector space of observables that is closed under the Poisson bracket (i.e. they form a Lie subalgebra). In that case, the groups of transformations of the observables combined generate a group $\mathcal{G}_{\mathcal{C}}$ [Hall, 2015, Thm. 5.20]. This group is parameterized by a vector of symmetry times $\boldsymbol{\tau} \in \mathbb{R}^K$. The corresponding flow is $\Phi^{\boldsymbol{\tau}}_{\mathcal{C}} = \Phi^1_{\sum_i \tau_i C_i}$. To make an observable f invariant to the symmetries of all conserverved quantities $\mathcal{C}$, equivalently to the group $\mathcal{G}_{\mathcal{C}}$, we symmetrize

$$\hat{f}(\boldsymbol{x}) = \int_{\mathbb{R}^K} f(\Phi^{\boldsymbol{\tau}}_{\mathcal{C}}(\boldsymbol{x}))\mu(\boldsymbol{\tau}). \tag{5}$$

with some measure μ over $\mathbb{R}^K$. As before, if this induces a uniform measure over $\mathcal{G}_{\mathcal{C}}$, then this symmetrization indeed makes $\hat{f}$ invariant to $\mathcal{G}_{\mathcal{C}}$.

However, a probability measure $\mu(\boldsymbol{\tau})$ that gives a uniform distribution over $\mathcal{G}_{\mathcal{C}}$ might not exist, for example when the group contains a non-compact group of translations. Even when such a measure does exist, it may be hard to construct, and the symmetrisation integral in Equation (5) may be intractable to compute. So instead, in practice, we approximate this by choosing $\mu(\boldsymbol{\tau})$ to be a unit normal distribution $\mathcal{N}(0, \boldsymbol{I}_K)$ or uniform distribution. This results in a relaxed notion of symmetry

https://doi.org/10.52202/079017-4318

in $\hat{f}$ which can be interpreted as a form of robustness to actions of the symmetry group implied by the conserved quantity, by smoothing the function in this direction around data, in contrast to strict invariance by definition closed under group actions along the full orbit. Finally, we approximate the integral by an unbiased Monte Carlo estimate with S samples.

4 Automatic Symmetry Discovery using Noether's Razor

Now that we have a way of parameterising symmetry differentiably as conservation laws through Noether's theorem, we need an objective function that is capable of selecting the right symmetry. Unfortunately, regular training objectives that only rely on data fit can not necessarily distinguish the correct inductive bias, as noted in prior work [van der Wilk et al., 2018, Immer et al., 2022, van der Ouderaa and van der Wilk, 2021]. This is because, even if train data originates from a symmetric distribution, there can be both non-symmetric and symmetric solutions that fit the train data equally well, given a sufficiently flexible model. Consequently, the regular maximum likelihood objective that only measures train data fit will not necessarily favour a symmetric model, even if we expect this to generalise best on test data. Instead of having to resort to cross-validation to select the right symmetry inductive bias, we propose to use an approximate marginal likelihood on the train data. This has the additional benefit of being differentiable, allowing symmetrisation to be learned with back-propagation along with regular parameters in a single training procedure. In our case, we use *Noether's theorem* to parameterise symmetries in our prior through conserved quantities, which we can optimise with back-propagation using a differentiable lower bound on the marginal likelihood. This quantity, also known as the 'evidence', differs distinctly from maximum likelihood in that it balances both train fit as well as model complexity. The *Occam's razor* effect encourages symmetry and leverages the symmetrisation process to 'cut away' prior density over Hamiltonian functions that are not symmetric, if this does not result in a worse data fit. The resulting posterior predictions automatically becomes symmetric if observed data obeys a symmetry (high evidence for symmetry), but can become non-symmetric if this does not match the data (low evidence for symmetry). Hence, the name of our proposed method for automatic inductive bias selection is *Noether's razor*.

4.1 Probabilistic model with symmetries embedded in the prior.

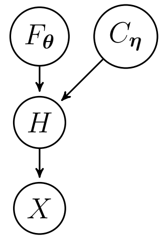

Figure 1: Graphical probabilistic model. Trajectory data X depends on a symmetrised Hamiltonian H induced by non-symmetrised observable F and conservation laws C.

To be more explicit about our probabilistic model, we can introduce four variables, namely a non-symmetrised observable $F_{\boldsymbol{\theta}}$, a set of conserved quantities $C_{\boldsymbol{\eta}}$, which induce a symmetrised Hamiltonian H generating the observed trajectory data X. We treat trajectory data an an observed variable, consider the conserved quantities as part of an empirical prior as we optimise over them, and integrate out the Hamiltonian as latent. The construction can be interpreted a placing a sophisticated prior over the functions that the symmetrised Hamiltonian H can represent, which is the variable of primary interest. The underlying non-symmetrised $F_{\boldsymbol{\theta}}$ does not have a direct physical meaning as H does, but defines a prior over neural networks to flexibly define a density over a rich class of possible functions. The conserved quantities $C_{\boldsymbol{\eta}}$ control the amount of symmetry in the effective prior over symmetrised Hamiltonians H. Empirically optimising $C_{\boldsymbol{\eta}}$ through Bayesian model selection allows us to 'cut away' density in the prior over H that correspond to functions that are not symmetric - as the symmetrisation averages functions in $F_{\boldsymbol{\theta}}$ that lie in the same orbit and thereby increases the relative density of symmetric functions in H. We hypothesise that we will not over-fit conserved quantities as $\boldsymbol{\eta}$ is relatively low-dimensional, only representing quadratic functions, while we integrate out the high-dimensional neural network model parameters $\boldsymbol{\theta}$ that parameterises the observable $F_{\boldsymbol{\theta}}$. In future work, it would be interesting to explore a richer function classes for conserved quantities, such as neural networks, although we do expect this to be more difficult and to require additional priors or regularisation techniques to avoid over-fitting.

https://doi.org/10.52202/079017-4318

4.2 Bayesian model selection for symmetry discovery

To learn the right symmetry from data, we propose to use Bayesian model selection through optimisation of the marginal likelihood. In the previous sections, we have phrased symmetries parameterised by $\boldsymbol{\eta}$ as part of the prior over Hamiltonians. The symmetry parameters $\boldsymbol{\eta}$ parameterise the space of possible 'models' that we consider, whereas the model parameters $\boldsymbol{\theta}$ parameterise the weights of a single model. To perform Bayesian model selection on the symmetries, we are interested in computing the marginal likelihood:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{6}$$

which requires integrating (marginalising) the likelihood over model parameters $\boldsymbol{\theta}$ weighted by the prior, and is sometimes referred to as the 'evidence' for a particular model. Unlike maximum likelihood, the marginal likelihood has an Occam's razor effect [Smith and Spiegelhalter, 1980, Rasmussen and Ghahramani, 2000] that balancing both data fit and model complexity, allowing optimisation of symmetry parameters $\boldsymbol{\eta}$. Although the marginal likelihood is typically intractable, certain approximate Bayesian inference techniques can provide differentiable estimates. In the next sections, we will use variational inference to derive a tractable and differentiable lower bound to the marginal likelihood that can be used to find a posterior over $\boldsymbol{\theta}$ and optimise symmetries $\boldsymbol{\eta}$.

Why the marginal likelihood can learn symmetry To understand why the marginal likelihood objective is capable of learning the right symmetry (to learn $\boldsymbol{\eta}$), Sec. 3.2 [van der Wilk et al., 2018] proposed to decompose it through the product rule:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = p(\boldsymbol{x}_1 \mid \boldsymbol{\eta}) p(\boldsymbol{x}_2|\boldsymbol{x}_1, \boldsymbol{\eta}) p(\boldsymbol{x}_3 \mid \boldsymbol{x}_{1:2}, \boldsymbol{\eta}) \prod\nolimits_{c=4}^{C} p(\boldsymbol{x}_c \mid \boldsymbol{x}_{1:c-1}, \boldsymbol{\eta}) \tag{7}$$

which shows that the marginal likelihood measures how much parts of the dataset predict other parts of the data - a measure of generalisation that does not require cross-validation. Given a perfect data fit, the marginal likelihood will be higher when the right symmetry is selected, as parts of the dataset will result in better and more certain predictions on other part of the data. This is unlike the maximum likelihood, which is always maximised with perfect data fit, with or without the right symmetry. For some posterior approximations, such as linearised Laplace approximations, it can be analytically shown that symmetry maximises the approximate marginal likelihood (App. G.2 of Immer et al. [2022]). Our method is very similar, but uses more expressive variational inference which can optimise the posterior globally, rather than relying on a local Taylor expansion.

4.3 Lower bounding the marginal likelihood

The marginal likelihood of an Hamiltonian neural network is typically not tractable in closed-form. However, we can derive a lower bound to the marginal likelihood using variational inference (VI):

$$\begin{aligned} \log p(\boldsymbol{x} \mid \boldsymbol{\eta}) &\geq \mathbb{E}_{\boldsymbol{\theta}} \left[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) \right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta})) \\ &\geq \mathbb{E}_{\boldsymbol{\theta}} \left[\mathbb{E}_{\boldsymbol{\tau}} \left[\sum\nolimits_{i=1}^{N} \log \mathcal{N}(\boldsymbol{x}_{t'}^i \mid \widehat{H}_{\boldsymbol{\theta},\boldsymbol{\eta}}^{\boldsymbol{\tau}}(\boldsymbol{x}_t^i), \sigma_{\text{data}}^2 \boldsymbol{I}) \right] \right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{0}, \sigma_{\text{prior}}^2 \boldsymbol{I})) \end{aligned} \tag{8}$$

where $\widehat{H}_{\boldsymbol{\theta},\boldsymbol{\eta}}^{\boldsymbol{\tau}}(\boldsymbol{x}_t^i) = \frac{1}{S} \sum\nolimits_{s=1}^{S} H_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{\Phi}_{\boldsymbol{\eta}}^{\boldsymbol{\tau}^{(s)}}(\boldsymbol{x}_t^i))$ and $\widehat{H}$ is an unbiased S-sample Monte Carlo estimator of the symmetrised Hamiltonian. We write $\mathbb{E}_{\boldsymbol{\theta}} := \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{m},\boldsymbol{S}}}$ and $\mathbb{E}_{\boldsymbol{\tau}} = \mathbb{E}_{\boldsymbol{\tau} \sim \prod_{s=1}^{S} \mu(\boldsymbol{\tau})}$ for which we can obtain an unbiased estimate by taking Monte Carlo samples. The first inequality is the standard VI lower bound. The second inequality follows from applying Jensen's inequality (again) which uses the fact that the log likelihood is a convex function. Similar lower bounds to invariant models that average over a symmetry group have recently appeared in prior work [van der Ouderaa and van der Wilk, 2021, Schwöbel et al., 2022, Nabarro et al., 2022]. Full derivation in Appendix A.1.

4.4 Improved variational inference for scalable Bayesian model selection

Variational inference is a common tool to perform Bayesian inference on models with intractable marginal likelihoods, including neural networks. In deep learning literature, however, its use is typically limited to better predictive uncertainty estimation and rarely for Bayesian model selection. Meanwhile, linearised Laplace approximations have recently been successfully applied to Bayesian

https://doi.org/10.52202/079017-4318

model selection [Immer et al., 2021] and symmetry learning in specific [Immer et al., 2022, van der Ouderaa et al., 2024], with a few reported cases of model selection using VI only in single neural network layers [van der Ouderaa and van der Wilk, 2021, Schwöbel et al., 2021]. Optimising Bayesian neural networks with variational inference is much less established than training regular neural networks, for which many useful heuristics are available. This work, however, provides evidence that it is also possible to perform approximate Bayesian model selection using VI in deep neural networks, which we deem an interesting observation in its own right. To make sure the lower bound on the marginal likelihood is sufficiently tight, we employ a series of techniques, including a richer non-mean field family of matrix normal posteriors [Louizos and Welling, 2016], and closed-form updates of the prior precision and output variances derived with expectation maximisation. Details on how we train a Bayesian neural network using variational inference can be found in Appendix D.

5 Results

In this section, we will discuss how the learned symmetries are analysed and then list our experiments and results.

5.1 Analyzing learned symmetries

In our experiments, we will find a set of K conserved quantities $C_k : \mathcal{M} \to \mathbb{R}$. As we consider quadratic conserved quantities in particular, we can equivalently analyze the resulting generators of the associated symmetries $\hat{G}_k(\boldsymbol{x}) = J\nabla C_k$ which are affine and thus representable with a matrix $\hat{G}_k \in \mathbb{R}^{(M+1)\times(M+1)}$ on homogeneous coordinates $(\boldsymbol{x}, 1)$. In Appendix B, we list for each system the L ground truth conserved quantities generators $G_l^\star$. The learned and ground truth generators can be stacked in to the matrices $\hat{\boldsymbol{G}} \in \mathbb{R}^{K\times(M+1)^2}, \boldsymbol{G}^* \in \mathbb{R}^{L\times(M+1)^2}$ respectively. As we can identify the symmetries only up to linear combinations, we have learned the correct symmetries if the learned generators span a linear subspace of $\mathbb{R}^{(M+1)^2}$ that coincides with the space spanned by the ground truth generators. To verify this, we test two properties. First, we show that the matrix $\hat{\boldsymbol{G}}$ has L non-zero singular values. Secondly, for the first L right singular vectors $v_i \in \mathbb{R}^{(M+1)^2}$, we decompose $v_i = v_i^{\parallel} + v_i^{\perp}$ in a vector in ground truth subspace, and one orthogonal to it. The learned v_i is a correct conserved quantity if $v_i^{\perp} = 0$, or equivalently, because the singular vectors are normalized, if $\|v_i^{\parallel}\| = 1$. We call this measure the "parallelness".

5.2 Simple Harmonic Oscillator

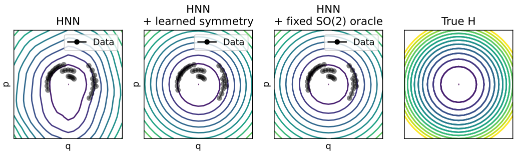

Figure 2: Learned Hamiltonians on phase space of simple harmonic oscillator by HNN models.

We start with a demonstration on the simple harmonic oscillator. This text book example has a 2-dimensional phase space, making learned Hamiltonians amenable to visualisation. Further, it has a clear rotational symmetry SO(2), relating to the conserved phase. On a finite set of generated train data, we model the Hamiltonian using a vanilla HNN, our symmetry learning method, and a model with true symmetry built-in as reference oracle (experimental details in Appendix B.1). In Figure 2, we find that our symmetry learning method results in a rotationally invariant Hamiltonian that matches the fixed rotational SO(2) symmetry. Further away from the origin, the learned Hamiltonian differs from the ground truth Hamiltonian, as there is no data in that region. In Table 1, we find that the learned symmetry has a better ELBO on the train set and matches the improved predictive performance of the model with the correct symmetry built-in. The symmetry learning method outperforms the vanilla model in terms of predictive performance on the test set.

Table 1: Learning Hamiltonian dynamics of the simple harmonic oscillator. We compare a vanilla HNN, our symmetry learning method, and a model with the correct SO(2) symmetry built-in as reference oracle. Our method achieves reference oracle performance, indicating correct symmetry learning, and outperforms the vanilla model by improving predictive performance on the test set.

Learned dynamics:		Train data				Test data
simple harmonic oscillator		Train MSE	NLL/N	KL/N	-ELBO/N (↓)	Test MSE (↓)
HNN		0.005	0.3667	3314.374	3314.741	0.005
HNN + learned symmetry	(**ours**)	0.002	-2.618	3304.754	**3302.136**	**0.002**
HNN + fixed SO(2)	(reference oracle)	0.002	-3.213	3298.357	**3295.144**	**0.002**

5.3 $n-$Harmonic Oscillators

Now, we consider $n-$harmonic oscillators. This system has as symmetry group the unitary Lie group U(n) of dimensionality of n^2 (see Appendix B.2). We sample random trajectories from phase space and train a HNN neural network without and with symmetry learning using variational inference. Again, we find improved ELBO and test performance for learned symmetries Table 2. Following the protocol from Section 5.1, we analyze the learned symmetries. In Figure 3 (right), we see that for varying n, we indeed find that the matrix of learned symmetries has n^2 nonzero singular values. Furthermore, the first n^2 singular vectors lie in the ground truth subspace of generators with measured parallelness $\|v_i^{\|}\| > 0.99$, as seen in Figure 3 (left). This shows that the $U(n)$ symmetry is corectly learned.

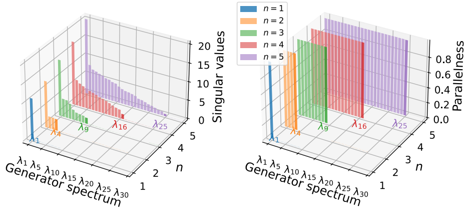

Figure 3: Singular value and parallelness of the singular vectors of the learned generators, for n oscilators. U(n) is correctly learned.

Table 2: Learning Hamiltonian dynamics of 3−fold harmonic oscillators. We compare HNN with symmetry learning to a vanilla HNN without symmetry learning and to the correct U(3) symmetry built-in as fixed reference oracle. We find that our method can discover the correct symmetry, achieves reference oracle performance, and outperforms vanilla training in both ELBO and test performance.

Learned dynamics:		Train data				Test data
simple harmonic oscillator		Train MSE	NLL/N	KL/N	-ELBO/N (↓)	Test MSE (↓)
HNN		0.00106	-12.04	5.27	-6.77	0.00002141
HNN + learned symmetry	(**ours**)	0.00102	-12.16	2.53	**-9.63**	**0.00000994**
HNN + fixed symmetry U(n)	(reference oracle)	0.00102	-12.15	2.21	**-9.94**	**0.00000898**

5.4 n-Body System

To investigate performance of our method on more interesting systems, we consider learning the Hamiltonian of an n-body system with gravitational interaction. We use 3 bodies in 2 dimensions so that trajectories and generators remain easy to visualise. As the Hamiltonian depends only on the norm of the momenta and on positions via the relative distances of the bodies, the three dimensional group SE(2) of rototranslations is an invariance of the ground truth Hamiltonian. However, as explained in Appendix B.3, the Hamiltonian has four more quadratic conserved quantities. They generate a 7-dimensional Lie group $\mathcal{G}$ of symmetries. This group has the same orbits on the phase space as SE(2). Therefore, a function being invariant to SE(2) is equivalent to it being invariant to $\mathcal{G}$. We'll find that Noether's razor discovers not just SE(2), but all seven generators of $\mathcal{G}$.

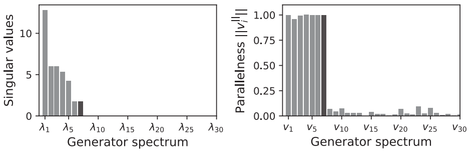

Figure 4: Singular value and parallelness of the singular vectors of the learned generators for three body system in two dimensions. The 7-dimensional Lie group $\mathcal{G}$ of quadratic conserved quantities is correctly learned.

https://doi.org/10.52202/079017-4318

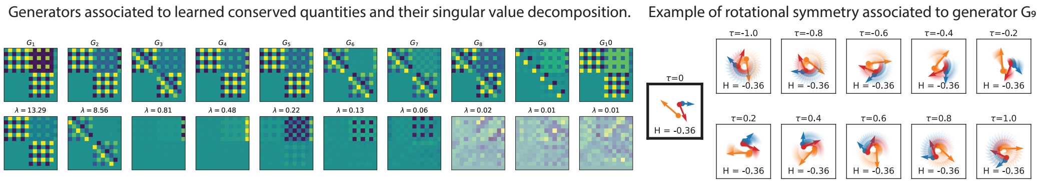

Figure 5: Learned generators associated by conserved quantities and their singular value decomposition. We find a subspace spanned by the 7 linear generators that correspond to the correct symmetries (see Appendix B.3): (1) rotation of the center of mass R^{COM}, (2) rotation around the origin R^{ABS}, (4+5) translation, (5+6+7) momentum-dependent translations P, Q, (8+9+10) inactive ($\lambda < 0.05$). The first 7 singular vectors lie in the ground truth subspace of generators with measured parallelness $||v_i^{||}|| > 0.95$.

In Table 3, we compare performance of a vanilla variational HNN with our symmetry learning approach and a model that has the appropriate SE(2) symmetry of rototranslations built-in as an reference oracle. We find that our method is able to *automatically discover* the conserved quantities and associated generators that span the symmetry group. The model achieves the same performance as the model with the symmetry built-in as reference oracle, but without having required the prior knowledge. Compared to the vanilla baseline, our approach improves test accuracy on both in-distribution as out-of-distribution test sets.

Table 3: Learning Hamiltonian dynamics of 2d 3-body system with variational Hamiltonian neural networks (HNN). We compare our symmetry learning method to a vanilla model without symmetry learning and a model with the correct SE(2) symmetry built-in as a reference oracle. Our method capable of discoverying symmetry achieves the oracle performance, outperforming the vanilla method.

Learned dynamics:		Train data				Test data	Test data (moved)	Test data (wider)
2d 3-body system		Train MSE	NLL/N	KL/N	-ELBO/N (↓)	Test MSE (↓)	Test MSE (↓)	Test MSE (↓)
HNN		0.0028	-13.87	13.34	-9.52	0.0016	0.0035	0.0016
HNN + learned symmetry	(**ours**)	0.0017	-20.09	7.28	**-12.81**	**0.0006**	**0.0004**	**0.0006**
HNN + fixed SE(2)	(reference oracle)	0.0019	-19.27	7.96	**-11.32**	**0.0006**	**0.0006**	**0.0006**

After training, we can analyse the learned conserved quantities and implied symmetries by inspecting their associated generators. In Figure 5, we plot these generators as well as their singular value decomposition. We see that our method correctly learns 7 singular values with $\lambda_i > 0.05$ and the associated singular vectors lie in the ground truth subspace with $\|v_i^{||}\| > 0.95$. This indicates that our method is in fact capable of inferring the right symmetries from train data, beyond merely improving generalisation by improving predictive performance on the test set.

6 Conclusion

In this work, we propose to use *Noether's theorem* to parameterise symmetries in machine learning models of dynamical systems in terms of conserved quantities. Secondly, we propose to leverage the *Occam's razor* effect of Bayesian model selection by phrasing symmetries implied by conserved quantities in the prior and learning them by optimising an approximate marginal likelihood directly on train data, which does not require validation data or explicit regularisation of the conserved quantities. Our approach, dubbed *Noether's razor*, encourages symmetries by balancing both data fit and model complexity. We derive a variational lower bound on the marginal likelihood providing a concrete objective capable of jointly learning the neural network as well as the conserved quantities that symmetrise the Hamiltonian. As far as we know, this is also the first time differentiable Bayesian model selection using variational inference has been demonstrated on deep neural networks. We demonstrate our approach on n-harmonic oscillators and n-body systems. We find that our method learns the correct conserved quantities by analysing the singular values and correctness of the subspace spanned by the generators implied by learned conserved quantitites. Further, we find that our method performs on-par with models with the true symmetries built-in explicitly and we outperform vanilla model, improving generalisation and predictive accuracies on test data.

https://doi.org/10.52202/079017-4318

References

Ferran Alet, Dylan Doblar, Allan Zhou, Josh Tenenbaum, Kenji Kawaguchi, and Chelsea Finn. Noether networks: meta-learning useful conserved quantities. Advances in Neural Information Processing Systems, 34, 2021.

Jean-Pierre Amiet and Stefan Weigert. Commensurate harmonic oscillators: Classical symmetries. J. Math. Phys., 43(8):4110–4126, August 2002. URL http://dx.doi.org/10.1063/1.1488672.

V I Arnold. Mathematical methods of classical mechanics, 1989.

John C Baez. Getting to the bottom of noether's theorem, 2020. URL http://arxiv.org/abs/2006.14741.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks. arXiv preprint arXiv:2010.11882, 2020.

Roberto Bondesan and Austen Lamacraft. Learning symmetries of classical integrable systems. arXiv preprint arXiv:1906.04645, 2019.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in neural information processing systems, 31, 2018.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In International conference on machine learning, pages 2990–2999. PMLR, 2016.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. arXiv preprint arXiv:2003.04630, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501, 2018.

Marco David and Florian Méhats. Symplectic learning for hamiltonian neural networks. Journal of Computational Physics, 494:112495, 2023.

Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. Advances in Neural Information Processing Systems, 34, 2021.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. Advances in neural information processing systems, 32, 2019.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In International Conference on Machine Learning, pages 573–582. PMLR, 2016.

Brian C Hall. Lie Groups, Lie Algebras, and Representations. Springer International Publishing, 2015. URL https://link.springer.com/book/10.1007/978-3-319-13467-3.

Paul R Halmos. Measure Theory. Springer New York, 1950. URL https://link.springer.com/book/10.1007/978-1-4684-9440-2.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR (Poster), 3, 2017.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. Journal of Machine Learning Research, 2013.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In International Conference on Machine Learning, pages 4563–4573. PMLR, 2021.

Alexander Immer, Tycho F. A. van der Ouderaa, Vincent Fortuin, Gunnar Rätsch, and Mark van der Wilk. Invariance learning in deep neural networks with differentiable laplace approximations, 2022.

Patrick Kidger. On neural differential equations. arXiv preprint arXiv:2202.02435, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In International conference on machine learning, pages 1708–1716. PMLR, 2016.

Kaitlin Maile, Dennis George Wilson, and Patrick Forré. Equivariance-aware architectural optimization of neural networks. In The Eleventh International Conference on Learning Representations, 2022.

Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM review, 45(1):3–49, 2003.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in bayesian neural networks and the cold posterior effect. In Uncertainty in Artificial Intelligence, pages 1434–1444. PMLR, 2022.

E. Noether. Invariante variationsprobleme. Königlich Gesellschaft der Wissenschaften Göttingen Nachrichten Mathematik-physik Klasse, 1918.

Carl Rasmussen and Zoubin Ghahramani. Occam's razor. Advances in neural information processing systems, 13, 2000.

Danilo Jimenez Rezende, Sébastien Racanière, Irina Higgins, and Peter Toth. Equivariant hamiltonian flows. arXiv preprint arXiv:1909.13739, 2019.

Magnus Ross and Markus Heinonen. Learning energy conserving dynamics efficiently with hamiltonian gaussian processes. arXiv preprint arXiv:2303.01925, 2023.

Pola Schwöbel, Martin Jørgensen, Sebastian W Ober, and Mark van der Wilk. Last layer marginal likelihood for invariance learning. arXiv preprint arXiv:2106.07512, 2021.

Pola Schwöbel, Martin Jørgensen, Sebastian W Ober, and Mark Van Der Wilk. Last layer marginal likelihood for invariance learning. In International Conference on Artificial Intelligence and Statistics, pages 3542–3555. PMLR, 2022.

Adrian FM Smith and David J Spiegelhalter. Bayes factors and choice criteria for linear models. Journal of the Royal Statistical Society: Series B (Methodological), 42(2):213–220, 1980.

Yusuke Tanaka, Tomoharu Iwata, et al. Symplectic spectrum gaussian processes: Learning hamiltonians from noisy and sparse data. Advances in Neural Information Processing Systems, 35: 20795–20808, 2022.

Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. arXiv preprint arXiv:1909.13789, 2019.

Tycho van der Ouderaa, Alexander Immer, and Mark van der Wilk. Learning layer-wise equivariances automatically using gradients. Advances in Neural Information Processing Systems, 36, 2024.

Tycho FA van der Ouderaa and Mark van der Wilk. Learning invariant weights in neural networks. In Workshop in Uncertainty & Robustness in Deep Learning, ICML, 2021.

Tycho FA van der Ouderaa, David W Romero, and Mark van der Wilk. Relaxing equivariance constraints with non-stationary continuous filters. arXiv preprint arXiv:2204.07178, 2022.

Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. arXiv preprint arXiv:1808.05563, 2018.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. arXiv preprint arXiv:1909.12077, 2019.

Allan Zhou, Tom Knowles, and Chelsea Finn. Meta-learning symmetries by reparameterization. arXiv preprint arXiv:2007.02933, 2020.

https://doi.org/10.52202/079017-4318

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper proposes a new method that can automatically discover correct symmetries and improved predictive performance on hold-out test data. This is justified by experiments that measure the correctness of symmetries and evaluate test performance.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper states modelling assumptions in terms of used likelihood and prior. Further, the paper is restricted to fairly small scale experiments.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-4318

Answer: [Yes]

Justification: All theoretical results provides required assumptions and a complete proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes. For each dataset, we provide a detailed description how data points were generated in Appendix C. For each experiment, we provide training details and important hyperparameters in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-4318

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Code will be published upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: Yes.

Justification: See Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: We report negative log likelihoods, ELBO, test predictive performance over datasets after optimising the ELBO for thousands of epochs. Train and test scores are computed over the full dataset, but not repeated for multiple seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-4318

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on a single NVIDIA RTX 4090 GPU with 24GiB of GPU memory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper respects the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is a foundational paper not tied to a particular application.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: This paper does not pose such risk and experiments only consider very simple text book problems.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The datasets used in this model are new, but details are clearly described in App. C.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-4318

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-4318

A Mathematical derivations

A.1 ELBO of Hamiltonian Neural Network

We can find a lower bound on the marginal likelihood $\log p(\boldsymbol{x} \mid \boldsymbol{\eta})$ through variational inference,

$$
\begin{aligned}
\log p(\boldsymbol{x} \mid \boldsymbol{\eta}) &\geq \mathbb{E}_{\boldsymbol{\theta}} \left[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\eta})\right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta})) && (9) \\
&= \mathbb{E}_{\boldsymbol{\theta}} \left[\sum\nolimits_{i=1}^{N} \log \mathcal{N}(\boldsymbol{x}_{t'}^{i} \mid H_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_t^i), \sigma^2_{\text{data}}\boldsymbol{I})\right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{0}, \sigma^2_{\text{prior}}\boldsymbol{I})) && (10) \\
&= \mathbb{E}_{\boldsymbol{\theta}} \left[\sum\nolimits_{i=1}^{N} \log \mathcal{N}(\boldsymbol{x}_{t'}^{i} \mid \mathbb{E}_{\boldsymbol{\tau}} \left[\widehat{H}^{\boldsymbol{\tau}}_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_t^i)\right], \sigma^2_{\text{data}}\boldsymbol{I})\right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{0}, \sigma^2_{\text{prior}}\boldsymbol{I})) && (11) \\
&\geq \mathbb{E}_{\boldsymbol{\theta}} \left[\mathbb{E}_{\boldsymbol{\tau}} \left[\sum\nolimits_{i=1}^{N} \log \mathcal{N}(\boldsymbol{x}_{t'}^{i} \mid \widehat{H}^{\boldsymbol{\tau}}_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_t^i), \sigma^2_{\text{data}}\boldsymbol{I})\right]\right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{0}, \sigma^2_{\text{prior}}\boldsymbol{I})) && (12) \\
&\approx \frac{1}{M} \sum_{i=1}^{M} \left[\sum\nolimits_{i=1}^{N} \log \mathcal{N}(\boldsymbol{x}_{t'}^{i} \mid \frac{1}{S} \sum_{s=1}^{S} H_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{\Phi}^{\boldsymbol{\tau}^{(s)}}_{\mathcal{C}_{\boldsymbol{\eta}}}(\boldsymbol{x}_t^i)), \sigma^2_{\text{data}}\boldsymbol{I})\right] - \mathrm{KL}(q_{\boldsymbol{m},\boldsymbol{S}}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{0}, \sigma^2_{\text{prior}}\boldsymbol{I})) && (13)
\end{aligned}
$$

where we use M samples to obtain an unbiased estimate of $\mathbb{E}_{\boldsymbol{\theta}} := \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{m},\boldsymbol{S}}}$ and a single sample $\mathbb{E}_{\boldsymbol{\tau}} = \mathbb{E}_{\boldsymbol{\tau} \sim \prod_{s=1}^{S} p(\boldsymbol{\tau})}$ and use an S-sampled Monte Carlo estimate of the symmetrised Hamiltonian:

$$
\widehat{H}_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{x}_t^i) = \frac{1}{S} \sum_{s=1}^{S} H_{\boldsymbol{\theta},\boldsymbol{\eta}}(\boldsymbol{\Phi}^{\boldsymbol{\tau}^{(s)}}_{\mathcal{C}_{\boldsymbol{\eta}}}(\boldsymbol{x}_t^i)) \text{ with samples } \boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}, \ldots, \boldsymbol{\tau}^{(S)} \sim \mu(\boldsymbol{\tau}) \qquad (14)
$$

with the fact that this yields an unbiased estimator of the true symmetrised Hamiltonian $\mathbb{E}_{\boldsymbol{\tau}} \left[\widehat{H}_{\boldsymbol{\theta},\boldsymbol{\eta}}(\cdot)\right] = \widehat{H}_{\boldsymbol{\theta},\boldsymbol{\eta}}(\cdot)$. where we obtained an unbiased estimate of expectations through S sampled symmetry transformations and M sampled parameters. The first inequality is the standard VI lower bound. The second inequality follows from applying Jensen's inequality (again), using the fact that the log likelihood is convex. Similar bounds to symmetrisation by averaging over orbits have appeared in prior work [van der Ouderaa and van der Wilk, 2021, Schwöbel et al., 2022, Nabarro et al., 2022].

B Ground truth conserved quantities

In this section, we'll discuss the conserved quantities present in the ground-truth Hamiltonians of the systems we discuss.

As stated in Section 2.3, we can combine conserved quantities into new ones by linear combinations, products, and Poisson brackets. Thus we'll speak of the generating set of conserved quantities, which combine into all conserved quantities.

B.1 Simple harmonic oscillator

For the simple harmonic oscillator, the phase space is $\mathbb{R} \times \mathbb{R}$. The ground truth Hamiltonian is

$$
H(p, q) = \frac{p^2}{2m} + \frac{kq^2}{2}.
$$

We choose $m = k = 1$, so that $H(p, q) = (p^2 + q^2)/2$. Time evolution is a rotation of phase space. The Hamiltonian itself generates all conserved quantities.

B.2 $n-$Simple harmonic oscillators

The phase space is $\mathbb{R}^{2n}$. We choose all $k = m = 1$, so that the Hamiltonian is

$$H(p, q) = \|p\|^2/2 + \|q\|^2/2.$$

Time evolution rotates each pair (q_i, p_i). The conserved quantities are generated by the following set, for $i, j = 1, ..., n, i \neq j$.

$$\begin{aligned} H_i &= (q_i^2 + p_i^2)/2 \\ R_{ij} &= q_i p_j - q_j p_i \\ F_{ij} &= q_i q_j + p_i p_j \end{aligned}$$

The conserved quantity H_i rotates the pair (q_i, p_i). R_{ij} rotates both pairs (q_i, q_j) and (p_i, p_j) and F_{ij} rotates both pairs (q_i, p_j) and (q_j, p_i).

Alternatively, we can interpret the phase space as $\mathbb{C}^n$ [Arnold, 1989, Sec. 41E], with the positions being the real part and the momenta the imaginary part. In that case, the symplectic form J becomes simply the complex number $-i$. Then $H(\boldsymbol{x}) = \boldsymbol{x}^\dagger \boldsymbol{x}/2$. Time evolution is multiplication by the complex number e^{-it}. Conserved quantities are $H_i = x_i^* x_i/2$, which is real, and $C_{ij} = x_i^* x_j$, whose real part corresponds to F_{ij} and imaginary part to R_{ij}. The conserved quantities H_i and C_{ij} are quadratic and thus their symmetries are generated by linear matrices, which are all skew-Hermitian. In fact, all skew-Hermitian matrices are spanned by these generators. This shows that the combined symmetry group is in fact $U(n)$ [Amiet and Weigert, 2002].

B.3 n-body

In D spatial dimensions, with n bodies, the phase space is $\mathbb{R}^{2nD}$ and the Hamiltonian is

$$H(p, q) = \sum_i \frac{\|p_i\|^2}{2m_i} + \sum_{i \neq j} \frac{G m_i m_j}{\sqrt{\|q_i - q_j\|^2 + \epsilon^2}}$$

with a small ϵ to make it smooth.

The main conserved quantities are, for $d, d' = 1, ...D$, $d \neq d'$,

$$\begin{aligned} T_d &= \sum_i p_{id} \\ R^{\text{ABS}}_{dd'} &= \sum_i (q_i \wedge p_i)_{dd'} \\ R^{\text{COM}}_{dd'} &= (q_{\text{COM}} \wedge p_{\text{COM}})_{dd'} \end{aligned}$$

with $q_{\text{COM}} = \sum_i m_i q_i / \sum_i m_i$ and $p_{\text{COM}} = \sum_i m_i p_i / \sum_i m_i$.

These generate further conserved quantities of interest:

$$\begin{aligned} P_d &= T_d^2 \\ Q_{dd'} &= T_d T_{d'} \\ R^{\text{REL}}_{dd'} &= \sum_i ((q_i - q_{\text{COM}}) \wedge (p_i - p_{\text{COM}})_{dd'} = R^{\text{ABS}}_{dd'} - 2R^{\text{COM}}_{dd'} \end{aligned}$$

As T_d is linear, we can still learn P_d and $Q_{dd'}$ as quadratic conserved quantities. As the R^{REL} is a linear combination of other conserved quantities, we disregard it in our analysis of the learned symmetries.

The corresponding symmetries are: T_d translates all bodies in the d direction. $R^{\text{COM}}_{dd'}$ rotates the center of mass of all the bodies in the plane dd', while preserving the positions relative to the center of mass. R^{ABS} rotates all bodies relative to the origin. R^{REL} rotates all bodies relative to the center of mass. P_d translates in the d direction proportional to its COM momentum. $Q_{dd'}$ translates in direction d proportional to $p_{\text{COM},d'}$ and vice versa.

These symmetries together generate a group we'll call $\mathcal{G}$. This group has as a subgroup SE(D), which is generated by T and R^{ABS}. The group $\mathcal{G}$ has the same orbits as SE(D), as each element in $\mathcal{G}$ can be seen as a rototranslation conditional on some property of phase space. Because the orbits are the same, for any observable $f : \mathcal{M} \to \mathbb{R}$, we have that f invariant to $\mathcal{G}$ is true if and only if f is invariant to SE(2).

This system can have further conserved quantities, such as the Laplace–Runge–Lenz vector for $n = 2$. However, these are not expressible as a quadratic polynomial. As far as we know, the conserved quantities listed above are all that are expressible as a quadratic polynomial.

C Experimental details

All experiments were run on a single NVIDIA RTX 4090 GPU with 24GiB of GPU memory.

C.1 Simple harmonic oscillator experiment

Data For the training data, we sampled 7 initial conditions from unit Gaussian and simulated 4 datapoints with $\Delta t = 0.2$ apart. For test data, we sampled 100 initial conditions from unit Gaussian and simulated 20 timesteps with $\Delta t = 0.2$ from each initial condition.

Training We use an MLP with 2 hidden layers, each consisting of 200 hidden neurons and a linear exponential unit activation function with $\alpha = 2$. For symmetrisation, we use $S = 200$ samples from a uniform measure for $\mu(\tau)$. We use 20 Euler steps for time integration. We use fixed output noise and closed-form prior variance (Appendix D). We optimise the ELBO in full batch with Adam [Kingma and Ba, 2014] $(\beta_1 = 0.9, \beta_2 = 0.999)$ trained for 2000 epochs with a learning rate of 0.001, cosine annealed to 0.

C.2 $n-$Simple harmonic oscillators experiment

Data For training data, we randomly sampled 200 initial conditions independently from a unit normal. For each initial condition, we simulated a trajectory consisting of 50 data points at 0.3 time units apart.

Training We use an MLP with 3 hidden layers with 200 hidden units and exponential linear activation functions with $\alpha = 1$. We optimise the ELBO in mini batches of $B = 20$ trajectories, using $S = 100$ symmetrisation samples, 20 Euler steps for time integration, and $M = 2$ weight samples using Adam [Kingma and Ba, 2014] $(\beta_1 = 0.9, \beta_2 = 0.999)$ for 2000 epochs with a learning rate starting from 0.001, cosine annealed to 0.

C.3 n-body experiment

Data For training data, we randomly sampled 200 initial conditions by independently sampling positions from a unit normal, shifted by a normal with a standard deviation of 3. From each initial condition, we simulated trajectories consisting of 50 data points 0.3 time units apart.

Training We use an MLP with 4 hidden layers with 250 hidden unit units and exponential linear unit activation functions with $\alpha = 1$. We optimise the ELBO batches of $B = 20$ trajectories, $S = 100$ symmetrisation samples, 20 Euler steps for time integration, and $M = 2$ weight samples using Adam [Kingma and Ba, 2014] $(\beta_1 = 0.9, \beta_2 = 0.999)$ for 2000 epochs with a learning rate starting from 0.001, cosine annealed to 0.

https://doi.org/10.52202/079017-4318

D On training a neural network with variational inference

D.1 Matrix normal variational posterior

Naively, the covariance of a Gaussian posterior over weights grows quadratically with the number of parameters $|\boldsymbol{\theta}|^2$. It is therefore common to disregard all correlations between weights, resulting in a diagonal or mean-field posterior. Although the ELBO remains a lower bound for any choice of approximate family, more crude approximations can increase the slack in the bound, possibly making it harder to use estimates for Bayesian model selection. We, therefore, propose to use matrix normal posteriors [Louizos and Welling, 2016] factorised per layer,

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{N} q(\boldsymbol{\theta}_l), \text{ with} \qquad\qquad q(\boldsymbol{\theta}_l) = \mathcal{N}(\boldsymbol{\theta}_l \; |, \boldsymbol{m}_l, \boldsymbol{S}_l \otimes \boldsymbol{A}_l) \tag{15}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L)$ denote the weights of each layer l. If we denote the parameters in terms of weight and bias matrices of each layer with in_l input and out_l output dimensions, $\text{vec}(\boldsymbol{\theta}_l) = [\boldsymbol{W}_l \quad \boldsymbol{b}_l] \in \mathbb{R}^{\text{out}_l \times (\text{in}_l + 1)}$, we can equivalently write this posterior as a factorised matrix normal distribution:

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{N} q(\boldsymbol{\theta}_l), \text{ with} \qquad\qquad q(\boldsymbol{\theta}_l) = \mathcal{MN}([\boldsymbol{W}_l \quad \boldsymbol{b}_l] \; |, \boldsymbol{M}_l, \boldsymbol{S}_l \otimes \boldsymbol{A}_l) \tag{16}$$

The variational parameters $\{\boldsymbol{W}_l, \boldsymbol{S}_l, \boldsymbol{A}_l\}$ provide the mean $\boldsymbol{M}_l$ as well as correlations between layer inputs and bias $\boldsymbol{A}_l \in \mathbb{R}^{(\text{in}_l+1),(\text{in}_l+1)}$ and layer outputs $\boldsymbol{S}_l \in \mathbb{R}^{\text{out}_l,\text{out}_l}$. For L hidden layers of width H, the number of variational parameters scales quadratically $\mathcal{O}(LH^2)$ compared to the quartic number of variational parameters $\mathcal{O}(LH^4)$ we would need to represent the full covariance. This strikes a practical balance between taking important correlations into account while avoiding having to make a mean-field assumption. Further, we note that the matrix gaussian posterior of [Louizos and Welling, 2016] is the same approximate distribution as used in Kronecker-factored Laplace approximations [Grosse and Martens, 2016]. In Laplace approximations the covariance is the inverse Hessian, whereas in variational inference the covariance is optimised using the ELBO. Layer-factored matrix normal distributions have been succesfully applied to perform approximate Bayesian model selection based on the Laplace approximation in [Immer et al., 2021, 2022, van der Ouderaa et al., 2024]. This work provides evidence that variational inference can also be used to obtain approximate posteriors of this form and obtain a lower bound on the marginal likelihood that is sufficiently tight to perform Bayesian model selection in deep neural networks.

D.2 Closed-form output variance

It can be shown that the output variance that maximises the marginal likelihood $\hat{\sigma}^2_{\text{data}}$ is the empirical variance of the output. We, therefore, either fixing the output variance $\hat{\sigma}^2_{\text{data}}$ - typically to a very small number in noise-free settings, or setting the output variance to an empirical output variance. An exponentially weighted average of the empirical variance over mini-batches can be used.

On downscaling the KL term by a $\beta-$scalar Many deep learning papers that use variational inference down-scale the KL term by a $\beta-$parameter [Higgins et al., 2017]. We note that, for standard Gaussian likelihoods, scaling the output variance is equivalent to inversely scaling the KL term. We do advice against downscaling of the KL term, as it makes it less clear that the resulting objective is still a lower bound to the marginal likelihood, and hides the fact that the lower bound corresponds to a changed model with altered output variance. In MAP estimation under a Gaussian likelihood, the output variance is arbitrary as it only scales the objective not effecting the optimum, and the objective is often simplified as the mean squared error. In variational inference, the output variance does play an important role of balancing the relative importance between the log likelihood (data fit) and KL term (pull to prior). In this setting, using half mean squared error effectively corresponds to an output variance of $\sigma^2_{\text{data}} = 1$. In practice, this value is often too high because common machine learning datasets have little label noise. As a result, the log likelihood term is too weak and the KL term is too strong. We hypothesise that this has led to practitioners to down-weighting the KL term to obtain sensible posterior predictions, without necessarily realising that they were effectively altering the output variance of the model. Using automatic output variance, the optimal $\hat{\beta}$ can be set to (a running estimate of) the inverse of the empirical variance, also known as the empirical prior precision.

D.3 Closed-form prior variance minimising inverse KL

Consider the setting of a D-dimensional Gaussian q, parameterised by mean $\boldsymbol{m}$ and covariance $\boldsymbol{S}$, and a zero mean Gaussian p_v with scalar variance v in each of the equally many dimensions:

$$q = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S}), \qquad\qquad p_v = \mathcal{N}(\boldsymbol{0}, v\boldsymbol{I})$$

where $\boldsymbol{0}$ denotes a zero vector and $\boldsymbol{I}$ an identity matrix. As the log likelihood does not depend on v, we can find v that optimises the marginal likelihood by finding the minimiser of the inverse KL:

$$\begin{aligned}
\arg\min_v \text{KL}\left[q \,||\, p_v\right] &= \arg\min_v \frac{1}{2}\left[\log\frac{|v\boldsymbol{I}|}{|\boldsymbol{S}|} - D + \text{Tr}((v\boldsymbol{I})^{-1}\boldsymbol{S}) + (\boldsymbol{0}-\boldsymbol{m})^T(v\boldsymbol{I})^{-1}(\boldsymbol{0}-\boldsymbol{m})\right] \\
&= \arg\min_v \frac{1}{2}\left[\log\frac{|v\boldsymbol{I}|}{|\boldsymbol{S}|} - D + \text{Tr}(\boldsymbol{S})/v + \boldsymbol{m}^T\boldsymbol{m}/v\right] \\
&= \arg\min_v \left[D\log(v) + \text{Tr}(\boldsymbol{S})/v + \boldsymbol{m}^T\boldsymbol{m}/v\right]
\end{aligned}$$

Setting the derivative to zero:

$$\begin{aligned}
0 &= \frac{\partial}{\partial v}\left[D\log(v) + \frac{1}{v}\text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}/v\right] \\
0 &= -\frac{-Dv + \text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}}{v^2} \\
v_* &= \frac{\text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}}{D} \qquad (17)
\end{aligned}$$

We found KL-minimising variance v_* in closed-form as a function of $\boldsymbol{m}$ and $\boldsymbol{S}$. Verified numerically.

D.4 Plugging minimising variance into KL

Plugging v_* back into the Gaussian p_{v_*} and computing the KL:

$$\begin{aligned}
\text{KL}\left[q \,||\, p_{v_*}\right] &= \frac{1}{2}\left[\log\frac{|v_*\boldsymbol{I}|}{|\boldsymbol{S}|} - D + \text{Tr}((v_*\boldsymbol{I})^{-1}\boldsymbol{S}) + (\boldsymbol{0}-\boldsymbol{m})^T(v_*\boldsymbol{I})^{-1}(\boldsymbol{0}-\boldsymbol{m})\right] \\
&= \frac{1}{2}\left[\log\frac{|v_*\boldsymbol{I}|}{|\boldsymbol{S}|} - D + \frac{D\text{Tr}(\boldsymbol{S})}{\text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}} + \frac{D\boldsymbol{m}^T\boldsymbol{m}}{\text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}}\right] = \frac{1}{2}\left[\log\frac{|v_*\boldsymbol{I}|}{|\boldsymbol{S}|}\right] \\
&= \frac{1}{2}\left[D\log\left(\frac{\text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}}{D}\right) - \log|\boldsymbol{S}|\right]
\end{aligned}$$

shows that the resulting KL is only measuring the relative volume $\frac{1}{2}\log\frac{|q|}{|p|}$ between the prior and the posterior, which can be further simplified as

$$\text{KL}\left[q \,||\, p_{v_*}\right] = \frac{1}{2}\left[D\log(\text{Tr}(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m}) - D\log(D) - \log|\boldsymbol{S}|\right]$$

In practice, we might reparameterise $\boldsymbol{S} = \boldsymbol{L}\boldsymbol{L}^T$ in terms of its triangular Cholesky factor $\boldsymbol{L}$ and use

$$\begin{aligned}
&\log|\boldsymbol{S}| = \log|\boldsymbol{L}\boldsymbol{L}^T| = \log|\boldsymbol{L}|^2 = 2\log|\boldsymbol{L}| = 2\log\prod_i \boldsymbol{L}_{ii} = 2\sum_i \log\boldsymbol{L}_{ii} \\
&\text{Tr}(\boldsymbol{S}) = \text{Tr}(\boldsymbol{L}\boldsymbol{L}^T) = \sum_{i,j}\boldsymbol{L}_{ij}^2
\end{aligned}$$

This gives the final expression of the KL

$$\text{KL}\left[q \,||\, p_{v_*}\right] = \frac{1}{2}\left[D\log\left(\sum_{i,j}\boldsymbol{L}_{ij}^2 + \sum_i \boldsymbol{m}_i^2\right) - D\log(D) - 2\sum_i \log\boldsymbol{L}_{ii}\right] \qquad (18)$$

which implicitly uses the derived optimal variance $v_* = \frac{\sum_{i,j}\boldsymbol{L}_{ij}^2 + \sum_i \boldsymbol{m}_i}{D}$.

E Code and Questions

The code is available at `https://github.com/tychovdo/noethers-razor`.
For any questions, please contact the corresponding author, Tycho van der Ouderaa, by email.

https://doi.org/10.52202/079017-4318