EQUIVARIANT MANIFOLD FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tractably modelling distributions over manifolds has long been an important goal in the natural sciences. Recent work has focused on developing general machine learning models to learn such distributions. Though these are significant advances, many problems require the learned distribution to obey symmetries inherent to the manifold. Up until now, the theory of invariant learning on manifolds was underdeveloped—models were unable to incorporate these symmetries while learning. In this paper, we lay the theoretical foundations for learning symmetry invariant distributions on arbitrary manifolds via equivariant manifold flows. We demonstrate the efficacy of our approach in the context of quantum field theory in learning gauge invariant densities over $SU(n)$.

1 INTRODUCTION

Density learning over manifolds has a broad array of applications, ranging from quantum field theory in physics (Wirnsberger et al., 2020) to motion estimation in robotics (Feiten et al., 2013) to protein-structure prediction in computational biology (Hamelryck et al., 2006). Recent work (Lou et al., 2020b) has extended the powerful framework of continuous normalizing flows (Chen et al., 2018; Grathwohl et al., 2019) to the setting of Riemannian manifolds, lifting the utility of these models for learning complex probability distributions to a more general setting.

Although tractable and principled learning of manifold distributions was a considerable step forward, this is insufficient for some problems in the natural sciences. Several applications, for example sampling coupled particle systems in physical chemistry (Köhler et al., 2020) or sampling for $SU(n)$ lattice gauge theories in theoretical physics (Boyda et al., 2020), require distribution symmetries that are nontrivial to enforce. Typically, manifold density structure and symmetries are enforced in an ad hoc way, using properties specific to the manifold in order to make density learning tractable. In contrast, our paper presents a fully general way to learn flows that induce symmetry invariant distributions.

2 RELATED WORK

Normalizing Flows on Manifolds Normalizing flows on manifolds have received a considerable amount of attention, both in terms of manifold-specific and general constructions. Rezende et al. (2020) introduced constructions specific to tori and spheres, while Bose et al. (2020) introduced constructions for hyperbolic space. Following this work, Lou et al. (2020b); Mathieu & Nickel (2020) introduced a fully general construction by extending Neural ODEs (Chen et al., 2018) to the setting of Riemannian manifolds.

Equivariant Machine Learning Equivariance has been recently discussed in the context of machine learning (Cohen & Welling, 2016; Cohen et al., 2018; 2019; Kondor & Trivedi, 2018), and in particular, Köhler et al. (2020) introduced equivariant normalizing flows for Euclidean space. Boyda et al. (2020), introduced equivariant flows for $SU(n)$ via a manifold-specific construction. In contrast, the equivariant manifold flows in our paper are fully general and are applicable to arbitrary Riemannian manifolds.

3 BACKGROUND

In this section, we provide a terse overview of necessary concepts. For a more detailed introduction, we refer the reader to a text such as Lee (2013); Kobyzev et al. (2020).

3.1 RIEMANNIAN GEOMETRY

Recall that a Riemannian manifold $(\mathcal{M}, h)$ is an n-dimensional manifold with a smooth collection of inner products $(h_x)_{x\in\mathcal{M}}$ for every tangent space $T_x\mathcal{M}$. The Riemannian metric h induces a distance d_h on the manifold.

A diffeomorphism $f : \mathcal{M} \to \mathcal{M}$ is called an isometry if $h(D_x f(u), D_x f(v)) = h(u, v)$ for all tangent vectors $u, v \in T_x\mathcal{M}$ where $D_x f$ is the differential of f. Note that isometries preserve the manifold distance function. The collection of all isometries forms a group G, which we call the isometry group of the manifold $\mathcal{M}$.

Riemannian metrics also allow for a natural analogue of gradients on $\mathbb{R}^n$. For a function $f : \mathcal{M} \to \mathbb{R}$, we define the Riemannian gradient $\nabla_x f$ to be the vector on $T_x\mathcal{M}$ such that $h(\nabla_x f, v) = D_x f(v)$ for $v \in T_x\mathcal{M}$.

3.2 NORMALIZING FLOWS ON MANIFOLDS

Let $(\mathcal{M}, h)$ be a Riemannian manifold. A normalizing flow on $\mathcal{M}$ is a diffeomorphism $f_\theta : \mathcal{M} \to \mathcal{M}$ (parametrized by θ) that transforms a prior density ρ to target density ρ_{f_θ}. The target distribution can be computed via the change of variables equation

$$\rho_{f_\theta}(x) = \rho\left(f_\theta^{-1}(x)\right)\left|\det \frac{\partial f_\theta^{-1}(x)}{dx}\right| = \rho\left(f_\theta^{-1}(x)\right)\left|\det J_{f_\theta^{-1}}(x)\right|.$$

3.3 EQUIVARIANCE

We say that a function $f : X \to Y$ is equivariant if, for symmetries $g_x : X \to X$ and $g_y : Y \to Y$, $f \circ g_x = g_y \circ f$. We say a function $f : X \to Y$ is invariant if $f \circ g_x = f$. When X and Y are manifolds, the symmetries g_x and g_y are isometries.

4 THEORETICAL DERIVATIONS

In this section, we derive the necessary theorems for flows equivariant to isometries on the manifold. In particular, we show how equivariant flows induce an invariant density, and we present a way of constructing equivariant flows from invariant functions. We defer the proofs of the theorems to the appendix.

4.1 EQUIVARIANT FLOWS

Invariance of Density For a group G, a density ρ on a manifold $\mathcal{M}$ is G-invariant if, for all $g \in G$ and $x \in \mathcal{M}$, $\rho(R_g x) = \rho(x)$, where R_g is the action of g on x.

Equivariant Flows A flow f on a manifold $\mathcal{M}$ is G-equivariant if it commutes with actions from G, i.e. we have $R_g \circ f = f \circ R_g$.

We first show that isometry equivariant flows induce isometry invariant densities. Note that we require the group to be an isometry in order to control the distribution of ρ_f, and the following theorem does not hold for general diffeomorphism groups.

Theorem 1. *Let $(\mathcal{M}, h)$ be a Riemannian manifold, and G be its isometry group (or one of its subgroups). If ρ is a G-invariant density on $\mathcal{M}$, and f is a G-equivariant diffeomorphism, then $\rho_f(x)$ is also G-invariant.*

4.2 CONSTRUCTING EQUIVARIANT FLOWS ON MANIFOLDS

To actually construct manifold equivariant flows, we will use tools from manifold ordinary differential equations (ODEs) and continuous normalizing flows (CNFs).

Equivariant Vector Field Let $X : \mathcal{M} \times [0, \infty) \to T\mathcal{M}$, $X(m, t) \in T_m\mathcal{M}$ be a time-dependent vector field on manifold $\mathcal{M}$, with base point $x_0 \in \mathcal{M}$.

X is a G-equivariant vector field if $\forall (m, t) \in \mathcal{M} \times [0, \infty)$, $X(R_g m, t) = (D_m R_g)X(m, t)$.

Manifold Continuous Normalizing Flows A manifold continuous normalizing flow with base point z is a function $\gamma : [0, \infty) \to \mathcal{M}$ that satisfies the manifold ODE

$$\frac{d\gamma(t)}{dt} = X(\gamma(t), t)\ , \gamma(0) = z$$

We define $F_{X,T} : \mathcal{M} \to \mathcal{M}$, $z \mapsto F_{X,T}(z)$ to map any base point $z \in \mathcal{M}$ to the value of the CNF starting at z, evaluated at time T. This function is known as the (vector field) flow of X.

There exists a natural correspondence between equivariant flows and equivariant vector fields.

Theorem 2. *Let $(\mathcal{M}, h)$ be a Riemannian manifold, and G be its isometry group (or one of its subgroups). Let X be any time-dependent vector field on $\mathcal{M}$, and $F_{X,T}$ be the flow of X. Then X is a G-equivariant vector field if and only if $F_{X,T}$ is a G-equivariant flow.*

4.3 EQUIVARIANT GRADIENT OF POTENTIAL

To design an equivariant vector field X as stated above, it is sufficient to set the vector field dynamics of X as the gradient of some G-invariant potential function $\Phi : \mathcal{M} \to \mathbb{R}$.

Theorem 3. *Let $(\mathcal{M}, h)$ be a Riemannian manifold and G be its group of isometries (or an isometry subgroup). If $\Phi : \mathcal{M} \to \mathbb{R}$ is a smooth G-invariant function, then the following diagram commutes for any $g \in G$:*

$$\begin{array}{ccc} \mathcal{M} & \xrightarrow{R_g} & \mathcal{M} \\ \downarrow{\scriptstyle \nabla\Phi} & & \downarrow{\scriptstyle \nabla\Phi} \\ T\mathcal{M} & \xrightarrow{DR_g} & T\mathcal{M} \end{array}$$

or $\nabla_{R_g u}\Phi = D_u R_g(\nabla_u \Phi)$. Hence $\nabla\Phi$ is a G-equivariant vector field. This condition is also tight in the sense that it only occurs if G is the group of isometries.

5 EQUIVARIANT FLOWS ON $SU(n)$

For many applications in physics (specifically gauge theory and lattice quantum field theory), one works with the Lie Group $SU(n)$ — the group of unitary matrices with determinant 1. In particular, when modelling probability distributions on $SU(n)$, the desired distribution must be invariant under conjugation by $SU(n)$ (Boyda et al., 2020). Conjugation is an isometry on $SU(n)$, so we can model probability distributions invariant under this action with our developed theory.

5.1 INVARIANT POTENTIAL PARAMETERIZATION

Our previous derivations have reduced our problem of modelling G-equivariant flows to modelling G-invariant potential functions $\Phi : SU(n) \to \mathbb{R}$. Note that matrix conjugation preserves eigenvalues. Thus, for a function $\Phi : SU(n) \to \mathbb{R}$ to be invariant to matrix conjugation, it has to act on the eigenvalues of $x \in SU(n)$ as a multi-set.

We can parameterize such potential functions Φ by the DeepSet network from Zaheer et al. (2017). DeepSet is a permutation invariant neural network that acts on the eigenvalues, so the mapping of $x \in SU(n)$ is $\Phi(x) = \hat{\Phi}(\{\lambda_1(x), \ldots, \lambda_n(x)\})$ for some set function $\hat{\Phi}$.

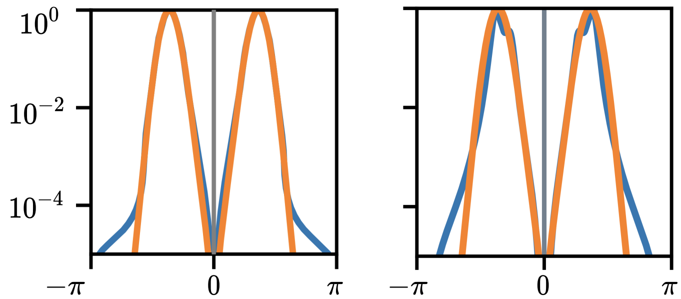

Figure 1: Learned densities on $SU(2)$ from [Left] Boyda et al. (2020) model and [Right] our model. The target density is in orange, while the model densities are in blue. All densities are normalized to have maximum value 1. The x-axis is θ for the eigenvalue $e^{i\theta}$ of a matrix in $SU(2)$ (note the other eigenvalue is determined as $e^{-i\theta}$). Our model has favorable behavior in low-density regions.

With this matrix conjugation invariant potential Φ, we can model a matrix conjugation equivariant vector field on $SU(n)$ with $\nabla\Phi$.

5.2 Prior Distributions

For the prior distribution of the flow, we use the Haar measure on $SU(n)$, which is given for an $x \in SU(n)$ as $\text{Haar}(x) = \prod_{i<j} |\lambda_i(x) - \lambda_j(x)|^2$ (Boyda et al., 2020). Note that this distribution is invariant with respect to matrix conjugation, so we may use it in our model. We can sample from and compute the log probabilities with respect to this distribution efficiently with standard matrix computations (Mezzadri, 2007).

6 Experiments

As mentioned previously, the learning of such an invariant density over a manifold has myriad applications, ranging from sampling coupled particle systems in physical chemistry (Köhler et al., 2020), to sampling for $SU(n)$ lattice gauge theories in theoretical physics (quantum field theory) (Kanwar et al., 2020; Boyda et al., 2020). Boyda et al. (2020) is particularly pertinent since they construct flows on $SU(n)$ that are invariant to conjugation by $SU(n)$. Hence we test in this context and compare our general model (with the above DeepSet potential) to their manifold-specific construction.

6.1 $SU(n)$ Gauge Equivariant Neural Network Flows

With our equivariant flows and invariant base distribution, our model learns densities on $SU(n)$ that are invariant to matrix conjugation. Figure 1 displays learned densities for our model and the model of Boyda et al. (2020) in the case of a particular density on $SU(2)$ described in Appendix C.2. Training details are given in Appendix C.1. While both models match the target distribution well in high-density regions, we find that our model shows an improvement in lower-density regions, where the tails of our learned distribution decay faster. Hence, our model appears to inherit the strength of Neural Manifold ODEs in modelling such regions (Lou et al., 2020a).

7 Conclusion

In this work, we introduce equivariant manifold flows in a fully general context and provide the necessary theory to ensure a principled construction. We also demonstrate the efficacy of our approach in the context of learning a conjugation invariant density over $SU(n)$, which is an important task for sampling $SU(n)$ lattice gauge theories in quantum field theory.

REFERENCES

Joey Bose, Ariella Smofsky, Renjie Liao, Prakash Panangaden, and Will Hamilton. Latent variable modelling with hyperbolic normalizing flows. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1045–1055, 2020.

Denis Boyda, Gurtej Kanwar, Sébastien Racanière, Danilo Jimenez Rezende, Michael S Albergo, Kyle Cranmer, Daniel C Hackett, and Phiala E Shanahan. Sampling using $su(n)$ gauge equivariant flows. *arXiv preprint arXiv:2008.05456*, 2020.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6571–6583, 2018.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2990–2999, 2016.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1321–1330, 2019.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

Wendelin Feiten, Muriel Lang, and Sandra Hirche. Rigid motion estimation using mixtures of projected gaussians. *Proceedings of the 16th International Conference on Information Fusion*, pp. 1465–1472, 2013.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019.

Thomas Hamelryck, John T. Kent, and Anders Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2, 2006.

Gurtej Kanwar, Michael S Albergo, Denis Boyda, Kyle Cranmer, Daniel C Hackett, Sébastien Racaniere, Danilo Jimenez Rezende, and Phiala E Shanahan. Equivariant flow-based sampling for lattice gauge theory. *Physical Review Letters*, 125(12):121601, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Jonas Köhler, Leon Klein, and Frank Noe. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5361–5370, 2020.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2747–2755, 2018.

J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York, 2013.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the fréchet mean. In *International Conference on Machine Learning*, 2020a.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M De Sa. Neural manifold ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17548–17558, 2020b.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in Neural Information Processing Systems*, 33, 2020.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54:592–604, 2007.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

Danilo Jimenez Rezende, George Papamakarios, Sebastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8083–8092, 2020.

Peter Wirnsberger, Andrew J. Ballard, George Papamakarios, Stuart Abercrombie, Sébastien Racanière, Alexander Pritzel, Danilo Jimenez Rezende, and Charles Blundell. Targeted free energy estimation via learned mappings. *arXiv: Computational Physics*, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

A Proof of Theorems

In this section, we restate and prove the theorems in Section 4. These give the theoretical foundations that we use to build our models.

A.1 Proof of Theorem 1

Theorem 1. *Let $(\mathcal{M}, h)$ be a Riemannian manifold, and G be its isometry group (or one of its subgroups). If ρ is a G-invariant density on $\mathcal{M}$, and f is a G-equivariant diffeomorphism, then $\rho_f(x)$ is also G-invariant.*

Proof. We wish to show $\rho_f(x)$ is also G-invariant, i.e. $\rho_f(R_g x) = \rho_f(x)$ for all $g \in G, x \in \mathcal{M}$.

We first recall the definition of ρ_f:

$$\rho_f(x) = \rho\left(f^{-1}(x)\right)\left|\det \frac{\partial f^{-1}(x)}{dx}\right| = \rho\left(f^{-1}(x)\right)\left|\det J_{f^{-1}}(x)\right|.$$

Since $f \in C^1(\mathcal{M}, \mathcal{M})$ is G-equivariant, we have $f \circ R_g = R_g \circ f$ for any $g \in G$. Also, since ρ is G-invariant, we have $\rho \circ R_g = \rho$. Combining these properties, we see that:

$$\begin{aligned}
\rho_f(R_g x) &= \rho_f(R_g x)\frac{|\det J_{R_g}(x)|}{|\det J_{R_g}(x)|} = \frac{\rho_{R_{g^{-1}} \circ f}(x)}{|\det J_{R_g}(x)|} && \text{(expanding definition of } \rho_f\text{)} \\
&= \frac{\rho_{f \circ R_{g^{-1}}}(x)}{|\det J_{R_g}(x)|} = \rho\left((R_g \circ f^{-1})(x)\right)\frac{|\det J_{R_g \circ f^{-1}}(x)|}{|\det J_{R_g}(x)|} && \text{(G-equivariance of } f\text{)} \\
&= (\rho \circ R_g \circ f^{-1})(x)\frac{|\det J_{R_g}(f^{-1}(x))J_{f^{-1}}(x)|}{|\det J_{R_g}(x)|} && \text{(expanding Jacobian)} \\
&= (\rho \circ f^{-1})(x)\frac{|\det J_{R_g}(f^{-1}(x))||\det J_{f^{-1}}(x)|}{|\det J_{R_g}(x)|} && \text{(G-invariance of } \rho\text{)} \\
&= \rho(f^{-1}(x))|\det J_{f^{-1}}(x)| \cdot \frac{|\det J_{R_g}(f^{-1}(x))|}{|\det J_{R_g}(x)|} && \text{(rearrangement)} \\
&= \rho_f(x) \cdot \frac{|\det J_{R_g}(f^{-1}(x))|}{|\det J_{R_g}(x)|} && \text{(expanding definition of } \rho_f\text{)}
\end{aligned}$$

Now note that G is contained in the isometry group, and thus R_g is an isometry. This means $|\det J_{R_g}(x)| = 1$ for any $x \in \mathcal{M}$, so RHS above is simply $\rho_f(x)$, which proves the theorem. □

A.2 Proof of Theorem 2

Theorem 2. *Let $(\mathcal{M}, h)$ be a Riemannian manifold, and G be its isometry group (or one of its subgroups). Let X be any time-dependent vector field on $\mathcal{M}$, and $F_{X,T}$ be the flow of X. Then X is an G-equivariant vector field if and only if $F_{X,T}$ is a G-equivariant flow for any $T \in [0, +\infty)$.*

Proof. **G-equivariant $X \Rightarrow G$-equivariant $F_{X,T}$.** We invoke the following lemma from (Lee, 2013, Corollary 9.14):

Lemma 1. *Let $F : \mathcal{M} \to \mathcal{N}$ be a diffeomorphism. If $X \in \mathfrak{X}(\mathcal{M})$ and θ is the flow of X, then the flow of $F_* X$ is $\eta_t = F \circ \theta_t \circ F^{-1}$, with domain $N_t = F(M_t)$ for each $t \in \mathbb{R}$.*

Examine R_g and its action on X. Since X is G-equivariant, we have for any $(x, t) \in \mathcal{M} \times [0, +\infty)$,

$$((R_g)_* X)(x, t) = (D_{R_g^{-1}(x)} R_g) X(R_g^{-1}(x), t) = X(R_g \circ R_g^{-1}(x), t) = X(x, t)$$

so it follows that $(R_g)_* X = X$. Applying the lemma above, we get

$$F_{(R_g)_*X,T} = R_g \circ F_{X,T} \circ R_g^{-1}$$

and, by simplifying, we get that $F_{X,T} \circ R_g = R_g \circ F_{X,T}$, as desired.

***G*-equivariant $X \Leftarrow$ *G*-equivariant $F_{X,T}$.** This direction follows from the chain rule. If $F_{X,T}$ is G-equivariant, then at all times we have:

$$\begin{aligned}
(D_m R_g)\big(X(F_{X,t}(m), t\big) &= (D_m R_g)\left(\frac{d}{dt} F_{X,T}(m)\right) && \text{(definition)} \\
&= \frac{d}{dt}(R_g \circ F_{X,T})(m) && \text{(chain rule)} \\
&= \frac{d}{dt} F_{X,T}(R_g m) && \text{(equivariance)} \\
&= X(R_g(F_{X,t}(m)), t) && \text{(definition)}
\end{aligned}$$

This concludes the proof of the backward direction. □

A.3 Proof of Theorem 3

Theorem 3. *Let $(\mathcal{M}, h)$ be a Riemannian manifold and G be its group of isometries (or an isometry subgroup). If $\Phi : \mathcal{M} \to \mathbb{R}$ is a smooth G-invariant function, then the following diagram commutes for any $g \in G$:*

$$\begin{array}{ccc}
\mathcal{M} & \xrightarrow{R_g} & \mathcal{M} \\
\Big\downarrow{\scriptstyle \nabla\Phi} & & \Big\downarrow{\scriptstyle \nabla\Phi} \\
T\mathcal{M} & \xrightarrow{DR_g} & T\mathcal{M}
\end{array}$$

or $\nabla_{R_g u}\Phi = D_u R_g(\nabla_u \Phi)$. This is condition is also tight in the sense that it only occurs if G is the group of isometries.

Proof. We first recall the Riemannian gradient chain rule:

$$\nabla_u(\Phi \circ R_g) = (D_u R_g)^\top(\nabla_{R_g u}\Phi)$$

where $(D_u R_g)^\top : T_{R_g u}\mathcal{M} \to T_u\mathcal{M}$ is the "adjoint" given by

$$h\left(D_u R_g(v), w\right) = h\Big(v, (D_u R_g)^\top(w)\Big).$$

Since R_g is an isometry, we also have

$$h(x, y) = h\left(D_u R_g(x), D_u R_g(y)\right).$$

Combining the above two equations gives

$$h(x, y) = h(DuR_g(x), D_u R_g(y)) = h\Big(x, (D_u R_g)^\top\left(D_u R_g(y)\right)\Big),$$

which implies for all y,

$$h\Big(x, y - (D_u R_g)^\top(D_u R_g(y))\Big) = 0.$$

Since h is a Riemannian metric (even pseudo-metric works due to non-degeneracy), we must have that $(D_u R_g)^\top \circ (D_u R_g) = I$.

To complete the proof, we recall that $\Phi = \Phi \circ R_g$, and this combined with chain rule gives

$$\nabla_u \Phi = \nabla_u(\Phi \circ R_g) = (D_u R_g)^\top(\nabla_{R_g u}\Phi).$$

Now applying $D_u R_g$ on both sides gives

$$\nabla_{R_g u}\Phi = D_u R_g \nabla_u \Phi$$

which is exactly what we want to show.

We see that this is an "only if" condition because we must necessarily get that the adjoint is the inverse, which must imply that R_g is an isometry. □

A.4 PROOF OF COROLLARY 1 AND 2

Recall that both Corollary 1 and 2 specializes Theorems 1 and 3 to the case $\mathcal{M} = SU(n)$ and G inducing the matrix conjugation group action. To successfully apply these theorems, the one remaining condition that we need to verify is that G is an isometry subgroup.

Lemma 2. *Let G be the group action of conjugation by $SU(n)$, and let each R_g represent the corresponding action of conjugation by $g \in SU(n)$. Then G is an isometry subgroup.*

Proof. We first show that the matrix conjugation action of $SU(n)$ is unitary. For $R, X \in SU(n)$, note that the action of conjugation is given by $\mathrm{vec}(RXR^{-1}) = (R^{-T} \otimes R)\mathrm{vec}(X)$. We have $R^{-T} \otimes R$ being unitary because:

$$
\begin{aligned}
&(R^{-T} \otimes R)^*(R^{-T} \otimes R) \\
&= (\overline{R^{-1}} \otimes R^*)(R^{-T} \otimes R) && \text{(conjugate transposes distribute over } \otimes) \\
&= (\overline{R^{-1}} R^{-T}) \otimes (R^* R) && \text{(mixed-product property of } \otimes) \\
&= (R^T R^{-T}) \otimes (I) = (I) \otimes (I) = I_{N^2 \times N^2} && \text{(simplification)}
\end{aligned}
$$

Now choose an orthonormal frame $X_1, \ldots, X_n$ of $T\mathcal{M}$. Note that $T\mathcal{M}$ locally consists of traceless skew-Hermitian matrices. We show G is an isometry subgroup by noting that when it acts on the frame, the resulting frame is orthonormal. Let $g \in G$, and consider the result of action of g on the frame, namely $R_g X_1, \ldots, R_g X_n$. Then we have:

$$(R_g X_i)^*(R_g X_j) = X_i^* R_g^* R_g X_j = X_i^* X_j$$

Note for $i \neq j$, we have $X_i^* X_j = 0$ and for $i = j$ we see $X_i^* X_i = 1$. Hence the resulting frame is orthonormal and G is an isometry subgroup. ☐

B DIFFERENTIATING THROUGH EIGENDECOMPOSITION

In this section, we reconstruct the steps of differentiation through eigendecomposition from (Boyda et al., 2020, Appendix C) that allows efficient computation in our use-case. For our matrix-conjugation-invariant $SU(n)$ flow, we need only differentiate the eigenvalues with respect to the input $U \in SU(n)$.

For an input $U \in SU(n)$, let its eigendecomposition be $U = PDP^*$, where $w = \mathrm{diag}(D) \in \mathbb{C}^n$ contains its eigenvalues, and $P = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix} \in \mathbb{C}^{n \times n}$ with $p_i \in \mathbb{C}^n$ as its eigenvectors. Let L denote our loss function, write the downstream gradients in row vector format:

$$g = \begin{bmatrix} \frac{\partial L}{\partial \mathrm{Re} w} & \frac{\partial L}{\partial \mathrm{Im} w} \end{bmatrix} = \begin{bmatrix} g^{(1)} & g^{(2)} \end{bmatrix}.$$

Then following similar steps as in Boyda et al. (2020), we can compute the gradient of L with respect to the real and imaginary parts of U as follows:

$$\frac{\partial L}{\partial \mathrm{Re} U} = \sum_{i=1}^{n} g_i^{(1)} \mathrm{Re}\left(\overline{p_i} p_i^\top\right) + \sum_{i=1}^{n} g_i^{(2)} \mathrm{Im}\left(\overline{p_i} p_i^\top\right)$$

$$\frac{\partial L}{\partial \mathrm{Im} U} = -\sum_{i=1}^{n} g_i^{(1)} \mathrm{Im}\left(\overline{p_i} p_i^\top\right) + \sum_{i=1}^{n} g_i^{(2)} \mathrm{Re}\left(\overline{p_i} p_i^\top\right)$$

If we define

$$Q^{(1)} = \begin{bmatrix} g_1^{(1)} \overline{p_1} & \ldots & g_n^{(1)} \overline{p_n} \end{bmatrix} \qquad Q^{(2)} = \begin{bmatrix} g_1^{(2)} \overline{p_1} & \ldots & g_n^{(2)} \overline{p_n} \end{bmatrix}$$

Then we can write the gradients in terms of efficient matrix computations:

$$
\begin{aligned}
\frac{\partial L}{\partial \mathrm{Re} U} &= \mathrm{Re}\left(Q^{(1)} P^\top\right) + \mathrm{Im}\left(Q^{(2)} P^\top\right) \\
\frac{\partial L}{\partial \mathrm{Im} U} &= -\mathrm{Im}\left(Q^{(1)} P^\top\right) + \mathrm{Re}\left(Q^{(2)} P^\top\right).
\end{aligned}
$$

C EXPERIMENT DETAILS

C.1 TRAINING DETAILS

Our DeepSet network (Zaheer et al., 2017) consists of a feature extractor and regressor. The feature extractor is a 1 layer tanh network with 32 hidden channels. We concatenate the time component to the sum component of the feature extractor before feeding the resulting 33 size tensor into a 1 layer tanh regressor network.

To train our flows, we minimize the KL divergence between our model distribution and the target distribution (Papamakarios et al., 2019), as is done in Boyda et al. (2020). In a training iteration, we draw a batch of samples uniformly from $SU(2)$, map them through our flow, and compute the gradients with respect to the batch KL divergence between our model probabilities and the target density probabilities. We use the Adam stochastic optimizer for gradient-based optimization (Kingma & Ba, 2015). The graph shown in Figure 1 was trained for 300 iterations with a batch size of 8192 and weight decay setting of 0.01; the starting learning rate for Adam was 0.01, and a multi-step learning rate schedule that decreased the learning rate by a factor of 10 every 100 epochs.

C.2 TARGET DISTRIBUTIONS FROM BOYDA ET AL. (2020) FOR $SU(2)$

Boyda et al. (2020) define a family of matrix-conjugation-invariant densities on $SU(n)$ as:

$$p_{toy}(U) = \frac{1}{Z} e^{\frac{\beta}{n} \mathrm{Re}\,\mathrm{tr}\left(\sum_k c_k U^k\right)},$$

which is parameterized by scalars c_k and β. The normalizing constant Z is not particularly important for tasks of density estimation. Note that in the case of $n = 2$, since the eigenvalues of a matrix $U \in SU(2)$ are $e^{i\theta}, e^{-i\theta}$, we have that $\mathrm{tr}(U) = e^{i\theta} + e^{-i\theta} = 2\cos(\theta)$, so that densities with three components take the form:

$$p_{toy}(U) = \frac{1}{Z} e^{c_1 \beta \cos\theta} \cdot e^{c_2 \beta \cos(2\theta)} \cdot e^{c_3 \beta \cos(3\theta)}.$$

We test on one instance of these densities that is also used in Boyda et al. (2020), with $c_1 = .98$, $c_2 = -.63$, $c_3 = -.21$, and $\beta = 9$.