Data Augmentation vs. Equivariant Networks: A Theory of Generalization on Dynamics Forecasting

Rui Wang ¹ Robin Walters ² Rose Yu ¹

Abstract

Exploiting symmetry in dynamical systems is a powerful way to improve the generalization of deep learning. The model learns to be invariant to transformation, and hence is more robust to distribution shift. Data augmentation and equivariant networks are two major approaches to inject symmetry into learning. However, their exact role in improving generalization is not well understood. In this work, we derive the generalization bounds for data augmentation and equivariant networks, characterizing their effect on learning in a unified framework. Unlike most prior theories for the i.i.d setting, we focus on non-stationary dynamics forecasting with complex temporal dependencies.

1. Introduction

Symmetry plays an important role in the success of deep learning; incorporating symmetries into layers or training deep neural nets can improve generalizability and robustness (Bronstein et al., 2021; Shorten & Khoshgoftaar, 2019; Lopes et al., 2019; Wang et al., 2021c). There are two main techniques to train models that preserve symmetries. In data augmentation, one add samples to the training set which are transformed versions of other samples. This enables the model to learn invariance to symmetry transformations and noise (Hernández-García & König, 2018a; Dao et al., 2019; Hernández-García & König, 2018b; Rajput et al., 2019; Ratner et al., 2017; Zhou et al., 2022; Perez & Wang, 2017; Wen et al., 2020). The other line of work is the design of equivariant neural networks, which have also achieved remarkable success in learning image data (Worrall & Welling, 2019; Cohen & Welling, 2016b; Weiler & Cesa, 2019; Cohen et al., 2018) and physical dynamics (Wang et al., 2021c; 2022; Shi et al., 2021).

Here, we consider the problem of learning dynamical systems, where the data is non-i.i.d and the symmetry is rarely perfect. A perfectly equivariant model may have trouble learning partial or approximated symmetries in real-world data. Thus, some work has recently explored the idea of building approximately equivariant models and empirically demonstrated the benefits of it in modeling real-world data (van der Ouderaa et al., 2022; Romero & Lohit, 2021; Finzi et al., 2021). For example, (Wang et al., 2022) designed approximately equivariant models by relaxing the weight sharing schemes in the equivariant convolution networks.

Most prior works on data augmentation techniques and equivariant networks are purely empirical. There is no theory that characterizes and compares their behavior. Even though both approaches exploit the symmetry in the prediction task, their exact role to improve generalization is not well understood. Furthermore, it is not clear when these approaches are beneficial or in what regime one approach is preferred over the other. While there exist some theories of generalization for data augmentation, all of them are under the i.i.d. assumption (Rajput et al., 2019; Ratner et al., 2017; Sannai et al., 2021). For equivariant networks, the theory of generalization is even more scarce.

In this work, we present a theory of generalization for dynamics forecasting, where the data are non-stationary and non-mixing time series. We theoretically analyze and compare the generalization strength of data augmentation versus equivariant networks. We show that when the underlying dynamics is symmetric, equivariant networks achieve a tighter generalization bound than data augmentation. Furthermore, when the symmetries in the data are only approximate, the generalization bound for approximately equivariant networks (Wang et al., 2022) is further improved.

In summary, our contributions include:

• We formally characterize the behavior of dynamic forecasting with deep learning under the assumption that the underlying dynamical system preserves a certain amount of symmetry.

• We derive the generalization bounds for data augmentation and equivariant networks, including both the
strict and approximately equivariant networks for non-stationary and non-mixing time series.

- We prove that equivariant networks have a tighter generalization upper bound than data augmentation. When the data do not have perfect symmetries, approximately equivariant models tend to have better generalizability than the other two approaches.

2. Background

2.1. Statistical Learning Theory for Time Series.

We consider forecasting deterministic dynamics where the learner receives a data sample of the form \{X^{(1)}, ..., X^{(N)}\} with length \(T\) of a dynamical system (Wang et al., 2021a,b). Each time series \(X^{(i)}\) is a sample from a dynamical system where the system parameters are drawn i.i.d. from a given distribution. Even though the system parameters are independently sampled, each time series can be highly non-stationary and exhibit complex dependencies.

Denote \(Z_t^{(i)} = (X_{i,k−1:t−1}^{(i)}, X_t^{(i)})\) as the training sample (a subsequence of time series \(i\) at time \(t\)). \(X_{i,k−1:t−1}^{(i)}\) and \(X_t^{(i)}\) are the input and output of a forecasting model. For a loss function \(L : X' \times X' \rightarrow [0, \infty)\) and a hypothesis set \(\mathcal{F}\) of functions that map from \(X'\) to \(X'\), we want to minimize its empirical risk:

\[
R_n(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(f_\theta(X_{i,k−1:t−1}^{(i)}), X_t^{(i)})
\]

where \(\theta\) represents the parameters in \(f\). For simplicity, we use \(L(\theta, Z_t^{(i)})\) to denote \(L(f_\theta(X_{i,k−1:t−1}^{(i)}), X_t^{(i)})\). Note that \(q_1, ..., q_T\) are real numbers, which in the standard statistical learning scenarios are chosen to be all equal to \(\frac{1}{T}\). We follow the time series forecasting setting in (Kuznetsov & Mohri, 2020). For non-stationary dynamics, different \(Z_t\) may follow different distributions, and thus distinct weights could be assigned to the errors made on different sample points, depending on their relevance to forecasting the future \(Z_{t+1}\). The learning objective is to find a \(\theta\) that achieves a small test error, \(E[L(\theta, Z_{T+1})]\).

To derive the generalization bound, Bousquet et al. (2003) and Rakhlin & Sridharan (2014) generalize the classic Rademacher Complexity (Gnecco & Sanguineti, 2008) to time series learning, as defined below:

**Definition 2.1** (Sequential Rademacher Complexity, Bousquet et al. (2003); Rakhlin & Sridharan (2014)). Given a function class \(G \subset \mathbb{R}^Z\), we define the sequential Rademacher complexity of class \(G\) as:

\[
\mathcal{R}_T^G(G) = \mathbb{E}_z \mathbb{E}_\sigma [\sup_{g \in G} \sum_{t=1}^{T} \sigma_t q_t g(z_t(\sigma))]
\]

where \(z\) is a real-valued complete binary tree that is a sequence \((z_1, ..., z_T)\) of \(T\) mappings \(z_t : \{\pm 1\}^{T−1} \rightarrow \mathbb{R}\) for \(t \in [1, ..., T]\), and \(\sigma\) is a sequence of Rademacher random variables, which is also a path in the tree \(\sigma = (\sigma_1, ..., \sigma_{T−1}) \in \{\pm 1\}^{T−1}\).

In our forecasting setting, the sequential Rademacher complexity of a loss class can be defined more specifically as:

\[
\mathcal{R}_T^{sq}(L \circ \Theta) = \mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t L(\theta, Z_t^{(i)}) \right]
\]

2.2. Equivariance and Invariance.

Symmetry is often described through the equivariance or invariance of a given function. In this subsection, we give the formal definitions of data augmentation, equivariance and approximate equivariance in the deterministic dynamics forecasting setting.

**Equivariant Functions** A function \(f\) respecting the symmetry coming from a group \(G\) is said to be equivariant.

**Definition 2.2** (G-equivariant function). Assume a group representation \(\rho_{in}\) of \(G\) acting on \(X\) and \(\rho_{out}\) acting on \(Y\). We say a function \(f : X \rightarrow Y\) is **G-equivariant** if

\[
f(\rho_{in}(g)(x)) = \rho_{out}(g)f(x)
\]

for all \(x \in X\) and \(g \in G\). The function \(f\) is **G-invariant** if \(f(\rho_{in}(g)(x)) = f(x)\) for all \(x \in X\) and \(g \in G\). This is a special case of equivariance for the case \(\rho_{out}(g) = 1\). Equivariant neural networks (Cohen & Welling, 2016b; Weiler & Cesa, 2019; Cohen & Welling, 2016a) learn equivariant functions through weight-sharing and weight-tying.

We define equivariance error, which quantifies the amount of symmetry the function \(f\) contains.

**Definition 2.3** (Equivariance Error). Let \(f : X \rightarrow Y\) be a function and \(G\) be a group. Assume that \(G\) acts on \(X\) and \(Y\) via representation \(\rho_{in}\) and \(\rho_{out}\). Then the **equivariance error** of \(f\) is

\[
\|f\|_{EE} = \sup_{x,g} \|f(\rho_{in}(g)(x)) - \rho_{out}(g)f(x)\|.
\]

For strictly equivariant functions, we have \(\epsilon = 0\). But for real-world dynamics, the symmetry is often approximately equivariant, defined below:

**Definition 2.4** (Approximate Equivariance). \(f : X \rightarrow Y\) is \(\epsilon\)-approximately equivariant if and only if \(\|f\|_{EE} < \epsilon\).

Several recent work have designed approximately equivariant networks (Wang et al., 2022; van der Ouderaa et al., 2022; Finzi et al., 2021) to learn the approximate functions.
In this work, we assume the equivariance errors of trained approximately equivariant models is less or equal to the true data equivariance errors.

**Data Aug. Introduces Symmetry.** Consider a finite group $G$ that acts on the observed time series, we assume that for any $g \in G$, there is a certain amount of symmetry in the distribution, that is $Z_t^{(i)} \approx_d gZ_t^{(i)}$, $Z_t^{(i)} \sim \mathbb{P}$. We assume the group transformations are norm-preserving, i.e. $\|g\| = 1 \forall g \in G$, such as rotation and translation.

**Definition 2.5** (Data Augmentation). Given a finite group $G$, we assume the augmented samples are the original samples applied with transformations uniformly sampled from the group. In other words, for every sample $Z_t^{(i)}$ in the original training set, we add samples $\{gZ_t^{(i)}, g \in G\}$. Then the augmented training set is the $|G|$ times bigger than the original training set.

### 3. Data Augmentation vs. Equivariant Nets

We derive generalization bounds for data augmentation and equivariant networks. We show that the strictly equivariant networks can outperform data augmentation. When the underlying dynamics are approximately symmetric, approximately equivariant estimator can outperform both data augmentation estimator and strictly equivariant networks.

#### 3.1. Population and Empirical Risk Minimizers

We first define the population and the empirical risk minimizers for data augmentation, perfectly equivariant models and approximately equivariant models based on the dynamic forecasting setting defined in the previous section.

- Population minimizer: $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}[L(\theta, Z)]$
- Empirical minimizer:
  $$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_t^{(i)})$$
- Empirical minimizer for data augmentation:
  $$\hat{\theta}_G = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t E_G[L(\theta, gZ_t^{(i)})]$$
  $$\hat{\theta}_G = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \bar{L}(\theta, Z_t^{(i)})$$
  where $\bar{L}$ is the orbit-averaging loss because of data augmentation based on the definition in (Chen et al., 2020).
- Empirical minimizer for perfectly equivariant models:
  $$\hat{\theta}_E = \arg\min_{\theta \in \Theta_E} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_t^{(i)}),$$
  $$\Theta_E = \{\theta: f_\theta(\rho_{in}(g)(x)) = \rho_{out}(g)f_\theta(x), \forall g \in G\}$$
- Empirical minimizer for approximately equivariant nets:
  $$\hat{\theta}_{AE} = \arg\min_{\theta \in \Theta_{AE}} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_t^{(i)}),$$
  $$\Theta_{AE} = \{\theta: \sup_{g \in G}\|f_\theta(\rho_{in}(g)(x))\| = \rho_{out}(g)f_\theta(x)\| \leq \epsilon\}$$

where $\Theta$ is the parameter space without symmetry inductive biases imposed, $\Theta_E$ is the parameter space of all equivariant functions, and $\Theta_{AE}$ is the parameter space of all approximately equivariant functions.

### 3.2. Generalization Bound for Dynamics Forecasting

(Kuznetsov & Mohri, 2020) presented a data-dependent learning bound for the general scenario of non-stationary non-mixing stochastic processes. Yet, our focus is forecasting deterministic dynamics. Since the dynamics is non-stationary, we define a discrepancy measure to characterize the distributional shift between the training and test sets:

**Definition 3.1** (Discrepancy Measure). We use $\text{disc}_T(q)$ to denote the discrepancy between the target distribution and the distribution of the training samples.

$$\text{disc}_T(q) = \sup_{\theta \in \Theta} \left| \mathbb{E} \left[ \sum_{t=1}^{T} q_t L(\theta, Z_t) \right] - \mathbb{E}L(\theta, Z_{T+1}) \right|$$

We prove that the upper bound of the generalization error on dynamics forecasting is controlled by the sequential Rademacher complexity and the discrepancy measure of the temporal distributional shift.

**Theorem 3.2.** For any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $\hat{\theta} \in \Theta$ and all $\alpha = \|q\|_2/2 > 0$:

$$\mathbb{E}L(\hat{\theta}, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \leq 2\text{disc}_T(q) + 64\sqrt{4\pi \log \text{TR}_N} \|L \circ \Theta\| + \sqrt{\frac{2 \log(2/\delta)}{N}} + \|q\|_2(M \sqrt{8\log(1/\alpha) + 1})$$

Full proof can be found in the Appendix A.1. Note that our result is consistent with the conclusion in Kuznetsov & Mohri (2020) for stochastic dynamics.

### 3.3. Effect of Symmetry

We derive generalization bounds for forecasting non-stationary dynamics with data augmentation, perfectly equivariant networks, and approximately equivariant networks based on Theorem 3.2. Following Chen et al. (2020), we use the Wasserstein distance to measure the closeness of the original distribution to the distribution under group transformations. The full proofs of the following corollary 3.3.3, 3.5, 3.6 can be found in Appendix A.2.
We generalize Theorem 3.4 in Chen et al. (2020) from the i.i.d case to non-stationary dynamics forecasting:

**Corollary 3.3** (Data Augmentation). Let \( L(\theta, \cdot) \) be uniformly Lipschitz w.r.t. \( \theta \) with Lipschitz constant \( \|L\|_{\text{Lip}} \). For any \( \delta > 0 \), with probability at least \( 1 - \delta \), the following inequality holds:

\[
\mathbb{E}L(\hat{\theta}_G, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \\
\leq 2\text{disc}_T(q) + 6M\sqrt{4\pi \log T} \mathcal{R}_T^{eq}(L \circ \Theta) + \sigma \\
+ \max_{i} ||L||_{\text{Lip}} \cdot \mathbb{E}_G[\mathcal{W}(Z_{T+1}, gZ_{T+1}) + q_i \mathcal{W}(Z_{T+1}^{(i)}, gZ_{T+1}^{(i)})]
\]

where \( \sigma = \sqrt{\frac{2\log(2/\delta)}{N}} + ||q||_2(M\sqrt{8\log\frac{1}{\delta}} + 1) \).

We can see that the performance gain of data augmentation is governed by a bias term \( \max_{i} ||L||_{\text{Lip}} \cdot E_G[\mathcal{W}(Z_{T+1}, gZ_{T+1}) + q_i \mathcal{W}(Z_{T+1}^{(i)}, gZ_{T+1}^{(i)})] \), which vanishes under exact symmetry and the sequential Rademacher complexity reduction because of the group orbit averaging over the loss function.

The difference in sequential Rademacher complexity between the data augmentation estimator and regular estimator can further be bounded as:

\[
\mathcal{R}_T^{eq}(\hat{\theta}_G) - \mathcal{R}_T^{eq}(L \circ \Theta) \\
\leq \Delta + \max_{i} ||L||_{\text{Lip}} \cdot \mathbb{E}_G[q_i \mathcal{W}(Z_{T+1}^{(i)}, gZ_{T+1}^{(i)})]
\]

where \( \Delta = \mathbb{E}_\sigma[sup_{\theta \in \Theta} \sum_{t=1}^{T} \sigma_t q_t \mathbb{E}_G L(\theta, gZ_t)] - \mathbb{E}_\sigma \mathbb{E}_G[sup_{\theta \in \Theta} \sum_{t=1}^{T} \sigma_t q_t L(\theta, gZ_t)] \leq 0 \).

Here \( \Delta \) corresponds to the “variance reduction term” defined in Chen et al. (2020). When \( \Delta \) is small, data augmentation has a strong effect on improving generalizability.

To compare the generalization bounds of data augmentation and equivariant networks, we first need to prove the following lemma.

**Lemma 3.4.** \( \mathcal{R}_T^{eq}(L \circ \Theta_E) \leq \mathcal{R}_T^{eq}(\hat{\theta}_G) \)

The proof can be found at the end of Appendix A.2.

Next, we derive the generalization bound for strictly equivariant networks.

**Corollary 3.5** (Equivariant Networks). Let \( L(\theta, \cdot) \) be uniformly Lipschitz w.r.t. \( \theta \) with Lipschitz constant \( \|L\|_{\text{Lip}} \). For any \( \delta > 0 \), with probability at least \( 1 - \delta \), the following inequality holds:

\[
\mathbb{E}L(\hat{\theta}_E, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \\
\leq 2\text{disc}_T(q) + 6M\sqrt{4\pi \log T} \mathcal{R}_T^{eq}(L \circ \Theta_E) + \sigma \\
+ ||L||_{\text{Lip}} \cdot \mathbb{E}_G[\mathcal{W}(Z_{T+1}, gZ_{T+1})]
\]

where \( \sigma = \sqrt{\frac{2\log(2/\delta)}{N}} + ||q||_2(M\sqrt{8\log\frac{1}{\delta}} + 1) \).

From Lemma 3.4, we have \( \mathcal{R}_T^{eq}(L \circ \Theta_E) \leq \mathcal{R}_T^{eq}(\hat{\theta}_G) \).

Hence, Corollary 3.5 indicates that equivariant networks have a tighter generalization bound than data augmentation. In particular, the generalization bound of data augmentation in Corollary 3.3 has an additional bias term \( \max_{i} ||L||_{\text{Lip}} \cdot \mathbb{E}_G[q_i \mathcal{W}(Z_{T+1}^{(i)}, gZ_{T+1}^{(i)})] \). This term vanishes when the data are perfectly symmetric.

However, in real-world scenarios, the data are very rarely perfectly symmetric. We further analyze the generalization behavior of a class of approximate equivariant models:

**Corollary 3.6** (Approximate Equivariance). Let \( L(\theta, \cdot) \) be uniformly Lipschitz w.r.t. \( \theta \) with a Lipschitz constant \( \|L\|_{\text{Lip}} \). We assume \( \|\hat{\theta}_{AE}\|_{EE} \leq \|\theta^*\|_{EE} \) and \( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_i L(\theta^*, Z_{T+1}^{(i)}) \leq \xi \). For any \( \delta > 0 \), with probability at least \( 1 - \delta \), the following inequality holds:

\[
\mathbb{E}L(\hat{\theta}_{AE}, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \\
\leq 2\text{disc}_T(q) + 6M\sqrt{4\pi \log T} \mathcal{R}_T^{eq}(L \circ \Theta_{AE}) + \sigma \\
+ ||L||_{\text{Lip}} \cdot \mathbb{E}_G[\mathcal{W}(Z_{T+1}, gZ_{T+1})] - ||q||_1\|\hat{\theta}_{AE}\|_{EE} + 2\xi
\]

where \( \sigma = \sqrt{\frac{2\log(2/\delta)}{N}} + ||q||_2(M\sqrt{8\log\frac{1}{\delta}} + 1) \).

To put it simply, when the data do not have perfect symmetries, approximately equivariant models may have better prediction performance than data augmentations and perfectly equivariant models because of the term \(-||q||_1\|\hat{\theta}_{AE}\|_{EE}\) in the bound. The empirical error of the population minimizer \( \xi \) can be small enough to be ignored.

If approximately equivariant estimators can learn the correct amount of symmetry in the data, which means that \( \|\hat{\theta}_{AE}\|_{EE} \) is big and close to the true equivariance error in the data \( \|\theta^*\|_{EE} \), then they tend to have better generalizability. On the contrary, the estimators trained on a uniformly augmented training set and perfectly equivariant estimators maintain zero equivariance error even when data are not perfectly symmetric, which is overly restricted.

**4. Conclusion**

We take the first steps in the theoretical understanding of data augmentation and equivariant networks on the task of non-stationary dynamics forecasting. We derive the generalization bounds and show that strictly equivariant networks have a tighter upper bound than data augmentation, and that approximately equivariant estimators can outperform both data augmentation and perfectly equivariant networks on modeling imperfectly symmetric dynamics. A limitation of this work is that our theoretical comparison is only for upper bounds, which can be arbitrarily loose in practice. Future work includes improving the generalizing bounds with Pac-Bayesian analysis and deriving lower bounds for these approaches characterizing the hardness of.
learning for different model classes.

References

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Summer school on machine learning*, pp. 169–207. Springer, 2003.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478*, 2021.

Chen, S., Dobriban, E., and Lee, J. A group-theoretic framework for data augmentation. *Advances in Neural Information Processing Systems*, 33:21321–21333, 2020.

Cohen, T. S. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning (ICML)*, pp. 2990–2999, 2016a.

Cohen, T. S. and Welling, M. Steerable CNNs. *arXiv preprint arXiv:1612.08498*, 2016b.

Cohen, T. S., Geiger, M., Kohler, J., and Welling, M. Spherical CNNs. In *International Conference on Learning Representations (ICLR)*, 2018.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pp. 1528–1537. PMLR, 2019.

Finzi, M. A., Benton, G., and Wilson, A. G. Residual pathway priors for soft equivariance constraints. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=k505ekjMzww.

Gnecco, G. and Sanguineti, M. Approximation error bounds via rademacher complexity. 2008.

Hernández-García, A. and König, P. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018a.

Hernández-García, A. and König, P. Further advantages of data augmentation on convolutional neural networks. In *International Conference on Artificial Neural Networks*, pp. 95–103. Springer, 2018b.

Kuznetsov, V. and Mohri, M. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Annals of Mathematics and Artificial Intelligence*, 88(4): 367–399, 2020.

Lopes, R. G., Yin, D., Poole, B., Gilmer, J., and Cubuk, E. D. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiou, D. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pp. 5321–5330. PMLR, 2019.

Rakhlin, A. and Sridharan, K. Statistical learning and sequential prediction. *Book Draft*, 2014.

Ratner, A. J., Ehrenberg, H. R., Hussain, Z., Dunnmon, J., and Ré, C. Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, 30:3239, 2017.

Romero, D. W. and Lohit, S. Learning equivariances and partial equivariances from data. *arXiv preprint arXiv:2110.10211*, 2021.

Sannai, A., Imaiizumi, M., and Kawano, M. Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. In *Uncertainty in Artificial Intelligence*, pp. 771–780. PMLR, 2021.

Shi, C., Luo, S., Xu, M., and Tang, J. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, pp. 9558–9568. PMLR, 2021.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

van der Ouderaa, T. F., Romero, D. W., and van der Wilk, M. Relaxing equivariance constraints with non-stationary continuous filters. *arXiv preprint arXiv:2204.07178*, 2022.

Wang, R., Maddix, D., Faloutsos, C., Wang, Y., and Yu, R. Bridging physics-based and data-driven modeling for learning dynamical systems. In *Learning for Dynamics and Control*, pp. 385–398. PMLR, 2021a.

Wang, R., Walters, R., and Yu, R. Meta-learning dynamics forecasting using task inference. *arXiv preprint arXiv:2102.10271*, 2021b.

Wang, R., Walters, R., and Yu, R. Incorporating symmetry into deep dynamics models for improved generalization. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=wta_8Hx2KD.
Wang, R., Walters, R., and Yu, R. Approximately equivariant networks for imperfectly symmetric dynamics. arXiv preprint arXiv:2201.11969, 2022.

Weiler, M. and Cesa, G. General E(2)-equivariant steerable CNNs. In Advances in Neural Information Processing Systems (NeurIPS), pp. 14334–14345, 2019.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., and Xu, H. Time series data augmentation for deep learning: A survey. arXiv preprint arXiv:2002.12478, 2020.

Worrall, D. and Welling, M. Deep scale-spaces: Equivariance over scale. In Advances in Neural Information Processing Systems (NeurIPS), pp. 7364–7376, 2019.

Zhou, A., Tajwar, F., Robey, A., Knowles, T., Pappas, G. J., Hassani, H., and Finn, C. Do deep networks transfer invariances across classes? In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=Fn7i_r5rR0q.
A. Appendix

Theorem A.1. For any δ > 0, with probability at least 1 − δ, the following inequality holds for all \( \hat{\theta} \in \Theta \) and all \( \alpha = \|q\|_2/2 > 0 \):

\[
\mathbb{E}(\hat{\theta}, Z_{T+1}) - \mathbb{E}(\theta^*, Z_{T+1}) \leq 2\text{disc}_T(q) + \sqrt{\frac{2\log(2/\sigma)}{N}} + 6M \sqrt{\frac{4\log T R^{\text{eq}}_T(L \circ \Theta)}{N}} + \|q\|_2 \sqrt{8\log \frac{1}{\delta} + 1}
\]

Proof.

\[
\mathbb{E}(\hat{\theta}, Z_{T+1}) - \mathbb{E}(\theta^*, Z_{T+1}) = I + II + III + IV
\]

\[
I = \mathbb{E}(\hat{\theta}, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}, Z_{t}^{(i)})
\]

\[
II = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_{t}^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_{t}^{(i)}) \leq 0 \text{ (the model does not underfit the data)}
\]

\[
III = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_{t}^{(i)}) - \mathbb{E} \sum_{t=1}^{T} q_t L(\theta^*, Z_{t}) \leq \sqrt{\frac{2\log(2/\sigma)}{N}} \text{ (time series are i.i.d sampled)}
\]

\[
IV = \mathbb{E} \sum_{t=1}^{T} q_t L(\theta^*, Z_{t}) - \mathbb{E} L(\theta^*, Z_{T+1}) \leq \sup_{\theta \in \Theta} \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) - \mathbb{E} L(\theta, Z_{T+1}) = \text{disc}_T(q)
\]

Now we only need to bound the first term \( I = \mathbb{E}(\hat{\theta}, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}, Z_{t}^{(i)}) \)

\[
P(I - \text{disc}_T(q) > \epsilon)
\]

\[
= P(\mathbb{E}(\hat{\theta}, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}, Z_{t}^{(i)}) - \sup_{\theta \in \Theta} \mathbb{E} L(\theta, Z_{T+1}) - \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) > \epsilon)
\]

\[
\leq P(\sup_{\theta \in \Theta} \mathbb{E} L(\theta, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_{t}^{(i)}) - \sup_{\theta \in \Theta} \mathbb{E} L(\theta, Z_{T+1}) - \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) > \epsilon)
\]

\[
\leq P(\sup_{\theta \in \Theta} \mathbb{E} L(\theta, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_{t}^{(i)}) + \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) - \mathbb{E} L(\theta, Z_{T+1}) > \epsilon)
\]

\[
\leq P(\sup_{\theta \in \Theta} \mathbb{E} L(\theta, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_{t}^{(i)}) + \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) - \mathbb{E} L(\theta, Z_{T+1}) > \epsilon)
\]

\[
= P(\exp(\lambda \sup_{\theta \in \Theta} \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_{t}^{(i)}))) > \exp(\lambda \epsilon)
\]

\[
\leq \exp(-\lambda \epsilon) \mathbb{E} \exp(\lambda \sup_{\theta \in \Theta} \mathbb{E} \sum_{t=1}^{T} q_t L(\theta, Z_{t}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_{t}^{(i)}))) \text{ (by Markov’s inequality)}
\]
Thus,

\[ \mathbb{E}\left[ \exp(\lambda \sup_{\theta \in \Theta} \left( \mathbb{E} \left( \sum_{t=1}^{T} q_t L(\theta, Z_t) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta, Z_t^{(i)}) \right) \right) \right] \]

\[ = \mathbb{E}\left[ \exp(\lambda \sup_{\theta \in \Theta} \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t (L(\theta, Z_t) - L(\theta, Z_t^{(i)})) \right) \right] \]

\[ = \mathbb{E}\left[ \exp(\lambda \sup_{\theta \in \Theta} \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t (\mathbb{E}[L(\theta, Z_t)|Z_0] - L(\theta, Z_t^{(i)})) \right) \right] \]

\[ \leq \mathbb{E}\left[ \mathbb{E}_{Z_0 \sim \mathcal{X} \times \mathcal{Y}} \exp(\lambda \sup_{\theta \in \Theta} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t (L(\theta, Z_t) - L(\theta, Z_t^{(i)})) \right) | Z_0 \right] \right] \]

\[ = \mathbb{E}\left[ \exp(\lambda \sup_{\theta \in \Theta} \left[ \sum_{t=1}^{T} q_t (L(\theta, Z_t) - L(\theta, Z_t^{(i)})) \right) \right] \]

\[ = \mathbb{E}\left[ \exp(\lambda \sup_{\theta \in \Theta} \left[ \sum_{t=1}^{T} q_t (L(\theta, Z_t) - L(\theta, Z_t^{(i)})) \right) \right] \]

\[ = \mathbb{E}\left[ \exp(\lambda \sup_{\theta \in \Theta} \left[ \sum_{t=1}^{T} q_t (L(\theta, Z_t) - L(\theta, Z_t^{(i)})) \right) \right] \]

\[ = \mathbb{E}\left[ \mathbb{E}_{\sigma} \exp(\lambda \sup_{\theta \in \Theta} \left[ \sum_{t=1}^{T} \sigma_t q_t (L(\theta, Z_t) - L(\theta, Z_t^{(i)})) \right) \right] \]

\[ = \mathbb{E}_{\sigma} \left[ \exp(2\lambda \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \sigma_t q_t f(z_t^{\star}(\sigma))) \right] \]

Given \( z^{\star} \), let \( C \) be the minimal \( \alpha \)-cover of \( \mathcal{F} \) on \( z^{\star} \),

\[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \sigma_t q_t f(z_t^{\star}(\sigma)) \leq \max_{c \in C} \sum_{t=1}^{T} \sigma_t q_t c_t(\sigma) + \alpha \]

Thus,

\[ \mathbb{E}_{\sigma} \left[ \exp(2\lambda \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \sigma_t q_t f(z_t^{\star}(\sigma))) \right] \]

\[ \leq \exp(2\lambda \alpha) \mathbb{E}_{\sigma} \left[ \exp(2\lambda \max_{c \in C} \sum_{t=1}^{T} \sigma_t q_t c_t(\sigma)) \right] \]

\[ \leq \exp(2\lambda \alpha) \max_{c \in C} \mathbb{E}_{\sigma} \left[ \exp(2\lambda \sum_{t=1}^{T-1} \sigma_t q_t c_t(\sigma)) \right] \]

\[ = \exp(2\lambda \alpha) \max_{c \in C} \mathbb{E}_{\sigma} \left[ \exp(2\lambda \sum_{t=1}^{T-1} \sigma_t q_t c_t(\sigma)) \right] \mathbb{E}_{\sigma_{T:T-1}} \left[ \exp(2\lambda \sigma_T q_T c_T(\sigma)) | \sigma_{1:T-1} \right] \]

\[ \leq \exp(2\lambda \alpha) \max_{c \in C} \mathbb{E}_{\sigma} \left[ \exp(2\lambda \sum_{t=1}^{T-1} \sigma_t q_t c_t(\sigma)) \right] \exp(2\lambda^2 q_T^2 M^2) \]

\[ \leq \exp(2\lambda \alpha) \exp(2\lambda^2 \| q \|^2 M^2) \] (Iterate the last inequality over \( t \))

Then we have

\[ \mathbb{P}(I - \text{disc}_T(q) > \epsilon) \leq \mathbb{E}_z [\mathcal{N}_1(\alpha, \Theta, z)] \exp(2\lambda \alpha - \lambda \epsilon + 2\lambda^2 \| q \|^2 M^2) \]
Corollary A.2. Let $L(\theta, \cdot)$ be Lipschitz uniformly over $\theta$, with a Lipschitz constant $\|L\|_{lip}$. Assume
\[ \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_t^{(i)}) \leq \xi. \] For any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $\alpha = \|q\|_2/2 > 0$:
\[
\mathbb{E}L(\hat{\theta}_G, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \leq 2\text{disc}_T(q) + 6M \sqrt{\frac{\log T}{N}} \max_{t,i} \|L\|_{lip} \cdot \mathbb{E}_{G}[W(Z_{T+1}, gZ_{T+1}) + q_t W(Z_t^{(i)}, gZ_t^{(i)})]
\]
\[
\mathbb{E}L(\hat{\theta}_E, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \leq 2\text{disc}_T(q) + 6M \sqrt{\frac{\log T}{N}} \max_{t,i} \|L\|_{lip} \cdot \mathbb{E}_{G}[W(Z_{T+1}, gZ_{T+1}) + q_t W(Z_t^{(i)}, gZ_t^{(i)})]
\]
\[
\mathbb{E}L(\hat{\theta}_{AE}, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) \leq 2\text{disc}_T(q) + 6M \sqrt{\frac{\log T}{N}} \max_{t,i} \|L\|_{lip} \cdot \mathbb{E}_{G}[W(Z_{T+1}, gZ_{T+1}) + q_t W(Z_t^{(i)}, gZ_t^{(i)})] + \Delta
\]

where $\Delta = \sqrt{\frac{2\log(2/\delta)}{N}} + \|q\|_2(M \sqrt{\frac{8\log 4}{\delta}} + 1)$.

Proof. When $\hat{\theta} = \hat{\theta}_G$, we only need to derive a bound for $I$ in the previous proof.

\[ I = \mathbb{E}L(\hat{\theta}, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}, Z_t^{(i)}) = A + B + C \]

\[ A = \mathbb{E}L(\hat{\theta}, Z_{T+1}) - \mathbb{E}G L(\hat{\theta}, gZ_{T+1}) \leq \mathbb{E}G |\mathbb{E}L(\hat{\theta}_G, Z_{T+1}) - \mathbb{E}L(\hat{\theta}_G, gZ_{T+1})| \leq \|L\|_{lip} \cdot \mathbb{E}_{G}[W(Z_{T+1}, gZ_{T+1})] \]

\[ B = \mathbb{E}G L(\hat{\theta}, gZ_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \mathbb{E}G L(\hat{\theta}, gZ_t^{(i)}) \leq \text{disc}_T(q) + 6M \sqrt{\frac{\log T}{N}} \max_{t,i} \|L\|_{lip} \cdot \mathbb{E}_{G}[W(Z_{T+1}, gZ_{T+1})] + \Delta \]

\[ C = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \mathbb{E}G L(\hat{\theta}, gZ_t^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}, Z_t^{(i)}) \leq \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t [\mathbb{E}G L(\hat{\theta}_G, gZ_t^{(i)}) - L(\hat{\theta}_G, Z_t^{(i)})] \]

\[
\leq \max_{t,i} \|L\|_{lip} \cdot \mathbb{E}_{G}[W(Z_t^{(i)}, gZ_t^{(i)})]
\]

When $\hat{\theta} = \hat{\theta}_{AE}$:
\[
\mathbb{E}L(\hat{\theta}_{AE}, Z_{T+1}) - \mathbb{E}L(\theta^*, Z_{T+1}) = I + II + III + IV + V + VI
\]
\[ I = EL(\hat{\theta}_{AE}, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}_{AE}, Z_t^{(i)}) \leq \text{disc}_T(q) + 6M \sqrt{4\pi \log T} R_T^{\alpha}(L \circ \Theta_{AE}) + \Delta \]

\[ II = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}_{AE}, Z_t^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_t^{(i)}) \leq 0 \text{ (The model does not underfit the data)} \]

\[ IV = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) - \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t) \leq \sqrt{\frac{2 \log(2/\sigma)}{N}} \]

\[ V = \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) - \mathbb{E}_G L(\theta^*, gZ_{T+1}) = \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) - \mathbb{E}_G L(\theta^*, gZ_{T+1}) \]

\[ \leq \sup_{\theta \in \Theta} \left| E \left[ \sum_{t=1}^{T} q_t \mathcal{L}(\theta, Z_t) \right] - \mathbb{E}_G \mathcal{L}(\theta, Z_{T+1}) \right| \leq \text{disc}_T(q) \]

\[ VI = \mathbb{E}_G L(\theta^*, gZ_{T+1}) - EL(\theta^*, Z_{T+1}) \leq \|L\|_{\text{Lip}} \cdot \mathbb{E}_G \mathcal{W}(Z_{T+1}, gZ_{T+1}) \]

\[ III = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_t^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) \ 	ext{(let } x_t^{(i)} = X_{t-1:t-1}^{(i)} \text{ and } y_t^{(i)} = X_t^{(i)}) \]

When \( \hat{\theta} = \hat{\theta}_E \):

\[ EL(\hat{\theta}_E, Z_{T+1}) - EL(\theta^*, Z_{T+1}) = I + II + III + IV + V + VI \]

\[ I = EL(\hat{\theta}_E, Z_{T+1}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}_E, Z_t^{(i)}) \leq \text{disc}_T(q) + 6M \sqrt{4\pi \log T} R_T^{\alpha}(L \circ \Theta_E) + \Delta \]

\[ II = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\hat{\theta}_E, Z_t^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_t^{(i)}) \leq 0 \text{ (The model does not underfit the data)} \]

\[ III = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t L(\theta^*, Z_t^{(i)}) - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) = 0 \]

\[ IV = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) - \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t) \leq \sqrt{\frac{2 \log(2/\sigma)}{N}} \]

\[ V = \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) - \mathbb{E}_G L(\theta^*, gZ_{T+1}) = \sum_{t=1}^{T} q_t \mathbb{E}_G L(\theta^*, gZ_t^{(i)}) - \mathbb{E}_G L(\theta^*, gZ_{T+1}) \]

\[ \leq \sup_{\theta \in \Theta} \left| E \left[ \sum_{t=1}^{T} q_t \mathcal{L}(\theta, Z_t) \right] - \mathbb{E}_G \mathcal{L}(\theta, Z_{T+1}) \right| \leq \text{disc}_T(q) \]

\[ VI = \mathbb{E}_G L(\theta^*, gZ_{T+1}) - EL(\theta^*, Z_{T+1}) \leq \|L\|_{\text{Lip}} \cdot \mathbb{E}_G \mathcal{W}(Z_{T+1}, gZ_{T+1}) \]

Combining the bounds for the six terms gives the desired result.
Moreover,

\[ R_{T}^{eq}(L \circ \Theta_E) - R_{T}^{eq}(\bar{L} \circ \Theta) \]
\[ \leq E_{\sigma} [\sup_{\theta \in \Theta_E} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t L(\theta_E, Z_i)] - E_{\sigma} [\sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t \mathbb{E}_{G}(L(\theta, gZ_i))] \]
\[ \leq E_{\sigma} [\sup_{\theta \in \Theta_E} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t \mathbb{E}_{G}(L(\theta_E, gZ_i))] - E_{\sigma} [\sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t \mathbb{E}_{G}(L(\theta, gZ_i))] \]
\[ = E_{\sigma} \mathbb{E}_{G} [\sup_{\theta \in \Theta_E} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t L(\theta_E, gZ_i)] - \sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t L(\theta, gZ_i)] \]
\[ \leq E_{\sigma} \mathbb{E}_{G} [\sup_{\theta \in \Theta_E} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t L(\theta_E, gZ_i)] - \sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_t q_t L(\theta, gZ_i)] \]
\[ = 0 \]