Summary of Mouli and Ribeiro's work [MR21]

and

Testing for group invariance using kernel hypothesis tests

STAT 548 Qualifying Paper

Kenny Chiu

October 23, 2021

1 Conceptual summary

The paper by Mouli and Ribeiro [MR21] examines the problem of extrapolating patterns learned from single-environment training data in a supervised setting to data from other environments. This problem context falls under the topic of *domain adaptation* that has been explored in recent literature [Far+20]. However, a key assumption in Mouli and Ribeiro's work that distinguishes it from much of the previous work in the literature is that the training data come from a single environment as opposed to multiple environments. Several previously proposed methods for domain adaptation—such as *Invariant Risk Minimization* [Arj+20] (IRM)—rely on training data from multiple environments and therefore would fail under this problem context. Mouli and Ribeiro take a different approach by viewing extrapolation as counterfactual reasoning in a specified structural causal model (SCM) and assuming that potential differences between environments can be described in terms of known linear transformation groups acting on the data. Under this formulation, Mouli and Ribeiro introduce a neural network learning framework for the single-environment problem that is able to learn the group invariances that do not contradict the data. In this conceptual summary, we discuss how the context and work of Mouli and Ribeiro [MR21] differ from previous work in the literature, review the key contributions of their work, and highlight the limitations of their approach.

1.1 Related work

Various methods for domain adaptation have been proposed in the literature, but the majority of these methods are not appropriate for the single-environment problem described by Mouli and Ribeiro [MR21]. For example, methods based on the *Independent Causal Mechanisms* (ICM) principle [Par+18] and other causal-based methods are generally based on learning some internal representation of the data that is invariant to non-causal environment information. The invariance in the representation is learned from the training data, which is assumed to come from multiple environments. When the data come from a single environment, the representation cannot distinguish which aspects of the data are environment-specific and so the learned representation is unlikely to extrapolate to new environments. The learning framework proposed by Mouli and Ribeiro works with single-environment data and has an advantage over existing methods in these settings.

Another common approach to domain adaptation is based on data augmentation [CDL20] where training is done with not only the original data but also proper transformations of the data. By augmenting the training data with seemingly irrelevant transformations, the aim is to desensitize the representation to these transformations and therefore learn invariance. Mouli and Ribeiro explain that data augmentation is a type of *forced group invariance* (i.e., forced *G-invariance*) where certain transformations of the data may actually introduce contradictions (e.g., trying to enforce rotation invariance in images of digits, but digits 6 and 9 are not invariant to 180^o rotations). Like in data augmentation, Mouli and Ribeiro's proposed framework starts with an a priori set of potential invariances (in the form of known groups rather than data), but the framework differs in that it then "unlearns" the invariances that contradict the training data.

While the single-environment problem is not entirely novel in the domain adaptation literature, the context of the problem and the proposed approaches to solve it vary greatly across works. For example, Kumar et al. [Kum+20] study reinforcement learning in the setting where only a single training Markov decision process is available. The *single-source unsupervised domain adaptation* literature examines problems where labeled data is only available from a single source and labels for data from other sources have to be predicted [Zha+20]. Mouli and Ribeiro's work fits into this literature but differs from most others in terms of its problem formulation and setup.

1.2 Main contributions

The main contributions of Mouli and Ribeiro [MR21] include a formulation of the single-environment extrapolation problem, a learning framework for neural networks that aims to learn the non-contradicting invariances, and an empirical evaluation of standard neural networks versus neural networks trained using

the proposed learning framework.

Mouli and Ribeiro's formulation of the single-environment extrapolation problem is based on the ICM principle and involves a SCM describes the causal and non-causal relationships between the variables [Sch19]. Extrapolation is then viewed as counterfactual reasoning where being able to extrapolate to different environments is tied to the output being invariant to interventions on non-causal environment variables. Mouli and Ribeiro extend this idea by assuming that differences between environments can be described in terms of known linear automorphism groups that act on the variables. Being able to extrapolate a representation is then equivalent to the representation being counterfactually group-invariant (i.e., *CG-invariant*) to the groups that act on non-causal variables. This additional assumption is the crux of the formulation that allows the proposed framework to work with only single-environment data.

The learning framework aims to learn an internal representation that is CG-invariant to the groups that do not contradict the training data. While G-invariances are easier to work with in practice, Mouli and Ribeiro [MR21] show that CG-invariance is stronger than G-invariance (Theorem 1). However, they also show that when the subset of groups acting on the non-causal variables is a normal subgroup of the overgroup acting on all variables, then G-invariance also implies CG-invariance (Theorem 2). These results establish the conditions under which it is sufficient for the model to learn G-invariances in place of CG-invariances, and it is for these reasons that Mouli and Ribeiro also assume that the subgroup acting on non-causal variables is normal to the overgroup on all variables.

The challenge in learning the G-invariances that do not contradict the training data is due to the fact that the subset of non-causal variables among all variables is unknown. To learn the invariances for the unknown set, Mouli and Ribeiro require the groups to be finite linear automorphisms. The *Reynolds operator*—a group-invariant transformation—can then be constructed by averaging over members of the particular group (Lemma 1). The Reynolds operator is a projection operator with eigenvalues 1 and 0. The left eigenspace spanned by eigenvectors with eigenvalue 1 represents the space of vectors that are invariant to transformations of the group (Lemma 2). To construct the subspace that is invariant to transformations of a specific set of groups, the intersection of the 1-eigenspaces for all groups in the set is taken, and the projection of the intersection onto the subspace of all overgroups is then removed from the intersection (Theorem 3). The invariant subspace is computed for each set in the power set of groups, and the invariant subspaces are partially ordered by their invariance strength (i.e., the number of groups that the subspace is invariant to). A basis for each subspace is then computed and encoded into a neural network where the learned parameters are neuron weights representing the coefficients for each basis. The framework's optimization objective then includes a regularization term that encourages the network representation to use the strongest G-invariance (i.e., have a non-zero weight) that does not significantly contradict the data, and to avoid invariances (i.e., have zero weights) that are lower-order or contradicting. The key aspects of Mouli and Ribeiro's proposed framework include needing to specify known groups, requiring the groups to be finite linear automorphisms and, in doing so, being able to learn the G-invariances that do not contradict the data.

Mouli and Ribeiro [MR21] evaluated neural networks trained using their proposed learning framework on various image tasks and array tasks. Their results broadly suggest that

1. standard neural networks do well when interpolating but not when extrapolating,
2. neural networks trained with forced G-invariances do poorly when interpolating but do well when extrapolating, and
3. neural networks trained with their learning framework generally do well when interpolating and when extrapolating.

Based on these conclusions, there appears to be merit in their proposed framework, and their approach may be worth further exploring in future work.

1.3 Limitations

The main limitations of the framework proposed by Mouli and Ribeiro [MR21] are the very specific assumptions required for the framework to work. To allow extrapolation of the model trained on single-environment data to different environments, the framework requires that the invariance groups acting on the data are specified a priori. Furthermore, to enable automatic learning of invariances that do not contradict the training data, the groups are restricted to be finite linear automorphisms. These restrictions imply that invariance groups that were not initially specified are unable to be learned. The framework also cannot be used if the differences between environments could not be expressed in terms of linear transformation groups that act on the data. These limitations naturally point to future work in the form of an extended framework that allows one or more of these assumptions to be violated.

2 Technical summary

The main technical aspects of the paper by Mouli and Ribeiro [MR21] include the proposed neural network learning framework and the theoretical results that justify its usage in the described problem setting. In this technical summary, we introduce the formulation and notation of the single-environment extrapolation problem, discuss the assumptions that are made and why, and explain how the proposed learning framework is used with a neural network.

Note that the definitions of terms and acroynyms used in this technical summary are defined in the conceptual summary in Section 1.

2.1 Single-environment extrapolation setting

In the setting of single-environment extrapolation described by Mouli and Ribeiro [MR21], the goal is to train a prediction model (under a supervised learning setup) that is able to perform well (i.e., extrapolate) across different environments when given only data from a single environment at training time. It is assumed that input data include causal and non-causal variables in relation to the output, and that environments differ only in the non-causal variables. In theory, the model should be able to extrapolate if it depends only on the causal variables. The challenge is then learning which information is causal and which is not given only training data from a single environment. To simplify the problem, Mouli and Ribeiro assume that differences in the data can be described by a given set of finite linear transformation groups $\mathcal{G}_1, \ldots, \mathcal{G}_m$ acting on the data. The objective is then to determine which of these groups correspond to non-causal information and to learn an internal representation of the data that is invariant to these non-causal groups.

Mouli and Ribeiro's formulation of the problem follows the ICM principle [Par+18] and assumes the SCM in Figure 1.

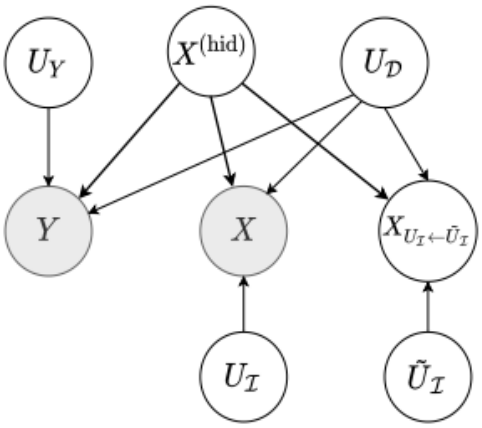

Figure 1: Assumed SCM in single-environment extrapolation. Grey nodes are observed variables. Arrows denote deterministic causal relationships where the target is dependent on the source. Figure from [MR21].

The variables in the SCM in Figure 1 are defined as follows:

- $\mathcal{D}$, $\mathcal{I}$: unknown disjoint sets of indices that refer to the sets of transformation groups that are relevant and irrelevant to the output, respectively. $\mathcal{D} \cup \mathcal{I} = \{1, \ldots, m\}$.
- U_Y, $U_{\mathcal{I}}$, $U_{\mathcal{D}}$, $\widetilde{U}_{\mathcal{I}}$: independent latent variables that model the stochastic components in the SCM. $U_{\mathcal{I}}$ and $U_{\mathcal{D}}$ in particular represent the collective randomness in their respective set of groups.

- $X^{(\text{hid})}$: unknown canonical form of the observed input $X \in \mathcal{X}$. It is assumed that given $U_{\mathcal{D}}$ and $U_{\mathcal{I}}$, X was obtained from an ordered sequence of transformations $T_{U_{\mathcal{D}},U_{\mathcal{I}}}$ applied to the canonical form, i.e.,

$$X = T_{U_{\mathcal{D}},U_{\mathcal{I}}} \circ X^{(\text{hid})} .$$

  The overgroup $\mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$ consists of transformations of the form

$$T_{U_{\mathcal{D}},U_{\mathcal{I}}} = T_{\mathcal{I}}^{(1)} \circ T_{\mathcal{D}}^{(1)} \circ T_{\mathcal{I}}^{(2)} \circ \dots$$

  where $T_{\mathcal{D}}^{(j)}$ is a transformation in group $\mathcal{G}_j$ from the overgroup $\mathcal{G}_{\mathcal{D}} = \langle \cup_{j\in\mathcal{D}} \mathcal{G}_j \rangle$, and similarly $T_{\mathcal{I}}^{(i)} \in \mathcal{G}_i \subset \mathcal{G}_{\mathcal{I}} = \langle \cup_{i\in\mathcal{I}} \mathcal{G}_i \rangle$. Note that Mouli and Ribeiro also assume that $\mathcal{G}_{\mathcal{I}}$ is a normal subgroup of $\mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$ (denoted $\mathcal{G}_{\mathcal{I}} \trianglelefteq \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$).

- Y: observed output assumed to be generated by

$$Y = h(X^{(\text{hid})}, U_{\mathcal{D}}, U_Y)$$

  where h is a deterministic function.

- $X_{U_{\mathcal{I}}\leftarrow\widetilde{U}_{\mathcal{I}}}$: counterfactual variable to X where $U_{\mathcal{I}}$ has been replaced by $\widetilde{U}_{\mathcal{I}}$, i.e.,

$$X_{U_{\mathcal{I}}\leftarrow\widetilde{U}_{\mathcal{I}}} = T_{U_{\mathcal{D}},\widetilde{U}_{\mathcal{I}}} \circ X^{(\text{hid})} .$$

Given the SCM, the goal is to learn a representation $\Gamma : \mathcal{X} \to \mathbb{R}^d$, $d \geq 1$, that is CG-invariant, i.e.,

$$\Gamma(X) = \Gamma(X_{U_{\mathcal{I}}\leftarrow\widetilde{U}_{\mathcal{I}}})$$

where the equality implies $\Gamma(X_{U_{\mathcal{I}}\leftarrow u}) = \Gamma(X_{U_{\mathcal{I}}\leftarrow u'})$ for all $u \in \text{supp}(U_{\mathcal{I}})$, $u' \in \text{supp}(\widetilde{U}_{\mathcal{I}})$. The learned representation Γ is fed into a learned link function $g : \mathbb{R}^d \to \text{Im}P(Y = y|X)$, $\text{Im}P(\bullet)$ being the image of $P(\bullet)$, which produces the prediction of the model, i.e.,

$$\hat{Y}|X \sim g(\Gamma(X)) .$$

For training data $X^{(\text{tr})}$, if

$$Y|X^{(\text{tr})} \overset{\text{d}}{=} \hat{Y}|X^{(\text{tr})} \sim g_{\text{true}}(\Gamma_{\text{true}}(X^{(\text{tr})}))$$

and $\Gamma_{\text{true}}(X) = \Gamma_{\text{true}}(X_{U_{\mathcal{I}}\leftarrow\widetilde{U}_{\mathcal{I}}})$, then $g_{\text{true}} \circ \Gamma_{\text{true}}$ extrapolates to test data $X^{(\text{te})}$ in the sense that

$$Y|X^{(\text{te})} \overset{\text{d}}{=} \hat{Y}|X^{(\text{te})} \sim g_{\text{true}}(\Gamma_{\text{true}}(X^{(\text{te})})) .$$

2.2 Assumptions and restrictions in single-environment extrapolation

Mouli and Ribeiro [MR21] make a number of assumptions in the setup described in the previous section in order to simplify the extrapolation problem and to allow for a feasible learning framework. The main assumptions that distinguish the current work from previous work in the literature are the ones revolving around the groups acting on the data.

Unlike previous work that assumes the availability of training data from multiple environments, the problem setting considered by Mouli and Ribeiro specifically considers data from a single environment. Without additional information that suggests how data from different environments may differ, it is difficult to learn which pieces of information are environment-specific and irrelevant to the output. Mouli and Ribeiro deal with this issue by assuming a priori knowledge of how environments may differ in the form of transformation groups. The assumed groups specify the potential ways data from different environments may differ, and it is left to the learning framework to "unlearn" the groups that contradict the training data.

Furthermore, Mouli and Ribeiro assume that the subset $\mathcal{G}_{\mathcal{I}}$ of groups is a normal subgroup of the overgroup $\mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$. This assumption is a consequence of Theorems 1 and 2, which together state that CG-invariances are G-invariances, i.e.,

$$\Gamma(X) = \Gamma(T_{\mathcal{I}} \circ X)$$

for all $T_{\mathcal{I}} \in \mathcal{G}_{\mathcal{I}}$, but G-invariances are CG-invariances only when $\mathcal{G}_{\mathcal{I}} \trianglelefteq \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$. As it is easier to check G-invariances than CG-invariances in practice due to its simpler definition, the normal subgroup assumption is made in order to make learning G-invariances sufficient for the objective. The proof of Theorem 1 (CG-invariance $\Rightarrow$ G-invariance) relies on the fact that for any transformation $T_{\mathcal{I}} \in \mathcal{G}_{\mathcal{I}}$, we can rewrite

$$T_{\mathcal{I}} \circ X = T_{\mathcal{I}} \circ T_{U_{\mathcal{D}}, U_{\mathcal{I}} \leftarrow u} \circ X^{(\text{hid})} = T_{U_{\mathcal{D}}, U_{\mathcal{I}} \leftarrow \tilde{u}} \circ X^{(\text{hid})}$$

where $T_{U_{\mathcal{D}}, \bullet} \in \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$ and $u, \tilde{u} \in U_{\mathcal{I}}$. The result then follows from repeated applications of CG-invariance and G-invariance definitions for a representation Γ. To show that not all G-invariances are CG-invariances, the counterexample in Figure 2 is given. The proof of Theorem 2 (G-invariance $\Rightarrow$ CG-invariance when $\mathcal{G}_{\mathcal{I}} \trianglelefteq \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$) uses the fact that under the normal subgroup assumption, $T_{\mathcal{D}} \circ T_{\mathcal{I}} \circ T_{\mathcal{D}}^{-1} \in \mathcal{G}_{\mathcal{I}}$ for all $T_{\mathcal{D}} \in \mathcal{G}_{\mathcal{D}}$, $T_{\mathcal{I}} \in \mathcal{GI}$. This also implies that

$$T_{\mathcal{D}} \circ T_{\mathcal{I}} = T'_{\mathcal{I}} \circ T_{\mathcal{D}}$$

for some $T'_{\mathcal{I}} = T_{\mathcal{D}} \circ T_{\mathcal{I}} \circ T_{\mathcal{D}}^{-1}$. Therefore for any transformation $T_{\mathcal{I}}^{(1)} \circ T_{\mathcal{D}}^{(1)} \circ T_{\mathcal{I}}^{(2)} \circ \ldots \in \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$, we can show CG-invariance for a representation Γ by using G-invariance to remove the leading $\mathcal{G}_{\mathcal{I}}$ transformation, applying the above fact to swap the order of the new leading $\mathcal{G}_{\mathcal{D}}$ and $\mathcal{G}_{\mathcal{I}}$ transformations, and repeating the procedure until only $\mathcal{G}_{\mathcal{D}}$ transformations remain. CG-invariance then follows from definition. It is worth noting that Theorems 1 and 2 do not make additional assumptions on the groups themselves, and so these results generalize to groups beyond finite linear automorphisms.

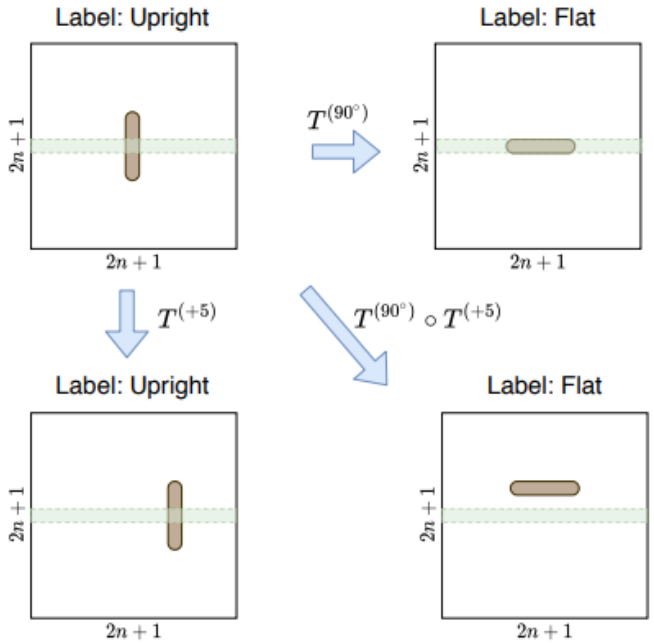

Figure 2: A counterexample that shows not all G-invariances are CG-invariances. The goal of the example problem is to determine the orientation of an upright or flat rod in an image. A representation that sums the middle row of the image is (G-)invariant to horizontal translations of the image but not invariant to 90^o rotations. Applying a translation before a rotation may result in a representation different from just applying a rotation, and so the representation is not CG-invariant. Figure from [MR21].

While a set of transformation groups are assumed a priori, it is not known which groups represent information causally related to the output. Additionally, some of the groups may actually specify transformations that contradict the training data. Mouli and Ribeiro address this problem in their proposed framework by restricting the groups to be finite linear automorphisms. Under this restriction, the Reynolds operator given by

$$\bar{T} = \frac{1}{|\mathcal{G}|} \sum_{T \in \mathcal{G}} T$$

can be computed for each group $\mathcal{G}_1, \ldots, \mathcal{G}_m$. The Reynolds operator is a projection operator (Lemma 1) and so it only has 1 and 0 eigenvalues. The 1-eigenspace $\mathcal{W}$ of the Reynolds operator for a group $\mathcal{G}$ corresponds to the subspace of linear transformations $\gamma(x; w, b) = w^T x + b$, $b \in \mathbb{R}$, that is invariant to all transformations $T \in \mathcal{G}$, i.e.,

$$\gamma(Tx; w, b) = \gamma(x; w, b)$$

if and only if $w \in \mathcal{W}$ (Lemma 2). The invariant subspace $\mathcal{B}_M$ for a set of groups can then be computed by taking the intersection of the invariant subspaces $\mathcal{W}_i$ for group $\mathcal{G}_i$ in the set, i.e.,

$$\widetilde{B}_M = \bigcap_{i \in M} \mathcal{W}_i \ ,$$

and removing from the intersection its projection onto the subspace $\mathcal{B}_{\supsetneq M} = \bigoplus_{N \supsetneq M} \mathcal{B}_N$ formed by direct sums of the invariant subspace of all overgroups (Theorem 3). Removal of the projection implies that the resulting subspace contains vectors only invariant to all the groups in the set and no more. Thus, there is a partial ordering on the invariant subspaces where a subspace corresponding to a set of groups is defined to have "stronger" invariance than a subspace corresponding to a smaller subset of groups. Mouli and Ribeiro's framework is then based on finding the strongest invariant subspace that does not contradict the training data.

Lemma 1 (Reynolds operator is a projection operator) is a well-known result that applies to any group with a finite measure [MFK94]. However, by restricting the groups to be finite linear automorphisms, the Reynolds operator itself takes the form of a computable matrix. Therefore, its eigenspaces can be found through standard linear algebraic methods, which is beneficial for practical use. Lemma 2 (linear transformation is group-invariant if and only if its inner product vector is in the 1-eigenspace of the Reynolds operator) is a stepping stone to Theorem 3, and is mainly used to show that considerations can be restricted to the 1-eigenspace of the Reynolds operator if the goal is group invariance. Its proof follows from direct logical derivations in both directions where sufficiency is shown using the projection property of the Reynolds operator as well as the definition of eigenvectors, and necessity is shown by reasoning that any vector in the invariant subspace is an eigenvector of the Reynolds operator. Theorem 3 (invariant subspaces can be partially ordered by invariance strength) is the main result that provides rationale for the proposed learning framework: because the invariant subspaces can be partially ordered by some notion of strength, take the strongest invariant subspace that excludes only invariances to groups that lead to contradictions in the data. Its proof consists of proving each substatement in the theorem, and each subproof is generally based on direct logical reasoning. The proof first shows that there is a hierarchy of subspaces $\mathcal{B}_M$ and $\mathcal{B}_{\supsetneq M}$ using induction on M starting with $M = \{1, \ldots, m\}$. Using properties of the subspace construction and Lemma 2, the proof then proceeds to show that the subspaces $\mathcal{B}_M$ and $\mathcal{B}_{\supsetneq M}$ are orthogonal for any M, and that the vectors in a particular subspace are invariant to only groups that the subspace was constructed from.

2.3 Learning framework for single-environment extrapolation

Under the context and assumptions described in the previous sections, the framework proposed by Mouli and Ribeiro [MR21] aims to learn a CG-invariant representation Γ and a link function g, both of which are in the form of a neural network. The representation Γ is a neural network layer with $H \geq 1$ neurons. The

h-th neuron has the form

$$\Gamma^{(h)}(x) = \sigma\left(x^T\left(\sum_{i=1}^{B}\mathbf{B}_{M_i}\boldsymbol{\omega}_{M_i,h}\right) + b_h\right)$$

where $\sigma(\bullet)$ is a non-linear activation function, b_h is a bias parameter, $\mathbf{B}_{M_i}$ is a matrix whose columns are the orthogonal basis of the invariant subspace $\mathcal{B}_{M_i}$ built from the set of groups indexed by M_i, and $\boldsymbol{\omega}_{M_i,h}$ are the learnable parameters which correspond to the linear combination coefficients of the basis. The parameters are collected in a neuron weight matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\omega}_{M_1,1} & \dots & \boldsymbol{\omega}_{M_1,H} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\omega}_{M_B,1} & \dots & \boldsymbol{\omega}_{M_B,H} \end{bmatrix}$$

where $M_1, \dots, M_B$, $B \leq \dim(\mathcal{X})$, are sets of indices corresponding to different subsets of groups. The optimization objective is then

$$\widehat{\boldsymbol{\Omega}}, \widehat{\mathbf{b}}, \widehat{\mathbf{W}}_g = \underset{\boldsymbol{\Omega},\mathbf{b},\mathbf{W}_g}{\arg\min} \sum_{(y^{(\mathrm{tr})},x^{(\mathrm{tr})})\in\mathcal{D}^{(\mathrm{tr})}} \mathcal{L}\left(y^{(\mathrm{tr})}, g(\Gamma(x^{(\mathrm{tr})};\boldsymbol{\Omega},\mathbf{b});\mathbf{W}_g)\right) + \lambda R(\boldsymbol{\Omega})$$

where $\mathbf{W}_g$ is the parameters of g, $\lambda > 0$ is a regularization parameter, $R(\bullet)$ is the regularization penalty given by

$$R(\boldsymbol{\Omega}) = |\{M_i : |M_i| > \ell, 1 \leq i \leq B\}| + \sum_{i:|M_i|=\ell, 1\leq i\leq B} \mathbf{1}\{\|\boldsymbol{\omega}_{M_i,\bullet}\|_2^2 > 0\}\,,$$

and $\ell = \min\{|M_i| \cdot \mathbf{1}\{\|\boldsymbol{\omega}_{M_i,\bullet}\|_2^2 > 0\} : 1 \leq i \leq B\}$. The integer ℓ describes the size of the smallest set of groups that is used by at least one neuron (i.e., non-zero weight) across all sets of groups M_i. The penalty $R(\bullet)$ then counts the number of sets that are larger or are equal in size to the smallest set. Therefore, this objective encourages Γ to use the strongest subspace, i.e., one that is invariant to more groups.

To use the learning framework, bases for the invariant subspaces for the power set of groups must first be computed. The procedure starts with the set of all groups and iterates through sets of decreasing size. While the procedure only needs to be run once for a particular initial set of groups, the runtime is technically exponential as the subspace needs to be computed for every set in the power set. This can be made the worst-case runtime by setting the procedure to terminate early once a subspace equal in size to the space of the input is found. Mouli and Ribeiro [MR21] comment that it is unclear if the worst-case runtime occurs in practice.

Once the bases have been computed, the neural network can be trained with the regularized objective $R(\bullet)$ using standard supervised learning algorithms. Note that $R(\bullet)$ is discrete, and so differentiable approximation is used during training where the indicator function $\mathbf{1}\{z > 0\}$ is approximated by

$$\widetilde{\mathbf{1}}\{z > 0\} = \frac{\tau z}{\tau z + 1}$$

with $\tau \geq 1$ being a temperature hyperparameter.

2.4 Analysis of technical work

We summarize our analysis of the technical work by Mouli and Ribeiro [MR21] from the previous sections and add a few additional comments.

The main theoretical innovations include the counterfactual formulation of the single-environment extrapolation problem and the results that rationalize the proposed learning framework under the setting. The SCM setup leads to a formalization of extrapolation that can reasoned with, and the group assumption on

the environment differences enables working with single-environment data. However, the group assumption is also a potential limitation of the framework where invariances for groups that are not initially specified cannot be learned.

The theoretical results all use a similar proof technique based on logical derivations that follow from applying properties and definitions. In consideration of Theorems 1 and 2, the problem setting should generalize to groups beyond finite linear automorphisms. However, Lemma 2 and Theorem 3 do not directly apply to non-linear groups and so developing analogous results for other groups would be necessary in order to justify using them with the framework. Extending the setting to non-groups or even just to non-normal subgroups for which Theorem 2 does not apply would likely require a different learning framework altogether as G-invariances are no longer CG-invariances in this case.

With regards to computation, generalizing the setting to non-finite groups (with a Haar measure) may pose a practical challenge where the Reynolds operator (or orbit-averaging operator in general) may not be computable exactly. The exponential-time procedure for computing invariant subspaces may perform poorly when a large number of groups is given even with the early termination condition. Mouli and Ribeiro [MR21] also comment that using the early termination condition with the regularized objective may lead to inoptimal learning of the invariances, and that encouraging sparsity through an additional entropy regularization term may help with the issue (which is left for future work).

3 Mini-proposals

3.1 Proposal 1: Learning Counterfactual G-invariances from Single Environments via Multiple Kernel Learning

The domain adaptation learning framework proposed by Mouli and Ribeiro [MR21] can be restrictive in that the specified groups are required to be finite. Furthermore, neural networks, while powerful for prediction, can also be challenging to work with if interpretability is desirable or if there is available domain-knowledge to incorporate. We propose an adaptation to their framework based on multiple kernel learning [GA11] that addresses these (and potentially other) restrictions. Such a framework would also have access to the additional benefits that kernels may have to offer, such as being able to use specially-designed kernels and a potentially infinite-dimensional feature space. We note that the details described in this proposal were considered as part of the conceptual planning for the project and may be subject to change.

Our proposed adaptation has the same goal as Mouli and Ribeiro's framework, which is to be able to extrapolate a model trained on single-environment data to different environments by learning invariances for a subset of given groups that describe non-causal information. The SCM setup and assumptions in our proposed adaptation is mostly identical to that of the original framework. However, unlike the original framework, we allow for continuous groups $\mathcal{G}_1, \ldots, \mathcal{G}_m$ at the expense of restricting the output space $\mathcal{Y} = \mathbb{R}$. We also assume that the groups are compact to ensure the existence of Haar measures $\lambda_1, \ldots, \lambda_m$. We retain the assumption that the non-causal subset of groups $\mathcal{G}_\mathcal{I}$ is a normal subgroup of the overgroup $\mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$ to take advantage of Theorem 2 in [MR21], which allows us to work directly with G-invariances rather than CG-invariances.

The main difference between the original framework and our proposed adaptation is the model itself. While Mouli and Ribeiro [MR21] propose to learn the invariances by encoding their respective invariant eigenspaces into neuron weights in a neural network layer, we propose to encode invariances as distinct invariant kernels in a prediction function that takes the form

$$f(x) = \alpha_0 + \sum_{j=1}^{p} \alpha_j \sum_{i=1}^{n} \beta_i \bar{k}_j(x, x_i)$$

where $p \approx 2^m$, $\bar{k}_j$ are kernels constructed to be invariant to different subsets of groups, β_i are learned weights on the training data x_i, and α_j are learned weights on the kernels. Learning the invariances then corresponds to learning the weights α of the kernels in the model. The optimization objective in the adaptation still uses a regularization term that encourages a greater weight on the strongest invariances that do not contradict the training data.

The main tasks of this proposed project would include the following:

1. Show how to construct a kernel that is invariant to a specific set of groups. Haasdonk, Vossen, and Burkhardt [HVB05] show that for a given kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a group $\mathcal{G}$, integrating the kernel inputs over the (normalized) Haar measure λ of the group produces the $\mathcal{G}$-invariant kernel

$$\bar{k}(x, x') = \int_\mathcal{G} \int_\mathcal{G} k(gx, g'x') \mathrm{d}\lambda(g) \mathrm{d}\lambda(g') \, .$$

   However, given a set of groups $\mathcal{G}_\mathcal{I}$ where $\mathcal{I} \subset \{1, \ldots, m\}$, it may not be obvious how to construct a function that is invariant to only transformations $g \in \mathcal{G}_\mathcal{I}$ and not $g' \in \mathcal{G}_{\supset\mathcal{I}}$ where $\mathcal{I}$ is a strict subset of $\supset \mathcal{I}$. Mouli and Ribeiro [MR21] deal with this problem in the context of finite linear groups by computing the intersection of the invariant eigenspaces for the groups in a set and removing their projection onto invariant eigenspaces for overgroups (Theorem 3). We expect that a kernel invariant to a specific set of groups may be constructed (or at least described) by drawing on summation properties

of reproducing kernel Hilbert spaces (RKHS) as well as RKHS decomposition results from [Ele21] (Lemmas 3 and 4 in particular). The goal would be to provide theoretical results analogous to Lemma 2 and Theorem 3 of [MR21].

2. If (1) is not possible or requires unreasonable assumptions, then identify the key obstacles that prevent the construction of such a kernel.

3. Formulate a kernel-based version of the framework by Mouli and Ribeiro. The main technical considerations would be computing the invariant kernels needed (which depends on the feasibility of (1)), adapting the regularization term in the optimization objective to work with kernel weights rather than neuron weights, and designing a feasible algorithm for learning the weights.

4. Determine the tractability of learning and using a model of the form given above. The value p in the above function is the number of subsets in the power set of the groups, and n is the number of training examples. This implies that just using the given model is of order $O(2^m n)$. The algorithm used by Mouli and Ribeiro [MR21] is also exponential, but early termination conditions are included and so it is unclear how rarely the worst-case runtime occurs in practice. It remains to be seen whether similar strategies can be applied for our proposed model (e.g., excluding kernels for low-order subsets).

If the tasks described above are completed successfully, the expected major contributions of this project would be as follows:

1. Further development of invariant kernel theory. Formal theory regarding the construction of a kernel that is invariant to exactly a specific set of groups does not seem to have been explored in existing literature and is an interesting idea in its own right.

2. A CG-invariance learning framework for single-environment domain adaptation based on kernels that works with (possibly non-linear) continuous groups. In addition to being a potentially useful method, understanding whether other methods aside from the one proposed by Mouli and Ribeiro are feasible is critical for drawing attention to the relatively new single-environment domain adaptation literature.

3. Empirical results that evaluate how the adapted framework compares to standard domain adaptation methods and the one proposed by Mouli and Ribeiro.

4 Project report

Testing for group invariance using kernel hypothesis tests

Abstract. TODO

4.1 Introduction

(TODO: make this section logically flow better)

Various recent works in the domain adaptation literature have framed learning to generalize across domains as learning conditional invariances to the transformation of a group (e.g., [MR21; Sch+21] TODO). Often, these works assume a group (or set of groups) of relevance a priori and encourage learning of these invariances through data using, for example, data augmentation and inductive biases in the model architecture (TODO). However, rather than trying to learn invariances from the data, it may be of interest to determine whether assuming the underlying generating process of the data being invariant would contradict what is observed in the data. This objective is particularly useful in the case where only data from a single domain are available as in the problem setting described by Mouli and Ribeiro [MR21].

In this project, we examine the problem of detecting potential invariances from a kernel hypothesis testing perspective. The objective and intuition behind the standard hypothesis testing framework appears to well-match the objective of our problem as it answers the question, *if the underlying generating distribution were invariant, is there evidence in the observed data that contradicts that assumption?* For practically conducting a test of invariance, we propose using kernel-based methods for their flexibility and ability to work with high-dimensional data. The contributions of this project are as follows:

1.

This project report is organized as follows: Section 4.2 highlights related work in the literature and how our work differs; Section 4.3 provides additional background for our proposed tests and introduces notation; Section 4.4 and Section 4.5 present our proposed method in the context of two different settings; and Section 4.6 concludes the report with a discussion and reflection of this project.

4.2 Related work

Learning invariance to unknown groups from single-domain data. This project is inspired by the work of Mouli and Ribeiro [MR21] who proposed a method for learning counterfactual group-invariances in neural networks given only data from a single domain. Their method is based on having a specified set of potential groups to be invariant to, and using a regularized optimization objective to "unlearn" the groups that contradict the data. In this work, we frame the context of determining whether being invariant to a specified group contradicts the given dataset as a hypothesis testing problem.

Domain adaptation based on learning invariances. The work by Mouli and Ribeiro [MR21] falls more generally under the domain adaptation literature. This literature is broad and consist of works involving varying problem settings and approaches. Recent work that reduces domain adaptation to learning invariances include (but are not limited nor mutually exclusive to) those based on invariant neural networks [Li+18; Zha+19; Sch+21], invariant kernels [Li+18; MGZ19; EZ21; Ele21], and causal reasoning [Mag+17; CB20]. The majority of these approaches involve learning invariances from available training data that come from multiple domains. Our work also shares the perspective of reducing domain adaptation to invariances. However, rather than learning invariances from data, our hypothesis testing approach determines whether a specified invariance is compatible with the data (which may come from only a single domain).

Kernel hypothesis testing. Our setup of testing for invariances builds on the existing literature of kernel hypothesis testing, which include kernel methods for common one-sample problems [Zha+11; Dor+14; KC15; CSG16; JKS20] and two-sample problems [Gre+07; Gre+12]. While the proposed approaches differ depending on the context and goal of their respective problems, most approaches are based on mapping empirical distributions (and non-empirical distributions in the case of some one-sample problems) to a kernel mean embedding in some reproducing kernel Hilbert space and comparing embeddings through the *maximum mean discrepancy* (MMD) test statistic [Har+13]. The MMD test statistic is convenient for kernel-based hypothesis tests as its null distribution can be estimated using standard methods such as bootstrap [Gre+12]. Our setup of the detecting invariance problem makes use of some of these existing kernel-based hypothesis tests.

Hypothesis testing for invariances. Our hypothesis testing setup is used to detect possible invariances in the data to actions of a mathematical group. From this perspective, our work is closely related to the literature on testing for symmetries in data [HKM03; Nga09; PP15]. However, while the literature addresses the problem of determining whether there is some aspect of invariance in the data, our work addresses the different problem of determining whether assuming the underlying distribution being invariant would contradict the data.

TODOadd bit about invariant hypothesis tests?

4.3 Background and notation

We first provide some additional background that will be useful for understanding our proposed tests as well as the mathematical notation that we will use in this report. In both of our formulations, we use $\mathcal{X}$ and $\mathcal{Y}$ to represent the (potentially high-dimensional) space of inputs and outputs, respectively. For a random variable X with support $\mathcal{X}$, we overload the notation $\mathbb{P}_X$ to represent both its distribution and the corresponding probability measure.

Transformation groups. Let $\mathcal{G}$ be a transformation group that measurably acts on the input space $\mathcal{X}$. For $g \in \mathcal{G}$, we denote the action of g on $x \in \mathcal{X}$ given by the measurable map $\Phi : \mathcal{G} \times \mathcal{X} \to \mathcal{X}$ as $g \cdot x = \Phi(g, x)$. For a set $A \subset \mathcal{X}$, we write $g \cdot A = \{g \cdot x : x \in A\}$. To refer to the distribution of $g^{-1} \cdot X$ for some random variable X on $\mathcal{X}$, we use the image measure denoted $\mathbb{P}_{gX} = \mathbb{P}_X \circ g^{-1}$, i.e., $\mathbb{P}_{gX}(dx) = \mathbb{P}_X(g^{-1} \cdot dx)$. We assume that $\mathcal{G}$ is compact and therefore has an unique (normalized) Haar measure λ (TODOcite?).

Kernels and RKHS. Let $\mathcal{H}_\mathcal{X}$ be a reproducing kernel Hilbert space (RKHS) of functions $f : \mathcal{X} \to \mathbb{R}$ on which the evaluation functional $\delta_x : \mathcal{H}_\mathcal{X} \to \mathbb{R}$ is continuous for all $x \in \mathcal{X}$, i.e., $\delta_x[f] = f(x)$ is bounded. Then by the Riesz representation theorem, there is a unique function $k_x \in \mathcal{H}_\mathcal{X}$ with the reproducing property, i.e., $f(x) = \langle f, k_x \rangle_{\mathcal{H}_\mathcal{X}}$ for all $f \in \mathcal{H}_\mathcal{X}$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{X}}$ is the inner product on $\mathcal{H}_\mathcal{X}$. The function $k_x = k(x, \cdot)$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite and symmetric with $k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}_\mathcal{X}}$, is the reproducing kernel of $\mathcal{H}_\mathcal{X}$. For $x \in \mathcal{X}$, the function $k_x = \varphi(x)$ can be viewed as a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ where d may potentially be infinite. Therefore, evaluations of the reproducing kernel $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ may be viewed as inner products on the mapped feature space. For joint spaces $\mathcal{X} \times \mathcal{Y}$, we consider the product kernel $k_{\mathcal{X}\mathcal{Y}}((x, y), (x', y')) = k_\mathcal{X}(x, x')k_\mathcal{Y}(y, y')$ that has feature map $\varphi_\mathcal{X} \otimes \varphi_\mathcal{Y}$ where $\otimes$ denotes the tensor product. Note that we will drop the context space subscript in our notation when the context space is obvious, e.g., write $\mathcal{H}$ instead of $\mathcal{H}_\mathcal{X}$, k instead of $k_{\mathcal{X}\times\mathcal{Y}}$, φ instead of $\varphi_\mathcal{Y}$, etc. (TODOcite?)

Kernel embeddings. For a distribution $\mathbb{P}_X \in \mathcal{H}$, the *kernel mean embedding* of $\mathbb{P}_X$ is defined as $\mu_X = \mathbb{E}[\varphi(X)] = \int_\mathcal{X} \varphi(x)\mathbb{P}_X(dx)$. The kernel embedding μ_X is the unique element of $\mathcal{H}$ such that $\mathbb{E}_X[f(X)] = \langle f, \mu_X \rangle_\mathcal{H}$ for all $f \in \mathcal{H}$. Let $\mathbb{P}_\mathcal{X}$ denote the set of distributions on $\mathcal{X}$. If the reproducing kernel k is characteristic, the kernel embedding $\mu_X : \mathbb{P}_\mathcal{X} \to \mathcal{H}$ is injective, i.e., each distribution corresponds to an unique point in the RKHS. In this work, we assume that the kernels on the spaces $\mathcal{X}$ and $\mathcal{Y}$ are characteristic and translation-invariant, i.e., $k(x, x') = \Psi(x - x')$ for some bounded continuous positive definite function $\Psi : \mathbb{R}^d \to \mathbb{R}$. Then under mild assumptions, the product kernel $k_{\mathcal{X}\mathcal{Y}} = k_\mathcal{X}k_\mathcal{Y}$ is also characteristic

[Sri+10, Corollary 11] and so the joint distribution kernel embedding $\mu_{XY} : \mathbb{P}_{\mathcal{XY}} \to \mathcal{H}$ is injective. (TODO)

Kernel hypothesis tests and MMD. Let $\mathcal{D}^{(1)} = \{x_i^{(1)}\}_{i=1}^{n_1}$ and $\mathcal{D}^{(2)} = \{x_i^{(2)}\}_{i=1}^{n_2}$ be two datasets. Two-sample kernel hypothesis tests are generally set up to determine whether the distributions $\mathbb{P}_X^{(1)}$ and $\mathbb{P}_X^{(2)}$ that generated $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, respectively, are the same. These tests compare $\mathbb{P}_X^{(1)}$ and $\mathbb{P}_X^{(2)}$ by comparing their kernel embeddings through the *maximum mean discrepancy* (MMD) test statistic [Gre+12], which is given by

$$\mathrm{MMD}(\mathbb{P}_X^{(1)}, \mathbb{P}_X^{(2)}) = \left\| \mu_X^{(1)} - \mu_X^{(2)} \right\|_{\mathcal{H}} .$$

The MMD test statistic is particularly convenient in the two-sample case as its squared form can be empirically estimated through only kernel evaluations using the (biased) estimator

$$\begin{aligned} \widehat{\mathrm{MMD}}^2(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}) &= \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi(x_i^{(1)}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \varphi(x_i^{(2)}) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n_1^2} \mathbf{1}_{n_1}^T \mathbf{K}^{(1)} \mathbf{1}_{n_1} + \frac{1}{n_2^2} \mathbf{1}_{n_2}^T \mathbf{K}^{(2)} \mathbf{1}_{n_2} - \frac{2}{n_1 n_2} \mathbf{1}_{n_1}^T \mathbf{K}^{(12)} \mathbf{1}_{n_2} \end{aligned}$$

where $(\mathbf{K}^{(1)})_{ij} = k(x_i^{(1)}, x_j^{(1)})$, $(\mathbf{K}^{(2)})_{ij} = k(x_i^{(2)}, x_j^{(2)})$ and $(\mathbf{K}^{(12)})_{ij} = k(x_i^{(1)}, x_j^{(2)})$. Under the null where $\mathbb{P}_X^{(1)} = \mathbb{P}_X^{(2)}$, the distribution of $\widehat{\mathrm{MMD}}$ can be estimated several ways. For example, one method is based on using a Gamma approximation where the parameters of the distribution are defined in terms of moments of the biased estimator [Gre+09]. Estimating the moments has a computational complexity of $O(n_1 n_2)$, which is generally lower than that of other methods based on Pearson curves or bootstrapping (which has cubic cost) [Gre+09]. Depending on how the distribution of $\widehat{\mathrm{MMD}}$ is estimated, the rejection region can be computed either parametrically or through bootstrap. The kernel hypothesis testing procedure is summarized in Algorithm 1.

Algorithm 1: Kernel hypothesis test

Input : datasets $\mathcal{D}^{(1)}$, $\mathcal{D}^{(2)}$; significance level α
Output: reject (1) or no reject (0)

1 compute approximation $\hat{\mathbb{P}}_0$ of null distribution $\mathbb{P}_0$;
2 compute critical value q^* such that $\hat{\mathbb{P}}_0(\{q \in \mathbb{R} : q \geq q^*\}) = 1 - \alpha$;
3 compute estimate $\widehat{\mathrm{MMD}}(\mathcal{D}^{(1)}, \mathcal{D}^{(2)})$;
4 **if** $\widehat{\mathrm{MMD}}(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}) > q^*$ **then**
5 | return 1;
6 **end**
7 return 0;

In this project, we present two formulations of testing for invariance and discuss how to compute $\widehat{\mathrm{MMD}}$ (line 4 of Algorithm 1) in each formulation. Once the test statistic has been estimated, the testing procedure outlined in Algorithm 1 can then be applied.

4.4 Testing for invariances via conditional independence tests

- Context: suppose we have a dataset $\mathcal{D} = \{(x_i, g_i, y_i)\}_{i=1}^n$ collected from a single environment where $x_i \in \mathcal{X}$ are inputs (in some canonical form), $y_i \in \mathcal{Y}$ are outputs, and $g_i \in \mathcal{G}$ for some known group $\mathcal{G}$ acting on $\mathcal{X}$ are observable transformations on the inputs (e.g., image orientations, image backgrounds, etc.). Assume that the joint distribution factorizes as $\mathbb{P}_{XYG} = \mathbb{P}_G \mathbb{P}_X \mathbb{P}_{Y|X,G}$. Assume that $\mathbb{P}_X$ and $\mathbb{P}_{Y|X,G}$ are fixed across environments but $\mathbb{P}_G$ may vary.

- Objective: our goal is to determine whether the outputs Y_i being conditionally invariant to the transformations G_i, i.e.,
$$\mathbb{P}_{XYG} = \mathbb{P}_G \mathbb{P}_X \mathbb{P}_{Y|X,G} = \mathbb{P}_G \mathbb{P}_X \mathbb{P}_{Y|X} \ ,$$
would contradict what is observed in the given dataset.

- Proposed approach: we set up this problem as a kernel conditional independence test where the null hypothesis says the outputs are conditionally invariant and the alternative hypothesis says otherwise.

  We follow the permutation-based approach described in [Dor+14]. The idea is that under the null, we should be able to permute the transformations g_i in the dataset without disturbing the joint distribution. We can then transform the conditional independence test into a two-sample test by splitting the dataset in half, permuting the transformations in one half, and then comparing the empirical distributions of the two halves. It is expected that unlike in [Dor+14], our setup does not require learning the permutation matrix (although we may want to maximize the number of samples that have different transformations after permutation).

  For comparing the distributions, we can use the well-studied MMD test statistic for two-sample kernel tests [Gre+12]. Let k_x, k_y and k_g be characteristic, positive-definite kernel functions defined on their respective spaces, e.g., $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and where k_x has a corresponding feature map $\phi_x : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}$ such that $k_x(x, x') = \langle \phi_x(x), \phi_x(x') \rangle$ with $\mathcal{H}_{\mathcal{X}}$ being the RKHS of k_x. Define the product kernel $k_{xyg}((x, y, g), (x', y', g')) = k_x(x, x') k_y(y, y') k_g(g, g')$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$ with feature map $\phi_x \otimes \phi_y \otimes \phi_g$ ($\otimes$ being the tensor product). Because the kernels are characteristic, the kernel mean embedding μ (estimated by $\widehat{\mu}$) maps distributions to unique functions in the RKHS of their respective spaces. Let $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ be the empirical distributions of the split dataset and let $\mathbf{P}$ be the permutation matrix. The MMD is then computed as
$$\mathrm{MMD}(\mathbb{P}^{(1)}, \mathbf{P}\mathbb{P}^{(2)}) = \left\| \widehat{\mu}(\mathbb{P}^{(1)}) - \widehat{\mu}(\mathbf{P}\mathbb{P}^{(2)}) \right\|_{\mathcal{H}}^2 \ .$$
  For small datasets, the null distribution of the MMD test statistic can be estimated via bootstrap where the original dataset is randomly split in half multiple times. If the null is not rejected, then the conditional invariance assumption does not significantly contradict the given data.

- Limitations: the transformations g_i must be observable and canonical forms of the inputs must be computable.

- Other considerations: the setup does not seem to restrict the input and output spaces nor does it make assumptions on the transformations. The set of possible transformations may not necessarily need to be a group (reduces problem to standard conditional independence testing of three variables)?

- Multiple environments: if we had data from two environments $\mathcal{D}^{(j)} = \{(x_i^{(j)}, g_i^{(j)}, y_i^{(j)})\}_{i=1}^n$ for $j \in \{1, 2\}$, a two-sample permutation test is possible by just merging the datasets and randomly splitting in two.

4.5 Testing for invariances via two-sample tests

Given datasets from two environment, we can give an alternative formulation:

- Context: suppose we have datasets $\mathcal{D}^{(j)} = \{(x_i^{(j)}, y_i^{(j)})\}_{i=1}^{n_j}$ collected from two different environments $j \in \{1, 2\}$ where $x_i \in \mathcal{X}$ are inputs and $y_i \in \mathcal{Y}$ are outputs. Suppose that different environments can be represented by different transformations g from some unknown group $\mathcal{G}^*$ acting on $\mathcal{X}$. Assume that the joint distribution factorizes as $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_{Y|X}$ where $\mathbb{P}_{Y|X}$ is $\mathcal{G}^*$-invariant, i.e.,
$$\mathbb{P}_{Y|X}(dy|x) = \mathbb{P}_{Y|X}(dy|g \cdot x)$$

for all $g \in \mathcal{G}^*$. Assume that $\mathbb{P}_{Y|X}$ is fixed across environments, and $\mathbb{P}_X$ is fixed but equivariant across environments in the sense that $\mathbb{P}_X^{(1)}(dx) = \mathbb{P}_X^{(2)}(g^{-1} \cdot dx)$ where $\mathbb{P}_X^{(1)}$ and $\mathbb{P}_X^{(2)}$ correspond to the marginals of different environments. For convenience of notation, we write $\mathbb{P}_{gX}(dx) = \mathbb{P}_X \circ g^{-1}(dx) = \mathbb{P}_X(g^{-1} \cdot dx)$.

- Notation: image of $\mathbb{P}_X$ under g: $\mathbb{P}_{gX} = \mathbb{P}_X \circ g^{-1}$.

$$\int \mathbb{P}_{gX}(dx) f(x) = \int \mathbb{P}_X(dx) f(gx)$$

- Objective: our goal is to determine whether a specified $\mathcal{G}$ with a Haar measure λ is the "correct" group, i.e., $\mathcal{G} = \mathcal{G}^*$.
- Proposed approach: we set up this problem as a kernel hypothesis test for comparing distributions where the null hypothesis says that $\mathcal{G} = \mathcal{G}^*$ and the alternative hypothesis says otherwise.

Let $\mathbb{P}_X^{(1)}$ and $\mathbb{P}_X^{(2)}$ denote the marginal distributions for the two environments where $\mathbb{P}_{g^*X}^{(1)} = \mathbb{P}_X^{(2)}$ for some $g^* \in \mathcal{G}^*$. By assumptions from the setup and from working under the null, the orbit-averaged joint distributions should be equivalent, i.e.,

$$\begin{aligned}
\int_{\mathcal{G}} \mathbb{P}_{(gX)Y}^{(1)} \lambda(dg) &= \mathbb{P}_{Y|X} \int_{\mathcal{G}} \mathbb{P}_{gX}^{(1)} \lambda(dg) \\
&= \mathbb{P}_{Y|X} \int_{\mathcal{G}} \mathbb{P}_{gg^*X}^{(1)} \lambda(dg) \\
&= \mathbb{P}_{Y|X} \int_{\mathcal{G}} \mathbb{P}_{gX}^{(2)} \lambda(dg) = \int_{\mathcal{G}} \mathbb{P}_{(gX)Y}^{(2)} \lambda(dg) \, .
\end{aligned}$$

Let $\mathcal{X}_{\mathcal{G}}$ denote the space of $\mathcal{G}$-orbits of $\mathcal{X}$, which is a measurable space assuming that $\mathcal{G}$ and $\mathcal{X}$ are measurable (TODO). Define $\mathbb{P}_{X_{\mathcal{G}}Y} = \int_{\mathcal{G}} \mathbb{P}_{(gX)Y} \lambda(dg)$ to be the probability measure on $\mathcal{X}_{\mathcal{G}} \times \mathcal{Y}$ obtained by orbit-averaging the joint measure $\mathbb{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$.

Let k be the product kernel $k(x \otimes y, x' \otimes y') = k_x(x, x') k_y(y, y')$ where k_x, k_y are bounded characteristic kernels on their respective spaces. The kernel embedding of the orbit-averaged measure corresponds to the kernel embedding of $\mathbb{P}_{XY}$ using a Haar-integration kernel [HVB05] invariant in one tensor dimension given by

$$\bar{k}(x \otimes y, x' \otimes y') = \int_{\mathcal{G}} \int_{\mathcal{G}} k(g \cdot x \otimes y, g' \cdot x' \otimes y') \lambda(dg) \lambda(dg') \, .$$

Define $k_{\mathcal{G}}$ be the kernel in the RKHS of $\mathcal{X}_{\mathcal{G}} \times \mathcal{Y}$ corresponding to $\bar{k}$, i.e.,

$$k_{\mathcal{G}}(x_{\mathcal{G}} \otimes y, \cdot) = \bar{k}(x \otimes y, \cdot) \, .$$

The equivalence of the embeddings then follows as

$$\begin{aligned}
\mu_{X_{\mathcal{G}}Y} &= \int_{\mathcal{X}_{\mathcal{G}} \times \mathcal{Y}} k_{\mathcal{G}}(x_{\mathcal{G}} \otimes y, \cdot) \mathbb{P}_{X_{\mathcal{G}}Y}(dx_{\mathcal{G}}, dy) \\
&= \int_{\mathcal{X} \times \mathcal{Y}} \bar{k}(x \otimes y, \cdot) \int_{\mathcal{G}} \mathbb{P}_{XY}(g^{-1} \cdot dx, dy) \lambda(dg) \\
&= \int_{\mathcal{G}} \int_{\mathcal{X} \times \mathcal{Y}} \bar{k}(g \cdot x \otimes y, \cdot) \mathbb{P}_{XY}(dx, dy) \lambda(dg) \\
&= \int_{\mathcal{X} \times \mathcal{Y}} \bar{k}(x \otimes y, \cdot) \mathbb{P}_{XY}(dx, dy) \\
&= \int_{\mathcal{G}} \int_{\mathcal{X} \times \mathcal{Y}} k(g \cdot x \otimes y, \cdot) \mathbb{P}_{XY}(dx, dy) \lambda(dg) \\
&= \int_{\mathcal{G}} \mu_{(gX)Y} \lambda(dg) \, .
\end{aligned}$$

The MMD test statistic is based on comparing the orbit-averaged kernel mean embeddings. The MMD test statistic is

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}^{(1)}_{X_{\mathcal{G}}Y}, \mathbb{P}^{(2)}_{X_{\mathcal{G}}Y}) &= \left\| \int_{\mathcal{G}} \mu^{(1)}_{(gX)Y} \lambda(dg) - \int_{\mathcal{G}} \mu^{(2)}_{(gX)Y} \lambda(dg) \right\|^2_{\mathcal{H}} \\ &= \left\| \int_{\mathcal{G}} \mu^{(1)}_{(gX)Y} - \mu^{(2)}_{(gX)Y} \lambda(dg) \right\|^2_{\mathcal{H}} \\ &= \left\| \int_{\mathcal{G}} \mathbb{E}_{\mathbb{P}^{(1)}_{gXY}} \left[\varphi(X^{(1)}, Y^{(1)}) \right] - \mathbb{E}_{\mathbb{P}^{(2)}_{gXY}} \left[\varphi(X^{(2)}, Y^{(2)}) \right] \lambda(dg) \right\|^2_{\mathcal{H}} .\end{aligned}$$

The MMD test statistic is empirically estimated by

$$\begin{aligned}\widehat{\mathrm{MMD}}(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}) &= \left\| \int_{\mathcal{G}} \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi(g \cdot x_i^{(1)}, y_i^{(1)}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \varphi(g \cdot x_i^{(2)}, y_i^{(2)}) \lambda(dg) \right\|^2_{\mathcal{H}} \\ &= \int_{\mathcal{G}} \int_{\mathcal{G}} \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(g \cdot x_i^{(1)} \otimes y_i^{(1)}, g \cdot x_j^{(1)} \otimes y_j^{(1)}) \\ &\quad + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} k(g' \cdot x_i^{(2)} \otimes y_i^{(2)}, g' \cdot x_j^{(2)} \otimes y_j^{(2)}) \\ &\quad - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(g \cdot x_i^{(1)} \otimes y_i^{(1)}, g' \cdot x_j^{(2)} \otimes y_j^{(2)}) \lambda(dg) \lambda(dg') \\ &= \int_{\mathcal{G}} \int_{\mathcal{G}} \frac{1}{{n_1}^2} \mathbf{1}_{n_1}^T \mathbf{K}^{(1)} \mathbf{1}_{n_1} + \frac{1}{n_2^2} \mathbf{1}_{n_2}^T \mathbf{K}^{(2)} \mathbf{1}_{n_2} - \frac{2}{n_1 n_2} \mathbf{1}_{n_1}^T \mathbf{K}^{(12)} \mathbf{1}_{n_2} \lambda(dg) \lambda(dg')\end{aligned}$$

where

$$\begin{aligned}\mathbf{K}^{(1)} &= \mathbf{K}^{(1)}_{gx} \odot \mathbf{K}^{(1)}_{y} , \\ \mathbf{K}^{(2)} &= \mathbf{K}^{(2)}_{g'x} \odot \mathbf{K}^{(2)}_{y} , \\ \mathbf{K}^{(12)} &= \mathbf{K}^{(12)}_{gx,g'x} \odot \mathbf{K}^{(12)}_{y,y}\end{aligned}$$

with $\odot$ denoting the Hadamard product, $(\mathbf{K}^{(\ell)}_x)_{ij} = k_x(x_i^{(\ell)}, x_j^{(\ell)})$ (similarly for $\mathbf{K}^{(\ell)}_y$), and $(\mathbf{K}^{(12)}_{x,x})_{ij} = k_x(x_i^{(1)}, x_j^{(2)})$ (similarly for $\mathbf{K}^{(12)}_{y,y}$). If $\mathcal{G}$ is finite and small enough, $\widehat{\mathrm{MMD}}$ may be computed exactly by summing over $\mathcal{G}$ and weighting by the discrete Haar measure. If $\mathcal{G}$ is too large or not finite, then $\widehat{\mathrm{MMD}}$ must be estimated by sampling $g, g' \in \mathcal{G}$, i.e.,

$$\begin{aligned}\widehat{\widehat{\mathrm{MMD}}}(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \{g_i, g_i'\}_{i=1}^m) &= \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\frac{1}{{n_1}^2} \mathbf{1}_{n_1}^T \mathbf{K}_i^{(1)} \mathbf{1}_{n_1} + \frac{1}{n_2^2} \mathbf{1}_{n_2}^T \mathbf{K}_j^{(2)} \mathbf{1}_{n_2} - \frac{2}{n_1 n_2} \mathbf{1}_{n_1}^T \mathbf{K}_{ij}^{(12)} \mathbf{1}_{n_2} \right) \\ &= \frac{1}{m n_1^2} \sum_{i=1}^{m} \mathbf{1}_{n_1}^T \mathbf{K}_i^{(1)} \mathbf{1}_{n_1} + \frac{1}{m n_2^2} \sum_{j=1}^{m} \mathbf{1}_{n_2}^T \mathbf{K}_j^{(2)} \mathbf{1}_{n_2} - \frac{2}{m^2 n_1 n_2} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{1}_{n_1}^T \mathbf{K}_{ij}^{(12)} \mathbf{1}_{n_2}\end{aligned}$$

for $m \geq 1$ and where

$$\begin{aligned}\mathbf{K}_i^{(1)} &= \mathbf{K}^{(1)}_{g_i x} \odot \mathbf{K}^{(1)}_{y} , \\ \mathbf{K}_j^{(2)} &= \mathbf{K}^{(2)}_{g_j' x} \odot \mathbf{K}^{(2)}_{y} , \\ \mathbf{K}_{ij}^{(12)} &= \mathbf{K}^{(12)}_{g_i x, g_j' x} \odot \mathbf{K}^{(12)}_{y,y} .\end{aligned}$$

As $\widehat{\mathrm{MMD}}$ can be viewed as comparing distributions on the space $\mathcal{X}_{\mathcal{G}} \times \mathcal{Y}$, the null distribution of $\widehat{\mathrm{MMD}}$ can be approximated using standard methods (e.g., Gamma approximation, bootstrap, etc.) [Gre+09]. The null distribution of $\widehat{\widehat{\mathrm{MMD}}}$ is the same but is now a second-layer approximation to the original distribution.

Note that if the kernel k_x satisfies $k_x(g \cdot x, x') = k_x(x, g^{-1} \cdot x')$, then $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$ are independent of g and g' respectively and so $\widehat{\mathrm{MMD}}$ simplifies to

$$\widehat{\mathrm{MMD}}(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}) = \frac{1}{{n_1}^2} \mathbf{1}_{n_1}^T \mathbf{K}^{(1)} \mathbf{1}_{n_1} + \frac{1}{n_2^2} \mathbf{1}_{n_2}^T \mathbf{K}^{(2)} \mathbf{1}_{n_2} - \frac{2}{n_1 n_2} \mathbf{1}_{n_1}^T \left(\int_{\mathcal{G}} \int_{\mathcal{G}} \mathbf{K}^{(12)} \lambda(dg) \lambda(dg') \right) \mathbf{1}_{n_2} .$$

Correspondingly, $\widehat{\widehat{\mathrm{MMD}}}$ simplifies to

$$\widehat{\widehat{\mathrm{MMD}}}(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \{g_i, g_i'\}_{i=1}^m) = \frac{1}{n_1^2} \mathbf{1}_{n_1}^T \mathbf{K}^{(1)} \mathbf{1}_{n_1} + \frac{1}{n_2^2} \mathbf{1}_{n_2}^T \mathbf{K}^{(2)} \mathbf{1}_{n_2} - \frac{2}{m^2 n_1 n_2} \sum_{i=1}^m \sum_{j=1}^m \mathbf{1}_{n_1}^T \mathbf{K}_{ij}^{(12)} \mathbf{1}_{n_2} .$$

- Conceptual idea: the joint distribution $\mathbb{P}_{XY}$ can be reduced to $\mathbb{P}_{X_{\mathcal{G}}Y}$. The conditional distribution of Y is invariant on each orbit.
- An approach based on one dataset and comparing the kernel mean embedding to its orbit-averaged mean embedding would not work because the marginal distributions would be different.

4.6 Discussion and reflection

TODOchallenges: conditional embeddings and joints

References

[Arj+20] M. Arjovsky et al. *Invariant Risk Minimization*. 2020. arXiv: 1907.02893 [stat.ML].

[CDL20] S. Chen, E. Dobriban, and J. Lee. "A Group-Theoretic Framework for Data Augmentation". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 21321–21333. URL: https://proceedings.neurips.cc/paper/2020/file/f4573fc71c731d5c362f0d7860945b88-Paper.pdf.

[CB20] Y. Chen and P. Bühlmann. "Domain Adaptation Under Structural Causal Models". In: *arXiv preprint arXiv:2010.15764* (2020).

[CSG16] K. Chwialkowski, H. Strathmann, and A. Gretton. "A Kernel Test of Goodness of Fit". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 2606–2615.

[Dor+14] G. Doran et al. "A Permutation-Based Kernel Conditional Independence Test". In: UAI'14. Quebec City, Quebec, Canada: AUAI Press, 2014, pp. 132–141. ISBN: 9780974903910.

[Ele21] B. Elesedy. *Provably Strict Generalisation Benefit for Invariance in Kernel Methods*. 2021. arXiv: 2106.02346 [stat.ML].

[EZ21] B. Elesedy and S. Zaidi. *Provably Strict Generalisation Benefit for Equivariant Models*. 2021. arXiv: 2102.10333 [stat.ML].

[Far+20] A. Farahani et al. *A Brief Review of Domain Adaptation*. 2020. arXiv: 2010.03978 [cs.LG].

[GA11] M. Gönen and E. Alpaydin. "Multiple Kernel Learning Algorithms". In: *Journal of Machine Learning Research* 12.64 (2011), pp. 2211–2268. URL: http://jmlr.org/papers/v12/gonen11a.html.

[Gre+07] A. Gretton et al. "A Kernel Approach to Comparing Distributions". In: *AAAI*. 2007.

[Gre+09] A. Gretton et al. "A Fast, Consistent Kernel Two-Sample Test". In: *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio et al. Vol. 22. Curran Associates, Inc., 2009. URL: https://proceedings.neurips.cc/paper/2009/file/9246444d94f081e3549803b928260f56-Paper.pdf.

[Gre+12] A. Gretton et al. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773. URL: http://jmlr.org/papers/v13/gretton12a.html.

[HVB05] B. Haasdonk, A. Vossen, and H. Burkhardt. "Invariance in Kernel Methods by Haar-Integration Kernels". In: *SCIA*. 2005.

[Har+13] Z. Harchaoui et al. "Kernel-Based Methods for Hypothesis Testing: A Unified View". In: *IEEE Signal Processing Magazine* 30.4 (2013), pp. 87–97. DOI: 10.1109/MSP.2013.2253631.

[HKM03] N. Henze, B. Klar, and S. Meintanis. "Invariant Tests for Symmetry About an Unspecified Point Based on the Empirical Characteristic Function". In: *Journal of Multivariate Analysis* 87.2 (2003), pp. 275–297. ISSN: 0047-259X. DOI: https://doi.org/10.1016/S0047-259X(03)00044-7. URL: https://www.sciencedirect.com/science/article/pii/S0047259X03000447.

[JKS20] W. Jitkrittum, H. Kanagawa, and B. Schölkopf. "Testing Goodness of Fit of Conditional Density Models with Kernels". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 221–230.

[KC15] J. Kellner and A. Celisse. *A One-Sample Test for Normality with Kernel Methods*. 2015. arXiv: 1507.02904 [math.ST].

[Kum+20] S. Kumar et al. *One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL*. 2020. arXiv: 2010.14484 [cs.LG].

[Li+18] Y. Li et al. "Deep Domain Generalization via Conditional Invariant Adversarial Networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[MGZ19] Y. Ma, V. Ganapathiraman, and X. Zhang. "Learning Invariant Representations with Kernel Warping". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 1003–1012. URL: https://proceedings.mlr.press/v89/ma19a.html.

[Mag+17] S. Magliacane et al. "Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions". In: *arXiv preprint arXiv:1707.06422* (2017).

[MR21] S. C. Mouli and B. Ribeiro. *Neural Networks for Learning Counterfactual G-Invariances from Single Environments*. 2021. arXiv: 2104.10105 [cs.LG].

[MFK94] D. Mumford, J. Fogarty, and F. C. Kirwan. *Geometric Invariant Theory*. English. 3rd enl. Vol. 34;3. Folge, Bd. 34.; New York;Berlin; Springer-Verlag, 1994.

[Nga09] J. Ngatchou-Wandji. "Testing for Symmetry in Multivariate Distributions". In: *Statistical Methodology* 6.3 (2009), pp. 230–250. ISSN: 1572-3127. DOI: https://doi.org/10.1016/j.stamet.2008.09.003. URL: https://www.sciencedirect.com/science/article/pii/S1572312708000634.

[Par+18] G. Parascandolo et al. *Learning Independent Causal Mechanisms*. 2018. arXiv: 1712.00961 [cs.LG].

[PP15] C. Partlett and P. Patil. "Measuring Asymmetry and Testing Symmetry". In: *Annals of the Institute of Statistical Mathematics* 69 (Dec. 2015). DOI: 10.1007/s10463-015-0547-4.

[Sch19] B. Schölkopf. *Causality for Machine Learning*. 2019. arXiv: 1911.10500 [cs.LG].

[Sch+21] P. E. Schwöbel et al. "Last Layer Marginal Likelihood for Invariance Learning". In: *arXiv preprint arXiv:2106.07512* (2021).

[Sri+10] B. K. Sriperumbudur et al. "Hilbert Space Embeddings and Metrics on Probability Measures". In: *J. Mach. Learn. Res.* 11 (Aug. 2010), pp. 1517–1561. ISSN: 1532-4435.

[Zha+11] K. Zhang et al. "Kernel-Based Conditional Independence Test and Application in Causal Discovery". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. UAI'11. Barcelona, Spain: AUAI Press, 2011, pp. 804–813. ISBN: 9780974903972.

[Zha+19] H. Zhao et al. *On Learning Invariant Representation for Domain Adaptation*. 2019. arXiv: 1901.09453 [cs.LG].

[Zha+20] S. Zhao et al. *A Review of Single-Source Deep Unsupervised Visual Domain Adaptation*. 2020. arXiv: 2009.00155 [cs.CV].