TODO

Kenny Chiu

October 3, 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, existing causal-based methods such as IRM and *Independent Causal Mechanisms* [Par+18] (ICM) 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 literature where 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 known. 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 simplfy 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 borrows from the ICM literature [Par+18] where the SCM in Figure 1 is assumed.

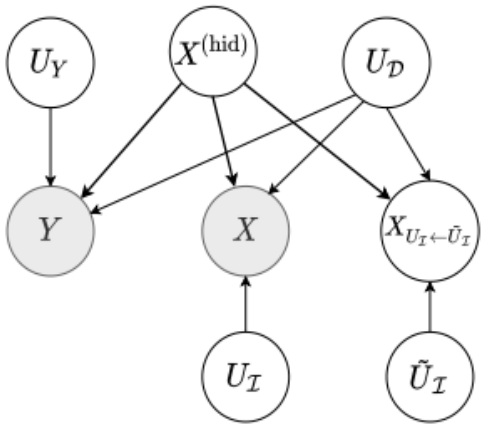

Figure 1: Assumed SCM in single-environment extrapolation. Grey nodes are observed variables. Arrows denote deterministic causal relationships where the target is dependent on the source. Figure from [MR21].

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 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.

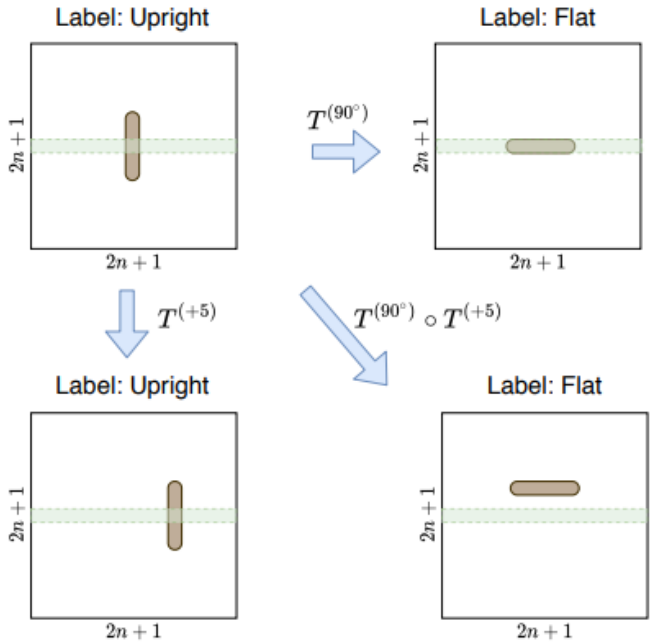

Figure 2: A counterexample that shows not all G-invariances are CG-invariances. The goal of the example problem is to determine the orientation of an upright or flat rod in an image. A representation that sums the middle row of the image is (G-)invariant to horizontal translations of the image but not 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 causal information 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 $\bigoplus_{N \supset M} \mathcal{B}_N$ 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, the set of invariant subspaces that are each computed from a subset in the power set of groups are disjoint. This result also implies a partial ordering on the invariant subspaces where a subspace corresponding to a superset of groups is defined to have "stronger" invariance than a subspace corresponding to a 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 [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 important for practical use. Lemma 2 (linear transformation is group-invariant if and only if the inner product vector is in the 1-eigenspace of the Reynolds operator) is mainly used to show that we can restrict considerations to the 1-eigenspace of the Reynolds operator if invariance is what we are after. Its proof follows from direct logical derivations in both directions where sufficiency is shown by 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 TODO

2.3 Learning framework for single-environment extrapolation

Suppose that $\mathcal{G}_1, \ldots, \mathcal{G}_m$ are known linear automorphisms. 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 neural networks). 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 orthogonal basis. The parameters are collected in a neuron weight matrix

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

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

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

where $\lambda > 0$ is a regularization parameter, $R(\bullet)$ is the regularization penalty given by

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

and $\ell = \min\{|M_i| \cdot \mathbf{1}\{\|\boldsymbol{\omega}_{M_i,\bullet}\|_2^2 > 0\} : 1 \leq i \leq B\}$. The number ℓ describes the smallest size across all sets of groups M_i that are used by at least one neuron. The penalty $R(\bullet)$ then counts the number of sets that are larger or are equal in size to the smallest set. This objective encourages Γ to use a subspace that is invariant to more groups. Note that while $R(\bullet)$ is discrete, a differentiable approximation is available for optimization.

To use the learning framework, the subspaces for the power set of groups must first be computed. While the procedure only needs to be run once for a particular set of groups, the runtime is technically exponential as the subspace needs to be computed for every set in the power set. The procedure can be set to terminate early once a subspace equal in size to the space of the input is found, and Mouli and Ribeiro comment that it is unclear if the worst-case runtime occurs in practice.

2.4 Analysis

TODO[MR21] requires specifying known linear groups of transformations.

- What implications are there with infinite linear groups? Learning framework finds invariant subspace of group using properties (idempotency) of projection operator.
  - Reynolds operator changes to integral over normalized Haar measure of (locally compact) group?
  - Provably Strict Generalisation Benefit for Invariance in Kernel Methods (Elesedy 2021). Lemma 3: RKHS decomposes into space of G-invariant functions and space of functions that vanish when averaged over an orbit. Both spaces are RKHS with different kernels.
  - Relation to normal subgroups and quotient spaces?
- If considering non-normal subgroups, then Theorem 2 does not hold. Have to work with CG-invariances rather than G-invariances? What exactly does this mean?
- Can the framework be adapted for some equivariant objective? i.e., learning equivariant groups rather than invariant groups.
- Framework breaks down with non-groups. Can a non-group approach work? Can non-group structures be specified/linear? Can a variant of Theorem 2 be derived?
- Can something be done without assuming known groups? Consider largest possible linear overgroup and automatically learn the strongest invariances?

- Does the chosen regularizing term make sense? Is it optimal? Does its differentiable approximation affect these properties?
- The proposed algorithm is exponential as it involves the power set. Can it be reduced? Appendix D: "...the algorithm stops after finding all the basis for the space vec($\mathcal{X}$), it is unclear if the worst-case runtime occurs in practice."

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. It is well-known that for a given function f and a group $\mathcal{G}$, averaging the function over the Haar measure λ of the group produces the $\mathcal{G}$-invariant function
$$\bar{f}(x) = \int_{\mathcal{G}} f(gx) \mathrm{d}\lambda(g) \, .$$
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

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.

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

[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].

[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. ISBN: 9783540569633;3540569634;0387569634;97

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

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

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