TODO

Kenny Chiu

September 26, 2021

1 Conceptual summary

The paper by Mouli and Ribeiro [MR21] examines the problem of extrapolating patterns learned from training data from a single environment in a supervised setting to data from other environments. This problem context falls under the idea 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 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 known (linear automorphism) group structures on the non-causal mechanisms. Under this formulation, Mouli and Ribeiro introduce a learning framework for the single-environment context that is able to learn invariances that do not contradict the data. In this conceptual summary, we review the key contributions of the paper by Mouli and Ribeiro [MR21] and discuss the strengths and weaknesses of their approach.

1.1 Key differences from previous work

Various methods for domain adaptation have been proposed in the literature, and how the work by Mouli and Ribeiro [MR21] relates to these methods are highlighted in their paper. For example, methods based on causal inference such as IRM and *Independent Causal Mechanisms* [Par+18] (ICM) broadly involve learning some internal representation of the data that is invariant to environment non-causal mechanisms. The invariant representation is learned from the training data which come from multiple environments. When the data come from a single environment, the representation cannot determine which aspects of the data are environment-specific and so the 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 well-known 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. Mouli and Ribeiro explains 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). Mouli and Ribeiro's proposed learning framework aims to learn only the invariances that do not contradict the training data.

1.2 Main contributions

The main contributions of the paper by Mouli and Ribeiro [MR21] include a formulation of the single-environment extrapolation problem, a learning framework based on the formulation 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 is used to describe the input variables that influence or are irrelevant to the internal representation across environments (TODOcite?). Extrapolation is then seen as counterfactual reasoning where being able to extrapolate to different environments is tied to the representation being invariant to interventions on non-causal environment variables. Mouli and Ribeiro extends this idea by assuming known linear automorphism groups acting on the non-causal variables, in which case extrapolation of a representation is equivalent to the representation being counterfactually invariant to a group (i.e., *CG-invariant*). This extension is the crux of the formulation that allows the proposed learning framework to work with single-environment data.

The learning framework proposed by Mouli and Ribeiro aims to learn an internal representation that is CG-invariant to specified groups that do not contradict the training data. Mouli and Ribeiro show that CG-invariance is stronger than G-invariance (Theorem 1), but when the group acting on the non-causal variables is a normal subgroup of the group acting on all variables, then G-invariance also implies CG-invariance (Theorem 2). These results establish the group conditions under which it is sufficient for the model to learn G-invariances in place of CG-invariances.

The challenge in learning the G-invariances that do not contradict the training data is due to the fact that the set 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. A group-invariant transformation can be constructed by averaging over members of the group (the *Reynolds operator*, Lemma 1). For linear transformations, the averaged transformation is a projection operator with eigenvalues 1 and 0. The left eigenspace spanned by eigenvectors with eigenvalue 1 represents the space of transformations that are invariant to the group (Lemma 2). Mouli and Ribeiro exploit this property by computing the intersection of these subspaces for all subsets of the set of groups, and the set of invariant subspaces can then be partially ordered by the size of their corresponding subset (i.e., the *strength* of the invariance, Theorem 3). The objective optimized in the learning framework then includes a regularization term that encourages learning a representation with the strongest G-invariance that does not significantly contradict the data. The key aspects of this learning framework include needing to specify known groups, requiring the groups to be linear automorphisms and, in doing so, being able to automatically learn the G-invariances that do not (significantly) contradict the data.

Mouli and Ribeiro 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.

1.3 Limitations

The main limitations of the learning framework proposed by Mouli and Ribeiro [MR21] originate from the assumptions made. To allow for single-environment data, the framework requires that the invariance groups acting on the data are known and specified. Furthermore, to enable automatic learning of invariances that do not contradict the training data, the groups are also restricted to be linear automorphisms. This framework cannot be used if no linear group structure could placed on the transformations that act on the data.

These limitations naturally point to directions for future work where non-linear groups or even non-groups of transformations are considered. It is also worth investigating if a method for single-environment extrapolation that does not require known groups is possible.

2 Technical summary

The technical components of the paper by Mouli and Ribeiro [MR21] include the proposed learning framework and the theoretical results that justify its usage in the given setting. In this technical summary, we introduce the formulation and notation of the single-environment extrapolation problem, discuss the assumptions made in the formulation, and describe the proposed learning framework for use under the formulated setting.

2.1 Single-environment extrapolation setting

In the context of single-environment extrapolation described by Mouli and Ribeiro [MR21], the goal is to learn (in a supervised learning setup) a prediction model where the output only depends on information that is relevant across different environments. The challenge is learning which information is relevant given only training data from a single environment. To simplfy this problem, Mouli and Ribeiro assume a known set of groups $\mathcal{G}_1, \ldots, \mathcal{G}_m$ of linear transformations acting on the data, and the objective is to learn an internal representation of the data that is invariant to an unknown subset of groups assumed to be irrelevant to the output. Mouli and Ribeiro works under an ICM setup where the SCM in Figure 1 is assumed.

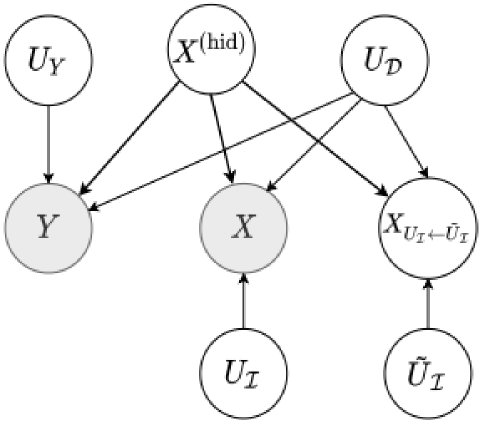
U_Y
$X^{(\text{hid})}$
$U_{\mathcal{D}}$
Y
X
$X_{U_{\mathcal{I}} \leftarrow \tilde{U}_{\mathcal{I}}}$
$U_{\mathcal{I}}$
$\tilde{U}_{\mathcal{I}}$

Figure 1: Structural causal model assumed in single-environment extrapolation. Grey nodes are observed variables. Figure taken from [MR21].

The variables in the assumed SCM are defined as follows:

- $\mathcal{D}$, $\mathcal{I}$: (unknown) disjoint sets of indices that describe the groups of transformations 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 influence the value of the variable(s) that they point to.
- $X^{(\text{hid})}$: some unknown canonical form of the observed input X. It is assumed that given $U_{\mathcal{D}}$ and $U_{\mathcal{I}}$, X was obtained from an ordered sequence of transformations on the canonical form, i.e.,

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

  where transformations

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

  make up the overgroup $\mathcal{G}_{\mathcal{D} \cup \mathcal{I}}$, $T_{\mathcal{D}}^{(j)}$ 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 $T_{\mathcal{I}}^{(i)} \in \mathcal{G}_i \subset \mathcal{G}_{\mathcal{I}} = \langle \cup_{i \in \mathcal{I}} \mathcal{G}_i \rangle$. Note that $\mathcal{G}_{\mathcal{I}}$ is also assumed to be a normal subgroup of $\mathcal{G}_{\mathcal{D} \cup \mathcal{I}}$.

- 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 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. Compared to previous work in the literature, the unconventional assumptions involve the transformation groups acting on the data.

Unlike previous work that assumes the availability of training data from multiple environments, the problem context 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 likely impossible to learn what pieces of information are environment-specific and irrelevant to the output. Mouli and Ribeiro get around 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, but G-invariances are CG-invariances only when $\mathcal{G}_{\mathcal{I}} \trianglelefteq \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$. The assumption is made as in practice, it is easier to work with G-invariances ($\forall T_{\mathcal{I}} \in \mathcal{G}_{\mathcal{I}}$, $\Gamma(X) = \Gamma(T_{\mathcal{I}} \circ X)$) than CG-invariances due to its simpler definition. The proof of Theorem 1 (CG-invariance ⇒ 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 the definitions of CG-invariance and G-invariance 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 ⇒ CG-invariance when $\mathcal{G}_{\mathcal{I}} \trianglelefteq \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$) uses the fact that under the required assumption, any $T_\mathcal{D} \circ T_\mathcal{I} \in \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$ can be written as $T_\mathcal{D} \circ T_\mathcal{I} = T'_\mathcal{I} \circ T_\mathcal{D}$ where $T_\mathcal{I}, T'_\mathcal{I} \in \mathcal{G}_\mathcal{I}$, $T_\mathcal{D} \in \mathcal{G}_\mathcal{D}$. Showing CG-invariance given any sequence of transformations $T \in \mathcal{G}_{\mathcal{D}\cup\mathcal{I}}$ then reduces to repeatedly rewriting the sequence with a leading $T'_\mathcal{I}$ that is then removed using the G-invariance of the representation Γ. CG-invariance is then shown after only transformations from $\mathcal{G}_\mathcal{D}$ remain in the sequence.

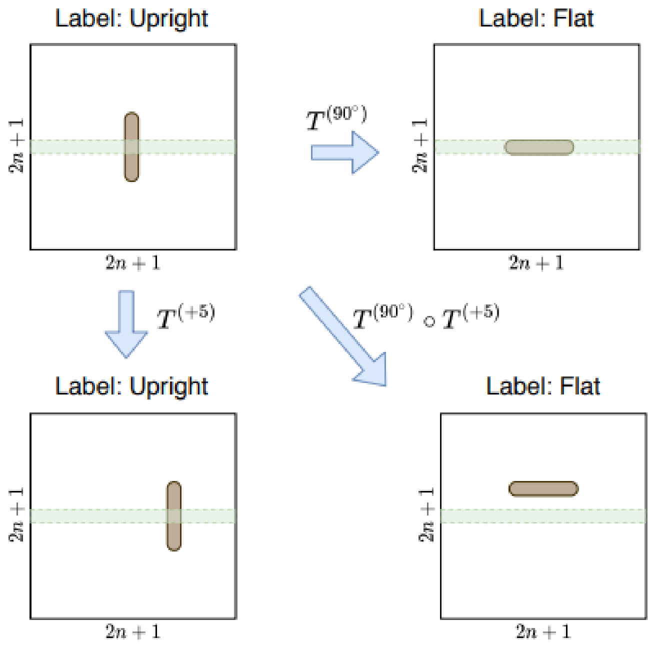

Figure 2: A counterexample for showing not all G-invariances are CG-invariances. The goal is to determine the orientation of an upright or flat rod in an image. A representation that sums the middle row of the image is (G-)invariant to horizontal translations of the image but not invariant to 90^o rotations. 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 taken from [MR21].

While a set of transformation groups are assumed a priori, it is not known which groups actually influence the output. Additionally, some of the groups may specify invariances that contradict the training data. Mouli and Ribeiro deal with this problem in their learning framework by obtaining a partial ordering on the invariant subspaces in terms of their "strength" and encouraging learning of only the strongest invariances through a regularized objective. In order to identify the invariant subspaces and obtain a partial ordering, Mouli and Ribeiro require the transformation groups to be linear automorphisms. This restriction implies that the Reynolds operator given by

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

is a G-invariant projection operator (Lemma 1). The invariant subspace for a group $\mathcal{G}$ is then precisely the left eigenspace corresponding to eigenvalue 1 of the Reynolds operator (Lemma 2). To construct the invariant subspace for a set of groups indexed by a set M (Theorem 3), the intersection of the 1-eigenspaces for the individual groups is taken. However, the intersection may also contain vectors that are invariant to some overgroup, and so the group's projection onto the subspace of the overgroup is removed from the intersection. The subspace $\mathcal{B}_M$ that remains describes the vectors that are invariant to transformations only from the set of groups that it was built from. (TODOproof) Mouli and Ribeiro's learning framework relies on being able to identify the invariant subspace for every subset of groups. Once the subspaces have been identified, a cost can be assigned to each subspace that encourages adopting the subspace invariant to the most number of groups.

2.3 Learning framework for single-environment extrapolation

Suppose that $\mathcal{G}_1, \dots, \mathcal{G}_m$ 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

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

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

TODOis this required in the summary?

3 Mini-proposals

3.1 Proposal 1: MY PROPOSAL TITLE

3.2 Proposal 2: MY OTHER PROPOSAL TITLE

[MR21] requires specifying known linear groups of transformations.

- Any finite group is linear? What implications are there with infinite linear groups?
- 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 something be down without assuming known groups? Consider largest possible linear overgroup and automatically learn the strongest invariances?
- 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?
- Learning framework finds invariant subspace of group using properties (idempotency) of projection operator. When not working in linear spaces, is there an analagous property of non-linear operators?
- Does the chosen regularizing term make sense? Is it optimal? Does its differentiable approximation affect these properties?
- Can the framework be adapted for some equivariant objective? i.e., learning equivariant groups rather than invariant groups.
- 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.”

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.

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

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