

1 Proposal scratch notes

- \mathcal{D}, \mathcal{I} : (unknown) disjoint sets of indices that describe the groups of transformations that are relevant and irrelevant to the output, respectively. $\mathcal{D} \cup \mathcal{I} = \{1, \dots, m\}$.
- $U_Y, U_{\mathcal{I}}, U_{\mathcal{D}}, \tilde{U}_{\mathcal{I}}$: 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

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

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.

- $X_{U_{\mathcal{I}} \leftarrow \tilde{U}_{\mathcal{I}}}$: counterfactual variable to X where $U_{\mathcal{I}}$ has been replaced by $\tilde{U}_{\mathcal{I}}$, i.e.,

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

- Want CG-invariant representation

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

When $\mathcal{G}_{\mathcal{I}}$ is a normal subgroup of $\mathcal{G}_{\mathcal{D} \cup \mathcal{I}}$, G-invariant representation

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

for all $T_{\mathcal{I}} \in \mathcal{G}_{\mathcal{I}}$ is sufficient.

- Because groups are finite linear automorphisms, each transformation T is just a linear function and so Reynolds operator can be applied directly to the group actions

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

For continuous linear groups, orbit-average over a Haar measure λ

$$\bar{T} = \int_{\mathcal{G}} T \lambda(T)$$

TODO: need to estimate the operator. Assume uniform Haar and sample?

2 Kernel hypothesis test notes

- Let X be a r.v. on domain \mathcal{X} . A kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ induces a RKHS \mathcal{H} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ where for $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$\langle f, k(x, \cdot) \rangle = f(x)$$

(reproducing property). $k(x, \cdot) = \phi_x$ can also be considered an implicit feature map ($\phi_x : \mathcal{X} \rightarrow \mathbb{R}$) where

$$\langle \phi_x, \phi_{x'} \rangle = k(x, x')$$

is a measure of similarity.

- Kernel methods work on inner products of feature maps of observations in the RKHS associated with kernel. Inner products may be computed without explicitly computing the high-dimensional feature map (“kernel trick”).
- Main challenge of designing kernel-based hypothesis tests is deriving large-sample distribution of test statistic under null.
- Gram matrix should be positive semidefinite. Satisfied if kernel is symmetric and positive semidefinite.
- Let X be a r.v. with distribution \mathbb{P} . [Mean element](#)

$$\mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[\phi_X]$$

associated with X is unique element of RKHS \mathcal{H} s.t. for all $f \in \mathcal{H}$,

$$\langle \mu_{\mathbb{P}}, f \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$$

Covariance operator $\Sigma_{\mathbb{P}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ associated with X is unique operator s.t. for all $f, g \in \mathcal{H}$,

$$\langle f, \Sigma_{\mathbb{P}} g \rangle = \text{Cov}(f(X), g(X)) = \mathbb{E}_{\mathbb{P}}[f(X)g(X)] - \langle \mu_{\mathbb{P}}, f \rangle \langle \mu_{\mathbb{P}}, g \rangle$$

Empirical estimates of inner products that lead to estimates of element/operator are available.

- Kernel is characteristic if mean embedding $\mu : \mathbb{P} \rightarrow \mathcal{H}$ is injective. Each distribution can be uniquely represented in the RKHS and all statistical features of distributions are preserved ([TODO](#)) by a characteristic kernel.
- If $\dim(\mathcal{H}) = \infty$, $\mu_{\mathbb{P}}$ has more significance than in classical statistics.
- (Kellner, 2015) MMD(? or related quantity) is a pseudo-metric. If restricting space of functions (e.g., r.v. space) to unit ball of RKHS with positive semi-definite characteristic kernel, MMD is a metric.

3 Detecting conditional invariances in single training environment via hypothesis testing

- Case of invariance of output to single known group acting on inputs. Following (Mouli, 2021) in that we specify potential group of invariant transformations and testing if dataset contradicts invariance to the group, rather than if dataset appears to be invariant to group.

[TODO](#) Contradict in hard sense (e.g., certain transformation impossible) or distributional sense (e.g., certain transformations unlikely)?

- Other assumptions?

- Possible two-sample test procedure: generate transformed inputs as “second” sample and compare conditional mean embeddings (e.g., Gretton, 2007). How should embeddings be estimated? (Song, 2013) Conditional embedding operator is a family of points in RKHS. Only when conditioned on a fixed value is it a single point.

Given dataset $\{(x_i, y_i)\}_{i=1}^N$ assumed to be from some joint distribution $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_{Y|X}$, want to determine if $\mathbb{P}_{Y|X} \stackrel{d}{=} \mathbb{P}_{Y|\mathcal{G} \cdot X}$ for group \mathcal{G} where $\mathbb{P}_{Y|\mathcal{G} \cdot X}(y|x) = \mathbb{P}_{Y|\mathcal{G} \cdot X}(y|g \cdot x)$ for all $g \in \mathcal{G}, x \in \mathcal{X}$. (Elesedy, 2021) obtains invariant functions by orbit-averaging. We want a metric of some form **TODO**

$$\text{MMD}(\mathbb{P}, \mathbb{P}') = \|\mu_{Y|X} - \mu_{Y|\mathcal{G} \cdot X}\|_{\mathcal{H}}^2$$

How to handle X and $\mu_{Y|X}$ being a family of functions? $\mu_{Y|\mathcal{G} \cdot X}$ estimated by generating/summing/integration transformed inputs?

Comparing joint distributions only makes sense if it is assumed $\mathbb{P}(X) = \mathbb{P}(\mathcal{G} \cdot X)$? This does not make sense for extrapolation context ($Y|X$ does not have to be invariant then?).

- If conditional is invariant to group, then conditional embedding should be same for orbit of conditioned value. Does that imply anything for conditional embedding operator? Projection? Invariant conditional embedding operator conditioned on orbits?
- Hard contradictions vs soft contradictions?
- One-sample test? Via parametric bootstrap (Kellner, 2015)?

TODO

Context 1:

- Assume that we have a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, from a single environment, i.e., $x_i \sim \mathbb{P}_X$ and $y_i \sim \mathbb{P}_{Y|X}$. Let $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_{Y|X}$ be the joint distribution corresponding to that environment.
- Assume that the differences in environments can be modeled by some unknown group \mathcal{G} acting on inputs $X \in \mathcal{X}$, e.g., $\mathbb{P}(X)$ and $\mathbb{P}(g \cdot X)$ for $\text{id}() \neq g \in \mathcal{G}$ are not equal and correspond to different environments. Assume that there is a true $\mathbb{P}_{Y|X}$ that holds across environments.
- We try to determine if $\mathbb{P}_{Y|X}$ being invariant to transformations $g \in \mathcal{G}$ on X for some specified \mathcal{G} would contradict the dataset, i.e., whether $\mathbb{P}_{Y|X} \stackrel{d}{=} \mathbb{P}_{Y|g \cdot X}$ is plausible for \mathcal{D} .
- Proposed approach 1: bootstrap samples from different environments by applying transformations to X then do kernel hypothesis test (via MMD) to compare conditional embeddings. **TODO**Problem1: if $\mathbb{P}(X)$ and $\mathbb{P}(g \cdot X)$ are not equal, how to bootstrap sample from a different environment? **TODO**Problem2: even if invariance does not “contradict”, embedded distributions may be far apart? **TODO**Problem3: how to compare conditional embedding operator?

Context 2: observed g

- Assume that we have a dataset $\mathcal{D} = \{(x_i, g_i, y_i)\}_{i=1}^n$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $g_i \in \mathcal{G}$ for some known group \mathcal{G} acting on \mathcal{X} , from a single environment, i.e., $x_i \sim \mathbb{P}_X$, $g_i \sim \mathbb{P}_G$, $y_i \sim \mathbb{P}_{Y|X, G}$. Let $\mathbb{P}_{XYG} = \mathbb{P}_X \mathbb{P}_G \mathbb{P}_{Y|X, G}$ be the joint distribution corresponding to that environment. Examples of when g is observed: image orientation, image background.
- Assume that \mathbb{P}_G differs between environments.
- We try to determine if $\mathbb{P}_{Y|X, G} = \mathbb{P}_{Y|X}$, i.e., if $\mathbb{P}_{XYG} = \mathbb{P}_X \mathbb{P}_{Y|X} \mathbb{P}_G$. **TODO**This can be done through permutation-based kernel conditional independence tests?

- **TODO**If non-diverse g observed, should not be a problem because does not contradict (test not rejected).

Context 3:

- Assume that we have a dataset $\mathcal{D} = \{(x_i, g_i, y_i)\}_{i=1}^n$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $g_i \in \mathcal{G}$ for some known group \mathcal{G} acting on \mathcal{X} , from a single environment, i.e., $x_i \sim \mathbb{P}_X$, $g_i \sim \mathbb{P}_G$, $y_i \sim \mathbb{P}_{Y|X,G}$. Let $\mathbb{P}_{XYG} = \mathbb{P}_X \mathbb{P}_G \mathbb{P}_{Y|X,G}$ be the joint distribution corresponding to that environment. Examples of when g is observed: image orientation, image background.
- Assume that \mathbb{P}_G differs between environments.
- We try to determine if $\mathbb{P}_{Y|X,G} = \mathbb{P}_{Y|X}$. **TODO**By Lemma 1 of [Elesedy:2021], \mathbb{P}_{XY} is invariant if and only if $\mathbb{P}_{XY} \stackrel{d}{=} \int_{\mathcal{G}} \mathbb{P}_{XYG} d\lambda$. We can then do standard two-sample kernel tests where second sample is obtained by orbit-averaged? **TODO**This is testing if distributions are consistent, not whether invariance contradicts.

TODOtwo-sample approach involving original and predicted datasets with invariant predictor?