The Crossover Process: Learnability meets Protection from Inference Attacks

Richard Nock
Data61 & The Australian National University
richard.nock@data61.csiro.au

Giorgio Patrini
The Australian National University & Data61
giorgio.patrini@anu.edu.au

Finnian Lattimore
The Australian National University & Data61
finnian.lattimore@nicta.com.au

Tiberio Caetano
Ambiata, The Australian National University & The University of New South Wales
tiberio.caetano@gmail.com

Abstract

It is usual to consider data protection and learnability as conflicting objectives. This is not always the case: we show how to jointly control causal inference — seen as the attack — *and* learnability by a noise-free process that mixes training examples, the Crossover Process (CP). One key point is that the CP is typically able to alter joint distributions *without* touching on marginals, nor altering the sufficient statistic for the class. In other words, it saves (and sometimes improves) generalization for supervised learning, but can alter the relationship between covariates — and therefore fool statistical measures of (nonlinear) independence and causal inference into misleading *ad-hoc* conclusions. Experiments on a dozen readily available domains validate the theory.

1 Introduction

There are at least two good reasons to alter the learning sample of a supervised learning problem. One is to help the process of learning: feature bagging [20, 37], importance weighting in features, boosting [28], denoising autoencoders [41], independent noisification, dropout [40], all are methods that optimise the presentation of examples to a learner with the objective to improve its generalisation abilities. The second is privacy, a growing concern in the public sphere [3, 9, 12]. Two leading mechanisms for the private release of data are differential privacy and k-anonymity [8, 9, 27, 38]. They guarantee *individual level* protection, *i.e.* identifiability, and rely on low-level modifications, typically touching marginal distributions, mostly with noise.

Nevertheless, as pointed out in [3], "even when individuals are not *identifiable* they may still be *reachable* [...] and subject to *consequential inferences* and predictions taken on that basis". The reference is to the possibility of performing *causal inference attacks* by a malicious agent willing to uncover *causal relationships*, or even just measure *statistical independence*, between sensitive covariates. In supervised learning, the task is the prediction of a single feature, yet the total number of observation variables available for the task is blowing up in mainstream datasets. Each subset of these observation variables is a potential target for attacks, and even when true causality is sometimes considered "a research field in its infancy" [12], it is hard to exaggerate the recent burst in causal inference techniques [6, 7, 14, 15, 16, 17, 21, 23], as well as the threats this may pose on privacy [2, 3, 9, 19]. For example, in a medical diagnosis data which gives a sickness state as a function of genetic, behavioural, habits and infection history, we may want to make causal inference between specific traits and behaviour harder, or we may want to hide gender-prone infections [32], *while* making sure that the utility of the dataset for predicting the sickness state remains unaltered. Ultimately, rather than making it harder, we may just look to *reverse* the observable causal direction between specific observation variables, and thereby fool causal inference or causal rule mining into misleading *ad-hoc* conclusions. In short, we target two different levels: the dataset's utility for the black-box supervised prediction task remains within control, but it is surgically altered against fined-grained specific causal inference attacks among features.

As we show, this task is within reach. Even when coping with the level of protection to statistical independence and causal inference attacks may require wrangling the complete data, this may be done with a tight *explicit* control of its utility for supervised learning, and it may even yield *better* models for prediction. Although counterintuitive, this last fact should not come with great surprise considering the success of sophisticated noisification methods, *e.g.* dropout, to enhance learning.

— This is our main contribution: to achieve our two goals, we introduce the (single-point) *Crossover Process*, CP. An analogy may be done with the biological crossover: a population of DNA strands gets mixed with a crossover, but there is a single zone for chiasma (*i.e.* contact point) for the whole population. In the same way as DNA strands exchange genetic material during recombination, feature values get mixed between observations during a CP , although in a more general way than in genetic recombination. The key to learning is that the CP may be done without changing the sufficient statistic for the class, nor touching class-marginals. The key to interfering with measures of independence and causal calculus is that the CP is able to surgically alter joint distributions. Our contribution is therefore twofolds: (i) we introduce the CP and show how it drives the generalisa-

tion abilities of linear and non-linear classifiers by the introduction of a new statistical complexity measure, the Rademacher CP complexity (RCP). We show that the RCP can be very significantly smaller than the standard empirical Rademacher complexity, thereby being a lightweight player — and a tractable knob — for generalisation. Then, (ii) we show how the components of a CP may be chosen to alter the powerful Hilbert-Schmidt independence criterion [15], how it may be devised to blow-up causal estimation errors [6], and finally how it can interfere with identifiable causal queries on a causal graph in the *do* calculus framework [30, 35].

Organisation of the paper — Section §2 gives general definitions. §3 presents the Crossover Process and its relationships with learnability. §4 shows the impact of the CP on measures of independence and causal queries. §5 details related experiments. A last Section discusses and concludes. An Appendix, starting page 17, provides all proofs, additional results and experiments performed. A movie[1], presented in Subsection 7.3.2, shows the effects of the CP on a popular domain for causal discovery [16].

2 General notations and definitions

Learning setting — We let $[m] \doteq \{1, 2, ..., m\}$ and $\Sigma_m \doteq \{\boldsymbol{\sigma} \in \{-1, 1\}^m\}$. $\mathcal{X} \subseteq \mathbb{R}^d$ is a domain of observations. Examples are couples (observation, label) $\in \mathcal{X} \times \Sigma_1$, sampled i.i.d. according to some unknown but fixed distribution $\mathcal{D}$. We denote $\mathcal{F} \doteq [d]$ the set of observation attributes (or features). $\mathcal{S} \doteq \{(\boldsymbol{x}_i, y_i), i \in [m]\} \sim \mathcal{D}_m$ is a training sample of $|\mathcal{S}| = m$ examples. For any vector $\boldsymbol{z} \in \mathbb{R}^d$, z_j denotes its coordinate j. Finally, notation $x \sim X$ for X a set denotes uniform sampling in X, and the mean operator is $\boldsymbol{\mu}_{\mathcal{S}} \doteq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{S}}[y \cdot \boldsymbol{x}]$ [29].

In supervised learning, the task is to learn a classifier $\mathcal{H} \ni h : \mathcal{X} \to \mathbb{R}$ from $\mathcal{S}$ with good generalisation properties, that is, having a small *true risk* $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[L_{0/1}(y, h(\boldsymbol{x}))]$, with $L_{0/1}(z, z') \doteq 1_{zz'\leq 0}$ the 0/1 loss ($1_{\cdot}$ is the indicator variable). In general, this is achieved by minimising over $\mathcal{S}$ a *φ-risk* $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{S}}[\varphi(yh(\boldsymbol{x}))] = (1/m) \cdot \sum_i \varphi(y_i h(\boldsymbol{x}_i))$, where $\varphi(z) \geq 1_{z\leq 0}$ is a *surrogate* of the 0/1 loss. In this paper, φ is any differentiable proper symmetric (PS) loss [26, 29] (symmetric meaning that there is no class-dependent misclassification cost). The logistic, square and Matsushita losses are examples of PS losses. Set $\mathcal{H}$ is a predefined set of classifiers, such as linear separators, decision trees, etc. .

Matrix quantities — The set of unnormalised column stochastic matrices, $\mathcal{M}_n \subset \mathbb{R}^{n\times n}$, is the superset of column stochastic matrices for which we drop the non-negativity constraint, thus keeping the sole constraint of unit per-*column* sums. We let $S_n \subset \mathcal{M}_n$ denote the symmetric group of order n. For any $\mathrm{A}, \mathrm{B} \in \mathbb{R}^{n\times n}$ and $\mathrm{M} \in \mathcal{M}_n$, we let

$$\langle \mathrm{A}, \mathrm{B} \rangle_{\mathrm{M}} \quad \doteq \quad \mathrm{tr}\left((\mathrm{I}_n - \mathrm{M})^\top \mathrm{A} (\mathrm{I}_n - \mathrm{M}) \mathrm{B}\right)$$

denote the *centered inner product* of A and B with respect to M. It is a generalisation of the centered inner product used in kernel statistical tests of independence [14], for which $\mathrm{M} = (1/n)\mathbf{1}\mathbf{1}^\top$.

Without loss of generality, we shall assume that indices in $\mathcal{S}$ cover first the positive class: $(y_i = +1 \wedge y_{i'} = -1) \Rightarrow i < i'$. A key subset of matrices of $\mathbb{R}^{m\times m}$ consists of block matrices whose coordinates on indices corresponding to different classes in $\mathcal{S}$ are zero: block-class matrices.

[1] Available at `http://users.cecs.anu.edu.au/~u5647716/cp/abalone_movie.mp4`

Definition 1 $\mathrm{A} \in \mathbb{R}^{m \times m}$ *is a* ***block-class matrix*** *iff* $(y_i \cdot y_{i'} = -1) \Rightarrow \mathrm{A}_{ii'} = 0, \forall i, i'$.

An asterisk exponent in a subset of matrices indicates the intersection of the set with block class matrices, such as for $\mathcal{M}_n^* \subset \mathcal{M}_n$ and $S_n^* \subset S_n$. Finally, matrix entries are noted with double indices like $\mathrm{M}_{ii'}$; replacing an index by a dot, ".", indicates a sum over the index, like $\mathrm{M}_{i\cdot} \doteq \sum_{i'} \mathrm{M}_{ii'}$.

3 The Crossover Process and learnability

The Crossover process (CP) transforms $\mathcal{S}$ in two steps: the split and the shuffle step. In the split step, a bi-partition of the features set $\mathcal{F}$ is computed: $\mathcal{F} = \mathcal{F}_\mathrm{a} \cup \mathcal{F}_\mathrm{s}$. $\mathcal{F}_\mathrm{a}$ is the *a*nchor set and $\mathcal{F}_\mathrm{s}$ is the *s*huffle set. To perform the shuffle step, we need additional notations. Without loss of generality, we assume $\mathcal{F}_\mathrm{a} \doteq [d_\mathrm{a}]$ and $\mathcal{F}_\mathrm{s} \doteq \{d_\mathrm{a} + j, j \in [d_\mathrm{s}]\}$, $d_\mathrm{a} > 0, d_\mathrm{s} > 0, d_\mathrm{a} + d_\mathrm{s} = d$. So, $\mathcal{F}_\mathrm{a}$ contains the first d_a features and $\mathcal{F}_\mathrm{s}$ contains the last d_s features. Let I_d be the identity matrix, and $[\mathrm{F}^\mathrm{a}|\mathrm{F}^\mathrm{s}] = \mathrm{I}_d$ a vertical block partition where $\mathrm{F}^\mathrm{a} \in \mathbb{R}^{d \times d_\mathrm{a}}$ ($\mathrm{F}^\mathrm{s} \in \mathbb{R}^{d \times d_\mathrm{s}}$) has columns representing the features of $\mathcal{F}_\mathrm{a}$ ($\mathcal{F}_\mathrm{s}$) — we use notation $[.]$ both for integer sets and block matrices without ambiguity. Finally, we define the (*row*-wise) observation matrix $\mathrm{S} \in \mathbb{R}^{m \times d}$ with $(\mathrm{S})_{ij} \doteq x_{ij}$. Let $\mathbf{1}_i$ be the i^{th} canonical basis vector.

Definition 2 *For any block partition* $[\mathrm{F}^\mathrm{a}|\mathrm{F}^\mathrm{s}] = \mathrm{I}_d$ *and any* ***shuffle matrix*** $\mathrm{M} \in \mathcal{M}_n$*, the Crossover process* $\mathscr{T} \doteq \mathrm{CP}(\mathcal{S}; \mathrm{F}^\mathrm{a}, \mathrm{F}^\mathrm{s}, \mathrm{M})$ *returns* m*-sample* $\mathcal{S}^\mathscr{T}$ *such that its observation matrix is* $\mathrm{S}^\mathrm{M} \doteq [\mathrm{SF}^\mathrm{a}|\mathrm{MSF}^\mathrm{s}]$*, and each example* $\mathcal{S}^\mathscr{T} \ni (\boldsymbol{x}_i^\mathrm{M}, y_i) \doteq ((\mathrm{S}^\mathrm{M})^\top \mathbf{1}_i, y_i)$.

We consider M fixed beforehand. Figure 1 (top) presents the CP on a toy data with M a permutation matrix. Figure 1 (bottom) presents another example with M block-uniform.

Learnability — We now explore the effect of the CP on generalisation. We need two assumptions on $\mathcal{H}$ and φ. The first is a weak linearity condition on $\mathcal{H}$:

(i) $\forall h \in \mathcal{H}$, $\exists$ classifiers $h_\mathrm{a}, h_\mathrm{s}$ over $\mathcal{F}_\mathrm{a}, \mathcal{F}_\mathrm{s}$ s. t. $h(\boldsymbol{x}) = h_\mathrm{a}((\mathrm{F}^\mathrm{a})^\top \boldsymbol{x}) + h_\mathrm{s}((\mathrm{F}^\mathrm{s})^\top \boldsymbol{x})$.

$(\mathrm{F}^\mathrm{s})^\top \boldsymbol{x}$ picks the features of $\boldsymbol{x}$ in $\mathcal{F}_\mathrm{s}$. Such an assumption is also made in the feature bagging model [37]. Any linear classifier satisfies **(i)**, but also any linear combination of arbitrary classifiers, each learnt over one of $\mathcal{F}_\mathrm{a}$ and $\mathcal{F}_\mathrm{s}$. We let $\mathcal{H}_\mathrm{s}$ denote the set of all h_s. The second assumption postulates that key quantities are bounded [4]:

(ii) $0 \leq \varphi(z) \leq K_\varphi, \forall z$ and $|h_\mathrm{s}((\mathrm{F}^\mathrm{s})^\top \boldsymbol{x})| \leq K_\mathrm{s}, \forall \boldsymbol{x} \in \mathcal{X}, \forall h_\mathrm{s} \in \mathcal{H}_\mathrm{s}$.

Let $\mathrm{R}_\mathcal{S}(\mathcal{H}) \doteq \mathbb{E}_{\boldsymbol{\sigma} \sim \Sigma_m} [\sup_{h \in \mathcal{H}} |(1/m) \cdot \sum_i \sigma_i h(\boldsymbol{x}_i)|]$ be the empirical Rademacher complexity of $\mathcal{H}$. Additionally, we coin the Rademacher CP complexity, RCP.

Definition 3 *The Rademacher* CP *complexity (*RCP*) of* $\mathcal{H}$ *with respect to* $\mathscr{T} \doteq \mathrm{CP}(\mathcal{S}; \mathrm{F}^\mathrm{a}, \mathrm{F}^\mathrm{s}, \mathrm{M})$ *is:*

$$\mathrm{RCP}_\mathscr{T}(\mathcal{H}) \doteq \mathbb{E}_{\boldsymbol{\sigma} \sim \Sigma_m} \left[\sup_{h \in \mathcal{H}_\mathrm{s}} \left| \frac{1}{m} \sum_i \sigma_i \left(h((\mathrm{SF}^\mathrm{s})^\top \mathbf{1}_i) - h((\mathrm{MSF}^\mathrm{s})^\top \mathbf{1}_i) \right) \right| \right] . \quad (1)$$

Notice that the RCP is computed over the shuffle set of features only, and $(\mathrm{SF}^\mathrm{s})^\top \mathbf{1}_i = (\mathrm{F}^\mathrm{s})^\top \boldsymbol{x}_i$. The next Theorem expresses a generalisation bound *wrt* the CP.

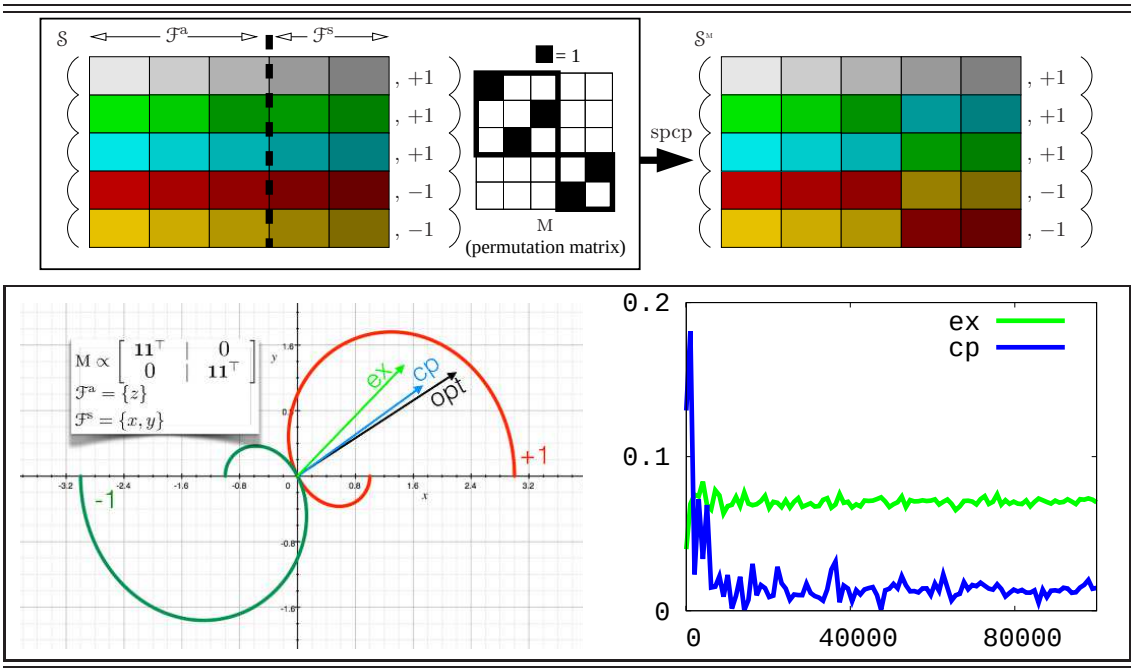


Figure 1: *Top*: example of CP with M a block-class permutation matrix (the two blocks are in bold). *Bottom*: toy domain where $d = 3$, but all examples have zero z-coordinate (not shown). The CP uniformly mixes examples by class. The domain consists of two spirals (red for positive, green for negative examples) with $\mathcal{D}$ = uniform distribution. Arrows depict respectively the optimal direction (black), and the directions learned by minimizing φ = square loss over $\mathcal{S}$ (light green, "ex") and $\mathcal{S}^{\mathscr{T}}$ (blue, "CP"). The right plot displays test errors (y-scale) on uniform sampling of datasets of different sizes (x-scale). The effect of the CP is to produce in $\mathcal{S}^{\mathscr{T}}$ *two* distinct examples that *average* the positive / negative examples, and yield a better approximation of the optimum.

Theorem 4 *Consider any $\mathcal{H}$, φ and split $\mathcal{F} = \mathcal{F}_{\rm a} \cup \mathcal{F}_{\rm s}$ such that* ***(i)*** *and* ***(ii)*** *hold. For any m and any $\delta > 0$, with probability $\geq 1 - \delta$ over i.i.d. m-sample $\mathcal{S}$, we have:*

$$\mathbb{E}_{\mathcal{D}}\left[L_{0/1}(y, h(\boldsymbol{x}))\right] \;\leq\; \mathbb{E}_{\mathcal{S}^{\mathscr{T}}}\left[\varphi(yh(\boldsymbol{x}))\right] + \mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) + \frac{4}{b_\varphi} \cdot \mathrm{R}_{\mathcal{S}}(\mathcal{H}) + (2K_\varphi + K_{\rm s}) \cdot \sqrt{\frac{2}{m} \log \frac{3}{\delta}} \;,$$

for every classifier h and every $\mathscr{T} \doteq \mathrm{CP}(\mathcal{S}; \mathrm{F}^{\rm a}, \mathrm{F}^{\rm s}, \mathrm{M})$ such that $\mathrm{M} \in \mathcal{M}^*_m$. Here, $b_\varphi > 0$ is a constant depending on φ.

(proof in Subsection 7.2.1) Notice that Theorem 4 requires that M is a block-class matrix. A key to the proof is the invariance of the mean operator: $\boldsymbol{\mu}_{\mathcal{S}} = \boldsymbol{\mu}_{\mathcal{S}^{\mathscr{T}}}$. Theorem 4 says that a key to good generalisation is the control of $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})$. We would typically want it to be small compared to the Rademacher complexity penalty. The rest of this Section shows that (and when) this is indeed achievable.

Upperbounds on $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})$ — We consider different configurations of $\mathcal{H}$ and / or $\mathscr{T}$:

Setting (A): Classifiers $h^{\rm s}$ and $h^{\rm a}$ in **(i)** above are linear;

Setting (B): $\mathrm{M} \in S_m^*$.

The following Lemma establishes a first bound on $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})$.

Lemma 5 *if $\mathscr{T}$ satisfies the conditions of Theorem 4, then $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) \leq 2 \cdot \mathrm{R}_{\mathcal{S}'}(\mathcal{H}_{\mathrm{s}})$, for $\mathrm{S}' \doteq (\mathrm{I}_m - \mathrm{M})\mathrm{SF}^{\mathrm{s}}$ in Setting (A), and $\mathrm{S}' \doteq \mathrm{SF}^{\mathrm{s}}$ in Setting (B). S' is the row-wise observation matrix of $\mathcal{S}'$.*

Proof (Sketch) Consider for example Setting (B). In this case, recalling that $(\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i = (\mathrm{F}^{\mathrm{s}})^\top \boldsymbol{x}_i$ and letting $\varsigma : [m] \to [m]$ denote the permutation that M represents, we have because of the triangle inequality:

$$\begin{aligned}
\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}_{\mathrm{s}}}\left|\frac{1}{m}\sum_i \sigma_i\left(h((\mathrm{F}^{\mathrm{s}})^\top \boldsymbol{x}_i) - h((\mathrm{F}^{\mathrm{s}})^\top \boldsymbol{x}_{\varsigma(i)})\right)\right|\right] \qquad (2)\\
&\leq \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}_{\mathrm{s}}}\left|\frac{1}{m}\sum_i \sigma_i h((\mathrm{F}^{\mathrm{s}})^\top \boldsymbol{x}_i)\right|\right] + \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}_{\mathrm{s}}}\left|\frac{1}{m}\sum_i \sigma_i h((\mathrm{F}^{\mathrm{s}})^\top \boldsymbol{x}_{\varsigma(i)})\right|\right]\\
&= 2 \cdot \mathrm{R}_{\mathcal{S}'}(\mathcal{H}_{\mathrm{s}}) \ ,
\end{aligned}$$

as claimed. The case of Setting (A) follows the same path. ■

Lemma 5 says that $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})$ is at most twice a Rademacher complexity over the *shuffle set*. This bound is however loose since many terms can cancel in the sum of eq. (2), and the inequality does not take this into account. In particular,

Theorem 6 *Under Setting (A), suppose any h_{s} is of the form $h_{\mathrm{s}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$ with $\|\boldsymbol{\theta}\|_2 \leq r_{\mathrm{s}}$, for some $r_{\mathrm{s}} > 0$. Let $\mathrm{K}^{\mathrm{s}} \doteq \mathrm{SF}^{\mathrm{s}}(\mathrm{SF}^{\mathrm{s}})^\top$. Then $\exists u \in (0,1)$ depending only in $\mathcal{S}$ such that for any $\mathrm{M} \in \mathcal{M}_m$,*

$$\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) \leq (ur_{\mathrm{s}}/m) \cdot \sqrt{\langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}} \rangle_{\mathrm{M}}} \ . \qquad (3)$$

Notice that K^{s} is a Gram matrix in the shuffle feature space. The proof technique (Subsection 7.2.3) relies on a data-dependent expression for u which depends on the cosines of angles between the observations in $\mathcal{S}$. It can be used to refine and improve a popular bound on the empirical Rademacher complexity of linear classifiers [18] (we give the proof in Theorem 15 in the Appendix). We now investigate an upperbound on Setting (B) in which classifiers in $\mathcal{H}^{\mathrm{s}}$ are (rooted) directed acyclic graph (DAG), like decision trees, with bounded real valued predictions (say, $K_{\mathrm{s}} > 0$) at the leaves. Each classifier h_{s} defines a partition over $\mathcal{X}$. We let $\mathcal{H}^{\mathrm{s}}_+$ be the subset of $\mathcal{H}^{\mathrm{s}}$ in which all leaves have in absolute value the largest magnitude, *i.e.*, K_{s}. Remark that we may have $|\mathcal{H}^{\mathrm{s}}_+| \ll \infty$ while $|\mathcal{H}^{\mathrm{s}}| = \infty$ in general.

Theorem 7 *Under Setting (B), suppose $\mathcal{H}^{\mathrm{s}}$ is* DAG *and assumption **(ii)** is satisfied. Suppose that $\log|\mathcal{H}^{\mathrm{s}}_+| \geq (4\varepsilon/3) \cdot m$ for some $\varepsilon > 0$. Then, letting $\mathrm{odd\_cycle}(\mathrm{M})$ denote the set of odd cycles (excluding fixed points) of* M*, we have:*

$$\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) \leq K_{\mathrm{s}} \cdot \sqrt{\frac{2}{m} \cdot \log \frac{|\mathcal{H}^{\mathrm{s}}_+|}{(1+\varepsilon)^{|\mathrm{odd\_cycle}(\mathrm{M})|}}} \ . \qquad (4)$$

(proof in Subsection 7.2.4) The assumption on $\mathcal{H}^{\mathrm{s}}_{+}$ is not restrictive and would be met by decision trees, branching programs, etc. (and subsets). Usual bounds on the Rademacher complexity of decision trees would roughly be the right-hand side of (4) *without* the denominator in the $\log$ (see for example [34, Chapter 5]). Hence, the RCP may be significantly smaller than the Rademacher complexity for more "involved" CPs. The number of cycles is not the only relevant parameter of the CP on which relies non-trivial bounds on $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})$: the Appendix presents, for the interested reader, a proof that the number of fixed points is another parameter which can decrease significantly the *expected* RCP (by a factor $\sqrt{1-|\mathrm{fixed\_points}|/m}$), when CPs are picked at random (see Theorem 17 and discussion in Subsection 7.2.4).

At last, we notice that Theorem 4 gives a perhaps counterintuitive rationale for the CP that goes beyond our framework to machine learning at large: *learning over a* CP*'ed* $\mathcal{S}$ *may improve generalisation over* $\mathcal{D}$ *as well*. By means of words, learning over transformed data may improve generalisation over the initial domain. Figure 1 (bottom) gives a toy example for which this holds. It is also not hard to exhibit domains for which we even have:

$$\min_{h} \mathbb{E}_{\mathcal{S}^{\mathscr{T}}} \left[\varphi(yh(\boldsymbol{x}))\right] + \mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) \quad < \quad \min_{h} \mathbb{E}_{\mathcal{S}} \left[\varphi(yh(\boldsymbol{x}))\right] \quad . \tag{5}$$

In order not to load the paper's body, we present such an example in the Appendix (Subsection 7.2.2). Having discussed learning guarantees, we are ready to dive into applications of the Crossover Process.

4 The Crossover Process versus statistical independence and causality

With CP, we are able to destroy or fake statistical measures of independence and causal relationships in the data. We develop two frameworks: Hilbert-Schmidt independence criterion and *do* calculus. In the former one, our results exploit the design of the shuffle matrix M to alter independence; in the latter one, our results exploit the split step of the CP to interfere with causal inference.

The Crossover Process and measures of independence — Here, we assume that $\mathcal{S}$ is subject to quantitative tests of independence, that is, assessing $\mathcal{U} \perp\!\!\!\perp \mathcal{V}$ for some $\mathcal{U}, \mathcal{V} \subset \mathcal{X}$. We compute CPs such that $\mathcal{U} \subseteq \mathcal{F}_{\mathrm{a}}$ and $\mathcal{V} \subseteq \mathcal{F}_{\mathrm{s}}$, so that the CP alters the measure of independence. One popular criterion to determine (conditional) (in)dependence is Hilbert-Schmidt Independence Criterion [7, 14, 15].

Definition 8 *Let* $\mathcal{U} \subset [d]$ *and* $\mathcal{V} \subset [d]$ *be non-empty and disjoint. Let* K^u *and* K^v *be two kernel functions over* $\mathcal{U}$ *and* $\mathcal{V}$ *computed using* $\mathcal{S}$*. The (unnormalised) Hilbert-Schmidt Independence Criterion (*HSIC*) between* $\mathcal{U}$ *and* $\mathcal{V}$ *is defined as* $\mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v) \doteq \langle \mathrm{K}^u, \mathrm{K}^v \rangle_{(1/m)\mathbf{1}\mathbf{1}^\top}$.

(We choose not to normalise the HSIC: various exist but they mainly rely on a multiplicative factor depending on m only, so they do not affect the results to come.) The choice of M in the CP directly influences the value the result of HSIC; therefore, we can design a search strategy aimed to alter it. Our first result shows that this is also algorithmic friendly: (a) while storing kernels requires $O(m^2)$ space, controlling the evolution of the HSIC requires only *linear*-space information about kernels, and (b) this information can be computed beforehand, and can be efficiently approximated from

low-rank approximations of the kernels [1]. Hereafter, we restrict ourselves to the scenario of the *decrease* of the HSIC, yet all our results would apply to the opposite polarity for the modification. We use for sake of readability the shorthand HSIC for $\text{HSIC}(\mathrm{K}^u, \mathrm{K}^v)$.

Theorem 9 *Let* HSIC *and* $\text{HSIC}_{\mathscr{J}}$ *denote the* HSIC *before and after applying the* CP $\mathscr{T}$ *to* $\mathcal{S}$*. Let* $\tilde{\boldsymbol{u}} \doteq (1/m)\sum_i \lambda_i(\mathbf{1}^\top \boldsymbol{u}_i)\boldsymbol{u}_i$*,* $\tilde{\boldsymbol{v}} \doteq (1/m)\sum_i \mu_i(\mathbf{1}^\top \boldsymbol{v}_i)\boldsymbol{v}_i$*, where* $\{\lambda_i, \boldsymbol{u}_i\}_{i\in[d]}$*,* $\{\mu_i, \boldsymbol{v}_i\}_{i\in[d]}$ *are respective eigensystems of* K^u *and* K^v*. Then*

$$\text{HSIC}_{\mathscr{J}} < \text{HSIC} \quad \textit{iff} \quad \tilde{\boldsymbol{u}}^\top(\mathrm{I}_m - \mathrm{M})\tilde{\boldsymbol{v}} < 0 \ .$$

(proof in Subsection 7.2.5) In the following Theorem, we compose CP processes with T different elementary permutation shuffling matrices. Notice that since the composition of permutation matrices is a permutation matrix, when the matrix of the final process $\text{HSIC}_{\mathscr{T}_T}$ is block class, Theorem 4 can be applied *directly* to $\text{HSIC}_{\mathscr{T}_T}$. We also let $\mathscr{R}^{u,v} \doteq m\,(1 - (\mathrm{K}^u_{\cdot\cdot} + \mathrm{K}^v_{\cdot\cdot})/(2m^2))$.

Theorem 10 *Suppose* $\mathcal{S}$ *is shuffled by a sequence of* $T = \epsilon m$ *elementary permutation (*$\epsilon > 0$*) and the kernels* K^u *and* K^v *have unit diagonal. Suppose that the initial* $\text{HSIC} > \mathscr{R}^{u,v}$*. Then there exists such a sequence such that* $\mathscr{T}_T$ *satisfies*

$$\text{HSIC}_{\mathscr{T}_T} \ \leq \ \mathscr{R}^{u,v} + \alpha \cdot (\text{HSIC} - \mathscr{R}^{u,v}) \ ,$$

where $\alpha \doteq \exp(-8\epsilon)$.

The proof (Subsection 7.2.6) states a more general result, not restricted to unit diagonal kernels. Theorem 10 is a worst-case result: some sequences of permutations may be much more efficient in decreasing HSIC. If we compare this bound to Theorem 3 in [15], then $\mathscr{R}^{u,v}$ may be *below* the expectation of the HSIC, and so Theorem 10 guarantees efficient jamming of dependence.

Remark: it is in fact possible to kill two birds with one stone, namely trick statistical tests into keeping independence *and* then incur arbitrarily large errors in estimating causal effects. Our basis is the Cornia-Mooij (CM) model [6] (see Appendix, Subsection 7.2.7) which, for space considerations, we defer to the Appendix as well as for how we can use the CP to achieve both goals. One interesting feature of this particular causal graph is that it is so simple that it may be found as subgraph of real-world domains, thus for which the results we give would directly transfer.

The Crossover Process and causal effects — We now consider the case where the causal directed acyclic graph is known, and the goal is to interfere with the inference of causal effects between covariates. The key challenge for causal inference is the existence of confounding variables that are causes of both the exposure and outcome variables. For example, suppose impact of hormone replacement therapy on women's health was captured by the causal DAG $I \to T \to H, I \to H$, where I represents income, T represents taking the treatment and H is the health outcome of interest. If wealthier women are more likely to see a doctor for treatment and also have generally better health, then $\Pr[H|T]$ will be more positive than the true causal effect $\Pr[H|do(T)]$.

Adjusting for such nuisance or confounding variables, either by matching [13, 33] or regression [10], is a central tool in economics and social sciences [24]. The back-door criterion [30] clarifies which variables it is appropriate to condition on in order to achieve unbiased estimates of causal effects. The CP can be designed to interfere with obtaining causal estimates via such adjustments.

Let $G = (\mathcal{U} \cup \mathcal{F}, \mathcal{A})$ be a causal directed acyclic graph over observable vertices $\mathcal{F}$, latent variables $\mathcal{U}$ and arcs $\mathcal{A}$ [30]. We are given a set $\mathcal{Q} \doteq \{(x_i, x_i'), i \in [q]\}$ of q causal queries, each of which represents the estimation of $\Pr[x_i|\mathrm{do}(x_i')]$.

An a (covariate) adjustment for a query (x_i, x_i') is a set $\mathcal{Z}_i \subset \mathcal{F}$ such that $x_i', x_i \notin \mathcal{Z}_i$ and

$$\Pr[x_i|\mathrm{do}(x_i')] \;=\; \sum_{z \sim \mathcal{Z}_i} \Pr[x_i|x_i', z] \Pr[z] \;\;, \tag{6}$$

An adjustment is not guaranteed to exist. In our example, for the query (H, T), there is no adjustment if $I \in \mathcal{U}$. An adjustment is minimal iff it does not contain any other adjustment as proper subset. Note that $\mathcal{Z}_i$ can be the empty set.

We say that the split $\mathcal{F}_\mathrm{a}$ and $\mathcal{F}_\mathrm{s}$ *interferes* with causal inference via adjustment for a query (x_i, x_i') if there exists a shuffle matrix M such that the solution to (6) differs between the datasets $\mathcal{S}$ and $\mathcal{S}^{\mathscr{T}}$, with $\mathscr{T} \doteq \textsc{cp}(\mathcal{S}; \mathrm{F}^\mathrm{a}, \mathrm{F}^\mathrm{s}, \mathrm{M})$. We put no constraint on the magnitude of the change, so interfering with causal queries is essentially a matter of biasing the distributions involved in the right-hand side of (6). Let Z_i denote the set of minimal adjustments for query (x_i, x_i').

Lemma 11 *Let $\mathcal{V}_i = x_i' \cup x_i \cup \mathcal{Z}_i$. The split $\mathcal{F}_\mathrm{a}$ and $\mathcal{F}_\mathrm{s}$ interferes with causal inference via adjustment for the query (x_i, x_i') iff $\forall \mathcal{Z}_i \in \mathsf{Z}_i$, $\exists$ variables $v_\mathrm{a}, v_\mathrm{s} \in \mathcal{V}_i$ such that $v_\mathrm{a} \in \mathcal{F}_\mathrm{a}$ and $v_\mathrm{s} \in \mathcal{F}_\mathrm{s}$.*

The proof is a direct consequence of eq. (6) and the fact that the shuffle matrix M alters joint distributions between variables that do not belong to the same split set, without touching marginals. We can always interfere with a single query (x_i, x_i') by ensuring x_i and x_i' are in different splits. To simultaneously interfere with the set of queries $\mathcal{Q}$, we must first find the set of minimal adjustments $\mathsf{Z}_i, \forall i \in [q]$, then select a split that satisfies Lemma 11 for every query. This involves heavy combinatorics. Enumerating the adjustments can be done with cubic *delay* per adjustment [39](the set of minimal adjustments for a given query can grow exponentially with d). The second step subsumes the infamous Set Splitting problem [11], which is NP-Complete. In practice, causal graphs must often be constructed by humans so this approach can still be computationally feasible for a small set of queries. Exploring the addition of constraints on the graph (such as sparsity) to develop more efficient algorithms is an interesting avenue for future research.

A causal query is *identifiable* if we can obtain an expression for it purely in terms of distributions over the observable variables $\mathcal{F}$. The existence of an adjustment is sufficient but not necessary for identifiability. In theory, a causal query for which we have interfered with any adjustments, could still be identified via another approach. The Do-Calculus [30] and Identify Algorithm [35] provide a complete framework for determining if a query is identifiable and computing an expression for it. In principle, we could utilize the CP to interfere with all routes to identifiability. This would require an algorithm that could enumerate the expressions for a causal query. We are not aware of such an algorithm in the literature. In practice, expressions that are not of the form of 6 are rarely used.

5 Experiments

Algorithm 1 Crossover Learning ($\mathcal{S}, T, \boldsymbol{\theta}_0, \mathrm{M}_0, \mathrm{F}^{\mathrm{a}}, \mathrm{F}^{\mathrm{s}}; \mathscr{G}_1, [\mathscr{G}_2]$)

Input Sample $\mathcal{S}$, iterations T, classifier $\boldsymbol{\theta}_0$, initial CP $\mathscr{T}$ (matrices $\mathrm{M}_0 \in S_m^*$, $[\mathrm{F}^{\mathrm{a}}|\mathrm{F}^{\mathrm{s}}] = \mathrm{I}_m$);
Step 1 : **for** $t = 1, 2, ..., T$
  Step 1.1 : $\mathrm{M} \leftarrow \arg\min_{\mathrm{M}' \in S_m^{\mathrm{e}*}} \mathscr{G}_1(\mathrm{M}' \circ \mathrm{M}_{t-1}[|\boldsymbol{\theta}_{t-1}])$;
    // finds update of shuffle matrix
  Step 1.2 : $\mathrm{M}_t \leftarrow \mathrm{M} \circ \mathrm{M}_{t-1}$;
    // updates shuffle matrix
  [Step 1.3 : $\boldsymbol{\theta}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathscr{G}_2(\boldsymbol{\theta}|\mathrm{M}_t)$;]
    // (optionally) updates classifier
Return classifier $\boldsymbol{\theta}_T$ and / or CP'ed dataset $\mathcal{S}^{\mathscr{T}(\mathrm{M}_T)}$

Our applications use the same meta-level algorithm (Algorithm 1) which operates in Setting (A) ∩ Setting (B) (M in S_m^*, linear classifiers), iteratively composing block-class elementary permutations. Here, $S_m^{\mathrm{e}*} \subset S_m^*$ is the set of block-class elementary permutations. The iteration step minimises a criterion $\mathscr{G}_1$ over $S_m^{\mathrm{e}*}$ and potentially, after the update of the CP matrix, a criterion $\mathscr{G}_2$ over $\mathcal{H}$. The optimization of $\mathscr{G}_1$ is performed by a simple greedy search in the space of $S_m^{\mathrm{e}*}$. The experimental setup (a dozen readily available domains) and results are provided *in extenso* in the Appendix (Section 7.3.3); Table 1 summarises them. The split step and the choice of $\mathcal{F}_{\mathrm{a}}$ are highly domain and task dependent: to keep experiments of reasonable length, unless otherwise stated, we put in $\mathcal{F}_{\mathrm{a}}$ the first half of features. In Abalone2D and Digoxin, this jams a particular ground truth (see below). Also, φ=logistic loss.

5.1 Disrupting dependence and causality

General experiments — We run Algorithm 1 without step 1.3, and let $\mathscr{G}_1$ be HSIC. As a proof of concept, we show that we can destroy the significance of statistical tests for independence, commonly as base for causal inference; in particular, we measure the change in the p-value computed on top of HSIC as in [15]. We use two Gaussian kernels for K^u and K^v, each computed over its full subset of features (Subsection 7.3.1) — hence, we do not seek to alter specifically the dependence between two features, but between the two sets of features defined by the anchor and shuffle sets. On most domains (Table 1, top row), the p-value of the independence test starts close to zero at the beginning of the CP, which implies that in general both anchor and shuffle sets are (predictably) dependent. In general, we achieve a good control of the RCP and manage in several cases to decrease the true error as well through the process.

Specific dependences — The general experiments revealed that we manage to blow-up the p-value, even for domains for which the ground truth clearly *implies the alternative hypothesis* H_1. To dive into this phenomenon, we have considered two sets of experiments on which HSIC is computed over two specific features that are known to have a causal relationship. To disrupt the dependence, we thus put one of the features in the anchor set and one in the shuffle set of features (*i.e.*, after we have split the feature set in two, if both features belong to the same set, we switch one with a randomly chosen feature of the other set).

In the first set of experiments, we consider datasets with $d = 2$ features, so that this surgical disruption embeds kernels measured over the complete set of features. Experiments are reported in Table 1 for domains Abalone2D and Digoxin (Subsection 7.3.1). In Abalone2D for example,

a gold standard for dependence [16], the final p-values is more than *ten billion* times the initial value. For Digoxin domain, another popular domain [7] with ground truth, p is very small at the beginning (which corresponds to the ground truth $D \not\perp\!\!\!\perp U$; D = digoxin clearance, U = urin flow). After shuffling, we obtain $p > 0.4$, which easily brings $D \perp\!\!\!\perp U$, while ground truth is $D \perp\!\!\!\perp U|C$ (C = creatinine clearance). A rather surprising fact is that in both cases, the effect on test error is minimal, considering that Abalone2D and Digoxin have $d = 2$ attributes only.

In the second set of experiments, we consider several domains with a larger d, between 7 and 279. These domains belong to the benchmarks of [23] in which specific pairs are known to have specific causal relationships, referred to as "causal tasks", indicated in Subsection 7.3.1. We have targeted one causal task for each domain. Table 2 summarizes the results obtained. In all domains, the p-value is blown up at almost no expense in test error. Quite remarkably, the initial value is indeed $p = 0$ (up to *sixteen* digits), while we manage at the end of the process to get p that exceeds 1‰, which would be quite sufficient to raise doubts about the causal relationships for sensitive domains. For example, in `pair0016`, the final $p > 2$‰ might lead us to keep the independence assumption between horsepower and acceleration. In Arrhythmia, we would keep the (obviously fake) independence between age and weight. In the Liver disorder domain, our experiment has the following interesting consequence. Causal task `pair0034` is the causality relationship between alcohol consumption and the measure of alkaline phosphatase (ALP, [23]). It is known that ALP elevation may be caused by heavy alcohol consumption. By keeping the independence assumption for such a value of p, one may just discourage specific blood tests related to alcohol consumption if they were to be designed from this domain. Again, the variation of the test error is minimal (if any) for all these domains.

5.2 Data optimisation for efficient learning

In the previous subsection, we showed how a CP may be carried out to target directly the disruption of causal relationships. In this subsection, we analyse how (and when) it can be devised to improve the test performances of classifiers. We perform Algorithm 1 with $\mathrm{M}_0 = \mathrm{I}_m$, $\mathscr{G}_1(\mathrm{M}|\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{S}^{\mathscr{T}}(\mathrm{M})}\left[\varphi(y\boldsymbol{\theta}^\top \boldsymbol{x})\right]$ (Theorem 4) and $\mathscr{G}_2(\boldsymbol{\theta}|\mathrm{M}) = \mathbb{E}_{\mathcal{S}^{\mathscr{T}}}\left[\varphi(y\boldsymbol{\theta}^\top \boldsymbol{x})\right] + \lambda\|\boldsymbol{\theta}\|_2^2$ where λ is learnt through cross-validation. The algorithm returns classifier $\boldsymbol{\theta}_T$. The bottom row in Table 1, and Appendix (Subsection 7.3.3) shows how we almost always find some permutations that reduce the test error compared to the initial data, even when a specific data optimisation should care for a risk of over-fitting, which seems to occur for problems with a very small number of features (see Ionosphere and Abalone2D, Subsection 7.3.3). It appears also that when the (bound on the) RCP flattens, it may indicate a regime where substantial reductions can be obtained on test error as witnessed by *e.g.* Synthetic, Heart, Glass (see also BreastWisc in Table 10). We also remarked that the regime where the (bound on the) RCP flattens can be associated with a peak or decrease of the number of odd cycles as seen in Table 13, which, interestingly, can be used to provide upper-bounds on the RCP as well (Theorem 7). Finally, the results on Glass display that some domains (predictably) make it possible to kill to birds in one shots at little effort: in this case, the CP both reduces the test error and increases the p-value for HSIC computed as in the general experiments of Subsection 5.1.

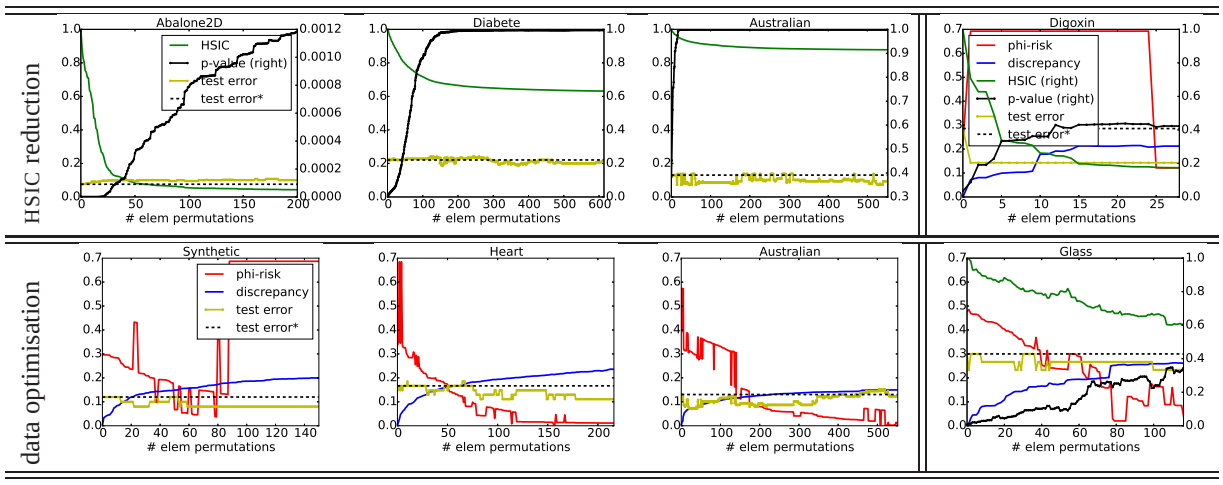


Table 1: Experiments performed with CP. Top row: reduction in HSIC task; bottom row: data optimisation task. References to domain names are provided in Appendix, Subsection 7.3.1. "test-error*" is test error over initial, non shuffled data. "discrepancy" is an upperbound on the RCP provided by Theorem 6. The rightmost column aggregates the information of all curves for the task in the row, for two different domains.

6 Discussion and conclusion

This paper introduces the Crossover Process (CP), a mechanism that cross-modifies data using a generalisation of stochastic matrices. This process can be used to cope with data optimisation for supervised learning, as well as for the problem of handling a process-level protection on data: causal inference attacks on a supervised learning dataset. In this case, the CP allows to release data with spotless low-level description (variable names, observed values, marginals), substantial utility (learnability), but disclosing dependences and causal effects under control, and thus that could even be crafted to be conflicting with a ground truth to protect[2]. Note that causality in "big data" may still be in its infancy [12], it is however rapidly growing with a variety of techniques and lots of promises. We have chosen to focus here on two major components of the actual trends, namely statistical measures of causal inference and causal queries. In these two directions, there are some very interesting and non-trivial avenues for future research, like for example the control of combinatorial blow-up in the worst case for causal queries, ideally as a function of the causal graph structure. There are also more applications of the CP in the field of causal discovery. Suppose for example that description features denote transactions. Since we modify joint distributions without touching on marginals, our technique has direct applications in causal rule mining, with the potential to fool any level-wise association rule mining algorithms, that is, any spawn of Apriori [21].

The theory we develop fo CP introduces a new complexity measure of the process, the Rademacher CP complexity. We do believe that the CP is also a good contender in the pool of methods optimising data for learning, and it may provide new metrics, algorithms and tools to devise improved

[2]Note that the initial data may not be lost, as opposed to differential privacy: knowing the noise parameters does not allow to revert differential privacy protection, while a CP protection is reversible when $\mathbf{M}$ is invertible.

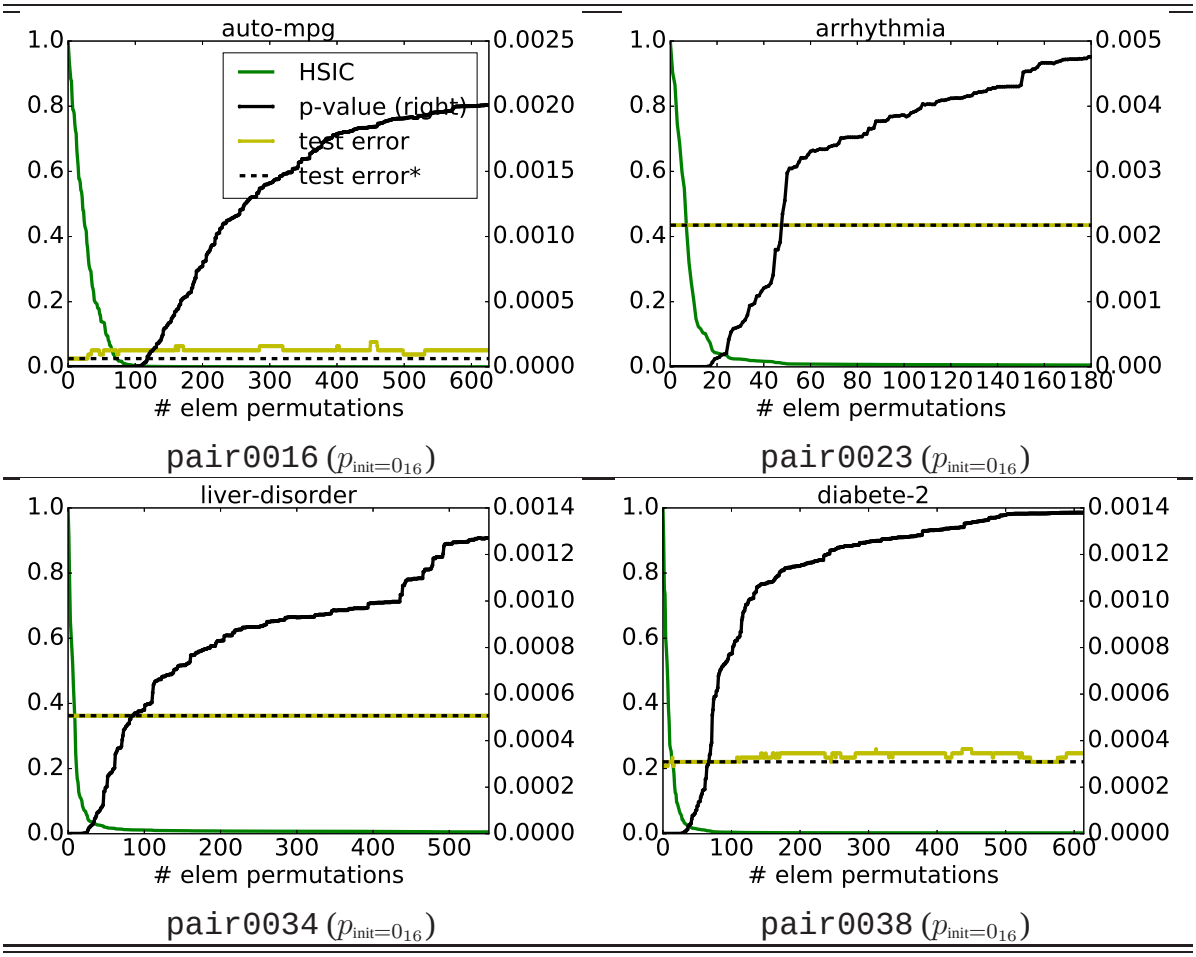


Table 2: HSIC reduction on specific causal tasks [23] (referred to as `pair00XX`); datasets indicated on pictures; "p_{init}" is the initial p-value, 0_N indicating zero up to N^{th} digit (see Table 1 for additional notations, and text for details).

solutions that fit to challenging domains not restricted to optimizing learning or data privacy.

Acknowledgments

Work carried out in NICTA which was supported by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Center of Excellence Program.

References

[1] F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *26 th COLT*, pages 185–209, 2013.

[2] S. Barocas and H. Nissenbaum. Big data's end run around procedural privacy protection. *Communications of the ACM*, 57:31–33, 2014.

[3] S. Barocas and H. Nissenbaum. Big data's end run around anonymity and consent. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, editors, *Privacy, Big Data, and the Public Good*, pages 44–75. Cambridge University Press, 2014.

[4] P.-L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[5] S. Chatterjee. Stein's method for concentration inequalities. *Probability Theory and Related Fields*, 138:305–321, 2007.

[6] N. Cornia and J.-M. Mooij. Type-II errors of independence tests can lead to arbitrarily large errors in estimated causal effects: An illustrative example. In *30th UAI Workshops*, pages 35–42, 2014.

[7] G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *30 th UAI*, 2014.

[8] C. Dwork and A. Roth. The algorithmic foudations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407, 2014.

[9] M. Enserink and G. Chin. The end of privacy. *Science*, 347:490–491, 2015.

[10] R.-A Fisher. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935.

[11] M.R. Garey and D.S. Johnson. *Computers and Intractability, a guide to the theory of NP-Completeness*. Bell Telephone Laboratories, 1979.

[12] S. Graham, W. Press, S.-J. Gates, M. Gorenberg, J.-P. Holden, E. Lander, C. Mundie, M. Savitz, and E. Schmidt. Big data and privacy: a technological perspective. CreateSpace Independent Publishing Platform, 2015. — President's Council of Advisors on Science and Technology.

[13] E. Greenwood. *Experimental sociology: A study in method*. King's crown Press, 1945.

[14] A. Gretton, O. Bousquet, A.-J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *16th ALT*, pages 63–77, 2005.

[15] A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *NIPS*20*, pages 585–592, 2007.

[16] P.-O. Hoyer, D. Janzing, J.-M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*21*, pages 689–696, 2008.

[17] D. Janzing, J.-M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

[18] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*21*, pages 793–800, 2008.

[19] M.-J. Kusner, Y. Sun, K. Sridharan, and K.-Q. Weinberger. Inferring the causal direction privately. In *19th AISTATS*, 2016.

[20] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *11th KDD*, pages 157–166, 2005.

[21] J. Li, T.-D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. From observational studies to causal rule mining. *ACM Trans. IST*, 7:1–27, 2016.

[22] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 1–54. Springer Verlag, 1998.

[23] J.-M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR*, 2016.

[24] S.-L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

[25] R. Nock and F. Nielsen. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pages 1201–1208, 2008.

[26] R. Nock and F. Nielsen. Bregman divergences and surrogates for learning. *IEEE Trans.PAMI*, 31:2048–2059, 2009.

[27] R. Nock, G. Patrini, and A. Friedman. Rademacher observations, private data, and boosting. *32 nd ICML*, 2015.

[28] J. O'Sullivan, J. Langford, R. Caruana, and A. Blum. Featureboost: A meta-learning algorithm that improves model robustness. In *17 th ICML*, pages 703–710, 2000.

[29] G. Patrini, R. Nock, P. Rivera, and T. Caetano. (Almost) no label no cry. In *NIPS*27*, 2014.

[30] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011.

[32] S.-J. Rizvi and J.-R. Haritsa. Maintaining data privacy in association rule mining. In *Proc. of the 28 th VLDB*, pages 682–693, 2002.

[33] D.-B. Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

[34] R.-E. Schapire and Y. Freund. *Boosting: foundations and algorithms*. MIT press, 2012.

[35] I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[36] L. Song, A. Smola, A. Gretton, J. Bedo, and K.-M. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

[37] C.-A. Sutton, M. Sindelar, and A. McCallum. Reducing weight undertraining in structured discriminative learning. In *10^{th} HLT-NAACL*, 2006.

[38] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertainty, Fuzziness and Knowledge-based systems*, 10:571–588, 2002.

[39] J. Textor and M. Liskiewicz. Adjustment criteria in causal diagrams: An algorithmic perspective. *CoRR*, abs/1202.3764, 2012.

[40] L. van der Maaten, M. Chen, S. Tyree, and K.-Q. Weinberger. Learning with marginalized corrupted features. In *30 th ICML*, pages 410–418, 2013.

[41] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *25 th ICML*, pages 1096–1103, 2008.

7 Appendix

7.1 Table of contents

7.2 Appendix — Proofs

We shall use the following notations and shorthands. We shall sometimes replace notation $\boldsymbol{x}_i$ by $\boldsymbol{x}_i^b$ for $b \in \{-,+\}$, indicating explicitly an observation from class $b1$. We also let m_b denote the number of examples in class $b1$ for $b \in \{-,+\}$ ($m = m_+ + m_-$). Furthermore, $[m]_{m'} \doteq \{m' + i : i \in [m]\}$ ($m \in \mathbb{N}_*, m' \in \mathbb{N}$). Matrix $\mathrm{U}_m \doteq (1/m)\mathbf{1}\mathbf{1}^\top$ denotes the uniform Markov chain.

7.2.1 Proof of Theorem 4

The first steps of the proof are the same as [4] (Theorems 5, 8). We sketch them. First, for any CP $\mathscr{T}$,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[L_{0/1}(y,h(\boldsymbol{x}))\right] &\leq \mathbb{E}_{\mathcal{D}}\left[\varphi(yh(\boldsymbol{x}))\right] \\ &\leq \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right] + \sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{D}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\} \quad . \end{aligned} \tag{7}$$

Then, since $\varphi(z) \in [0, K_\varphi]$ (assumption **(ii)**), the use of the independent bounded differences inequality [22] yield for any sample $\mathcal{S}$ sampled from $\mathcal{D}$, and any $\varsigma \in S_m$, and any δ_1, we have with probability $\geq 1 - \delta_1$:

$$\begin{aligned}&\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{D}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\} \\ &\quad\leq \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{D}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right] + K_\varphi \cdot \sqrt{\frac{2}{m}\log\frac{1}{\delta_1}} \quad . \end{aligned} \tag{8}$$

We also have, because of the convexity of $\sup$ (see [4]),

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{D}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right] \\ &\quad\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{S}'}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right] \quad . \end{aligned}$$

The proof now takes a fork compared to [4], as we integrate new steps to upperbound the right-hand side. We split the right supremum in two, one which involves different datasets of size m not being subject to CP, and one which involves the same dataset with and without CP:

$$\begin{aligned}&\mathbb{E}_{\mathcal{S},\mathcal{S}'\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{S}'}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right] \\ &= \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\left(\mathbb{E}_{\mathcal{S}'}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right]\right) + \left(\mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right)\right\}\right] \\ &\leq \underbrace{\mathbb{E}_{\mathcal{S},\mathcal{S}'\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{S}'}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right]}_{\doteq A} \\ &\quad + \underbrace{\mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right]}_{\doteq B} \quad . \end{aligned} \tag{9}$$

We handle A and B separately.

Upperbound on A. Handling A is achieved in the usual way [4]. Following the usual symmetrisation trick [4] (Theorem 8), and the fact [4] (Theorem 12.4) that φ is $1/b_\varphi$-Lipschitz [25] for some $b_\varphi > 0$, we obtain that with probability $\geq 1 - \delta_1$, we have:

$$\begin{aligned} & \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{S}'}\left[\varphi(yh(\boldsymbol{x}))\right]-\mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right] \\ \leq\ & \frac{4}{b_\varphi}\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_i \sigma_i h\left(\boldsymbol{x}_i\right)\right|\right]+K_\varphi\cdot\sqrt{\frac{2}{m}\log\frac{1}{\delta_1}}\ , \end{aligned} \tag{10}$$

$\forall \delta_1 > 0$.

Upperbound on B. This penalty appears when $\mathrm{M} \neq \mathrm{I}_m$. The trick is because φ is proper symmetric, there is a simple way to make appear Rademacher variables and a particular Rademacher complexity, which follows from the fact that [29]:

$$\mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right] \ = \ \frac{b_\varphi}{2m}\sum_{\sigma\in\Sigma_1}\sum_i \varphi(\sigma h(\boldsymbol{x}_i)) - \frac{\overline{h}\left(\mathcal{S}\right)}{2}\ , \tag{11}$$

where

$$\overline{h}\left(\mathcal{S}\right) \ \doteq \ \frac{1}{m}\cdot\sum_i y_i h(\boldsymbol{x}_i) \tag{12}$$

is the *h-mean-operator*, a statistics which can be proven to be minimally sufficient for classes given h [29]. Now, assumption **(i)** yields the invariance of $\overline{h}\left(\mathcal{S}\right)$ under permutation operation. This is proved in the following Lemma.

Lemma 12 *(mean-operator consistency of $\mathscr{T}$) Under the conditions of Theorem 4,*

$$\overline{h}\left(\mathcal{S}\right) \ = \ \overline{h}\left(\mathcal{S}^{\mathscr{T}}\right)\ ,\forall h,\forall\mathcal{S},\forall\mathscr{T}\ . \tag{13}$$

Proof To prove it, we let $\oplus$ denote the vector concatenation operation over the features of $\mathcal{F}_{\mathrm{a}}$ and $\mathcal{F}_{\mathrm{s}}$. We now first write using Assumption **(i)**:

$$\begin{aligned} m\cdot\overline{h}\left(\mathcal{S}\right) \ &= \ \sum_i y_i h(\boldsymbol{x}_i) \\ &= \ \sum_i y_i\cdot h((\mathrm{SF}^{\mathrm{a}})^\top\mathbf{1}_i\oplus(\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i) \\ &= \ \sum_i y_i\cdot h_{\mathrm{a}}((\mathrm{SF}^{\mathrm{a}})^\top\mathbf{1}_i)+\sum_i y_i\cdot h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)\ , \end{aligned} \tag{14}$$

and also, for the same reason,

$$m\cdot\overline{h}\left(\mathcal{S}^{\mathscr{T}}\right) \ = \ \sum_i y_i\cdot h_{\mathrm{a}}((\mathrm{SF}^{\mathrm{a}})^\top\mathbf{1}_i)+\sum_i y_i\cdot h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i)\ . \tag{15}$$

Therefore, we need to prove an equality that depends only upon the features of $\mathcal{F}_{\mathrm{s}}$:

$$\sum_i y_i \cdot h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i) \;=\; \sum_i y_i \cdot h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i) \;. \tag{16}$$

We have two cases to consider to prove eq. (16). Notice that since it is a block-class matrix, M admits the following block matrix decomposition:

$$\mathrm{M} \;=\; \left[\begin{array}{c|c} \mathrm{M}_+ & 0 \\ 0 & \mathrm{M}_- \end{array}\right] \;, \tag{17}$$

with $\mathrm{M}_b \in \mathbb{R}^{m_b \times m_b}$. We distinguish two cases.
Case 1 — Setting (A). In this case,

$$\begin{aligned}
\sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i) &= \sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}\left(\bigoplus_{k\in[d_{\mathrm{s}}]_{d_{\mathrm{a}}}} \sum_{l\in[m]} \mathrm{M}_{il}\mathrm{S}_{lk}\right) \\
&= \sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}\left(\sum_{l\in[m]} \mathrm{M}_{il} \bigoplus_{k\in[d_{\mathrm{s}}]_{d_{\mathrm{a}}}} \mathrm{S}_{lk}\right) \\
&= \sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}\left(\sum_{l\in[m]} \mathrm{M}_{il}(\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_l\right) \\
&= \sum_{i\in[m]} y_i \sum_{l\in[m]} \mathrm{M}_{il} \cdot h_{\mathrm{s}}\left((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_l\right) &(18)\\
&= \sum_{l\in[m]} \left(\sum_{i\in[m]} \mathrm{M}_{il}\right) y_l \cdot h_{\mathrm{s}}\left((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_l\right) &(19)\\
&= \sum_{l\in[m]} y_l \cdot h_{\mathrm{s}}\left((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_l\right) \;. &(20)
\end{aligned}$$

Here, $\oplus$ denotes the concatenation operator. Eq. (18) holds because of Setting (A), eq. (19) holds because of the class-consistency assumption (eq. (17)), and eq. (20) holds because $\mathrm{M} \in \mathcal{M}_m$.

Case 2 — Setting (B). We have $\mathrm{M} \in S^*_m$ (and no further assumption on h_{s}). In this case, letting

$\varsigma : [m] \to [m]$ represent the (block-class) permutation, we have:

$$\begin{aligned}
\sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i) &= \sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}\left(\bigoplus_{k\in[d_{\mathrm{s}}]_{d_{\mathrm{a}}}} \sum_{l\in[m]} \mathrm{M}_{il}\mathrm{S}_{lk}\right) \\
&= \sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}\left(\bigoplus_{k\in[d_{\mathrm{s}}]_{d_{\mathrm{a}}}} \mathrm{S}_{\varsigma(i)k}\right) \\
&= \sum_{i\in[m]} y_i \cdot h_{\mathrm{s}}\left((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_{\varsigma(i)}\right) \\
&= \sum_{i\in[m]} y_{\varsigma(i)} \cdot h_{\mathrm{s}}\left((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_{\varsigma(i)}\right) \qquad (21) \\
&= \sum_{l\in[m]} y_l \cdot h_{\mathrm{s}}\left((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_l\right) \ . \qquad (22)
\end{aligned}$$

Eq. (21) holds because M is a block-class matrix (eq. (17), and eq. (22) holds because ς is a permutation. This ends the proof of Lemma 12 ■

As a remark, when $h_{\mathrm{s}}(\boldsymbol{x}) = \boldsymbol{\theta}_{\mathrm{s}}^\top \boldsymbol{x}_{\mathrm{s}}$, with $\boldsymbol{\theta}_{\mathrm{s}} \in \mathbb{R}^{d_{\mathrm{s}}}$, eq. (13) shows the invariance of the mean operator

$$\mu_{\mathcal{S}\mathscr{T}} = \mu_{\mathcal{S}} \ \doteq \ \frac{1}{m} \cdot \sum_i y_i \boldsymbol{x}_i \ (\forall \mathscr{T}) \ , \qquad (23)$$

as minimal sufficient statistic for the classes [29]. Using eqs. (11) and (13) yield the first following identity ($\forall \mathcal{S}, \mathscr{T}, h$):

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right] - \mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right] \\
&= \frac{b_\varphi}{2m} \left\{ \begin{array}{c} \sum_{\sigma\in\Sigma_1} \sum_i \varphi(\sigma h((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i \oplus (\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i)) \\ - \\ \sum_{\sigma\in\Sigma_1} \sum_i \varphi(\sigma h((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i \oplus (\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i)) \end{array} \right\} \\
&= \frac{b_\varphi}{2m} \cdot \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m} \left[\begin{array}{c} \sum_i \varphi(\sigma_i h((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i \oplus (\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i)) \\ - \\ \sum_i \varphi(\sigma_i h((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i \oplus (\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i)) \end{array} \right] \qquad (24) \\
&\leq \frac{1}{2m} \cdot \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m} \left[\left| \sum_i \sigma_i h((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i \oplus (\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i) - \sigma_i h((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i \oplus (\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i) \right| \right] \qquad (25) \\
&= \frac{1}{2m} \cdot \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m} \left[\left| \sum_i \sigma_i \left\{ \begin{array}{c} (h_{\mathrm{a}}((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i) - h_{\mathrm{a}}((\mathrm{SF}^{\mathrm{a}})^\top \mathbf{1}_i)) \\ + \\ (h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i) - h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i)) \end{array} \right\} \right| \right] \qquad (26) \\
&= \frac{1}{2m} \cdot \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m} \left[\left| \sum_i \sigma_i (h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i) - h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i)) \right| \right] \ . \qquad (27)
\end{aligned}$$

Eq. (24) holds because $\boldsymbol{\sigma}$ is Rademacher. Ineq. (25) holds because F_φ is $(1/b_\varphi)$-Lipschitz [25]. Eq. (26) holds because of assumption **(i)**. We thus get the following upperbound for B in ineq. (9):

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{S}}\left[\varphi(yh(\boldsymbol{x}))\right]-\mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\right\}\right]\\
&\leq\ \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\sup_{h_{\mathrm{s}}}\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\left|\frac{1}{m}\sum_i\sigma_i(h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)-h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i))\right|\right]\right]\\
&\leq\ \mathbb{E}_{\mathcal{S}\sim\mathcal{D},\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h_{\mathrm{s}}}\left|\frac{1}{m}\sum_i\sigma_i(h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)-h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i))\right|\right]\ , \qquad (28)
\end{aligned}$$

where ineq. (28) holds because of the convexity of $\sup$. Using assumption **(ii)** ($h_{\mathrm{s}}(.)\in[0,K_{\mathrm{s}}]$), another use of the independent bounded differences inequality [22] yield with probability $\geq 1-\delta_2$:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}\sim\mathcal{D},\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h_{\mathrm{s}}}\left|\frac{1}{m}\sum_i\sigma_i(h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)-h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i))\right|\right]\\
&\leq\ \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h_{\mathrm{s}}}\left|\frac{1}{m}\sum_i\sigma_i(h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)-h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i))\right|\right]+K_{\mathrm{s}}\cdot\sqrt{\frac{2}{m}\log\frac{1}{\delta_2}}\ . \qquad (29)
\end{aligned}$$

We now put altogether ineqs. (8), (10) and (29) and obtain that with probability $\geq 1-(2\delta_1+\delta_2)$, we shall have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[L_{0/1}(y,h(\boldsymbol{x}))\right]\ \leq\ &\mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\\
&+\frac{4}{b_\varphi}\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_i\sigma_i h\left(\boldsymbol{x}_i\right)\right|\right]\\
&+\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h_{\mathrm{s}}}\left|\frac{1}{m}\sum_i\sigma_i(h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)-h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i))\right|\right]\\
&+2K_\varphi\cdot\sqrt{\frac{2}{m}\log\frac{1}{\delta_1}}+K_{\mathrm{s}}\cdot\sqrt{\frac{2}{m}\log\frac{1}{\delta_2}}\ . \qquad (30)
\end{aligned}$$

To simplify this expression, we fix $\delta_1=\delta_2=\delta/3$ and get with probability $\geq 1-\delta$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[L_{0/1}(y,h(\boldsymbol{x}))\right]\ \leq\ &\mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]\\
&+\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h_{\mathrm{s}}}\left|\frac{1}{m}\sum_i\sigma_i(h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i)-h_{\mathrm{s}}((\mathrm{MSF}^{\mathrm{s}})^\top\mathbf{1}_i))\right|\right]\\
&+\frac{4}{b_\varphi}\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_i\sigma_i h\left(\boldsymbol{x}_i\right)\right|\right]+(2K_\varphi+K_{\mathrm{s}})\cdot\sqrt{\frac{2}{m}\log\frac{3}{\delta}}\\
&=\mathbb{E}_{\mathcal{S}\mathscr{T}}\left[\varphi(yh(\boldsymbol{x}))\right]+\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})+\frac{4}{b_\varphi}\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_i\sigma_i h\left(\boldsymbol{x}_i\right)\right|\right]\\
&+(2K_\varphi+K_{\mathrm{s}})\cdot\sqrt{\frac{2}{m}\log\frac{3}{\delta}}\ ,
\end{aligned}$$

from which we obtain the statement of Theorem 4.

7.2.2 Example of domain and CP for which (min risk over $\mathcal{S}^{\mathscr{T}}$ + Rademacher CP complexity) is strictly smaller than (min risk over $\mathcal{S}$)

We exhibit a toy domain which shows that

$$\min_h \mathbb{E}_{\mathcal{S}^{\mathscr{T}}} \left[\varphi(yh(\boldsymbol{x}))\right] + \text{RCP}_{\mathscr{T}}(\mathcal{H}) \quad < \quad \min_h \mathbb{E}_{\mathcal{S}} \left[\varphi(yh(\boldsymbol{x}))\right] \quad , \tag{31}$$

for φ = square loss. Let $\mathcal{X} = \mathbb{R}^2$, with $\mathcal{S}$ consisting of 2 copies of observation $(0,0)$ (positive), 2 copies of observation $(1,1)$ (positive), and 1 copy of observation $(-1,-1)$ (negative). We enumerate the examples in $\mathcal{S}$ in this order. The CP satisfies $\text{F}^{\text{a}} \doteq \{x\}$, $\text{F}^{\text{s}} \doteq \{y\}$ and

$$\text{M} \quad \doteq \quad \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad ,$$

so that $\mathcal{S}^{\mathscr{T}}$ consists of two copies of observation $(1,0)$ (positive), two copies of $(0,1)$ (positive) and one copy of $(-1,-1)$ (the same observation as in $\mathcal{S}$). Let $h \doteq \boldsymbol{\theta}$, with its coordinates denoted x and y. The square loss L over $\mathcal{S}$ equals:

$$L \quad = \quad \frac{1}{5} \cdot \left(2 + 2(1-x-y)^2 + (1-x-y)^2\right) \quad , \tag{32}$$

which is minimized for $x = y = 1/2$ and yields $L = 2/5$. The square loss $L^{\mathscr{T}}$ over $\mathcal{S}^{\mathscr{T}}$ equals:

$$L^{\mathscr{T}} \quad = \quad \frac{1}{5} \cdot \left(2(1-x)^2 + 2(1-y)^2 + (1-x-y)^2\right) \quad , \tag{33}$$

which is minimized for $x = y = 3/4$ and yields $L^{\mathscr{T}} = 1/10$. Assuming all linear separators have ℓ_∞ norm bounded by $3/4$ (which allows to have both solutions above), the RCP is

$$\begin{aligned} \text{RCP}_{\mathscr{T}}(\mathcal{H}) \quad &\doteq \quad \mathbb{E}_{\boldsymbol{\sigma} \sim \Sigma_m} \left[\sup_{h \in \mathcal{H}_{\text{s}}} \left| \frac{1}{m} \sum_i \sigma_i \left(h((\text{SF}^{\text{s}})^\top \mathbf{1}_i) - h((\text{MSF}^{\text{s}})^\top \mathbf{1}_i) \right) \right| \right] \\ &= \quad \frac{1}{5} \cdot \frac{3}{4} \cdot \frac{1}{16} \cdot \sum_{\boldsymbol{\sigma} \in \Sigma_4} |-\sigma_1 - \sigma_2 + \sigma_3 + \sigma_4| \\ &= \quad \frac{3}{20} \cdot \frac{1}{16} \cdot \sum_{\boldsymbol{\sigma} \in \Sigma_4} \left| \mathbf{1}^\top \boldsymbol{\sigma} \right| \\ &= \quad \frac{3}{20} \cdot \frac{1}{16} \cdot (4 \cdot 2 + 2 \cdot 8 + 0 \cdot 6) \qquad (34) \\ &= \quad \frac{3}{20} \cdot \frac{24}{16} = \frac{9}{40} \quad . \qquad (35) \end{aligned}$$

We then check that

$$L^{\mathscr{T}} + \text{RCP}_{\mathscr{T}}(\mathcal{H}) = \frac{13}{40} \quad < \quad \frac{2}{5} = L \quad , \tag{36}$$

as claimed.

7.2.3 Proof of Theorem 6

The proof of Theorem 6 follows from the proof of a more general Theorem that we prove here. We say that $(\mathrm{M}, \mathrm{K}^{\mathrm{s}})$ satisfies the (γ, δ)-correlation assumption for some $0 < \delta, \gamma \leq 1$ iff the following two assumptions hold:

(a) $((\mathrm{I}_m - \mathrm{M})\mathrm{K}^{\mathrm{s}}(\mathrm{I}_m - \mathrm{M})^\top)_{ii} \geq (1-\delta) \cdot (1/m) \cdot \mathrm{tr}\left((\mathrm{I}_m - \mathrm{M})\mathrm{K}^{\mathrm{s}}(\mathrm{I}_m - \mathrm{M})^\top\right), \forall i \in [m]$;

(b) $|((\mathrm{I}_m - \mathrm{M})\mathrm{K}^{\mathrm{s}}(\mathrm{I}_m - \mathrm{M})^\top)_{ii'} / \sqrt{((\mathrm{I}_m - \mathrm{M})\mathrm{K}^{\mathrm{s}}(\mathrm{I}_m - \mathrm{M})^\top)_{ii}((\mathrm{I}_m - \mathrm{M})\mathrm{K}^{\mathrm{s}}(\mathrm{I}_m - \mathrm{M})^\top)_{i'i'}}| \geq 1 - \gamma, \forall i, i' \in [m]$.

Regardless of $\mathcal{S}$, there always exist $0 < \delta, \gamma \leq 1$ for which this holds, but the bound may be quantitatively better when at least one is small.

Theorem 13 *Using notations of Theorem 6, there exists $\epsilon > 0$ such that for any $\mathcal{S}$ and $\mathscr{T}$ for which $(\mathrm{M}, \mathrm{K}^{\mathrm{s}})$ satisfies the (γ, δ)-correlation assumption, we have*

$$\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) \quad \leq \quad u \cdot \frac{r_{\mathrm{s}}}{\sqrt{m}} \cdot \sqrt{\frac{1}{m} \cdot \langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}} \rangle_{\mathrm{M}} \mathrm{I}_m} \ . \tag{37}$$

with

$$u \quad \doteq \quad \frac{1}{m} + \kappa(\epsilon)\left(1 - \frac{1}{m}\right) \ , \tag{38}$$

and $\kappa(\epsilon) = 1 - ((1-\delta)(1-\epsilon)(1-\gamma))^2 \in (0,1)$.

Furthermore,

$$\frac{1}{m} \cdot \langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}} \rangle_{\mathrm{M}} \quad = \quad 2 \cdot \sum_{j \in [d_{\mathrm{s}}]} \mathbb{V}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j)(1 - \rho(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j, \mathrm{MSF}^{\mathrm{s}} \mathbf{1}_j)) \ . \tag{39}$$

Proof We observe that $\forall \boldsymbol{\sigma} \in \Sigma_m$,

$$\begin{aligned} &\arg \sup_{\boldsymbol{\theta} \in \mathbb{R}^{d_{\mathrm{s}}} : \|\boldsymbol{\theta}\|_2 \leq r_{\mathrm{s}}} \left| \frac{1}{m} \sum_i \sigma_i \boldsymbol{\theta}^\top \left(((\mathrm{I}_m - \mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i \right) \right| \\ &= \quad \frac{r_{\mathrm{s}}}{\| \sum_i \sigma_i ((\mathrm{I}_m - \mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i \|_2} \sum_i \sigma_i ((\mathrm{I}_m - \mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i \ , \end{aligned}$$

and so:

$$\begin{aligned}
\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h\in\mathcal{H}_{\mathrm{s}}}\left|\frac{1}{m}\sum_i \sigma_i\left(h((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i) - h((\mathrm{MSF}^{\mathrm{s}})^\top \mathbf{1}_i)\right)\right|\right] \qquad (40)\\
&= \frac{r_{\mathrm{s}}}{m}\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\left\|\sum_i \sigma_i((\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i\right\|_2\right]\\
&= \frac{r_{\mathrm{s}}}{m}\cdot\sqrt{\sum_i\|((\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i\|_2^2}\\
&\quad\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sqrt{1+\frac{\sum_{i\neq i'}\sigma_i\sigma_{i'}\mathbf{1}_i^\top(\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}}(\mathrm{SF}^{\mathrm{s}})^\top(\mathrm{I}_m-\mathrm{M})^\top\mathbf{1}_{i'}}{\sum_i\|((\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i\|_2^2}}\right]\\
&\doteq \frac{r_{\mathrm{s}}}{m}\cdot\sqrt{\sum_i\|((\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i\|_2^2}\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sqrt{1+u(\boldsymbol{\sigma})}\right]\ , \qquad (41)
\end{aligned}$$

with

$$u(\boldsymbol{\sigma}) \doteq \frac{\sum_{i\neq i'}\sigma_i\sigma_{i'}\mathbf{1}_i^\top(\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}}(\mathrm{SF}^{\mathrm{s}})^\top(\mathrm{I}_m-\mathrm{M})^\top\mathbf{1}_{i'}}{\sum_i\|((\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i\|_2^2}\ . \qquad (42)$$

Let us call for short $\boldsymbol{\delta}_i \doteq ((\mathrm{I}_m-\mathrm{M})\mathrm{SF}^{\mathrm{s}})^\top\mathbf{1}_i$, so that eq. (42) can be simplified to $u(\boldsymbol{\sigma}) = (\sum_i\|\boldsymbol{\delta}_i\|_2^2)^{-1}\sum_{i\neq i'}\sigma_i\sigma_{i'}\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_{i'}$. $\forall n\in\mathbb{N}_*$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[u^n(\boldsymbol{\sigma})\right]\\
&= \frac{1}{\left(\sum_i\|\boldsymbol{\delta}_i\|_2^2\right)^n}\cdot\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sum_{i_1\neq i_1'}\sum_{i_2\neq i_2'}\cdots\sum_{i_n\neq i_n'}\prod_{k=1}^n\sigma_{i_k}\sigma_{i_k'}\boldsymbol{\delta}_{i_k}^\top\boldsymbol{\delta}_{i_k'}\right]\\
&= \frac{1}{\left(\sum_i\|\boldsymbol{\delta}_i\|_2^2\right)^n}\cdot\sum_{i_1\neq i_1'}\sum_{i_2\neq i_2'}\cdots\sum_{i_n\neq i_n'}\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\prod_{k=1}^n\sigma_{i_k}\sigma_{i_k'}\boldsymbol{\delta}_{i_k}^\top\boldsymbol{\delta}_{i_k'}\right]\\
&= \frac{1}{\left(\sum_i\|\boldsymbol{\delta}_i\|_2^2\right)^n}\cdot\sum_{i_1\neq i_1'}\sum_{i_2\neq i_2'}\cdots\sum_{i_n\neq i_n'}\prod_{(i,i')\in\{(i_k,i_k')\}_{k=1}^n}\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[(\sigma_i\sigma_{i'})^{n(i,i')}\right]\left(\boldsymbol{\delta}_i^\top\boldsymbol{\delta}_{i'}\right)^{n(i,i')} \qquad (43)
\end{aligned}$$

with $n(i,i') \doteq |\{k : (i,i') = (i_k,i_k')\}|$ satisfying $\sum n(i,i') = n$. Whenever $n(i,i')$ is odd, $\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[(\sigma_i\sigma_{i'})^{n(i,i')}\right] = 0$ (because $\boldsymbol{\sigma}$ is Rademacher), and it is 1 otherwise. We get, if n is even:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[u^n(\boldsymbol{\sigma})\right]\\
&= \frac{1}{\left(\sum_i\|\boldsymbol{\delta}_i\|_2^2\right)^n}\cdot\underbrace{\sum_{0<\ell\leq n}\ \sum_{\substack{\{n_k\}_{k=1}^\ell\subset\mathbb{N}_*\\ \text{s.t. } 2\sum n_k = n}}\ \sum_{\substack{\{(i_k,i_k')\}_{k=1}^\ell\\ \text{s.t. } i_k\neq i_k',\forall k}}\prod_{k=1}^\ell\left(\boldsymbol{\delta}_{i_k}^\top\boldsymbol{\delta}_{i_k'}\right)^{2n_k}}_{\doteq\zeta(n)}\ , \qquad (44)
\end{aligned}$$

and $\mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[u^n(\boldsymbol{\sigma})\right] = 0$ if n is odd. Since

$$\sqrt{1+x} = 1+\sum_{n\in\mathbb{N}_*}\frac{1}{2^n n!}\cdot\prod_{k=0}^{n-1}(1-2k)x^n\ , \qquad (45)$$

we get after combining with eqs (41) and (44) and using the definition of $\zeta(.)$ in eq. (44):

$$\begin{aligned}\text{RCP}_{\mathscr{T}}(\mathcal{H}) &= \frac{r_{\text{s}}}{m}\cdot\sqrt{\sum_i \|\boldsymbol{\delta}_i\|_2^2}\cdot\left(1-\sum_{n\in\mathbb{N}_*}\frac{\prod_{k=1}^{2n-1}(2k-1)\cdot\zeta(2n)}{(2n)!\left(2\sum_i\|\boldsymbol{\delta}_i\|_2^2\right)^{2n}}\right)\\ &= \frac{r_{\text{s}}}{m}\cdot\sqrt{\sum_i \|\boldsymbol{\delta}_i\|_2^2}\cdot\left(1-\sum_{n\in\mathbb{N}_*}\frac{\prod_{k=1}^{2n-1}(2k-1)}{(2m)^{2n}(2n)!}\cdot\tilde{\zeta}(2n)\right)\ , \end{aligned} \tag{46}$$

with:

$$\begin{aligned}&\tilde{\zeta}(2n)\\ &\doteq \sum_{0<\ell\leq 2n}\ \sum_{\substack{\{n_k\}_{k=1}^{\ell}\subset\mathbb{N}_*\\ \text{s.t. }\sum n_k=n}}\ \sum_{\substack{\{(i_k,i'_k)\}_{k=1}^{\ell}\\ \text{s.t. } i_k\neq i'_k,\forall k}}\ \prod_{k=1}^{\ell}\left(\begin{array}{c}\frac{\left\|\boldsymbol{\delta}_{i_k}\right\|_2^2}{\frac{1}{m}\cdot\sum_p\|\boldsymbol{\delta}_p\|_2^2}\\ \cdot\frac{\left\|\boldsymbol{\delta}_{i'_k}\right\|_2^2}{\frac{1}{m}\cdot\sum_p\|\boldsymbol{\delta}_p\|_2^2}\cdot\gamma_{i_k,i'_k}\end{array}\right)^{2n_k}\ , \end{aligned} \tag{47}$$

and $\gamma_{i,i'}\doteq\cos(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{i'})$. Remark that eq. (46) is an equality. We now use assumption **(a)** and obtain

$$\tilde{\zeta}(2n) \ \geq\ (1-\delta)^{2n}\sum_{0<\ell\leq 2n}\ \sum_{\substack{\{n_k\}_{k=1}^{\ell}\subset\mathbb{N}_*\\ \text{s.t. }\sum n_k=n}}\ \sum_{\substack{\{(i_k,i'_k)\}_{k=1}^{\ell}\\ \text{s.t. } i_k\neq i'_k,\forall k}}\ \prod_{k=1}^{\ell}\left(\gamma_{i_k,i'_k}\right)^{2n_k}\ . \tag{48}$$

Denote for short $\mathcal{U}(n)$ the set of eligible triples (n_k,i_k,i'_k) in the summation. We get because of assumption **(b)** $\tilde{\zeta}(2n)\geq|\mathcal{U}(n)|((1-\delta)(1-\gamma))^{2n}$, and so, using the shorthand

$$\mathscr{L}\ \doteq\ \frac{1}{m}\cdot\sum_i\|\boldsymbol{\delta}_i\|_2^2\ , \tag{49}$$

we obtain our first upperbound,

$$\text{RCP}_{\mathscr{T}}(\mathcal{H})\ \leq\ \frac{r_{\text{s}}}{\sqrt{m}}\cdot\sqrt{\mathscr{L}}\cdot\left(1-\sum_{n\in\mathbb{N}_*}|\mathcal{U}(n)|\left(\frac{(1-\delta)(1-\gamma)}{m}\right)^{2n}\cdot\frac{\prod_{k=1}^{2n-1}(2k-1)}{2^{2n}(2n)!}\right) \tag{50}$$

Lemma 14 *There exists a constant $\epsilon>0$ such that $\forall n\in\mathbb{N}_*$,*

$$\frac{\prod_{k=1}^{2n-1}(2k-1)}{(2n)!}\ \geq\ (2(1-\epsilon))^{2n}\ . \tag{51}$$

Proof We proceed by induction, letting $g_\epsilon(n)\doteq(2(1-\epsilon))^{2n}$ and

$$f(n)\ \doteq\ \frac{\prod_{k=1}^{2n-1}(2k-1)}{(2n)!}\ , \tag{52}$$

and for $\epsilon = \epsilon_* \doteq 1 - 1/(2\sqrt{2})$. We remark that $g_{\epsilon_*}(1) = f(1)$. Furthermore,

$$\begin{aligned} f(n+1) = \frac{\prod_{k=1}^{2(n+1)-1}(2k-1)}{(2(n+1))!} &= \frac{(2n-1)(2n+1)}{(n+1)(n+2)} \cdot \frac{\prod_{k=1}^{2n-1}(2k-1)}{(2n)!} \\ &\geq \frac{(2n-1)(2n+1)}{(n+1)(n+2)} \cdot (2(1-\epsilon))^{2n} \qquad (53) \\ &= \frac{(2n-1)(2n+1)}{(n+1)(n+2)(2(1-\epsilon))^2} \cdot g_\epsilon(n+1) \ , \end{aligned}$$

where ineq. (53) uses the induction hypothesis. We prove the Lemma once we prove that the factor on the right is at least 1, that is, for $\epsilon = \epsilon_*$, we need to prove

$$\frac{(2n-1)(2n+1)}{(n+1)(n+2)} \geq \frac{1}{2} \ , \qquad (54)$$

which is indeed the case since the left function is strictly increasing over $\mathbb{N}$ and equals the right-hand side for $n = 1$. ■

So we get from ineq. (50):

$$\begin{aligned} \text{RCP}_{\mathscr{T}}(\mathcal{H}) &\leq \frac{r_{\text{s}}}{\sqrt{m}} \cdot \sqrt{\mathscr{L}} \cdot \left(1 - \sum_{n \in \mathbb{N}_*} |\mathcal{U}(n)| \left(\frac{(1-\delta)(1-\gamma)(1-\epsilon)}{m}\right)^{2n}\right) \\ &\leq \frac{r_{\text{s}}}{\sqrt{m}} \cdot \sqrt{\mathscr{L}} \cdot \left(1 - u_*(m) \sum_{n \in \mathbb{N}_*} \left(\frac{(1-\delta)(1-\gamma)(1-\epsilon)}{m}\right)^{2n}\right) \ , \qquad (55) \end{aligned}$$

where $u_*(m)$ satisfies $u_*(m) \leq \min_n |\mathcal{U}(n)|$. We finally get:

$$\begin{aligned} \text{RCP}_{\mathscr{T}}(\mathcal{H}) &\leq \frac{r_{\text{s}}}{\sqrt{m}} \cdot \sqrt{\mathscr{L}} \cdot \left(1 - (1-\kappa(\epsilon)) \cdot \frac{u_*(m)}{m^2 - (1-\kappa(\epsilon))}\right) \\ &\leq \frac{r_{\text{s}}}{\sqrt{m}} \cdot \sqrt{\mathscr{L}} \cdot \left(1 - (1-\kappa(\epsilon)) \cdot \frac{u_*(m)}{m^2}\right) \\ &\leq \frac{r_{\text{s}}}{\sqrt{m}} \cdot \sqrt{\mathscr{L}} \cdot \left(\frac{1}{m} + \kappa(\epsilon)\left(1 - \frac{1}{m}\right)\right) \ , \qquad (56) \end{aligned}$$

with $\kappa(\epsilon) = 1 - ((1-\delta)(1-\epsilon)(1-\gamma))^2 > 0$, since $u_*(m) \geq m(m-1)$ (obtained for $\ell = 2n$ in eq. (48)). We finish the proof by remarking that $\mathscr{L}$ in eq. (49) satisfies

$$\begin{aligned} \mathscr{L} &= \frac{1}{m} \cdot \sum_i \mathbf{1}_i^\top (\mathrm{I}_m - \mathrm{M}) \mathrm{S}\mathrm{F}^{\text{s}} (\mathrm{S}\mathrm{F}^{\text{s}})^\top (\mathrm{I}_m - \mathrm{M})^\top \mathbf{1}_i \\ &= \frac{1}{m} \cdot \mathrm{tr}\left((\mathrm{I}_m - \mathrm{M}) \mathrm{S}\mathrm{F}^{\text{s}} (\mathrm{S}\mathrm{F}^{\text{s}})^\top (\mathrm{I}_m - \mathrm{M})^\top\right) \\ &= \frac{1}{m} \cdot \mathrm{tr}\left((\mathrm{I}_m - \mathrm{M}) \mathrm{K}^{\text{s}} (\mathrm{I}_m - \mathrm{M})^\top\right) \\ &= \frac{1}{m} \cdot \mathrm{tr}\left((\mathrm{I}_m - \mathrm{M})^\top \mathrm{I}_m (\mathrm{I}_m - \mathrm{M}) \mathrm{K}^{\text{s}}\right) \\ &= \frac{1}{m} \cdot \langle \mathrm{I}_m, \mathrm{K}^{\text{s}} \rangle_{\mathrm{M}} \ . \qquad (57) \end{aligned}$$

Ineq. (56) and eq. (57) allow to conclude the proof of Theorem 13 with

$$u \doteq \frac{1}{m} + \kappa(\epsilon)\left(1 - \frac{1}{m}\right) , \tag{58}$$

which, since $\kappa(\epsilon) < 1$, satisfies indeed $u \in (0,1)$. This achieves the main part of the proof of Theorem 13. To prove eq. (39), we just have to write (letting $\varsigma : [m] \to [m]$ represent the corresponding permutation),

$$\begin{aligned}
&\frac{1}{m} \cdot \langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}} \rangle_{\mathbf{M}} \\
&= \frac{1}{m} \sum_i \left\| \boldsymbol{x}_i^{\mathrm{s}} - \boldsymbol{x}_{\varsigma(i)}^{\mathrm{s}} \right\|_2^2 \\
&= \sum_{j \in [d_{\mathrm{s}}]} \frac{1}{m} \cdot \sum_i (x_{ij}^{\mathrm{s}} - x_{\varsigma(i)j}^{\mathrm{s}})^2 \\
&= 2 \cdot \sum_{j \in [d_{\mathrm{s}}]} \left\{ \frac{1}{m} \cdot \sum_i (x_{ij}^{\mathrm{s}})^2 - \frac{1}{m} \cdot \sum_i x_{ij}^{\mathrm{s}} x_{\varsigma(i)j}^{\mathrm{s}} \right\} \qquad (59) \\
&= 2 \cdot \sum_{j \in [d_{\mathrm{s}}]} \left\{ \frac{1}{m} \cdot \sum_i (x_{ij}^{\mathrm{s}})^2 - \left(\frac{1}{m} \cdot \sum_i x_{ij}^{\mathrm{s}} \right)^2 + \left(\frac{1}{m} \cdot \sum_i x_{ij}^{\mathrm{s}} \right)^2 - \frac{1}{m} \cdot \sum_i x_{ij}^{\mathrm{s}} x_{\varsigma(i)j}^{\mathrm{s}} \right\} \\
&= 2 \cdot \sum_{j \in [d_{\mathrm{s}}]} \{ \mathbb{V}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j) - \mathbb{C}\mathrm{ov}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j, \mathrm{MSF}^{\mathrm{s}} \mathbf{1}_j) \} \\
&= 2 \cdot \sum_{j \in [d_{\mathrm{s}}]} \mathbb{V}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j) \left\{ 1 - \frac{\mathbb{C}\mathrm{ov}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j, \mathrm{MSF}^{\mathrm{s}} \mathbf{1}_j)}{\sqrt{\mathbb{V}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j)} \sqrt{\mathbb{V}(\mathrm{MSF}^{\mathrm{s}} \mathbf{1}_j)}} \right\} \qquad (60) \\
&= 2 \cdot \sum_{j \in [d_{\mathrm{s}}]} \mathbb{V}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j)(1 - \rho(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j, \mathrm{MSF}^{\mathrm{s}} \mathbf{1}_j)) ,
\end{aligned}$$

where eq. (60) follows from the fact that $\mathbb{V}(\mathrm{SF}^{\mathrm{s}} \mathbf{1}_j) = \mathbb{V}(\mathrm{MSF}^{\mathrm{s}} \mathbf{1}_j)$. ■

Remark: it is worthwhile remarking that the proof of Theorem 6 can also be applied to upperbound the empirical Rademacher complexity of linear functions, without modifications, except for the handling of $\mathscr{L}$. In this case, the proof improves the upperbound known [18] (Theorem 1) by factor u in eq. (58). This is due to the fact that the proof in [18] takes into account only the maximum norm in the observations of $\mathcal{S}$, and not the angles between the observations. We now state the corresponding Theorem.

Theorem 15 *Following [15], we let $\mathbf{i}_r^m$ denote the set of r-tuples drawn without replacement drawn from $[m]$. Suppose $\mathcal{S}$ satisfies the following for some $\delta, \gamma > 0$ and $r_x > 0$:*

(a) $\|\boldsymbol{x}_i\|_2^2 \geq (1-\delta) \cdot (1/m) \sum_{i'} \|\boldsymbol{x}_{i'}\|_2^2, \forall i \in [m]$*;*

(b) $\mathbb{E}_{(i,i') \sim \mathbf{i}_2^m}[|\cos(\boldsymbol{x}_i, \boldsymbol{x}_{i'})|] \geq 1 - \gamma$*;*

(c) $\|\boldsymbol{x}_i\|_2 \leq r_x, \forall i \in [m]$.

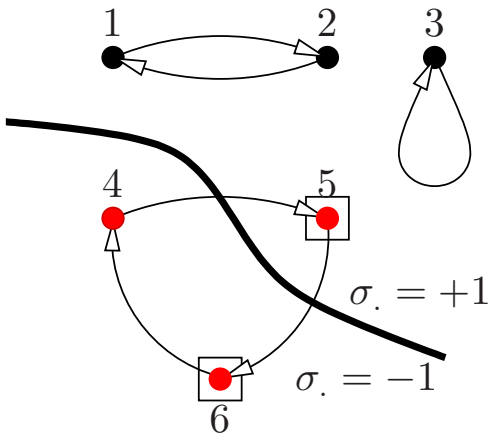


Figure 2: A permutation ς defines an oriented graph whose vertices are examples and permutation ς defines arcs (here, $\varsigma(1) = 2$ for example; black dots are positive examples, red dots are negative examples). Each term in the $\sup$ of eq. (62) is a weighted cut (one weighted cut for each $\boldsymbol{\sigma} \in \{-1, 1\}^m$): the squares depict the examples for which $w(.) \neq 0$ in Lemma 16 for the $\boldsymbol{\sigma}$ displayed. In this example of $\boldsymbol{\sigma}$, only two examples out of the six would bring a non-zero weight $w(.)$.

Then, assuming that $\mathcal{H}$ contains linear classifiers of the form $\boldsymbol{\theta}^\top \boldsymbol{x}$ with $\|\boldsymbol{\theta}\|_2 \leq r_\theta$, there exists $\epsilon > 0$ such that the empirical Rademacher complexity of $\mathcal{H}$ satisfies:

$$\mathrm{R}_{\mathcal{S}}(\mathcal{H}) \quad \leq \quad \left(\frac{1}{m} + \kappa(\epsilon)\left(1 - \frac{1}{m}\right)\right) \cdot \frac{r_x r_\theta}{\sqrt{m}} \ , \tag{61}$$

with $\kappa(\epsilon) = 1 - ((1-\delta)(1-\epsilon)(1-\gamma))^2 \in (0, 1)$.

[18]'s proof relies on **(c)**. Since **(a)** and **(b)** can always be satisfied for some $\delta, \gamma > 0$, ineq. (37) holds under their setting as well; however, it becomes better than theirs as both δ, γ are small, so in particular in the case where observations start to be heavily correlated and be of approximately the same norm. Indeed, in this case, $\sum_i \sigma_i \boldsymbol{x}_i$ will often have *small* magnitude, because $\Sigma_m \ni \boldsymbol{\sigma} \sim \{-1, 1\}$ and thus many vectors will approximately cancel through the sum in many draws of $\boldsymbol{\sigma}$.

7.2.4 Proof of Theorem 7

We first start by a Lemma which shows that indeed $\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})$ can be significantly smaller than a Rademacher complexity.

Lemma 16 *Suppose setting (B) holds in Theorem 4. Then*

$$\mathrm{RCP}_{\mathscr{T}}(\mathcal{H}) \quad = \quad 2 \cdot \mathbb{E}_{\boldsymbol{\sigma} \sim \Sigma_m}\left[\sup_{h_{\mathrm{s}} \in \mathcal{H}_{\mathrm{s}}} \left| \frac{1}{m} \sum_i w(i) h_{\mathrm{s}}((\mathrm{SF}^{\mathrm{s}})^\top \mathbf{1}_i) \right| \right] \ , \tag{62}$$

where $w(i) \doteq (1/2) \cdot (\sigma_i - \sigma_{\varsigma^{-1}(i)}) \in \{-1, 0, 1\}$.

The proof of this Lemma is straightforward. What is interesting is, since $w(.)$ can take on zero values, to what extent $\textsc{rcp}_{\mathscr{T}}(\mathcal{H})$ can be smaller than the corresponding Rademacher complexity in which $w(.)$ would be replaced by $\boldsymbol{\sigma} \in \{-1,1\}^m$, and what drives this reduction. Figure 2 displays qualitatively this intuition on a simple example. We now investigate a quantitative derivation of the reduction for DAG classifiers.

We let $\mathrm{cut}_\varsigma(\boldsymbol{\sigma}) \doteq \{i : \sigma_i \neq \sigma_{\varsigma^{-1}(i)}\}$. We simplify notations in the proof and drop notation b so that notation $h^i_{\rm s} \doteq h_{\rm s}((\mathrm{SF}^{\rm s})^\top \mathbf{1}_i)$ where $i \in [m]$. We let

$$\mu \doteq \mathbb{E}_{\boldsymbol{\sigma}\sim\Sigma_m}\left[\sup_{h_{\rm s}\in\mathcal{H}^{\rm s}} \sum_{i\in\mathrm{cut}_\varsigma(\boldsymbol{\sigma})} \sigma_i h^i_{\rm s}\right] \ , \tag{63}$$

which is a generalisation of $m \cdot \textsc{rcp}_{\mathscr{T}}(\mathcal{H})$ to any permutation $\varsigma \in S_m$, and not just a block-class permutation in S^*_m as assumed in Setting (B). We assume basic knowledge of Massart's finite class Lemma's proof. Using Jensen's inequality, we arrive, after the same chain of derivations, for any $t > 0$, to:

$$\exp(t\mu) \ \leq \ \frac{1}{2^m} \sum_{\boldsymbol{\sigma}\in\{-1,1\}^m} \sup_{h_{\rm s}\in\mathcal{H}^{\rm s}} \exp\left(t \cdot \sum_{i\in\mathrm{cut}_\varsigma(\boldsymbol{\sigma})} \sigma_i h^i_{\rm s}\right) \ . \tag{64}$$

The proof (of Massart's Lemma) now involves replacing the $\sup$ by a sum. Remark that when h is DAG, the $\sup$ implies that each $h^j_{\rm s}$ is in fact $\in \{\pm K_{\rm s}\}$. So let us use $\mathcal{H}^{\rm s}_+ \subseteq \mathcal{H}^{\rm s}$, the set of classifiers whose output is in $\{\pm K_{\rm s}\}$. We get:

$$\exp(t\mu) \ \leq \ \sum_{h_{\rm s}\in\mathcal{H}^{\rm s}_+} \frac{1}{2^m} \sum_{\boldsymbol{\sigma}\in\{-1,1\}^m} \exp\left(t \cdot \sum_{i\in\mathrm{cut}_\varsigma(\boldsymbol{\sigma})} \sigma_i h^i_{\rm s}\right) \ . \tag{65}$$

To identify better permutations, we name in this proof $\varsigma \in S^*_m$ the permutation represented by M, so that we also have

$$\mathrm{odd\_cycle}(\mathrm{M}) \ \doteq \ \mathrm{odd\_cycle}(\varsigma) \ .$$

Since each coordinate of $\boldsymbol{\sigma}$ is chosen uniformly at random, cycles in a permutation are disjoint and $\exp(a+b) = \exp(a)\exp(b)$, the inner sigma in ineq. (65) factors over the cycles of permutation ς:

$$\begin{aligned}
&\sum_{h_{\rm s}\in\mathcal{H}^{\rm s}_+} \frac{1}{2^m} \sum_{\boldsymbol{\sigma}\in\{-1,1\}^m} \exp\left(t \cdot \sum_{i\in\mathrm{cut}_\varsigma(\boldsymbol{\sigma})} \sigma_i h^i_{\rm s}\right) \\
&= \ \frac{1}{2^m} \sum_{h_{\rm s}\in\mathcal{H}^{\rm s}_+} \prod_{U\in\mathrm{cycle}(\varsigma)} \left(\sum_{\boldsymbol{\sigma}\in\{-1,1\}^{|U|}} \exp\left(t \cdot \sum_{i\in\mathrm{cut}_\varsigma(\boldsymbol{\sigma})\cap U} \sigma_i h^U_i\right)\right) \ ,
\end{aligned} \tag{66}$$

where h^U indicates coordinates of h in U and $\mathrm{cycle}(\varsigma)$ is the set of cycles *without 1-cycles* (*i.e.* fixed points) — we have let m_ς denote the number of fixed points of ς. We have used the fact that cycles define a partition of $[m]$.

Now, whenever U contains an *odd* number of indexes, whatever $\boldsymbol{\sigma}$, the sum over $\mathrm{cut}_\varsigma(\boldsymbol{\sigma}) \cap U$ cannot cover the sum over U: there always remains at least one vertex which does not belong to the sum (In Figure 2, the red dot cycle displays this fact on an example). Let us denote $i_{\boldsymbol{\sigma}} \in [|U|]$ this vertex. Since $\exp(x) + \exp(-x) \geq 2 + x^2$ and $h_{\mathrm{s}} \in \mathcal{H}^{\mathrm{s}}_+$, we get:

$$\begin{aligned}
&\sum_{\boldsymbol{\sigma} \in \{-1,1\}^{|U|}} \exp\left(t \cdot \sum_{i \in \mathrm{cut}_\varsigma(\boldsymbol{\sigma}) \cap U} \sigma_i h_i^U \right) \\
&\leq \frac{1}{2 + t^2 K_{\mathrm{s}}^2} \cdot \sum_{\boldsymbol{\sigma} \in \{-1,1\}^{|U|}} \left\{ \exp\left(t \cdot \sum_{i \in \mathrm{cut}_\varsigma(\boldsymbol{\sigma}) \cap U} \sigma_i h_i^U \right) \cdot \left(\exp(t h_{i_{\boldsymbol{\sigma}}}^U) + \exp(-t h_{i_{\boldsymbol{\sigma}}}^U) \right) \right\} \quad (67) \\
&\leq \frac{2}{2 + t^2 K_{\mathrm{s}}^2} \cdot \sum_{\boldsymbol{\sigma} \in \{-1,1\}^{|U|}} \exp\left(t \cdot \sum_{i \in U} \sigma_i h_i^U \right) \ , \quad (68)
\end{aligned}$$

since the multiplication in ineq. (67) duplicates part of the terms in (68). We now plug this bound in eq. (65 — 66) and finish the derivation following Massart's finite class Lemma:

$$\begin{aligned}
\exp(t\mu) &\leq \left(\frac{2}{2 + t^2 K_{\mathrm{s}}^2} \right)^{|\mathrm{odd_cycle}(\varsigma)|} \cdot \frac{1}{2^m} \sum_{h_{\mathrm{s}} \in \mathcal{H}^{\mathrm{s}}_+} \sum_{\boldsymbol{\sigma} \in \{-1,1\}^m} \exp\left(t \cdot \sum_i \sigma_i h_i \right) \\
&= \left(\frac{2}{2 + t^2 K_{\mathrm{s}}^2} \right)^{|\mathrm{odd_cycle}(\varsigma)|} \cdot \sum_{h_{\mathrm{s}} \in \mathcal{H}^{\mathrm{s}}_+} \prod_i \left(\frac{\exp(t h_{\mathrm{s}}^i) + \exp(-t h_{\mathrm{s}}^i)}{2} \right) \\
&\leq \left(\frac{2}{2 + t^2 K_{\mathrm{s}}^2} \right)^{|\mathrm{odd_cycle}(\varsigma)|} \cdot |\mathcal{H}^{\mathrm{s}}_+| \exp\left(\frac{t^2}{2} \cdot \sum_i \left(\max_{h \in \mathcal{H}^{\mathrm{s}}_+} h_{\mathrm{s}}^i \right)^2 \right) \quad (69) \\
&\leq \left(\frac{2}{2 + t^2 K_{\mathrm{s}}^2} \right)^{|\mathrm{odd_cycle}(\varsigma)|} \cdot |\mathcal{H}^{\mathrm{s}}_+| \exp\left(\frac{t^2 m K_{\mathrm{s}}^2}{2} \right) \ .
\end{aligned}$$

Ineq. (69) holds because $(\exp(x) + \exp(-x))/2 \leq \exp(x^2/2)$. Taking logs and rearranging yields:

$$\mu \leq \frac{1}{t} \cdot \log \frac{|\mathcal{H}^{\mathrm{s}}_+|}{\left(1 + \frac{t^2 K_{\mathrm{s}}^2}{2}\right)^{|\mathrm{odd\_cycle}(\varsigma)|}} + \frac{t m K_{\mathrm{s}}^2}{2} \ . \quad (70)$$

Now, suppose that we can choose t such that

$$\frac{t^2 K_{\mathrm{s}}^2}{2} \geq \varepsilon \ , \quad (71)$$

for some $\varepsilon > 0$. In this case, ineq. (70) implies

$$\mu \leq \frac{1}{t} \cdot \log \frac{|\mathcal{H}^{\mathrm{s}}_+|}{(1+\varepsilon)^{|\mathrm{odd\_cycle}(\varsigma)|}} + \frac{t m K_{\mathrm{s}}^2}{2} \ , \quad (72)$$

which is of the form $\mu \leq A/t + Bt$ with $B > 0$. Taking $t = \sqrt{A/B}$ yields, using the fact that each classifier in $\mathcal{H}^{\mathrm{s}}_+$:

$$\mu \leq K_{\mathrm{s}} \cdot \sqrt{2m \log \frac{|\mathcal{H}^{\mathrm{s}}_+|}{(1+\varepsilon)^{|\mathrm{odd\_cycle}(\varsigma)|}}} \ . \quad (73)$$

Dividing by m gives the statement of Theorem 7. We need however to check that ineq. (71) holds, which, since $t = \sqrt{A/B}$, yields that we must have, after simplification:

$$\log|\mathcal{H}^{\mathrm{s}}_{+}| \;\;\geq\;\; \varepsilon m + |\mathrm{odd\_cycle}(\varsigma)|\log(1+\varepsilon)\;. \tag{74}$$

Since we have excluded fixed points, $|\mathrm{odd\_cycle}(\varsigma)| \leq m/3$, and since $\log(1+x) \leq x$, a sufficient condition is $\log|\mathcal{H}^{\mathrm{s}}_{+}| \geq 4\varepsilon m/3$, which is the Theorem's assumption.

We now show a bound on the expected RCP when M in $\mathscr{T}$ is picked uniformly at random, with or without the class consistency requirement, as a function of the non-fixed points in the permutations. Following [15], we let $\mathbf{i}_r^m$ denote the set of r-tuples drawn without replacement drawn from $[m]$. We also let $\mathbf{i}_r^{m,b}$ denote the set of r-tuples drawn without replacement drawn from $[m] \cap \{i : y_i = b1\}$, for $b \in \{-,+\}$.

Theorem 17 *Under the joint Settings of Theorem 6 and Setting (B), let S_m^k denotes the set of permutations with exactly k non-fixed points, and S_m^{kb} its subset of block-class permutations with non-fixed points in class $b \in \{-,+\}$. Then for $S \in \{S_m^k, S_m^{k-}, S_m^{k+}\}$ the following holds over the uniform sampling of permutations:*

$$\mathbb{E}_{\mathbf{M}\sim S}[\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})] \;\;\leq\;\; u\cdot\frac{r_{\mathrm{s}}}{\sqrt{m}}\cdot\sqrt{\frac{k}{m}\cdot\mathscr{Q}}\;, \tag{75}$$

where $\mathscr{Q} \doteq \mathbb{E}_{(i,i')\sim\mathbf{i}_2^m}[\|\boldsymbol{x}_i^{\mathrm{s}} - \boldsymbol{x}_{i'}^{\mathrm{s}}\|_2^2]$ if $S = S_m^k$, and $\mathscr{Q} \doteq \mathbb{E}_{(i,i')\sim\mathbf{i}_2^{m,b}}[\|\boldsymbol{x}_i^{\mathrm{s}} - \boldsymbol{x}_{i'}^{\mathrm{s}}\|_2^2]$ if $S = S_m^{kb}$ ($b \in \{-,+\}$).

Proof We make the proof for $S = S_m^k$. The two other cases follow in the same way. We have from Theorem 6, because of Jensen inequality:

$$\begin{aligned}\mathbb{E}_{\mathbf{M}\sim S_m^k}\left[\mathrm{RCP}_{\mathscr{T}}(\mathcal{H})\right] &\leq u\cdot\frac{r_{\mathrm{s}}}{\sqrt{m}}\cdot\mathbb{E}_{\mathbf{M}\sim S_m^k}\left[\sqrt{\frac{1}{m}\cdot\langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}}\rangle_{\mathbf{M}}}\right]\\ &\leq u\cdot\frac{r_{\mathrm{s}}}{\sqrt{m}}\cdot\sqrt{\frac{1}{m}\cdot\mathbb{E}_{\mathbf{M}\sim S_m^k}\left[\langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}}\rangle_{\mathbf{M}}\right]}\;.\end{aligned} \tag{76}$$

We now decompose the expectation inside and first condition on the set of permutations whose set of non fixed points are the same set of k examples, say for $i \in [k]$. Let us call S_m^{k*} this subset of

S_m^k. In this case, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\mathrm{M}\sim S_m^{k*}}\left[\langle \mathrm{I}_m, \mathrm{K}^{\mathrm{s}}\rangle_{\mathrm{M}}\right] \\
&= \mathbb{E}_{\mathrm{M}\sim S_m^{k*}}\left[\mathrm{tr}\left((\mathrm{I}_m-\mathrm{M})\mathrm{K}^{\mathrm{s}}(\mathrm{I}_m-\mathrm{M})^\top\right)\right] \\
&= \mathrm{tr}\left(\mathrm{K}^{\mathrm{s}}\right)+\mathbb{E}_{\mathrm{M}\sim S_m^{k*}}\left[\mathrm{tr}\left(\mathrm{M}\mathrm{K}^{\mathrm{s}}\mathrm{M}^\top\right)\right]-2\cdot\mathbb{E}_{\mathrm{M}\sim S_m^{k*}}\left[\mathrm{tr}\left(\mathrm{M}\mathrm{K}^{\mathrm{s}}\right)\right] \qquad (77)\\
&= 2\cdot\left(\mathrm{tr}\left(\mathrm{K}^{\mathrm{s}}\right)-\mathrm{tr}\left(\mathbb{E}_{\mathrm{M}\sim S_m^{k*}}\left[\mathrm{M}\right]\mathrm{K}^{\mathrm{s}}\right)\right) \\
&= 2\cdot\left(\sum_{i\in[m]}\mathrm{K}^{\mathrm{s}}_{ii}-\sum_{i\in[k]}\frac{1}{k}\cdot\sum_{i'\in[k]}\mathrm{K}^{\mathrm{s}}_{ii'}-\sum_{i\in[m]\setminus[k]}\mathrm{K}^{\mathrm{s}}_{ii}\right) \qquad (78)\\
&= 2\cdot\left(\frac{k-1}{k}\cdot\sum_{i\in[k]}\mathrm{K}^{\mathrm{s}}_{ii}-\frac{1}{k}\cdot\sum_{(i,i')\in\mathbf{i}_2^k}\mathrm{K}^{\mathrm{s}}_{ii'}\right) \\
&= \frac{2}{k}\cdot\left(\sum_{(i,i')\in\mathbf{i}_2^k}\frac{\mathrm{K}^{\mathrm{s}}_{ii}+\mathrm{K}^{\mathrm{s}}_{i'i'}}{2}-\mathrm{K}^{\mathrm{s}}_{ii'}\right) \\
&= \frac{1}{k}\cdot\sum_{(i,i')\in\mathbf{i}_2^k}\|\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_{i'}^{\mathrm{s}}\|_2^2\ . \qquad (79)
\end{aligned}$$

In eq. (77) we use the fact that K^{s} is symmetric. Eq. (78) uses the fact that

$$\mathbb{E}_{\mathrm{M}\sim S_m^{k*}}\left[\mathrm{M}\right] = \left[\begin{array}{c|c} \mathrm{U}_k & 0 \\ 0 & \mathrm{I}_{m-k}\end{array}\right],$$

where we recall that $\mathrm{U}_k \doteq \frac{1}{k}\cdot\mathbf{1}\mathbf{1}^\top$ (main file, Definition 8). There remains to average eq. (79) over the set of all permutations whose set of fixed points is a different $(m-k)$-subset of $[m]$ and the statement of Theorem 17 is proven for $S=S_m^k$. ■

The key point in the bound is factor k/m, which implies that when permutations have lots of fixed points, say $(1-\Omega(1))m$, then the RCP may just vanish (as m increases) wrt the Rademacher complexity, whose dependency on m is $\Omega(1/\sqrt{m})$ [18].

7.2.5 Proof of Theorem 9

The proof stems from the following Theorem, which just assumes that K^u and K^v can be diagonalized (hence, it is applies to a more general setting than kernel functions).

Theorem 18 *Let K^u and K^v be two diagonalisable matrices with respective eigendecomposition $\{\lambda_i,\boldsymbol{u}_i\}_{i\in[d]}$ and $\{\mu_i,\boldsymbol{v}_i\}_{i\in[d]}$, with eigenvalues eventually duplicated up to their algebraic multiplicity. Letting $\overline{a}\doteq(1/m)\mathbf{1}^\top\boldsymbol{a}$ denote the average coordinate in $\boldsymbol{a}$, the difference in Hilbert-Schmidt Independence Criterion with respect to shuffling* M *satisfies:*

$$\mathrm{HSIC}(\mathrm{K}^u,\mathrm{K}^v)-\mathrm{HSIC}(\mathrm{K}^u,\mathrm{M}\mathrm{K}^v\mathrm{M}^\top) = -2m\cdot\left(\sum_i\lambda_i\overline{u}_i\boldsymbol{u}_i\right)^\top(\mathrm{I}_m-\mathrm{M})\left(\sum_i\mu_i\overline{v}_i\boldsymbol{v}_i\right) \qquad (80)$$

Hence, if $\mathrm{M} \in S_m^e$ *permutes* ℓ *and* ℓ' *in* $[m]$*, then* $\text{HSIC}(\mathrm{K}^u, \mathrm{MK}^v\mathrm{M}^\top) > \text{HSIC}(\mathrm{K}^u, \mathrm{K}^v)$ *iff:*

$$\left(\sum_i \lambda_i \overline{u}_i (u_{i\ell} - u_{i\ell'})\right)\left(\sum_i \mu_i \overline{v}_i (v_{i\ell} - v_{i\ell'})\right) \;>\; 0 \;. \tag{81}$$

Proof Being symmetric, K^u and K^v can be diagonalized as $\mathrm{K}^u = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ and $\mathrm{K}^v = \sum_i \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$. We use definition $\mathrm{U}_m \doteq (1/m)\mathbf{1}\mathbf{1}^\top$ for short. The following is folklore or can be can be checked after analytic derivations that we omit:

$$\begin{aligned}
\operatorname{tr}(\mathrm{U}_m\mathrm{K}^u\mathrm{K}^v) &= m\left(\sum_i \lambda_i \overline{u}_i \boldsymbol{u}_i\right)\left(\sum_i \mu_i \overline{v}_i \boldsymbol{v}_i\right) \\
&= \operatorname{tr}(\mathrm{K}^u\mathrm{U}_m\mathrm{K}^v) \\
\operatorname{tr}(\mathrm{U}_m\mathrm{K}^u\mathrm{U}_m\mathrm{K}^v) &= m^2\left(\sum_i \lambda_i (\overline{u}_i)^2\right)\left(\sum_i \mu_i (\overline{v}_i)^2\right) \\
\operatorname{tr}(\mathrm{K}^u\mathrm{K}^v) &= \sum_i \lambda_i \mu_i \;.
\end{aligned} \tag{82}$$

We thus get

$$\begin{aligned}
\text{HSIC}(\mathrm{K}^u, \mathrm{K}^v) &= \langle \mathrm{K}^u, \mathrm{K}^v \rangle_{\mathrm{U}_m} \\
&= \operatorname{tr}((\mathrm{I}_m - \mathrm{U}_m)\mathrm{K}^u(\mathrm{I}_m - \mathrm{U}_m)\mathrm{K}^v) \\
&= \operatorname{tr}(\mathrm{K}^u\mathrm{K}^v) - \operatorname{tr}(\mathrm{U}_m\mathrm{K}^u\mathrm{K}^v) - \operatorname{tr}(\mathrm{K}^u\mathrm{U}_m\mathrm{K}^v) + \operatorname{tr}(\mathrm{U}_m\mathrm{K}^u\mathrm{U}_m\mathrm{K}^v) \\
&= \sum_i \lambda_i\mu_i - 2m \cdot \left(\sum_i \lambda_i \overline{u}_i \boldsymbol{u}_i\right)^\top \left(\sum_i \mu_i \overline{v}_i \boldsymbol{v}_i\right) \\
&\quad + m^2 \cdot \left(\sum_i \lambda_i (\overline{u}_i)^2\right) \cdot \left(\sum_i \mu_i (\overline{v}_i)^2\right) \\
&= m^2\left(\frac{1}{m^2} \cdot \sum_i \lambda_i\mu_i - \frac{2}{m}\sum_{i,j} \lambda_i \overline{u}_i \boldsymbol{u}_i^\top \mu_j \overline{v}_j \boldsymbol{v}_j + \sum_{i,j} \lambda_i\mu_j (\overline{u}_i)^2 (\overline{v}_j)^2\right) \\
&= m^2\left(\frac{1}{m^2} \cdot \sum_i \lambda_i\mu_i - \frac{1}{m^2} \cdot \sum_{i,j} \lambda_i\mu_j (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 + \sum_{i,j} \lambda_i\mu_j \left(\frac{1}{m}\boldsymbol{u}_i^\top \boldsymbol{v}_j - \overline{u}_i\overline{v}_j\right)^2\right) \\
&= m^2\left(\frac{1}{m^2} \cdot \boldsymbol{\lambda}^\top (\mathrm{I} - \mathrm{C})\,\boldsymbol{\mu} + \boldsymbol{\lambda}^\top \mathrm{J} \boldsymbol{\mu}\right) \qquad (83) \\
&= \boldsymbol{\lambda}^\top \left((\mathrm{I} - \mathrm{C}) + m^2 \cdot \mathrm{J}\right)\boldsymbol{\mu} \;. \qquad (84)
\end{aligned}$$

We have used here the square cosine matrix C with $\mathrm{C}_{ij} \doteq \cos^2(\boldsymbol{u}_i, \boldsymbol{u}_j)$, and the square correlation matrix J with $\mathrm{J}_{ij} \doteq ((1/m)\boldsymbol{u}_i^\top \boldsymbol{v}_j - \overline{u}_i\overline{v}_j)^2$. Now, suppose we perform CP $\mathscr{T}$ with shuffling matrix M. K^v and its eigendecomposition become after shuffling

$$\mathrm{MK}^v\mathrm{M}^\top \;=\; \sum_i \mu_i (\mathrm{M}\boldsymbol{v}_i)(\mathrm{M}\boldsymbol{v}_i)^\top \;. \tag{85}$$

Remark that shuffling affects the order in the coordinate of *all* eigenvectors. So the difference between the two Hilbert-Schmidt Independence Criteria (before - after shuffling) is:

$$\begin{aligned}
&\text{HSIC}(\mathrm{K}^u, \mathrm{K}^v) - \text{HSIC}(\mathrm{K}^u, \mathrm{M}\mathrm{K}^v\mathrm{M}^\top) \\
&= \sum_{i,j} \lambda_i\mu_j((\boldsymbol{u}_i^\top \mathrm{M}\boldsymbol{v}_j)^2 - (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2) \\
&\quad - \sum_{i,j} \lambda_i\mu_j \left\{ \left(\boldsymbol{u}_i^\top \mathrm{M}\boldsymbol{v}_j - m\overline{u}_i\overline{v}_j\right)^2 - \left(\boldsymbol{u}_i^\top \boldsymbol{v}_j - m\overline{u}_i\overline{v}_j\right)^2 \right\} \\
&= \sum_{i,j} \lambda_i\mu_j \cdot \boldsymbol{u}_i^\top(\mathrm{M} - \mathrm{I}_m)\boldsymbol{v}_j \cdot \boldsymbol{u}_i^\top(\mathrm{M} + \mathrm{I}_m)\boldsymbol{v}_j \\
&\quad - \sum_{i,j} \lambda_i\mu_j \left\{ \boldsymbol{u}_i^\top(\mathrm{M} - \mathrm{I}_m)\boldsymbol{v}_j \cdot (\boldsymbol{u}_i^\top(\mathrm{M} + \mathrm{I}_m)\boldsymbol{v}_j - 2m\overline{u}_i\overline{v}_j) \right\} \\
&= 2m \cdot \sum_{i,j} (\lambda_i\overline{u}_i)\boldsymbol{u}_i^\top(\mathrm{M} - \mathrm{I}_m)(\mu_j\overline{v}_j) \\
&= 2m \cdot \left(\sum_i \lambda_i\overline{u}_i\boldsymbol{u}_i\right)^\top (\mathrm{M} - \mathrm{I}_m) \left(\sum_i \mu_i\overline{v}_i\boldsymbol{v}_i\right) \quad . \qquad (86)
\end{aligned}$$

We now remark that whenever $\mathrm{M} \in S_m^{\mathrm{e}}$, if it permutes ℓ and ℓ' in $[m]$, then $\boldsymbol{a}(\mathrm{M} - \mathrm{I}_m)\boldsymbol{b} = a_\ell(b_{\ell'} - b_\ell) + a_{\ell'}(b_\ell - b_{\ell'}) = -(a_\ell - a_{\ell'})(b_\ell - b_{\ell'})$, so we get:

$$\begin{aligned}
&\text{HSIC}(\mathrm{K}^u, \mathrm{K}^v) - \text{HSIC}(\mathrm{K}^u, \mathrm{M}\mathrm{K}^v\mathrm{M}^\top) \\
&\quad = -2m \left(\sum_i \lambda_i\overline{u}_i(u_{i\ell} - u_{i\ell'})\right)\left(\sum_i \mu_i\overline{v}_i(v_{i\ell} - v_{i\ell'})\right) \quad , \qquad (87)
\end{aligned}$$

and we get ineq. (81). ■

This ends the proof of Theorem 9.

7.2.6 Proof of Theorem 10

The Theorem is a direct consequence of the following Theorem.

Theorem 19 *Let K^u and K^v be two kernel functions over $\mathcal{S}$. Then for any elementary permutation $\mathrm{M} \in S_m^{\mathrm{e}}$ that permutes ℓ and ℓ' in $[m]$,*

$$\text{HSIC}(\mathrm{K}^u, \mathrm{M}\mathrm{K}^v\mathrm{M}^\top) - \text{HSIC}(\mathrm{K}^u, \mathrm{K}^v) \quad = \quad -2m \cdot \mathbb{C}\mathrm{ov}(\boldsymbol{\delta}_{\ell\ell'}^u, \boldsymbol{\delta}_{\ell\ell'}^v) + \mathscr{R}_{\ell\ell'}^{u,v} \quad , \qquad (88)$$

with

$$\mathscr{R}_{\ell\ell'}^{u,v} \quad \doteq \quad (\mathrm{K}_{\ell\ell}^u - \mathrm{K}_{\ell'\ell}^u)(\mathrm{K}_{\ell\ell}^v - \mathrm{K}_{\ell'\ell}^v) + (\mathrm{K}_{\ell'\ell'}^u - \mathrm{K}_{\ell'\ell}^u)(\mathrm{K}_{\ell'\ell'}^v - \mathrm{K}_{\ell'\ell}^v) \quad . \qquad (89)$$

Furthermore, the uniform sampling of elementary permutations in S_m^{e} satisfies:

$$\mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}\left[\text{HSIC}(\mathrm{K}^u, \mathrm{M}\mathrm{K}^v\mathrm{M}^\top)\right] \quad = \quad \left(1 - \frac{8}{m-1}\right) \cdot \text{HSIC}(\mathrm{K}^u, \mathrm{K}^v) + \frac{8}{m-1} \cdot \mathscr{R}^{u,v} \quad , \qquad (90)$$

with

$$\mathscr{R}^{u,v} \doteq \sum_i \mathrm{K}^u_{ii}\mathrm{K}^v_{ii} - \frac{1}{m}\cdot\left(\frac{\sum_i \mathrm{K}^u_{ii}\mathrm{K}^v_{.i} + \sum_i \mathrm{K}^u_{.i}\mathrm{K}^v_{ii}}{2}\right) . \tag{91}$$

Here, when replacing an index notation by a point, ".", we denote a sum over all possible values of this index.

Proof We first decompose $\mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v)$:

$$\mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v) = \sum_{i,i'} \mathrm{K}^u_{ii'}\mathrm{K}^v_{ii'} - \frac{2}{m}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} + \frac{1}{m^2}\mathrm{K}^u_{..}\mathrm{K}^u_{..} . \tag{92}$$

for any $(\ell,\ell') \in \mathbf{i}_2^m$. For any $\mathrm{M} \in S_m^{\mathrm{e}}$ denoting an elementary permutation ς of the features in $\mathcal{F}^{\mathrm{s}}$ such that $\mathcal{V} \subseteq \mathcal{F}^{\mathrm{s}}$ and $\varsigma(\ell) = \ell'$, $\varsigma(\ell') = \ell$, we obtain:

$$\begin{aligned}
&\mathrm{HSIC}(\mathrm{K}^u, \mathrm{M}\mathrm{K}^v\mathrm{M}^\top) - \mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v) \\
&= 2\cdot\left(\sum_{i\neq\ell,\ell'} \mathrm{K}^u_{\ell i}\mathrm{K}^v_{\ell' i} + \sum_{i\neq\ell,\ell'} \mathrm{K}^u_{\ell' i}\mathrm{K}^v_{\ell i} - \sum_{i\neq\ell,\ell'} \mathrm{K}^u_{\ell i}\mathrm{K}^v_{\ell i} - \sum_{i\neq\ell,\ell'} \mathrm{K}^u_{\ell' i}\mathrm{K}^v_{\ell' i}\right) \\
&\quad -\frac{2}{m}\cdot\mathrm{K}^u_{\ell.}\mathrm{K}^v_{\ell'.} - \frac{2}{m}\cdot\mathrm{K}^u_{\ell'.}\mathrm{K}^v_{\ell.} + \frac{2}{m}\cdot\mathrm{K}^u_{\ell.}\mathrm{K}^v_{\ell.} + \frac{2}{m}\cdot\mathrm{K}^u_{\ell'.}\mathrm{K}^v_{\ell'.} \\
&= -2\left(\sum_i (\mathrm{K}^u_{\ell i} - \mathrm{K}^u_{\ell' i})(\mathrm{K}^v_{\ell i} - \mathrm{K}^v_{\ell' i}) - \frac{1}{m}\cdot(\mathrm{K}^u_{\ell.} - \mathrm{K}^u_{\ell'.})(\mathrm{K}^v_{\ell.} - \mathrm{K}^v_{\ell'.})\right) \\
&\quad +(\mathrm{K}^u_{\ell\ell} - \mathrm{K}^u_{\ell'\ell})(\mathrm{K}^v_{\ell\ell} - \mathrm{K}^v_{\ell'\ell}) + (\mathrm{K}^u_{\ell'\ell'} - \mathrm{K}^u_{\ell'\ell})(\mathrm{K}^v_{\ell'\ell'} - \mathrm{K}^v_{\ell'\ell}) \\
&= -2m\cdot\mathbb{C}\mathrm{ov}(\boldsymbol{\delta}^u_{\ell\ell'}, \boldsymbol{\delta}^v_{\ell\ell'}) + \mathscr{R}^{u,v}_{\ell\ell'} ,
\end{aligned}$$

which is eq. (88). We also have:

$$\begin{aligned}
\mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}\left[\sum_i (\mathrm{K}^u_{\ell i} - \mathrm{K}^u_{\ell' i})(\mathrm{K}^v_{\ell i} - \mathrm{K}^v_{\ell' i})\right] &= \frac{4}{m}\cdot\sum_{i,i'} \mathrm{K}^u_{ii'}\mathrm{K}^v_{ii'} - \frac{4}{m(m-1)}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} \\
&\quad +\frac{4}{m(m-1)}\cdot\sum_{i,i'} \mathrm{K}^u_{ii'}\mathrm{K}^v_{ii'} \\
&= \frac{4}{m-1}\cdot\sum_{i,i'} \mathrm{K}^u_{ii'}\mathrm{K}^v_{ii'} - \frac{4}{m(m-1)}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} ,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}\left[(\mathrm{K}^u_{\ell.} - \mathrm{K}^u_{\ell'.})(\mathrm{K}^v_{\ell.} - \mathrm{K}^v_{\ell'.})\right] &= \frac{4}{m}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} - \frac{4}{m(m-1)}\cdot\sum_{i\in[m]} \mathrm{K}^u_{i.}\sum_{i'\in[m]\setminus\{i\}} \mathrm{K}^v_{i'.} \\
&= \frac{4}{m}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} - \frac{4}{m(m-1)}\cdot\sum_i \mathrm{K}^u_{i.}\cdot(\mathrm{K}^v_{..} - \mathrm{K}^v_{i.}) \\
&= \frac{4}{m}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} - \frac{4}{m(m-1)}\mathrm{K}^u_{..}\mathrm{K}^v_{..} + \frac{4}{m(m-1)}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} \\
&= \frac{4}{m-1}\cdot\sum_i \mathrm{K}^u_{i.}\mathrm{K}^v_{i.} - \frac{4}{m(m-1)}\mathrm{K}^u_{..}\mathrm{K}^v_{..} ,
\end{aligned}$$

and finally

$$\begin{aligned}\mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}[\mathscr{R}_{\ell\ell'}^{u,v}] &= \frac{8}{m-1}\cdot\left[\sum_i \mathrm{K}_{ii}^u\mathrm{K}_{ii}^v - \frac{1}{m}\cdot\left(\frac{\sum_i \mathrm{K}_{ii}^u\mathrm{K}_{\cdot i}^v + \sum_i \mathrm{K}_{\cdot i}^u\mathrm{K}_{ii}^v}{2}\right)\right] \\ &\doteq \frac{8}{m-1}\cdot\mathscr{R}^{u,v}\ , \end{aligned} \tag{93}$$

since

$$\begin{aligned}&\mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}[(\mathrm{K}_{\ell\ell}^u - \mathrm{K}_{\ell'\ell}^u)(\mathrm{K}_{\ell\ell}^v - \mathrm{K}_{\ell'\ell}^v)] \\ &= \frac{4}{m(m-1)}\sum_i\sum_{i'\neq i}\mathrm{K}_{ii}^u\mathrm{K}_{ii}^v - \frac{2}{m(m-1)}\sum_i\sum_{i'\neq i}\mathrm{K}_{ii}^u\mathrm{K}_{i'i}^v - \frac{2}{m(m-1)}\sum_i\sum_{i'\neq i}\mathrm{K}_{i'i}^u\mathrm{K}_{ii}^v \\ &= \frac{4}{m}\sum_i\mathrm{K}_{ii}^u\mathrm{K}_{ii}^v - \frac{2}{m(m-1)}\sum_i\mathrm{K}_{ii}^u(\mathrm{K}_{\cdot i}^v - \mathrm{K}_{ii}^v) - \frac{2}{m(m-1)}\sum_i(\mathrm{K}_{i\cdot}^u - \mathrm{K}_{ii}^u)\mathrm{K}_{ii}^v \\ &= \frac{4}{m}\sum_i\mathrm{K}_{ii}^u\mathrm{K}_{ii}^v + \frac{4}{m(m-1)}\sum_i\mathrm{K}_{ii}^u\mathrm{K}_{ii}^v - \frac{2}{m(m-1)}\sum_i\mathrm{K}_{ii}^u\mathrm{K}_{\cdot i}^v - \frac{2}{m(m-1)}\sum_i\mathrm{K}_{i\cdot}^u\mathrm{K}_{ii}^v \\ &= \frac{4}{m-1}\sum_i\mathrm{K}_{ii}^u\mathrm{K}_{ii}^v - \frac{2}{m(m-1)}\sum_i\mathrm{K}_{ii}^u\mathrm{K}_{\cdot i}^v - \frac{2}{m(m-1)}\sum_i\mathrm{K}_{i\cdot}^u\mathrm{K}_{ii}^v \\ &= \mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}[(\mathrm{K}_{\ell'\ell'}^u - \mathrm{K}_{\ell'\ell}^u)(\mathrm{K}_{\ell'\ell'}^v - \mathrm{K}_{\ell'\ell}^v)]\ . \end{aligned}$$

So we obtain

$$\begin{aligned}&\mathbb{E}_{\mathrm{M}\sim S_m^{\mathrm{e}}}\left[\mathrm{HSIC}(\mathrm{K}^u, \mathrm{M}\mathrm{K}^v\mathrm{M}^\top) - \mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v)\right] \\ &= -\frac{8}{m-1}\cdot\sum_{i,i'}\mathrm{K}_{ii'}^u\mathrm{K}_{ii'}^v + \frac{8}{m(m-1)}\cdot\sum_i\mathrm{K}_{i\cdot}^u\mathrm{K}_{i\cdot}^v \\ &\quad + \frac{8}{m(m-1)}\cdot\sum_i\mathrm{K}_{i\cdot}^u\mathrm{K}_{i\cdot}^v - \frac{8}{m^2(m-1)}\mathrm{K}_{\cdot\cdot}^u\mathrm{K}_{\cdot\cdot}^v + \frac{8}{m-1}\cdot\mathscr{R}^{u,v} \\ &= -\frac{8}{m-1}\cdot\mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v) + \frac{8}{m-1}\cdot\mathscr{R}^{u,v}\ . \end{aligned}$$

This ends the proof of Theorem 19. ■

When kernel functions have unit diagonal (such as for the Gaussian kernel), eq. (91) simplifies to:

$$\mathscr{R}^{u,v} \doteq m\left(1 - \frac{1}{m^2}\cdot\left(\frac{\mathrm{K}_{\cdot\cdot}^u + \mathrm{K}_{\cdot\cdot}^v}{2}\right)\right)\ . \tag{94}$$

Hence, provided we perform $T \doteq \epsilon m$ elementary permutations, there exists a sequence of such permutations such that the composition $\mathrm{M}_* \doteq \mathrm{M}_T\mathrm{M}_{T-1}\cdots\mathrm{M}_1$ satisfies:

$$\begin{aligned}\mathrm{HSIC}(\mathrm{K}^u, \mathrm{M}_*\mathrm{K}^v\mathrm{M}_*^\top) &\leq \left(1 - \frac{8}{m-1}\right)^{\epsilon m}\cdot\mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v) + \left[1 - \left(1 - \frac{8}{m-1}\right)^{\epsilon m}\right]\cdot\mathscr{R}^{u,v} \\ &\leq \alpha(\epsilon)\cdot\mathrm{HSIC}(\mathrm{K}^u, \mathrm{K}^v) + (1-\alpha(\epsilon))\cdot\mathscr{R}^{u,v}\ , \end{aligned} \tag{95}$$

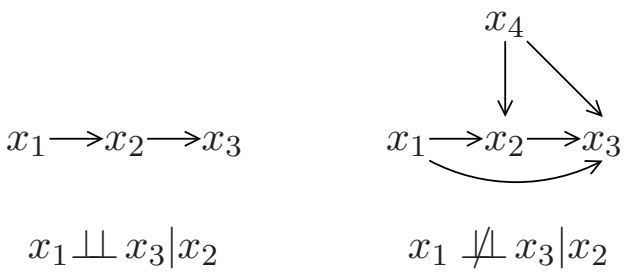


Figure 3: The Cornia-Mooij model [6]. Left: belief, right: true model.

with

$$\alpha(\epsilon) \doteq \exp(-8\epsilon) , \tag{96}$$

as long as $\text{HSIC}(\text{K}^u, \text{K}^v) \geq \mathscr{R}^{u,v}$. We have used the fact that

$$\begin{aligned} \left(1 - \frac{8}{m-1}\right)^{\epsilon m} &\leq \exp\left(-\frac{8\epsilon m}{m-1}\right) \\ &\leq \exp(-8\epsilon) . \end{aligned} \tag{97}$$

This achieves the proof of Theorem 10.

7.2.7 The Cornia-Mooij model and results

We now show how to trick statistical tests into keeping independence *and* then incur arbitrarily large errors in estimating causal effects. The model we refer to is the Cornia-Mooij (CM) model [6], shown in Figure 3. In the CM model, there are $d = 3$ observation variables, and a true model which relies on a weak conditional dependence $x_1 \not\perp\!\!\!\perp x_3 | x_2$. [6] show that *if* one keeps the independence assumption H_0 that $x_1 \perp\!\!\!\perp x_3 | x_2$, this can lead to very high causal estimation errors, as measured by $|\mathbb{E}[x_3|x_2] - \mathbb{E}[x_3|\text{do}(x_2)]|/|x_2|$ [6]. We show that the CP is precisely able to trick statistics into keeping H_0.

There is a hidden confounder x_4, which is assumed to be independent from x_1. The true model makes the following statistical dependence assumptions:

- $x_1 \not\perp\!\!\!\perp x_2$,
- $x_2 \not\perp\!\!\!\perp x_3$,
- and the most important one, which we scramble through the CP, $x_1 \not\perp\!\!\!\perp x_3 | x_2$.

We chose this simple model because (a) it belongs to the few worst-case models for causality analysis, and (b) it shows, in addition to jamming (non)linear correlations, how CP can also jam partial correlations. In the CM model, there are $d = 3$ observation variables, and a true model which relies on a weak conditional dependence $x_1 \not\perp\!\!\!\perp x_3 | x_2$. [6] show that *if* one keeps the independence assumption H_0 that $x_1 \perp\!\!\!\perp x_3 | x_2$, this can lead to very high causal estimation errors[3]. We show that it is

[3] As measured by $|\mathbb{E}[x_3|x_2] - \mathbb{E}[x_3|\text{do}(x_2)]|/|x_2|$ [6].

possible, through a CP, to trick statistics into *keeping H_0 as well*. In the following, $\mathcal{F}_{\rm s} = \{x_3\}$. It is shown in [6] that $x_1 \perp\!\!\!\perp x_3 | x_2$ iff the partial correlation $\rho_{(13)\cdot 2} \doteq (\rho_{13} - \rho_{12}\rho_{23})/\sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)}$ vanishes. Assuming $\rho_{(13)\cdot 2}$ is large enough in the dataset we have (so that we would reject H_0 from observing $\mathcal{S}$), we show how to reduce it through a sequence of CPs, using a similar strategy as in Theorem 10, the main difference being that we rely on block-class permutations. For any $\varsigma \in S_m$, notation 3^{ς} indicates column variable 3 shuffled. We assume $\rho_{(13)\cdot 2} > 0$ (the same analysis can be done if $\rho_{(13)\cdot 2} < 0$).

Theorem 20 *Suppose that there exists $\epsilon > 0$ such that $\rho_{12}^2 \leq 1-\epsilon$ and $\rho_{23^{\varsigma}}^2 \leq 1-\epsilon$ for any $\varsigma \in S_m^*$. Then there exists $T > 0$ and a sequence of T elementary permutations in S_m^* such that $\rho_{(13^{\varsigma})\cdot 2}$ is strictly decreasing in the sequence and meets at the end $\rho_{(13^{\varsigma})\cdot 2} \leq \mathscr{R}$ with*

$$\mathscr{R} \doteq (1-\epsilon)^{-1} \cdot p_+(1-p_+) \cdot (\tilde{\mu}_1 - \rho_{12} \cdot \tilde{\mu}_2) \cdot \tilde{\mu}_3 \ ,$$

and $\tilde{\mu}_j \doteq (1/(2\sqrt{v_j})) \cdot \sum_{y'} y' \mathbb{E}_{(\boldsymbol{x},y)\sim \mathcal{S}}[x_j | y = y']$.

Proof

The Theorem is a direct consequence of the following Lemma.

Lemma 21 *Let c_{jk} denote the covariance between columns j and k, $\mu_l^b \doteq (1/m_b) \sum_{i:y_i=b} x_{il}$ and $p \doteq m_+/m$. Suppose that $j \in \mathcal{F}_{\rm a}$ and $k \in \mathcal{F}_{\rm s}$. Then as long as*

$$c_{jk} > p(1-p) \cdot (\mu_j^+ - \mu_j^-) \cdot (\mu_k^+ - \mu_k^-) \ , \tag{98}$$

there always exist $\varsigma \in S_m^*$ such that $\rho_{jk^{\varsigma}} < \rho_{jk}$, where k^{ς} denote column variable k shuffled according to ς in the corresponding CP.

Proof Let us denote for short $\boldsymbol{c}_j \in \mathbb{R}^m$ the j^{th} feature column. We have because of the fact that $\mu_{k^{\varsigma}} = \mu_k$ and $v_{k^{\varsigma}} = v_k$ ($v_{\cdot}$ being the variance):

$$\rho_{jk^{\varsigma}} - \rho_{jk} = \frac{1}{m\sqrt{v_j v_k}} \cdot \left(\boldsymbol{c}_j^{\top} \left(\mathrm{M}_{\varsigma} - \mathrm{I}_m \right) \boldsymbol{c}_k \right) \ , \tag{99}$$

where M_{ς} is the shuffling matrix of permutation ς. Matrix $\mathrm{M}_{\varsigma} - \mathrm{I}_m$ has only four non-zero coordinates: in (ℓ,ℓ) and (ℓ',ℓ') (both -1), and in (ℓ,ℓ') and (ℓ',ℓ) (both 1), so we get:

$$\begin{aligned} \rho_{jk^{\varsigma}} - \rho_{jk} &= \frac{1}{m\sqrt{v_j v_k}} \cdot (x_{\ell j}(x_{\ell' k} - x_{\ell k}) + x_{\ell' j}(x_{\ell k} - x_{\ell' k})) \\ &= -\frac{1}{m\sqrt{v_j v_k}} \cdot ((x_{\ell j} - x_{\ell' j})(x_{\ell k} - x_{\ell' k})) \ . \end{aligned} \tag{100}$$

Hence, $\rho_{jk^{\varsigma}} < \rho_{jk}$ iff the sign of $x_{\ell j} - x_{\ell' j}$ is the same as the sign of $x_{\ell k} - x_{\ell' k}$. Let $\pi^b(\ell)$ the predicate $\rho_{jk^{\varsigma}} \geq \rho_{jk}$, for any elementary permutation $\varsigma \in S_m^*$ that changes ℓ to index ℓ' of the same class b ($\varsigma(\ell) = \ell'$, $\varsigma(\ell') = \ell$). If $\pi^b(\ell)$ is true, then, averaging over all such permutations, we obtain:

$$\begin{aligned} 0 \geq & -\frac{1}{m\sqrt{v_j v_k}} \cdot \frac{1}{m_b} \sum_{\ell'} \left((x_{\ell j}^b - x_{\ell' j}^b)(x_{\ell k}^b - x_{\ell' k}^b) \right) \\ &= -\frac{1}{m\sqrt{v_j v_k}} \cdot \left(x_{\ell j}^b x_{\ell k}^b - x_{\ell j}^b \mu_k^b - x_{\ell k}^b \mu_j^b + \mu_{jk}^b \right) \\ &= -\frac{1}{m\sqrt{v_j v_k}} \cdot \left(c_{jk}^b + (x_{\ell j}^b - \mu_j^b)(x_{\ell k}^b - \mu_k^b) \right) \ , \end{aligned}$$

i.e. we have:

$$(x^b_{\ell j} - \mu^b_j)(x^b_{\ell k} - \mu^b_k) \;\geq\; -c^b_{jk} \;. \tag{101}$$

Assume now that $\pi^b(\ell)$ holds over any $\ell \in [m_b]$. As long as c^b_{jk} is strictly positive, we thus obtain, averaging ineq. (101) over all $\ell \in [m_b]$,

$$\begin{aligned} c^b_{jk} &\doteq \frac{1}{m_b}\sum_\ell (x^b_{\ell j} - \mu^b_j)(x^b_{\ell k} - \mu^b_k) \\ &\leq -c^b_{jk} \\ &< 0 \;, \end{aligned} \tag{102}$$

a contradiction. Hence, as long as $c^b_{jk} > 0$, there must exist $(\ell, \ell') \in \mathbf{i}_2^{m_b}$ such that the elementary permutation $\varsigma(\ell) = \ell'$, $\varsigma(\ell') = \ell$ satisfies

$$c^b_{jk^\varsigma} \;<\; c^b_{jk} \;, \tag{103}$$

and this holds for $b \in \{-,+\}$. Now remark that

$$\begin{aligned} c_{jk} &= p\mu^+_{jk} + (1-p)\mu^-_{jk} - (p\mu^+_j + (1-p)\mu^-_j)(p\mu^+_k + (1-p)\mu^-_k) \\ &= pc^+_{jk} + (1-p)c^-_{jk} + p(1-p)(\mu^+_j - \mu^-_j)(\mu^+_k - \mu^-_k) \;, \end{aligned} \tag{104}$$

and so as long as whichever $c^+_{jk} > 0$ or $c^-_{jk} > 0$, we can always find an elementary permutation that decreases the one chosen. When no more elementary permutations achieve that, $c_{jk} \leq p(1-p)(\mu^+_j - \mu^-_j)(\mu^+_k - \mu^-_k)$, which yields the statement of the Lemma. ■

To prove the Theorem, remark that

$$\begin{aligned} \rho_{13^\varsigma} - \rho_{12}\rho_{23^\varsigma} &= \frac{1}{\sqrt{v_1 v_3}}\left(\frac{1}{m}\cdot\sum_i \left(x_{i1} - \frac{c_{12}}{v_2}\cdot x_{i2}\right) x_{\varsigma(i)3}\right) - \left(\frac{\mu_1\mu_3}{\sqrt{v_1 v_3}} - c_{12}\frac{\mu_2\mu_3}{v_2\sqrt{v_1 v_3}}\right) \\ &= \frac{1}{\sqrt{v_1 v_3}}\left(\frac{1}{m}\cdot\sum_i \left(x_{i1} - \frac{c_{12}}{v_2}\cdot x_{i2}\right) x_{\varsigma(i)3} - \left(\mu_1 - \frac{c_{12}\mu_2}{v_2}\right)\mu_3\right) \end{aligned} \tag{105}$$

We apply Lemma 21 to linearly transformed column $\boldsymbol{c}' \doteq \boldsymbol{c}_1 - \frac{c_{12}}{v_2}\boldsymbol{c}_2$ and column $\boldsymbol{c}_3$ and obtain that as long as

$$\rho_{13} - \rho_{12}\rho_{23} \;>\; \frac{p(1-p)}{\sqrt{v_1 v_3}}\left((\mu^+_1 - \mu^-_1) - \frac{c_{12}}{v_2}(\mu^+_2 - \mu^-_2)\right)(\mu^+_3 - \mu^-_3) \;, \tag{106}$$

there always exist a block-class elementary permutation ς that is going to make $\rho_{13^\varsigma} - \rho_{12}\rho_{23^\varsigma} < \rho_{13} - \rho_{12}\rho_{23}$. When no such permutation exist anymore, we have, letting ς_T denote the composition of all elementary permutations performed so far and $\varsigma_* \doteq \arg\max_{\varsigma \in S^*_m} \rho^2_{23^\varsigma}$,

$$\begin{aligned} \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{1-\rho^2_{12}}\sqrt{1-\rho^2_{23^{\varsigma_T}}}} &\leq \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{1-\rho^2_{12}}\sqrt{1-\rho^2_{23^{\varsigma_*}}}} \\ &= \frac{p(1-p)}{\sqrt{v_1 v_2 - c^2_{12}}\sqrt{v_2 v_3 - c^2_{23^{\varsigma_*}}}} \\ &\cdot \left(v_2(\mu^+_1 - \mu^-_1) - c_{12}(\mu^+_2 - \mu^-_2)\right)(\mu^+_3 - \mu^-_3) \\ &= \frac{p(1-p)}{1-\epsilon}\left(\frac{\mu^+_1 - \mu^-_1}{\sqrt{v_1}} - \rho_{12}\cdot\frac{\mu^+_2 - \mu^-_2}{\sqrt{v_2}}\right)\cdot\frac{\mu^+_3 - \mu^-_3}{\sqrt{v}_3} \end{aligned} \tag{107}$$

as long as $c_{12}^2 \leq (1-\epsilon)v_1v_2$ and $c_{23^{\varsigma*}}^2 \leq (1-\epsilon)v_2v_3$. We just have to use the fact that

$$\tilde{\mu}_j \;=\; \frac{\mu_j^+ - \mu_j^-}{\sqrt{v_j}} \tag{108}$$

using the main file notation to conclude (End of the proof of Theorem 20). ■

Remark that the proof also shows that the conditions to blow up the type-II error (Corollary 2.1 in [6]) are not affected by the CP. To see that it is possible to still make the Type II error blow, up, in the CM model, the Type II error can be made at least K/v_2 [6] where the coefficient K does not depend on ς ([6], Corollary 2.1). Since v_2 is also not altered by the permutations, the Type II error can still be blown up following [6]'s construction.

We now show that the iterative process is actually not necessary if one has enough data: sampling $\varsigma \sim S_m^*$ jams $\rho_{(13^\varsigma)\cdot 2}$ up to bounds competitive with Theorem 20 with high probability. Such good concentration results also hold for HSIC [36].

Theorem 22 *For any $\delta > 0$, provided $m = \Omega((1/\delta)\log(1/\delta))$, the uniform sampling of ς in S_m^* satisfies*

$$\mathbb{P}_{\varsigma\sim S_m^*}[\rho_{(13^\varsigma)\cdot 2} \;\leq\; \mathscr{R} + \delta] \geq 1-\delta \;,$$

where $\mathscr{R}$ is defined in Theorem 20.

Proof We detail first the sampling process of ς. It relies on the fundamental property that a permutation uniquely factors as a product of disjoint cycles, and so a block-class permutation ς factors uniquely as two permutations ς_+ and ς_-, each of which acts in one of the two classes. Therefore, sampling uniformly each of ς_+ and ς_- results in an uniform sampling of a block-class ς.

The Theorem stems from the following Lemma, whose notations follow Lemma 21.

Lemma 23 *For any $q > 0$, as long as*

$$m \;=\; \Omega\left(\frac{1}{q}\log\frac{1}{q}\right) \;, \tag{109}$$

there is probability $\geq 1-q$ that a randomly chosen block-class permutation ς shall bring

$$c_{jk^\varsigma} \;\in\; \left[p(1-p)(\mu_j^+ - \mu_j^-)(\mu_k^+ - \mu_k^-) - q, p(1-p)(\mu_j^+ - \mu_j^-)(\mu_k^+ - \mu_k^-) + q\right] \;. \tag{110}$$

Proof Let $S_m^{*b} \subset S_m^*$ denote the set of block-class permutations whose set of fixed points contains all examples from class $\neq b1$, $\forall b \in \{-,+\}$. We have

$$\mathbb{E}_{\varsigma\sim S_m^{*b}}\left[\sum_{l:y_l=b} x_{lj}^b x_{\varsigma(l)k}^b\right] = m_b\mu_j^b\mu_k^b \;,$$

and since $(1/m_b)\sum_{l:y_l=b} x_{lj}^b x_{\varsigma(l)k}^b - \mu_j^b\mu_k^b = c_{jk^\varsigma}^b$, if we sample uniformly at random $\varsigma \sim S_m^{*b}$, then we get from [5] (Proposition 1.1):

$$\begin{aligned}\mathbb{P}_{\varsigma\sim S_m^{*b}}[|c_{jk^\varsigma}^b| \geq t] \;&=\; \mathbb{P}_{\varsigma\sim S_m^{*b}}\left[\left|\sum_{l:y_l=b} x_{lj}^b x_{\varsigma(l)k}^b - m_b\mu_j^b\mu_k^b\right| \geq m_b t\right] \\ &\leq\; 2\exp\left(-\frac{m_b t^2}{4\mu_j^b\mu_k^b + 2t}\right) \;. \end{aligned} \tag{111}$$

We want the right hand side to be no more than some δ_b; equivalently, we want

$$t^2 - \left(\frac{2}{m_b}\log\frac{2}{\delta_b}\right)t - \frac{2}{m_b}\log\frac{2}{\delta_b} \;\geq\; 0 \;, \tag{112}$$

which holds provided

$$t \;\geq\; \frac{2(1+o(1))}{m_b}\log\frac{2}{\delta_b} \;, \tag{113}$$

where the little-oh is measured wrt m_b. Since a block-class permutation factors as two fully determined permutations from S_m^{*+} and S_m^{*-}, if we fix $\delta_+ = \delta_- = \delta/2$, we get that if we sample uniformly at random these two permutations $\varsigma_+ \sim S_m^{*+}$ and $\varsigma_- \sim S_m^{*-}$, then we shall have simultaneously

$$|c^b_{jk^\varsigma}| \;\leq\; \frac{2(1+o(1))}{m_b}\log\frac{4}{\delta} \;, \forall b \in \{-,+\} \;, \tag{114}$$

which implies for the factored permutation ς,

$$\begin{aligned} |c_{jk^\varsigma} - p(1-p)(\mu_j^+ - \mu_j^-)(\mu_k^+ - \mu_k^-)| \;&\leq\; \sum_b \frac{m_b}{m}\cdot\frac{2(1+o(1))}{m_b}\log\frac{4}{\delta} \\ &= \frac{2(1+o(1))}{m}\log\frac{4}{\delta} \end{aligned} \tag{115}$$

from eq. (104). Hence, if

$$m \;=\; \Omega\left(\frac{1}{\delta}\log\frac{1}{\delta}\right) \;, \tag{116}$$

there will be probability $\geq 1-\delta$ that c_{jk^ς} is within additive δ from $p(1-p)(\mu_j^+ - \mu_j^-)(\mu_k^+ - \mu_k^-)$. ■

We get that with probability $\geq 1-\delta$, a randomly chosen block-class permutation ς shall make

$$\frac{\rho_{13^\varsigma} - \rho_{12}\rho_{23^\varsigma}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23^\varsigma}^2}} \;\leq\; \frac{p(1-p)}{1-\epsilon}\left(\frac{\mu_1^+ - \mu_1^-}{\sqrt{v_1}} - \rho_{12}\cdot\frac{\mu_2^+ - \mu_2^-}{\sqrt{v_2}}\right)\cdot\frac{\mu_3^+ - \mu_3^-}{\sqrt{v_3}} + \delta \;, \tag{117}$$

and the Theorem is proven (End of the proof of Theorem 22). ■

7.3 Appendix — Experiments

7.3.1 Domains and setup

Domain characteristics are described in Table 3. In particular, the process for train/test split is there given in detail for each dataset. Some domains deserve more comments.

Similarly to [16], we consider only two features of the *abalone* datasets; those are *rings* (the age) and *length*, which are provably causally linked –*age* causes *length*–, and hence correlated.

name	$m^{(*)}$	d	source	notes
digoxin	35	2	[7]	features are cond. independent given label
glass	146	9	UCI	
abalone-2D	200, 567	2	UCI, [16]	subsample, {rings, length} predict diameter
synthetic	200	13	scikit-learn	
heart	270	13	UCI	
liver disorders	345	7	UCI, [23]	predict mcv $< 30^{th}$ percentile, task `pair0034`
ionosphere	351	34	UCI	
auto+mpg	398	8	UCI, [23]	predict mpg $<$ mean, task `pair0016` (feature vectors with missing values removed)
arrhythmia	452	279	UCI, [23]	task `pair0023`, missing values replaced by 0
breastw	683	10	UCI	
australian	690	14	UCI	
diabete(-2)	768	8	UCI	Pima domain (-2 = task `pair0038`, [23], half of the dataset used for training)

Table 3: Domains considered. $^{(*)}$ When only one number appears in the column, 1/5 of m was hold out at random for test; when two numbers are present, the first is training set size, while the second is test size, that is fixed by the dataset description.

We predict the attribute *diameter* (reasonably caused by *age* as well); to turn this into a binary classification problem, we classify if the *diameter* is above or below the average one. (We also exclude abalone examples which have missing *sex* attribute.) For the experiments, we train with 200 examples and held out 567, both picked at random. The *digoxin* domain [7] is already defined by only two features, *digoxin* and *urine*, which are conditionally independent given *creatine*. From those, we predict if the level of *creatine* is above or below average. Domains Liver disorder, Auto+MPG, Arrhythmia and Diabete are part of the benchmark of domains of [23].

The *synthetic* dataset is generated by the function `datasets.make_classification` of the *scikit-learn* python library [31], with 6 features, 3 informative for the class prediction, and 3 more that are linear combinations of the formers. The rational of this toy domain is to craft two feature subspaces highly correlated.

All training sets are standardized, and the same transformation is then applied to the respective test sets. The partition of the feature space is defined by the first split $F = \lfloor d/2 \rfloor$, and its complement; features are taken in the order defined by the datasets.

Unless stated differently, models are trained with L_2 regularisation by *scikit-learn*'s `linear_model.LogisticRegression`. The hyper-parameter λ is optimized by 5-folds cross validation on the grid $\{10^{-5}, 10^{-4}, \dots, 10^{4}\}$.

7.3.2 Explanation of the movie

Along with this SM comes a movie displaying the impact on the p-value of DT, in the context of the decrease of the HSIC. The movie shows 200 iterations of DT on the Abalone2D domain, along with the modification of the point cloud (classes are red / blue). The p-value is indicated. Notice

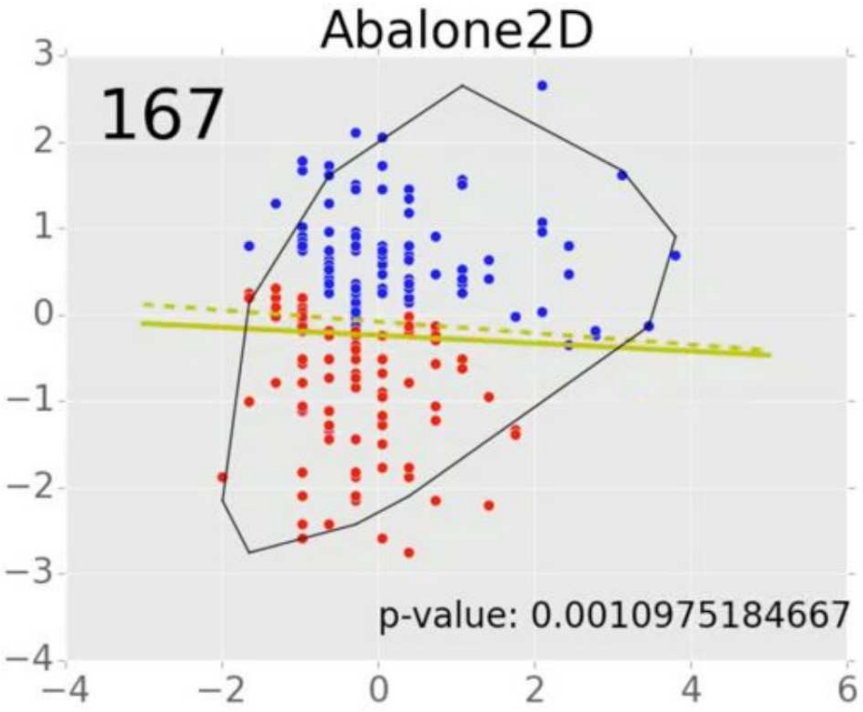
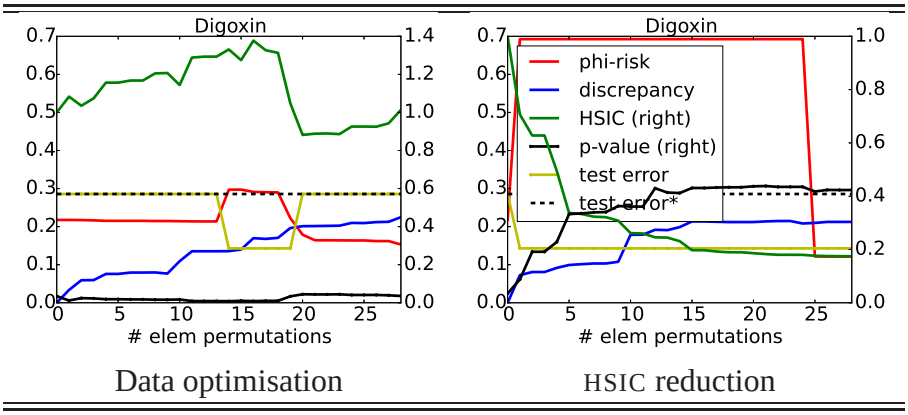


Figure 4: Crop of the movie (see text for explanation).

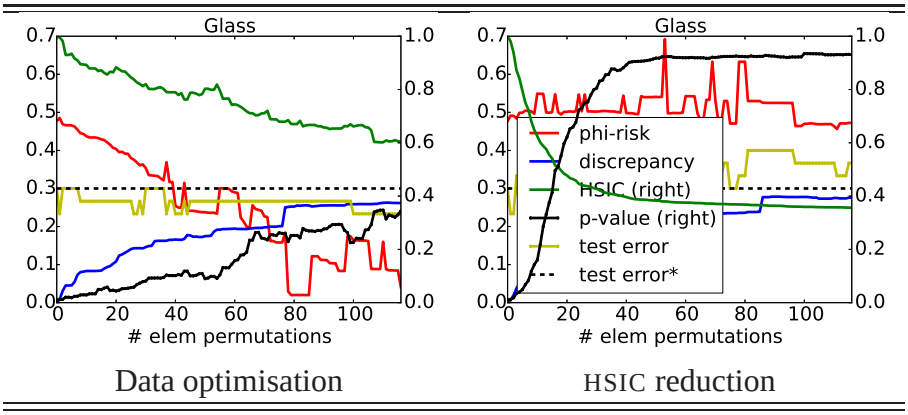
that is begins at value 0 up to 13 digits before DT starts. The two lines indicate the classifier learnt over the current data (plain gold line) and compare with the initial classifier learnt over the data before running DT (dashed gold line). The big number (167 in Figure 4) is the iteration number. Finally, the polygon displayed is the convex envelope of the initial data.

7.3.3 Complete experimental results



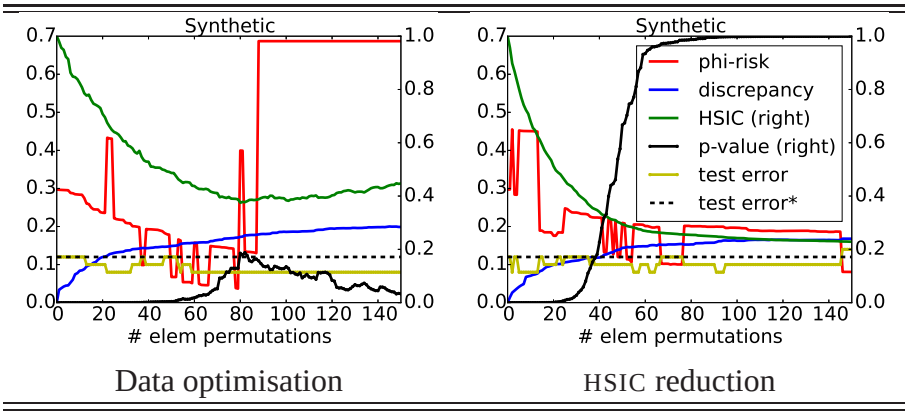
Data optimisation HSIC reduction

Table 4: Results on domain Digoxin. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. See text for details.



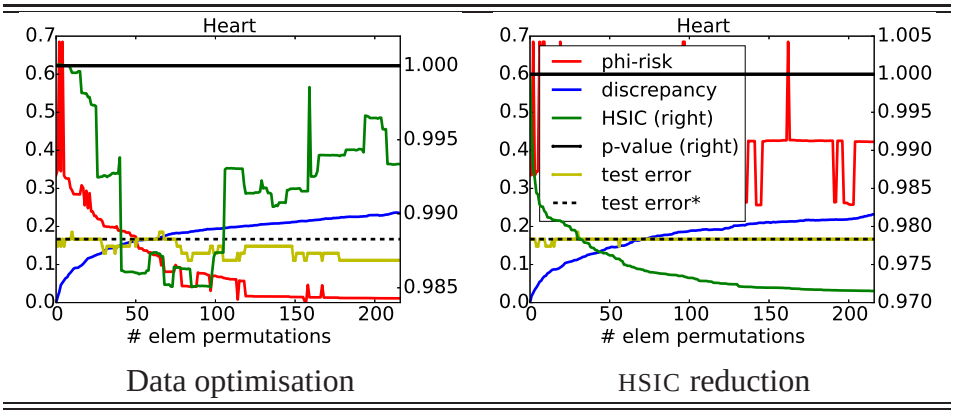
Data optimisation HSIC reduction

Table 5: Results on domain Glass. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. Color codes are the same on all plots. See text for details.



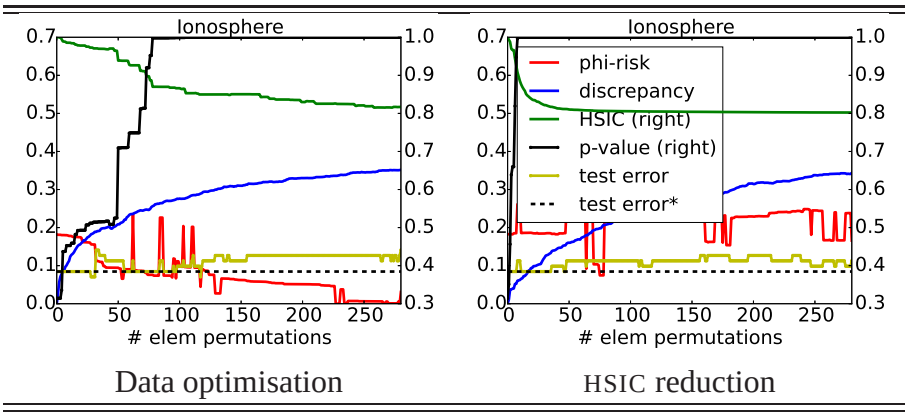
Data optimisation HSIC reduction

Table 6: Results on domain Synthetic. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. Color codes are the same on all plots. See text for details.



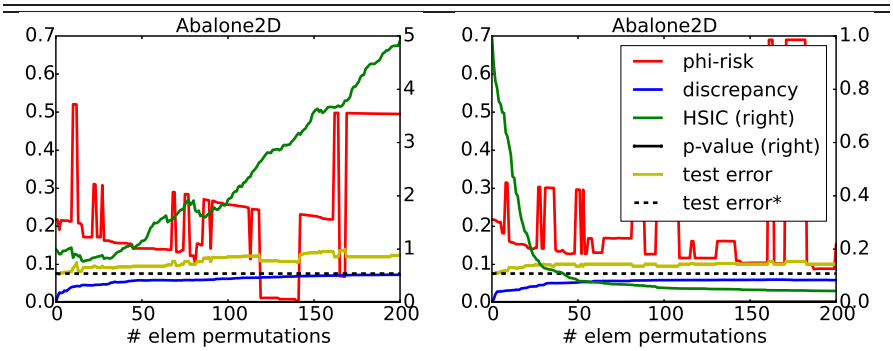
Data optimisation HSIC reduction

Table 7: Results on domain Heart. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. See text for details.



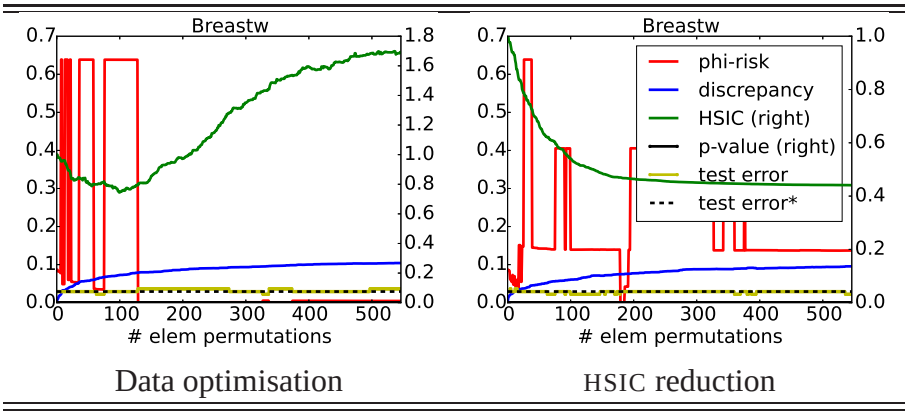
Data optimisation | HSIC reduction

Table 8: Results on domain Ionosphere. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. See text for details.



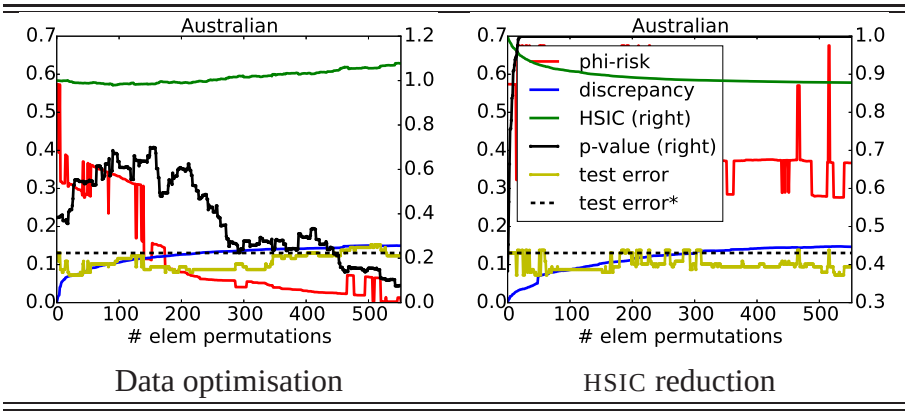
Data optimisation | HSIC reduction

Table 9: Results on domain Abalone. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. The scale of the p-value curve is not the same as in the main file: here, its scale is the same as for the HSIC curve, which explains why it seems to be flat while the value for the first iterations is the zero-machine and the values for the last exceed one per thousand. See text for details.



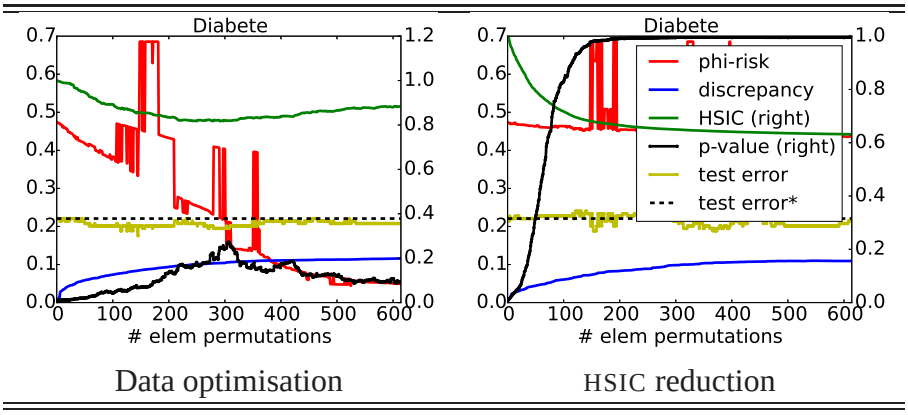
Data optimisation HSIC reduction

Table 10: Results on domain BreastWisc. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. See text for details.



Data optimisation HSIC reduction

Table 11: Results on domain Australian. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. See text for details.



Data optimisation HSIC reduction

Table 12: Results on domain Diabete_scale. Left: Data optimisation; right: HSIC reduction. Color codes are the same on all plots. See text for details.

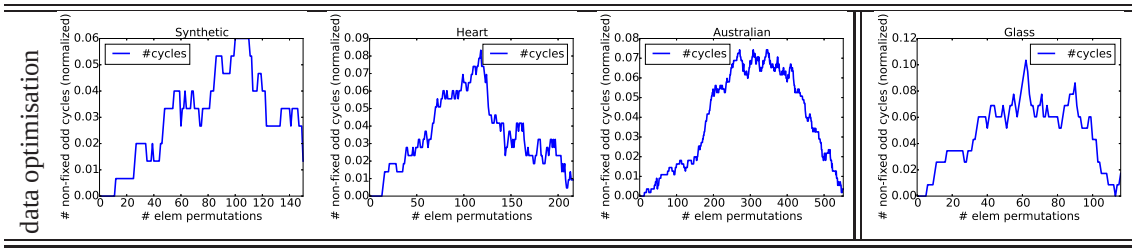


Table 13: Number of odd cycles (excluding fixed points, normalized by m) for the data optimization experiments in Table 1.

7.3.4 Comparisons of block-class vs arbitrary permutations

We now compare DT as in Algorithm 1 to the one where we relax the constraint that permutations must be block-class (implying the invariance of the mean operator). See Tables 14, 15. The results are a clear advocacy for the constraint, as relaxing it brings poor results, from both the φ-risk and test error standpoints.

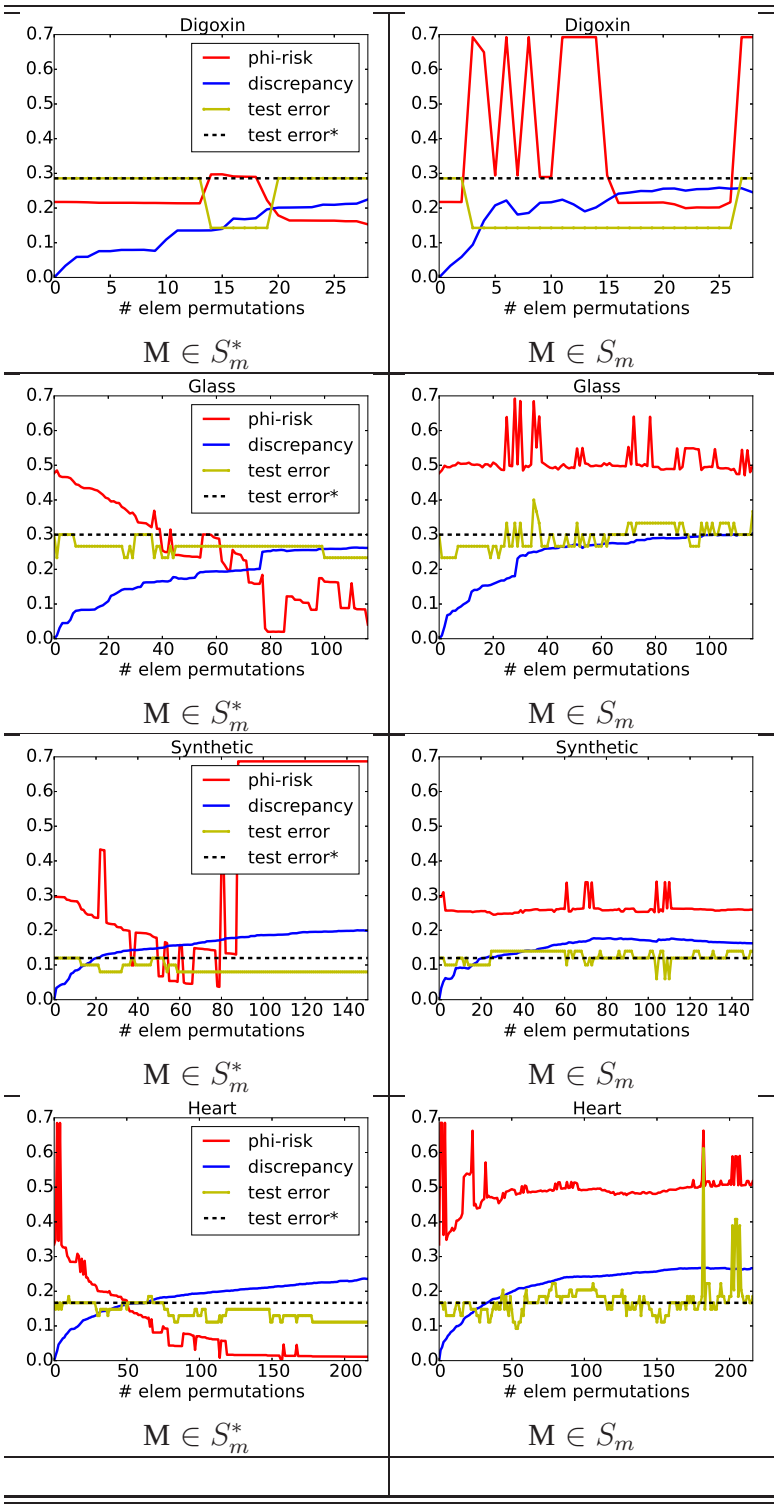


Table 14: Comparison, for data optimisation, of algorithm DT in which elementary permutation matrices are constrained to be block-class (left), and *not* constrained to be block-class (right).

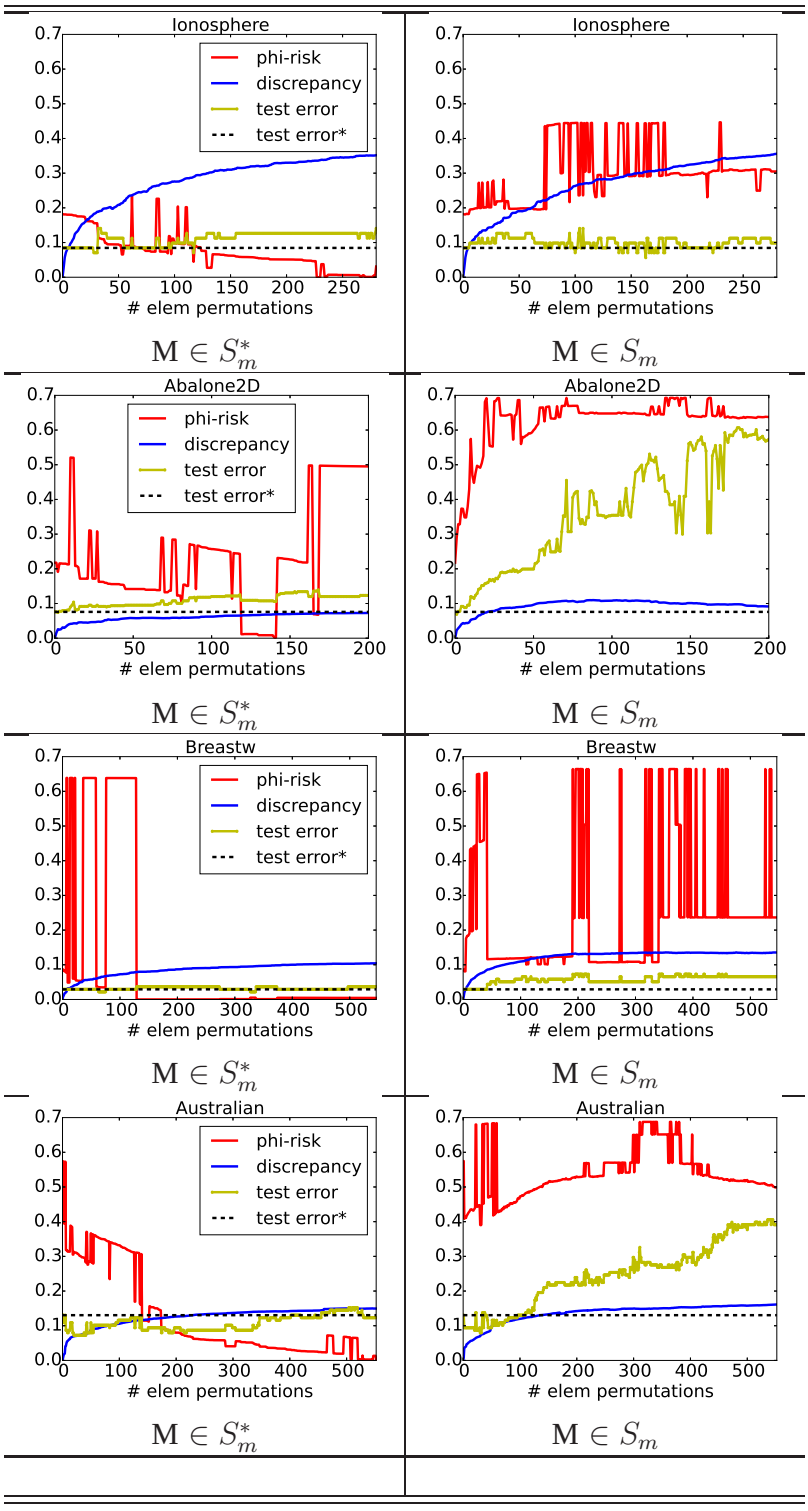


Table 15: Comparison (cont'd), for data optimisation, of algorithm DT in which elementary permutation matrices are constrained to be block-class (left), and *not* constrained to be block-class (right).