Supplementary to Domain Generalization via Invariant Feature Representation

Krikamol Muandet

Max Planck Institute for Intelligent Systems
Spemannstraße 38, 72076 Tübingen, Germany
krikamol@tuebingen.mpg.de

David Balduzzi

Department of Computer Science, ETH Zurich
Universitätstrasse 6, 8092 Zurich, Switzerland
david.balduzzi@inf.ethz.ch

Bernhard Schölkopf

Max Planck Institute for Intelligent Systems,
Spemannstraße 38, 72076 Tübingen, Germany
bs@tuebingen.mpg.de

1 Domain Generalization and Related Frameworks

The most fundamental assumption in machine learning is that the observations are independent and identically distributed (i.i.d.). That is, each observation comes from the same probability distribution as the others and all are mutually independent. However, this assumption is often violated in practice, in which case the standard machine learning algorithms do not perform well. In the past decades, many techniques have been proposed to tackle scenarios where there is a mismatch between training and test distributions. These include domain adaptation [Bickel et al., 2009], multitask learning [Caruana, 1997], transfer learning [Pan and Yang, 2010], covariate/dataset shift [Quionero-Candela et al., 2009] and concept drift [Widmer and Kurat, 1996]. To better understand domain generalization, we briefly discuss how it relates to some of these approaches.

Transfer learning (see e.g., Pan and Yang [2010] and references therein). Transfer learning aims at transferring knowledge from some previous tasks to a target task when the latter has limited training data. That is, although there may be few labeled examples, "knowledge" obtained in related tasks may be available. Transfer learning focuses on improving the learning of the target

Table 1: Comparison of domain generalization with other well-known frameworks. Note that the domain generalization is closely related to multi-task learning and domain adaptation. The difference of domain generalization is that one does not observe the target domains in which a classifier will be applied without retraining the classifier.

Framework	Distribution Mismatch	Multiple Sources	Target Domain
Standard Setup	✗	✗	✗
Transfer Learning	✓	✗	✓
Multi-task Learning	✓	✓	✗
Domain Adaptation	✓	✓	✓
Domain Generalization	✓	✓	✗

predictive function using the knowledge in the source task. Although not identical, domain generalization can be viewed as a transfer learning when knowledge of the target task is unavailable during training.

Multitask learning (see e.g., Caruana [1997] and references therein). The goal of multitask learning is to learn multiple tasks simultaneously – especially when training examples in each task are scarce. By learning all tasks simultaneously, one expects to improve generalization on individual tasks. An important assumption is therefore that all the tasks are related. Multitask learning differs from domain generalization because learning the new task often requires retraining.

Domain adaptation (see e.g., Bickel et al. [2009] and references therein). Domain adaptation, also known as covariate shift, deals primarily with a mismatch between training and test distributions. Domain generalization deals with a broader setting where training instances may have been collected from multiple source domains. A second difference is that in domain adaptation one observes the target domain during the training time whereas in domain generalization one does not.

Table 1 summarizes the main differences between the various frameworks.

2 Proof of Theorem 1

Lemma 6. *Given a set of distributions* $\mathcal{P} = \{\mathbb{P}^1, \mathbb{P}^2 \ldots, \mathbb{P}^N\}$*, the distributional variance of* $\mathcal{P}$ *is* $\mathbb{V}_{\mathcal{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N} \|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2$ *where* $\mu_{\bar{\mathbb{P}}} = (1/N)\sum_{i=1}^{N} \mu_{\mathbb{P}^i}$ *and* $\bar{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}^i$*.*

Proof. Let $\bar{\mathbb{P}}$ be the probability distribution defined as $(1/N)\sum_{i=1}^{N} \mathbb{P}^i$, i.e., $\bar{\mathbb{P}}(x) = (1/N)\sum_{i=1}^{N} \mathbb{P}^i(x)$. It follows from the linearity of the expectation that $\mu_{\bar{\mathbb{P}}} = (1/N)\sum_{i=1}^{N} \mu_{\mathbb{P}^i}$. For brevity, we will denote $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ by $\langle \cdot, \cdot \rangle$. Then, expanding (3) gives

$$\begin{aligned} \mathbb{V}_{\mathcal{H}}(\mathcal{P}) &= \frac{1}{N}\mathrm{tr}(\Sigma) = \frac{1}{N}\mathrm{tr}(G) - \frac{1}{N^2}\sum_{i,j=1}^{N} G_{ij} \\ &= \frac{1}{N}\sum_{i=1}^{N}\langle \mu_{\mathbb{P}^i}, \mu_{\mathbb{P}^i}\rangle - \frac{1}{N^2}\sum_{i,j=1}^{N}\langle \mu_{\mathbb{P}^i}, \mu_{\mathbb{P}^j}\rangle \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\left[\sum_{i=1}^{N}\langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^i}\rangle - \frac{2}{N}\sum_{i,j=1}^{N}\langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^j}\rangle + \frac{1}{N}\sum_{i,j=1}^{N}\langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^j}\rangle\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^i}\rangle - 2\sum_{i=1}^{N}\left\langle\mu_{\mathbb{P}^i},\frac{1}{N}\sum_{j=1}^{N}\mu_{\mathbb{P}^j}\right\rangle + N\left\langle\frac{1}{N}\sum_{i=1}^{N}\mu_{\mathbb{P}^i},\frac{1}{N}\sum_{j=1}^{N}\mu_{\mathbb{P}^j}\right\rangle\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^i}\rangle - 2\sum_{i=1}^{N}\langle\mu_{\mathbb{P}^i},\mu_{\bar{\mathbb{P}}}\rangle + N\langle\mu_{\bar{\mathbb{P}}},\mu_{\bar{\mathbb{P}}}\rangle\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(\langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^i}\rangle - 2\cdot\langle\mu_{\mathbb{P}^i},\mu_{\bar{\mathbb{P}}}\rangle + \langle\mu_{\bar{\mathbb{P}}},\mu_{\bar{\mathbb{P}}}\rangle\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N}\|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2 \ ,
\end{aligned}
$$

which completes the proof. ■

Theorem 1 *For a characteristic kernel k, $\mathbb{V}_{\mathcal{H}}(\mathcal{P}) = 0$ if and only if $\mathbb{P}^1 = \mathbb{P}^2 = \cdots = \mathbb{P}^N$.*

Proof. Since k is characteristic, $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2$ is a metric and is zero iff $\mathbb{P} = \mathbb{Q}$ for any distributions $\mathbb{P}$ and $\mathbb{Q}$ [Sriperumbudur et al., 2010]. By Lemma 6, $\mathbb{V}_{\mathcal{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N}\|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2$. Thus, $\|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2 = 0$ iff $\mathbb{P}^i = \bar{\mathbb{P}}$. Consequently, if $\mathbb{V}_{\mathcal{H}}(\mathcal{P})$ is zero, this implies that $\mathbb{P}^i = \bar{\mathbb{P}}$ for all i, meaning that $\mathbb{P}^1 = \cdots = \mathbb{P}^\ell$. Conversely, if $\mathbb{P}^1 = \cdots = \mathbb{P}^\ell$, then $\|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2 = 0$ is zero for all i and thereby $\mathbb{V}_{\mathcal{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N}\|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2$ is zero. ■

3 Proof of Theorem 2

Theorem 2 *The empirical estimator $\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{\Sigma}) = \mathrm{tr}(KQ)$ obtained from Gram matrix*

$$
\widehat{G}_{ij} := \frac{1}{n_i \cdot n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} k(x_k^{(i)}, x_l^{(j)})
$$

is a consistent estimator of $\mathbb{V}_{\mathcal{H}}(\mathcal{P})$.

Proof. Recall that

$$
\mathbb{V}_{\mathcal{H}}(\mathcal{P}) = \frac{1}{N}\mathrm{tr}(G) - \frac{1}{N^2}\sum_{i,j=1}^{N} G_{ij} \quad \text{and} \quad \widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{G}) - \frac{1}{N^2}\sum_{i,j=1}^{N}\widehat{G}_{ij}
$$

where

$$
\begin{aligned}
G_{ij} &= \langle\mu_{\mathbb{P}^i},\mu_{\mathbb{P}^j}\rangle_{\mathcal{H}} = \iint k(x,z)\,\mathrm{d}\mathbb{P}^i(x)\,\mathrm{d}\mathbb{P}^j(z) \\
\widehat{G}_{ij} &= \langle\hat{\mu}_{\mathbb{P}^i},\hat{\mu}_{\mathbb{P}^j}\rangle_{\mathcal{H}} = \frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} k(x_k^{(i)}, x_l^{(j)})
\end{aligned}
$$

By Theorem 15 in Altun and Smola [2006], we have a fast convergence of $\hat{\mu}_{\mathbb{P}}$ to $\mu_{\mathbb{P}}$. Consequently, we have $\widehat{G} \to G$, which implies that $\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{S}) \to \mathbb{V}_{\mathcal{H}}(\mathcal{P})$. Hence, $\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{S})$ is a consistent estimator of $\mathbb{V}_{\mathcal{H}}(\mathcal{P})$. ■

4 Derivation of Eq. (8)

DICA employs the covariance of inverse regressor $\mathbb{V}(\mathbb{E}[\phi(X)|Y])$, which can be written in terms of covariance operators. Let $\mathcal{H}$ and $\mathcal{F}$ be the RKHSes of X and Y endowed with reproducing kernels k and l, respectively. Let Σ_{xx}, Σ_{yy}, Σ_{xy}, and Σ_{yx} be the covariance operators in and between the corresponding RKHSes of X and Y. We define the conditional covariance operator of X given Y, denoted by $\Sigma_{\mathrm{xx|y}}$, as

$$\Sigma_{\mathrm{xx|y}} \triangleq \Sigma_{\mathrm{xx}} - \Sigma_{\mathrm{xy}}\Sigma_{\mathrm{yy}}^{-1}\Sigma_{\mathrm{yx}} \ . \tag{1}$$

The following theorem from Fukumizu et al. [2004] states that, under mild conditions, $\Sigma_{\mathrm{xx|y}}$ equals the expected conditional variance of $\phi(X)$ given Y.

Theorem 7. *For any $f \in \mathcal{H}$, if there exists $g \in \mathcal{F}$ such that $\mathbb{E}[f(X)|Y] = g(Y)$ for almost every Y, then $\Sigma_{\mathrm{xx|y}} = \mathbb{E}[\mathbb{V}(\phi(X)|Y)]$.*

Using the *E-V-V-E* identity[1], the covariance $\mathbb{V}(\mathbb{E}[\phi(X)|Y])$ can be expressed in terms of the conditional covariance operators as follow:

$$\mathbb{V}(\mathbb{E}[\phi(X)|Y]) = \mathbb{V}(\phi(X)) - \mathbb{E}[\mathbb{V}(\phi(X)|Y)], \tag{2}$$

assuming that the inverse regressor $\mathbb{E}[f(x)|y]$ is a smooth function of y for any $f \in \mathcal{H}$.

By virtue of Theorem 7, the second term in the r.h.s. of (2) is $\Sigma_{\mathrm{xx|y}}$. Since $\mathbb{V}(\phi(X)) = \mathrm{Cov}(\phi(x), \phi(x)) = \Sigma_{\mathrm{xx}}$, it follows from (1) that the covariance of the inverse regression $\mathbb{V}(\mathbb{E}[\phi(X)]|Y)$ can be expressed as

$$\mathbb{V}(\mathbb{E}[\phi(X)|Y]) = \Sigma_{\mathrm{xy}}\Sigma_{\mathrm{yy}}^{-1}\Sigma_{\mathrm{yx}} \ . \tag{3}$$

The covariance (3) can be estimated from finite samples $(x_1, y_1), \ldots, (x_n, y_n)$ by $\widehat{\mathbb{V}}(\mathbb{E}[\phi(X)|Y]) = \widehat{\Sigma}_{\mathrm{xy}}\widehat{\Sigma}_{\mathrm{yy}}^{-1}\widehat{\Sigma}_{\mathrm{yx}}$ where $\widehat{\Sigma}_{\mathrm{xy}} = \frac{1}{n}\Phi_x\Phi_y^\top$ and $\Phi_x = [\phi(x_1), \ldots, \phi(x_n)]$ and $\Phi_y = [\varphi(y_1), \ldots, \varphi(y_n)]$. Let K and L denote the kernel matrices computed over samples $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$, respectively. We have

$$\begin{aligned}
\widehat{\mathbb{V}}(\mathbb{E}[\phi(X)|Y]) &= \left(\frac{1}{n}\Phi_x\Phi_y^\top\right)\left(\frac{1}{n}(\Phi_y\Phi_y^\top + n\varepsilon\mathcal{I})\right)^{-1}\left(\frac{1}{n}\Phi_y\Phi_x^\top\right) \\
&= \frac{1}{n}\Phi_x\Phi_y^\top\Phi_y\left(\Phi_y^\top\Phi_y + n\varepsilon I_n\right)^{-1}\Phi_x^\top \\
&= \frac{1}{n}\Phi_x L\left(L + n\varepsilon I_n\right)^{-1}\Phi_x^\top
\end{aligned} \tag{4}$$

where $L = \Phi_y^\top\Phi_y$ and $\mathcal{I}$ is the identity operator. The second equation is obtained by applying the fact that $(\Phi_y\Phi_y^\top + n\varepsilon\mathcal{I})\Phi_y = \Phi_y(\Phi_y^\top\Phi_y + n\varepsilon I_n)$.

[1] $\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|Y)] + \mathbb{V}(\mathbb{E}[X|Y])$ for any X, Y.

Finally, using $\widehat{\Sigma}_{\text{xx}} = \frac{1}{n}\Phi_x\Phi_x^\top$ and recalling that $K = \Phi_x^\top\Phi_x$, we obtain

$$\begin{aligned}
\mathbf{b}_k^\top \widehat{\Sigma}_{\text{xx}}^{-1}\widehat{\mathbb{V}}(\mathbb{E}[X|Y])\widehat{\Sigma}_{\text{xx}}\mathbf{b}_k &= \mathbf{b}_k^\top \left(\frac{1}{n}\Phi_x\Phi_x^\top\right)^{-1}\left(\frac{1}{n}\Phi_x L\left(L+n\varepsilon I_n\right)^{-1}\Phi_x^\top\right)\left(\frac{1}{n}\Phi_x\Phi_x^\top\right)\mathbf{b}_k \\
&= \frac{1}{n}\boldsymbol{\beta}_k^\top\Phi_x^\top\left(\Phi_x\Phi_x^\top\right)^{-1}\Phi_x L\left(L+n\varepsilon I_n\right)^{-1}\Phi_x^\top\left(\Phi_x\Phi_x^\top\right)\Phi_x\boldsymbol{\beta}_k \\
&= \frac{1}{n}\boldsymbol{\beta}_k^\top\Phi_x^\top\Phi_x\left(\Phi_x^\top\Phi_x\right)^{-1}L\left(L+n\varepsilon I_n\right)^{-1}\Phi_x^\top\left(\Phi_x\Phi_x^\top\right)\Phi_x\boldsymbol{\beta}_k \\
&= \frac{1}{n}\boldsymbol{\beta}_k^\top L(L+n\varepsilon I)^{-1}K^2\boldsymbol{\beta}_k
\end{aligned}$$

and

$$\mathbf{b}_k^\top\mathbf{b}_k = \boldsymbol{\beta}_k^\top\Phi_x^\top\Phi_x\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^\top K\boldsymbol{\beta}_k$$

as desired.

5 Derivation of Lagrangian (10)

Observe that optimization

$$\max_{B\in\mathbb{R}^{n\times m}} \frac{\operatorname{tr}\left(B^\top XB\right)}{\operatorname{tr}\left(B^\top YB\right)} \tag{5}$$

is invariant to rescaling $B \mapsto \alpha\cdot B$. Optimization (5) is therefore equivalent to

$$\begin{aligned}
&\max_{B\in\mathbb{R}^{n\times m}} \operatorname{tr}\left(B^\top XB\right) \\
&\text{subject to: } \operatorname{tr}\left(B^\top YB\right) = 1,
\end{aligned}$$

which yields Lagrangian

$$\mathcal{L} = \operatorname{tr}\left(B^\top XB\right) - \operatorname{tr}\left(\left(B^\top YB - I\right)\Gamma\right). \tag{6}$$

6 Proof of Theorem 5

We consider a scenario where distributions $\mathbb{P}^i$ are drawn according to $\mathscr{P}^*$ with probability μ_i. Introduce shorthand $\tilde{X}_{ij}$ for $(\mathbb{P}^{(i)}, X_{ij})$ for a distribution on $\mathfrak{P}_{\mathcal{X}}$ and a corresponding random variable on $\mathcal{X}$.

The quantity of interest is the difference between the expected and empirical loss of a classifier $f : \mathfrak{P}_{\mathcal{X}}\times\mathcal{X}\to\mathcal{Y}$ under loss function $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_+$.

Assumptions. The loss function $\ell : \mathbb{R}\times\mathcal{Y}\to\mathbb{R}_+$ is ϕ_ℓ-Lipschitz in its first variable and bounded by U_ℓ. The kernel $k_{\mathcal{X}}$ is bounded by $U_{\mathcal{X}}$. Assume that all distributions in $\mathscr{P}^*$ are mapped into a ball of size $U_{\mathfrak{P}}$ by $\Psi_{\mathfrak{P}}$. Finally, since $k_{\mathfrak{P}}$ is a is a square exponential, there is a constant $L_{\mathfrak{P}}$ such that

$$\|\Phi_{\mathfrak{P}}(v) - \Phi_{\mathfrak{P}}(w)\| \le L_{\mathfrak{P}}\|v-w\| \text{ for all } v, w.$$

Recall that N is the number of sampled domains, n_i is the number of samples in domain i, and $n = \sum_{i=1}^N n_i$ is the total number of samples. The proof assumes $n_i = n_j$ for all i, j.

Theorem 5. *Assumes the conditions above hold. Then with probability at least* $1-\delta$

$$\sup_{\|f\|_{\mathcal{H}}\leq 1}\left|\mathbb{E}^*_{\mathscr{P}}\mathbb{E}_{\mathbb{P}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)-\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)\right|^2$$
$$\leq c_1\frac{1}{N}tr(B^\intercal KLKB)+tr(B^\top KB)\left(c_2\frac{N\cdot(\log\delta^{-1}+2\log N)}{n}+c_3\frac{\log\delta^{-1}}{N}+\frac{c_4}{N}\right).$$

Remark 1. *Recall that* $\Phi_x=[\phi(x_1),\ldots,\phi(x_n)]$. *The composition* $x_t\mapsto \mathbf{k}_t\cdot B$, *where* $\mathbf{k}_t=[k(x_1,x_t),\ldots,k(x_n,x_t)]$, *can therefore be rewritten as* $\phi(x_t)\cdot\mathcal{B}=\phi(x_t)\cdot\Phi_x\cdot B$.

Proof. The proof modifies the approach taken in Blanchard et al. [2011] to handle the preprocessing via transform $\mathcal{B}$, and the fact that we work with *squared* errors. Parts of the proof that pass through largely unchanged are omitted.

We repeatedly apply the inequality $|a+b|^2\leq 2|a|^2+2|b|^2$. However, we only incur the multiplication-by-2 penalty once since $|a_1+\cdots+a_n|^2\leq 2|a_1|^2+\cdots+2|a_n|^2$.

Decompose

$$\begin{aligned}
&\sup_{\|f\|_{\mathcal{H}}\leq 1}\left|\mathbb{E}^*_{\mathscr{P}}\mathbb{E}_{\mathbb{P}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)-\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)\right|^2\\
&\leq \sup_{\|f\|_{\mathcal{H}}\leq 1}\frac{2}{N}\sum_{i=1}^N\left|\mathbb{E}^*_{\mathscr{P}}\mathbb{E}_{\mathbb{P}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)-\mathbb{E}_{\mathbb{P}^i}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)\right|^2\\
&+\sup_{\|f\|_{\mathcal{H}}\leq 1}\frac{2}{N}\sum_{i=1}^N\left|\mathbb{E}_{\mathbb{P}^i}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)-\mathbb{E}_{\widehat{\mathbb{P}}^i}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)\right|^2\\
&+\sup_{\|f\|_{\mathcal{H}}\leq 1}\frac{2}{N}\sum_{i=1}^N\left|\mathbb{E}_{\widehat{\mathbb{P}}^i}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)-\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)\right|^2\\
&=(A)+(B)+(C)\ .
\end{aligned}$$

Control of (C):

$$\begin{aligned}
(C)&=\sup_{\|f\|_{\mathcal{H}}\leq 1}\frac{2}{N}\sum_{i=1}^N\left|\mathbb{E}_{\widehat{\mathbb{P}}^i}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)-\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}\mathcal{B}),Y_i)\right|^2\\
&\leq\phi_\ell^2\sup_{\|f\|_{\mathcal{H}}\leq 1}\frac{2}{N}\sum_{i=1}^N\left|\mathbb{E}_{\widehat{\mathbb{P}}^i}f(\tilde{X}_{ij}\mathcal{B})-\mathbb{E}_{\hat{\mathbb{P}}}f(\tilde{X}_{ij}\mathcal{B})\right|^2\\
&=\phi_\ell^2\cdot\frac{2}{N}\sum_{i=1}^N\left\|\Psi_{\mathfrak{P}}(\widehat{\mathbb{P}^i})\otimes\mu_{\widehat{\mathbb{P}}^i}\mathcal{B}-\Psi_{\mathfrak{P}}(\widehat{\mathbb{P}})\otimes\mu_{\widehat{\mathbb{P}}}\mathcal{B}\right\|^2
\end{aligned}$$

Note that $\|\Psi_{\mathfrak{P}}(\mu(\mathbb{P}))\|^2\leq L_{\mathfrak{P}}\cdot\|\mu_{\mathbb{P}}\|^2\leq L_{\mathfrak{P}}U_{\mathfrak{P}}$. Therefore,

$$(C)\leq\phi_\ell^2L_{\mathfrak{P}}U_{\mathfrak{P}}\frac{2}{N}\sum_{i=1}^N\left\|\mu_{\widehat{\mathbb{P}}^i}\mathcal{B}-\mu_{\widehat{\mathbb{P}}}\mathcal{B}\right\|^2.$$

By the proof of Theorem 1 and since $\Phi_x^\top \mathcal{B} = KB$, we have

$$(C) \leq 2\phi_\ell^2 L_{\mathfrak{P}} U_{\mathfrak{P}} \frac{1}{N} \mathrm{tr}(KBB^\intercal KL).$$

Control of (B): Similarly,

$$\begin{aligned}(B) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \frac{2}{N} \sum_{i=1}^{N} \left| \mathbb{E}_{\mathbb{P}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\widehat{\mathbb{P}}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\ &\leq 2\phi_\ell^2 L_{\mathfrak{P}} U_{\mathfrak{P}} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| \mu_{\mathbb{P}^i}\mathcal{B} - \mu_{\widehat{\mathbb{P}}^i}\mathcal{B} \right\|^2 \\ &\leq 2\phi_\ell^2 L_{\mathfrak{P}} U_{\mathfrak{P}} \cdot \|\mathcal{B}\|_{HS}^2 \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| \mu_{\mathbb{P}^i} - \mu_{\widehat{\mathbb{P}}^i} \right\|^2\end{aligned}$$

Here we follow the strategy applied by Blanchard et al. [2011] to control their term (I) in Theorem 5.1. Assume $n_i = n_j$ for all i, j and recall $n = \sum_{i=1}^{N} n_i$ so $n_i = n/N$ for all i.

By Hoeffding's inequality in Hilbert space, with probability greater than $1 - \delta$ the following inequality holds

$$\left\| \frac{1}{n_i} \sum_{j=1}^{n_i} \mu(\hat{X}_{ij}) - \mathbb{E}_{\mathbb{P}^{(i)}} \mu(X_{ij}) \right\|^2 \leq 9U_{\mathcal{X}} \frac{N \cdot \log 2\delta^{-1}}{n}.$$

Applying the union bound obtains

$$(Ib) \leq 18\phi_\ell^2 L_{\mathfrak{P}} U_{\mathfrak{P}} U_{\mathcal{X}} \cdot \|\mathcal{B}\|_{HS}^2 \cdot \frac{N \cdot (\log \delta^{-1} + 2 \log N)}{n}.$$

Control of (A):

$$(A) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \frac{2}{N} \sum_{i=1}^{N} \left| \mathbb{E}^*_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\mathbb{P}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2$$

Following the strategy used by Blanchard et al. [2011] to control (II) in Theorem 5.1, we obtain

$$(A) \leq c_3 \frac{\phi_\ell^2 U_{\mathcal{X}}^2 U_{\mathfrak{P}} + U_\ell \log \delta^{-1}}{N} \cdot \|\mathcal{B}\|_{HS}^2.$$

End of proof: We have that K is invertible since $\widehat{\Sigma}_{xx}$ is assumed to be invertible. It follows that the trace $\mathrm{tr}(B^\top KB)$ defines a norm which coincides with the Hilbert-Schmidt norm $\|\mathcal{B}\|_{HS}^2$. Combining the three inequalities above concludes the proof. ■

7 Leave-one-out accuracy

Figure 1 depicts the leave-one-out accuracies of different approaches evaluated on each subject in the dataset. Average leave-one-out accuracies are reported in Table 2. The distributional SVM outperforms the pooling SVM in this setting, possibly because of the relatively large number of training subjects, i.e., 29 subjects. Using the invariant features learnt by DICA also gives higher accuracies than other approaches.

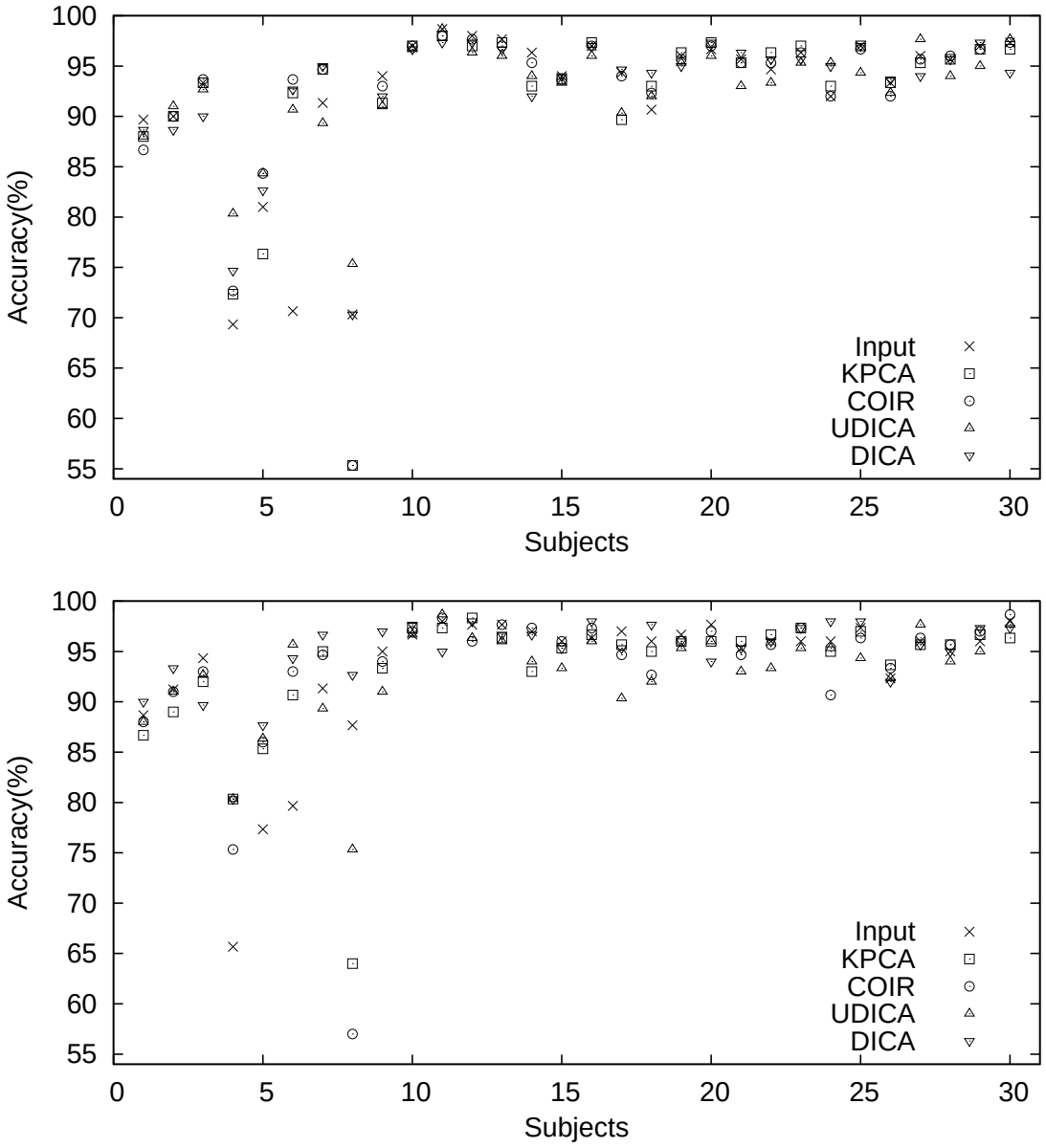

Figure 1: The leave-one-out accuracy of different methods evaluated on each subject in the GvHD dataset. The top figure depicts the pooling setting, whereas the bottom figure depicts the distributional setting.

References

Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proc. of Conf. on Learning Theory (COLT)*, 2006.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, Dec. 2009. ISSN 1532-4435.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems 24*, pages 2178–2186, 2011.

R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

G. Widmer and M. Kurat. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23:69–101, 1996.