7 Appendix

In this appendix we provide proofs for some theoretical results:

7.1 Proofs of Theorem 1 and Proposition 1

1. Proof for Theorem 1

Proof: *If the conditions in the theorem hold true, we can rewrite the expected loss as follows:*

$$\begin{aligned} & l_{P_{XY}}[f] \triangleq \mathbb{E}_{P_{XY}}[l(f(X), Y)] \\ & = \int l(f(x), y) p(y|x) p(x) \, dx \, dy \\ & = \int l_h(\tilde{f}(h(x)), y) p(y|x, h(x)) p(x|h(x)) p(h(x)) \, dh \, dx \, dy \\ & = \int l_h(\tilde{f}(h(x)), y) p(y|h(x)) p(h(x)) \, dh \, dy. \end{aligned}$$

Since $p^{te}(y|x) = p^{tr}(y|x)$, implied by the covariate shift setting, and $p(y|h(x)) = p(y|h(x), x) = p(y|x)$, we have $p^{te}(y|h(x)) = p^{tr}(y|h(x))$. Let $\beta(h) \triangleq \frac{p^{te}(h(x))}{p^{tr}(h(x))}$, and the expected loss in the target domain, $l_{P^{te}_{XY}}[f]$, further becomes

$$\begin{aligned} l_{P^{te}_{XY}}[f] &= \int l_h(\tilde{f}(h(x)), y) p^{te}(y|h(x)) p^{te}(h(x)) \beta(h) \, dh \, dy \\ &= \int l_h(\tilde{f}(h(x)), y) p^{tr}(y|h(x)) p^{tr}(h(x)) \beta(h) \, dh \, dy \\ &= \mathbb{E}_{P^{tr}_{XY}}[l_h(\tilde{f}(h(x)), y) \beta(h)]. \end{aligned}$$

□

2. We also present a proof for Proposition 1

Proof: *For the case when Y is binary, we have: $p(Y = 1 \,|\, X) = p(Y = 1 \,|\, X, h(X)) = h(X) = p(Y = 1 \,|\, h(X))$, and similarly $p(Y = 0 \,|\, X) = p(Y = 0 \,|\, X, h(X)) = 1 - h(X) = p(Y = 0 \,|\, h(X))$. For the case when Y is continuous, it follows trivially because $f(X) = \mathbb{E}[Y|X]$, and $Y \perp\!\!\!\perp f(X) | f(X)$, which implies that $Y \perp\!\!\!\perp X | f(X)$ due to the fact that X and ϵ are independent.* □

7.2 Proof of Theorem 2

We shall first introduce some relevant quantities:

- Expected risk in target domain $R^{te}(l) = \mathbb{E}_{y|x}[\frac{1}{n_{te}} \sum_{i=1}^{n_{te}} l(x_i^{te}, y_i^{te}, \theta)]$, the optimal function $l^* = \arg\min_{l \in \mathcal{G}} R^{te}(l)$
- Expected risk in the target domain with projected features $R^{te}_W(l) = \mathbb{E}_{y|x}[\frac{1}{n_{te}} \sum_{i=1}^{n_{te}} l(W^\intercal x_i^{te}, y_i^{te}, \theta)]$, the optimal function $l^*_W = \arg\min_{l \in \mathcal{G}} R^{te}_W(l)$
- Expected risk in the weighted source domain with original features $R^{tr}_\beta(l) = \mathbb{E}_{y|x}[\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_i l(x_i^{tr}, y_i^{tr}, \theta)]$, the estimated function $l^*_\beta = \arg\min_{l \in \mathcal{G}} R^{tr}_\beta(l)$
- Expected risk in the weighted source domain with projected features $R^{tr}_{\beta_W}(l) = \mathbb{E}_{y|x}[\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_{W,i} l(W^\intercal x_i^{tr}, y_i^{tr}, \theta)]$, the estimated function $l^*_{\beta_W} = \arg\min_{l \in \mathcal{G}} R^{tr}_{\beta_W}(l)$
- Empirical risk in the weighted source domain with true weights $\hat{R}^{tr}_\beta(l) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_i l(x_i^{tr}, y_i^{tr}, \theta)$, the estimated function $l_\beta = \arg\min_{l \in \mathcal{G}} \hat{R}^{tr}_\beta(l)$
- Empirical risk in the weighted source domain with projected features and true weights $\hat{R}^{tr}_{\beta_W}(l) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_{W,i} l(W^\intercal x_i^{tr}, y_i^{tr}, \theta)$, the estimated function $l_{\beta_W} = \arg\min_{l \in \mathcal{G}} \hat{R}^{tr}_{\beta_W}(l)$

- Empirical risk in the weighted source domain with original features and estimated weights $\hat{R}^{tr}_{\hat{\beta}}(l) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \hat{\beta}_i l(x_i^{tr}, y_i^{tr}, \theta)$, the estimated function $l_{\hat{\beta}} = \arg\min_{l \in \mathcal{G}} \hat{R}^{tr}_{\hat{\beta}}(l)$
- Empirical risk in the weighted source domain with projected features and estimated weights $\hat{R}^{tr}_{\hat{\beta}_W}(l) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \hat{\beta}_{W,i} l(W^\intercal x_i^{tr}, y_i^{tr}, \theta)$, the estimated function $l_{\hat{\beta}_W} = \arg\min_{l \in \mathcal{G}} \hat{R}^{tr}_{\hat{\beta}_W}(l)$

Before we proceed to the proof of Theorem 2, we present some lemmata that are required to analyze the generalization in the target domain. We we also need a general variant of **A3**:

A3': The features of X, given by $X_1, ..., X_D$ are independent.

Both of them are proven assuming we have features X, in with D dimensions.

We first introduce a lemma which is a modification of the result by Gretton et al., in which we analyze the impact of dimensionality on the variance of the empirical weighted risk in the source domain, under the assumptions made in this study. Please note that we prove the lemmata assuming an input feature space of D, regardless if it is an original feature space of observations or a result of a transformation.

Lemma 1 (Adapted from Corollary 1.9 in Gretton et al): *Assume **A1**, **A2** and **A3'** hold. Then the following bound on the reweighted risk in the source domain holds with probability at least* $1 - \delta$:

$$\sup_{l \in \mathcal{G}} |\hat{R}^{tr}_{\beta}(l) - R^{te}(l)| = |\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_i l(x_i^{tr}, y_i^{tr}, \theta) - \mathbb{E}_{Y|X}[\frac{1}{n_{te}} \sum_{i=1}^{n_{tr}} l(x_i^{te}, y_i^{te}, \theta)]| \leq \tag{5}$$

$$\frac{(2 + \sqrt{4 \log(1/\delta)}) C U^D}{\frac{n_{tr}}{\sqrt{Q^D}}} + C(1 + \sqrt{4 \log(1/\delta)}) U^D \sqrt{T^{2D}/n_{tr} + 1/n_{te}} \tag{6}$$

Proof: The proof will examine the constants R, B and provide an upper bound on $||\beta||^2$, which can then be used to modify the bound in the corollary by Gretton et al. Assume for now that all the dimensions of X are independent: $p(x) = \prod_{i=1}^{D} p(x^{(i)})$. This means that the weights on all the features $\beta(x) = \beta(x^{(1)})\beta(x^{(2)})...\beta(x^{(D)})$. Therefore, the squared norm of the weights is: $||\beta(x)||_2^2 = \int \beta(x)^2 dx = \int \prod_{i=1}^{D} \beta(x^{(i)})^2 dx = \prod_{i=1}^{D} \int \beta(x^{(i)})^2 dx = \prod_{i=1}^{D} ||\beta(x^{(i)})||_2^2 \leq Q^D$ where the third equality is due to the independence of the features.

Furthermore, from the assumptions on the feature transform Φ and its corresponding reproducing kernel k, we see that $||\Phi(X_i)||^2 \leq R^2 \implies k(x_i, x_i) \leq R^2 \; \forall i \in \{1, ..., D\} \implies k(\mathbf{x}, \mathbf{x}) \leq U^{2D} \implies ||\Phi(\mathbf{x})|| \leq U^D$, meaning we can substitute R with U^d in the bound. By the same derivation, $||\Psi(\mathbf{x}, \mathbf{y})|| \leq U^D$ as well.

Finally, it can be easily seen that from the assumption of independence of the features, $\beta(x_i) \leq T \; \forall i \in \{1, ..., D\} \implies \beta(\mathbf{X}) \leq T^D$.

Plugging the bound on $||\beta(x)||_2^2 \leq Q^D$, U^D for R, and T^D for B in the bound by Gretton et al. yields the result. □

In addition to the variance of the empirical weighted risk in the source domain, another component that is important for analysis of the generalization in the target domain is the estimation error of the weights obtained by KMM, given by $\hat{\beta}$. For this purpose, we analyze the role of the dimensionality on the estimation error as studied by Theorem 4 in Cortes et al. [19]. We first restate the formal definition of admissibility given by Cortes et al. [19]

Definition 1:[19]*Let H be a hypothesis set. The loss l is σ-admissible if there exists $\sigma \in \mathbb{R}_+$ such that for any two hypotheses h, $h' \in H$ and for all $(x, y) \in X \times Y$,*

$$|l(x, y, h) - l(x, y, h')| \leq \sigma |h(x) - h'(x)|$$

We now present Theorem 4 proved by Cortes et al [19], before we analyze it in terms of the dimensionality of the dataset:

Theorem 4[19] *Let k be a strictly positive definite symmetric universal kernel suck that $k(x, x) \leq \kappa < \infty$. Let h_β be the hypothesis returned by a kernel-based regularization algorithm using a weighted sample S_β using weights β, and let $\hat{h_\beta}$ be a hypothesis obtained the same way using a weighted sample $S_{\hat{\beta}}$ with weights $\hat{\beta}$. Let l_β be the*

loss on a hypothesis function h_β and let $l_{\hat{\beta}}$ be the loss on hypothesis function $h_{\hat{\beta}}$, where the loss function $l(\cdot,\cdot,h)$ is σ-admissible. Then, for any $\delta > 0$ with probability at least $1-\delta$, the difference in generalization error of the hypotheses is bounded as:

$$|R(l_\beta) - R(l_{\hat{\beta}})| = \frac{\sigma^2\kappa^2}{\lambda}\left(\frac{\xi B}{\sqrt{n_{tr}}} + \frac{\kappa^{\frac{1}{2}}}{\lambda_{min}^{\frac{1}{2}}(\mathbf{K})}\sqrt{\frac{B^2}{n_{tr}} + \frac{1}{n_{te}}}\left(1 + \sqrt{2\log\left(\frac{2}{\delta}\right)}\right)\right) \tag{7}$$

Now we can make use of this estimation error result and analyze it in terms of dimensinality in the following lemma :

Lemma 2: *Let the σ-admissibility assumption hold on the loss function $l(\cdot,\cdot,\theta)$, and let assumptions **A1**, **A2** and **A3** hold. Let k be a kernel function that satisfies **A2**, and such that $||\Phi(X_j)|| \leq U\ \forall j \in 1,...,d$. Let $\mathbf{K}$ be the kernel Gram matrix for kernel k, $\mathbf{K}_1$, $\mathbf{K}_2$,...,$\mathbf{K}_d$ be the kernel Gram matrices of $k(x^{(1)},y^{(1)}),..,k(x^{(d)},y^{(d)})$ respectively, and let $\tilde{\lambda}$ be the smallest among the minimum eigenvalues $\lambda_{min}(\boldsymbol{K}_1),...,\lambda_{min}(\boldsymbol{K}_d)$. Then the following bound on the estimation error of the risk in the test domain holds:*

$$|R^{te}(l_{\hat{\beta}}) - R^{te}(l_\beta)| \leq \frac{\sigma^2\kappa^2}{\lambda}\left(\frac{\xi T^d}{\sqrt{n_{tr}}} + \frac{\kappa^{\frac{1}{2}}}{\tilde{\lambda}^{\frac{d}{2}}}\sqrt{\frac{T^{2d}}{n_{tr}} + \frac{1}{n_{te}}}\left(1 + \sqrt{2\log\left(\frac{2}{\delta}\right)}\right)\right) \tag{8}$$

Proof: Since k satisfies **A1**, from the positive-semidefinite property of the kernel gram matrix, the following holds:

$$\lambda_{min}(\mathbf{K}) \geq \lambda_{min}(\mathbf{K}_1...\mathbf{K}_d) \geq \prod_{i=1}^{d}\lambda_{min}(\mathbf{K}_i) \geq \tilde{\lambda}^d$$

Thus, we can insert $\tilde{\lambda}^d$ for $\lambda_{min}(\mathbf{K})$ in the bound by Cortes et al [19] to obtain the result. Similarly, we can insert T^d for B with the same argument used in the proof of Lemma 1. □

Now that we have the bounds on the variance of the empirical risk in the source domain and the estimation error of β, we can combine them to prove Theorem 2 which we presented in the main text, and we restate here:

Theorem 2: *Assume that **A1**, **A2** and **A3** hold and let for each feature i, $||\beta_W(x^{(i)})||_2^2 \leq Q, \beta_W(x^{(i)}) \leq T$ $\forall i \in 1,...,d$. Furthermore, let the importance weights β_W be a result of the KMM procedure using a feature map $\Phi : \mathcal{X} \to \mathcal{H}$ which corresponds to a kernel function k that satisfies **A1**, and such that $||\Phi(X_j)|| \leq U$ $\forall j \in 1,...,d$. Let $\mathbf{K}$ be the kernel Gram matrix for kernel k, $\mathbf{K}_1$, $\mathbf{K}_2$,...,$\mathbf{K}_d$ be the kernel Gram matrices of $k(x^{(1)},y^{(1)}),..,k(x^{(d)},y^{(d)})$ respectively, and let $\tilde{\lambda}$ be the smallest among the minimum eigenvalues $\lambda_{min}(\boldsymbol{K}_1),...,\lambda_{min}(\boldsymbol{K}_d)$. Then the generalization bound in the target domain satisfies:*

$$\begin{aligned}|R^{te}(l_{\hat{\beta}_W}) - R^{te}(l^*)| \leq & |R^{te}(l_W^*) - R^{te}(l^*)| + \frac{(2+\sqrt{2\log(6/\delta)})CU^d}{\frac{n_{tr}}{\sqrt{Q^d}}} + C(1+\sqrt{2\log(6/\delta)})U^d\sqrt{T^{2d}/n_{tr} + 1/n_{te}} \\ & + \frac{\sigma^2\kappa^2}{\lambda}\left(\frac{\xi T^d}{\sqrt{n_{tr}}} + \frac{\kappa^{\frac{1}{2}}}{\tilde{\lambda}^{d/2}}\sqrt{\frac{T^{2d}}{n_{tr}} + \frac{1}{n_{te}}}\left(1 + \sqrt{2\log(6/\delta)}\right)\right)\end{aligned}$$

Proof: By expanding the risk in the target domain and using triangle inequality, we have:

$$\begin{aligned}
&|R^{te}(l_{\hat{\beta}_W}) - R^{te}(l^*)| \\
=&|R^{te}(l_{\hat{\beta}_W}) - R^{te}(l_{\beta_W}) + R^{te}(l_{\beta_W}) - R^{tr}_{\beta_W}(l_{\beta_W}) + R^{tr}_{\beta_W}(l_{\beta_W}) - R^{tr}_{\beta_W}(l^*_{\beta_W}) + R^{tr}_{\beta_W}(l^*_{\beta_W}) - R^{te}(l^*_W) + R^{te}(l^*_W) - R^{te}(l^*)| \\
\leq&|R^{te}(l_{\hat{\beta}_W}) - R^{te}(l_{\beta_W})| + |R^{te}(l_{\beta_W}) - R^{tr}_{\beta_W}(l_{\beta_W})| + |R^{tr}_{\beta_W}(l_{\beta_W}) - R^{tr}_{\beta_W}(l^*_{\beta_W})| + |R^{tr}_{\beta_W}(l^*_{\beta_W}) - R^{te}(l^*_W)| + |R^{te}(l^*_W) - R^{te}(l^*)| \\
\leq&|R^{te}(l_{\hat{\beta}_W}) - R^{te}(l_{\beta_W})| + 2\sup_{l\in\mathcal{G}}|\hat{R}^{tr}_{\beta_W}(l) - R^{tr}_{\beta_W}(l)| + 2\sup_{l\in\mathcal{G}}|R^{tr}_{\beta_W}(l) - R^{te}(l)| + |R^{te}(l^*_W) - R^{te}(l^*)| \\
\leq&|R^{te}(l^*_W) - R^{te}(l^*)| + |R^{te}(l_{\hat{\beta}_W}) - R^{te}(l_{\beta_W})| + 2\sup_{l\in\mathcal{G}}|\hat{R}^{tr}_{\beta_W}(l) - R^{te}(l)|
\end{aligned} \tag{9}$$

Substituting the bound of Lemma 1 for the second term in the RHS, and the bound of Lemma 2 in the first term of the RHS yields the final result. □

8 Covariance Operators

In the main text we discussed estimating the trace of the conditional covariance operator. It can be empirically estimated as (using notation as defined in the main text):

$$\begin{aligned}
&\mathrm{Tr}[\hat{\mathcal{U}}_{YY|\hat{h}(X)}] \\
&= \mathrm{Tr}[\hat{\mathcal{U}}_{YY}] - \mathrm{Tr}[\hat{\mathcal{U}}_{Y,\hat{h}(X)}\hat{\mathcal{U}}^{-1}_{\hat{h}(X),\hat{h}(X)}\hat{\mathcal{U}}_{\hat{h}(X),Y}] \\
&= \frac{1}{n_{tr}}\mathrm{Tr}[\rho(\mathbf{Y})\rho(\mathbf{Y})^T] - \frac{1}{n_{tr}}\mathrm{Tr}[\rho(\mathbf{Y})\phi(\hat{h}(\mathbf{X}))^T(\phi(\hat{h}(\mathbf{X}))\phi(\hat{h}(\mathbf{X}))^T + n_{tr}\epsilon\mathbf{I})^{-1}\phi(\hat{h}(\mathbf{X}))\rho(\mathbf{Y})^T],
\end{aligned}$$

where ϵ is a small number to prevent ill conditioning of the matrix, and fixed to 0.01.

9 Additional Tables

Table for pseudo-real dataset which includes the standard errors:

	Unweighted	KMM-all	KLIEP-all	RuLSIF-all	LHSS	KMM-PCA	KLIEP-PCA	RuLSIF-PCA	KMM-W	D-W	D-original	p-value
Ailerons	2.26(0.1)	2.01(0.1)	2.09(0.1)	2.18(0.1)	2.06(0.1)	2.72(0.17)	2.74(0.15)	2.81(0.18)	0.92(0.07)	9	40	0.001
Bank32NH	0.79(0.03)	0.79(0.02)	0.82(0.02)	0.78(0.03)	0.73(0.03)	0.91(0.02)	0.92(0.02)	0.91(0.03)	0.62(0.05)	11	32	0.0527
Bank8FM	0.81(0.06)	0.78(0.05)	0.84(0.05)	0.79(0.06)	0.74(0.06)	0.92(0.06)	0.99(0.05)	0.99(0.05)	0.32(0.08)	1	8	0.001
Abalone	0.99(0.15)	0.87(0.13)	0.9(0.14)	0.95(0.15)	0.85(0.13)	1.27(0.2)	1.05(0.16)	0.96(0.14)	0.87(0.13)	7	7	0.7842
Elevators	1.14(0.02)	1.02(0.02)	1.07(0.01)	1.12(0.02)	1.02(0.01)	1.5(0.05)	1.1(0.01)	1.12(0.02)	1.02(0.02)	16	18	0.4229
CPU-Act	1.56(0.3)	1.29(0.26)	1.46(0.25)	1.44(0.26)	1.36(0.2)	2.12(0.5)	2.1(0.5)	2.22(0.6)	0.53(.13)	13	21	0.001
California	0.99(0.02)	0.93(0.02)	1.0(0.02)	0.99(0.03)	0.94(0.03)	1.21(0.03)	1.55(0.3)	1.23(0.02)	0.78(0.03)	5	8	0.0049
Puma8NH	0.45(0.008)	0.38(0.007)	0.43(0.008)	0.44(0.008)	0.38(0.007)	0.74(0.06)	0.74(0.06)	0.74(0.06)	0.33(0.008)	3	8	0.0049

Table 4: NMSE results on the baselines and the proposed method, on the pseudo-synthetic datasets. The methods with the suffix "-all" use all the features to calculate importance weights. The "-PCA" suffix means that PCA was used to represent the data in lower dimensions before estimating the weights $\hat{\beta}$; the suffix "-W" means that the proposed low-dimensional representation given by $\mathbf{W}^T\mathbf{X}$.