A. Proof of Theorem 1

We use a result implicit in the proof of Theorem 2 of Cortes & Mohri (2014), for the case where $\mathcal{H}$ is the set of linear hypotheses over a fixed representation $\Phi$. Cortes & Mohri (2014) state their result for the case of domain adaptation: in our case, the factual distribution is the so-called “source domain”, and the counterfactual distribution is the “target domain”.

Theorem A.1. [Cortes & Mohri (2014)] Using the notation and assumptions of Theorem 1, for both $Q = P^F$ and $Q = P^{CF}$:

$$
\frac{\lambda}{\mu r} (\mathcal{L}_Q(\tilde{\beta}^F(\Phi)) - \mathcal{L}_Q(\tilde{\beta}^{CF}(\Phi)))^2 \leq \text{disc}_{\mathcal{H}_i}(\hat{P}^F, \hat{P}^{CF}) + \min_{\hat{h} \in \mathcal{H}_i} \frac{1}{n} \left( \sum_{i=1}^{n} |\tilde{y}^F_i(\Phi, \hat{h}) - y^F_i| + |\tilde{y}^{CF}_i(\Phi, \hat{h}) - y^{CF}_i| \right)
$$

(1)

In their work, Cortes & Mohri (2014) assume the $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) for a universal kernel, and they do not consider the role of the representation $\Phi$. Since the RKHS hypothesis space they use is much stronger than the linear space $\mathcal{H}_i$, it is often reasonable to assume that the second term in the bound 1 is small. We however cannot make this assumption, and therefore we wish to explicitly bound the term $\min_{\hat{h} \in \mathcal{H}_i} \frac{1}{n} \left( \sum_{i=1}^{n} |\tilde{y}^F_i(\Phi, \hat{h}) - y^F_i| + |\tilde{y}^{CF}_i(\Phi, \hat{h}) - y^{CF}_i| \right)$, while using the fact that we have control over the representation $\Phi$.

Lemma 1. Let $\{(x_i, t_i, y^F_i)\}_{i=1}^{n}$, $x_i \in \mathcal{X}$, $t_i \in \{0,1\}$ and $y^F_i \in \mathcal{Y} \subseteq \mathbb{R}$. We assume that $\mathcal{X}$ is a metric space with metric $d$, and that there exist two function $Y_0(x)$ and $Y_1(x)$ such that $y^F_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$, and in addition we define $y^{CF}_i = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$. We further assume that the functions $Y_0(x)$ and $Y_1(x)$ are Lipschitz continuous with constants $K_0$ and $K_1$ respectively, such that $d(x_a, x_b) \leq c \implies |Y_1(x_a) - Y_1(x_b)| \leq K_1 c$.

Proof. By the triangle inequality, we have that:

$$
|b - y_i^{CF}| \leq |b - y_i^{F}| + |y_i^{F} - y_i^{CF}|
$$

By the Lipschitz assumption on $Y_1 - t_i$, and since $d(x_i, x(j)(i)) \leq d_{i,j}(i)$, we obtain that

$$
|y_i^{F} - y_i^{CF}| = |Y_1 - t_i(x(j)(i)) - Y_1 - t_i(x_i)| \leq d_{i,j}(i) K_1 - t_i.
$$

By definition $y_i^{CF} = Y_1 - t_i(x_i)$. In addition, by definition of $j(i)$, we have $d_{i,j}(i) = 1 - t_i$, and therefore $y_i^{CF} = Y_1 - t_i(x(j)(i))$, proving the equality. The inequality is an immediate consequence of the Lipschitz property. \(\square\)

We restate Theorem 1 and prove it.

Theorem 1. For a sample $\{(x_i, t_i, y^F_i)\}_{i=1}^{n}$, $x_i \in \mathcal{X}$, $t_i \in \{0,1\}$ and $y^F_i \in \mathcal{Y} \subseteq \mathbb{R}$, recall that $y^F_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$, and in addition define $y^{CF}_i = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$. For a given representation function $\Phi : \mathcal{X} \rightarrow \mathbb{R}^d$, let $\hat{P}^F = (\Phi(x_1), t_1), \ldots, (\Phi(x_n), t_n)$, $\hat{P}^{CF} = (\Phi(x_1), 1 - t_1), \ldots, (\Phi(x_n), 1 - t_n)$. We assume that $\mathcal{X}$ is a metric space with metric $d$, and that the potential outcome functions $Y_0(x)$ and $Y_1(x)$ are Lipschitz continuous with constants $K_0$ and $K_1$ respectively, such that $d(x_a, x_b) \leq c \implies |Y_1(x_a) - Y_1(x_b)| \leq K_1 c$.
Let $H_t \subset \mathbb{R}^{d+1}$ be the space of linear functions, and for $\beta \in H_t$, let $L_P(\beta) = \mathbb{E}_{(x,t,y) \sim P} [L(\beta(x,t),y)]$ be the expected loss of $\beta$ over distribution $P$. Let $r = \max (\mathbb{E}_{(x,t) \sim P} [\|\Phi(x),t\|_2], \mathbb{E}_{(x,t) \sim P_{CF}} [\|\Phi(x),t\|_2])$.

For $\lambda > 0$, let $\hat{\beta}^F(\Phi) = \arg \min_{\beta \in H_t} L_{\hat{P}_F}(\beta) + \lambda \|\beta\|_2^2$, and $\hat{\beta}^{CF}(\Phi)$ similarly for $\hat{P}_{CF}$, i.e. $\hat{\beta}^F(\Phi)$ and $\hat{\beta}^{CF}(\Phi)$ are the ridge regression solutions for the factual and counterfactual empirical distributions, respectively.

Let $\hat{y}_i^F(\Phi, h) = h^\top[\Phi(x_i), t_i]$ and $\hat{y}_i^{CF}(\Phi, h) = h^\top[\Phi(x_i), 1 - t_i]$ be the outputs of the hypothesis $h \in H_t$ over the representation $\Phi(x_i)$ for the factual and counterfactual settings of $t_i$, respectively. Finally, for each $i \in \{1 \ldots n\}$, let $j(i) \in \arg \min_{j \in \{1 \ldots n\}, t_j = 1 - t_i} d(x_j, x_i)$ be the nearest neighbor of $x_i$ among the group that received the opposite treatment from unit $i$. Let $d_{i,j} = d(x_i, x_j)$.

Then for both $Q = P^F$ and $Q = P^{CF}$ we have:

$$ \frac{\lambda}{\mu r} (L_Q(\hat{\beta}^F(\Phi)) - L_Q(\hat{\beta}^{CF}(\Phi)))^2 \leq (2) $$

$$ \text{disc}_{H_t}(\hat{P}_F^\Phi, \hat{P}_{CF}^\Phi) + \min_{h \in H_t} \frac{1}{n} \sum_{i=1}^{n} (|\hat{y}_i^F(\Phi, h) - y_i^F| + |\hat{y}_i^{CF}(\Phi, h) - y_i^{CF}|) \leq (3) $$

$$ \text{disc}_{H_t}(\hat{P}_F^\Phi, \hat{P}_{CF}^\Phi) + \min_{h \in H_t} \frac{1}{n} \sum_{i=1}^{n} (|\hat{y}_i^F(\Phi, h) - y_i^F| + |\hat{y}_i^{CF}(\Phi, h) - y_{j(i)}^{CF}|) + \frac{K_0}{n} \sum_{i: t_i = 1} d_{i,j(i)} + \frac{K_1}{n} \sum_{i: t_i = 0} d_{i,j(i)}.$$

**Proof.** Inequality (2) is immediate by Theorem A1. In order to prove inequality (3), we apply Lemma 1, setting $b = y_i^{CF}$ and summing over the $i$. \qed

**References**