Supplementary Material for: Learning Representations for Counterfactual Inference

Fredrik D. Johansson* FREJOHK@CHALMERS.SE
CSE, Chalmers University of Technology, Göteborg, SE-412 96, Sweden

Uri Shalit* SHALIT@CS.NYU.EDU
David Sontag DSONTAG@CS.NYU.EDU
CIMS, New York University, 251 Mercer Street, New York, NY 10012 USA

* **Equal contribution**

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 Φ. 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 A1. *[Cortes & Mohri (2014)] Using the notation and assumptions of Theorem 1, for* both $Q = P^F$ *and* $Q = P^{CF}$*:*

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

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 Φ. Since the RKHS hypothesis space they use is much stronger than the linear space $\mathcal{H}_l$, 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 $\min_{h\in\mathcal{H}_l} \frac{1}{n}\left(\sum_{i=1}^n |\hat{y}^F_i(\Phi,h) - y^F_i| + |\hat{y}^{CF}_i(\Phi,h) - y^{CF}_i|\right)$, while using the fact that we have control over the representation Φ.

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* $\mathrm{d}(x_a, x_b) \leq c \implies |Y_t(x_a) - Y_t(x_b)| \leq K_t c$. *Define* $j(i) \in \arg\min_{j\in\{1\ldots n\} \text{ s.t. } t_j = 1-t_i} \mathrm{d}(x_j, x_i)$ *to be the nearest neighbor of* x_i *among the group that received the opposite treatment from unit i, for all* $i \in \{1\ldots n\}$. *Let* $\mathrm{d}_{i,j} = \mathrm{d}(x_i, x_j)$

For any $b \in \mathcal{Y}$ *and* $h \in \mathcal{H}$*:*

$$|b - y^{CF}_i| \leq |b - y^F_{j(i)}| + K_{1-t_i}\mathrm{d}_{i,j(i)}$$

Proof. By the triangle inequality, we have that:

$$|b - y^{CF}_i| \leq |b - y^F_{j(i)}| + |y^F_{j(i)} - y^{CF}_i|.$$

By the Lipschitz assumption on Y_{1-t_i}, and since $\mathrm{d}(x_i, x_{j(i)}) \leq \mathrm{d}_{i,j(i)}$, we obtain that

$$|y^F_{j(i)} - y^{CF}_i| = |Y_{1-t_i}(x_{j(i)}) - Y_{1-t_i}(x_i)| \leq \mathrm{d}_{i,j(i)} K_{1-t_i}.$$

By definition $y^{CF}_i = Y_{1-t_i}(x_i)$. In addition, by definition of $j(i)$, we have $t_{j(i)} = 1 - t_i$, and therefore $y^F_{j(i)} = Y_{1-t_i}(x_{j(i)})$, proving the equality. The inequality is an immediate consequence of the Lipschitz property. □

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* $y_i \in \mathcal{Y}$*, 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* $\Phi : \mathcal{X} \to \mathbb{R}^d$*, let* $\hat{P}^F_\Phi = (\Phi(x_1), t_1), \ldots, (\Phi(x_n), t_n)$, $\hat{P}^{CF}_\Phi = (\Phi(x_1), 1-t_1), \ldots, (\Phi(x_n), 1-t_n)$. *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* $\mathrm{d}(x_a, x_b) \leq c \implies |Y_t(x_a) - Y_t(x_b)| \leq K_t c$.

Let $\mathcal{H}_l \subset \mathbb{R}^{d+1}$ be the space of linear functions, and for $\beta \in \mathcal{H}_l$, let $\mathcal{L}_P(\beta) = \mathbb{E}_{(x,t,y)\sim P}\left[L(\beta(x,t),y)\right]$ be the expected loss of β over distribution P. Let $r = max\left(\mathbb{E}_{(x,t)\sim P^F}\left[\|[\Phi(x),t]\|_2\right], \mathbb{E}_{(x,t)\sim P^{CF}}\left[\|[\Phi(x),t]\|_2\right]\right)$. For $\lambda > 0$, let $\hat{\beta}^F(\Phi) = \arg\min_{\beta\in\mathcal{H}_l} \mathcal{L}_{\hat{P}^F_\Phi}(\beta) + \lambda\|\beta\|_2^2$, and $\hat{\beta}^{CF}(\Phi)$ similarly for $\hat{P}^{CF}_\Phi$, 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 $\hat{y}^F_i(\Phi,h) = h^\top[\Phi(x_i),t_i]$ and $\hat{y}^{CF}_i(\Phi,h) = h^\top[\Phi(x_i),\,1-t_i]$ be the outputs of the hypothesis $h\in\mathcal{H}_l$ 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 $j(i) \in \arg\min_{j\in\{1\ldots n\}\text{ s.t. } t_j=1-t_i} \mathrm{d}(x_j,x_i)$ be the nearest neighbor of x_i among the group that received the opposite treatment from unit i. Let $\mathrm{d}_{i,j} = \mathrm{d}(x_i,x_j)$.

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

$$\frac{\lambda}{\mu r}(\mathcal{L}_Q(\hat{\beta}^F(\Phi)) - \mathcal{L}_Q(\hat{\beta}^{CF}(\Phi)))^2 \leq \tag{2}$$

$$disc_{\mathcal{H}_l}(\hat{P}^F_\Phi, \hat{P}^{CF}_\Phi) + \min_{h\in\mathcal{H}_l} \frac{1}{n}\sum_{i=1}^n \left(|\hat{y}^F_i(\Phi,h) - y^F_i| + |\hat{y}^{CF}_i(\Phi,h) - y^{CF}_i|\right) \leq \tag{3}$$

$$disc_{\mathcal{H}_l}(\hat{P}^F_\Phi, \hat{P}^{CF}_\Phi) + \min_{h\in\mathcal{H}_l} \frac{1}{n}\sum_{i=1}^n \left(|\hat{y}^F_i(\Phi,h) - y^F_i| + |\hat{y}^{CF}_i(\Phi,h) - y^F_{j(i)}|\right) + \frac{K_0}{n}\sum_{i:t_i=1} \mathrm{d}_{i,j(i)} + \frac{K_1}{n}\sum_{i:t_i=0} \mathrm{d}_{i,j(i)}.$$

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

References

Cortes, Corinna and Mohri, Mehryar. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.