Supplemental Materials for:
Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit [*] [1] Fredrik D. Johansson [*] [2] David Sontag [2] [3]

A. Proofs

A.1. Definitions, assumptions, and auxiliary lemmas

We first define the necessary distributions and prove some simple results about them. We assume a joint distribution function $p(x, t, Y_0, Y_1)$, such that $(Y_1, Y_0) \perp\!\!\!\perp t|x$, and $0 < p(t = 1|x) < 1$ for all x. Recall that we assume *Consistency*, that is we assume that we observe $y = Y_1|(t = 1)$ and $y = Y_0|(t = 0)$.

Definition A1. *The treatment effect for unit x is:*

$$\tau(x) := \mathbb{E}\left[Y_1 - Y_0|x\right].$$

We first show that under consistency and strong ignorability, the ITE function $\tau(x)$ is identifiable:

Lemma A1. *We have:*

$$\begin{aligned}
&\mathbb{E}\left[Y_1 - Y_0|x\right] = \\
&\mathbb{E}\left[Y_1|x\right] - \mathbb{E}\left[Y_0|x\right] = && (1)\\
&\mathbb{E}\left[Y_1|x, t=1\right] - \mathbb{E}\left[Y_0|x, t=0\right] = && (2)\\
&\mathbb{E}\left[y|x, t=1\right] - \mathbb{E}\left[y|x, t=0\right].
\end{aligned}$$

Equality (1) is because we assume that Y_t and t are independent conditioned on x. Equality (2) follows from the consistency assumption. Finally, the last equation is composed entirely of observable quantities and can be estimated from data since we assume $0 < p(t = 1|x) < 1$ for all x.

Definition A2. *Let $p^{t=1}(x) := p(x|t = 1)$, and $p^{t=0}(x) := p(x|t = 0)$ denote respectively the treatment and control distributions.*

Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a *representation function*. We will assume that Φ is differentiable.

Assumption A1. *The representation function Φ is one-to-one. Without loss of generality we will assume that $\mathcal{R}$ is the image of $\mathcal{X}$ under Φ, and define $\Psi : \mathcal{R} \to \mathcal{X}$ to be the inverse of Φ, such that $\Psi(\Phi(x)) = x$ for all $x \in \mathcal{X}$.*

Definition A3. *For a representation function $\Phi : \mathcal{X} \to \mathcal{R}$, and for a distribution p defined over $\mathcal{X}$, let p_Φ be the distribution induced by Φ over $\mathcal{R}$. Define $p_\Phi^{t=1}(r) := p_\Phi(r|t = 1)$, $p_\Phi^{t=0}(r) := p_\Phi(r|t = 0)$, to be the treatment and control distributions induced over $\mathcal{R}$.*

For a one-to-one Φ, the distribution p_Φ over $\mathcal{R} \times \{0, 1\}$ can be obtained by the standard change of variables formula, using the determinant of the Jacobian of $\Psi(r)$. See (Ben-Israel, 1999) for the case of a mapping Φ between spaces of different dimensions.

Lemma A2. *For all $r \in \mathcal{R}$, $t \in \{0, 1\}$:*

$$\begin{aligned}
p_\Phi(t|r) &= p(t|\Psi(r))\\
p(Y_t|r) &= p(Y_t|\Psi(r)).
\end{aligned}$$

Proof. Let $J_\Psi(r)$ be the absolute of the determinant of the Jacobian of $\Psi(r)$.

$$p_\Phi(t|r) = \frac{p_\Phi(t, r)}{p_\Phi(r)} \overset{(a)}{=} \frac{p(t, \Psi(r))J_\Psi(r)}{p(\Psi(r))J_\Psi(r)} = \frac{p(t, \Psi(r))}{p(\Psi(r))} = p(t|\Psi(r)),$$

where equality (a) is by the change of variable formula. The proof is identical for $p(Y_t|r)$. □

Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function, e.g. the absolute loss or squared loss.

Definition A4. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a representation function. Let $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ be an hypothesis defined over the representation space $\mathcal{R}$. The expected loss for the unit and treatment pair (x, t) is:*

$$\ell_{h,\Phi}(x, t) = \int_{\mathcal{Y}} L(Y_t, h(\Phi(x), t))p(Y_t|x)dY_t$$

Definition A5. *The expected factual loss and counterfactual losses of h and Φ are, respectively:*

$$\epsilon_F(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(x, t)\, p(x, t)\, dxdt$$

[*] Equal contribution [1] CIMS, New York University, New York, NY 10003 [2] IMES, MIT, Cambridge, MA 02142 [3] CSAIL, MIT, Cambridge, MA 02139. Correspondence to: Uri Shalit <shalit@cs.nyu.edu>, Fredrik D. Johansson <fredrikj@mit.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

Notation:
$p(x, t)$: distribution on $\mathcal{X} \times \{0, 1\}$
$u = p(t = 1)$: the marginal probability of treatment.
$p^{t=1}(x) = p(x|t = 1)$: treated distribution. $p^{t=0}(x) = p(x|t = 0)$: control distribution.
Φ: representation function mapping from $\mathcal{X}$ to $\mathcal{R}$.
Ψ: the inverse function of Φ, mapping from $\mathcal{R}$ to $\mathcal{X}$.
$p_\Phi(r, t)$: the distribution induced by Φ on $\mathcal{R} \times \{0, 1\}$.
$p_\Phi^{t=1}(r)$, $p_\Phi^{t=0}(r)$: treated and control distributions induced by Φ on $\mathcal{R}$.
$L(\cdot, \cdot)$: loss function, from $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R}_+$.
$\ell_{h,\Phi}(x, t)$: the expected loss of $h(\Phi(x), t)$ for the unit x and treatment t.
$\epsilon_F(h, \Phi)$, $\epsilon_{CF}(h, \Phi)$: expected factual and counterfactual loss of $h(\Phi(x), t)$.
$\tau(x) := \mathbb{E}[Y_1 - Y_0|x]$, the expected treatment effect for unit x.
$\epsilon_{\text{PEHE}}(f)$: expected error in estimating the individual treatment effect of a function $f(x, t)$.
$\text{IPM}_{\text{G}}(p, q)$: the integral probability metric distance induced by function family G between distributions p and q.

$$\epsilon_{CF}(h, \Phi) = \int_{\mathcal{X} \times \{0,1\}} \ell_{h,\Phi}(x, t)\, p(x, 1 - t)\, dxdt.$$

When it is clear from the context, we will sometimes use $\epsilon_F(f)$ and $\epsilon_{CF}(f)$ for the expected factual and counterfactual losses of an arbitrary function $f : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$.

Definition A6. *The expected* treated *and* control *losses are:*

$$\epsilon_F^{t=1}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(x, 1)\, p^{t=1}(x)\, dx$$

$$\epsilon_F^{t=0}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(x, 0)\, p^{t=0}(x)\, dx$$

$$\epsilon_{CF}^{t=1}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(x, 1)\, p^{t=0}(x)\, dx$$

$$\epsilon_{CF}^{t=0}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(x, 0)\, p^{t=1}(x)\, dx.$$

The four losses above are simply the loss conditioned on either the control or treated set. Let $u := p(t = 1)$ be the proportion of treated in the population. We then have the immediate result:

Lemma A3.

$$\epsilon_F(h, \Phi) = u \cdot \epsilon_F^{t=1}(h, \Phi) + (1 - u) \cdot \epsilon_F^{t=0}(h, \Phi)$$

$$\epsilon_{CF}(h, \Phi) = (1 - u) \cdot \epsilon_{CF}^{t=1}(h, \Phi) + u \cdot \epsilon_{CF}^{t=0}(h, \Phi).$$

The proof is immediate, noting that $p(x, t) = u \cdot p^{t=1}(x) + (1 - u) \cdot (x)$, and from the Definitions A4 and A6 of the losses.

Definition A7. *Let* G *be a function family consisting of functions* $g : \mathcal{S} \to \mathbb{R}$*. For a pair of distributions* p_1, p_2 *over* $\mathcal{S}$*, define the* Integral Probability Metric*:*

$$\mathit{IPM}_{\text{G}}(p_1, p_2) = \sup_{g \in \text{G}} \left| \int_{\mathcal{S}} g(s)\, (p_1(s) - p_2(s))\; ds \right|$$

$\text{IPM}_{\text{G}}(\cdot, \cdot)$ defines a pseudo-metric on the space of probability functions over $\mathcal{S}$, and for sufficiently large function families, $\text{IPM}_{\text{G}}(\cdot, \cdot)$ is a proper metric (Müller, 1997). Examples of sufficiently large functions families includes the set of bounded continuous functions, the set of 1-Lipschitz functions, and the set of unit norm functions in a universal Reproducing Norm Hilbert Space. The latter two give rise to the Wasserstein and Maximum Mean Discrepancy metrics, respectively (Gretton et al., 2012; Sriperumbudur et al., 2012). We note that for function families G such as the three mentioned above, for which $g \in \text{G} \implies -g \in \text{G}$, the absolute value can be omitted from definition A7.

A.2. General IPM bound

We now state and prove the most important technical lemma of this section.

Lemma A4 (Lemma 1, main text)**.** *Let* $\Phi : \mathcal{X} \to \mathcal{R}$ *be an invertible representation with* Ψ *its inverse. Let* $p_\Phi^{t=1}, p_\Phi^{t=0}$ *be defined as in Definition A3. Let* $u = p(t = 1)$*. Let* G *be a family of functions* $g : \mathcal{R} \to \mathbb{R}$*, and denote by* $\mathit{IPM}_{\text{G}}(\cdot, \cdot)$ *the integral probability metric induced by* G*. Let* $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ *be an hypothesis. Assume there exists a constant* $B_\Phi > 0$*, such that for* $t = 0, 1$*, the function* $g_{\Phi,h}(r, t) := \frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r), t) \in \text{G}$*. Then we have:*

$$\begin{aligned} &\epsilon_{CF}(h, \Phi) \leq \\ &(1 - u)\epsilon_F^{t=1}(h, \Phi) + u\epsilon_F^{t=0}(h, \Phi) + \\ &B_\Phi \cdot \mathit{IPM}_{\text{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right). \end{aligned} \tag{3}$$

Proof.

$$\begin{aligned} &\epsilon_{CF}(h, \Phi) - \left[(1 - u) \cdot \epsilon_F^{t=1}(h, \Phi) + u \cdot \epsilon_F^{t=0}(h, \Phi)\right] = \\ &\left[(1 - u) \cdot \epsilon_{CF}^{t=1}(h, \Phi) + u \cdot \epsilon_{CF}^{t=0}(h, \Phi)\right] - \\ &\left[(1 - u) \cdot \epsilon_F^{t=1}(h, \Phi) + u \cdot \epsilon_F^{t=0}(h, \Phi)\right] = \\ &(1 - u) \cdot \left[\epsilon_{CF}^{t=1}(h, \Phi) - \epsilon_F^{t=1}(h, \Phi)\right] + \\ &u \cdot \left[\epsilon_{CF}^{t=0}(h, \Phi) - \epsilon_F^{t=0}(h, \Phi)\right] = \end{aligned} \tag{4}$$

$$(1-u)\int_{\mathcal{X}} \ell_{h,\Phi}(x,1)\;\left(p^{t=0}(x) - p^{t=1}(x)\right)\,dx +$$

$$u\int_{\mathcal{X}} \ell_{h,\Phi}(x,0)\;\left(p^{t=1}(x) - p^{t=0}(x)\right)\,dx = \tag{5}$$

$$(1-u)\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(r),1)\;\left(p_{\Phi}^{t=0}(r) - p_{\Phi}^{t=1}(r)\right)\,dr +$$

$$u\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(r),0)\;\left(p_{\Phi}^{t=1}(r) - p_{\Phi}^{t=0}(r)\right)\,dr =$$

$$B_{\Phi}\cdot(1-u)\int_{\mathcal{R}} \frac{1}{B_{\Phi}}\ell_{h,\Phi}(\Psi(r),1)\;\left(p_{\Phi}^{t=0}(r) - p_{\Phi}^{t=1}(r)\right)\,dr +$$

$$B_{\Phi}\cdot u\int_{\mathcal{R}} \frac{1}{B_{\Phi}}\ell_{h,\Phi}(\Psi(r),0)\;\left(p_{\Phi}^{t=1}(r) - p_{\Phi}^{t=0}(r)\right)\,dr \leq \tag{6}$$

$$B_{\Phi}\cdot(1-u)\sup_{g\in \mathrm{G}}\left|\int_{\mathcal{R}} g(r)\;\left(p_{\Phi}^{t=0}(r) - p_{\Phi}^{t=1}(r)\right)\,dr\right| +$$

$$B_{\Phi}\cdot u\sup_{g\in \mathrm{G}}\left|\int_{\mathcal{R}} g(r)\;\left(p_{\Phi}^{t=1}(r) - p_{\Phi}^{t=0}(r)\right)\,dr\right| = \tag{7}$$

$$B_{\Phi}\cdot \mathrm{IPM}_{\mathrm{G}}(p_{\Phi}^{t=0}, p_{\Phi}^{t=1}). \tag{8}$$

Equality (4) is by Definition A6 of the treated and control loss, equality (5) is by the change of variables formula and Definition A3 of $p_{\Phi}^{t=1}$ and $p_{\Phi}^{t=0}$, inequality (6) is by the premise that $\frac{1}{B_{\Phi}}\cdot\ell_{h,\Phi}(\Psi(r),t)\in \mathrm{G}$ for $t=0,1$, and (7) is by Definition A7 of an IPM. □

The essential point in the proof of Lemma A4 is inequality 6. Note that on the l.h.s. of the inequality, we need to evaluate the expectations of $\ell_{h,\Phi}(\Psi(r),0)$ over $p_{\Phi}^{t=1}$ and $\ell_{h,\Phi}(\Psi(r),1)$ over $p_{\Phi}^{t=0}$. Both of these expectations are in general unavailable, since they require us to evaluate treatment outcomes on the control, and control outcomes on the treated. We therefore upper bound these unknowable quantities by taking a supremum over a function family which includes $\ell_{h,\Phi}(\Psi(r),0)$ and $\ell_{h,\Phi}(\Psi(r),1)$. The upper bound ignores most of the details of the outcome, and amounts to measuring a distance between two distributions we have samples from: the control and treated distribution. Note that for a randomized trial (i.e. when $t \perp\!\!\!\perp x$) with we have that $\mathrm{IPM}(p_{\Phi}^{t=1}, p_{\Phi}^{t=0}) = 0$. Indeed, it is straightforward to show that in that case we actually have an equality: $\epsilon_{CF}(h,\Phi) = (1-u)\cdot\epsilon_{F}^{t=1}(h,\Phi) + u\cdot\epsilon_{F}^{t=0}(h,\Phi)$.

The crucial condition in Lemma A4 is that the function $g_{\Phi,h}(r) := \frac{1}{B_{\Phi}}\ell_{h,\Phi}(\Psi(r),t)$ is in G. In subsections A.3 and A.4 below we look into two specific function families G, and evaluate what does this inclusion condition entail, and in particular we will derive specific bounds for B_{Φ}.

Definition A8. *For $t=0,1$ define:*

$$m_t(x) := \mathbb{E}\left[Y_t|x\right].$$

Obviously for the treatment effect $\tau(x)$ we have $\tau(x) = m_1(x) - m_0(x)$.

Let $f : \mathcal{X}\times\{0,1\}\to\mathcal{Y}$ by an hypothesis, such that $f(x,t) = h(\Phi(x),t)$ for a representation Φ and hypothesis h defined over the output of Φ.

Definition A9. *The treatment effect estimate for unit x is:*

$$\hat{\tau}_f(x) = f(x,1) - f(x,0).$$

Definition A10. *The expected Precision in Estimation of Heterogeneous Effect (PEHE) loss of g is:*

$$\epsilon_{PEHE}(f) = \int_{\mathcal{X}} \left(\hat{\tau}_f(x) - \tau(x)\right)^2\, p(x)\,dx.$$

Definition A11. *The expected variance of Y_t with respect to a distribution $p(x,t)$:*

$$\sigma_{Y_t}^2(p(x,t)) = \int_{\mathcal{X}\times\mathcal{Y}} \left(Y_t - m_t(x)\right)^2 p(Y_t|x)p(x,t)\,dY_t dx.$$

We define:

$$\begin{aligned}\sigma_{Y_t}^2 &= \min\{\sigma_{Y_t}^2(p(x,t)), \sigma_{Y_t}^2(p(x,1-t))\},\\ \sigma_{Y}^2 &= \min\{\sigma_{Y_0}^2, \sigma_{Y_1}^2\}.\end{aligned}$$

If Y_t are deterministic functions of x, then $\sigma_Y^2 = 0$.

We now show that $\epsilon_{\text{PEHE}}(f)$ is upper bounded by $2\epsilon_F + 2\epsilon_{CF} - 2\sigma_Y^2$ where ϵ_F and ϵ_{CF} are w.r.t. to the squared loss. An analogous result can be obtained for the absolute loss, using mean absolute deviation.

Lemma A5. *For any function $f : \mathcal{X}\times\{0,1\}\to\mathcal{Y}$, and distribution $p(x,t)$ over $\mathcal{X}\times\{0,1\}$:*

$$\begin{aligned}&\int_{\mathcal{X}} \left(f(x,t) - m_t(x)\right)^2\, p(x,t)\,dxdt =\\ &\quad \epsilon_F(f) - \sigma_{Y_t}^2(p(x,t)),\\ &\int_{\mathcal{X}} \left(f(x,t) - m_t(x)\right)^2\, p(x,1-t)\,dxdt =\\ &\quad \epsilon_{CF}(f) - \sigma_{Y_t}^2(p(x,1-t)),\end{aligned}$$

where $\epsilon_F(f)$ and $\epsilon_{CF}(f)$ are w.r.t. to the squared loss.

Proof. For simplicity we will prove for $p(x,t)$ and $\epsilon_F(f)$.

The proof for $p(x, 1-t)$ and ϵ_{CF} is identical.

$$\epsilon_F(f) =$$

$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f(x,t) - Y_t)^2\, p(Y_t|x)p(x,t)\, dY_t dx dt =$$

$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f(x,t) - m_t(x))^2\, p(Y_t|x)p(x,t)\, dY_t dx dt +$$

$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (m_t(x) - Y_t)^2\, p(Y_t|x)p(x,t)\, dY_t dx dt + \tag{9}$$

$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f(x,t) - m_t(x))\,(m_t(x) - Y_t)\, p(Y_t|x)p(x,t)\, dY_t dx dt = \tag{10}$$

$$\int_{\mathcal{X}\times\{0,1\}} (f(x,t) - m_t(x))^2\, p(x,t)\, dx dt +$$

$$\sigma^2_{Y_0}(p(x,t)) + \sigma^2_{Y_1}(p(x,t)) + 0,$$

where the equality (10) is by the Definition A11 of $\sigma^2_{Y_t}(p)$, and because the integral in (9) evaluates to zero, since $m_t(x) = \int_{\mathcal{X}} Y_t p(Y_t|x)\, dx$. □

Theorem 1. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function, with inverse Ψ. Let $p_\Phi^{t=1}, p_\Phi^{t=0}$ be defined as in Definition A3. Let $u = p(t=1)$. Let G be a family of functions $g : \mathcal{R} \to \mathbb{R}$, and denote by $IPM_{\mathrm{G}}(\cdot,\cdot)$ the integral probability metric induced by G. Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be an hypothesis. Let the loss $L(y_1, y_2) = (y_1 - y_2)^2$. Assume there exists a constant $B_\Phi > 0$, such that for $t \in \{0,1\}$, the functions $g_{\Phi,h}(r,t) := \frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r), t) \in \mathrm{G}$. We then have:*

$$\epsilon_{PEHE}(h,\Phi) \leq$$

$$2\big(\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) - 2\sigma_Y^2\big) \leq$$

$$2\big(\epsilon_F^{t=0}(h,\Phi) + \epsilon_F^{t=1}(h,\Phi) + B_\Phi IPM_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - 2\sigma_Y^2\big),$$

where ϵ_F and ϵ_{CF} are with respect to the squared loss.

Proof. We will prove the first inequality, $\epsilon_{\text{PEHE}}(f) \leq 2\epsilon_{CF}(h,\Phi) + 2\epsilon_F(h,\Phi) - 2\sigma_Y^2$. The second inequality is then immediate by Lemma A4. Recall that we denote $\epsilon_{\text{PEHE}}(f) = \epsilon_{\text{PEHE}}(h,\Phi)$ for $f(x,t) = h(\Phi(x), t)$.

$$\epsilon_{\text{PEHE}}(f) =$$

$$\int_{\mathcal{X}} \big((f(x,1) - f(x,0)) - (m_1(x) - m_0(x))\big)^2 p(x)\, dx =$$

$$\int_{\mathcal{X}} \big((f(x,1) - m_1(x)) + (m_0(x) - f(x,0))\big)^2 p(x)\, dx \leq \tag{11}$$

$$2\int_{\mathcal{X}} \Big((f(x,1) - m_1(x))^2 + (m_0(x) - f(x,0))^2\Big)\, p(x)\, dx = \tag{12}$$

$$2\int_{\mathcal{X}} (f(x,1) - m_1(x))^2\, p(x, t=1)\, dx +$$

$$2\int_{\mathcal{X}} (m_0(x) - f(x,0))^2\, p(x, t=0)\, dx +$$

$$2\int_{\mathcal{X}} (f(x,1) - m_1(x))^2\, p(x, t=0)\, dx +$$

$$2\int_{\mathcal{X}} (m_0(x) - f(x,0))^2\, p(x, t=1)\, dx =$$

$$2\int_{\mathcal{X}} (f(x,t) - m_t(x))^2\, p(x,t)\, dx dt +$$

$$2\int_{\mathcal{X}} (f(x,t) - m_t(x))^2\, p(x, 1-t)\, dx dt \leq \tag{13}$$

$$2(\epsilon_F - \sigma_Y^2) + 2(\epsilon_{CF} - \sigma_Y^2).$$

where (11) is because $(x+y)^2 \leq 2(x^2 + y^2)$, (12) is because $p(x) = p(x, t=0) + p(x, t=1)$ and (13) is by Lemma A5 and Definition A5 of the losses ϵ_F, ϵ_{CF} and Definition A11 of σ_Y^2. Having established the first inequality in the Theorem statement, we now show the second. We have by Lemma A4 that:

$$\epsilon_{CF}(h,\Phi) \leq$$

$$(1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + B_\Phi \cdot \mathrm{IPM}_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right).$$

We further have by Lemma A3 that:

$$\epsilon_F(h,\Phi) = u\epsilon_F^{t=1}(h,\Phi) + (1-u)\epsilon_F^{t=0}(h,\Phi).$$

Therefore

$$\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) \leq$$

$$\epsilon_F^{t=1}(h,\Phi) + \epsilon_F^{t=0}(h,\Phi) + B_\Phi \mathrm{IPM}_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right).$$

□

The upper bound is in terms of the standard generalization error on the treated and control distributions separately. Note that in some cases we might have very different sample sizes for treated and control, and that will show up in the finite sample bounds of these generalization errors.

We also note that the upper bound can be easily adapted to the case of the absolute loss PEHE $|\hat{\tau}(x) - \tau(x)|$. In that case the upper bound in the Theorem will have a factor 1 instead of the 2 stated above, and the standard deviation σ_Y^2 replaced by mean absolute deviation. The proof is straightforward where one simply applies the triangle inequality in inequality (11).

We will now give specific upper bounds for the constant B_Φ in Theorem 1, using two function families G in the IPM: the family of 1-Lipschitz functions, and the family of 1-norm reproducing kernel Hilbert space functions. Each one will have different assumptions about the distribution $p(x, t, Y_0, Y_1)$ and about the representation Φ and hypothesis h.

A.3. The family of 1-Lipschitz functions

For $\mathcal{S} \subset \mathbb{R}^d$, a function $f : \mathcal{S} \to \mathbb{R}$ has Lipschitz constant K if for all $x, y \in \mathcal{S}$, $|f(x) - f(y)| \leq K\|x - y\|$. If f is differentiable, then a sufficient condition for K-Lipschitz constant is if $\|\frac{\partial f}{\partial s}\| \leq K$ for all $s \in \mathcal{S}$.

For simplicity's sake we assume throughout this subsection that the true labeling functions the densities $p(Y_t|x)$ and the loss L are differentiable. However, this assumption could be relaxed to a mere Lipschitzness assumption.

Assumption A2. *There exists a constant $K > 0$ such that for all $x \in \mathcal{X}$, $t \in \{0, 1\}$, $\|\frac{p(Y_t|x)}{\partial x}\| \leq K$.*

Assumption A2 entails that each of the potential outcomes change smoothly as a function of the covariates (context) x.

Assumption A3. *The loss function L is differentiable, and there exists a constant $K_L > 0$ such that $\left|\frac{dL(y_1,y_2)}{dy_i}\right| \leq K_L$ for $i = 1, 2$. Additionally, there exists a constant M such that for all $y_2 \in \mathcal{Y}$, $M \geq \int_{\mathcal{Y}} L(y_1, y_2)\, dy_1$.*

Assuming $\mathcal{Y}$ is compact, loss functions which obey Assumption A3 include the log-loss, hinge-loss, absolute loss, and the squared loss.

When we let G in Definition A7 be the family of 1-Lipschitz functions, we obtain the so-called 1-*Wasserstein* distance between distributions, which we denote $\text{Wass}(\cdot, \cdot)$. It is well known that $\text{Wass}(\cdot, \cdot)$ is indeed a metric between distributions (Villani, 2008).

Definition A12. *Let $\frac{\partial \Phi(x)}{\partial x}$ be the Jacobian matrix of Φ at point x, i.e. the matrix of the partial derivatives of Φ. Let $\sigma_{max}(A)$ and $\sigma_{min}(A)$ denote respectively the largest and smallest singular values of a matrix A. Define $\rho(\Phi) = \sup_{x \in \mathcal{X}} \sigma_{max}\left(\frac{\partial \Phi(x)}{\partial x}\right) / \sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)$.*

It is an immediate result that $\rho(\Phi) \geq 1$.

Definition A13. *We will call a representation function $\Phi : \mathcal{X} \to \mathcal{R}$* Jacobian-normalized *if $\sup_{x \in \mathcal{X}} \sigma_{max}\left(\frac{\partial \Phi(x)}{\partial x}\right) = 1$.*

Note that any non-constant representation function Φ can be Jacobian-normalized by a simple scalar multiplication.

Lemma A6. *Assume that Φ is a Jacobian-normalized representation, and let Ψ be its inverse. For $t = 0, 1$, the Lipschitz constant of $p(Y_t|\Psi(r))$ is bounded by $\rho(\Phi)K$, where K is from Assumption A2, and $\rho(\Phi)$ as in Definition A12.*

Proof. Let $\Psi : \mathcal{R} \to \mathcal{X}$ be the inverse of Φ, which exists by the assumption that Φ is one-to-one. Let $\frac{\partial \Phi(x)}{\partial x}$ be the Jacobian matrix of Φ evaluated at x, and similarly let $\frac{\partial \Psi(r)}{\partial r}$ be the Jacobian matrix of Ψ evaluated at r. Note that $\frac{\partial \Psi(r)}{\partial r} \cdot \frac{\partial \Phi(x)}{\partial x} = I$ for $r = \Phi(x)$, since $\Psi \circ \Phi$ is the identity function on $\mathcal{X}$. Therefore for any $r \in \mathcal{R}$ and $x = \Psi(r)$:

$$\sigma_{max}\left(\frac{\partial \Psi(r)}{\partial r}\right) = \frac{1}{\sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)}, \tag{14}$$

where $\sigma_{max}(A)$ and $\sigma_{min}(A)$ are respectively the largest and smallest singular values of the matrix A, i.e. $\sigma_{max}(A)$ is the spectral norm of A.

For $x = \Psi(r)$ and $t \in \{0, 1\}$, we have by the chain rule:

$$\|\frac{\partial p(Y_t|\Psi(r))}{\partial r}\| = \|\frac{\partial p(Y_t|\Psi(r))}{\partial \Psi(r)} \frac{\partial \Psi(r)}{\partial r}\| \leq \tag{15}$$

$$\|\frac{\partial \Psi(r)}{\partial r}\| \|\frac{\partial p(Y_t|\Psi(r))}{\partial \Psi(r)}\| = \tag{16}$$

$$\frac{1}{\sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)} \|\frac{\partial p(Y_t|x)}{\partial x}\| \leq \tag{17}$$

$$\frac{K}{\sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)} \leq \rho(\Phi)K, \tag{18}$$

where inequality (15) is by the matrix norm inequality, equality (16) is by (14), inequality (17) is by assumption A2 on the norms of the gradient of $p(Y_t|x)$ w.r.t x , and inequality (18) is by Definition A12 of $\rho(\Phi)$, the assumption that Φ is Jacobian-normalized, and noting that singular values are necessarily non-negative.

□

Lemma A7. *Under the conditions of Lemma A4, further assume that for $t = 0, 1$, $p(Y_t|x)$ has gradients bounded by K as in A2, that h has bounded gradient norm bK, that the loss L has bounded gradient norm K_L, and that Φ is Jacobian-normalized. Then the Lipschitz constant of $\ell_{h,\Phi}(\Psi(r), t)$ is upper bounded by $K_L \cdot K\left(M\rho(\Phi) + b\right)$ for $t = 0, 1$.*

Proof. Using the chain rule, we have that:

$$\|\frac{\partial \ell_{h,\Phi}(\Psi(r),t)}{\partial r}\| = \|\frac{\partial}{\partial r}\int_{\mathcal{Y}} L(Y_t,h(r,t))p(Y_t|r)dY_t\| =$$

$$\|\int_{\mathcal{Y}} \frac{\partial}{\partial r}\left[L(Y_t,h(r,t))p(Y_t|r)\right]\, dY_t\| =$$

$$\|\int_{\mathcal{Y}} p(Y_t|r)\frac{\partial}{\partial r}L(Y_t,h(r,t)) + L(Y_t,h(r,t))\frac{\partial}{\partial r}p(Y_t|r)dY_t\| \leq$$

$$\int_{\mathcal{Y}} p(Y_t|r)\|\frac{\partial}{\partial r}L(Y_t,h(r,t))\|\,dY_t +$$

$$\int_{\mathcal{Y}} L(Y_t,h(r,t))\frac{\partial}{\partial r}p(Y_t|r)\,dY_t \leq \tag{19}$$

$$\int_{\mathcal{Y}} p(Y_t|r)\|\frac{\partial L(Y_t,h(r,t))}{\partial h(r,t)}\frac{\partial h(r,t)}{\partial r}\|\,dY_t +$$

$$\int_{\mathcal{Y}} L(Y_t,h(r,t))\frac{\partial}{\partial r}p(Y_t|r)\,dY_t \leq \tag{20}$$

$$\int_{\mathcal{Y}} p(Y_t|r)K_L \cdot b \cdot K + M \cdot \rho(\Phi) \cdot K, \tag{21}$$

where inequality 19 is due to Assumption A3 and inequality 20 is due to Lemma A6. □

Lemma A8. *Let $u = p(t = 1)$ be the marginal probability of treatment, and assume $0 < u < 1$. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one, Jacobian-normalized representation function. Let K be the Lipschitz constant of the functions $p(Y_t|x)$ on $\mathcal{X}$. Let K_L be the Lipschitz constant of the loss function L, and M be as in Assumption A3. Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis with Lipschitz constant bK. Then:*

$$\begin{aligned}&\epsilon_{CF}(h,\Phi) \leq \\ &(1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + \\ &2\left(M\rho(\Phi) + b\right) \cdot K \cdot K_L \cdot \mathit{Wass}(p_\Phi^{t=1}, p_\Phi^{t=0}).\end{aligned} \tag{22}$$

Proof. We will apply Lemma A4 with $\mathrm{G} = \{g : \mathcal{R} \to \mathbb{R} \text{ s.t. } f \text{ is 1-Lipschitz}\}$. By Lemma A7, we have that for $B_\Phi = (M\rho(\Phi) + b) \cdot K \cdot K_L$, the function $\frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. Inequality (22) then holds as a special case of Lemma A4. □

Theorem 2. *Under the assumptions of Lemma A8, using the squared loss for ϵ_F, we have:*

$$\begin{aligned}&\epsilon_{PEHE}(h,\Phi) \leq \\ &2\epsilon_F^{t=0}(h,\Phi) + 2\epsilon_F^{t=1}(h,\Phi) - 4\sigma_Y^2 + \\ &2\left(M\rho(\Phi) + b\right) \cdot K \cdot K_L \cdot \mathit{Wass}(p_\Phi^{t=1}, p_\Phi^{t=0}).\end{aligned}$$

Proof. Plug in the upper bound of Lemma A8 into the upper bound of Theorem 1. □

We examine the constant $(M\rho(\Phi) + b) \cdot K \cdot K_L$ in Theorem A8. K, the Lipschitz constant of m_0 and m_1, is not under our control and measures an aspect of the complexity of the true underlying functions we wish to approximate. The terms K_L and M depend on our choice of loss function and the size of the space $\mathcal{Y}$. The term b comes from our assumption that the hypothesis h has norm bK. Note that smaller b, while reducing the bound, might force the factual loss term $\epsilon_F(h,\Phi)$ to be larger since a small b implies a less flexible h. Finally, consider the term $\rho(\Phi)$. The assumption that Φ is normalized is rather natural, as we do not expect a certain scale from a representation. Furthermore, below we show that in fact the Wasserstein distance is positively homogeneous with respect to the representation Φ. Therefore, in Lemma A8, we can indeed assume that Φ is normalized. The specific choice of *Jacobian-normalized* scaling yields what is in our opinion a more interpretable result in terms of the inverse condition number $\rho(\Phi)$. For twice-differentiable Φ, $\rho(\Phi)$ is minimized if and only if Φ is a linear orthogonal transformation (mat).

Lemma A9. *The Wasserstein distance is positive homogeneous for scalar transformations of the underlying space. Let p, q be probability density functions defined over $\mathcal{X}$. For $\alpha > 0$ and the mapping $\Phi(x) = \alpha x$, let p_α and q_α be the distributions on $\alpha\mathcal{X}$ induced by Φ. Then:*

$$\mathit{Wass}\left(p_\alpha, q_\alpha\right) = \alpha \mathit{Wass}\left(p, q\right).$$

Proof. Following (Villani, 2008; Kuang & Tabak, 2016), we use another characterization of the Wasserstein distance. Let $\mathcal{M}_{p,q}$ be the set of mass preserving maps from $\mathcal{X}$ to itself which map the distribution p to the distribution q. That is, $\mathcal{M}_{p,q} = \{M : \mathcal{X} \to \mathcal{X} \text{ s.t. } q(M(S)) = p(S) \text{ for all measurable bounded } S \subset \mathcal{X}\}$. We then have that:

$$\mathrm{Wass}(p,q) = \inf_{M \in \mathcal{M}_{p,q}} \int_{\mathcal{X}} \|M(x) - x\| p(x)\,dx. \tag{23}$$

It is known that the infimum in (23) is actually achievable (Villani, 2008, Theorem 5.2). Denote by $M^* : \mathcal{X} \to \mathcal{X}$ the map achieving the infimum for $\mathrm{Wass}(p,q)$. Define $M^*_\alpha : \alpha\mathcal{X} \to \alpha\mathcal{X}$, by $M^*_\alpha(x') = \alpha M^*(\frac{x'}{\alpha})$, where $x' = \alpha x$. M^*_α maps p_α to q_α, and we have that $\|M^*_\alpha(x') - x'\| = \alpha\|M^*(x) - x\|$. Therefore M^*_α achieves the infimum for the pair (p_α, q_α), and we have that $\mathrm{Wass}\left(p_\alpha, q_\alpha\right) = \alpha\mathrm{Wass}\left(p, q\right)$. □

A.4. Functions in the unit ball of a RKHS

Let $\mathcal{H}_x, \mathcal{H}_r$ be a reproducing kernel Hilbert space, with corresponding kernels $k_x(\cdot,\cdot)$, $k_r(\cdot,\cdot)$. We have for all $x \in \mathcal{X}$ that $k_x(\cdot, x)$ is its Hilbert space mapping, and similarly $k_r(\cdot, r)$ for all $r \in \mathcal{R}$.

Recall that the major condition in Lemma A4 is that $\frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. The function space G we use here is $\mathrm{G} = \{g \in \mathcal{H}_r \text{ s.t. } \|g\|_{\mathcal{H}_r} \leq 1\}$.

We will focus on the case where L is the squared loss, and we will make the following two assumptions:

Assumption A4. *There exist $f_0^Y, f_1^Y \in \mathcal{H}_x$ such that $m_t(x) = \left\langle f_t^Y, k_x(x,\cdot)\right\rangle_{\mathcal{H}_x}$, i.e. the mean potential outcome functions m_0, m_1 are in $\mathcal{H}_x$. Further assume that $\|f_t^Y\|_{\mathcal{H}_x} \leq K$.*

Definition A14. *Define $\eta_{Y_t}(x) := \sqrt{\int_{\mathcal{Y}} \left(Y_t - m_t(x)\right)^2 p(Y_t|x)}$. $\eta_{Y_t}(x)$ is the standard deviation of $Y_t|x$.*

Assumption A5. *There exists $f_0^\eta, f_1^\eta \in \mathcal{H}_x$ such that $\eta_{Y_t}(x) = \langle f_t^\eta, k_x(x,\cdot)\rangle_{\mathcal{H}_x}$, i.e. the conditional standard deviation functions of $Y_t|x$ are in $\mathcal{H}_x$. Further assume that $\|f_t^\eta\|_{\mathcal{H}_x} \leq M$.*

Assumption A6. *Let $\Phi : \mathcal{X} \to \mathcal{Y}$ be an invertible representation function, and let Ψ be its inverse. We assume there exists a bounded linear operator $\Gamma_\Phi : \mathcal{H}_r \to \mathcal{H}_x$ such that $\left\langle f_t^Y, k_x(\Psi(r),\cdot)\right\rangle_{\mathcal{H}_x} = \left\langle f_t^Y, \Gamma_\Phi k_r(r,\cdot)\right\rangle_{\mathcal{H}_x}$. We further assume that the Hilbert-Schmidt norm (operator norm) $\|\Gamma_\Phi\|_{HS}$ of Γ_Φ is bounded by K_Φ.*

The two assumptions above amount to assuming that Φ can be represented as one-to-one linear map between the two Hilbert spaces $\mathcal{H}_x$ and $\mathcal{H}_r$.

Under Assumptions A4 and A6 about m_0, m_1, and Φ, we have that $m_t(\Psi(r)) = \left\langle \Gamma_\Phi^* f_t^Y, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$, where Γ_Φ^* is the adjoint operator of Γ_Φ (Grunewalder et al., 2013).

Lemma A10. *Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis, and assume that there exist $f_t^h \in \mathcal{H}_r$ such that $h(r,t) = \left\langle f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$, and such that $\|f_t^h\|_{\mathcal{H}_r} \leq b$. Under Assumption A4 about m_0, m_1, we have that $\ell_{h,\Phi}(\Psi(r),t) = \int_{\mathcal{Y}} \left(Y_t - h(r,t)\right)^2 p(Y_t|r) dY_t$ is in the tensor Hilbert space $\mathcal{H}_r \otimes \mathcal{H}_r$. Moreover, the norm of $\ell_{h,\Phi}(\Psi(r),t)$ in $\mathcal{H}_r \otimes \mathcal{H}_r$ is upper bounded by $4\left(K_\Phi^2 K^2 + b^2\right)$.*

Proof. We first decompose $\int_{\mathcal{Y}} \left(Y_t - h(r,t)\right)^2 p(Y_t|x) dY_t$ into a noise and mean fitting term, using $r = \Phi(x)$:

$$\ell_{h,\Phi}(\Psi(r),t) =$$

$$\int_{\mathcal{Y}} \left(Y_t - h(r,t)\right)^2 p(Y_t|r)\, dY_t =$$

$$\int_{\mathcal{Y}} \left(Y_t - m_t(x) + m_t(x) - h(\Phi(x),t)\right)^2 p(Y_t|x)\, dY_t =$$

$$\int_{\mathcal{Y}} \left(Y_t - m_t(x)\right)^2 p(Y_t|x)\, dY_t +$$

$$\left(m_t(x) - h(\Phi(x),t)\right)^2 +$$

$$2\int_{\mathcal{Y}} \left(Y_t - m_t(x)\right)\left(m_t(x) - h(\Phi(x),t)\right) p(Y_t|x) dY_t = \tag{24}$$

$$\eta_{Y_t}^2(x) + \left(m_t(x) - h(\Phi(x),t)\right)^2 + 0, \tag{25}$$

where equality (24) is by Definition A14 of η, and because $\int_{\mathcal{Y}} \left(Y_t - m_t(x)\right) p(Y_t|x)\, dY_t = 0$ by definition of $m_t(x)$.

Moving to $\mathcal{R}$, recall that $r = \Phi(x)$, $x = \Psi(r)$. By linearity of the Hilbert space, we have that $m_t(\Psi(r)) - h(r,t) = \left\langle \Gamma_\Phi^* f_t^Y, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r} - \left\langle f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r} = \left\langle \Gamma_\Phi^* f_t^Y - f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$. By a well known result (Steinwart & Christmann, 2008, Theorem 7.25), the product $\left(Y_t(\Psi(r)) - h(r,t)\right) \cdot \left(Y_t(\Psi(r)) - h(r,t)\right)$ lies in the tensor product space $\mathcal{H}_r \otimes \mathcal{H}_r$, and is equal to $\left\langle (\Gamma_\Phi^* f_t^Y - f_t^h) \otimes (\Gamma_\Phi^* f_t^Y - f_t^h), k_r(r,\cdot) \otimes k_r(r,\cdot)\right\rangle_{\mathcal{H}_r \otimes \mathcal{H}_r}$. The norm of this function in $\mathcal{H}_r \otimes \mathcal{H}_r$ is $\|\Gamma_\Phi^* f_t^Y - f_t^h\|_{\mathcal{H}_r}^2$. This is the general Hilbert space version of the fact that for a vector $w \in \mathbb{R}^d$ one has that $\|ww^\top\|_F = \|w\|_2^2$, where $\|\cdot\|_F$ is the matrix Frobenius norm, and $\|\cdot\|_2^2$ is the square of the standard Euclidean norm. We therefore have a similar result for $\eta_{Y_t}^2$, using Assumption A5: $\eta_{Y_t}^2(x) = \eta_{Y_t}^2(\Psi(r)) = \langle \Gamma_\Phi^* f_t^\eta \otimes \Gamma_\Phi^* f_t^\eta, k_r(r,\cdot) \otimes k_r(r,\cdot)\rangle_{\mathcal{H}_r \otimes \mathcal{H}_r}$. The norm of this function in $\mathcal{H}_r \otimes \mathcal{H}_r$ is $\|\Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r}^2$. Overall this leads us to conclude, using Equation (25) that $\ell_{h,\Phi}(\Psi(r),t) \in \mathcal{H}_r \otimes \mathcal{H}_r$. Now we have, using (25):

$$\|\ell_{h,\Phi}(\Psi(r),t)\|_{\mathcal{H}_r \otimes \mathcal{H}_r} =$$

$$\|(\Gamma_\Phi^* f_t^Y - f_t^h) \otimes (\Gamma_\Phi^* f_t^Y - f_t^h) + \Gamma_\Phi^* f_t^\eta \otimes \Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r \otimes \mathcal{H}_r} \leq \tag{26}$$

$$\|\Gamma_\Phi^* f_t^Y - f_t^h\|_{\mathcal{H}_r}^2 + \|\Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r}^2 \leq \tag{27}$$

$$2\|\Gamma_\Phi^* f_t^Y\|_{\mathcal{H}_r}^2 + 2\|f_t^h\|_{\mathcal{H}_r}^2 + \|\Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r}^2 \leq \tag{28}$$

$$\|\Gamma_\Phi^*\|_{HS}^2 \left(2\|f_t^Y\|_{\mathcal{H}_x}^2 + \|f_t^\eta\|_{\mathcal{H}_x}^2\right) + 2\|f_t^h\|_{\mathcal{H}_r}^2 = \tag{29}$$

$$\|\Gamma_\Phi\|_{HS}^2 \left(2\|f_t^Y\|_{\mathcal{H}_x}^2 + \|f_t^\eta\|_{\mathcal{H}_x}^2\right) + 2\|f_t^h\|_{\mathcal{H}_r}^2 \leq \tag{30}$$

$$2K_\Phi^2(K^2 + M^2) + 2b^2.$$

Inequality (26) is by the norms given above and the triangle inequality. Inequality (27) is because for any Hilbert space $\mathcal{H}$, $\|a - b\|_{\mathcal{H}}^2 \leq 2\|a\|_{\mathcal{H}}^2 + 2\|b\|_{\mathcal{H}}^2$. Inequality (28) is by the definition of the operator norm. Equality (29) is because the norm of the adjoint operator is equal to the norm of the original operator, where we abused the notation $\|\cdot\|_{HS}$ to mean both the norm of operators from $\mathcal{H}_x$ to $\mathcal{H}_r$ and vice-versa. Finally, inequality (30) is by Assumptions A4, A5 and A6, and by the Lemma's premise on the norm of f_T^h. □

Lemma A11. *Let $u = p(t = 1)$ be the marginal probability of treatment, and assume $0 < u < 1$. Assume the distribution of Y_t conditioned on x follows Assumptions A5 with constant M. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function which obeys Assumption A6 with corresponding operator Γ_Φ with operator norm K_Φ. Let the functions Y_0, Y_1 obey Assumption A4, with bounded Hilbert space norm K . Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis, and assume that there exist $f_t^h \in \mathcal{H}_r$ such that $h(r,t) = \left\langle f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$, such that $\|f_t^h\|_{\mathcal{H}_r} \leq b$. Assume that ϵ_F and ϵ_{CF} are defined with respect to L being the* *squared loss. Then:*

$$
\begin{aligned}
&\epsilon_{CF}(h,\Phi) \leq \\
&\quad (1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + \\
&\quad 2\left(K_\Phi^2(K^2+M^2)+b^2\right)\cdot \textit{MMD}(p_\Phi^{t=1}, p_\Phi^{t=0}),
\end{aligned}
\tag{31}
$$

where ϵ_{CF} and ϵ_F use the squared loss.

Proof. We will apply Lemma A4 with G $=$ $f \in \mathcal{H}_r \otimes \mathcal{H}_r$ s.t. $\|f\|_{\mathcal{H}_r\otimes\mathcal{H}_r} \leq 1$. By Lemma A10, we have that for $B_\Phi = 2\left(K_\Phi^2(K^2+M^2)+b^2\right)$ and L being the squared loss, $\frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t) \in$ G. Inequality (31) then holds as a special case of Lemma A4. □

Theorem 3. *Under the assumptions of Lemma A11, using the squared loss for ϵ_F, we have:*

$$
\begin{aligned}
&\epsilon_{\textit{PEHE}}(h,\Phi) \leq \\
&2\epsilon_F^{t=0}(h,\Phi) + 2\epsilon_F^{t=1}(h,\Phi) - 4\sigma_Y^2 + \\
&4\left(K_\Phi^2(K^2+M^2)+b^2\right)\cdot \textit{MMD}(p_\Phi^{t=1}, p_\Phi^{t=0}).
\end{aligned}
$$

Proof. Plug in the upper bound of Lemma A11 into the upper bound of Theorem 1. □

B. Algorithmic details

We give details about the algorithms used in our framework.

B.1. Minimizing the Wasserstein distance

In general, computing (and minimizing) the Wasserstein distance involves solving a linear program, which may be prohibitively expensive for many practical applications. Cuturi (2013) showed that an approximation based on entropic regularization can be obtained through the Sinkhorn-Knopp matrix scaling algorithm, at orders of magnitude faster speed. Dubbed Sinkhorn distances, the approximation is computed using a fixed-point iteration involving repeated multiplication with a kernel matrix K. We can use the algorithm of Cuturi (2013) in our framework. See Algorithm 1 for an overview of how to compute the gradient g_1 in Algorithm **??**. When computing g_1, disregarding the gradient $\nabla_{\mathbf{W}} T^*$ amounts to minimizing an upper bound on the Sinkhorn transport. More advanced ideas for stochastic optimization of this distance have recently proposed by Aude et al. (2016), and might be used in future work.

While our framework is agnostic to the parameterization of Φ, our experiments focus on the case where Φ is a neural network. For convenience of implementation, we may represent the fixed-point iterations of the Sinkhorn algorithm as a recurrent neural network, where the states u_t evolve according to

$$
u_{t+1} = n_t./(n_c K(1./(u_t^\top K)^\top))\ .
$$

Here, K is a kernel matrix corresponding to a metric such as the euclidean distance, $K_{ij} = e^{-\lambda\|\Phi(x_i)-\Phi(x_j)\|_2}$, and n_c, n_t are the sizes of the control and treatment groups. In this way, we can minimize our entire objective with most of the frameworks commonly used for training neural networks, out of the box.

Algorithm 1 Computing the stochastic gradient of the Wasserstein distance

```
Input: Factual $(x_1,t_1,y_1),\ldots,(x_n,t_n,y_n)$, representation network $\Phi_{\mathbf{W}}$ with current weights by $\mathbf{W}$
Randomly sample a mini-batch with $m$ treated and $m'$ control units $(x_{i_1},0,y_{i_1}),\ldots,$
    $(x_{i_m},0,y_{i_m}),(x_{i_{m+1}},1,y_{i_{m+1}}),\ldots,(x_{i_{2m}},1,y_{i_{2m}})$
Calculate the $m\times m$ pairwise distance matrix between all treatment and control pairs $M(\Phi_{\mathbf{W}})$:
    $M_{kl}(\Phi) = \|\Phi_{\mathbf{W}}(x_{i_k}) - \Phi_{\mathbf{W}}(x_{i_{m+l}})\|$
Calculate the approximate optimal transport matrix $T^*$ using Algorithm 3 of Cuturi & Doucet (2014), with input $M(\Phi_{\mathbf{W}})$
Calculate the gradient:
    $g_1 = \nabla_{\mathbf{W}}\langle T^*, M(\Phi_{\mathbf{W}})\rangle$
```

B.2. Minimizing the maximum mean discrepancy

The MMD of treatment populations in the representation Φ, for a kernel $k(\cdot,\cdot)$ can be written as,

$$
\mathrm{MMD}_k(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^m, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m'}) = \tag{32}
$$

$$
\frac{1}{m(m-1)}\sum_{j=1}^m\sum_{k=1,k\neq j}^m k(\Phi_{\mathbf{W}}(x_{i_j}),\Phi_{\mathbf{W}}(x_{i_k})) \tag{33}
$$

$$
+\frac{2}{mm'}\sum_{j=1}^m\sum_{k=m}^{m+m'} k(\Phi_{\mathbf{W}}(x_{i_j}),\Phi_{\mathbf{W}}(x_{i_k})) \tag{34}
$$

$$
+\frac{1}{m'(1-m')}\sum_{j=1}^m\sum_{k=m,k\neq j}^{m'} k(\Phi_{\mathbf{W}}(x_{i_j}),\Phi_{\mathbf{W}}(x_{i_k})) \tag{35}
$$

The linear maximum-mean discrepancy can be written as a distance between means. In the notation of Algorithm **??**,

$$
\mathrm{MMD} = 2\left\|\frac{1}{m}\sum_{j=1}^m \Phi_{\mathbf{W}}(x_{i_j}) - \frac{1}{m'}\sum_{k=m+1}^{m'}\Phi_{\mathbf{W}}(x_{i_k})\right\|_2
$$

Let

$$
\mathbf{f}(\mathbf{W}) = \frac{1}{m}\sum_{j=1}^m \Phi_{\mathbf{W}}(x_{i_j}) - \frac{1}{m'}\sum_{k=m+1}^{m+m'}\Phi_{\mathbf{W}}(x_{i_k})
$$

Table 1. Hyperparameters and ranges.

Parameter	Range
Imbalance parameter, α	$\{10^{k/2}\}_{k=-10}^{6}$
Num. representation layers	$\{1, 2, 3\}$
Num. hypothesis layers	$\{1, 2, 3\}$
Dim. representation layers	$\{20, 50, 100, 200\}$
Dim. hypothesis layers	$\{20, 50, 100, 200\}$
Batch size	$\{100, 200, 500, 700\}$
Normalization	{projection, batch norm}
Weight decay (hypothesis)	$\{10^{-4}, 10^{-3}, 10^{-2}\}$

Then the gradient of the MMD with respect to $\mathbf{W}$ is,

$$g_1 = 2\frac{d\mathbf{f}(\mathbf{W})}{d\mathbf{W}} \frac{\mathbf{f}(\mathbf{W})}{\|\mathbf{f}(\mathbf{W})\|_\mathbf{2}} \ .$$

C. Experimental details

Our implementations of CFR and TARNet are based on Python and TensorFlow and are available at https://github.com/clinicalml/cfrnet. Both models were trained using stochastic gradient descent with Adam.

C.1. Hyperparameter selection

Standard methods for hyperparameter selection, such as cross-validation, are not generally applicable for estimating the PEHE loss since only one potential outcome is observed (unless the outcome is simulated). For real-world data, we may use the observed outcome $y_{j(i)}$ of the nearest neighbor $j(i)$ to i in the opposite treatment group, $t_{j(i)} = 1 - t_i$ as surrogate for the counterfactual outcome. We use this to define a nearest-neighbor approximation of the PEHE loss, $\epsilon_{\text{PEHE}_{nn}}(f) = \frac{1}{n}\sum_{i=1}^{n}\left((1-2t_i)(y_{j(i)} - y_i) - (f(x_i,1) - f(x_i,0))\right)^2$. On IHDP, we use the objective value on the validation set for early stopping in CFR, and $\epsilon_{\text{PEHE}_{nn}}(f)$ for hyperparameter selection. On the Jobs dataset, we use the policy risk on the validation set.

See Table 1 for a description of hyperparameters and search ranges.

C.2. Learned representations

Figure 1 show the representations learned by our CFR algorithm.

C.3. Absolute error for increasingly imbalanced data

Figure 2 shows the results of the same experiment as Figure 2 of the main paper, but in absolute terms.

References

MathOverflow: functions with orthogonal Jacobian. https://mathoverflow.net/questions/228964/functions-with-orthogonal-jacobian. Accessed: 2016-05-05.

Aude, Genevay, Cuturi, Marco, Peyré, Gabriel, and Bach, Francis. Stochastic optimization for large-scale optimal transport. *arXiv preprint arXiv:1605.08527*, 2016.

Ben-Israel, Adi. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.

Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 685–693, 2014.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012. ISSN 1532-4435.

Grunewalder, Steffen, Arthur, Gretton, and Shawe-Taylor, John. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1184–1192, 2013.

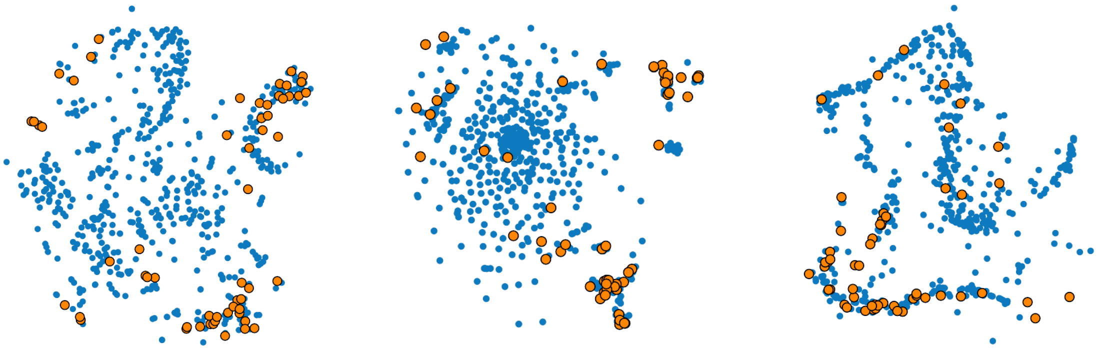
Kuang, Max and Tabak, Esteban. Preconditioning of optimal transport. *Preprint*, 2016.

Müller, Alfred. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pp. 429–443, 1997.

Sriperumbudur, Bharath K, Fukumizu, Kenji, Gretton, Arthur, Schölkopf, Bernhard, Lanckriet, Gert RG, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Steinwart, Ingo and Christmann, Andreas. *Support vector machines.* Springer Science & Business Media, 2008.

Villani, Cédric. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.



(a) Original data (b) Linear MMD (c) Wasserstein

Figure 1. t-SNE visualizations of the balanced representations of IHDP learned by our algorithms CFR, CFR MMD and CFR Wass. We note that the nearest-neighbor like quality of the Wasserstein distance results in a strip-like representation, whereas the linear MMD results in a ball-like shape in regions where overlap is small.

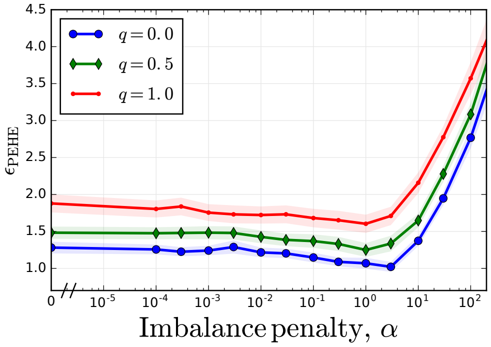


Figure 2. Out-of-sample error in estimated ITE, as a function of IPM regularization parameter for CFR Wass, on 500 realizations of IHDP, with high ($q = 1$), medium and low (artificial) imbalance between control and treated.