Counterfactual Cross-Validation: Stable Model Selection Procedure for Causal Inference Models– Appendix

A. Omitted Proofs

In this section, we denote $\tau(X)$, $\hat{\tau}(X)$, and $\tilde{\tau}(X,T,Y)$ as τ, $\hat{\tau}$, and $\tilde{\tau}$ for simplicity. We also denote $\tau(X_i)$, $\hat{\tau}(X_i)$, and $\tilde{\tau}(X_i,T_i,Y_i)$ as τ_i, $\hat{\tau}_i$, and $\tilde{\tau}_i$.

A.1. Proof of Proposition 1

Proof. First, the following equality holds:

$$\mathbb{E}\left[\widehat{\mathcal{R}}(\hat{\tau})\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i-\hat{\tau}_i)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\tau}-\tau+\tau-\hat{\tau})^2\right] = \mathbb{E}\left[(\tilde{\tau}-\tau)^2\right] - \frac{2}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}\left[(\hat{\tau}-\tau)(\tilde{\tau}-\tau)\right]}_{(a)} + \underbrace{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}_{\mathcal{R}_{true}(\hat{\tau})}$$

Then, we have $(a) = \mathbb{E}\left[(\hat{\tau}-\tau)(\tilde{\tau}-\tau)\right] = \mathbb{E}\left[\mathbb{E}\left[(\hat{\tau}-\tau)(\tilde{\tau}-\tau)\mid X\right]\right] = \mathbb{E}\left[(\hat{\tau}-\tau)(\mathbb{E}[\tilde{\tau}\mid X]-\tau)\right] = 0$.

Thus, we obtain $\mathbb{E}[\widehat{\mathcal{R}}\left(\hat{\tau}\right)] = \mathcal{R}_{true}\left(\hat{\tau}\right) + \mathbb{E}[(\tilde{\tau}-\tau)^2]$. □

A.2. Derivation of Eq. (5)

Proof. Following the same procedure as in the proof of Proposition 1, we have

$$\widehat{\mathcal{R}}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i-\hat{\tau}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i-\tau_i+\tau_i-\hat{\tau}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i-\tau_i)^2 - \frac{2}{n}\sum_{i=1}^{n}(\tilde{\tau}_i-\tau_i)(\hat{\tau}_i-\tau_i) + \frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i-\tau_i)^2$$

□

A.3. Proof of Proposition 3

Proof. We rewrite the DR plug-in tau in Eq. (8) as:

$$\begin{aligned}\tilde{\tau}_{DR}(X,T,Y) &= \frac{T}{e(X)}\left(Y-f_1(X)\right) - \frac{1-T}{1-e(X)}\left(Y-f_0(X)\right) + \left(f_1(X)-f_0(X)\right)\\ &= \tilde{\tau}_{DR_1}(X,T,Y) - \tilde{\tau}_{DR_0}(X,T,Y)\end{aligned}$$

where $\tilde{\tau}_{DR_1}(X,T,Y) = \frac{T}{e(X)}\left(Y-f_1(X)\right) + f_1(X)$ and $\tilde{\tau}_{DR_0}(X,T,Y) = \frac{1-T}{1-e(X)}\left(Y-f_0(X)\right) + f_0(X)$.

Then, the expectation of $\tilde{\tau}_{DR_1}$ is $\mathbb{E}\left[\tilde{\tau}_{DR_1}\mid X\right] = \mathbb{E}\left[\frac{T}{e(X)}\mid X\right]\mathbb{E}\left[\left(Y(1)-f_1(X)\right)\mid X\right] + f_1(X) = \mathbb{E}\left[Y(1)\mid X\right]$.

We also have $\mathbb{E}\left[\tilde{\tau}_{DR_0}\mid X\right] = \mathbb{E}\left[Y(0)\mid X\right]$ in the same way. Thus, we have, $\mathbb{E}\left[\tilde{\tau}_{DR}\mid X\right] = \mathbb{E}\left[Y(1)-Y(0)\mid X\right] = \tau$. □

A.4. Proof of Proposition 4

Proof. The second moment of $\tilde{\tau}_{DR_1}$ is

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{\tau}_{DR_1}\right)^2\mid X\right] &= \mathbb{E}\left[\left(\frac{T}{e(X)}\left(Y(1)-f_1(X)\right)+f_1(X)\right)^2\mid X\right]\\ &= \mathbb{E}\left[\left(\left(1-\frac{T}{e(X)}\right)\left(f_1(X)-Y(1)\right)+Y(1)\right)^2\mid X\right]\\ &= \mathbb{E}\left[\zeta_1\mid X\right] + \left(m_1(X)\right)^2 + w_1(X)\left(f_1(X)-m_1(X)\right)^2\end{aligned}$$

We also have the second moment of $\tilde{\tau}_{DR_0}$ in the same manner as follows:

$$\mathbb{E}\left[\left(\tilde{\tau}_{DR_0}\right)^2 \mid X\right] = \mathbb{E}\left[\zeta_0 \mid X\right] + \left(m_0(X)\right)^2 + w_0(X)\left(f_0(X) - m_0(X)\right)^2$$

where $\zeta_1 = (Y(1) - m_1(X))^2/e(X)$, $\zeta_0 = (Y(0) - m_0(X))^2/(1 - e(X))$. Note that $\mathbb{E}\left[\zeta_1 \mid X\right] = \mathbb{V}(Y(1) \mid X)/e(X)$ and $\mathbb{E}\left[\zeta_0 \mid X\right] = \mathbb{V}(Y(0) \mid X)/(1 - e(X))$.

Then, by using the result of Proposition 3, we obtain

$$\begin{aligned}
\mathbb{V}\left(\tilde{\tau}_{DR_1} \mid X\right) &= \mathbb{E}\left[\zeta_1 \mid X\right] + w_1(X)\left(f_1(X) - m_1(X)\right)^2 \\
\mathbb{V}\left(\tilde{\tau}_{DR_0} \mid X\right) &= \mathbb{E}\left[\zeta_0 \mid X\right] + w_0(X)\left(f_0(X) - m_0(X)\right)^2
\end{aligned}$$

In addition, from Lemma 6,

$$\begin{aligned}
\mathbb{V}\left(\tilde{\tau}_{DR} \mid X\right) &= \mathbb{V}\left(\tilde{\tau}_{DR_1} - \tilde{\tau}_{DR_0} \mid X\right) \\
&= \mathbb{V}\left(\tilde{\tau}_{DR_1} \mid X\right) - 2\mathrm{Cov}\left(\tilde{\tau}_{DR_1}, \tilde{\tau}_{DR_0} \mid X\right) + \mathbb{V}\left(\tilde{\tau}_{DR_0} \mid X\right) \\
&= \mathbb{E}\left[\zeta_1 + \zeta_0 \mid X\right] + w_1(X)\left(f_1(X) - m_1(X)\right)^2 \\
&\quad + w_0(X)\left(f_0(X) - m_0(X)\right)^2 + 2\left(f_1(X) - m_1(X)\right)\left(f_0(X) - m_0(X)\right) \\
&= \mathbb{E}\left[\zeta_1 + \zeta_0 \mid X\right] + \left(\sqrt{w_1(X)}\left(f_1(X) - m_1(X)\right) + \sqrt{w_0(X)}\left(f_0(X) - m_0(X)\right)\right)^2
\end{aligned}$$

where $w_1(X)w_0(X) = 1$. Hence, we have $\mathbb{E}_X[\mathbb{V}\left(\tilde{\tau}_{DR} \mid X\right)] = \zeta + \mathbb{E}_X\left[\left\{\sum_{t\in\mathcal{T}}\sqrt{w_t(X)}(f_t(X) - m_t(X))\right\}^2\right]$ where $\zeta = \mathbb{E}\left[\zeta_1 + \zeta_0\right]$. □

A.5. Proof of Theorem 2

Proof.

$$\begin{aligned}
&\mathbb{V}\left(2n^{-1}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)\right) \\
&= 4n^{-2}\,\mathbb{V}\left(\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)\right) \\
&= 4n^{-2}\,\mathbb{E}\left[\left(\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)\right)^2\right] \quad \because (a) = 0 \\
&= 4n^{-2}\,\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)(\hat{\tau}_j - \tau_j)(\tilde{\tau}_j - \tau_j)\right] \\
&= 4n^{-2}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\tau}_i - \tau_i)^2(\tilde{\tau}_i - \tau_i)^2\right] \quad \because \mathbb{E}_X[(\hat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)(\hat{\tau}_j - \tau_j)(\tilde{\tau}_j - \tau_j)] = 0, \forall i, j(i \neq j) \\
&\leq 4C_{\max}n^{-1}\,\mathbb{E}\left[(\tilde{\tau} - \tau)^2\right] \\
&= 4C_{\max}n^{-1}\,\mathbb{E}_X\left[\mathbb{E}[(\tilde{\tau} - \mathbb{E}[\tilde{\tau} \mid X])^2] \mid X\right] \\
&= 4C_{\max}n^{-1}\,\mathbb{E}_X\left[\mathbb{V}\left(\tilde{\tau} \mid X\right)\right]
\end{aligned}$$

□

A.6. Technical Lemmas

***Lemma* 6.** *The conditional covariance of $\tilde{\tau}_{DR_1}$ and $\tilde{\tau}_{DR_0}$ is:*

$$\mathrm{Cov}\left(\tilde{\tau}_{DR_1}, \tilde{\tau}_{DR_0} \mid X\right) = -\left(f_1(X) - m_1(X)\right)\left(f_0(X) - m_0(X)\right)$$

Proof.

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\tau}_{DR_1}, \tilde{\tau}_{DR_0} \mid X\right) &= \mathbb{E}\left[\tilde{\tau}_{DR_1} \cdot \tilde{\tau}_{DR_0} \mid X\right] - \mathbb{E}\left[\tilde{\tau}_{DR_1} \mid X\right] \cdot \mathbb{E}\left[\tilde{\tau}_{DR_0} \mid X\right] \\ &= \mathbb{E}\left[\tilde{\tau}_{DR_1} \cdot \tilde{\tau}_{DR_0} \mid X\right] - m_1(X) \cdot m_0(X)\end{aligned}$$

Then,

$$\begin{aligned}&\mathbb{E}\left[\tilde{\tau}_{DR_1} \cdot \tilde{\tau}_{DR_0} \mid X\right] \\ &= f_1(X)f_0(X) + f_1(X)\mathbb{E}\left[\frac{1-T}{1-e(X)}\left(Y(0) - f_0(X)\right) \mid X\right] + f_0(X)\mathbb{E}\left[\frac{T}{e(X)}\left(Y(1) - f_1(X)\right) \mid X\right] \\ &= f_1(X)f_0(X) + f_1(X)(m_0(X) - f_0(X)) + f_0(X)(m_1(X) - f_1(X))\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\tilde{\tau}_{DR_1} \cdot \tilde{\tau}_{DR_0} \mid X\right] - m_1(X)m_0(X) \\ &= f_1(X)f_0(X) + f_1(X)(m_0(X) - f_0(X)) + f_0(X)(m_1(X) - f_1(X)) - m_1(X)m_0(X) \\ &= -\left(f_1(X) - m_1(X)\right)\left(f_0(X) - m_0(X)\right)\end{aligned}$$

□

***Lemma* 7.** *(Similar to Lemma A.4 of (Shalit et al., 2017)) Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation with Ψ its inverse. Let G be a family of functions $g : \mathcal{R} \to \mathbb{R}_{\geq 0}$ and $h : \mathcal{R} \times \mathcal{T} \to \mathcal{Y}$ be a hypothesis. Assume that, for any given $t \in \mathcal{T}$ and $w : \mathcal{X} \times \mathcal{T} \to \mathcal{R}_{\geq 0}$, there exists a constant $B_\Phi > 0$, such that $\frac{1}{B_\Phi} \cdot \ell^w_{h,\Phi}(\Psi(r), t) \in \mathrm{G}$. Then we have:*

$$\epsilon^w_{CF_{1-t}}(h, \Phi) \leq \epsilon^w_{F_t}(h, \Phi) + B_\Phi \cdot \mathrm{IPM}_G\left(p^\Phi_t, p^\Phi_{1-t}\right)$$

Proof.

$$\begin{aligned}\epsilon^w_{CF_{1-t}}(h, \Phi) - \epsilon^w_{F_t}(h, \Phi) &= \int_{\mathcal{X}} \ell^w_{h,\Phi}(x, t)\left(p_{1-t}(x) - p_t(x)\right) dx \\ &= \int_{\mathcal{R}} \ell^w_{h,\Phi}(\Psi(r), t)\left(p^\Phi_{1-t} - p^\Phi_t\right) dr \quad \because \text{Lemma A.2 of (Shalit et al., 2017)} \\ &= B_\Phi \cdot \int_{\mathcal{R}} \frac{\ell^w_{h,\Phi}(\Psi(r), t)}{B_\Phi}\left(p^\Phi_{1-t} - p^\Phi_t\right) dr \\ &\leq B_\Phi \cdot \sup_{g \in G}\left|\int_{\mathcal{R}} g(r)\left(p^\Phi_{1-t} - p^\Phi_t\right) dr\right| \\ &= B_\Phi \cdot \mathrm{IPM}_G\left(p^\Phi_t, p^\Phi_{1-t}\right)\end{aligned}$$

where Lemma A.2 of (Shalit et al., 2017) states the standerd changes of variable formula: $p^\Phi(t \mid r) = p(t \mid \Psi(r))$ and $p^\Phi(Y(t) \mid r) = p(Y(t) \mid \Psi(r))$ for all $r \in \mathcal{R}$ and $t \in \mathcal{T}$. □

***Lemma* 8.** *(Similar to Lemma A.5 of (Shalit et al., 2017)) Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation and $h : \mathcal{R} \times \mathcal{T} \to \mathcal{Y}$ be a hypothesis. We also define a regression function as $f_t(x) = h(\Phi(x), t)$. Then, for any given $t \in \mathcal{T}$ and $w : \mathcal{X} \times \mathcal{T} \to \mathcal{R}_{\geq 0}$, the following equalities hold:*

$$\begin{aligned}\int_{\mathcal{X}} w(x)\left(f_t(x) - m_t(x)\right)^2 p_t(x) dx &= \epsilon^w_{F_t}(h, \Phi) - \sigma^2_{t,w}(p_t) \\ \int_{\mathcal{X}} w(x)\left(f_t(x) - m_t(x)\right)^2 p_{1-t}(x) dx &= \epsilon^w_{CF_{1-t}}(h, \Phi) - \sigma^2_{t,w}(p_{1-t})\end{aligned}$$

Proof.

$$\begin{aligned}
\epsilon^{w}_{F_t}(h,\Phi) &= \int_{\mathcal{X}} \ell^{w}_{h,\Phi}(x,t)p_t(x)dx \\
&= \int_{\mathcal{X}\times\mathcal{Y}} w(x)\left(f_t(x)-Y(t)\right)^2 p\left(Y(t)|x\right)p_t(x)dY(t)dx \\
&= \int_{\mathcal{X}} w(x)\left(f_t(x)-m_t(x)\right)^2 p_t(x)dx \\
&\quad - 2\int_{\mathcal{X}\times\mathcal{Y}} w(x)\left(f_t(x)-m_t(x)\right)\left(Y(t)-m_t(x)\right)p\left(Y(t),x\mid t\right)dY(t)dx \\
&\quad + \int_{\mathcal{X}\times\mathcal{Y}} w(x)\left(Y(t)-m_t(x)\right)^2 p\left(Y(t),x\mid t\right)dY(t)dx \\
&= \int_{\mathcal{X}} w(x)\left(f_t(x)-m_t(x)\right)^2 p_t(x)dx + \sigma^2_{t,w}(p_t)
\end{aligned}$$

Thus, we have,

$$\int_{\mathcal{X}} w(x)\left(f_t(x)-m_t(x)\right)^2 p_t(x)dx = \epsilon^{w}_{F_t}(h,\Phi) - \sigma^2_{t,w}(p_t)$$

We can derive the analogous equality for counterfactual loss in the same manner. □

A.7. Proof of Theorem 5

Proof.

$$\begin{aligned}
&\mathbb{E}_X[\{\sum_{t\in\mathcal{T}}\sqrt{w_t(X)}(f_t(X)-m_t(X))\}^2] \\
&= \mathbb{E}_X[\{\sqrt{w_1(X)}(f_1(X)-m_1(X)) + \sqrt{w_0(X)}(f_0(X)-m_0(X))\}^2] \\
&\le 2\int_{\mathcal{X}} \Big(w_1(X)\left(f_1(X)-m_1(X)\right)^2 + w_0(X)\left(f_0(X)-m_0(X)\right)^2\Big)\,p(x)\,dx \quad \because (x+y)^2 \le 2(x^2+y^2) \\
&= 2\pi_1\int_{\mathcal{X}} w_1(X)\left(f_1(X)-m_1(X)\right)^2 p_1(x)\,dx + 2\pi_0\int_{\mathcal{X}} w_1(X)\left(f_1(X)-m_1(X)\right)^2 p_0(x)\,dx \\
&\quad + 2\pi_1\int_{\mathcal{X}} w_0(X)\left(f_0(X)-m_0(X)\right)^2 p_1(x)\,dx + 2\pi_0\int_{\mathcal{X}} w_0(X)\left(f_0(X)-m_0(X)\right)^2 p_0(x)\,dx \\
&= 2\pi_1\left(\epsilon^{w_1}_{F_1}(h,\Phi) - \sigma^2_{t=1,w_1}(p_1)\right) + 2\pi_0\left(\epsilon^{w_1}_{CF_0}(h,\Phi) - \sigma^2_{t=1,w_1}(p_0)\right) \\
&\quad + 2\pi_1\left(\epsilon^{w_0}_{CF_1}(h,\Phi) - \sigma^2_{t=0,w_0}(p_1)\right) + 2\pi_0\left(\epsilon^{w_0}_{F_0}(h,\Phi) - \sigma^2_{t=0,w_0}(p_0)\right) \quad \because \text{Lemma 8} \\
&\le 2\epsilon^{w_1}_{F_1}(h,\Phi) + 2\epsilon^{w_0}_{F_0}(h,\Phi) + 2B_\Phi\cdot \mathrm{IPM}_G\left(p^\Phi_t,\, p^\Phi_{1-t}\right) - 4\sigma^2 \quad \because \text{Lemma 7}
\end{aligned}$$

□

B. Detailed Experimental Settings

B.1. Model Selection Experiment in Section 5.2

The weighted counterfactual regression model used as a regression function of our proposed metric has some hyperparameters itself. To tune the hyperparameters of this model, we used the simple μ-risk as described in (Schuler et al., 2018). Table 2 describes the hyperparameter search spaces and the resulting set of hyperparameters for the weighted counterfactual regression .

B.2. Hyperparameter Tuning Experiment in Section 5.3

Table 3 provides the hyperparameter search space of the Gradient Boosting Regressor used in the hyperparameter tuning experiment in Section 5.3.

Table 2. Hyperparameter search spaces and the selected values of the hyperparameters for the weighted counterfactual regression in our proposed CFCV. A set of hyperparameters optimzed the μ-risk (Schuler et al., 2018) was selected for the weighted CFR.

Hyperparameters	Search spaces	Selected values
Num. of hidden layers for h and Φ in Eq. (12)	$\{1, 2, 3\}$	3
Dim. of hidden layers for h and Φ in Eq. (12)	$\{20, 50, 100\}$	100
trade-off parameter α in Eq. (12)	$[0.01, 100]$	0.356
learning_rate	$[0.0001, 0.01]$	4.292×10^{-4}
batch_size	256 (fixed)	256 (fixed)
dropout rate	0.2 (fixed)	0.2 (fixed)

Table 3. Hyperparameter search space for Gradient Boosting Regressors.

Hyperparameters	Search spaces
n_estimators	100 (fixed)
max_depth	$[1, 20]$
min_samples_leaf	$[1, 20]$
learning_rate	$[10^{-5}, 10^{-1}]$
subsample	$\{0.1, 0.2, \ldots 1.0\}$