A. CATE Interval

Lemma 2. *The unbiased estimate of the expected potential outcome under hidden confounding, given in Equation (3) has the following equivalent characterization:*

$$\mathbb{E}\left[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{X}=\mathbf{x}\right]=\mu(\mathbf{x},\mathrm{t})+\frac{\int(\mathrm{y}-\mu(\mathbf{x},\mathrm{t}))w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}{\int w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}. \tag{12}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{X}=\mathbf{x}\right] &= \mu(w_{\mathrm{t}};\mathbf{x},\mathrm{t}) && (13a)\\
&= \frac{\int \mathrm{y}w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy}{\int w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy} && (13b)\\
&= \int \mathrm{y}\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y} && (13c)\\
&= \mu(\mathbf{x},\mathrm{t})+\int \mathrm{y}\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y}-\mu(\mathbf{x},\mathrm{t}) && (13d)\\
&= \mu(\mathbf{x},\mathrm{t})+\int \mathrm{y}\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y}-\mu(\mathbf{x},\mathrm{t})\int\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y} && (13e)\\
&= \mu(\mathbf{x},\mathrm{t})+\int \mathrm{y}\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y}-\int\mu(\mathbf{x},\mathrm{t})\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y} && (13f)\\
&= \mu(\mathbf{x},\mathrm{t})+\int(\mathrm{y}-\mu(\mathbf{x},\mathrm{t}))\frac{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})}{\int w_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}'\mid\mathbf{x})dy'}d\mathrm{y} && (13g)\\
&= \mu(\mathbf{x},\mathrm{t})+\frac{\int(\mathrm{y}-\mu(\mathbf{x},\mathrm{t}))w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})e_{\mathrm{t}}(\mathbf{x})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}{\int w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})e_{\mathrm{t}}(\mathbf{x})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy} && (13h)\\
&= \mu(\mathbf{x},\mathrm{t})+\frac{\int(\mathrm{y}-\mu(\mathbf{x},\mathrm{t}))w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}{\int w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}. && (13i)
\end{aligned}$$

□

Lemma 3. *The bounds for the conditional expected potential outcomes* $\underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$ *and* $\overline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$ *defined in equations (4) have the following equivalent characterization:*

$$\begin{aligned}
\underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t}) &= \inf_{\mathrm{y}^*\in\mathcal{Y}}\mu(\mathbf{x},\mathrm{t})+\frac{\int_{-\infty}^{\mathrm{y}^*}\mathrm{r}_{\mathrm{t}}(\mathrm{y})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x})+\mathrm{P}(\mathrm{Y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t})},\\
\overline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t}) &= \sup_{\mathrm{y}^*\in\mathcal{Y}}\mu(\mathbf{x},\mathrm{t})+\frac{\int_{\mathrm{y}^*}^{\infty}\mathrm{r}_{\mathrm{t}}(\mathrm{y})f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x})+\mathrm{P}(\mathrm{Y}>\mathrm{y}^*\mid\mathbf{x},\mathrm{t})},
\end{aligned}$$

where $\mathrm{r}_{\mathrm{t}}(\mathrm{y})=(\mathrm{y}-\mu(\mathbf{x},\mathrm{t}))$ *and* $\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x})=\frac{\alpha_{\mathrm{t}}(\mathbf{x};\Gamma)}{\beta_{\mathrm{t}}(\mathbf{x};\Gamma)-\alpha_{\mathrm{t}}(\mathbf{x};\Gamma)}$.

Proof. We prove the result for $\underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$ and the result for $\overline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$ can be proved analogously. From Kallus et al. (2019) Lemma 1,

$$\begin{aligned}
\underline{\mu}(\mathbf{x},\mathrm{t}) &= \inf_{w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})\in[\alpha_{\mathrm{t}}(\mathbf{x};\Gamma),\beta_{\mathrm{t}}(\mathbf{x};\Gamma)]}\frac{\int \mathrm{y}w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy}{\int w_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy}\\
&= \inf_{u\in\mathcal{U}^{ni}}\frac{\alpha_{\mathrm{t}}(\mathbf{x};\Gamma)\int \mathrm{y}f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy+(\beta_{\mathrm{t}}(\mathbf{x};\Gamma)-\alpha_{\mathrm{t}}(\mathbf{x};\Gamma))\int u(\mathrm{y})\mathrm{y}f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy}{\alpha_{\mathrm{t}}(\mathbf{x};\Gamma)\int f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy+(\beta_{\mathrm{t}}(\mathbf{x};\Gamma)-\alpha_{\mathrm{t}}(\mathbf{x};\Gamma))\int u(\mathrm{y})f_{\mathrm{t}}(\mathrm{y}\mid\mathbf{x})dy}, && (15)\\
&= \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})
\end{aligned}$$

where $\mathcal{U}^{ni} = \{u : \mathcal{Y} \to [0,1] \mid u(\mathrm{y}) \text{ is non-increasing}\}$. Therefore, from the equivalence in Equation (13),

$$\begin{aligned}
\underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t}) &= \inf_{u \in \mathcal{U}^{ni}} \mu(\mathbf{x},\mathrm{t}) + \frac{\alpha_{\mathrm{t}}(\mathbf{x}) \int (\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y} + (\beta_{\mathrm{t}}(\mathbf{x}) - \alpha_{\mathrm{t}}(\mathbf{x})) \int u(\mathrm{y})(\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}(\mathbf{x}) \int f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y} + (\beta_{\mathrm{t}}(\mathbf{x}) - \alpha_{\mathrm{t}}(\mathbf{x})) \int u(\mathrm{y}) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}} \\
&= \inf_{u \in \mathcal{U}^{ni}} \mu(\mathbf{x},\mathrm{t}) + \frac{\alpha_{\mathrm{t}}(\mathbf{x})(\mu(\mathbf{x},\mathrm{t}) - \mu(\mathbf{x},\mathrm{t})) + (\beta_{\mathrm{t}}(\mathbf{x}) - \alpha_{\mathrm{t}}(\mathbf{x})) \int u(\mathrm{y})(\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}(\mathbf{x}) \int f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y} + (\beta_{\mathrm{t}}(\mathbf{x}) - \alpha_{\mathrm{t}}(\mathbf{x})) \int u(\mathrm{y}) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}} \\
&= \inf_{u \in \mathcal{U}^{ni}} \mu(\mathbf{x},\mathrm{t}) + \frac{(\beta_{\mathrm{t}}(\mathbf{x};\Gamma) - \alpha_{\mathrm{t}}(\mathbf{x};\Gamma)) \int u(\mathrm{y})(\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}(\mathbf{x};\Gamma) + (\beta_{\mathrm{t}}(\mathbf{x};\Gamma) - \alpha_{\mathrm{t}}(\mathbf{x};\Gamma)) \int u(\mathrm{y}) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}} \\
&= \inf_{u \in \mathcal{U}^{ni}} \mu(\mathbf{x},\mathrm{t}) + \frac{\int u(\mathrm{y})(\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x}) + \int u(\mathrm{y}) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}} \\
&= \inf_{y^* \in \mathcal{Y}} \mu(\mathbf{x},\mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x}) + \int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}} \\
&= \inf_{y^* \in \mathcal{Y}} \mu(\mathbf{x},\mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x}) + \mathrm{P}(\mathrm{Y} \leq \mathrm{y}^* \mid \mathbf{x},\mathrm{t})} \\
&= \inf_{y^* \in \mathcal{Y}} \mu(\mathbf{x},\mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} \mathrm{r}_{\mathrm{t}}(\mathrm{y}) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x}) + \mathrm{P}(\mathrm{Y} \leq \mathrm{y}^* \mid \mathbf{x},\mathrm{t})}.
\end{aligned}$$

□

B. CATE Interval Estimator

Proof for Theorem 1. Here we prove that $\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) \xrightarrow{p} \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$, from which $\overline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) \xrightarrow{p} \overline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$ can be proved analogously. Note that $\xrightarrow{p}$ indicates convergence in probability. As a reminder

$$\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) = \inf_{\mathrm{y}^* \in \mathcal{Y}} \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t})) f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x},\mathrm{t}) + \int_{-\infty}^{\mathrm{y}^*} f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}, \tag{16}$$

and

$$\underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t}) = \inf_{\mathrm{y}^* \in \mathcal{Y}} \mu(\mathbf{x},\mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma}(\mathbf{x}) + \mathrm{P}(\mathrm{Y} \leq \mathrm{y}^* \mid \mathbf{x},\mathrm{t})}. \tag{17}$$

Further, our assumptions are

1. $n \to \infty$, and $\mathbf{x} \in \mathcal{D}$.
2. Y is a bounded random variable.
3. $f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x},\mathrm{t})$ converges in measure to $f(\mathrm{y} \mid \mathbf{x},\mathrm{t})$. Specifically, $\lim_{n\to\infty} P(\{\mathrm{y} \in \mathcal{Y} : |f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) - f_{\boldsymbol{\omega}|\mathcal{D}_n}(\mathrm{y} \mid \mathbf{x},\mathrm{t})| \geq \epsilon\}) = 0$, for every $\epsilon \geq 0$, where $\mathcal{D}_n$ is a dataset of size n. Convergence in measure is a generalization of convergence in probability.
4. $\widehat{e}_{\mathrm{t},\boldsymbol{\omega}}(\mathbf{x})$ and $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x})$ are consistent estimators of $\mathbb{E}[\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}]$ and $\mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$.
5. $e_{\mathrm{t}}(\mathbf{x},\mathrm{y})$ is bounded away from 0 and 1 uniformly over $\mathbf{x} \in \mathcal{X}$, $\mathrm{y} \in \mathcal{Y}$, and $\mathrm{t} \in \{0,1\}$.

We need to show that $\lim_{n\to\infty} P(|\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) - \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})| \geq \epsilon) = 0$, for all $\epsilon > 0$, where the parameters $\boldsymbol{\omega}$ are dependent on the size of the dataset n. First, define the following quantities:

$$\begin{aligned}
\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x},\mathrm{t};n) &= \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t})) f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}, \quad & I_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x},\mathrm{t}) &= \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \mu(\mathbf{x},\mathrm{t})) f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}, \\
\kappa^{\mathrm{y}^*}(\mathbf{x},\mathrm{t};n) &= \int_{-\infty}^{\mathrm{y}^*} f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y}, & I^{\mathrm{y}^*}(\mathbf{x},\mathrm{t}) &= \int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y} \mid \mathbf{x},\mathrm{t}) d\mathrm{y},
\end{aligned}$$

so that

$$\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) = \inf_{\mathrm{y}^*\in\mathcal{Y}} \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) + \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x},\mathrm{t};n)}{\alpha'^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) + \kappa^{\mathrm{y}^*}(\mathbf{x},\mathrm{t};n)}, \quad \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t}) = \inf_{\mathrm{y}^*\in\mathcal{Y}} \mu(\mathbf{x},\mathrm{t}) + \frac{I_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x},\mathrm{t})}{\alpha'^{\Gamma}_{\mathrm{t}}(\mathbf{x}) + I^{\mathrm{y}^*}(\mathbf{x},\mathrm{t})}$$

To start, we want to express $|\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) - \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})|$ in terms of the following 4 terms: $\Delta_1(\mathbf{x},\mathrm{t};n) = |\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) - \mu(\mathbf{x},\mathrm{t})|$; $\Delta_2(\mathbf{x},\mathrm{t};n) = \left|\alpha'^{\Gamma}_{\boldsymbol{\omega}} - \alpha'^{\Gamma}_{\mathrm{t}}\right|$; $\Delta_3(\mathbf{x},\mathrm{t};n) = \sup_{\mathrm{y}^*\in\mathcal{Y}} |\delta_3|$, where $\delta_3 = \kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}$; and $\Delta_4(\mathbf{x},\mathrm{t};n) = \sup_{\mathrm{y}^*\in\mathcal{Y}} |\delta_4|$, where $\delta_4 = \kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}$. To this end, we use Lemma 3 from Kallus et al. (2019) in line 1 below and define the following inequality:

$$\begin{aligned}
|\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) - \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})| &\leq \sup_{\mathrm{y}^*\in\mathcal{Y}} \left| \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) - \mu(\mathbf{x},\mathrm{t}) + \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x},\mathrm{t};n)}{\alpha'^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) + \kappa^{\mathrm{y}^*}(\mathbf{x},\mathrm{t};n)} - \frac{I_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x},\mathrm{t})}{\alpha'^{\Gamma}_{\mathrm{t}}(\mathbf{x}) + I^{\mathrm{y}^*}(\mathbf{x},\mathrm{t})} \right|, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \sup_{\mathrm{y}^*\in\mathcal{Y}} \left| \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}} - \frac{I_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}} \right|, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \sup_{\mathrm{y}^*\in\mathcal{Y}} \left\{ |\kappa_{\mathrm{y}}^{\mathrm{y}^*}| \frac{|(\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}) - (\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*})|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} - \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} \right\}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \sup_{\mathrm{y}^*\in\mathcal{Y}} \left\{ \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\boldsymbol{\omega}} - \alpha'^{\Gamma}_{\mathrm{t}}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} \right\}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\boldsymbol{\omega}} - \alpha'^{\Gamma}_{\mathrm{t}}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|}, \\
&= \Delta_1(\mathbf{x},\mathrm{t};n) + |\alpha'^{\Gamma}_{\boldsymbol{\omega}} - \alpha'^{\Gamma}_{\mathrm{t}}| \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} \\
&= \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*} + I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*} + I_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|}, \\
&= \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3 + I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3 + I_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\mathrm{t}} + I^{\mathrm{y}^*}|}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3 + I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3 + I_{\mathrm{y}}^{\mathrm{y}^*}||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\mathrm{t}}|}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3| + |I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{(|\delta_3| + |I_{\mathrm{y}}^{\mathrm{y}^*}|)|\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\mathrm{t}}|}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}} + \kappa^{\mathrm{y}^*}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\mathrm{t}}|}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3||\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\mathrm{t}}|}, \\
&\leq \Delta_1(\mathbf{x},\mathrm{t};n) + \Delta_2(\mathbf{x},\mathrm{t};n) \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3||\delta_4|}{|\alpha'^{\Gamma}_{\boldsymbol{\omega}}||\alpha'^{\Gamma}_{\mathrm{t}}|} + \sup_{\mathrm{y}^*\in\mathcal{Y}} \frac{|\delta_3|}{|\alpha'^{\Gamma}_{\mathrm{t}}|}, \\
&= \Delta_1(\mathbf{x},\mathrm{t};n) + \frac{\Delta_2(\mathbf{x},\mathrm{t};n)\Delta_3(\mathbf{x},\mathrm{t};n)}{\alpha'^{\Gamma}_{\boldsymbol{\omega}}\alpha'^{\Gamma}_{\mathrm{t}}} + \frac{\Delta_3(\mathbf{x},\mathrm{t};n)\Delta_4(\mathbf{x},\mathrm{t};n)}{\alpha'^{\Gamma}_{\boldsymbol{\omega}}\alpha'^{\Gamma}_{\mathrm{t}}} + \frac{\Delta_3(\mathbf{x},\mathrm{t};n)}{\alpha'^{\Gamma}_{\mathrm{t}}}.
\end{aligned}$$

So, we now need only prove that $\Delta_1(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$, $\Delta_2(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$, $\Delta_3(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$, and $\Delta_4(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$, when $n \to \infty$. Note that both $\Delta_1(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$ and $\Delta_2(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$ are covered by Assumption 4 of Theorem 1; namely, $\widehat{e}_{\mathrm{t},\boldsymbol{\omega}}(\mathbf{x})$ *and* $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x})$ *are consistent estimators of* $\mathbb{E}[\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}]$ *and* $\mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$.

First, we prove that $\Delta_4(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$.

Prove that $\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}\right| \xrightarrow{p} 0$

$$\begin{aligned}
\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}\right| &= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} f_{\boldsymbol{\omega}}(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy - \int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy\right| \\
&= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|P_{\boldsymbol{\omega}}(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t}) - P(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t})\right|
\end{aligned}$$

Convergence in probability implies convergence in distribution ($\lim_{n\to\infty} P_n(\mathbf{X}\leq\mathbf{x}) = P(\mathbf{X}\leq\mathbf{x})$), so by Assumption 3 in Theorem 1

$$\begin{aligned}
\lim_{n\to\infty} P\left(\left|\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\kappa^{\mathrm{y}^*} - I^{\mathrm{y}^*}\right|\right| \geq \epsilon\right) &= \lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|P_{\boldsymbol{\omega}}(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t}) - P(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t})\right| \geq \epsilon\right) \\
&= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|P(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t}) - P(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t})\right| \geq \epsilon\right) \\
&= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}} |0| \geq \epsilon\right) \\
&= P\left(0 \geq \epsilon\right) \\
&= 0
\end{aligned}$$

Finally, we prove $\Delta_3(\mathbf{x},\mathrm{t};n) \xrightarrow{p} 0$.

Prove that $\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right| \xrightarrow{p} 0$

$$\begin{aligned}
\sup_{\mathrm{y}^*\in\mathcal{Y}} |\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}| &= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}))f_{\boldsymbol{\omega}}(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy - \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - \mu(\mathbf{x},\mathrm{t}))f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy\right|, \\
&= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy + \mu(\mathbf{x},\mathrm{t})\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t})\int_{-\infty}^{\mathrm{y}^*} f_{\boldsymbol{\omega}}(\mathrm{y}\mid\mathbf{x},\mathrm{t})dy\right|, \\
&= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy + \mu\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y})dy - \widehat{\mu}_{\boldsymbol{\omega}}\int_{-\infty}^{\mathrm{y}^*} f_{\boldsymbol{\omega}}(\mathrm{y})dy\right|, \\
&= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy + (\mu - \widehat{\mu}_{\boldsymbol{\omega}} + \widehat{\mu}_{\boldsymbol{\omega}})\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y})dy - \widehat{\mu}_{\boldsymbol{\omega}}\int_{-\infty}^{\mathrm{y}^*} (f_{\boldsymbol{\omega}}(\mathrm{y}) - f(\mathrm{y}) + f(\mathrm{y}))dy\right|, \\
&= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy + (\mu - \widehat{\mu}_{\boldsymbol{\omega}})\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y})dy - \widehat{\mu}_{\boldsymbol{\omega}}\int_{-\infty}^{\mathrm{y}^*} (f_{\boldsymbol{\omega}}(\mathrm{y}) - f(\mathrm{y}))\,dy\right|, \\
&= \sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy + (\mu - \widehat{\mu}_{\boldsymbol{\omega}})\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y})dy - \widehat{\mu}_{\boldsymbol{\omega}}\left(P_{\boldsymbol{\omega}}(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t}) - P(\mathrm{y}\leq\mathrm{y}^*\mid\mathbf{x},\mathrm{t})\right)\right|.
\end{aligned}$$

As a first step, we can use the result for $\Delta_4(\mathbf{x},\mathrm{t};n)$ to remove the green term from the supremum and now we need to show that

$$\lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy + (\mu - \widehat{\mu}_{\boldsymbol{\omega}})\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y})dy\right| \geq \epsilon.\right) = 0$$

Next, under assumption 4 of Theorem 1 we have $\mu(\mathbf{x},\mathrm{t}) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) \xrightarrow{p} 0$, and we are left finally to show that

$$\lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}} \left|\int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy\right| \geq \epsilon.\right) = 0$$

Assumption 2 of Theorem 1 states that Y is a bounded random variable. As such, there exists a $g(\mathrm{y})$ such that $|\mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})| \leq g(\mathrm{y})$ for all n and $y \in \mathcal{Y}$. Therefore, in conjunction with Assumption 3, by Lebesgue's dominated convergence theorem we have $\lim_{n\to\infty} \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f_{\boldsymbol{\omega}}(\mathrm{y})dy = \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} f(\mathrm{y})dy$

$$\begin{aligned}\lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}f_{\boldsymbol{\omega}}(\mathrm{y})dy-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}f(\mathrm{y})dy\right|\geq\epsilon.\right) &= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\lim_{n\to\infty}\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}f_{\boldsymbol{\omega}}(\mathrm{y})dy-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}f(\mathrm{y})dy\right|\geq\epsilon.\right)\\ &= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}f(\mathrm{y})dy-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}f(\mathrm{y})dy\right|\geq\epsilon.\right)\\ &= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}|0|\geq\epsilon.\right)\\ &= 0\end{aligned}$$

Therefore, $\widehat{\underline{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) \xrightarrow{p} \underline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$, and $\widehat{\overline{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x},\mathrm{t}) \xrightarrow{p} \overline{\mu}^{\Gamma}(\mathbf{x},\mathrm{t})$ can be proved analogously, which concludes our proof that both $\widehat{\underline{\tau}}_{\boldsymbol{\omega}}(\mathbf{x};\Gamma) \xrightarrow{p} \underline{\tau}(\mathbf{x};\Gamma)$, and $\widehat{\overline{\tau}}_{\boldsymbol{\omega}}(\mathbf{x};\Gamma) \xrightarrow{p} \overline{\tau}(\mathbf{x};\Gamma)$. □

C. Datasets

C.1. Simulated Data

The simulated dataset presented by Kallus et al. (2019) is described by the following structural causal model (SCM):

$$\mathrm{u} := N_{\mathrm{u}}, \tag{22a}$$

$$\mathrm{x} := N_{\mathrm{x}}, \tag{22b}$$

$$\mathrm{t} := N_{\mathrm{t}}, \tag{22c}$$

$$\mathrm{y} := (2\mathrm{t}-1)\mathrm{x} + (2\mathrm{t}-1) - 2\sin(2(2\mathrm{t}-1)\mathrm{x}) - 2(2\mathrm{u}-1)(1+0.5\mathrm{x}) + N_{\mathrm{y}}, \tag{22d}$$

where $N_{\mathrm{u}} \sim \mathrm{Bern}(0.5)$, $N_{\mathrm{x}} \sim \mathrm{Unif}[-2,2]$, $N_{\mathrm{u}} \perp\!\!\!\perp N_{\mathrm{x}}$, $N_{\mathrm{t}} \sim \mathrm{Bern}(e(\mathrm{x},\mathrm{u}))$, $e(\mathrm{x},\mathrm{u}) = \frac{\mathrm{u}}{\alpha_{\mathrm{t}}(\mathrm{x};\Gamma^*)} + \frac{1-\mathrm{u}}{\beta_{\mathrm{t}}(\mathrm{x};\Gamma^*)}$, $e(\mathrm{x}) = \mathrm{sigmoid}(0.75\mathrm{x}+0.5)$, and $N_{\mathrm{y}} \sim \mathcal{N}(0,1)$.

Remember that only x, t, and y are observed. So the bias induced by hidden confounding at x is given by

$$\tilde{\tau}(\mathrm{x}) - \tau(\mathrm{x}) = 2(2+\mathrm{x})\left(P(\mathrm{u}=1 \mid \mathrm{T}=1, \mathrm{X}=\mathrm{x}) - P(\mathrm{u}=1 \mid \mathrm{T}=0, \mathrm{X}=\mathrm{x})\right), \tag{23}$$

where $\tilde{\tau}(\mathrm{x})$ is the confounded CATE estimate.

Each random realization of the simulated dataset generates 1000 training examples, 100 validation examples, and 1000 test examples. In the experiments we report results over 50 random realizations. The seeds for the random number generators are i, $i+1$, and $i+2$; $\{i \in [0,1,\ldots,49]\}$, for the training, validation, and test sets, respectively. Code is available in file /library/datasets/synthetic.py on github at `https://github.com/anndvision/quince`.

C.2. HC-MNIST

HC-MNIST is an extension of the above dataset with high-dimensional covariates $\mathbf{x}$. Specifically, $\mathbf{x}$ are MNIST digits. HC-MNIST is described by the following SCM:

$$\mathrm{u} := N_{\mathrm{u}}, \tag{24a}$$

$$\mathbf{x} := N_{\mathbf{x}}, \tag{24b}$$

$$\phi := \left(\text{clip}\left(\frac{\mu_{N_{\mathbf{x}}} - \mu_{\mathrm{c}}}{\sigma_{\mathrm{c}}}; -1.4, 1.4\right) - \text{Min}_{\mathrm{c}}\right)\frac{\text{Max}_{\mathrm{c}} - \text{Min}_{\mathrm{c}}}{1.4 - \text{-}1.4} \tag{24c}$$

$$\mathrm{t} := N_{\mathrm{t}}, \tag{24d}$$

$$\mathrm{y} := (2\mathrm{t}-1)\phi + (2\mathrm{t}-1) - 2\sin(2(2\mathrm{t}-1)\phi) - 2(2\mathrm{u}-1)(1+0.5\phi) + N_{\mathrm{y}}, \tag{24e}$$

where N_{u}, N_{t} (swapping x for ϕ), and N_{y} are as described in Appendix C.1. $N_{\mathbf{x}}$ is a sample of an MNIST image. The sampled image has a corresponding label $c \in [0,\ldots,9]$. $\mu_{N_{\mathbf{x}}}$ is the average intensity of the sampled image. μ_{c} and σ_{c} are the mean and standard deviation of the average image intensities over all images with label c in the MNIST training set.

In other words, $\mu_{\mathrm{c}} = \mathbb{E}[\mu_{N_{\mathbf{x}}} \mid \mathrm{c}]$ and $\sigma_{\mathrm{c}}^2 = \mathrm{Var}[\mu_{N_{\mathbf{x}}} \mid \mathrm{c}]$. To map the high dimensional images $\mathbf{x}$ onto a one-dimensional manifold ϕ with the same domain as $\mathrm{x} \in [-2, 2]$ above, we first clip the standardized average image intensity on the range $(-1.4, 1.4)$. Each digit class has its own domain in ϕ, so there is a linear transformation of the clipped value onto the range $[\mathrm{Min}_{\mathrm{c}}, \mathrm{Max}_{\mathrm{c}}]$. Finally, $\mathrm{Min}_{\mathrm{c}} = -2 + \frac{4}{10}\mathrm{c}$, and $\mathrm{Max}_{\mathrm{c}} = -2 + \frac{4}{10}(\mathrm{c} + 1)$.

For each random realization of the dataset, the MNIST training set is split into training ($n = 35000$) and validation ($n = 15000$) subsets using the scikit-learn function train_test_split(). The test set is generated using the MNIST test set ($n = 10000$). The random seeds are $\{i \in [0, 1, \dots, 19]\}$ for the 20 random realizations generated. Code to generate this dataset is available in file /library/datasets/hcmnist.py on github at `https://github.com/anndvision/quince`.

C.3. IHDP Hidden Confounding

The experimental data from the Infant Health and Development Program (IHDP) are used by Hill (2011) to generate simulated outcomes. The treatment group reveives "intensive high-quality child care and home visits from a trained provider." Hill (2011) uses "measurements on the child–birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status–as well as behaviors engaged in during pregnancy–smoked cigarettes, drank alcohol, took drugs–and measurements on the mother at the time she gave birth–age, marital status, educational attainment (did not graduate from high school, graduated from high school, attended some college but did not graduate, graduated from college), whether she worked during pregnancy, whether she received prenatal care–and the site (8 total) in which the family resided at the start of the intervention. There are 6 continuous covariates and 19 binary covariates." Hill (2011) excludes "a nonrandom portion of the treatment group: all children with nonwhite mothers," in order to simulate an observational study. Table 3 enumerates the included covariates. There are 139 examples in the treatment group and 608 examples in the control group, for a total of 747 examples.

Table 3. **IHDP Covariates** Binary covariates $x_9 - \mathrm{x}_{18}$ are attributes of the mother. Mother's education level "College" indicated by covariates $\mathrm{x}_{10} - \mathrm{x}_{12}$ all zero. Site 8 indicated by covariates $\mathrm{x}_{19} - \mathrm{x}_{25}$ all zero. We show the frequency of occurrence for each binary covariate $p(\mathrm{x} = 1)$, as well as the adjusted mutual information $I(\mathrm{x}; \mathrm{t})$ between the binary covariate and the treatment variable.

Continuous		Binary			
Covariate	Description	Covariate	Description	$I(\mathrm{x}; \mathrm{t})$	$p(\mathrm{x} = 1)$
x_1	birthweight	x_7	child's gender (female=1)	0.00	0.51
x_2	head circumference	x_8	is child a twin	0.00	0.09
x_3	number of weeks pre-term	x_9	married when child born	**0.02**	**0.52**
x_4	birth order	x_{10}	left High School	0.00	0.36
x_5	"neo-natal health index"	x_{11}	completed High School	0.00	0.27
x_6	mom's age	x_{12}	some College	0.00	0.22
		x_{13}	child is first born	0.00	0.36
		x_{14}	smoked cigarettes when pregnant	**0.01**	**0.48**
		x_{15}	consumed alcohol when pregnant	0.00	0.14
		x_{16}	used drugs when pregnant	0.00	0.96
		x_{17}	worked during pregnancy	**0.01**	**0.59**
		x_{18}	received any prenatal care	**0.01**	0.96
		x_{19}	site 1	0.00	0.14
		x_{20}	site 2	**0.01**	0.14
		x_{21}	site 3	0.00	0.16
		x_{22}	site 4	**0.01**	0.08
		x_{23}	site 5	**0.02**	0.07
		x_{24}	site 6	**0.01**	0.13
		x_{25}	site 7	**0.02**	0.16

Response surface B, designed by Hill (2011), is described by the following SCM:

$$
\begin{aligned}
\mathbf{x} &\coloneqq N_{\mathbf{x}}, && (25a)\\
\mathrm{t} &\coloneqq N_{\mathrm{t}}, && (25b)\\
\mathrm{y} &\coloneqq (\mathrm{t}-1)\left(\exp(\beta_{\mathbf{x}}(\mathbf{x}+\mathbf{w})) + N_{\mathrm{Y}^0}\right) + \mathrm{t}\left(\beta_{\mathbf{x}}\mathbf{x} - \omega^s + N_{\mathrm{Y}^1}\right), && (25c)
\end{aligned}
$$

where $(N_{\mathbf{x}}, N_{\mathrm{t}}) \sim p_{\mathcal{D}}(\{\mathrm{x}_1, \dots \mathrm{x}_{25}\}, \mathrm{t})$, $N_{\mathrm{Y}^0} \sim \mathcal{N}(0,1)$, and $N_{\mathrm{Y}^1} \sim \mathcal{N}(0,1)$. The coefficients $\beta_{\mathbf{x}}$ are a vector of randomly sampled values $(0.0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$. Hill (2011) describes ω^s as follows: "For the sth simulation, $[\boldsymbol{\omega}^s]$ is chosen in the overlap setting, where we estimate the effect of the treatment on the treated [(CATT)], such that CATT equals 4; similarly it was chosen in the incomplete setting, where we estimate the effect of the treatment on the controls [(CATC)], so that CATC equals 4." An offset vector $\mathbf{w}$, equal in dimension to $\mathbf{x}$, with every value set to 0.5, is added to $\mathbf{x}$.

To induce hidden confounding, we need to select a variable u that is associated with the treatment that will be hidden from the CATE interval estimator, and design a response surface where the outcome will always be affected by u. In Table 3, we list 3 potential candidates for u: x_9, x_{14}, and x_{17}. Each of these variable have a non-negligible association with the treatment, as indicated by the adjusted mutual information score $I(\mathrm{x};\mathrm{t})$, and have a frequency of taking the value 1 at around 0.5 (increasing the chances that we will have both positive and negative examples in each of the training, validation, and testing splits). Here we select x_9 and define the following SCM:

$$
\begin{aligned}
\mathrm{u} &\coloneqq N_{\mathrm{u}}, && (26a)\\
\mathbf{x} &\coloneqq N_{\mathbf{x}}, && (26b)\\
\mathrm{t} &\coloneqq N_{\mathrm{t}}, && (26c)\\
\mathrm{y} &\coloneqq (\mathrm{t}-1)(\exp(\beta_{\mathbf{x}}(\mathbf{x}+\mathbf{w}) + \beta_{\mathrm{u}}(\mathrm{u}+0.5)) + N_{\mathrm{Y}^0}) + \mathrm{t}(\beta_{\mathbf{x}}\mathbf{x} + \beta_{\mathrm{u}}\mathrm{u} - \omega^s + N_{\mathrm{Y}^1})), && (26d)
\end{aligned}
$$

where $(N_{\mathrm{u}}, N_{\mathbf{x}}, N_{\mathrm{t}}) \sim p_{\mathcal{D}}(\mathrm{x}_9, \{\mathrm{x}_1, \dots, \mathrm{x}_8, \mathrm{x}_{10}, \dots, \mathrm{x}_{25}\}, \mathrm{t})$, $N_{\mathrm{Y}^0} \sim \mathcal{N}(0,1)$, and $N_{\mathrm{Y}^1} \sim \mathcal{N}(0,1)$. The coefficient β_{u} is randomly sampled from $(0.1, 0.2, 0.3, 0.4, 0.5)$ with probabilities $(0.2, 0.2, 0.2, 0.2, 0.2)$. The remaining parameters–$\beta_{\mathbf{x}}$, ω^s, and $\boldsymbol{\omega}$–are given as above, taking into account u.

For each random realization of the dataset, the IHDP data is split into training ($n = 470$), validation ($n = 202$) and test ($n = 75$) subsets using the scikit-learn function train_test_split(). The random seeds for both splitting and outcome generation are $\{i \in [0, 1, \dots, 999]\}$ for the 1000 realizations generated. Code to generate this dataset is available in file /library/datasets/ihdp.py on github at `https://github.com/anndvision/quince`.

D. Implementation Details

Experiments for the Simulated and IHDP datasets were run using a single NVIDIA GeForce GTX 1080 ti, an Intel(R) Core(TM) i7-8700K, on a desktop computer with 16GB of RAM. Experiments for the HCMNIST dataset were run using 4 NVIDIA GeForce RTX 2080 ti GPUs, an Intel(R) Core(TM) i9-9900K, on a server with 64GB of RAM. Code is written in python. Packages used include PyTorch (Paszke et al., 2019), scikit-learn (Pedregosa et al., 2011), Ray (Moritz et al., 2018), NumPy, SciPy, and Matplotlib. We use ray tune (Liaw et al., 2018) with the hyperopt (Bergstra et al., 2013) search algorithm to optimize our network hyper-parameters. The hyper-parameters we consider are accounted for in Table 4. The hyper-paramter optimization objective for each dataset is the expected batch-wise log-likelihood of the validation data for a single dataset realization with random seed 1331.

Each experiment is replicated using the training, validation, and testing datasets described in the previous section.

Code to replicate these experiments is available at `https://github.com/anndvision/quince`.

D.1. Simulated Data

As a reminder, we need parametric models for the distribution over outcomes $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ and the nominal propensity $\widehat{e}_{\boldsymbol{\omega}}(\mathbf{x})$. For $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$, we use a 4 hidden layer mixture density network (MDN) (Bishop, 1994) with 5 mixture components. The 1D treatment variable t and 1D covariate $\mathbf{x}$ are concatenated to make a 2D network input. Each hidden layer is comprised of a 100 neuron linear transformation, followed by a ReLU activation function. The MDN parameters are inferred using a linear layer to predict the 5 means, a linear layer followed by a softplus activation to predict the square root of the 5 variances, and

Hyper-parameter	Search Space
hidden units	[50, 100, 200, 400]
network depth	[1, 2, 3, 4, 5]
negative slope	[ReLU, 0.1, 0.2, 0.3, 0.4, 0.5, ELU]
dropout rate	[0.00, 0.10, 0.15, 0.20, 0.25, 0.50]
spectral norm	[0.95, 1.0, 2.5, 3.0, 6.0, 12.0, 24.0]
batch size	[16, 32, 64, 100, 200]
learning rate	[5e-5, 1e-4, 2e-4, 5e-4, 1e-3]

Table 4. Hyper-parameter search space

Hyper-parameter	Simulated	HCMNIST	IHDP
hidden units	200	200	200
network depth	4	2	4
negative slope	ReLU	ReLU	LeakyReLU 0.3
dropout rate	0.10	0.15	0.5
spectral norm	6.0	3.0	6.0
batch size	32	200	200
learning rate	1e-3	5e-4	5e-4

Table 5. Final hyper-parameters for each dataset

a linear layer to predict the logits of the 5 mixture components. We use the pytorch MixtureSameFamily distribution, with mixture_distribution=Categorical(.), and component_distribution=Normal(.) (Paszke et al., 2019). The objective function optimized is the negative log likelihood for the label y of the above distribution with the parameters predicted from (x, t). Dropout is applied to the inputs of each layer after the input layer with a rate of 0.1. Spectral normalization is applied to the weights of the networks with value 6.0. For $\widehat{e}_{\boldsymbol{\omega}}(\mathbf{x})$, we use a 4 hidden layer neural network with Bernoulli likelihood. Each hidden layer is comprised of a 200 neuron linear transformation, followed by a ReLU activation function. Spectral normalization is applied to the weights of the networks with value 6.0. The objective function optimized is the negative log likelihood for the observed treatment t of the Bernoulli distribution with the logits predicted from x. For both models, We use Adam optimization with default pytorch parameters (Kingma & Ba, 2017). We use a batch size of 32. We use early stopping based on the objective function value on the validation set with a patience of 20 epochs and train for a maximum of 500 epochs. We train an ensemble of 10 models as an estimation of Bayesian model averaging. At test time, we do 10 MC samples, corresponding to a forward pass of each model in the ensemble for $\boldsymbol{\omega}$ and 100 MC samples for y, for each model under $\mathrm{t} = 0$ and $\mathrm{t} = 1$.

Hyper-parameter selection The hyper parameter search space is given in Table 4 and a summary of the final hyper-parameters used are given in Table 5 under the column Simulated. Because the hidden confounding is a binary variable, it induces a bi-modal distribution in y at $\mathbf{x}$, as shown in Figure 5. In practice, we would not know the form of the distribution of y at $\mathbf{x}$. To this end we select 5 mixture components for the MDN to show that we can over estimate the true modality, and still obtain sensible results. Alternatively, the validation set could be used to find the number of components that minimizes negative log likelihood of the data. The number of MC samples are chosen based on the stability of network predictions, i.e. we increase the number of MC samples until the variances with respect to $\boldsymbol{\omega}$ or y no longer change significantly.

D.2. HC-MNIST

For $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$, we use a ResNet CNN feature extractor with 2 residual blocks, followed by a 2 hidden layer MDN with 5 mixture components. The 1D treatment variable t and ResNet output are concatenated to make a 49 dimensional MDN input. Each hidden layer in the MDN is comprised of a 200 neuron linear transformation, followed by a ReLU activation function. The MDN parameters are inferred using in the same manner as for the Simulated data above. The objective function optimized is the negative log likelihood for the label y of the above distribution with the parameters predicted from (x, t). Dropout is applied to the inputs of each layer after the input layer with a rate of 0.15. Spectral normalization is applied to the weights of the network with value 3.0. For $\widehat{e}_{\boldsymbol{\omega}}(\mathbf{x})$, we use a ResNet CNN feature extractor with 2 residual

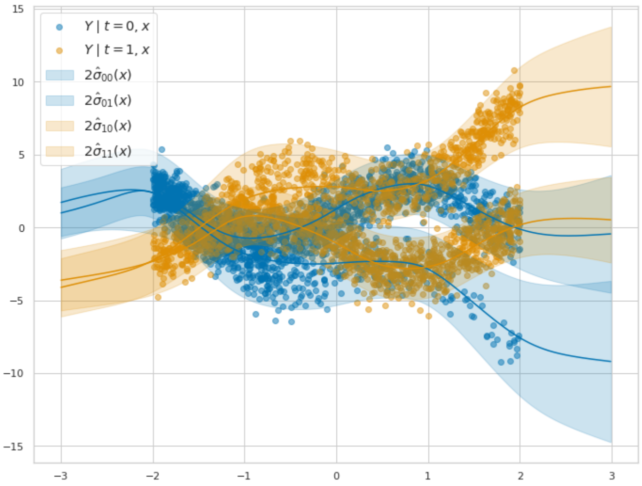

Figure 5. **Hidden confounding induces a multi-modal distribution in** y **at** x

blocks, followed by a 2 hidden layer neural network with Bernoulli likelihood. Each hidden layer of the Neural Network is comprised of a 200 neuron linear transformation, followed by a ReLU activation function. Dropout is applied to the inputs of each layer after the input layer with a rate of 0.15. Spectral normalization is applied to the weights of the network with value 3.0. The objective function optimized is the negative log likelihood for the observed treatment t of the Bernoulli distribution with the logits predicted from x. For both models, We use Adam optimization with a learning rate of 5e-4 (Kingma & Ba, 2017). We use a batch size of 200. We use early stopping based on the objective function value on the validation set with a patience of 20 epochs and train for a maximum of 500 epochs. We train an ensemble of 5 models as an estimation of Bayesian model averaging. At test time, we do 5 MC samples, corresponding to a forward pass of each model in the ensemble for $\boldsymbol{\omega}$ and 100 MC samples for y, for each model under $\mathrm{t} = 0$ and $\mathrm{t} = 1$.

D.3. IHDP Hidden Confounding

For $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$, we use a neural network feature extractor with 4 hidden layers, followed by a 2 hidden layer MDN with 5 mixture components. The 1D treatment variable t and feature extractor output are concatenated to make a 201 dimensional MDN input. Each hidden layer in the feature extractor and MDN is comprised of a 200 neuron linear transformation, followed by an LeakyReLU activation function. The MDN parameters are inferred using in the same manner as for the Simulated data above. The objective function optimized is the negative log likelihood for the label y of the above distribution with the parameters predicted from (x, t). Dropout is applied to the inputs of each layer after the input layer with a rate of 0.5. Spectral normalization is applied to the weights of the network with value 6.0. For $\widehat{e}_{\boldsymbol{\omega}}(\mathbf{x})$, we use a neural network feature extractor with 3 hidden layers, followed by a 2 hidden layer neural network with Bernoulli likelihood. Each hidden layer of the Neural Network is comprised of a 200 neuron linear transformation, followed by a ELU activation function. Dropout is applied to the inputs of each layer after the input layer with a rate of 0.5. Spectral normalization is applied to the weights of the network with value 6.0. The objective function optimized is the negative log likelihood for the observed treatment t of the Bernoulli distribution with the logits predicted from x. For both models, We use Adam optimization with a learning rate of 5e-4 (Kingma & Ba, 2017). We use a batch size of 200. We use early stopping based on the objective function value on the validation set with a patience of 20 epochs and train for a maximum of 500 epochs. We train an ensemble of 10 models as an estimation of Bayesian model averaging. At test time, we do 10 MC samples, corresponding to a forward pass of each model in the ensemble for $\boldsymbol{\omega}$ and 100 MC samples for y, for each model under $\mathrm{t} = 0$ and