A Theoretical Results

A.1 τ-BALD

Theorem 1. *Under the following assumptions:*

1. *Unconfoundedness* $(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$*;*
2. *Consistency* $\mathrm{Y} \mid \mathrm{T} = \mathrm{Y}^{\mathrm{t}}$*;*
3. Y^1 *and* Y^0*, when conditioned on realizations* $\mathbf{x}$ *of the r.v.* $\mathbf{X}$ *and* t *of the r.v.* T*, are independent-normally distributed or joint-normally distributed r.v.s.*
4. $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$ *is a consistent estimator of* $\mathbb{E}[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}]$

the information gain for $\boldsymbol{\Omega}$ *if we could observe a label for the difference in potential outcomes* $\mathrm{Y}^1 - \mathrm{Y}^0$ *given measured covariates* $\mathbf{x}$*, treatment* t *and a dataset of observations* $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$ *is approximated as*

$$\mathrm{I}(\mathrm{Y}^1 - \mathrm{Y}^0; \boldsymbol{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \approx \underset{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0)\right) \tag{10}$$

Proof.

$$\mathrm{I}(\mathrm{Y}^1 - \mathrm{Y}^0; \boldsymbol{\Omega} \mid \mathbf{x}, \mathcal{D}_{\text{train}}) = \mathrm{H}(\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \mathcal{D}_{\text{train}}) - \underset{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathbb{E}} \left[\mathrm{H}(\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega})\right] \tag{11a}$$

$$\approx \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \mathcal{D}_{\text{train}}) - \underset{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathbb{E}} \left[\mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega})\right] \tag{11b}$$

$$= \underset{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} \left(\mathbb{E}[\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}]\right) \tag{11c}$$

$$= \underset{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0)\right) \tag{11d}$$

In (11a) we adapt the result of Houlsby et al. [18] and express the information gain as the mutual information between the observable difference in potential outcomes $\mathrm{Y}^1 - \mathrm{Y}^0$ and the parameters $\boldsymbol{\Omega}$; given observed covariates $\mathbf{x}$, treatment t, and training data $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)\}_{i=1}^{n_{\text{train}}}$. In (11b) we apply lemma 1.1 to the r.h.s terms of (11a). We then use the result in Jesson et al. [20] and move from (11b) to (11c) by application of the law of total variance. Finally, under the consistency and unconfoundedness assumptions we express the information gain in terms of the identifiable difference in expected outcomes $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0)$. □

Lemma 1.1. *Under the following assumptions:*

1. $\mathrm{Y}^1, \mathrm{Y}^0$ *are independent-normally distributed or joint-normally distributed r.v.s;*
2. *With* $A = \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0)$*: let* $|A - 1| \leq 1$ *and* $A \neq 0$*. That is to say, the predictive variance must be greater than 0 and less than or equal to 2;*

$$\mathrm{H}(\mathrm{Y}^1 - \mathrm{Y}^0) \approx \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0) \tag{12}$$

Proof. By assumption 1, $\mathrm{Y}^1 - \mathrm{Y}^0$ is also a normally distributed random variable. By corollary 1.1,

$$\mathrm{H}(\mathrm{Y}^1 - \mathrm{Y}^0) = \frac{1}{2} + \frac{1}{2} \log(2\pi \, \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0)) \tag{13}$$

So given assumption 2, the first order Taylor polynomial of $\mathrm{H}(\mathrm{Y}^1 - \mathrm{Y}^0)$ is

$$\begin{aligned} \frac{1}{2} + \frac{1}{2} \log(2\pi \, \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0)) &\approx \frac{1}{2} + \frac{1}{2}(2\pi \, \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0) - 1) \\ &= \frac{1}{2} + \pi \, \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0) - \frac{1}{2} \\ &= \pi \, \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0) \\ &\propto \mathrm{Var}(\mathrm{Y}^1 - \mathrm{Y}^0) \end{aligned} \tag{14}$$

□

Corollary 1.1. *The entropy of a normally distributed random variable with variance σ^2 is $\frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2)$*

A.2 μ-BALD

Theorem 2. *Under the following assumptions:*

1. *Unconfoundedness* $(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$,
2. *Consistency* $\mathrm{Y} \mid \mathrm{T} = \mathrm{Y}^{\mathrm{t}}$,
3. Y *conditioned on* $\mathbf{x}$ *and* t *is a normally distributed random variable,*
4. $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$ *is a consistent estimator of* $\mathbb{E}[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}]$,

the information gain for $\boldsymbol{\Omega}$ when we observe a label for the potential outcome Y^{t} given measured covariates $\mathbf{x}$, treatment t and a dataset of observations $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$ can be approximated as is

$$\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \boldsymbol{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \approx \frac{1}{2} \log \left(\frac{\mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}})}{\mathbb{E}_{\boldsymbol{\omega}}[\mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\omega})]} \right), \tag{15}$$

or

$$\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \boldsymbol{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \approx \underset{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) \right). \tag{16}$$

Equation (15) expresses the information gain as the logarithm of a ratio between predictive and aleatoric uncertainty in the outcome. Whereas, equation (15) expresses the information gain as a direct estimate of the epistemic uncertainty.

Proof.

$$\begin{aligned}
\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \boldsymbol{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) &= \mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) - \mathop{\mathbb{E}}_{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} [\mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\omega})] && \text{(17a)} \\
&= \frac{1}{2} \log \left(2\pi \, \mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \right) - \mathop{\mathbb{E}}_{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} \frac{1}{2} \log \left(2\pi \, \mathrm{Var}(\mathrm{Y} \mid \boldsymbol{\omega}, \mathbf{x}, \mathrm{t}) \right) && \text{(17b)} \\
&\geq \frac{1}{2} \log \left(2\pi \, \mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \right) - \frac{1}{2} \log \left(2\pi \mathop{\mathbb{E}}_{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} \mathrm{Var}(\mathrm{Y} \mid \boldsymbol{\omega}, \mathbf{x}, \mathrm{t}) \right) && \text{(17c)} \\
&= \frac{1}{2} \log \left(\frac{\mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}})}{\mathbb{E}_{\boldsymbol{\omega}}[\mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\omega})]} \right) && \text{(17d)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \boldsymbol{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) &= \mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) - \mathop{\mathbb{E}}_{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} [\mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\omega})] && \text{(18a)} \\
&\approx \mathrm{Var}[\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} [\mathrm{Var}[\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\omega}]] && \text{(18b)} \\
&= \underset{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) \right) && \text{(18c)}
\end{aligned}$$

In (18a) we express the information gain as the mutual information between the observed potential outcome Y^{t} and the parameters $\boldsymbol{\Omega}$; given observed covariates $\mathbf{x}$, treatment t, and training data $\mathcal{D}_{\text{train}}$. By consistency, we can drop the superscript on the potential outcome. In (18b) we approximate the r.h.s terms of (18a) by application of Lemma 1.1. Finally, we can move from (18b) to (18c) by application of the law of total variance. □

Note that for discrete or categorical Y*, it is straightforward to evaluate Equation (18a) directly.*

A.3 ρ-BALD

Theorem 3. *Under the following assumptions*

1. $\{\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) : \mathrm{t} \in \{0,1\}\}$ *are instances of the independent-normally distributed or joint-normally distributed random variables* $\{\widehat{\mu}^{\mathrm{t}}_{\boldsymbol{\Omega}} = \mathbb{E}[\mathrm{Y} \mid \boldsymbol{\Omega}, \mathrm{T} = \mathrm{t}, \mathbf{x}] : \mathrm{t} \in \{0,1\}\}$,
2. $\mathrm{Var}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\mathrm{train}})}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}')) > 0$.

Let $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x})$ *be a realization of the random variable* $\widehat{\tau}_{\boldsymbol{\Omega}} = \widehat{\mu}^1_{\boldsymbol{\Omega}} - \widehat{\mu}^0_{\boldsymbol{\Omega}}$. *The information gain for* $\widehat{\tau}_{\boldsymbol{\Omega}}$ *if we observe the label for the potential outcome* Y^{t} *given measured covariates* $\mathbf{x}$*, treatment* t *and a dataset of observations* $\mathcal{D}_{\mathrm{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$ *is approximately*

$$\begin{aligned} \mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\mathrm{train}}) &\approx \frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}))}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}, \\ &= \frac{\mathrm{Var}_{\boldsymbol{\omega}}\left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})\right) - 2\mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}{\mathrm{Var}_{\boldsymbol{\omega}}\left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}')\right)} + 1, \end{aligned} \tag{19}$$

where for binary $\mathrm{T} = t$, $\mathrm{t}' = (1 - \mathrm{t})$.

Proof.

$$\begin{aligned} \mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}) &= \mathrm{H}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}) - \mathrm{H}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{Y}^{\mathrm{t}}, \mathbf{x}, \mathrm{t}, \mathcal{D}) && \text{(20a)} \\ &= \mathrm{H}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}) - \mathop{\mathbb{E}}_{\mathrm{y}^{\mathrm{t}} \sim p(\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})} \mathrm{H}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}}, \mathbf{x}, \mathrm{t}) && \text{(20b)} \\ &= \frac{1}{2}\log(2\pi\,\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}})) - \mathop{\mathbb{E}}_{\mathrm{y}^{\mathrm{t}} \sim p(\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})} \left[\frac{1}{2}\log(2\pi\,\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}}))\right] && \text{(20c)} \\ &\geq \frac{1}{2}\log(2\pi\,\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}})) - \frac{1}{2}\log(2\pi\,\mathbb{E}\left[\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}})\right]) && \text{(20d)} \\ &= \frac{1}{2}\log\left(\frac{\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}})}{\mathbb{E}\left[\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}})\right]}\right), && \text{(20e)} \end{aligned}$$

and we can further expand the fraction to

$$\begin{aligned} \frac{\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})}{\mathbb{E}\left[\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}})\right]} &= \frac{\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})}{\mathrm{Var}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D})}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))} && \text{(20f)} \\ &= \frac{\mathrm{Var}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D})}(\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}) \mid \mathrm{t})}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))} && \text{(20g)} \\ &= \frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0) \mid \mathrm{t})}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))} && \text{(20h)} \\ &= \frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))} && \text{(20i)} \\ &= \frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})) + \mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}')) - 2\,\mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))} && \text{(20j)} \\ &= \frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})) - 2\,\mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))} + 1, && \text{(20k)} \end{aligned}$$

where (20a) by definition of mutual information; (20a)-(20b) from the result of Houlsby et al. [18]; (20b)-(20c) by Assumption 1. and Corollary 1.1; (20c)-(20d) by Jensen's inequality; (20d)-(20e) by the logarithmic quotient identity; (20f) by Lemma 3.1; (20f)-(20g) by definition of the variance. (20g)-(20h) by definition of $\widehat{\tau}_{\boldsymbol{\omega}}$; (20h)-(20i) by symmetry of the variance of the difference of two random variables; (20i)-(20j) by the definition of the variance of the difference of two random variables; and (20j)-(20k) by cancelling terms. □

Lemma 3.1. *Under the following assumptions*

1. *Consistency* $\mathrm{Y} \mid \mathrm{T} = \mathrm{Y}^{\mathrm{t}}$*;*
2. *Unconfoundedness* $(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$*;*

$$\mathop{\mathbb{E}}_{\mathrm{y}^{\mathrm{t}} \sim p(\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})} \left[\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}}) \right] \approx \mathop{\mathbb{E}}_{\mathrm{y}^{\mathrm{t}} \sim p(\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})} \left[\mathop{\mathrm{Var}}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} (\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}')) \right], \tag{21}$$

where for binary $\mathrm{T} = \mathrm{t}$, $\mathrm{t}' = (1 - \mathrm{t})$.

Proof.

$$\mathbb{E}_{\mathrm{y}^{\mathrm{t}} \sim p(\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})} \left[\mathrm{Var}(\widehat{\tau}_{\boldsymbol{\Omega}} \mid \mathrm{y}^{\mathrm{t}}) \right] = \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\left(\widehat{\tau}_{\boldsymbol{\omega}} - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\widehat{\tau}_{\boldsymbol{\omega}} \mid \mathrm{y}^{\mathrm{t}} \right] \right)^2 \mid \mathrm{y}^{\mathrm{t}} \right] \right], \tag{22a}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\left(\mathbb{E}[\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] \right)^2 \mid \mathrm{y}^{\mathrm{t}} \right] \right], \tag{22b}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\left(\mathbb{E}[\mathrm{Y}^1 \mid \mathbf{x}, \boldsymbol{\omega}] - \mathbb{E}[\mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^1 \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] + \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] \right)^2 \mid \mathrm{y}^{\mathrm{t}} \right] \right], \tag{22c}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\left(\left(\mathbb{E}[\mathrm{Y}^1 \mid \mathbf{x}, \boldsymbol{\omega}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^1 \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] \right) - \left(\mathbb{E}[\mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^0 \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] \right) \right)^2 \mid \mathrm{y}^{\mathrm{t}} \right] \right], \tag{22d}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\left(\left(\mathbb{E}[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \boldsymbol{\omega}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] \right) - \left(\mathbb{E}[\mathrm{Y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathbb{E}[\mathrm{Y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega}] \mid \mathrm{y}^{\mathrm{t}} \right] \right) \right)^2 \mid \mathrm{y}^{\mathrm{t}} \right] \right], \tag{22e}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\left(\left(\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}}] \right] \right) - \left(\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}'}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}'}] \right] \right) \right)^2 \right] \right], \tag{22f}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\left(\underbrace{\left(\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}}] \right] \right)}_{\approx 0} - \left(\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}'}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}'}] \right] \right) \right)^2 \right] \right], \tag{22g}$$

$$\approx \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\left(\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}'}] - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}'} \mid \mathbf{x}, \boldsymbol{\omega})} [\mathrm{y}^{\mathrm{t}'}] \right] \right)^2 \right] \right], \tag{22h}$$

$$= \mathop{\mathbb{E}}_{p(\mathrm{y}^{\mathrm{t}})} \left[\mathop{\mathbb{E}}_{p(\boldsymbol{\omega} \mid \mathrm{y}^{\mathrm{t}})} \left[\left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}') - \mathop{\mathbb{E}}_{p(\boldsymbol{\omega})} \left[\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}') \right] \right)^2 \right] \right], \tag{22i}$$

$$= \mathop{\mathbb{E}}_{\mathrm{y}^{\mathrm{t}} \sim p(\mathrm{Y}^{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D})} \left[\mathop{\mathrm{Var}}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} (\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}')) \right], \tag{22j}$$

where (22a) by definition of variance; (22a)-(22b) by definition of $\widehat{\tau}_{\boldsymbol{\omega}}$; (22b)-(22c) by linearity of expectations; (22c)-(22d) by grouping terms; (22d)-(22e) by symmetry of the square; (22e)-(22f) by rewriting expectations in terms of densities; (22f)-(22g) the observed potential outcome does not have an effect on the expectation of the model for the counterfactual outcome; (22g)-(22h) we drop the term as an approximation as we cannot estimate here how much the expected outcome is going to change—the conservative assumption is that will not change; (22h)-(22i) by definition of $\widehat{\mu}_{\boldsymbol{\omega}}$; (22i)-(22j) by definition of variance; □

B Baselines

B.1 S-type error Information Gain

In their work, Sundin et al. [45] assume that the underlying model is a Gaussian Process (GP) and also that they have access to the counterfactual outcome. Although GPs are suitable for uncertainty estimation, they do not scale up to high dimensional datasets (e.g. images). We propose to use Deep Ensembles and DUE for alleviating the capabilities issues and we modified the objective to be more suitable for our architecture.

Following the formulation from Houlsby et al. [18], the acquisition strategy becomes $\arg\max_x \mathbb{H}[\gamma|x, D] - \mathbb{E}_{\mathbb{H}[p(\theta|D)}[\gamma|x, \theta]]$, where $\gamma(x) = \text{probit}^{-1}(-\frac{|\mathbf{E}_{p(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}[\tau]|}{\sqrt{\text{Var}(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}})$, $\text{probit}^{-1}(\cdot)$ is the cumulative distribution function of normal distribution and $p(\gamma|x, D) = \text{Bernoulli}(\gamma)$. With DUE (Deep Kernel Learning method) Deep Ensembles (samples from $p(\theta|D)$ we can compute those terms similarly to how we implemented our BALD objectives.

Below is an example of how this was implemented in PyTorch:

```
tau_mu = mu1s - mu0s
tau_var = var1s + var0s + 1e-07
gammas = torch.distributions.normal.Normal(0, 1).cdf(
    -tau_mu.abs() / tau_var.sqrt()
)
gamma = gammas.mean(-1)
predictive_entropy = dist.Bernoulli(gamma).entropy()
conditional_entropy = dist.Bernoulli(gammas).entropy().mean(-1)
# it can get negative very small number
# because of numerical instabilities
scores = (predictive_entropy - conditional_entropy).clamp_min(1e-07)
```

C Datasets

C.1 Synthetic Data

We modify the synthetic dataset presented by Kallus et al. [22]. Our dataset is described by the following structural causal model (SCM):

$$\mathbf{x} \coloneqq N_{\mathbf{x}}, \tag{23a}$$
$$\mathrm{t} \coloneqq N_{\mathrm{t}}, \tag{23b}$$
$$\mathrm{y} \coloneqq (2\mathrm{t}-1)\mathbf{x} + (2\mathrm{t}-1) - 2\sin(2(2\mathrm{t}-1)\mathbf{x}) + 2(1+0.5\mathbf{x}) + N_{\mathrm{y}}, \tag{23c}$$

where $N_{\mathbf{x}} \sim \mathcal{N}(0,1)$, $N_{\mathrm{t}} \sim \text{Bern}(\text{sigmoid}(2\mathbf{x}+0.5))$, and $N_{\mathrm{y}} \sim \mathcal{N}(0,1)$.

Each random realization of the simulated dataset generates 10000 pool set examples, 1000 validation examples, and 1000 test examples. In the experiments we report results over 40 random realizations. The seeds for the random number generators are i, $i+1$, and $i+2$; $\{i \in [0, 1, \ldots, 19]\}$, for the training, validation, and test sets, respectively.

C.2 IHDP Data.

Infant Health and Development Program (IHDP) is a semi-synthetic dataset [17, 42] commonly used in literature to study the performance of causal effect estimation methods. The dataset consists of 747 cases, out of which 139 are assigned in treatment group and 608 in control. Each unit is represented by 25 covariates describing different aspects of the infants and their mothers. We report results over 200 random realizations of response surface B described by Hill [17].

C.3 CMNIST Data.

Following the setup from [21], we use a simulated dataset based on MNIST [29]. CMNIST is described by the following SCM:

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

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

$$\text{t} := N_{\text{t}}, \tag{24c}$$

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

where 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, $\mu_{\text{c}} = \mathbb{E}[\mu_{N_{\mathbf{x}}} \mid \text{c}]$ and $\sigma_{\text{c}}^2 = \text{Var}[\mu_{N_{\mathbf{x}}} \mid \text{c}]$. To map the high dimensional images $\mathbf{x}$ onto a one-dimensional manifold ϕ with domain $[-3, 3]$ 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 $[\text{Min}_{\text{c}}, \text{Max}_{\text{c}}]$. Finally, $\text{Min}_{\text{c}} = -2 + \frac{4}{10}\text{c}$, and $\text{Max}_{\text{c}} = -2 + \frac{4}{10}(\text{c} + 1)$.

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 $\{i \in [0, 1, \ldots, 19]\}$ for the 10 random realizations generated.

D More Results

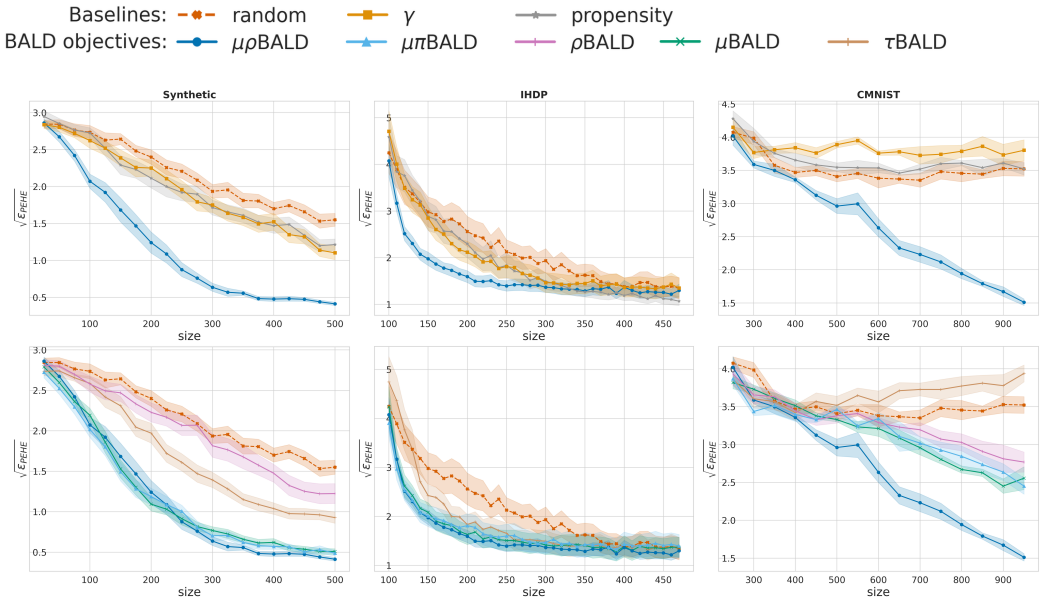

Figure 6: $\sqrt{\epsilon_{PEHE}}$ performance (shaded standard error) for Deep Ensembles based models. **(left to right) synthetic** (20 seeds), **IHDP** (50 seeds) and *CMNIST* (5 seeds) dataset results, **(top to bottom)** comparison with baselines, comparison between BALD objectives. We observe that BALD objectives outperform the **random**, γ and **propensity** acquisition functions significantly, suggesting that epistemic uncertainty aware methods that target reducible uncertainty can be more sample efficient.

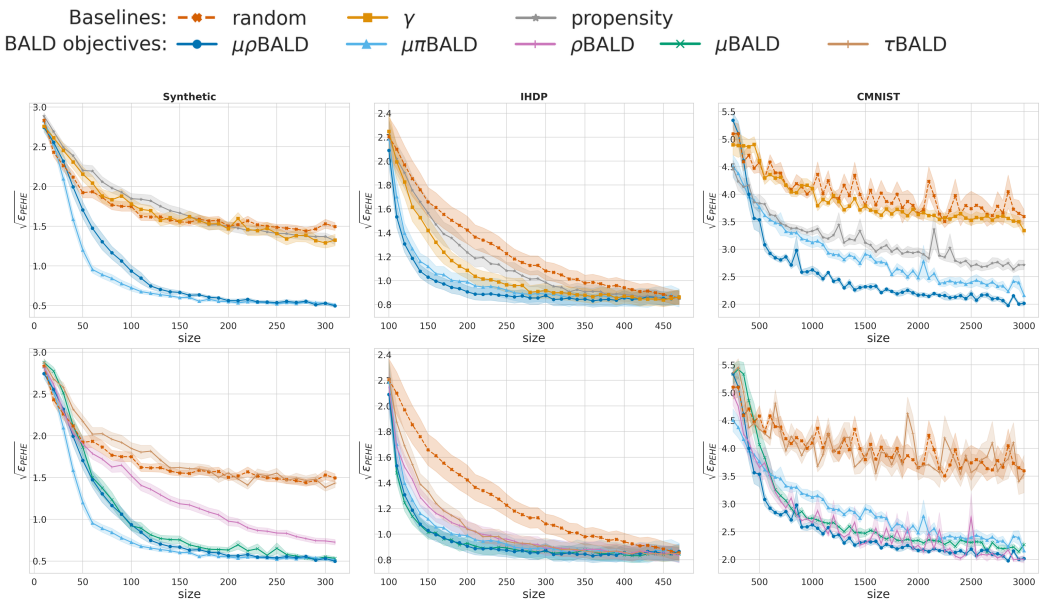

Figure 7: $\sqrt{\epsilon_{PEHE}}$ performance (shaded standard error) for DUE models. **(left to right) synthetic** (40 seeds), and **IHDP** (200 seeds). We observe that BALD objectives outperform the **random**, γ and **propensity** acquisition functions significantly, suggesting that epistemic uncertainty aware methods that target reducible uncertainty can be more sample efficient.

E Compute

We used a cluster of 8 nodes with 4 GPUs each (16 RTX 2080 and 16 Titan RTX). The total GPU hours is estimated to be:

8 baselines x (.5 + 1 + 1) days per dataset x (5 ensemble components * 0.25 GPU usage + 1 DUE * 0.3 GPU usage) x 24 hours = 744 GPU hours

Code is written in python. Packages used include PyTorch [35], scikit-learn [36], Ray [32], NumPy, SciPy, and Matplotlib.

F Model Architectures

For deep ensembles, we use an ensemble of TarNETs [42]. For Due, we append the treatment variable to the features extracted, then define the GP over that input. **Synthetic Architecture**

```
==========================================================
Layer (type:depth-idx)                             Output Shape
==========================================================
Sequential                                         --
├─NeuralNetwork: 1-1                               [64, 100]
│    └─Sequential: 2-1                             [64, 100]
│    │    └─Linear: 3-1                            [64, 100]
│    │    └─ResidualDense: 3-2                     [64, 100]
│    │    │    └─PreactivationDense: 4-1           [64, 100]
│    │    │    │    └─Sequential: 5-1              [64, 100]
│    │    │    │    │    └─Activation: 6-1         [64, 100]
│    │    │    │    │    └─Linear: 6-2             [64, 100]
│    │    │    └─Identity: 4-2                     [64, 100]
│    │    └─ResidualDense: 3-3                     [64, 100]
│    │    │    └─PreactivationDense: 4-3           [64, 100]
│    │    │    │    └─Sequential: 5-2              [64, 100]
│    │    │    │    │    └─Activation: 6-3         [64, 100]
│    │    │    │    │    └─Linear: 6-4             [64, 100]
│    │    │    └─Identity: 4-4                     [64, 100]
```

```
│    │    └─ResidualDense: 3-4                [64, 100]
│    │    │    └─PreactivationDense: 4-5       [64, 100]
│    │    │    │    └─Sequential: 5-3          [64, 100]
│    │    │    │    │    └─Activation: 6-5     [64, 100]
│    │    │    │    │    └─Linear: 6-6         [64, 100]
│    │    │    └─Identity: 4-6                 [64, 100]
│    │    └─Activation: 3-5                    [64, 100]
│    │    │    └─Sequential: 4-7               [64, 100]
│    │    │    │    └─Identity: 5-4            [64, 100]
│    │    │    │    └─LeakyReLU: 5-5           [64, 100]
│    │    │    │    └─Dropout: 5-6             [64, 100]
├─GMM: 1-2                                     [64, 5]
│    └─Linear: 2-2                             [64, 5]
│    └─Linear: 2-3                             [64, 5]
│    └─Sequential: 2-4                         [64, 5]
│    │    └─Linear: 3-6                        [64, 5]
│    │    └─Softplus: 3-7                      [64, 5]
==========================================================
Total params: 32,115
```

IHDP Architecture

```
==========================================================================
Layer (type:depth-idx)                                   Output Shape
==========================================================================
Sequential                                               --
├─TARNet: 1-1                                           [64, 400]
│    └─NeuralNetwork: 2-1                               [64, 400]
│    │    └─Sequential: 3-1                             [64, 400]
│    │    │    └─Linear: 4-1                            [64, 400]
│    │    │    └─ResidualDense: 4-2                     [64, 400]
│    │    │    │    └─PreactivationDense: 5-1           [64, 400]
│    │    │    │    │    └─Sequential: 6-1              [64, 400]
│    │    │    │    └─Identity: 5-2                     [64, 400]
│    │    │    └─ResidualDense: 4-3                     [64, 400]
│    │    │    │    └─PreactivationDense: 5-3           [64, 400]
│    │    │    │    │    └─Sequential: 6-2              [64, 400]
│    │    │    │    └─Identity: 5-4                     [64, 400]
│    └─Sequential: 2-2                                  [64, 400]
│    │    └─ResidualDense: 3-2                          [64, 400]
│    │    │    └─PreactivationDense: 4-4                [64, 400]
│    │    │    │    └─Sequential: 5-5                   [64, 400]
│    │    │    │    │    └─Activation: 6-3              [64, 401]
│    │    │    │    │    └─Linear: 6-4                  [64, 400]
│    │    │    └─Sequential: 4-5                        [64, 400]
│    │    │    │    └─Dropout: 5-6                      [64, 401]
│    │    │    │    └─Linear: 5-7                       [64, 400]
│    │    └─ResidualDense: 3-3                          [64, 400]
│    │    │    └─PreactivationDense: 4-6                [64, 400]
│    │    │    │    └─Sequential: 5-8                   [64, 400]
│    │    │    │    │    └─Activation: 6-5              [64, 400]
│    │    │    │    │    └─Linear: 6-6                  [64, 400]
│    │    │    └─Identity: 4-7                          [64, 400]
│    │    └─Activation: 3-4                             [64, 400]
│    │    │    └─Sequential: 4-8                        [64, 400]
│    │    │    │    └─Identity: 5-9                     [64, 400]
│    │    │    │    └─ELU: 5-10                         [64, 400]
│    │    │    │    └─Dropout: 5-11                     [64, 400]
├─GMM: 1-2                                              [64, 5]
│    └─Linear: 2-3                                      [64, 5]
│    └─Linear: 2-4                                      [64, 5]
│    └─Sequential: 2-5                                  [64, 5]
│    │    └─Linear: 3-5                                 [64, 5]
│    │    └─Softplus: 3-6                               [64, 5]
==========================================================================
```

CMNIST Architecture

```
==========================================================================
Layer (type:depth-idx)                                   Output Shape
==========================================================================
Sequential                                               --
├─TARNet: 1-1                                            [200, 100]
│    └─ResNet: 2-1                                       [200, 48]
│    │    └─Sequential: 3-1                              [200, 48, 1, 1]
│    │    │    └─Conv2d: 4-1                             [200, 12, 28, 28]
│    │    │    └─Identity: 4-2                           [200, 12, 28, 28]
│    │    │    └─ResidualConv: 4-3                       [200, 12, 28, 28]
│    │    │    │    └─Sequential: 5-1                    [200, 12, 28, 28]
│    │    │    │    │    └─PreactivationConv: 6-1        [200, 12, 28, 28]
│    │    │    │    │    └─PreactivationConv: 6-2        [200, 12, 28, 28]
│    │    │    │    └─Sequential: 5-2                    [200, 12, 28, 28]
│    │    │    │    │    └─Dropout2d: 6-3                [200, 12, 28, 28]
│    │    │    │    │    └─Conv2d: 6-4                   [200, 12, 28, 28]
│    │    │    └─ResidualConv: 4-4                       [200, 24, 14, 14]
│    │    │    │    └─Sequential: 5-3                    [200, 24, 14, 14]
│    │    │    │    │    └─PreactivationConv: 6-5        [200, 12, 28, 28]
│    │    │    │    │    └─PreactivationConv: 6-6        [200, 24, 14, 14]
│    │    │    │    └─Sequential: 5-4                    [200, 24, 14, 14]
│    │    │    │    │    └─Dropout2d: 6-7                [200, 12, 28, 28]
│    │    │    │    │    └─Conv2d: 6-8                   [200, 24, 14, 14]
│    │    │    └─ResidualConv: 4-5                       [200, 24, 14, 14]
│    │    │    │    └─Sequential: 5-5                    [200, 24, 14, 14]
│    │    │    │    │    └─PreactivationConv: 6-9        [200, 24, 14, 14]
│    │    │    │    │    └─PreactivationConv: 6-10       [200, 24, 14, 14]
│    │    │    │    └─Sequential: 5-6                    [200, 24, 14, 14]
│    │    │    │    │    └─Dropout2d: 6-11               [200, 24, 14, 14]
│    │    │    │    │    └─Conv2d: 6-12                  [200, 24, 14, 14]
│    │    │    └─ResidualConv: 4-6                       [200, 48, 7, 7]
│    │    │    │    └─Sequential: 5-7                    [200, 48, 7, 7]
│    │    │    │    │    └─PreactivationConv: 6-13       [200, 24, 14, 14]
│    │    │    │    │    └─PreactivationConv: 6-14       [200, 48, 7, 7]
│    │    │    │    └─Sequential: 5-8                    [200, 48, 7, 7]
│    │    │    │    │    └─Dropout2d: 6-15               [200, 24, 14, 14]
│    │    │    │    │    └─Conv2d: 6-16                  [200, 48, 7, 7]
│    │    │    └─ResidualConv: 4-7                       [200, 48, 7, 7]
│    │    │    │    └─Sequential: 5-9                    [200, 48, 7, 7]
│    │    │    │    │    └─PreactivationConv: 6-17       [200, 48, 7, 7]
│    │    │    │    │    └─PreactivationConv: 6-18       [200, 48, 7, 7]
│    │    │    │    └─Sequential: 5-10                   [200, 48, 7, 7]
│    │    │    │    │    └─Dropout2d: 6-19               [200, 48, 7, 7]
│    │    │    │    │    └─Conv2d: 6-20                  [200, 48, 7, 7]
│    │    │    └─ResidualConv: 4-8                       [200, 48, 7, 7]
│    │    │    │    └─Sequential: 5-11                   [200, 48, 7, 7]
│    │    │    │    │    └─PreactivationConv: 6-21       [200, 48, 7, 7]
│    │    │    │    │    └─PreactivationConv: 6-22       [200, 48, 7, 7]
│    │    │    │    └─Sequential: 5-12                   [200, 48, 7, 7]
│    │    │    │    │    └─Dropout2d: 6-23               [200, 48, 7, 7]
│    │    │    │    │    └─Conv2d: 6-24                  [200, 48, 7, 7]
│    │    │    └─AdaptiveAvgPool2d: 4-9                  [200, 48, 1, 1]
│    └─Sequential: 2-2                                   [200, 100]
│    │    └─ResidualDense: 3-2                           [200, 100]
│    │    │    └─PreactivationDense: 4-10                [200, 100]
│    │    │    │    └─Sequential: 5-13                   [200, 100]
│    │    │    │    │    └─Activation: 6-25              [200, 49]
│    │    │    │    │    └─Linear: 6-26                  [200, 100]
│    │    │    └─Sequential: 4-11                        [200, 100]
│    │    │    │    └─Dropout: 5-14                      [200, 49]
│    │    │    │    └─Linear: 5-15                       [200, 100]
│    │    └─ResidualDense: 3-3                           [200, 100]
│    │    │    └─PreactivationDense: 4-12                [200, 100]
```

```
│    │    │    │    └─Sequential: 5-16                  [200, 100]
│    │    │    │    │    └─Activation: 6-27             [200, 100]
│    │    │    │    │    └─Linear: 6-28                 [200, 100]
│    │    │    └─Identity: 4-13                         [200, 100]
│    │    └─Activation: 3-4                             [200, 100]
│    │    │    └─Sequential: 4-14                       [200, 100]
│    │    │    │    └─Identity: 5-17                    [200, 100]
│    │    │    │    └─LeakyReLU: 5-18                   [200, 100]
│    │    │    │    └─Dropout: 5-19                     [200, 100]
├─GMM: 1-2                                              [200, 5]
│    └─Linear: 2-3                                      [200, 5]
│    └─Linear: 2-4                                      [200, 5]
│    └─Sequential: 2-5                                  [200, 5]
│    │    └─Linear: 3-5                                 [200, 5]
│    │    └─Softplus: 3-6                               [200, 5]
==========================================================================
```

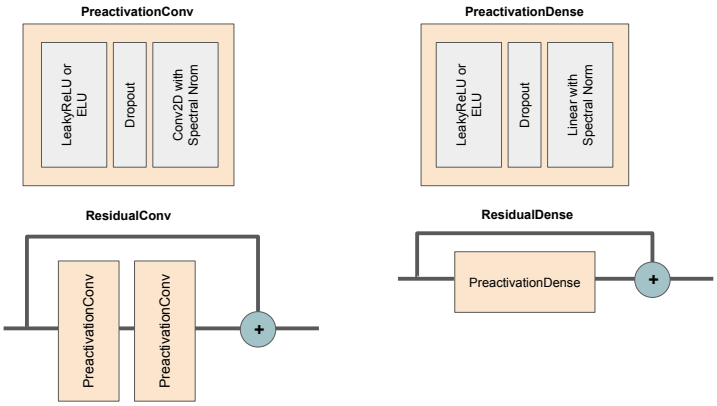

Figure 8: `PreactivationConv` is a convolution layer with LeakyReLU (or ELU when slope is negative) activation, dropout and spectral norm applied [15, 31]. Similarly, `PreactivationDense` is a dense layer with BatchNorm [19], LeakyReLU (or ELU when slope is negative) activation and spectral norm applied [15, 31]. `ResidualConv` is the residual convolution layer, defined as `PreactivationConv(PreactivationConv(x)) + SpectralNorm(1x1Conv(x))` and `ResidualDense` are residual dense layers, defined as `PreactivationDense(x)+x)`.

All experiments were trained using Adam optimizer [24].

F.1 Hyper-parameters

We use ray tune [30] with the hyperopt [6] search algorithm to optimize our network hyper-parameters. The hyper-parameter search spaces are given in Table 2 and Table 3. 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. The final hyper-parameters are given in Table 4 and Table 5.

Table 2: Hyper-parameter search space for **Deep Ensemble**

Hyper-parameter	Search Space
hidden units	[100, 200, 400]
network depth	[2, 3, 4]
negative slope	[ReLU [2], 0.1, 0.2, ELU [7]]
dropout rate	[0.05, 0.1, 0.2, 0.5]
spectral norm	[None, 0.95, 1.5, 3.0]
batch size	[32, 64, 100, 200]
learning rate	[2e-4, 5e-4, 1e-3]

Table 3: Hyper-parameter search space for **DUE**

Hyper-parameter	Search Space
kernel	[RBF, Matern]
ν (Matern)	[0.5, 1.5, 2.5]
inducing points	[20, 50, 100, 200]
hidden units	[100, 200, 400]
network depth	[2, 3, 4]
negative slope	[ReLU [2], 0.1, 0.2, ELU [7]]
dropout rate	[0.05, 0.1, 0.2, 0.5]
spectral norm	[None, 0.95, 1.5, 3.0]
batch size	[32, 64, 100, 200]
learning rate	[2e-4, 5e-4, 1e-3]

Table 4: Training hyper parameters for **Deep Ensemble** experiments

Parameter	**Synthetic**	**IHDP**	**CMNIST**
dim hidden	100	400	100
dropout	0.0	0.15	0.1
depth	4	3	3
spectral norm	12	0.95	24
learning rate	0.001	0.001	0.001
non-linearity	ReLU	ELU	ReLU

Table 5: Training hyper parameters for **DUE** experiments

Parameter	**Synthetic**	**IHDP**	**CMNIST**
kernel	RBF	Matern ($\nu = 1.5$)	RBF
inducing points	20	100	100
dim hidden	100	200	200
dropout	0.2	0.1	0.05
depth	3	3	2
batch size	200	100	64
spectral norm	0.95	0.95	3.0
learning rate	0.001	0.001	0.001
non-linearity	ReLU	ELU	ELU