Appendix for "Learning Overlapping Representations for the Estimation of Individualized Treatment Effects"

This Appendix provides full proofs for all theorems stated in the main body of this paper, it includes further experimental and implementation details, a performance analysis for the average treatment effect (as opposed to individualized treatment effect) and additional figures to illustrate our approach.

1 Theory

In this section, we provide the proof of Theorem 1 and 2 in the main text. For reader's convenience, we start by restating the notation introduced in Section 2 of the main text, Theorem 1 and 2 followed by a proof sketch.

Notation. Let $\mathcal{P}_{x,t}$ denote the input data distribution $p(x,t)$, $\mathcal{P}_t = p(y_t|x)p(x,t)$ the joint factual distribution of x and y_t, $\mathcal{P}_{1-t} = p(y_t|x)p(x,1-t)$ the joint counterfactual distribution of x and y_t. Each instance in the observed (factual) dataset $\mathcal{D}_t = \{(x_{i,t}, y_{i,t})\}_{i=1}^{N_t}$, is assumed to be sampled $i.i.d$ from $\mathcal{P}_t$. $\mathcal{D}_{t-1}$ will be used to denote the unobserved (counterfactual) data set that results from fliping the treatment assignment for each instance. While we assume ϕ to be deterministic, in the following section we let w_t be a vector of weights with prior distribution $\pi_t = \mathcal{N}(\mathbf{0}, \lambda_t^{-1}\mathbf{I})$, $\lambda_t > 0$. In this sense, π_t defines the hypothesis space $\mathcal{F}$ of f_t and we write $\hat{\rho}_t$ for the posterior distribution of f_t, itself a random variable. We write $\mu(x_i|\mathcal{D}_t, \Theta_t)$ and $\sigma^2(x_i|\mathcal{D}_t, \Theta_t)$ for its posterior mean and variance given context x_i. Θ_t includes all hyperparameters (both shared parameters (in ϕ) and specific parameters to each treatment group).

Theorem 1. *With the assumption that the squared loss function $L : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is sub-gaussian under π_t and $\mathcal{P}_t$, and using the notation introduced above, the following holds. With probability at least $1-\delta$ and for any posterior distribution $\hat{\rho}_t$ on $\mathcal{F}$, ϵ_{PEHE} is upper-bounded by*

$$\sum_{t=0}^{1}\left(2\tilde{D}_\infty L_{\hat{\rho}_t}(\mathbf{X}_t,\mathbf{Y}_t)+\big(\tilde{D}_\infty+1\big)\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_t)+\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_{1-t})+\frac{C_{t,1}(N_t,N_{1-t})}{N_tN_{1-t}}\,\mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{C_{t,2}(N_t,N_{1-t})}{N_tN_{1-t}}\frac{1}{\delta}+C_{t,3}\right)$$

where $C_{t,1}(N_t, N_{1-t})$, $C_{t,2}(N_t, N_{1-t})$ are linear function in their arguments, $C_{t,3}$ is constant, $\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X})$ is the posterior variance on $\mathbf{X}$, $\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\sigma^2(x_i|\mathcal{D}_t,\Theta_t)$, $L_{\hat{\rho}_t}(\mathbf{X},\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N}\big(\mu(x_i|\mathcal{D}_t,\Theta_t)-y_i\big)^2$ is the posterior prediction loss, $\tilde{D}_\infty = D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})+1$, $D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t}) = \sup_x \frac{p(x,1-t)}{p(x,t)}$, and finally $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence.

Proof. The proof consists of four steps: (a) Use the decomposition method in [14] to derive a upper bound of ϵ_{PEHE} which is expressed in terms of the expected factual risk and the expected counterfactual risk (Section 1.1); (b) Bound the expected factual risk using the PAC-Bayes bound of supervised learning (Section 1.2); (c) Derive the PAC-Bayes bound of the expected counterfactual risk (Section 1.3). (d) Substitute the bounds of the expected factual and counterfactual risk into the decomposition from (a) (Section 3.5), we have the PAC-Bayes bound in Equation 1. □

Theorem 2. *Assume the notation introduced above, for any posterior distribution $\hat{\rho}_t$ on $\mathcal{F}$, the expected counterfactual Gibbs risk*[1] *$R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$ is bounded above by,*

$$\frac{1}{2}D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t}) + D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})L_{\mathcal{P}_t}(G_{\hat{\rho}_t}). \tag{1}$$

where $L_{\mathcal{P}_t}(\hat{\rho}_t) = \mathbb{E}_{(x,y)\sim\mathcal{P}_t}\Big[\big(\mu(x|\mathcal{D}_t,\Theta_t)-y\big)^2 + \frac{1}{2}\sigma^2(x|\mathcal{D}_t,\Theta_t)\Big]$ is the expected factual loss, $D_{\mathcal{P}_{x,1-t}}(\hat{\rho}_t) = \mathbb{E}_{x\sim\mathcal{P}_{x,1-t}}\big[\sigma^2(x|\mathcal{D}_t,\Theta_t)\big]$ is the expected counterfactual variance, and finally $D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t}) = \sup_x \frac{p(x,1-t)}{p(x,t)}$.

Proof. The Theorem is a summary of the results in Lemma 5 and 6 in Section 1.3. Using the notation above, the expected disagreement loss and the expected ensemble loss in Lemma 5 are given as the expected counterfactual variance

[1]The expected counterfactual Gibbs risk is an upper bound of the expected Bayesian counterfactual risk, a detailed discussion of Gibbs and Bayesian risk is given in Appendix 1.2.

and the expected factual loss respectively in Theorem 2. Replacing the expected disagreement loss by the expected counterfactual variance, and the expected ensemble loss by the expected factual loss in Lemma 6, we obtain the result in Equation (1) . □

1.1 PEHE decompostion

The expected Precision in Estimation of Heterogeneous Effects, ϵ_{PEHE}, can be bounded using the expected factual risk and the expected counterfactual risk [14].

Proposition 3. *Given data distributions $\mathcal{P}_t$, Bayesian model $B_{\hat{\rho}_t}$, $t \in \{0, 1\}$, we have*

$$\epsilon_{\text{PEHE}} \leq 2 \sum_{t=0}^{1} (R_{\mathcal{P}_t}(B_{\hat{\rho}_t}) + R_{\mathcal{P}_{1-t}}(B_{\hat{\rho}_t}) - 2\beta_t^{-1}) \tag{2}$$

where $R_{\mathcal{P}_\tau}(B_{\hat{\rho}_\tau}) = \int \left(\hat{f}_t(x) - y_t\right)^2 p(y_t, x, \tau) dx dy_t$ is the expected risk under the distribution $p(y_t, x, \tau)$.

The proof of Proposition 3 is provided in Section 3.1. The goal of bounding ϵ_{PEHE} can be achieved by bounding the expected factual risk $R_{\mathcal{P}_t}(B_{\hat{\rho}_t})$ and the expected counterfactual risk $R_{\mathcal{P}_{1-t}}(B_{\hat{\rho}_t})$. To differentiate the two risks, the data distribution $\mathcal{P}_{1-t}$ and the model $B_{\hat{\rho}_t}$ are indexed differently in the expected counterfactual risk, indicating the counterfactual outcomes are missing.

1.2 PAC-Bayesian Bound for the expected factual risk

We first introduce Gibbs model and Bayesian model in PAC-Bayesian theory [12, 15]. Given a distribution $\hat{\rho}$ over the functions in a hypothesis space $\mathcal{F}$, **Gibbs model** $G_{\hat{\rho}}$ is defined as making prediction for every example by randomly sampling a function in $\mathcal{F}$. In other words, to predict the label of an example x, the Gibbs model first draws a function f from $\mathcal{F}$ according to $\hat{\rho}$, then return $f(x)$ as label. **Bayesian model** $B_{\hat{\rho}}$ is defined as making prediction on x by averaging all the functions in $\mathcal{F}$ with respect to $\hat{\rho}$. The expected risk of $G_{\hat{\rho}}$ is

$$R_{\mathcal{P}}(G_{\hat{\rho}}) = \mathbb{E}_{f\sim\hat{\rho}} \mathbb{E}_{(x,y)\sim\mathcal{P}} \big[l(f(x, y) \big] \tag{3}$$

and the expected risk of $B_{\hat{\rho}}$ is

$$R_{\mathcal{P}}(B_{\hat{\rho}}) = \mathbb{E}_{(x,y)\sim\mathcal{P}} \big[l(\mathbb{E}_{f\sim\hat{\rho}} \big[f(x) \big], y) \big] \tag{4}$$

where l is a loss function. In DKLITE, the loss function l is the squared loss $l(\hat{y}, y) = (\hat{y} - y)^2$. There are two well known connections between the two risks in Equation (3) and (4). First, $R_{\mathcal{P}}(B_{\hat{\rho}}) < 2R_{\mathcal{P}}(G_{\hat{\rho}})$ when l is a 0-1 loss. This is because if $B_{\hat{\rho}}$ classifies an example incorrectly, at least half of the classifiers in $\mathcal{F}$ under $\hat{\rho}$ will classify this example incorrectly. When loss function is convex e.g. the squared loss, Jensen's inequality gives that $R_{\mathcal{P}}(B_{\hat{\rho}}) \leq R_{\mathcal{P}}(G_{\hat{\rho}})$. Because of these connections, the bound of Gibbs risk is also applicable to the Bayesian risk when a suitable loss function is in use. The expected factual risk can be bounded using some existing PAC-Bayes theorems in supervised learning [2, 6]. Following the convention of PAC-Bayes literature, the bound is stated mainly on the Gibbs risk but implicitly applied to Bayesian risk.

Theorem 4. *With the assumption that the squared loss function $l : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is sub-gaussian with variance factor ξ_t^2 under π_t and $\mathcal{P}_t$, and using the notation introduced in Section 1. Then, for $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the random choice $S_t = (\mathbf{X}_t, \mathbf{Y}_t) \in \mathcal{P}^{N_t}$, for any posterior distribution $\hat{\rho}_t$ on $\mathcal{F}$, we have*

$$R_{\mathcal{P}_t}(G_{\hat{\rho}_t}) \leq R_{S_t}(G_{\hat{\rho}_t}) + \frac{1}{N_t} \big[\text{KL}(\hat{\rho}_t \| \pi_t) + \frac{1}{\delta} \big] + \frac{1}{2} \xi_t^2 \tag{5}$$

where $R_{\mathcal{P}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{P}} \big[l(f(x), y) \big]$ is the expected risk, $R_S(G_{\hat{\rho}}) = \frac{1}{N_t} \sum_{i=1}^{N_t} l(f(x_{i,t}), y)$ is the empirical risk on S_t.

Proof. Section 3.2. □

1.3 PAC-Bayesian Bound for the expected counterfactual risk

In this section, we show how to derive the PAC-Bayes bound for the expected counterfactual Gibbs risk $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$. The PAC-Bayes bound in supervised learning is not applicable to bound $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$ since we do not have the label data to construct $R_{S_{1-t}}(G_{\hat{\rho}_t})$. This issue can be addressed by applying some existing methods in the literature, such as importance weighting [3, 7] and Integral Probability Metrics [14, 17]. We first introduce a decomposition of the Gibbs risk, which gives rise to the counterfactual variance in Theorem 1 and 2. Then we show importance weighting [3, 7] and Integral Probability Metrics [14, 17] lead to two different upper bounds for the expected counterfactual Gibbs risk $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$.

Lemma 5. *Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, a hypothesis space $\mathcal{F}$, $\forall \hat{\rho}$ on $\mathcal{F}$, we have*

$$R_{\mathcal{P}}(G_{\hat{\rho}}) = \frac{1}{4} D_{\mathcal{P}_x}(G_{\hat{\rho}}) + L_{\mathcal{P}}(G_{\hat{\rho}}) \tag{6}$$

where $D_{\mathcal{P}_x}(\hat{\rho})$ is the expected disagreement loss,

$$D_{\mathcal{P}_x}(G_{\hat{\rho}}) = \mathbb{E}_{(f,f')\sim\hat{\rho}^2} \mathbb{E}_{x\sim\mathcal{P}_x} \left[\big(f(x) - f'(x)\big)^2 \right] \tag{7}$$

and $L_{\mathcal{P}}(\hat{\rho})$ is the expected ensemble loss,

$$L_{\mathcal{P}}(G_{\hat{\rho}}) = \mathbb{E}_{(f,f')\sim\hat{\rho}^2} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[\big(\frac{f(x) + f'(x)}{2} - y \big)^2 \right] \tag{8}$$

Proof. Section 3.3. □

The expected disagreement loss and the expected ensemble loss can be rewritten as the expected counterfactual variance and the expected factual loss respectively in Theorem 2 using the notation in Section 1. The proof for this can be found in the Section 3.5 where we derive the empirical estimate of the disagreement and ensemble loss. Let $\tilde{l}_{f,f'}(x, y) = \big(\frac{f(x)+f'(x)}{2} - y \big)^2$ denote the ensemble loss. We now introduce the upper bounds of the expected counterfactual Gibbs risk using importance weighting [3, 7] and Integral Probability Metrics [14, 17].

Lemma 6. *Given two distributions $\mathcal{P}_{1-t}$ and $\mathcal{P}_t$ over $\mathcal{X} \times \mathcal{Y}$, a hypothesis space $\mathcal{F}$, $\forall \hat{\rho}_t$ on $\mathcal{F}$, we have*

$$R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t}) \leq \frac{1}{4} D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t}) + D_\infty(\mathcal{P}_{x,1-t} \| \mathcal{P}_{x,t}) L_{\mathcal{P}_t}(G_{\hat{\rho}_t}) \tag{9}$$

where $D_\infty(\mathcal{P}_{x,1-t} \| \mathcal{P}_{x,t}) = \sup_x \frac{p(x,1-t)}{p(x,t)}$.

Proof. Section 3.4. □

Notice that the problem of missing counterfactuals no long exist since $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ is defined with the input data distribution $\mathcal{P}_{x,1-t}$, and $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$ is under the distribution $\mathcal{P}_t$ where we have access to the label data. We can make use of the PAC-Bayes theory in supervised learning to upperbound $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ and $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$. Then the expected counterfactual risk $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$ is easily bounded by substituting these upperbounds into Equation (9). The results are summarized in the following theorem.

Theorem 7. *With the assumption that the squared loss function $l : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is sub-gaussian with variance factor ξ_t^2 under π_t and $\mathcal{P}_t$, and using the notation introduced in Section 1. Then, for $\delta \in (0, 1]$, $\kappa > 0$, with probability at least $1 - \delta$ over the random choice $S_t = (\mathbf{X}_t, \mathbf{Y}_t) \in \mathcal{P}_t^{N_t}$ and $S_{x,1-t} = \mathbf{X}_{1-t} \in \mathcal{P}_{x,1-t}^{N_{1-t}}$, for any posterior distribution $\hat{\rho}_t$ on $\mathcal{F}$, we have*

$$L_{\mathcal{P}_t}(G_{\hat{\rho}_t}) \leq L_{S_t}(G_{\hat{\rho}_t}) + \frac{1}{N_t} \big[2\,\mathrm{KL}(\hat{\rho}_t \| \pi_t) + \frac{1}{\delta} \big] + \frac{1}{2} \xi_t^2 \tag{10}$$

where $L_{S_t}(G_{\hat{\rho}_t}) = \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big(\frac{f(x_{i,t})+f'(x_{i,t})}{2} - y_{i,t}\big)^2$. *For some constants* $\tilde{C}_{t,1}$ *and* $\tilde{C}_{t,2}$, *we have*

$$D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t}) \leq D_{S_{x,1-t}}(G_{\hat{\rho}_t}) + \frac{\tilde{C}_{t,1}}{N_{1-t}}\big[2\,\mathrm{KL}(\hat{\rho}_t\|\pi_t) + \frac{1}{\delta}\big] + \tilde{C}_{t,2} \tag{11}$$

where $D_{S_{x,1-t}}(G_{\hat{\rho}_t}) = \frac{1}{N_{1-t}}\sum_{i=1}^{N_{1-t}}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big(f(x_{i,1-t}) - f'(x_{i,1-t})\big)^2$. *For some linear function* $\tilde{C}_{t,3}(N_t, N_{1-t})$ *and constant* $\tilde{C}_{t,4}$, *we have*

$$R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t}) \leq \frac{1}{4}D_{S_{x,1-t}}(G_{\hat{\rho}_t}) + D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})L_{S_t}(G_{\hat{\rho}_t}) + \frac{\tilde{C}_{t,3}(N_t, N_{1-t})}{N_{1-t}N_t}\Big[2\,\mathrm{KL}(\hat{\rho}_t\|\pi_t) + \frac{1}{\delta}\Big] + \tilde{C}_{t,4} \tag{12}$$

Proof. Section 3.5. □

2 Regularized Bayes

An alternative view of our algorithm is through *Regularized Bayes* in [9, 20]. In Bayesian inference, variance and reconstruction loss minimization can be understood as constraints/regularization on the approximate posterior distribution. However, this view does not provide insight into how these additional regularizers can improve generalization.

Theorem 8. *Assume the notation introduced in Section 1. Let* $\mathcal{D}$ *be the union of dataset* $\mathcal{D}_0$ *and* $\mathcal{D}_1$, $\mathcal{H}$ *a space of the neural network parameters in* ϕ *, the minimization problem in Equation (13) in the main paper is equivalent to the Multitask RegBayes optimization problem as follows,*

$$\inf_{q(w_0,w_1,\phi|\mathcal{D})}\sum_{t=0}^{1}\mathcal{L}(q(w_t,\phi|\mathcal{D})) + \Omega(q(w_t,\phi|\mathcal{D}))$$
$$s.t.\ q(w_0,w_1,\phi|\mathcal{D}) \in \mathcal{P}_{prob}$$

with

$$\Omega(q(w_t,\phi|\mathcal{D})) = \alpha_t^{'}\sum_{i=1}^{N_{1-t}}\Big(\int_{w_t}\big[(w_t - \mathbb{E}[w_t])^\top\phi(x_{i,1-t})\big]^2\cdot\int_\phi p(w_t|\phi,\mathcal{D})q(\phi|\mathcal{D})d\phi dw_t\Big)$$

and

$$\mathcal{P}_{prob} = \big\{q : q(w_0,w_1,\phi|\mathcal{D}) = \delta_{\hat{\phi}}(\phi|\mathcal{D})\prod_{t=0}^{1}q(w_t|\phi,\mathcal{D})\big\}$$

where $\mathcal{L}(q(w_t,\phi|\mathcal{D}))$ *is the KL-divergence between the approximate posterior* $q(w_t,\phi|\mathcal{D})$ *and the true posterior* $p(w_t,\phi|\mathcal{D})$, *and* $\Omega(q(w_t,\phi|\mathcal{D}))$ *is a regularizer of the approximate posterior* $q(w_t,\phi|\mathcal{D})$, $\mathcal{H} = \{\phi \in \tilde{\mathcal{H}} : \exists\ \psi \in \tilde{\mathcal{H}}\ s.t.\ \psi = \phi^{-1}\}$, $\tilde{\mathcal{H}}$ *is an arbitrary parameter space of neural network models,* $\delta_{\hat{\phi}}(\phi|\mathcal{D})$ *is a Dirac measure concentrated at* $\hat{\phi} \in \mathcal{H}$ *and* $\alpha_t^{'} = \frac{\alpha_t N_t}{N_{1-t}}$.

Proof. See Theorem 1 in [9]. □

3 Detailed Proofs

3.1 Proof of Proposition 3

Proof.

$$
\begin{aligned}
\epsilon_{\text{PEHE}} &= \int \left(\hat{\tau}(x) - \tau(x)\right)^2 p(x)dx \\
&= \int \Big(\big(\hat{f}_1(x) - \hat{f}_0(x)\big) - \big(f_1^*(x) - f_0^*(x)\big) \Big)^2 p(x)dx \\
&= \int \Big(\big(\hat{f}_1(x) - f_1^*(x)\big) + \big(\hat{f}_0(x) - f_0^*(x)\big) \Big)^2 p(x)dx \\
&\leq 2 \sum_{t=0}^{1} \int \big(\hat{f}_t(x) - f_t^*(x)\big)^2 p(x)dx \\
&= 2 \sum_{t=0}^{1} \sum_{t'=0}^{1} \int \big(\hat{f}_t(x) - f_t^*(x)\big)^2 p(x,t')dx \\
&= 2 \sum_{t=0}^{1} \int \big(\hat{f}_t(x) - f_t^*(x)\big)^2 p(x,t)dx + 2 \sum_{t=0}^{1} \int \big(\hat{f}_t(x) - f_t^*(x)\big)^2 p(x,1-t)dx \\
&= 2 \sum_{t=0}^{1} (R_{\mathcal{P}_t}(B_{\hat{\rho}_t}) + R_{\mathcal{P}_{1-t}}(B_{\hat{\rho}_t}) - 2\beta_t^{-1})
\end{aligned}
\tag{13}
$$

The first inequality is achieved using the triangle inequality. The last equality is achieved as follows,

$$
\begin{aligned}
R_{\mathcal{P}_t}(B_{\hat{\rho}_t}) &= \int \big(\hat{f}_t(x) - y_t\big)^2 p(y_t|x)p(x,t)dxdy_t \\
&= \int \big(\hat{f}_t(x) - f_t^*(x) - \epsilon_t\big)^2 \mathcal{N}(\epsilon_t; 0, \beta_t^{-1}) p(x, T=t)dxdy_t \\
&= \int \Big(\big(\hat{f}_t(x) - f_t^*(x)\big)^2 - 2\epsilon_t \big(\hat{f}_t(x) - f_t^*(x)\big) + \epsilon_t^2 \Big) \mathcal{N}(\epsilon_t; 0, \beta_t^{-1}) p(x, T=t)dxdy_t \\
&= \int \big(\hat{f}_t(x) - f_t^*(x)\big)^2 p(x,t)dx + \beta_t^{-1}
\end{aligned}
$$

□

3.2 Proof of Theorem 4

The proof is based on the following PAC-Bayes theorem in supervised learning. We can bound the expected factual risk $R_{\mathcal{P}_t}(G_{\hat{\rho}})$ using Equation (16) since the factual risk minimization is a supervised learning problem where we have access to the label information.

Theorem 9. *(Alquier et al. [2]) Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, a hypothesis space $\mathcal{F}$, a loss function $l : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, a prior distribution π over $\mathcal{F}$, a $\delta \in (0, 1]$, and a real number $\kappa > 0$, with probability at least $1 - \delta$ over the random choice $S = (\mathbf{X}, \mathbf{Y}) \in \mathcal{P}^n$, $\forall \hat{\rho}$ on $\mathcal{F}$, we have*

$$
R_{\mathcal{P}}(G_{\hat{\rho}}) \leq R_S(G_{\hat{\rho}}) + \frac{1}{\kappa}\big[\,\mathrm{KL}(\hat{\rho}\|\pi) + \frac{1}{\delta} + n\zeta_{\pi,\mathcal{P}}(\kappa, n)\big] \tag{14}
$$

where

$$
\zeta_{\pi,\mathcal{P}}(\kappa, n) = \ln \mathbb{E}_{f\sim\pi} \mathbb{E}_{(x,y)\sim\mathcal{P}} \Big[\exp\Big(\frac{\kappa}{n} \big(R_{\mathcal{P}}(f) - l(f(x), y) \big) \Big) \Big], \tag{15}
$$

$R_{\mathcal{P}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\big[l(f(x), y)\big]$ is the expected risk and $R_S(G_{\hat{\rho}}) = \frac{1}{n}\sum_{(x,y)\in S} l(f(x), y)$ is the empirical risk on S.

Note that the inner expectation in $\zeta_{\pi,\mathcal{P}}(\kappa, n)$ is the moment generating function of a zero mean random variable $V = R_{\mathcal{P}} - l(f(x), y)$. If we assume the loss $l(f(x), y)$ is sub-gaussian variable with variance factor ξ^2 under the prior π and data distribution $\mathcal{P}$, we have $\zeta_{\pi,\mathcal{P}}(\kappa, n) \leq \frac{\kappa^2\xi^2}{2n^2}$ [6]. Because the moment generating function of a zero-mean sub-gaussian variable is upper bounded by the moment generating function of the zero-mean Gaussian variable with the same variance factor ξ^2, s.t. $\mathbb{E}_V\big[\exp(\frac{\kappa}{n}V)\big] \leq \exp(\frac{\kappa^2\xi^2}{2n^2})$. The sub-gaussian assumption of $l(f(x), y)$ is more realistic than the standard bounded assumption because the squared loss function is usually unbounded but with strong tail decay property. By choosing $\kappa := n$, the bound in Equation (14) converges to,

$$R_{\mathcal{P}}(G_{\hat{\rho}}) \leq R_S(G_{\hat{\rho}}) + \frac{1}{n}\big[\,\mathrm{KL}(\hat{\rho}\|\pi) + \frac{1}{\delta}\big] + \frac{1}{2}\xi^2 \tag{16}$$

The result in Theorem 4 is rewritten from Equation (18) using the notation in Section 1.

3.3 Proof of Lemma 5

Proof. $\forall\hat{\rho}$ on $\mathcal{F}$, we have

$$\begin{aligned}
R_{\mathcal{P}}(G_{\hat{\rho}}) =& \mathbb{E}_{f\sim\hat{\rho}}\mathbb{E}_{(x,y)\sim\mathcal{P}}\big[\big(f(x)-y\big)^2\big] \\
=& \frac{1}{2}\mathbb{E}_{(f,f')\sim\hat{\rho}^2}\mathbb{E}_{(x,y)\sim\mathcal{P}}\big[\big(f(x)-y\big)^2 + \big(f'(x)-y\big)^2\big] \\
=& \frac{1}{2}\mathbb{E}_{(f,f')\sim\hat{\rho}^2}\mathbb{E}_{(x,y)\sim\mathcal{P}}\big[f^2(x) + f'^2(x) - 2yf(x) - 2yf'(x) + 2y^2\big] \\
=& \frac{1}{2}\mathbb{E}_{(f,f')\sim\hat{\rho}^2}\mathbb{E}_{(x,y)\sim\mathcal{P}}\big[\frac{1}{2}\big(f^2(x) + f'^2(x) - 2f(x)f'(x)\big) + \frac{1}{2}\big(f^2(x) + f'^2(x) \\
& + 2f(x)f'(x) - 4y(f(x)+f'(x)) + 4y^2\big)\big] \\
=& \frac{1}{2}\mathbb{E}_{(f,f')\sim\hat{\rho}^2}\mathbb{E}_{(x,y)\sim\mathcal{P}}\big[\frac{1}{2}\big(f(x)-f'(x)\big)^2 + \frac{1}{2}\big(f(x)+f'(x)-2y\big)^2\big] \\
=& \frac{1}{2}\mathbb{E}_{(f,f')\sim\hat{\rho}^2}\mathbb{E}_{(x,y)\sim\mathcal{P}}\big[\frac{1}{2}\big(f(x)-f'(x)\big)^2 + 2\big(\frac{f(x)+f'(x)}{2} - y\big)^2\big] \\
=& \frac{1}{4}D_{\mathcal{P}_x}(G_{\hat{\rho}}) + L_{\mathcal{P}}(G_{\hat{\rho}})
\end{aligned} \tag{17}$$

□

3.4 Proof of Lemma 6

Proof. From Lemma 5, we have $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t}) = \frac{1}{4}D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t}) + L_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$. Using the importance weighting trick in [3, 7], we can split $L_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$ into two parts,

$$\begin{aligned}
L_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t}) =& \mathbb{E}_{(x,y)\sim\mathcal{P}_{1-t}}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big[\tilde{l}_{f,f'}(x,y)\big] \\
=& \mathbb{E}_{(x,y)\sim\mathcal{P}_t}\Big[\frac{\mathcal{P}_{1-t}}{\mathcal{P}_t}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big[\tilde{l}_{f,f'}(x,y)\big]\Big] \\
\leq& \left(\mathbb{E}_{x\sim\mathcal{P}_{x,t}}\Big[\Big(\frac{p(x,1-t)}{p(x,t)}\Big)^{\alpha}\Big]\right)^{\frac{1}{\alpha}}\Big[\mathbb{E}_{(x,y)\sim\mathcal{P}_{1-t}}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big[\tilde{l}_{f,f'}(x,y)\big]^{\frac{\alpha}{\alpha-1}}\Big]^{1-\frac{1}{\alpha}}
\end{aligned} \tag{18}$$

The second equality is achieved because the Overlap assumption in the Rubin-Neyman potential outcomes model [13] ensures that $\mathrm{SUPP}(\mathcal{P}_{1-t}) = \mathrm{SUPP}(\mathcal{P}_t)$. The first inequality is achieved by using Hölder's inequality. Note that the label distribution $p(y_t|x)$ is canceled out since it is shared by $\mathcal{P}_{1-t} = p(y_t, x, 1-t) = p(y_t|x)p(x, 1-t)$ and $\mathcal{P}_t = p(y_t, x, t) = p(y_t|x)p(x, t)$. Taking the positive constant $\alpha \to +\infty$, we obtain

$$R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t}) \leq \frac{1}{4}D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t}) + D_{\infty}(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})L_{\mathcal{P}_t}(G_{\hat{\rho}_t}). \tag{19}$$

□

3.5 Proof of Theorem 7

Proof. The proof is to show the terms $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$ and $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ can be bounded using the supervised PAC-Bayes bound in Theorem 9. The key of applying Theorem 9 is to show Equation (15) is bounded for the proposed Gibbs model. In $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ and $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$, the hypothesis pair (f, f') is drawn from a product posterior distribution $\hat{\rho}_t^2$ on $\mathcal{F}^2$. We also consider a product prior π_t^2 over $\mathcal{F}^2$. Recall that $\tilde{l}_{f,f'}(x,y) = l\big(\frac{f(x)+f'(x)}{2}, y\big) = \big(\frac{f(x)+f(x')}{2} - y\big)^2$. The Equation (15) for $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$ is given as

$$\zeta_{\pi_t^2,\mathcal{P}_t}(\kappa, N_t) = \ln \mathbb{E}_{(f,f')\sim\pi_t^2}\mathbb{E}_{(x,y)\sim\mathcal{P}_t}\Big[\exp\Big(\frac{\kappa}{N_t}\big(\mathbb{E}_{(x',y')\sim\mathcal{P}_t}\big[\tilde{l}_{f,f'}(x',y')\big] - \tilde{l}_{f,f'}(x,y)\big)\Big)\Big] \tag{20}$$

Recall that the prior $\pi_t = \mathcal{N}(w_t; \mathbf{0}, \lambda_t^{-1}\mathbf{I})$ is given as a Gaussian distribution in Section 1. For $(f, f') \sim \pi_t^2$, the ensemble model $\frac{f+f'}{2}$ in $\tilde{l}_{f,f'}(x,y)$ follows the same Gaussian prior distribution π_t. Let $\tilde{f} = \frac{f+f'}{2}$, we can rewrite Equation (20) as

$$\zeta_{\pi_t^2,\mathcal{P}_t}(\kappa, N_t) = \ln \mathbb{E}_{(x,y)\sim\mathcal{P}_t}\mathbb{E}_{\tilde{f}\sim\pi_t}\Big[\exp\Big(\frac{\kappa}{N_t}\big(\mathbb{E}_{(x',y')\sim\mathcal{P}_t}\big[l(\tilde{f}(x'), y')\big] - l(\tilde{f}(x), y)\big)\Big)\Big] \tag{21}$$

Under the sub-gaussian assumption of $l(\tilde{f}(x), y)$, we have $\zeta_{\pi^2,\mathcal{P}_t}(\kappa, N_t) \leq \frac{\kappa^2 s^2}{2N_t^2}$. Applying Equation (15) to $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$ gives the bound in Equation (10) where the double KL divergence is from $\mathrm{KL}(\hat{\rho}^2\|\pi^2) = 2\,\mathrm{KL}(\hat{\rho}\|\pi)$ (see Lemma 10). Recall that $f(x) = w_t^\top\phi(x)$, we define $\tilde{f}(x) = \tilde{w}_t^\top\phi(x) = (w_t - w_t')^\top\phi(x)$. Let $\tilde{\pi}_t$ denote the prior distribution $\mathcal{N}(\tilde{w}_t; \mathbf{0}, 2\lambda_t^{-1}\mathbf{I})$, $\tilde{f}(x) = f(x) - f'(x)$, $\tilde{d}_{\tilde{f}}(x) = \big(f(x) - f'(x)\big)^2$ and $\gamma_t = \mathbb{E}_{x\sim\mathcal{P}_{x,1-t}}\big[\|\phi(x)\|^2\big]$. Note that we do not assume $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ is sub-gaussian, we can upperbound the Equation (15) for $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ as follows,

$$\begin{aligned}
\zeta_{\tilde{\pi}_t,\mathcal{P}_{x,1-t}}(\kappa, N_{1-t}) &= \ln \mathbb{E}_{x\sim\mathcal{P}_{x,1-t}}\mathbb{E}_{\tilde{f}\sim\tilde{\pi}_t}\Big[\exp\Big(\frac{\kappa}{N_{1-t}}\big(\mathbb{E}_{x'\sim\mathcal{P}_{x,1-t}}\big[\tilde{d}_{\tilde{f}}(x')\big] - \tilde{d}_{\tilde{f}}(x)\big)\Big)\Big] \\
&\leq \ln \mathbb{E}_{\tilde{w}_t}\Big[\exp\Big(\frac{\kappa}{N_{1-t}}\mathbb{E}_{x\sim\mathcal{P}_{x,1-t}}\big[(\tilde{w}_t^\top\phi(x))^2\big]\Big)\Big] \\
&\leq \ln \mathbb{E}_{\tilde{w}_t}\Big[\exp\Big(\tilde{\gamma}\tilde{w}_t^\top\tilde{w}_t\Big)\Big] \\
&= \ln \int \frac{1}{\sqrt{(2\pi)^{d_\phi}(2\lambda_t^{-1})^{d_\phi}}}\exp\Big(-\frac{1}{2}(\frac{\lambda_t}{2} - 2\tilde{\gamma})\tilde{w}_t^\top\tilde{w}_t\Big) \\
&= -\frac{d_\phi}{2}\ln\Big(1 - \frac{4\tilde{\gamma}}{\lambda_t}\Big) \\
&\leq \frac{2d_\phi\tilde{\gamma}}{\lambda_t - 4\tilde{\gamma}}
\end{aligned} \tag{22}$$

where $\tilde{\gamma} = \frac{\kappa\gamma_t}{N_{1-t}}$, the first inequality is achieved by $\tilde{d}_{\tilde{f}}(x) \geq 0$ and the last inequality is achieved by $\ln(1-c) \geq \frac{-c}{1-c}$, $c \in [0, 1)$. We need to set $\kappa > 0$ s.t. $\lambda_t - 4\tilde{\gamma} > 0 \Rightarrow 0 < \kappa < \frac{N_{1-t}\lambda_t}{4\gamma_t}$. Let $\kappa = \frac{N_{1-t}\lambda_t}{C_\kappa\gamma_t}$ for some $C_\kappa > 4$. Then we have $\tilde{\gamma} = \frac{\lambda_t}{C_\kappa}$ and $\zeta_{\tilde{\pi}_t,\mathcal{P}_{x,1-t}}(\kappa, N_{1-t}) \leq \frac{2d_\phi}{C_\kappa - 4}$. Substituting $\kappa = \frac{N_{1-t}\lambda_t}{C_\kappa\gamma_t}$ and the upper bound of $\zeta_{\tilde{\pi}_t,\mathcal{P}_{x,1-t}}(\kappa, N_{1-t})$ into Equation (14), we obtain the bound for $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ in Equation (11), where $\tilde{C}_{t,1} = C_\kappa\lambda_t^{-1}\gamma_t$ and $\tilde{C}_{t,2} = 2d_\phi(C_\kappa - 4)^{-1}C_\kappa\lambda_t^{-1}\gamma_t$. Bounding $D_{\mathcal{P}_{x,1-t}}(G_{\hat{\rho}_t})$ and $L_{\mathcal{P}_t}(G_{\hat{\rho}_t})$ in Equation (9) with Equation (10) and (11) respectively, we obtain the bound for $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$ in Equation (12), where $\tilde{C}_{t,3}(N_t, N_{1-t}) = \frac{\tilde{C}_{t,1}}{4}N_t + D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})N_{1-t}$ and $\tilde{C}_{t,4} = \frac{1}{4}\tilde{C}_{t,2} + \frac{1}{2}D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})\xi_t^2$. □

Lemma 10. *Given any two distributions $\hat{\rho}$ and π over a hypothesis space $\mathcal{F}$. Suppose $\hat{\rho}^2(f, f') = \hat{\rho}(f)\hat{\rho}(f')$ and $\pi^2(f, f') = \pi(f)\pi(f')$, we have*

$$\mathrm{KL}(\hat{\rho}^2\|\pi^2) = 2\,\mathrm{KL}(\hat{\rho}\|\pi).$$

Proof. Given any two distributions $\hat{\rho}$ and π, we have

$$\begin{aligned}\mathrm{KL}(\hat{\rho}^2\|\pi^2) &= \int \hat{\rho}(f)\hat{\rho}(f')\log\frac{\hat{\rho}(f)\hat{\rho}(f')}{\pi(f)\pi(f')}df\,df' \\ &= \int \hat{\rho}(f)\log\frac{\hat{\rho}(f)}{\pi(f)}df + \int \hat{\rho}(f')\log\frac{\hat{\rho}(f')}{\pi(f')}df\,df' \\ &= 2\,\mathrm{KL}(\hat{\rho}\|\pi).\end{aligned} \tag{23}$$

□

3.6 Proof of Theorem 1

Proof. Substituting the bounds for $R_{\mathcal{P}_t}(G_{\hat{\rho}_t})$ and $R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t})$ in Equation (5) and (12) into Equation (2) gives that

$$\begin{aligned}\epsilon_{\text{PEHE}} \leq& 2\sum_{t=0}^{1}(R_{\mathcal{P}_t}(B_{\hat{\rho}_t}) + R_{\mathcal{P}_{1-t}}(B_{\hat{\rho}_t}) - 2\beta_t^{-1}) \\ \leq& 2\sum_{t=0}^{1}(R_{\mathcal{P}_t}(G_{\hat{\rho}_t}) + R_{\mathcal{P}_{1-t}}(G_{\hat{\rho}_t}) - 2\beta_t^{-1}) \\ \leq& \sum_{t=0}^{1}\Bigg(\frac{1}{2}D_{S_{x,1-t}}(G_{\hat{\rho}_t}) + 2D_\infty(\mathcal{P}_{x,1-t}\|\mathcal{P}_{x,t})L_{S_t}(G_{\hat{\rho}_t}) + 2R_{S_t}(G_{\hat{\rho}_t}) \\ &+ \frac{C_{t,1}(N_t,N_{1-t})}{N_tN_{1-t}}\,\mathrm{KL}(\hat{\rho}_t\|\pi_t) + \frac{C_{t,2}(N_t,N_{1-t})}{N_tN_{1-t}}\frac{1}{\delta} + C_{t,3}\Bigg)\end{aligned} \tag{24}$$

We express $C_{t,1}(N_t,N_{1-t})$, $C_{t,1}(N_t,N_{1-t})$, $C_{t,3}$ in terms of the linear function $\tilde{C}_{t,3}(N_t,N_{1-t})$ and constant $\tilde{C}_{t,4}$ defined in the proof Theorem 7 (See Section 3.5): $C_{t,1}(N_t,N_{1-t}) = 4\tilde{C}_{t,3}(N_t,N_{1-t}) + 2N_{1-t}$, $C_{t,2}(N_t,N_{1-t}) = 2\tilde{C}_{t,3}(N_t,N_{1-t}) + 2N_{1-t}$ and $C_{t,3}(N_t,N_{1-t}) = \xi_t^2 + 2\tilde{C}_{t,4} - 4\beta_t^2$. We now compute the explicit expression of the empirical terms in Equation (24) with the posterior distribution $\hat{\rho}_t = \mathcal{N}\big(\mu(x|\mathcal{D}_t,\Theta_t),\sigma^2(x|\mathcal{D}_t,\Theta_t)\big)$ where $\mu(x|\mathcal{D}_t,\Theta_t) = \mathbf{m}_{w_t}^\top\phi(x)$, and $\sigma^2(x|\mathcal{D}_t,\Theta_t) = \phi(x)^\top\mathbf{K}_{w_t}^{-1}\phi(x)$. The empirical disagreement loss can be expressed as the variance of the missing counterfactuals,

$$\begin{aligned}D_{S_{x,1-t}}(G_{\hat{\rho}_t}) &= \frac{1}{N_{1-t}}\sum_{i=1}^{N_{1-t}}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big(f(x_{i,1-t}) - f'(x_{i,1-t})\big)^2 \\ &= \frac{1}{N_{1-t}}\sum_{i=1}^{N_{1-t}}2\sigma^2(x_{i,1-t}|\mathcal{D}_t,\Theta_t) + 2\mu^2(x_{i,1-t}|\mathcal{D}_t,\Theta_t) - 2\mu^2(x_{i,1-t}|\mathcal{D}_t,\Theta_t) \\ &= 2\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_{1-t})\end{aligned}$$

The empirical ensemble loss is

$$\begin{aligned}L_{S_t}(G_{\hat{\rho}_t}) &= \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{E}_{(f,f')\sim\hat{\rho}_t^2}\big(\frac{f(x_{i,t}) + f'(x_{i,t})}{2} - y_{i,t}\big)^2 \\ &= \frac{1}{N_t}\sum_{i=1}^{N_t}\mu^2(x_{i,t}|\mathcal{D}_t,\Theta_t) + \frac{\sigma^2(x_{i,t}|\mathcal{D}_t,\Theta_t)}{2} - 2\mu(x_{i,t}|\mathcal{D}_t,\Theta_t)y_{i,t} + y_{i,t}^2 \\ &= \frac{1}{N_t}\sum_{i=1}^{N_t}\Big[\big(y_{i,t} - \mu(x_{i,t}|\mathcal{D}_t,\Theta_t)\big)^2 + \frac{\sigma^2(x_{i,t}|\mathcal{D}_t,\Theta_t)}{2}\Big] \\ &= L_{\hat{\rho}_t}(\mathbf{X}_t,\mathbf{Y}_t) + \frac{1}{2}\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_t)\end{aligned}$$

The empirical factual Gibbs risk is

$$\begin{aligned}
R_{S_t}(G_{\hat{\rho}_t}) &= \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{E}_{f\sim\hat{\rho}_t}\big(f(x_{i,t})-y_{i,t}\big)^2 \\
&= \frac{1}{N_t}\sum_{i=1}^{N_t}\Big[\big(\mu(x_{i,t}|\mathcal{D}_t,\Theta_t)-y_{i,t}\big)^2+\sigma^2(x_{i,t}|\mathcal{D}_t,\Theta_t)\Big] \\
&= L_{\hat{\rho}_t}(\mathbf{X}_t,\mathbf{Y}_t)+\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_t)
\end{aligned}$$

Substituting the above expressions of $D_{S_{x,1-t}}(G_{\hat{\rho}_t})$, $L_{S_t}(G_{\hat{\rho}_t})$ and $R_{S_t}(G_{\hat{\rho}_t})$ into Equation (24), we obtain the bound in Equation (1). □

Lemma 11. *In DKLITE, the negative log marginal likelihood $\mathcal{L}(\Theta_t)$ can be rewritten as,*

$$\mathcal{L}(\Theta_t)=\mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{N_t\beta_t}{2}\Big[\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_t)+L_{\hat{\rho}_t}(\mathbf{X}_t,\mathbf{Y}_t)\Big]$$

Proof.

$$\begin{aligned}
\mathcal{L}(\Theta_t) &= -\frac{d_\phi}{2}\ln\lambda_t-\frac{N_t}{2}\ln\beta_t+\frac{N_t}{2}\ln(2\pi)+\frac{\beta_t}{2}\|\mathbf{Y}_t-\mathbf{\Phi}_t\mathbf{m}_{w_t}\|^2+\frac{\lambda_t}{2}\mathbf{m}_{w_t}^\top\mathbf{m}_{w_t}+\frac{1}{2}\ln|\mathbf{K}_{w_t}| \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{d_\phi}{2}-\frac{\lambda_t}{2}\mathrm{Tr}(\mathbf{K}_{w_t}^{-1})+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{\beta_t}{2}\|\mathbf{Y}_t-\mathbf{\Phi}_t\mathbf{m}_{w_t}\|^2 \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{1}{2}\mathrm{Tr}(\mathbf{K}_{w_t}\mathbf{K}_{w_t}^{-1})-\frac{\lambda_t}{2}\mathrm{Tr}(\mathbf{K}_{w_t}^{-1})+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{\beta_t}{2}\|\mathbf{Y}_t-\mathbf{\Phi}_t\mathbf{m}_{w_t}\|^2 \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{1}{2}\mathrm{Tr}((\mathbf{K}_{w_t}-\lambda_t\mathbf{I})\mathbf{K}_{w_t}^{-1})+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{\beta_t}{2}\|\mathbf{Y}_t-\mathbf{\Phi}_t\mathbf{m}_{w_t}\|^2 \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{\beta_t}{2}\mathrm{Tr}((\mathbf{\Phi}_t^\top\mathbf{\Phi}_t\mathbf{K}_{w_t}^{-1})+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{\beta_t}{2}\|\mathbf{Y}_t-\mathbf{\Phi}_t\mathbf{m}_{w_t}\|^2 \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{\beta_t}{2}\mathrm{Tr}((\mathbf{\Phi}_t\mathbf{K}_{w_t}^{-1}\mathbf{\Phi}_t^\top)+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{\beta_t}{2}\sum_{i=1}^{N_t}\big(y_{i,t}-\mu(x_{i,t}|\mathcal{D}_t,\Theta_t)\big)^2 \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{\beta_t}{2}\sum_{i=1}^{N_t}\Big[\big(y_{i,t}-\mu(x_{i,t}|\mathcal{D}_t,\Theta_t)\big)^2+\sigma^2(x_{i,t}|\mathcal{D}_t,\Theta_t)\Big] \\
&= \mathrm{KL}(\hat{\rho}_t\|\pi_t)+\frac{N_t}{2}\ln(2\pi\beta_t^{-1})+\frac{N_t\beta_t}{2}\Big[\mathrm{Var}_{\hat{\rho}_t}(\mathbf{X}_t)+L_{\hat{\rho}_t}(\mathbf{X}_t,\mathbf{Y}_t)\Big]
\end{aligned}$$

where the second equality is achieved by

$$\mathrm{KL}(\hat{\rho}_t\|\pi_t)=\frac{\lambda_t}{2}\mathrm{Tr}(\mathbf{K}_{w_t}^{-1})+\frac{\lambda_t}{2}\mathbf{m}_{w_t}^\top\mathbf{m}_{w_t}-\frac{d_\phi}{2}-\frac{d_\phi}{2}\ln\lambda_t+\frac{1}{2}\ln|\mathbf{K}_{w_t}|$$

□

4 Hyperparameter selection

Hyperparameters	Range
Variance regularization parameter α_1	$\{0.001, 0.01, 0.1, 1, 10, 25, 50, 75, 100\}$
Reconstruction regularization parameter α_2	$\{0.001, 0.01, 0.1, 1, 10, 25, 50, 75, 100\}$
Number of hidden layers	$\{1, 2, 3\}$
Number of hidden units	$\{50, 100, 150, 200\}$
Dimension of the feature map	$\{25, 50, 75, 100\}$
Regression Form	$\{\text{Primal}, \text{Dual}\}$

Table 1: Hyperparameters and ranges of DKLITE

Due to the missing of counterfactual outcomes, standard methods for hyperparameter selection, such as cross-validation, are not generally applicable for estimating the PEHE loss. Following the hyperparameter optimization code of the work [14], we choose hyperparameters using random search over the hyperparameter space in Table 1. The missing counterfactual outcomes in the PEHE loss are approximated by the observed outcome of the nearest neighbor in the opposite group.

5 Further experimental details and additional figures

5.1 Data description

IHDP. *Counterfactual outcomes are randomly generated via a predefined probabilistic model* [1, 8, 14, 18, 19]. The objective is to estimate the effects of specialist home visits to individuals on their future cognitive test scores. Patient covariates X were collected from the actual randomized experiment but the overall cohort was made artificially imbalanced by removing a subset of the treated population. The dataset comprises 747 units (139 treated, 608 control) and 25 covariates measuring aspects of children and their mother[2].

• *Metric.* We evaluate performance on **IHDP** with the empirical precision in estimating heterogeneous effects,

$$\hat{\epsilon}_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}(\tau(x_i) - \hat{\tau}(x_i))^2 \tag{25}$$

Twins. *Outcomes are observed but the treatment assignment is simulated.* The objective is to predict the mortality of each of one of two twins in their first year. We consider the "treated" twin to be the one with higher weight at birth and, since we have records for both twins, we treat their outcomes as two potential outcomes, i.e. $y(1)$ and $y(0)$. Now, in order to simulate an observational study, we need to select one of the two twins for inclusion into our data, that is define $Pr(T|X)$. We do so by sampling from $t|x \sim \text{Bern}(\text{Sigmoid}(w^\top x + \epsilon))$ where $w^\top \sim \text{Uniform}((-0.1, 0.1)^{30\times 1})$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. The final data contains 11400 individuals with 30 measured covariates relating to their parents, the pregnancy and their birth.

• *Metric.* We evaluate performance on **Twins** with the observed precision in estimating heterogeneous effects,

$$\tilde{\epsilon}_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}(y_{i,1} - y_{i,0} - \hat{\tau}(x_i))^2 \tag{26}$$

Jobs. *Approximately real data* [4, 10, 14]. The objective is to measure the *policy risk* of a government job training program on future employment and income. The study includes 7 pre-treatment covariates such as demographics and previous employment details, on a total of 3112 individuals, 297 of them were enrolled in the program and 2915 of them

[2]Outcomes $Y(0)$ and $Y(1)$ are obtained by implementing the setting "A" in the NPCI package [5].

are considered the control sample (from a mixture of randomized and experimental samples, see [10] for a description of the data collection process).

• *Metric.* We evaluate performance on **Jobs** with the policy risk, it measures the expected loss if the treatment is taken according to the ITE policy prescribed by the algorithm. It is defined as:

$$\begin{aligned}\mathcal{R}_{pol} = 1 - \mathbb{E}[Y(1)|\pi(x) = 1]P(\pi(x) = 1) \\ + \mathbb{E}[Y(0)|\pi(x) = 0]P(\pi(x) = 0\end{aligned} \tag{27}$$

where $\pi(x) = 1$ if $\hat{y}(1) - \hat{y}(0) > 0$ and $\pi(x) = 0$, otherwise.

For the IHDP dataset, we average over 1000 realizations of the outcomes with $63/27/10\%$ train/validation/test split, as suggested in [14]. For Twins and Jobs dataset, each dataset is divided $56/24/20\%$ into training/validation/testing sets, and we report the results average over 100 realizations.

5.2 Estimation of Average Treatment Effects

The estimation of treatment effects on average over a population is much more widely studied. It is useful in many domains and typically a first step into understanding the impact of an intervention. In this section, we evaluate all algorithms on all benchmark data sets for the average treatment effect. Below with give the exact formulation of the metrics use and results.

Metrics. In the IHDP dataset, we use the empirical absolute error of average treatment effect $\hat{\epsilon}_{\text{ATE}}$,

$$\hat{\epsilon}_{\text{ATE}} = \big|\frac{1}{N}\sum_{i=1}^{N}\tau(x_i) - \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}(x_i)\big| \tag{28}$$

In the Twins dataset, we use the approximate absolute error of average treatment effect $\tilde{\epsilon}_{\text{ATE}}$,

$$\tilde{\epsilon}_{\text{ATE}} = \big|\frac{1}{N}\sum_{i=1}^{N}(y_{i,1} - y_{i,0}) - \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}(x_i)\big| \tag{29}$$

In the Jobs dataset, we use the true average treatment effect on the treated (ATT),

$$\text{ATT} = \frac{1}{|Q_1|}\sum_{x_i \in Q_1} y_{i,1} - \frac{1}{|Q_0|}\sum_{x_i \in Q_0} y_{i,0} \tag{30}$$

where $Q_j = W_j \cup E$, $j \in 0, 1$. The absolute error of ATT is given as,

$$\hat{\epsilon}_{\text{ATT}} = \big|\text{ATT} - \frac{1}{|Q_1|}\sum_{x_i \in Q_1}(f_1(x_i) - f_0(x_i))\big| \tag{31}$$

Results. Full performance results for the ATE problem are given in Table 2. Suggested by one of the reviewers, we add an additional experiment for top 3 benchmarks on the LBIDD dataset [11] in the IBM causal inference benchmarking framework [16], which enables us to run experiments on a much wider range of data generating processes. The results are given in Table 3.

5.3 Additional figures

Similarly to Figure 1 in the main body of the paper, we illustrate in Figure 1 the T-SNE visualizations of the learned embeddings for the treated potential outcomes $Y(1)$ of the IHDP dataset optimized for different loss functions. In Figure 2, we provide the corresponding T-SNE visualizations of the learned embeddings for the control group (in purple) and treated group (in yellow) of the IHDP dataset.

Dataset (Mean ± Std)	IHDP ($\hat{\epsilon}_{\text{ATE}}$)		Twins ($\tilde{\epsilon}_{\text{ATE}}$)		Jobs ($\hat{\epsilon}_{\text{ATT}}$)	
Metrics / Dataset	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample
OLS/LR$_1$	.73 ± .04	.94 ± .06	.0038 ± .0025	.0069 ± .0056	**.01 ± .00**	.08 ± .04
OLS/LR$_2$	.14 ± .01	.31 ± .02	.0039 ± .0025	.0070 ± .0059	**.01 ± .01**	.08 ± .03
BLR	.72 ± .04	.93 ± .05	.0057 ± .0036	.0334 ± .0092	**.01 ± .01**	.08 ± .03
K-NN	.14 ± .01	.90 ± .05	.0028 ± .0021	.0051 ± .0039	.21 ± .01	.13 ± .05
BART	.23 ± .01	.34 ± .02	.1206 ± .0236	.1265 ± .0234	.02 ± .00	.08 ± .03
R-FOREST	.73 ± .05	.96 ± .06	.0049 ± .0034	.0080 ± .0051	.03 ± .01	.09 ± .04
C-FOREST	.18 ± .01	.40 ± .03	.0286 ± .0035	.0335 ± .0083	.03 ± .01	.07 ± .03
BNN	.37 ±.03	.42 ± .03	.0056 ± .0032	.0203 ± .0071	.04 ± .01	.09 ± .04
TARNET	.26 ± .01	.28 ± .01	.0108 ± .0017	.0151 ± .0018	.05 ± .02	.11 ± .04
CAR$_{\text{WASS}}$	.25 ± .01	.27 ± .01	.0112 ± .0016	.0284 ± .0032	.04 ± .01	.09 ± .03
CMGP	.11 ± .10	.13 ± .12	.0124 ± .0051	.0143 ± .0116	.06± .06	.09 ± .07
DKLITE	**.08 ± .01**	**.10 ± .02**	**.0035 ± .0024**	**.0042 ± .0025**	.03 ± .01	.08 ± .02
DKLITE-U	**.08 ± .01**	**.09 ± .02**	**.0034 ± .0020**	**.0043 ± .0021**	.03 ± .01	.08 ± .01

Table 2: Performance of average treatment effect estimation on three real-world datasets

Metrics	PEHE		ATE	
Mean ± Std	In-sample	Out-sample	In-sample	Out-sample
CAR$_{\text{WASS}}$	27.08 ± 2.49	26.92 ± 2.48	25.44 ± 2.45	25.39 ± 2.46
CMGP	27.21 ± 2.48	27.24 ± 2.49	25.28± 2.45	25.30 ± 2.44
DKLITE	**25.41 ± 2.43**	**25.19 ± 2.44**	**23.28 ± 2.37**	**23.25 ± 2.38**

Table 3: Performanceof PEHE and ATE on the LBIDD dataset

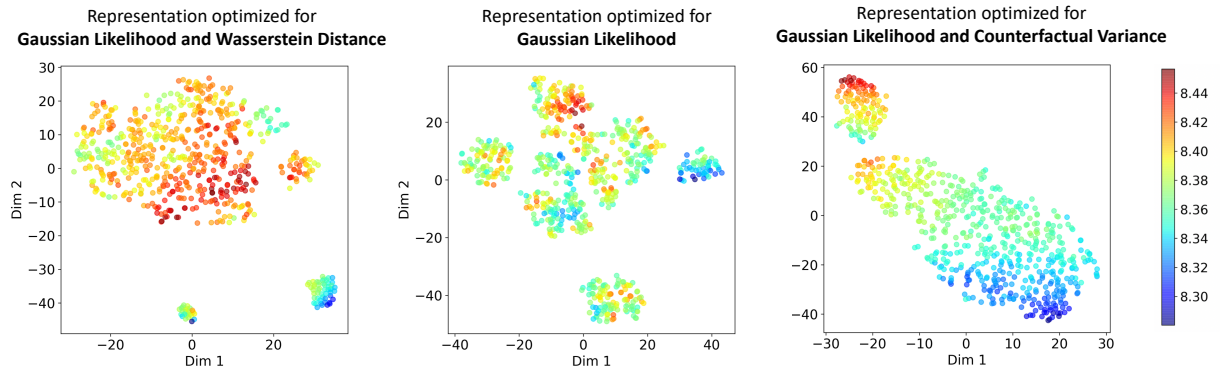

Figure 1: T-SNE visualizations of the learned embeddings for the treated potential outcomes $Y(1)$.

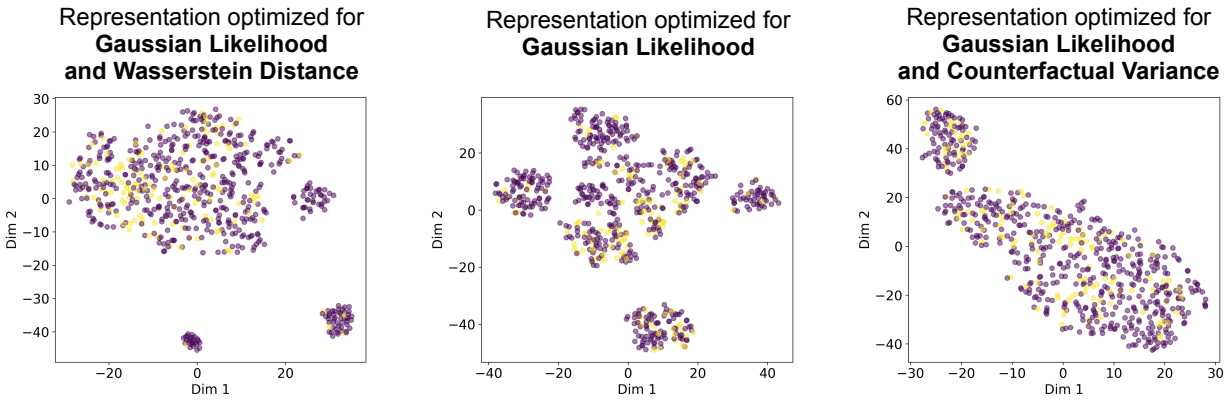

Figure 2: T-SNE visualizations of the learned embeddings for the control group (in purple) and treated group (in yellow) of the IHDP dataset.

References

[1] A. M. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[2] P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

[3] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[4] R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[5] V. Dorie. Npci: Non-parametrics for causal inference. *URL: https://github. com/vdorie/npci*, 2016.

[6] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

[7] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pages 738–746, 2013.

[8] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[9] N. Jean, S. M. Xie, and S. Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *Advances in Neural Information Processing Systems*, pages 5322–5333, 2018.

[10] R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[11] M. MacDorman and J. Atkinson. Infant mortality statistics from the linked birth/infant death data set-1995 period data: monthly vital statistics report. *Hyattsville (MD): National Center for Health Statistics*, 1998.

[12] D. A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[13] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[14] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[15] J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

[16] Y. Shimoni, C. Yanover, E. Karavani, and Y. Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

[17] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[18] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[19] J. Yoon, J. Jordon, and M. van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. 2018.

[20] J. Zhu, N. Chen, and E. P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.