Appendix A: Proofs

Theorem 1 (Counterfactual Generalization Bound)**.** *Let ϕ be an encoder $X \to Z$, and let f be an outcome function $Z \times T \to Y$. Under Assumptions 3 and 4, we have:*

$$\epsilon_{\text{cf}} \leq \epsilon_{\text{f}} + 2C * H(p_\phi(z,t) \,\|\, p_\phi(z)p(t)). \tag{1}$$

Proof. Before proceeding with the proof, we first introduce the following lemma.

Lemma 1. *[7, 4][Hierarchy with Hellinger] For probability measures P, Q with densities p, q, let*

$$\text{TV}(P,Q) = \tfrac{1}{2}\int |p-q|, \qquad H(P,Q) = \left[\int \left(\sqrt{p}-\sqrt{q}\right)^2 dx\right]^{1/2}$$

Then

$$\boxed{\text{TV}(P,Q) \;\leq H(P,Q) \leq \; \sqrt{\text{KL}(P\|Q)}}. \tag{2}$$

Let L denote the loss and ϵ_{cf} ϵ_{f} the counterfactual and factual losses, respectively. Following the techniques of [5] and [2], we write:

$$\begin{aligned}
\varepsilon^{\ell}_{cf} - \varepsilon^{\ell}_{f} &= \int_{\mathcal{T}}\int_{\mathcal{X}} \ell_{L,h,\phi}(x,t)\,[p(x)p(t) - p(x,t)]\,dxdt && (3)\\
&= \int_{\mathcal{T}}\int_{\mathcal{Z}} \ell_{L,h,\phi}(\psi(z),t)\,[p(\psi(z))p(t) - p(\psi(z),t)]\,J_\psi J_\psi^{-1}\,dzdt && (4)\\
&= \int_{\mathcal{T}}\int_{\mathcal{Z}} \ell_{L,h,\phi}(\psi(z),t)\,[p_\phi(z)p(t) - p_\phi(z,t)]\,dzdt && (5)\\
&\leq \int_{\mathcal{T}}\int_{\mathcal{Z}} C\,|p_\phi(z)p(t) - p_\phi(z,t)|\,dzdt && (6)\\
&\leq 2C * H(p_\phi(z,t) \,\|\, p_\phi(z)p(t)) && (7)\\
&\leq 2C * \sqrt{D_{\text{KL}}(p_\phi(z,t) \,\|\, p_\phi(z)p(t))}, && (8)
\end{aligned}$$

where the equality (5) holds by the reparameterization $x = \psi(z)$, inequality (6) holds by *Assumption 4* constraining the function ℓ, and the last two inequalities is by Lemma 1, $\int |p-q| = 2TV_D(p,q) \leq 2H(p,q) \leq 2\sqrt{KL(p,q)}$.

□

Proposition 1 (PEHE Error)**.** *Given an encoder ϕ and outcome prediction function f and a unit-loss function $\ell_{L,f,\phi}(\mathbf{x},\mathbf{t})$ that satisfies Assumption 4 and its associated L is squared error $\|\cdot\|^2$,*

$$\begin{aligned}
&\varepsilon_{pehe}(t_1,t_2) \leq \; \varepsilon^{\ell}_{\text{f}}(t_1) + \varepsilon^{\ell}_{\text{f}}(t_2)\\
&+ 2C * \left(H\left(p_\phi(\mathbf{z})\|p_\phi(\mathbf{z}|t_1)\right) + H\left(p_\phi(\mathbf{z})\|p_\phi(\mathbf{z}|t_2)\right)\right).
\end{aligned} \tag{9}$$

Proof.

$$\begin{aligned}
\varepsilon_{\text{pehe}}(t_1,t_2) &= \int_{\mathcal{X}} \Big[(\mu(x,t_1) - \mu(x,t_2)) - (h(\phi(x),t_1) - h(\phi(x),t_2))\Big]^2 p(x)\,dx && (10)\\
&\leq \int_{\mathcal{X}} (\mu(x,t_1) - h(\phi(x),t_1))^2 p(x)\,dx + \int_{\mathcal{X}} (\mu(x,t_2) - h(\phi(x),t_2))^2 p(x)\,dx && (11)\\
&= \varepsilon^{\ell}_{cf}(t_1) + \varepsilon^{\ell}_{cf}(t_2) && (12)\\
&\leq \varepsilon^{\ell}_{f}(t_1) + \varepsilon^{\ell}_{f}(t_2) + 2C[H\left(p_\phi(\mathbf{z})\|p_\phi(\mathbf{z}|t_1)\right) + H\left(p_\phi(\mathbf{z})\|p_\phi(\mathbf{z}|t_2)\right)] && (13)
\end{aligned}$$

where inequality (11) follows from the triangle inequality, and the last two lines hold by the definition of the counterfactual error and Theorem 1.

□

Appendix B: Algorithm

Algorithm 1 IBEX - Training Algorithm

Require: Dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{t}_i, y_i)\}_{i=1}^{N}$; epochs T; batch size B; predictor/encoder LR η_P; critic LR η_D;
penalty weight β; critic steps K; #shuffles S; score clip δ; tiny ε
Ensure: Trained parameters $(\theta_\phi, \theta_\tau, \theta_f, \psi)$
Initialize $(\theta_\phi, \theta_\tau, \theta_f, \psi)$ (e.g. AdamW)
for epoch $= 1$ to T **do**
  Shuffle $\mathcal{D}$ and split into $\lceil N/B \rceil$ mini-batches
  for all mini-batch $\mathcal{B} = \{(\mathbf{x}_j, \mathbf{t}_j, y_j)\}_{j=1}^{B}$ **do**
    Encode: $\mathbf{z}_j \leftarrow \phi_{\theta_\phi}(\mathbf{x}_j), \quad \tilde{\mathbf{t}}_j \leftarrow \tau_{\theta_\tau}(\mathbf{t}_j)$
    Predict: $\hat{y}_j \leftarrow f_{\theta_f}(\mathbf{z}_j, \tilde{\mathbf{t}}_j), \quad \mathcal{L}_{\text{MSE}} \leftarrow \frac{1}{B}\sum_{j=1}^{B}(y_j - \hat{y}_j)^2$
  (Inner: critic K steps, maximize J; detach z, t)
    for $k = 1$ to K **do**
      draw independent permutations $\{\pi_s\}_{s=1}^{S}$ on $\{1, \ldots, B\}$

$$J \leftarrow 2 - \frac{1}{S}\sum_{s=1}^{S}\left[\underbrace{\frac{1}{B}\sum_{j=1}^{B} D_\psi(\text{stop\_grad}(\mathbf{z}_j), \text{stop\_grad}(\tilde{\mathbf{t}}_j))}_{\widehat{\mathbb{E}}_{P_\phi}[D]} + \underbrace{\frac{1}{B}\sum_{j=1}^{B}\frac{1}{D_\psi(\text{stop\_grad}(\mathbf{z}_j), \text{stop\_grad}(\tilde{\mathbf{t}}_{\pi_s(j)}))}}_{\widehat{\mathbb{E}}_{Q_\phi}[D^{-1}]}\right]$$

      $\psi \leftarrow \psi + \eta_D \nabla_\psi J$ ▷ gradient *ascent*
    end for
  (Outer: one predictor/encoder step; critic frozen)
    draw fresh $\{\pi_s\}_{s=1}^{S}$ and compute J_{outer} as above (no detach)
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{MSE}} + \beta\,(-J_{\text{outer}})$
    $(\theta_\phi, \theta_\tau, \theta_f) \leftarrow (\theta_\phi, \theta_\tau, \theta_f) - \eta_P \nabla\mathcal{L}$
  end for
end for
return $\theta_\phi, \theta_\tau, \theta_f$

Appendix C: Dataset Descriptions

The News dataset [1] comprises bag-of-words representations of randomly sampled *New York Times* articles. It contains a 3,477-dimensional word count vector. Following prior work [3], we use a subset of 5,000 samples and synthetically generate continuous treatment and outcome variables to evaluate model performance. Moreover, following Schwab et al. [6] and Bica, Jordon, and van der Schaar [3], we train an initial topics model q over the covariates $q(\mathbf{x})$. The dataset is available at: `https://archive.ics.uci.edu/dataset/164/bag+of+words`. We use the same treatment mechanism as in Kazemi and Ester [5].

The TCGA (The Cancer Genome Atlas Program) [8] dataset contains gene expression data for cancer patients. We selected a total of 9000 patients and the 4000 most variable genes. The gene expression values were and scaled to fall within the $[0, 1]$ range. Additionally, features were normalized to have unit norm for each patient. The outcome can be understood as how likely cancer recurrence is. The dataset version used is the same as the one employed in DRNet, available at: `https://github.com/d909b/drnet`.

References

[1] Asuncion, A.; Newman, D.; et al. 2007. UCI machine learning repository.

[2] Bellot, A.; Dhir, A.; and Prando, G. 2022. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage. *arXiv preprint arXiv:2205.14692*.

[3] Bica, I.; Jordon, J.; and van der Schaar, M. 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33: 16434–16445.

[4] Cover, T. M.; and Thomas, J. A. 2006. *Elements of Information Theory*. John Wiley & Sons.

[5] Kazemi, A.; and Ester, M. 2024. Adversarially balanced representation for continuous treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13085–13093.

[6] Schwab, P.; Linhardt, L.; Bauer, S.; Buhmann, J. M.; and Karlen, W. 2020. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5612–5619.

[7] Wasserman, L. 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[8] Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10): 1113–1120.