Continuous Treatment Effect Estimation with Cauchy-Schwarz Divergence Information Bottleneck

Anonymous authors
Paper under double-blind review

Abstract

Estimating conditional average treatment effects (CATE) for continuous and multivariate treatments remains a fundamental yet underexplored problem in causal inference, as most existing methods are confined to binary treatment settings. In this paper, we make two key theoretical contributions. First, we derive a novel counterfactual error bound based on the Cauchy–Schwarz (CS) divergence, which is provably tighter than prior bounds derived from the Kullback–Leibler (KL) divergence. Second, we strengthen this bound by integrating the Information Bottleneck principle, introducing a compression regularization on latent representations to enhance generalization. Building on these insights, we propose a new neural framework that operationalizes our theory. Extensive experiments on three benchmarks show that our method consistently outperforms state-of-the-art baselines and remains robust under biased treatment assignments.

1 Introduction

Estimating individual causal effects from observational data is inherently difficult because counterfactual outcomes are never observed, making direct validation impossible (Imbens & Rubin, 2015). As such, it has been recognized that robust causal inference methods must exhibit strong theoretical properties to ensure that treatment-effect estimates are well-bounded (Shalit et al., 2017).

In the binary treatment setting (where a certain treatment is either prescribed or not prescribed), the main objective of much of the theoretical work is to achieve group rebalancing. Group rebalancing is a procedure which essentially aimed to counter any bias that may occur in observational data due to, for instance, treatment effect bias (e.g., certain groups of patients may receive more treatment T than others). As such, reducing the effect of confounding is a key step toward obtaining more reliable estimates from observational data. (Holland, 1986; Rubin, 2005).

While traditional causal inference methods such as inverse propensity score weighting (IPW) Austin & Stuart (2015) to directly re-weight the impact of treatment of the outcome, deep causal machine learning approaches, such as counterfactual regression methods (e.g., Shalit et al. (2017)), achieve bias reduction by learning a shared representation that balances the treatment groups in representation space. Under certain assumptions, it then becomes possible to derive generalization bounds on the conditional average treatment effect (CATE) estimates, limiting potential error and improving performance (Bellot et al., 2022; Shalit et al., 2017). The key idea behind these methods is that, by using distributional measures such as integral probability metrics (IPM), one can quantify the distributional shift between the treated and control groups.

Extending counterfactual regression with distributional methods paradigm to the continuous treatment setting, where treatments represent real-valued dosages or multivariate exposures, introduces additional complexity. First, CATEs are now functions over a continuum of treatments, not binary, discrete units. Second, confounding adjustment requires estimating generalized propensity *densities* rather than scores (Imbens, 2000). Finally, many existing architectures, such as DRNet (Schwab et al., 2020), scale poorly, requiring a separate output head per treatment stratum.

Recent work addresses some of these challenges using adversarial approaches (Bica et al., 2020; Kazemi & Ester, 2024), extending generative adversarial networks (GANs) for binary treatment effect prediction to the continuous setting. However, these approaches remain adaptations of binary-treatment methods, and they often struggle with instability, sensitivity to hyperparameters, or poor scalability to high-dimensional treatment spaces.

In contrast, *information bottleneck* (IB) methods offer a theoretically grounded alternative. IB aims to learn representations Z of covariates X that are maximally predictive of outcomes Y while discarding irrelevant information in X (Tishby et al., 2000). When applied to causal inference, IB has been shown to reduce confounding in binary settings (Parbhoo et al., 2020; Lu et al., 2022). Yet, despite their promising properties, there is a lack of strong theoretical analysis on how IB methods improve counterfactual generalization. Moreover, IB has never been applied to continuous or multivariate treatment effect estimation.

To explore the potential of IB in more complex treatment settings, we propose ***I**nformation **B**ottleneck for **E**stimating continuous e**X**posures* (**IBEX**), a novel framework for CATE estimation with continuous and multivariate treatments. IBEX minimizes the statistical dependence between the learned representation and the treatment variable using a tractable approximation of mutual information based on the Cauchy-Schwarz (CS) divergence (Yu et al., 2024). To further encourage invariance and generalization, we apply a dimensionality bottleneck to the latent space.

Contributions. ① We derive novel *counterfactual generalization bounds* for continuous treatments using the CS divergence, and show that these bounds are *tighter* than those based on the Kullback–Leibler divergence under mild assumptions. ② We design a modular architecture with separate covariate and treatment encoders, incorporating dimensionality regularization to control representation capacity. ③ We empirically validate IBEX across three benchmarks (MIMIC-IV, TCGA, News), showing *state-of-the-art* performance in terms of dose-response estimation and policy regret, and robustness under strong treatment-assignment bias.

2 Related Work

In this section, we provide a brief overview of relevant prior work on treatment effect estimation, with a particular focus on continuous treatments and representation learning approaches.

2.1 Information Bottleneck and its Application to Causal Inference

The Information Bottleneck (IB) principle, introduced by (Tishby et al., 2000), formulates representation learning as a trade-off between extracting information from the input variable $\mathbf{x}$ that is relevant for predicting the target variable y, and discarding nuisance factors in $\mathbf{x}$ that are irrelevant to y. Formally, the objective of IB is to learn a compressed representation $\mathbf{z}$ by minimizing the mutual information $I(\mathbf{x};\mathbf{z})$, ensuring the minimality and compactness of $\mathbf{z}$, while simultaneously maximizing $I(\mathbf{z};\mathbf{y})$, thereby preserving the predictive information to y. The optimization objective can thus be written as:

$$\min \quad I(\mathbf{x};\mathbf{z}) \; - \; \beta I(\mathbf{z};y), \tag{1}$$

where $\beta > 0$ controls the trade-off between compression and prediction. It has been theoretically shown that Z naturally constitutes the minimal sufficient representation (Gilad-Bachrach et al., 2003).

To make this objective tractable for high-dimensional data and deep learning models, variational approximations such as the variational information bottleneck (VIB) (Alemi et al., 2016) and the nonlinear information bottleneck (NIB) (Kolchinsky et al., 2019) have been proposed.

Recent works have applied the IB principle in discrete treatment settings (Kim et al., 2019; Parbhoo et al., 2020; Lu et al., 2022). In principle, these approaches share a common high-level idea: leveraging the IB principle to compress high-dimensional covariates $\mathbf{x}$ into a low-dimensional representation $\mathbf{z}$ that retains information relevant for treatment effects $\{y, t\}$. They typically formulate an IB objective of the form:

$$\min I(\mathbf{x};\mathbf{z}) - \beta I(\mathbf{z};y,t), \tag{2}$$

with some methodological differences. For instance, the causal effect information bottleneck (CEIB) (Parbhoo et al., 2020) learns discrete latent representations separately from $\mathbf{x}_0$ (covariates of untreated patients) and $\mathbf{x}_1$ (covariates of treated patients). In contrast, causal information bottleneck (CIB) (Kim et al., 2019) uses two separate heads to estimate $I(\mathbf{z}; y_0)$ and $I(\mathbf{z}; y_1)$, where y_0 and y_1 denote the control and treatment outcomes, respectively.

However, these methods are limited to binary treatments and rely heavily on variational approximations, which are known to suffer from loose bounds and biased estimates of mutual information.

In this paper, we extend the IB framework to continuous treatment effect estimation by introducing a novel IB objective that differs fundamentally from Eq. (2). What is more, our implementation avoids variational approximation entirely, thereby mitigating the issues of bound looseness and biased information estimation. Moreover, we provide a formal analysis of the generalization error bound, which, to the best of our knowledge, has not been addressed in prior IB-based works.

2.2 Continuous Treatment Effect Estimation

Recent work has focused on extending methods originally developed for binary treatment estimation to the more realistic setting of continuous treatments, where dosages or combinations of dosages are considered. Bellot et al. (2022) derive generalization bounds for continuous treatment effect estimation and make use of a HSIC-type regularization, extending prior literature. Tanimoto et al. (2021) propose a regret-minimization approach for handling large action spaces. Schweisthal et al. (2023) develop a conformal prediction framework for estimating generalized propensity scores. Schwab et al. (2020) introduce DRNet, a representation learning method inspired by counterfactual regression approaches such as Shalit et al. (2017). Another notable line of work using conformal prediction for continuous treatments is presented by Schröder et al. (2024). Bica et al. (2020) (SCIGAN) and Kazemi & Ester (2024) (ACFR) adopt adversarial approaches to address the intractability of posteriors and use a Kullback–Leibler regularizer to correct for distributional shift. In contrast, we aim to improve robustness by leveraging the more stable Cauchy–Schwarz divergence.

3 Theoretical Preliminaries

The objective of our approach is to estimate individualized treatment effects (ITE) in settings with multivariate, continuous treatments, such as personalized dosage recommendations in healthcare. In this section, we introduce the relevant formalizations.

Terminology. Let $\mathcal{D}_\mathrm{f}$ be the factual dataset that contains i.i.d. samples $(\mathbf{x}^i, \mathbf{t}_\mathrm{f}^i, y_\mathrm{f}^i)$ drawn from distribution $p_{\mathbf{X},\mathbf{T}_\mathrm{f},y_\mathrm{f}}$. Let $\mathbf{X}$ denote a covariate vector taking values $\mathbf{x} \in \mathcal{X}$ (e.g. age, weight, lab results), and $\mathbf{x}$ represents a realization of $\mathbf{X}$. The treatment variable is in the form $\mathbf{T}_\mathrm{f} = (W_\mathrm{f}, D_\mathrm{f}) \in \mathcal{T}$, where the discrete component $W_\mathrm{f} \in \mathcal{W} = \{w_1, \ldots, w_k\}$ denotes the treatment type (e.g. specific combination of medications) and $D_\mathrm{f} \in \mathcal{D}_{W_\mathrm{f}}$ denotes the associated dosage (e.g. a number in $[0, 1]$ indicating the amount of medication provided). We denote the factual outcome as $Y_\mathrm{f} = Y(\mathbf{T}_\mathrm{f})$ and the counterfactual (i.e. unobserved) outcome as Y_cf.

While there is only one pair of counterfactual treatment and outcome under the binary treatment setting, there are infinitely many of them in the case of continuous treatment. Therefore, we define the individual dose-response function.

Definition 1. *For any covariate vector $\mathbf{x} \in \mathcal{X}$, we define the dose–response function as*

$$\mu(\mathbf{t}, \mathbf{x}) := \mathbb{E}[Y(\mathbf{t})|\mathbf{X} = \mathbf{x}], \qquad \forall\, \mathbf{t} \in \mathcal{T}. \tag{3}$$

Definition 2. *The generalized propensity score (Imbens, 2000) is given by the conditional density*

$$e(\mathbf{x}) := p_{\mathbf{T}_\mathrm{f}|\mathbf{X}}(\mathbf{t}|\mathbf{x}), \qquad \forall\, \mathbf{x} \in \mathcal{X}, \tag{4}$$

where $\mathbf{t}$ may contain a continuous component.

The generalized propensity score generalizes the conventional propensity score to account for continuous treatment.

Definition 3. *We define the treatment effect for a treatment-effect pair $t_1, t_2 \in \mathcal{T}$ as*

$$\tau_{t_1,t_2}(\mathbf{x}) := \mu(\mathbf{x}, t_1) - \mu(\mathbf{x}, t_2), \qquad \forall \mathbf{x} \in \mathcal{X}. \tag{5}$$

Equation equation 5 measures the relative treatment effect between different medications administered to the same subject.

Assumption 1 (Ignorability Assumption)**.** *We assume that the potential outcome is independent from the treatment given the sufficient adjustment set $\mathbf{X}$, i.e. $\mathbf{X}$ blocks all non-causal paths between treatment and outcome,*

$$\{Y(w,d)\}_{(w,d)\in\mathcal{T}} \perp\!\!\!\perp \mathbf{T}_\mathrm{f} \mid \mathbf{X}. \tag{6}$$

Assumption 2 (Overlap Assumption)**.** *We assume that the conditional treatment distribution admits a density $p_{\mathbf{T}_\mathrm{f}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x})$ with respect to the Lebesgue measure, and that this density is strictly positive almost everywhere on $\mathcal{X} \times \mathcal{T}$, i.e.,*

$$p_{\mathbf{T}_\mathrm{f}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x}) > 0 \quad \textit{for almost every } (\mathbf{x}, \mathbf{t}) \in \mathcal{X} \times \mathcal{T}. \tag{7}$$

Going forward, we introduce a stochastic encoder q_ϕ which compresses the covariate space $\mathcal{X}$ into a low-dimensional latent space $\mathcal{Z}$ and a predictor model $f : \mathcal{Z} \times \mathcal{T} \to \mathcal{Y}$. Additionally, let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ be a loss function.

Definition 4. *Define the unit loss $\ell_{L,f,\phi} : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^+$ as*

$$\ell_{L,f,\phi}(\mathbf{x}, \mathbf{t}) = L(f(\phi(\mathbf{x}), \mathbf{t}), y). \tag{8}$$

Unit loss $\ell_{L,f,\phi}(\mathbf{x}, \mathbf{t})$ measures the loss between the predicted outcome $\hat{y} = f(\phi(\mathbf{x}), \mathbf{t})$ and the ground-truth outcome $y = \mu(\mathbf{x}, \mathbf{t})$.

Definition 5. *We define the factual and counterfactual errors at treatment $t \in \mathcal{T}$ respectively as*

$$\epsilon_\mathrm{f}^\ell(\mathbf{t}) := \int_\mathcal{X} \ell_{L,f,\phi}(\mathbf{x}, \mathbf{t})\, p(\mathbf{x}|\mathbf{t})\, d\mathbf{x}, \tag{9}$$

$$\begin{aligned} \epsilon_\mathrm{cf}^\ell(\mathbf{t}) &:= \int_{\mathcal{T}'=[0,1]\setminus\{\mathbf{t}\}} \int_\mathcal{X} \ell_{L,f,\phi}(\mathbf{x}, \mathbf{t})\, p(\mathbf{x}|\mathbf{t}')\, d\mathbf{x}\, d\mathbf{t}' \\ &= \int_\mathcal{X} \ell_{L,f,\phi}(\mathbf{x}, \mathbf{t})\, p(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{10}$$

Essentially, factual error $\epsilon_\mathrm{f}^\ell(\mathbf{t})$ is obtained by marginalizing over $p(\mathbf{x}|\mathbf{t})$ while the counterfactual error $\epsilon_\mathrm{cf}^\ell(\mathbf{t})$ is obtained by marginalizing over $p(\mathbf{x})$. Furthermore, we define $\epsilon_\mathrm{f} = \int_\mathcal{T} \epsilon_\mathrm{f}^\ell(\mathbf{t}) p(\mathbf{t}) d\mathbf{t}$ and $\epsilon_\mathrm{cf} = \int_\mathcal{T} \epsilon_\mathrm{cf}^\ell(\mathbf{t}) p(\mathbf{t}) d\mathbf{t}$.

Definition 6 (Cauchy-Schwarz Divergence (Principe et al., 2000; Jenssen et al., 2006))**.** *Let $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ be probability measures on a Borel subset $\mathcal{X} \in \mathbb{R}^d$. Assume μ and ν are absolutely continuous with respect to the Lebesgue measure* Leb*, and denote their density functions by $p = d\mu/d\mathrm{Leb}$ and $q = d\nu/d\mathrm{Leb}$. If $p, q \in L^2(\mathrm{Leb})$, then Cauchy-Schwarz inequality gives*

$$\Big(\int_\mathcal{X} p(\mathbf{x})q(\mathbf{x})\, d\mathbf{x}\Big)^2 \le \Big(\int_\mathcal{X} p^2(\mathbf{x})\, d\mathbf{x}\Big)\Big(\int_\mathcal{X} q^2(\mathbf{x})\, d\mathbf{x}\Big), \tag{11}$$

with equality holding if and only if p and q are colinear, almost everywhere on $\mathcal{X}$.

The CS divergence defines the distance between p and q by measuring the tightness (or gap) of the two sides of Eq. equation 11 using the logarithm of their ratio:

$$D_{CS}(p\|q) = -\log\left(\frac{\left(\int p(\mathbf{x})q(\mathbf{x})\, d\mathbf{x}\right)^2}{\int p(\mathbf{x})^2\, d\mathbf{x} \cdot \int q(\mathbf{x})^2\, d\mathbf{x}}\right). \tag{12}$$

The CS divergence possesses several appealing properties compared to the KL divergence, which is defined as

$$D_{\mathrm{KL}}(p\|q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}. \tag{13}$$

In particular, the CS divergence is symmetric and admits closed-form expressions for mixture-of-Gaussians (MoG) distributions (Tran et al., 2022). In the following subsection, we further demonstrate how the use of CS divergence leads to a tighter generalization error bound compared to its KL divergence counterpart.

4 Continuous Treatment Generalization Bounds

We begin by analyzing generalization in the continuous, multivariate treatment setting. Under standard regularity conditions, we derive a counterfactual error bound in terms of the CS divergence between the learned representation and the treatment variable. Compared with recent approaches (Bellot et al., 2022; Kazemi & Ester, 2024), the resulting bound is expressed directly in representation space and admits a tighter dependence characterization under multivariate treatments.

Beyond dependence control, we further investigate the role of representation capacity in generalization. In particular, we show that IB–inspired regularization provides an upper bound on the factual generalization error through capacity control.

Taken together, this section provides theoretical insights into two complementary aspects of counterfactual learning: (i) representation–treatment dependence, which governs counterfactual robustness; and (ii) representation capacity, which governs generalization from empirical to population risk. These two perspectives are analyzed separately and later combined to motivate the IBEX objective.

4.1 Bounding the Counterfactual Error via CS Divergence–Induced Regularization

Before presenting our main theoretical result, we first introduce Assumptions 3 and 4, following the framework of Kazemi & Ester (2024).

Assumption 3. *The encoder function $\phi : \mathcal{X} \to \mathcal{Z}$ is a twice-differentiable bijection. The representation space $\mathcal{Z}$ is the image of $\mathcal{X}$ under ϕ with the induced distribution $p_\phi(\mathbf{z})$.*

Assumption 3 is introduced as a technical condition to facilitate the density transformation and change-of-variable arguments required for deriving the counterfactual generalization bound in this section. In particular, the twice-differentiable bijection ensures that the induced distribution $p_\phi(\mathbf{z})$ is well-defined and allows tractable analysis of divergence terms Kazemi & Ester (2024).

Importantly, this assumption is not imposed in the practical model. When introducing the IB-inspired capacity-control regularization, we explicitly relax the bijectivity requirement and allow non-invertible encoders. In this relaxed regime, the IB regularization term $I(\mathbf{x}; \mathbf{z})$ is no longer constant and becomes finite and controllable Kawaguchi et al. (2023), making it a meaningful measure of representational capacity.

Assumption 4. *Let G be a class of functions with infinity norm less than 1, $G = \{g : \mathcal{Z} \times \mathcal{T} \to \mathbb{R}^+ \mid \|g\|_\infty \leq 1\}$. Then, there exists a constant $C > 0$ such that*

$$\frac{\ell_{L,g,\phi}(\mathbf{x}, \mathbf{t})}{C} \in G.$$

This means for any $(\mathbf{x}, \mathbf{t})$ we have

$$\frac{\ell_{L,g,\phi}(\mathbf{x}, \mathbf{t})}{C} \leq 1.$$

Theorem 1 (Counterfactual Generalization Bound, Gaussian Scenario)**.** *Let ϕ be an encoder $\mathcal{X} \to \mathcal{Z}$, and let f be an outcome function $\mathcal{Z} \times \mathcal{T} \to \mathcal{Y}$. Assume that the joint distribution $p(z, t)$ follows a multivariate Gaussian distribution:*

$$p(\mathbf{z}, \mathbf{t}) \sim \mathcal{N}\left(\begin{bmatrix}\mu_z \\ \mu_t\end{bmatrix}, \Sigma_1\right), \quad \textit{where} \quad \Sigma_1 = \begin{bmatrix}\Sigma_z & \Sigma_{z,t} \\ \Sigma_{z,t}^T & \Sigma_t\end{bmatrix}.$$

Let Σ_2 denote the covariance matrix of the product of marginals $p(\mathbf{z})p(\mathbf{t})$, i.e., the case where $\mathbf{z} \perp \mathbf{t}$. Then,

$$\Sigma_2 = \begin{bmatrix} \Sigma_z & 0 \\ 0 & \Sigma_t \end{bmatrix}.$$

Under Assumptions 3 and 4, we have:

$$\epsilon_{\text{cf}} \leq \epsilon_{\text{f}} + C\sqrt{2D_{CS}(p_\phi(\mathbf{z},\mathbf{t}) \,\|\, p_\phi(\mathbf{z})p(\mathbf{t}))}, \tag{14}$$

if

$$\sum_{i=1}^{d} \log\left(\frac{2+\lambda_i+1/\lambda_i}{4}\right) \geq 4,$$

where λ_i is the i-th eigenvalue of

$$\Sigma_2^{-1}\Sigma_1 = \begin{bmatrix} I & \Sigma_z^{-1}\Sigma_{z,t} \\ \Sigma_t^{-1}\Sigma_{z,t}^T & I \end{bmatrix}.$$

Proof. All the proofs can be found in Appendix B. □

The conditions in Theorem 1 are easy to satisfy. In our study, the joint dimension is $d = d_t + d_z = 130$, with treatment dimension $d_t = 2$ and latent dimension $d_z = 128$. Each term $\log\left(\frac{2+\lambda_i+1/\lambda_i}{4}\right)$ is non-negative, since $\lambda + 1/\lambda \geq 2$ for $\lambda \in (0, 1]$. Thus, when d is large, the total sum easily exceeds the threshold. Even if most λ_i values are close to 1 (indicating weak correlation between z and t), a small number of moderately deviating values (e.g., $\lambda_i \leq 0.7$) are sufficient to push the sum above the bound.

In fact, Theorem 1 can be extended to general joint distribution $p(z, t)$ without assuming Gaussianity.

Proposition 1. *Let ϕ be an encoder mapping $\mathcal{X} \to \mathcal{Z}$, and let f be an outcome function $\mathcal{Z} \times \mathcal{T} \to \mathcal{Y}$. Assume that $p(\mathbf{z}, \mathbf{t})$ is an arbitrary joint distribution. Then, we have*

$$\epsilon_{\text{cf}} \lesssim \epsilon_{\text{f}} + C\sqrt{2D_{CS}(p_\phi(\mathbf{z},\mathbf{t}) \,\|\, p_\phi(\mathbf{z})p(\mathbf{t}))}, \tag{15}$$

where $\lesssim$ denotes "less than or approximately equal to," and the precise conditions under which this inequality holds are discussed in Appendix B.

Theorem 1 and Proposition 1 imply that reducing counterfactual error requires not only minimizing the factual error (which is intuitive) but also encouraging independence between $\mathbf{z}$ and $\mathbf{t}$, since $p(\mathbf{z}, \mathbf{t}) = p(\mathbf{z})p(\mathbf{t})$ if and only if $\mathbf{z} \perp \mathbf{t}$.

Remark 1 (Tighter Bound)**.** *A similar counterfactual error bound based on KL divergence is presented in (Bellot et al., 2022; Kazemi & Ester, 2024):*

$$\epsilon_{\text{cf}} \leq \epsilon_{\text{f}} + C\sqrt{2D_{KL}(p_\phi(\mathbf{z},\mathbf{t}) \| p_\phi(\mathbf{z})p(\mathbf{t}))}. \tag{16}$$

In general, there is no universal ordering between D_{CS} and D_{KL} for arbitrary distributions. However, under mild regularity conditions, KL divergence admits a lower bound in terms of CS divergence. Specifically, let $\mathbf{u} = [\mathbf{z}; \mathbf{t}] \in \mathbb{R}^{d_\mathbf{z}+d_\mathbf{t}}$ and let p, q be density functions defined on a bounded integration domain $K \subset \mathbb{R}^{d_\mathbf{z}+d_\mathbf{t}}$ with $|K| < \infty$. Assume p, q are Riemann integrable on K, then (Yin et al., 2024):

$$C_1\Big(D_{CS}(p;q) - \log|K| + 2\log C_2\Big) \leq D_{KL}(p;q), \tag{17}$$

where $C_1 := \int_K p(\mathbf{u})\,\mathrm{d}\mathbf{u} > 0$ and $C_2 := C_1\left(\int_K p^2(\mathbf{u})\,\mathrm{d}\mathbf{u} \int_K q^2(\mathbf{u})\,\mathrm{d}\mathbf{u}\right)^{-1/4}$.

Equivalently,

$$D_{CS}(p;q) \leq \frac{1}{C_1} D_{KL}(p;q) + \log|K| - 2\log C_2. \tag{18}$$

Substituting this relation into Eq. (15) yields:

$$\begin{aligned}\epsilon_{\text{cf}} &\lesssim \epsilon_{\text{f}} + C\sqrt{2D_{CS}(p_\phi(\mathbf{z},\mathbf{t}) \,\|\, p_\phi(\mathbf{z})p(\mathbf{t}))} \\ &\leq \epsilon_{\text{f}} + C\sqrt{2\left(\frac{1}{C_1}D_{KL}(p_\phi(\mathbf{z},\mathbf{t})\|p_\phi(\mathbf{z})p(\mathbf{t})) + \log|K| - 2\log C_2\right)}.\end{aligned} \tag{19}$$

This establishes a theoretical comparison between the CS-based and KL-based formulations in the bounded continuous setting considered here. The constants C_1 and C_2 are distribution-dependent but finite under the stated assumptions.

It is important to note that a tighter bound does not necessarily imply improved empirical performance; rather, it guarantees only that, under the assumptions of the theory, the Cauchy-Schwarz divergence yields a tighter theoretical upper bound on counterfactual error than the KL divergence.

We can further provide a bound in terms of the precision estimation of heterogeneous effects (PEHE), a metric commonly used in causal inference to measure the treatment-effect error (Hassanpour & Greiner, 2019; Shalit et al., 2017).

Definition 7. *We define the expected precision of estimating heterogeneous effect (PEHE) between treatment pairs $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{T}$ as*

$$\begin{aligned}\varepsilon_{pehe}(\mathbf{t}_1,\mathbf{t}_2) := \int_{\mathcal{X}} \Big[&(\mu(\mathbf{x},\mathbf{t}_1) - \mu(\mathbf{x},\mathbf{t}_2)) \\ &- (f(\phi(\mathbf{x}),\mathbf{t}_1) - f(\phi(\mathbf{x}),\mathbf{t}_2))\Big]^2 p(\mathbf{x})\, d\mathbf{x}.\end{aligned} \tag{20}$$

Following Proposition 1, we can then easily derive the following bound.

Proposition 2 (PEHE Error Bound)**.** *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 $\|\cdot\|^2$, the following inequality holds:*

$$\begin{aligned}\varepsilon_{pehe}(\mathbf{t}_1,\mathbf{t}_2) \leq\ & \varepsilon_{\text{f}}^{\ell}(\mathbf{t}_1) + \varepsilon_{\text{f}}^{\ell}(\mathbf{t}_2) \\ &+ \sqrt{2D_{CS}\left(p_\phi(\mathbf{z})\|p_\phi(\mathbf{z}|\mathbf{t}_1)\right)} + \sqrt{2D_{CS}\left(p_\phi(\mathbf{z})\|p_\phi(\mathbf{z}|\mathbf{t}_2)\right)}.\end{aligned} \tag{21}$$

The bound effectively states that for any pair of treatments $\mathbf{t}_1, \mathbf{t}_2$, minimizing the divergence between the marginal $p_\phi(\mathbf{z})$ and the conditionals $p_\phi(\mathbf{z}|\mathbf{t}_i)$ reduces the dependence of the learned representation $\mathbf{z} = \phi(\mathbf{x})$ on the treatment assignment. This encourages the encoder ϕ to learn a *balanced representation*—that is, one where the distribution of $\mathbf{z}$ is approximately invariant across treatment groups:

$$p_\phi(\mathbf{z} \mid \mathbf{t}_1) \approx p_\phi(\mathbf{z} \mid \mathbf{t}_2) \approx p_\phi(\mathbf{z}),$$

which implies $\mathbf{z} \perp \mathbf{t}$, i.e., independence of the representation from the treatment assignment.

4.2 Bounding the Factual Error with IB Approach

We now analyze how representation complexity affects generalization of factual risk.

Theorem 2 (Information Bottleneck Bound on Factual Error (Kawaguchi et al., 2023))**.** *Let ϕ be a stochastic encoder mapping input $\mathbf{x}$ to a representation $\mathbf{z} = \phi(\mathbf{x})$, and let $\ell_{L,h,\phi}(\mathbf{x},t)$ be a loss function that is L-Lipschitz and bounded in $[0,1]$. Suppose the training data $\{(\mathbf{x}_i,t_i)\}_{i=1}^n$ are drawn i.i.d. from the joint distribution $p(\mathbf{x},t)$, where $t \in [0,1]$. Then, with probability at least $1-\eta$, the expected factual error satisfies:*

$$\epsilon_{\text{f}} \leq \hat{\epsilon}_{\text{f}} + B\sqrt{\frac{I(\mathbf{x};\mathbf{z})}{n}} + \frac{\delta}{\sqrt{n}},$$

where $\hat{\epsilon}_{\text{f}} = \frac{1}{n}\sum_{i=1}^n \ell_{L,h,\phi}(\mathbf{x}_i,t_i)$ is the empirical factual error, B is a constant depending on the Lipschitz constant of the loss function, δ is a vanishing term (e.g., $\mathcal{O}(\sqrt{\log(1/\eta)}/n^{1/4})$) as $n \to \infty$.

Remark 2 (Effect of IB on Counterfactual Error)**.** *Beyond its classical role in controlling the generalization gap between empirical and factual risk, IB-inspired regularization may also indirectly reduce the counterfactual error ϵ_{cf}. In practice, by applying a capacity-control regularization motivated by the Information Bottleneck principle, the learned representation $\mathbf{z} = \phi(\mathbf{x})$ is encouraged to discard task-irrelevant or treatment-specific features that may not generalize across treatment regimes. According to the data processing inequality (Cover & Thomas, 2006), we have*

$$I(\mathbf{z}; t) \leq I(\mathbf{x}; \mathbf{z}). \tag{22}$$

Since the counterfactual error bound includes a divergence term of the form (see Eq. (19))

$$\epsilon_{\text{cf}} \leq \epsilon_{\text{f}} + C\sqrt{2D_{\text{KL}}(p(\mathbf{z}, t) \| p(\mathbf{z})p(t))} = \epsilon_{\text{f}} + C\sqrt{2I(\mathbf{z}; t)},$$

reducing the dependence between the representation $\mathbf{z}$ and the treatment variable t tightens the gap between factual and counterfactual error. While Assumption 3 assumes a bijective encoder for theoretical tractability, in practical implementation this condition is relaxed. In the non-invertible regime, IB-inspired regularization effectively reduces $I(\mathbf{x}; \mathbf{z})$, which in turn constrains $I(\mathbf{z}; t)$ via data processing inequality, thereby promoting treatment-invariant representations and improving counterfactual robustness.

4.3 Complementary Roles of Dependence and Capacity

Theorem 1 and Theorem 2 address two distinct sources of error in counterfactual learning. First, Theorem 1 shows that counterfactual robustness is governed by the statistical dependence between the learned representation $\mathbf{z}$ and the treatment variable $\mathbf{t}$. In particular, reducing the divergence term $D_{\text{CS}}(p(\mathbf{z}, \mathbf{t}) \| p(\mathbf{z})p(\mathbf{t}))$ tightens the gap between factual and counterfactual error, thereby mitigating counterfactual shift.

Second, Theorem 2 establishes that the generalization of the factual risk depends on representation capacity. Controlling the mutual information $I(\mathbf{x}; \mathbf{z})$ limits representation complexity and improves the transition from empirical to population risk.

Although these two bounds are derived under different technical regimes: Theorem 1 assuming an invertible encoder for analytical tractability, and Theorem 2 allowing non-invertible capacity-limited encoders, they provide complementary design principles for practical model construction:

- Minimize representation-treatment dependence to reduce counterfactual shift;
- Control representation capacity to improve generalization.

In practice, our IBEX jointly optimizes a dependence penalty $D_{\text{CS}}(p(\mathbf{z}, \mathbf{t}) \| p(\mathbf{z})p(\mathbf{t}))$ motivated by the counterfactual bound, together with a capacity-control regularizer approximating $I(\mathbf{x}; \mathbf{z})$ motivated by the factual generalization bound. While we do not claim a single unified bound under non-invertible encoders, these independently established theoretical results jointly justify the structure of the IBEX objective.

5 Methodology

5.1 Structural overview

Before delving into the optimization objective, we first ground the reader in the structural causal model (SCM) that motivates IBEX.

Figure 1 contains the SCMs of the IBEX before and after intervention. In the conventional causal effect model, covariates $\mathbf{x}$ simultaneously influence the treatment t and the outcome y, while t also affects y. This layout conflates *all* information in $\mathbf{x}$ (both outcome–relevant and nuisance, treatment–specific factors), making generalisation difficult when the distribution of t shifts.

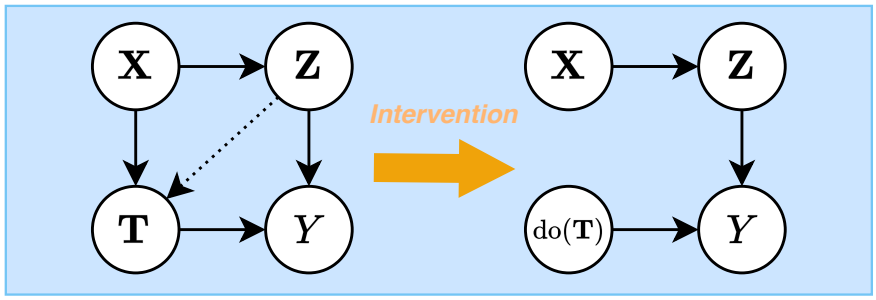

Figure 1: Structural causal models implementing the IBEX framework: pre-intervention scenario (left panel) and post-intervention outcome (right panel). The dotted arrow from $\mathbf{Z}$ to $\mathbf{T}$ indicates statistical dependence induced via the shared parent $\mathbf{X}$. Note that $\mathbf{Z}$ is a function of $\mathbf{X}$ via $\mathbf{Z} = f(\mathbf{X})$.

On the other hand, IBEX inserts a learned bottleneck variable $\mathbf{z} = f_\phi(\mathbf{x})$ between $\mathbf{x}$ and the rest of the system and explicitly regularizes two information pathways: ① We maximize $I((\mathbf{z}, \mathbf{t}); y)$ so that $\mathbf{z}$ keeps precisely the features of $\mathbf{x}$ needed, jointly with $\mathbf{t}$, to predict y. ② We minimize $I(\mathbf{x}; \mathbf{z})$ and drive $I(\mathbf{z}; \mathbf{t})$ toward zero via a CS divergence term, forcing the encoder to forget treatment-specific components that do not help predict y.

5.2 Objective function and Implementation

Guided by the complementary principles established in Section 4.3, IBEX is formulated as an information-theoretic objective that simultaneously promotes predictive sufficiency, limits representation capacity, and reduces representation–treatment dependence. The high-level objective is:

$$\max \quad I\big((\mathbf{z}, t); y\big) \; - \; \beta I(\mathbf{x}; \mathbf{z}) \; - \; \gamma I(\mathbf{z}; t), \tag{23}$$

where $\beta, \gamma > 0$ are trade-off hyperparameters.

The first term $I((\mathbf{z}, t); y)$ enforces *sufficiency*, ensuring that the representation retains information necessary for predicting the outcome jointly with the treatment. The second term $I(\mathbf{x}; \mathbf{z})$ controls representation capacity in the spirit of the IB principle. Reducing this mutual information encourages compression of $\mathbf{x}$ into a minimal predictive representation, thereby improving generalization as suggested by Theorem 2. The third term $I(\mathbf{z}; t)$ penalizes statistical dependence between the representation and treatment assignment. Minimizing this term mitigates treatment-specific confounding and reduces counterfactual shift, consistent with the dependence-based bound in Theorem 1.

In practice, the mutual information terms are approximated using tractable surrogates. Representation capacity is controlled either explicitly through dimensionality constraints on $\mathbf{z}$ (Tao et al., 2020) or via regularized mutual-information estimators. The dependence term $I(\mathbf{z}; t)$ is approximated using the CS divergence, which provides a stable and computationally efficient measure of representation–treatment dependence in multivariate settings.

5.2.1 Maximizing Expressiveness Term

We maximize $I(\mathbf{z}, \mathbf{t}); y)$, which states that the encoding-treatment pair needs to be expressive enough to predict the outcome Y. This is implemented via standard empirical risk minimization as (Kolchinsky et al., 2019):

$$\mathbb{E}_{(\mathbf{x},\mathbf{t},y)\sim p(\mathbf{x},\mathbf{t},y)} \left[\big(y - f(\phi(\mathbf{x}), \tilde{\mathbf{t}})\big)^2 \right], \tag{24}$$

where $\tilde{\mathbf{t}}$ is a learned treatment embedding from an embedding function $\tau : (w, d) \in \mathcal{T} \to \tilde{\mathcal{T}}$ and f an output predictor head with the mapping $\tilde{\mathcal{T}} \times \mathcal{Z} \to \mathcal{Y}$. We estimate this expectation using the empirical mean squared error (MSE) over the training data via $\frac{1}{N} \sum_{i=1}^{N} \big(y_i - f(\phi(\mathbf{x}_i), \tilde{\mathbf{t}}_i)\big)^2$.

5.2.2 Treatment-Compression Term

Minimizing $I(\mathbf{z};\mathbf{t})$ is achieved by estimating the CS divergence between $\mathbf{T}$ and $\mathbf{Z}$. Our approach is conceptually similar to existing HSIC-based methods (e.g., Bellot et al. (2022)) which enforce $\mathbf{t} \perp \mathbf{z}$, but replaces HSIC with the CS divergence. Formally, we minimize:

$$D_{\mathrm{CS}}(p(\mathbf{z},\mathbf{t})\|p(\mathbf{z})p(\mathbf{t})) = -\log\left(\frac{\left(\int p(\mathbf{z},\mathbf{t})\,p(\mathbf{z})p(\mathbf{t})\,d\mathbf{z}d\mathbf{t}\right)^2}{\int p(\mathbf{z},\mathbf{t})^2\,d\mathbf{z}d\mathbf{t}\cdot\int p(\mathbf{z})^2p(\mathbf{t})^2\,d\mathbf{z}d\mathbf{t}}\right). \tag{25}$$

Given a batch of data points $\{(\mathbf{x}_i,\mathbf{t}_i)\}_{i=1}^N \sim p(\mathbf{x},\mathbf{t})$, we compute representations via a deterministic encoder $\mathbf{z}_i = \phi(\mathbf{x}_i)$. This induces joint samples $\{(\mathbf{z}_i,\mathbf{t}_i)\}_{i=1}^N \sim p(\mathbf{z},\mathbf{t})$, which can be used to assess statistical dependence between $\mathbf{z}$ and $\mathbf{t}$. The empirical CS divergence can then be estimated as (Yu et al., 2024):

$$\begin{aligned}\widehat{I}_{\mathrm{CS}}(\mathbf{z};\mathbf{t}) &= \log\left(\frac{1}{N^2}\sum_{i,j}^N K_{ij}Q_{ij}\right) + \log\left(\frac{1}{N^4}\sum_{i,j,q,r}^N K_{ij}Q_{qr}\right) - 2\log\left(\frac{1}{N^3}\sum_{i,j,q}^N K_{ij}Q_{iq}\right)\\ &= \log\left(\frac{1}{N^2}\mathrm{tr}(KQ)\right) + \log\left(\frac{1}{N^4}\mathbb{1}^TK\mathbb{1}\mathbb{1}^TQ\mathbb{1}\right) - 2\log\left(\frac{1}{N^3}\mathbb{1}^TKQ\mathbb{1}\right),\end{aligned} \tag{26}$$

where $\mathbb{1}$ is a $N\times 1$ vector of ones, and K and Q denote the Gram matrices for variables z and t, respectively. Specifically, $K_{i,j} = \kappa(z_i, z_j)$ with κ is a positive-definite kernel, such as the Gaussian RBF kernel. The second equality of Eq. (26) reduces the complexity to $\mathcal{O}(N^2)$.

Note that our empirical estimator of I_{CS} is fully non-parametric and does not rely on any parametric distributional assumptions on $p(z,t)$, such as Gaussianity, even though our first theoretical result in Theorem 1 assumes such a form.

5.2.3 Approximating the Compression Term

To minimize $I(\mathbf{x};\mathbf{z})$, we limit the latent space via a fixed low-dimensional $\mathbf{z}\in\mathbb{R}^d$ and apply regularization. Group sparsity and entropy-based penalties (e.g., log-det covariance) further reduce $\mathbf{z}$'s capacity (Dai et al., 2018; Kawaguchi et al., 2023; Tishby et al., 2000). In particular, we have regularization term

$$R_{\mathrm{dim}}(\mathbf{z}) := \|\mathbf{z}^\top\|_{2,1} + \kappa\log\det\left(\Sigma_z + \epsilon I\right), \tag{27}$$

where $\|\mathbf{z}^\top\|_{2,1} = \sum_{j=1}^d\left(\sum_{i=1}^N z_{ij}^2\right)^{1/2}$ is the $(2,1)$-norm promoting column sparsity, and $\Sigma_z = \frac{1}{N}\sum_{i=1}^N(z_i - \bar{z})(z_i-\bar{z})^\top$ is the empirical covariance matrix. I denotes the identity matrix of dimension $d_z\times d_z$ with a scalar ϵ to ensure numerical stability, $\bar{z}$ the empirical mean of the batch and κ a hyperparameter which takes on values in $[0,1]$ and is used to balances the two terms.

5.3 Model Objectives and Architecture

Our model architecture builds upon existing methodologies such as VCNet Bellot et al. (2022), but separates the treatment embedding and covariate embedding layers to align with the modelling objectives and the observational structure. As shown in Figure 2, we learn an embedding of the covariate space via $\phi:\mathcal{X}\to\mathcal{Z}$. A treatment encoder $\tau(w,d)$ learns a mapping $\mathcal{T}\to\tilde{\mathcal{T}}$ is used to impose a bottleneck in the architecture, promoting sample efficiency. Lastly, f predicts outcomes given both $\tilde{\mathbf{t}}$ and the covariate representation $\mathbf{z}$. The network is optimized end-to-end using backpropagation, with gradients flowing through all components of the architecture.[1] The optimization objective consists of the empirical loss term, along with two theoretically

[1]Note that, following standard practice in continuous causal modeling (e.g., Bellot et al. (2022), Kazemi & Ester (2024)), we assume ϕ to be a bijection for theoretical analysis. In practical implementation, however, ϕ is parameterized as a neural network and is not required to be invertible.

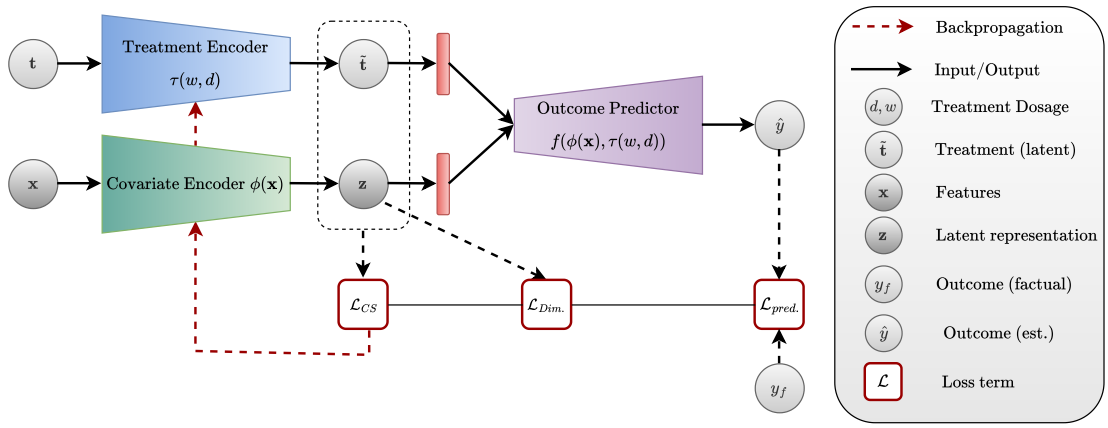

Figure 2: Illustration of the **IBEX** model architecture. IBEX jointly encodes treatment and covariate information using a *Treatment Encoder* $\tau(w, d)$ and a *Covariate Encoder* $\phi(x)$, where w and d denote the treatment identifier and dosage, respectively. The encoded representations are combined to predict the outcome $\hat{y}$ using an *Outcome Predictor* $f(\phi(x), \tau(w, d))$.

motivated regularization terms. Formally:

$$\begin{aligned}\mathcal{L}_{\text{IBEX}}(\phi, \tau, f) = & \underbrace{\mathbb{E}_{(\mathbf{x},\mathbf{t},y)\sim p(\mathbf{x},\mathbf{t},y)}\left[\left(y - f(\phi(\mathbf{x}), \tilde{\mathbf{t}})\right)^2\right]}_{\text{Prediction Term}} \\ + & \underbrace{\beta \cdot R_{\text{dim}}(\mathbf{z})}_{\text{Dimensionality Bottleneck}} + \underbrace{\gamma \cdot D_{\text{CS}}\left(p_\phi(\mathbf{z}, \mathbf{t}) \| p_\phi(\mathbf{z}) p(\mathbf{t})\right)}_{\text{Treatment compression Term}},\end{aligned} \tag{28}$$

where β and γ are hyperparameters which take on values in $[0, 1]$.

6 Experiments

We conduct a series of experiments following the setups used in previous comparative studies Bellot et al. (2022); Bica et al. (2020); Kazemi & Ester (2024). We run the experiments on a MacOS M4 system with a 10-core CPU, 32 GB unified RAM, and 120 GB/s memory bandwidth. Our implementation is available at `https://anonymous.4open.science/r/IBEX-D261`.

Baselines. We compare with: (1) **DRNet** (Schwab et al., 2020), with two variants—**HSIC** (Gretton et al., 2007) and **Wass** (Villani et al., 2008) regularization; (2) **SCIGAN** (Bica et al., 2020), a GAN-based counterfactual model; (3) Generalised Propensity Score **GPS** (Imbens, 2000); (4) a two-layer multilayer perceptron **MLP**; (5) **VCNet** (Bellot et al., 2022) with HSIC and Wass variants; (6) **ACFR** (Kazemi & Ester, 2024), which uses adversarial KL loss and attention; and (7) **GIKS** (Nagalapatti et al., 2024), which uses data-augmentation as a debiasing method.

Benchmark Datasets. We evaluate on three datasets: (1) **MIMIC-IV** (Johnson et al., 2023) contains records from 5,476 ICU patients who received mechanical ventilation. Treatment defined as a 2D continuous vector of ventilator settings (tidal volume and respiratory rate). (2) **News** (Asuncion et al., 2007) is a bag-of-words data set of New York Times articles. Lastly, (3) The Cancer Genome Atlas (**TCGA**) is a gene expression data for 9,000 cancer patients. We synthetically generate treatments and outcomes in accordance with previous work (e.g. Bica et al. 2020; Kazemi & Ester 2024; Schwab et al. 2020) (details in Appendix E).

Evaluation metrics. We report the square root of the **Mean Integrated Squared Error (MISE)**: $\frac{1}{N\,|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{i=1}^{N}\int_{\mathcal{D}_w}\left(y_i(w,d) - \hat{y}_i(w,d)\right)^2 \mathrm{d}d$, which averages squared errors between true and predicted outcomes over individuals, treatment types $\mathcal{W}$, and dosage ranges $\mathcal{D}_w$ and effectively compares the true

Method	News	MIMIC-IV	TCGA
SCIGAN	3.71 ± 0.05	2.09 ± 0.12	0.36 ± 0.04
DRNet$_{\mathbf{HSIC}}$	4.98 ± 0.12	4.45 ± 0.07	3.02 ± 0.28
DRNet$_{\mathbf{Wass}}$	5.07 ± 0.12	4.47 ± 0.12	1.73 ± 0.26
VCNet$_{\mathbf{HSIC}}$	3.41 ± 0.11	1.15 ± 0.02	0.95 ± 0.02
VCNet$_{\mathbf{Wass}}$	3.46 ± 0.04	1.23 ± 0.12	1.06 ± 0.01
GPS	6.97 ± 0.11	7.39 ± 0.00	6.12 ± 0.91
MLP	5.48 ± 0.16	5.34 ± 0.16	2.02 ± 0.33
ACFR	5.44 ± 0.14	2.19 ± 0.19	1.03 ± 0.12
GIKS	3.66 ± 0.04	**1.09** ± 0.06	1.25 ± 0.06
IBEX	**2.96** ± 0.16	**1.05** ± 0.02	**0.15** ± 0.09

Table 1: Out-of-sample performance of $\sqrt{\text{MISE}}$ on News, MIMIC, and TCGA datasets. Lower is better. Highest performer ($p < 0.05$ paired t-test) in bold face. Results are averaged over $n = 10$ random seeds. We assess significance using paired two-sided t-tests between our method and each baseline, with Holm correction for multiple comparisons.

Method	News	MIMIC-IV	TCGA
SCIGAN	3.90 ± 0.05	0.32 ± 0.05	0.23 ± 0.03
DRNet$_{\mathbf{HSIC}}$	4.17 ± 0.11	1.44 ± 0.05	1.24 ± 0.03
DRNet$_{\mathbf{Wass}}$	4.56 ± 0.12	1.37 ± 0.05	1.27 ± 0.05
VCNet$_{\mathbf{HSIC}}$	3.10 ± 0.21	0.63 ± 0.02	0.39 ± 0.01
VCNet$_{\mathbf{Wass}}$	2.99 ± 0.12	0.58 ± 0.03	0.44 ± 0.03
GPS	24.1 ± 0.55	20.2 ± 0.01	1.26 ± 0.01
MLP	6.45 ± 0.21	1.65 ± 0.05	1.13 ± 0.17
ACFR	5.11 ± 0.12	0.80 ± 0.02	1.11 ± 0.13
GIKS	2.15 ± 0.09	0.51 ± 0.02	0.95 ± 0.03
IBEX	**1.75** ± 0.09	**0.31** ± 0.04	**0.15** ± 0.03

Table 2: Out-of-sample performance of $\sqrt{\text{PE}}$. Lower is better. Highest performer ($p < 0.05$ paired t-test) in bold face.

outcome for a given treatment (and dosage) and the predicted outcome. We also report the square root of the **Policy Error (PE)**: $\frac{1}{N}\sum_{i=1}^{N}(y_i(w_i^*, d_i^*) - y_i(\hat{w}_i^*, \hat{d}_i^*))^2$, where (w_i^*, d_i^*) is the actual optimal treatment–dosage and $(\hat{w}_i^*, \hat{d}_i^*)$ is the predicted optimal treatment chosen by the model. PE quantifies regret from suboptimal policy choices.

6.1 Experimental Results

Tables 1-2 show that IBEX achieves the best performance on MIMIC with $\sqrt{\text{MISE}} = 1.61$ and $\sqrt{\text{PE}} = 0.12$, showing accurate outcome estimation and strong policy performance in a clinical setting with multivariate treatments, outperforming or matching VCNet-based approaches. On the News dataset, IBEX leads with the lowest $\sqrt{\text{MISE}} = 2.96$ and $\sqrt{\text{PE}} = 1.75$. On the TCGA dataset, IBEX again achieves the lowest ($\sqrt{\text{MISE}} = 0.15$), indicating that our approach allows for precise modelling of gene expression outcomes and reliable treatment policy learning.

Ablation Results. Table 3 reports MIMIC-IV results for four IBEX variants: using only the dimensionality regularizer ($\beta = 0.1$), only the Cauchy regularizer ($\gamma = 0.01$), neither regularizer, or both (full IBEX). Both regularizers independently improve performance over the unregularized baseline ($\sqrt{\text{MISE}} = 1.32$), with scores of 1.11 and 1.08, respectively, demonstrating the clear effectiveness of including the regularizers in the optimisation objective, as well as their complementary benefits.

Treatment bias robustness. We assess model robustness to treatment bias on the News and MIMIC-IV datasets. We vary the assignment bias α from 0 (none) to 10 (strong). As shown in Figure 3, IBEX maintains generally low $\sqrt{\text{MISE}}$ and $\sqrt{\text{PE}}$ even under strong bias, highlighting its resilience to confounding. In the

Method	$\sqrt{\text{MISE}}$	$\sqrt{\text{PE}}$
IBEX$_{\textbf{Dim. Reg. Only}}$	1.11 ± 0.01	0.57 ± 0.01
IBEX$_{\textbf{CS. Reg. Only}}$	1.08 ± 0.02	0.59 ± 0.02
IBEX$_{\textbf{No regularization}}$	1.32 ± 0.02	0.71 ± 0.01
IBEX	1.05 ± 0.16	0.12 ± 0.09

Table 3: Out-of-sample performance on various versions of the IBEX model on the MIMIC-IV dataset. Lower is better. In this setting $\gamma = 0.01$ and $\beta = 0.1$.

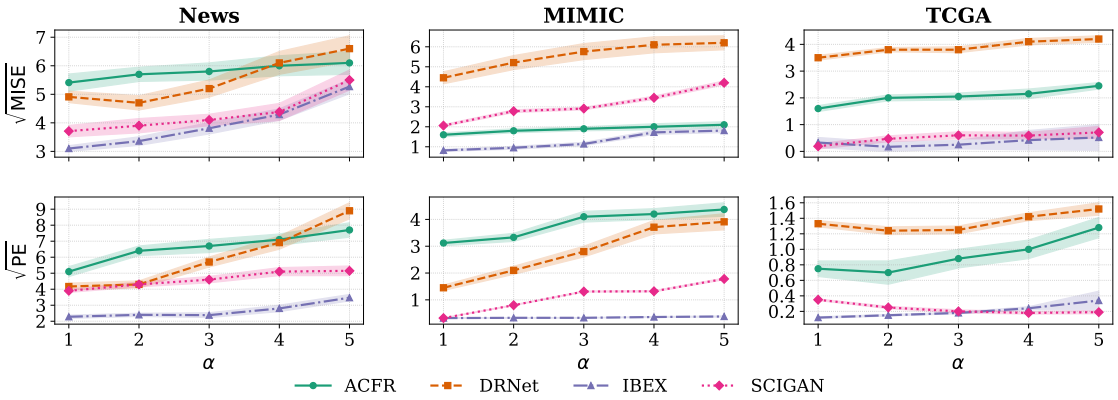

Figure 3: Results (out-of-sample) on News for Various Treatment-bias Values. We used DRNet$_{\textbf{HSIC}}$ in this setup.

MIMIC-IV setting on $\sqrt{\text{MISE}}$, ACFR also demonstrates a steady performance which eventually matches ours.

7 Discussion

IBEX outperforms all baselines across diverse datasets, demonstrating the strength of Cauchy–Schwarz (CS) divergence-based regularization for estimating continuous treatment effects. Its advantage is particularly evident under treatment-assignment bias, a common challenge in observational data. Beyond its theoretical grounding, IBEX performs well empirically across domains with different data characteristics, suggesting that the CS divergence provides a stable and expressive measure for enforcing representation invariance.

Nevertheless, like other causal inference approaches, IBEX relies on standard assumptions such as ignorability and overlap, which may not always hold in practice.

Future work could explore relaxing theoretical conditions, extending IBEX to dynamic treatment regimes (e.g., chronic care), and improving scalability. Integrating IBEX with conformal prediction or adversarial debiasing techniques may further enhance its utility in real-world settings with missing covariates or unmeasured confounding, broadening its impact in precision medicine and policy evaluation.

8 Conclusion

We introduced IBEX, a method for continuous and multivariate treatment effect estimation grounded in the information bottleneck principle. By deriving novel counterfactual generalization bounds and implementing them via two regularization objectives, IBEX improves out-of-sample performance while maintaining tractability, even with high-dimensional structured treatments.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

Alexis Bellot, Anish Dhir, and Giulia Prando. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage. *arXiv preprint arXiv:2205.14692*, 2022.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33: 16434–16445, 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

Bin Dai, Chen Zhu, Baining Guo, and David Wipf. Compressing neural networks using the variational information bottleneck. In *International Conference on Machine Learning*, pp. 1135–1144. PMLR, 2018.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop*, pp. 595–609, 2003.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960, 1986.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87 (3):706–710, 2000.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International Conference on Machine Learning*, pp. 16049–16096. PMLR, 2023.

Amirreza Kazemi and Martin Ester. Adversarially balanced representation for continuous treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13085–13093, 2024.

Sungyub Kim, Yongsu Baek, Sung Ju Hwang, and Eunho Yang. Reliable estimation of individual treatment effect with causal information bottleneck. *arXiv preprint arXiv:1906.03118*, 2019.

Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

Zhenyu Lu, Yurong Cheng, Mingjun Zhong, George Stoian, Ye Yuan, and Guoren Wang. Causal effect estimation using variational information bottleneck. In *International Conference on Web Information Systems and Applications*, pp. 288–296. Springer, 2022.

Lokesh Nagalapatti, Akshay Iyer, Abir De, and Sunita Sarawagi. Continuous treatment effect estimation using gradient interpolation and kernel smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14397–14404, 2024.

Sonali Parbhoo, Mario Wieser, Aleksander Wieczorek, and Volker Roth. Information bottleneck for estimating treatment effects with systematically missing covariates. *Entropy*, 22(4):389, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python, 2011. URL https://scikit-learn.org/. Accessed: April 28, 2023.

Jose C Principe, Dongxin Xu, Qun Zhao, and John W Fisher Iii. Learning from examples with information theoretic criteria. *Journal of VLSI signal processing systems for signal, image and video technology*, 26 (1):61–77, 2000.

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

Maresa Schröder, Dennis Frauen, Jonas Schweisthal, Konstantin Heß, Valentyn Melnychuk, and Stefan Feuerriegel. Conformal prediction for causal effects of continuous treatments. *arXiv preprint arXiv:2407.03094*, 2024.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5612–5619, 2020.

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Reliable off-policy learning for dosage combinations. *Advances in Neural Information Processing Systems*, 36:67900–67924, 2023.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Akira Tanimoto, Tomoya Sakai, Takashi Takenouchi, and Hisashi Kashima. Regret minimization for causal inference on large treatment space. In *International Conference on Artificial Intelligence and Statistics*, pp. 946–954. PMLR, 2021.

Ruo Yu Tao, Vincent François-Lavet, and Joelle Pineau. Novelty search in representational space for sample efficient exploration. *Advances in Neural Information Processing Systems*, 33:8114–8126, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Linh Tran, Maja Pantic, and Marc Peter Deisenroth. Cauchy–schwarz regularized autoencoder. *Journal of Machine Learning Research*, 23(115):1–37, 2022.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Wenzhe Yin, Shujian Yu, Yicong Lin, Jie Liu, Jan-Jakob Sonke, and Efstratios Gavves. Domain adaptation with cauchy-schwarz divergence. *arXiv preprint arXiv:2405.19978*, 2024.

Shujian Yu, Xi Yu, Sigurd Løkse, Robert Jenssen, and Jose C Principe. Cauchy-schwarz divergence information bottleneck for regression. *arXiv preprint arXiv:2404.17951*, 2024.