Do Regularization Methods for Shortcut Mitigation Work As Intended?

Haoyang Hong
Imperial College London

Ioanna Papanikolaou
Imperial College London

Sonali Parbhoo
Imperial College London

Abstract

Mitigating shortcuts, where models exploit spurious correlations in training data, remains a significant challenge for improving generalization. Regularization methods have been proposed to address this issue by enhancing model generalizability. However, we demonstrate that these methods can sometimes overregularize, inadvertently suppressing causal features along with spurious ones. In this work, we analyze the theoretical mechanisms by which regularization mitigates shortcuts and explore the limits of its effectiveness. Additionally, we identify the conditions under which regularization can successfully eliminate shortcuts without compromising causal features. Through experiments on synthetic and real-world datasets, our comprehensive analysis provides valuable insights into the strengths and limitations of regularization techniques for addressing shortcuts, offering guidance for developing more robust models.

1 Introduction

Despite successes of modern machine learning (ML), ML models frequently fail to generalize in scenarios where distribution shift occurs (Beery et al., 2018; Ilyas et al., 2019; Subbaswamy et al., 2019; Arjovsky et al., 2019) due to shortcut learning (Jabbour et al., 2020; Geirhos et al., 2018; Azulay and Weiss, 2019). Shortcut learning is defined as follows: *"A predictor relies on input features that are easy to represent (i.e., shortcuts) and are predictive of the label in the training data, but whose association with the label changes under relevant distribution shifts."*(Makar et al., 2022). If left unresolved, models tend to exploit these spurious correlations between input attributes and output to produce high accuracy predictions on training data (Hermann et al., 2024). However, when these models encounter out-of-distribution data (OOD) or data from a different distribution at test time, exploiting the incorrect information can negatively impact a model's generalization and robustness.

Several methods have been proposed to address a model's reliance on spurious attributes (Ye et al., 2024). Some of these methods assume that the spurious attributes are known apriori and develop regularization strategies (Veitch et al., 2021; Makar et al., 2022; Kaur et al., 2022; Wang et al., 2022), optimization techniques (Sagawa et al., 2019) or data augmentation strategies (Kaushik et al., 2021; Joshi and He, 2021) to explicitly make a model invariant to those specific attributes. Kumar et al. (2024) do not explicitly assume an attribute is either spurious or not, and develop a method for regularizing the effect of attributes proportional to their *average causal effect*. While Kumar et al. (2024) is significantly more general than former approaches, it relies on several specific assumptions about the data generating process that may not hold in practice or may not be available in advance.

Unlike previous works, this paper provides a comprehensive analysis (both theoretically and empirically) of the performance of regularization methods for shortcut mitigation and demonstrates that these methods can sometimes behave counterintuitively than indicated in existing literature. We identify several failure modes where issues remain unaddressed, particularly when shortcuts are unknown a priori or correlate with the outcome of interest. Specifically, we (1) derive the theoretical conditions for the success of these strategies and explore their practical limits; (2) demonstrate through synthetic and real experiments that spurious information can persist in concepts, allowing the model to exploit it for improved predictions even after regularization, especially when shortcuts correlate with input features or outputs; and (3) perform an in-depth analysis on strengths and limitations of regu-

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

larization techniques for addressing shortcut learning and provide practical guidance and considerations for developing models that are robust to shortcuts.

2 Related work

Model-based Methods for Shortcut Mitigation. Shortcuts occur as a result of distribution shift (Subbaswamy et al., 2019; Arjovsky et al., 2019). Invariant risk minimization (Arjovsky et al., 2019) aims to make models invariant to changes in the data distribution across environments, encouraging reliance on stable, causal features as oppose to shortcuts. Group-based methods (Sagawa et al., 2019) model multiple groups of the training data based on the spurious attribute and optimize for worst group accuracy. This requires knowledge of the spurious variables apriori. Other methods such as (Liu et al., 2021; Creager et al., 2021) relax this assumption to require the spurious attribute only for validation data. In contrast to these, our work considers how regularization methods perform for shortcut removal.

Regularization Methods for Credible Models. Many works have considered augmenting the data to decorrelate shortcuts with data (Cubuk et al., 2018) or regularizing the model to not consider the shortcut (Wang et al., 2022; Veitch et al., 2021; Kaur et al., 2022; Zwaan et al., 2012). However, these methods fail when shortcuts and features are correlated since without prior knowledge, these are indistinguishable. Hybrid methods such as Zhang et al. (2018); Ganin et al. (2016) combine model-based methods with regularization to consider reweighting the data or using domain-adversarial training to learn invariant representations and avoid shortcuts. These methods all assume shortcuts are known *apriori*. Zheng and Makar (2022) consider automatically discovering spurious attributes by building risk invariant predictors that are additionally regularized to reduce variance. Further, Pezeshki et al. (2021) proposed using the network logits as regularization to achieve better generalization performance in binary classification problems. Unlike these works, we demonstrate that regularization methods can behave in counterintuitive ways than expected which could have significant impact on the overall influence of shortcuts.

Causal Methods for Shortcut Mitigation. Causal methods such as Schölkopf et al. (2021) explicitly leverage causal structures via causal representation learning methods to avoid relying on shortcuts. Kumar et al. (2024) propose a technique for regularizing attributes proportional to their average causal effect. The authors use the fact that changing spurious attributes will not lead to a change in the task label and should thus have zero causal effect on the task label. However, the authors assume complete knowledge of the underlying causal structure of the data which may not always be available, particularly if certain key variables are unknown.

3 Setup and Preliminaries

Concept Bottleneck Models (CBMs) (Koh et al., 2020) are designed to map inputs to interpretable features that are known to influence the outputs. This design can mitigate shortcut effects with the captured concepts but is not guaranteed to. Specifically, CBMs can often underperform compared to end-to-end models when there is an incomplete set of known concepts resulting from limited domain knowledge or insufficient data, or if the concepts themselves contain shortcut information (Havasi et al., 2022). Combining a CBM with a feature extractor to capture both causal and non-causal features may help compensate for the predictive performance gap and enables regularization methods to be seamlessly integrated into the framework, as has been demonstrated in Wang et al. (2022).

Consider a dataset $D = \{\boldsymbol{x}_i, \boldsymbol{c}_i, y_i\}_{i=1}^n$ consisting of n samples, where $y_i \in \mathbb{R}$ represents the target output feature, and $\boldsymbol{x}_i \in \mathbb{R}^d$ are the input features. The vector $\boldsymbol{c}_i \in \mathbb{R}^c$ contains c binary or continuous **known concepts** (denoted by $\boldsymbol{C}$), which are known to cause y_i based on domain knowledge. A concept bottleneck model (CBM) (Koh et al., 2020) $f_{\theta_c} \circ f_{\theta_x} : X \to C \to Y$ parameterized by θ_c and θ_x can be trained accordingly by equation 1.

$$\hat{\theta_c}, \hat{\theta_x} = \arg\min_{\theta_c, \theta_x} \frac{1}{n} \sum_i^n [\mathcal{L}(f_{\theta_x}(\boldsymbol{x}_i), c_i) + \lambda_{cbm} \mathcal{L}(f_{\theta_c} \circ f_{\theta_x}(\boldsymbol{x}_i), y_i)] \quad (1)$$

where $\mathcal{L}$ is the loss function and λ_{cbm} is the hyperparameter that balances two losses. Beyond the predefined $\boldsymbol{C}$, we define the remaining concepts as **unknown concepts** ($\boldsymbol{U}$), assuming that $\boldsymbol{C}$ and $\boldsymbol{U}$ adhere to Assumption 1.

Assumption 1 (Completeness of concepts) *The known concepts $\boldsymbol{C}$ and unknown concepts $\boldsymbol{U}$ together constitute the complete set of features that causally influence the output feature Y.*

In practice, $\boldsymbol{U}$ can be derived from models pre-trained on unbiased datasets for similar tasks or large datasets like ImageNet (Deng et al., 2009), which capture general features that serve as unknown concepts. However, during feature extraction, there is a risk of identifying shortcut features ($\boldsymbol{S}$) that are spuriously correlated with the output feature Y but not causally

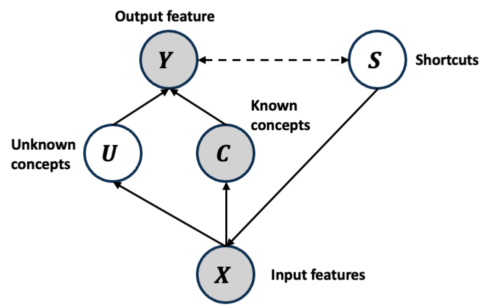

Figure 1: Causal graph for shortcut learning. $Y \in \mathbb{R}$ is the output feature. $\boldsymbol{X} \in \mathbb{R}^d$ represents the input features. $\boldsymbol{C} \in \mathbb{R}^c$ denotes the known concepts. Unknown concepts ($\boldsymbol{U} \in \mathbb{R}^u$) also cause Y but are not observed or known. Shortcuts ($\boldsymbol{S} \in \mathbb{R}^s$) that have spurious correlation with Y can be extracted from $\boldsymbol{X}$. $\boldsymbol{U}$ and $\boldsymbol{S}$ are mixed together. Observed variables are in gray. Dashed/solid edges represent correlation that is broken/unaffected under distribution shifts.

related. For instance, when inputting an image ($\boldsymbol{X}$) of a bird to determine whether it is a seabird or a landbird, known concepts ($\boldsymbol{C}$) such as the presence of wings or a beak can be extracted from the dataset. Unknown concepts ($\boldsymbol{U}$) are causally related features that are not measured in the dataset and can be extracted from pre-trained models. However, some extracted features, such as those representing the background color of the image, may be spurious concepts ($\boldsymbol{S}$). The causal relationships among all variables are illustrated in Figure 1.

After extracting $\boldsymbol{U}$ alongside the $\boldsymbol{S}$, we assume the linear relationships as outlined in Assumption 2 (2). This assumption is justified given that both $\boldsymbol{U}$ and $\boldsymbol{S}$ are in a latent space, meaning linear relationships are feasible, although this could also be relaxed (see supplement for details B.1.1).

Assumption 2 (Linear relationship) *Known concepts ($\boldsymbol{C} \in \mathbb{R}^c$) and unknown concepts ($\boldsymbol{U} \in \mathbb{R}^u$) have a linear relationship with output feature Y, i.e.*

$$Y = \sum_{i=1}^{c} \beta_{ci} C_i + \sum_{j=1}^{u} \beta_{uj} U_j + \epsilon, \ \ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{2}$$

Then the objective of mitigating shortcuts is equivalent to recovering $\beta_{ci}, i \in \{1, 2, \cdots, c\}$ and $\beta_{uj}, j \in \{1, 2, \cdots, u\}$ given input $\{\boldsymbol{C}, \boldsymbol{U}, \boldsymbol{S}\}$, while assigning 0 to all shortcut features. Since $\boldsymbol{S}$ and $\boldsymbol{U}$ are mixed from unknown concepts extraction, distinguishing between shortcut and causally relevant features prior to model input is not feasible. Therefore, it is impossible to get rid of shortcuts before feeding the data to the model. Consequently, regularization techniques applied to this linear layer aim to effectively mitigate the influence of shortcuts. Formally, the problem is defined as follows.

Problem Setting. *Given $\boldsymbol{X} \in \mathbb{R}^d$, $\boldsymbol{C} \in \mathbb{R}^c$ and $Y \in \mathbb{R}$ representing input features, known concepts, and output features respectively, train a CBM ($f_{\theta_c} \circ f_{\theta_x}$) based on $\boldsymbol{X}$, $\boldsymbol{C}$ and Y following Equation* (1). *Apply a pre-trained model ($f_{\theta_{us}} : \boldsymbol{X} \to \{\boldsymbol{U} \in \mathbb{R}^u, \boldsymbol{S} \in \mathbb{R}^s\}$) to $\boldsymbol{X}$ for $\{\boldsymbol{U}, \boldsymbol{S}\}$ extraction. Assume $\boldsymbol{C}$ and the extracted $\boldsymbol{U}$, $\boldsymbol{S}$ satisfy Assumption 1 and 2. Then, based on $\boldsymbol{H} = \{\boldsymbol{C}, \boldsymbol{U}, \boldsymbol{S}\} \in \mathbb{R}^h$ where $h = c + u + s$, a linear model* (3) *is trained by minimizing a loss function with regularization terms* (7) *to recover the true coefficients assigned to each variable. The targeted model is:*

$$\begin{aligned} \hat{Y} &= \sum_{i=1}^{h} \hat{\beta}_i H_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{H} & (3) \\ &= \hat{\boldsymbol{\beta}_{\boldsymbol{c}}}^T \mathbf{C} + \hat{\boldsymbol{\beta}_{\boldsymbol{us}}}^T [\mathbf{U} \ \ \mathbf{S}] & (4) \\ &= \hat{\boldsymbol{\beta}_{\boldsymbol{c}}}^T \mathbf{C} + \hat{\boldsymbol{\beta}_{\boldsymbol{u}}}^T \mathbf{U} + \hat{\boldsymbol{\beta}_{\boldsymbol{s}}}^T \mathbf{S} & (5) \\ &= \hat{\boldsymbol{\beta}_{\boldsymbol{c}}}^T (f_{\theta_x}(\boldsymbol{X})) + \left[\hat{\boldsymbol{\beta}_{\boldsymbol{u}}}^T \ \ \hat{\boldsymbol{\beta}_{\boldsymbol{s}}}^T\right] f_{\theta_{us}}(\boldsymbol{X}) & (6) \end{aligned}$$

The loss function with regularization is:

$$loss = \mathcal{L}(Y, \hat{Y}) + \lambda_{reg} \mathcal{R}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}) \tag{7}$$

where $\mathcal{L}$ represents the empirical risk, $\hat{Y}$ is the predicted output, and λ_{reg} denotes the regularization strength. $\mathcal{R}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}})$ represents the regularization term trained parameters. Bold symbols denote vectors or matrices, and symbols with $\hat{}$ indicate predicted values, while those without denote ground truth.

In this paper, we analyze five regularization methods for mitigating shortcut learning: Lasso (L1) regularization, Ridge (L2) regularization, causal regularization (Bahadori et al., 2017), Expert Yielded Estimates (EYE) regularization (Wang et al., 2018, 2022), and causal effect regularization (Kumar et al., 2024). Through both theoretical and empirical analyses, we evaluate these methods to assess their effectiveness under different conditions.

L1 regularization. L1 (lasso) regularization is commonly used to achieve sparse solutions (Vidaurre et al., 2013). We apply L1 regularization solely to the parameters $\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}$, as penalizing $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ (known concepts) is unnecessary, given that these concepts are established to causally influence the output.

$$\mathcal{R}_{L1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}) = ||\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}||_1 \tag{8}$$

where p-norm is defined as

$$||\boldsymbol{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

L2 regularization. L2 (ridge) regularization (Hoerl and Kennard, 1970b) is commonly used to mitigate the bias of coefficients (Hoerl and Kennard, 1970a). Similarly, we applied L2 regularization only to the parameters $\boldsymbol{\beta_u}$ and $\boldsymbol{\beta_s}$.

$$\mathcal{R}_{L2}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}) = ||\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}||_2^2 \tag{9}$$

EYE regularization. EYE regularization (Wang et al., 2018, 2022) is a combination of L1 and L2 regularizations. It is proposed to improve model credibility, encouraging denser coefficients for known concepts while promoting sparsity for other features. In our setting, it is formulated as

$$\mathcal{R}_{EYE}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}) = ||\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}||_1 + \sqrt{||\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2} \tag{10}$$

Causal & Causal effect regularization. Causal regularization (Bahadori et al., 2017) and causal effect regularization (Kumar et al., 2024) are variants of L2 regularization. The general form of these two regularization methods are

$$\mathcal{R}_{caus}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}) = \mathcal{R}_{caus}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{h} \lambda_i ||\hat{\boldsymbol{\beta}}_i||_2^2 \tag{11}$$

where λ_i is the hyperparameter assigned to the coefficient $\hat{\boldsymbol{\beta}}_i$. For causal regularization,

$$\lambda_i = 1/P(H_i \text{ is the cause of } Y) \tag{12}$$

For causal effect regularization,

$$\lambda_i = 1/\text{Estimated treatment effect of } (H_i) \tag{13}$$

In this paper, we treat these two regularization methods as equivalent because both assign coefficients that reflect the causal properties of features to the L2 regularization terms. Consequently, we use causal effect regularization as the representative method.

4 Theoretical Analysis of Shortcut Mitigation Methods

We first analyze the effectiveness of these regularization terms in shortcut mitigation. We conclude that L2, causal and causal effect regularization methods guarantee to alleviate the shortcut effects, i.e. shrink the absolute values of $\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}$ while the two other regularization methods cannot. The statements is formally structured as Proposition 1.

Proposition 1: *Follow the problem setting* (3)*, denote the trained parameters with regularization by*

$$\hat{\boldsymbol{\beta}}^{(\lambda)} = [\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{(\lambda)} \;\; \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{(\lambda)} \;\; \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}^{(\lambda)}]$$

where λ is the regularization strength. When $\lambda = 0$, the model is trained without regularization. (i) Training with L1 and EYE does not guarantee to alleviate the shortcut effects. (ii) Training with L2, causal and causal effect regularization can help mitigate shortcut to some extent, which is

$$|\hat{\boldsymbol{\beta}}_s^{reg}| \leq |\hat{\boldsymbol{\beta}}_s^{without\_reg}|$$

Proof sketch. (i) For L1 regularization, consider the case where U and S are perfectly correlated. Because L1 regularization will generate sparse results, the results can be either $\hat{\beta}_s = \beta_u$ and $\hat{\beta}_u = 0$ or vice versa. When the result is $\hat{\beta}_s = \beta_u$ and $\hat{\beta}_u = 0$, the shortcut effects are exacerbated rather than mitigated.

For EYE regularization, the coupling penalty on $\sqrt{||\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}$ creates a trade-off between $\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$. Hence, decreasing $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ to reduce the penalty may increase the need for larger $\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}$ to maintain the fit, which makes it possible to have $|\hat{\boldsymbol{\beta}}_s^{reg}| > |\hat{\boldsymbol{\beta}}_s^{without\_reg}|$.

(ii) The loss function with L2, causal and causal effect regularization can be formulated as

$$loss = \frac{1}{n}||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||_2^2 + \lambda||\boldsymbol{\mathcal{D}}_\lambda \hat{\boldsymbol{\beta}}_{us}||_2^2 \tag{14}$$

where $\mathcal{D}_\lambda$ is a diagonal matrix where the diagonal entities are the square root of the weights assigned to each coefficients in causal and causal effect regularization. For L2 regularization, $\mathcal{D}_\lambda = \mathbb{I}$ is an identity matrix.

Take derivative with respect to $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ and $\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}$,

$$\hat{\boldsymbol{\beta}_{\boldsymbol{us}}}^{(\lambda)} = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)^{-1} \boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{Y} \tag{15}$$

where

$$\boldsymbol{H}_{us} = [\boldsymbol{U} \;\; \boldsymbol{S}], \;\; \boldsymbol{\Pi}_c = \boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T$$

Then $\boldsymbol{H}_{us}^T(\mathbb{I} - \boldsymbol{\Pi}_c)\boldsymbol{H}_{us}$ is symmetric and normal, thus can be decomposed as

$$\boldsymbol{H}_{us}^T(\mathbb{I} - \boldsymbol{\Pi}_c)\boldsymbol{H}_{us} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \tag{16}$$

where $\boldsymbol{Q}$ is orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with non-negative eigenvalues Λ_i. Then

$$\boldsymbol{H}_{us}^T(\mathbb{I} - \boldsymbol{\Pi}_c)\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{z} \tag{17}$$

$$\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)^{-1}\boldsymbol{Q}\boldsymbol{z} \tag{18}$$

Then, for $i-$th entity,

$$\hat{\boldsymbol{\beta}}_{us,i} = \frac{n}{\frac{\Lambda_i}{n} + n\lambda\boldsymbol{\mathcal{D}}_{ii}^2} z_i' \tag{19}$$

Hence, increasing regularization strength λ increasing the positive denominator, which makes the absolute value of $\hat{\boldsymbol{\beta}}_{us,i}$ decrease. □. The entire proof is shown in the supplementary A.1.

Correlations between unknown concepts and shortcuts affect the regularization performance. The proof indicates that L1 and EYE regularization can sometimes amplify the influence of shortcut variables, particularly when these variables correlate with unknown concepts. This amplification occurs because the regularization process cannot distinguish between shortcuts and unknown concepts, making it challenging for the model to determine which features to prioritize. Furthermore, L1 regularization typically produces sparse models that focus on fewer features while maintaining predictive performance. However, when shortcuts are highly correlated with unknown concepts, especially when considering combinations of several unknown concepts, L1 regularization may disproportionately emphasize the shortcut features to achieve sparsity. This can exacerbate the problem of shortcut learning.

Regularization can over-regularize causally relevant features. The proof of Proposition 1 demonstrates that while L2-based regularization aims to mitigate the effects of shortcuts, it also diminishes the influence of unknown concepts, potentially degrading model performance. This suppression is influenced by the eigenvalues $\boldsymbol{\Lambda}$ of $\boldsymbol{H}_{us}^T(\mathbb{I} - \boldsymbol{\Pi}c)\boldsymbol{H}_{us}$ and the coefficients $\boldsymbol{\mathcal{D}}_{ii}$ assigned to each regularization term. For L2 regularization, $\boldsymbol{\Lambda}$ alone determines the resulting weights. For causal and causal effect regularization, accurate estimations of the probability of causal correlations and treatment effects are also critical. Inaccurate estimations can cause models to wrongly rely on shortcut variables, undermining model's robustness.

Additionally, we aim to investigate scenarios where regularization methods effectively eliminate shortcut learning. We detail instances where each regularization approach succeeds in shortcut mitigation, as outlined in Proposition 2.

Proposition 2: *Follow the problem setting, when $c = u = s = 1$, and S is correlated with Y by correlating with C and U, i.e. C and U are not co-linear, $S = \delta_c C + \delta_u U$, the regularization methods can eliminate the shortcuts under the following situations. (Without loss of generality, assume β_c and β_u are non negative, other cases can be derived in the same way.)*

(i) For L1 regularization: $\frac{\delta_c+\delta_u-1}{\delta_c} \leq 0$

(ii) For L2 regularization: $\frac{\beta_c+\beta_c\delta_u^2-2\beta_u\delta_c\delta_u}{\delta_c^2+\delta_u^2+1} \geq \beta_c$

(iii) For EYE regularization: $\frac{2\beta_c-2\frac{\delta_c}{\delta_u-1}\beta_u}{\left(\frac{\delta_c}{\delta_u-1}\right)^2+1} \geq \beta_c$

(iv) For causal and causal effect regularization, denote the coefficients assigned to C, U and S as λ_c, λ_u and λ_s, the condition is $\frac{\beta_c+\frac{\lambda_u}{\lambda_s}\delta_u^2\beta_c-\frac{\lambda_u}{\lambda_s}\delta_u\delta_c\beta_u}{\frac{\lambda_c}{\lambda_s}\delta_c^2+\frac{\lambda_u}{\lambda_s}\delta_u^2+1} = \beta_c$

Proof sketch. Given the simplified linear relationship between input variables and output as

$$Y_i = \beta_c C_i + \beta_u U_i + \beta_s S_i + \epsilon_i, i \in \{1, 2, \cdots, n\} \quad (20)$$

where the non-bold symbols indicate scalars rather than vectors. Choose the regularization strength λ so that after regularization, the loss for Y and $\hat{Y}$ is no less than without regularization. Further, for any $\hat{\beta}_c$, we can get the same empirical risk (empirical risk for noise terms) when $\hat{\beta}_u$ and $\hat{\beta}_s$ follow:

$$\hat{\beta}_u = \beta_u + \hat{\beta}_c\frac{\delta_u}{\delta_c} - \beta_c\frac{\delta_u}{\delta_c} \qquad \hat{\beta}_s = \frac{\beta_c - \hat{\beta}_c}{\delta_c} \quad (21)$$

where notations without ˆ represent the true weights and represent trained weights withˆ. Then for different regularization methods, the regularization term can be converted into a function with respect to $\hat{\beta}_c$. Taking the derivative with respect to $\hat{\beta}_c$ shows us when can the regularization be optimal. □

The complete proof are shown in supplementary A.2.

Data normalization enhances the effectiveness of regularization methods. According to the proposition, if all features, including the output feature Y, are standardized with zero mean and unit variance, then $\delta_c + \delta_u = 1$. Consequently, for $L1$ regularization, the optimal condition is satisfied. Moreover, for EYE regularization, the optimal condition can be simplified to:

$$\beta_c + \beta_u \geq \beta_c \quad (22)$$

This condition is inherently met, assuming β_c and β_u are non-negative, with β_u also being non-zero.

Accurate estimation of causal properties is essential for effectively mitigating shortcuts. Proposition 2 indicates that when the causal properties of shortcut features, especially the probability of a feature causing the output and its treatment effect, are accurately estimated to be close to zero, the shortcuts are successfully mitigated. This is because when $\lambda_s \to \infty$, the relationship in (iv) of Proposition 2 becomes $\beta_c = \beta_c$, which is inherently satisfied. Thus, precise estimation of causal properties is crucial for reducing shortcuts and uncovering the underlying causal structure of the data.

5 Experiments

In this section, we first explore the success and failure of regularization methods in mitigating shortcuts using synthetic datasets. Additionally, we demonstrate how these regularization methods work on three different real datasets to reduce shortcuts and the key factors that affect the performance. The results generate insights that align with our theoretical framework. The codes for experiments are available at `https://github.com/ai4ai-lab/regularization_for_shortcut_migtigation`.

Baselines. For each dataset, we initially train the model without any regularization to establish a baseline for using shortcut variables in prediction. Subsequently, we apply L1, L2, EYE, and causal effect regularization to assess their impact. We exclude causal regularization for two reasons. First, it performs similarly to causal effect regularization due to their analogous designs. Second, the dataset required to train the coefficients for each regularization term in causal regularization is not publicly available. Consequently, we use causal effect regularization as a representative method to assess performance.

Synthetic data. We generate three types of synthetic datasets (differing in terms of the relationship between shortcut variables and other variables) following:

1. Extract two concept variables ($\boldsymbol{C} = [C_1, C_2]$) and two unknown concepts ($\boldsymbol{U} = [U_1, U_2]$) from standard normal distribution $\mathcal{N}(0, 1)$.
2. Generate output feature (Y) based on a linear combination of $\boldsymbol{C}$ and $\boldsymbol{U}$ (independent of S).
3. Generate one shortcut S variable in three different ways:
   - Only correlating with $\boldsymbol{C}$
   - Correlating with $\boldsymbol{U}$ and partial of $\boldsymbol{C}$
   - Adding a random noise on Y

For each shortcut variable, correlations with input and output features exist only in the training data. In the test set, S is sampled from a standard normal distribution. For the synthetic dataset, we use a linear model with standardized inputs and apply each regularization method with a strength of 0.001. For causal effect regularization, the treatment effect of the shortcut is set to 0.001, while it is set to 1 for the other variables (details regarding the experimental settings can be found in supplementary B.1).

5.1 Results on Synthetic Data

To evaluate shortcut mitigation, we employ two metrics: trained weights and the estimated treatment effect of S across epochs. Using a synthetic dataset with known ground truth enables comparison between true and regularized weights, illustrating the effectiveness of regularization in reducing shortcut reliance.

To estimate the treatment effect, we adapt the T-learner approach. Since S is continuous, we randomly sample 100 instances from a standard normal distribution, input them into the model, and compute the standard deviation of the outputs as the estimated treatment effect. Effective mitigation should result in uniform outputs with a standard deviation of zero. The results are shown in Figure 2.

Regularization methods succeed when shortcuts correlate solely with known concepts. As shown in Figure 2, panel A demonstrates that when S correlates only with $\boldsymbol{C}$, all regularization methods perform effectively. This is because $\boldsymbol{C}$, which are not penalized by L1 and L2 regularization, contain all relevant information from S. Consequently, the model shifts its predictive reliance towards $\boldsymbol{U}$, reducing dependence on shortcuts. This prioritization allows regularization to effectively mitigate shortcut reliance. Additionally, the estimated treatment effects plot shows that causal effect regularization converges fastet than the other methods, highlighting its advantage.

Furthermore, we investigated the mitigation performance in relation to the relationship between the shortcut variable S and the unknown concepts $\boldsymbol{U}$. We conducted experiments on various correlations between S and $\boldsymbol{U}$, ranging from 0 to 0.5. The results indicate that as the correlation between S and $\boldsymbol{U}$ increases, the mitigation performance of the regularization methods declines rapidly. This highlights a limitation of regularization methods when the shortcut variable is highly correlated with unknown concepts. The detailed results are presented in Appendix B.1.3.

Regularization methods fail when shortcuts correlate with unknown concepts. Panel B shows that when S is partially correlated with $\boldsymbol{U}$, all regularization methods fail, except for causal effect regularization with accurately estimated treatment effects. The failure of the other methods aligns with the findings from Proposition 1. Specifically, L1 exacerbates shortcut learning by allocating all weights to S, favoring sparsity by prioritizing one S that provides similar predictive information as using the two unknown concepts. L2, on the other hand, distributes weights among both $\boldsymbol{U}$ and S, also fails to effectively mitigate shortcut reliance. The treatment effect plots reveal that the application of L1, L2, and EYE unexpectedly increases the model's dependence on S, which is contrary to the desired outcome.

Regularization methods fail when shortcuts are highly correlated with the output. Panel C reveals that when S is derived directly from the Y, specifically by adding random noise, all regularization methods except for causal effect regularization fail. In this case, the correlation between $\boldsymbol{U}$, $\boldsymbol{C}$ and S is the strongest among three scenarios, which might be the

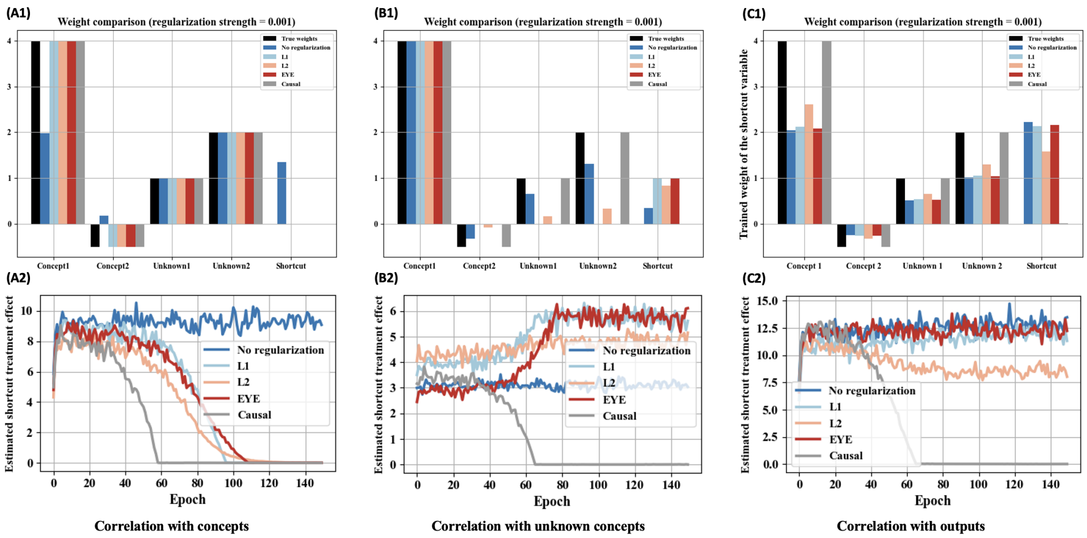

Figure 2: Comparison of trained weights and estimated treatment effects across training epochs. (A) Regularization are successful when S is only correalted with $\boldsymbol{C}$. (B) Regularization fail when S is correlated with $\boldsymbol{U}$. (C) Regularization fail when S is highly correlated with Y. In the weight comparison plots, black bars represent the true weights assigned to each variable, consistent across different shortcut scenarios. The treatment effect plots depict the estimated treatment effect of the shortcut variable throughout the training epochs.

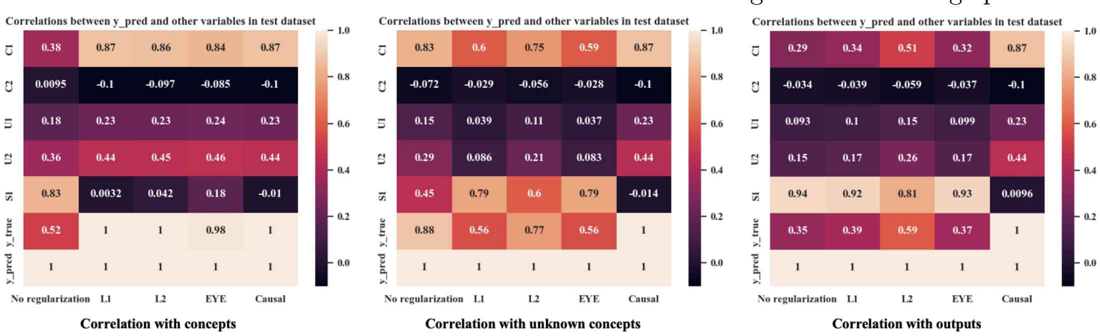

Figure 3: Correlation between predicted values and all variables in synthetic experiments, including true outputs in the test dataset, under different regularization methods. As the correlation between outputs and unknown concepts increases, regularization methods become less effective in mitigating shortcut dependencies.

reason to fail. Further, the treatment effect plot underscores this outcome, showing that the regularization approaches, aside from causal effect, fail to reduce the model's reliance on the shortcut variable, learning nothing to diminish this dependency during training.

Figure 3 illustrates the correlation between different regularization methods and predicted values. The results indicate that only when the shortcut variable is correlated with concepts do regularization methods effectively reduce its correlation with predictions to near zero. Otherwise, they provide only limited mitigation.

Complex model structures diminish the effectiveness of regularization methods. Furthermore, relaxing the assumption of a linear relationship (Assumption 2) allows us to apply regularization methods to neural networks. For instance, in a multi-layer network with a first layer consisting of 10 neurons and 5 input variables, the parameter matrix has a shape of $(10, 5)$. We can apply regularization to these parameters, designating the first two columns for known concepts (β_c) and the next two for unknown concepts (β_u). We conducted this experiment to investigate the application of regularization in non-linear models. The results of this experiment are presented in Figure 5. In comparison to Figure 2, when the shortcut is only correlated with known concepts, regularization methods are effective in linear models but not in neural networks. This discrepancy may be attributed to the deep architectures of neural networks, which can diminish the impact of regularization in mitigating shortcuts. Therefore, when applying regularization to address shortcuts, it is advisable to utilize simpler models and avoid overly complex structures. More details are shown in Appendix B.1.1.

In addition to the synthetic dataset, we explore the impact of regularization methods on real datasets and how varying the regularization strength affects model performance. We selected three real-world datasets, Colored-MNIST, MultiNLI and MIMIC-ICU for this analysis. We extracted concepts $\boldsymbol{C}$, unknown concepts $\boldsymbol{U}$, and shortcuts $\boldsymbol{S}$ from these datasets to demonstrate that the performance of the regularization methods align with the previous findings.

5.2 Results on Real Data.

Colored-MNIST. Following Arjovsky et al. (2019), we adapted the `MNIST` dataset (LeCun et al., 1998) by (i) binarizing the labels (digits 0-4 as one class, 5-9 as the other), and (ii) setting color labels s equal to y in the training set while reversing this in the test set. The image color (green or red) serves as a shortcut for predicting the binary label. We also created an unbiased dataset with random color assignments. Two models were trained: one for binary label prediction and one for color prediction, extracting concepts from the former and shortcuts from the latter. For unknown concept extraction, we used a pre-trained `resnet50` from `torchvision`.

MultiNLI. For MultiNLI dataset (Williams et al., 2017), the task is to classify whether a hypothesis is entailed or contradicts a premise. Previous work found a spurious correlation between contradictions and negation words (Gururangan et al., 2018). We focus on the binary task of distinguishing entailment from contradiction, introducing shortcuts by ensuring contradiction samples in the training set contain negation words. In the test set, entailment samples include negation words, breaking the correlation. We also created an unbiased dataset where negation and contradictions are independent. Two models were trained to extract concepts (true labels) and shortcuts (negation words), with unknown concept extraction using a pre-trained `BERT` model (Kenton and Toutanova, 2019). We trained models without regularization, then applied L1, L2, EYE, and causal effect regularization. Shortcut mitigation was evaluated using two metrics: (1) average absolute weights of shortcut variables, with values close to zero indicating reduced shortcut reliance, and (2) AUC on the test set.

MIMIC-ICU. We conducted further experiments using MIMIC-IV database (Johnson et al., 2023), focusing on predicting the length of stay (LOS) in intensive care units (ICU). This regression problem differs from the previous classification experiments. We identified 19 variables for predicting LOS, categorizing them into known and unknown concepts based on their correlations with LOS. Specifically, variables with correlations greater than 0.1 with LOS are classified as known concepts, while the others are considered unknown. This categorization is necessary due to the lack of generally pre-trained models for extracting information from this dataset. Finally, we got 12 known concepts and 7 unknown concepts. We also generated 10 shortcut variables based on LOS, all of which are highly correlated with LOS in the training dataset but randomly distributed in the test dataset. As the other experiments, we applied all regularization methods to train a regression model, using mean square error (MSE) to evaluate model performance. The results on real-world datasets are shown in Figure 4. The detailed experimental settings are shown in the Appendix B.2.1.

Mitigation performance depends on the correlation. Figure 4 shows that, in the synthetic dataset, all regularization strengths fail to mitigate shortcuts, except for causal effect regularization. In the Colored-MNIST dataset, proper regularization effectively mitigates shortcuts. However, in the MultiNLI dataset, while regularization strengths above 0.01 reduce shortcuts, they also harm predictive performance. This may be due to correlations within input features. As seen in Figure 7, the synthetic dataset has perfect correlation between the shortcut and output, preventing regularization success. In MultiNLI, shortcuts are only partially correlated with unknown concepts, leading to incomplete mitigation. In contrast, Colored-MNIST has shortcut-concept correlations, allowing regularization to work. Further, in MIMIC-ICU, shortcuts are highly correlated with the output, making it difficult to mitigate. Thus, shortcut mitigation success depends on the correlations between input features.

Choice of regularization strength is critical. The plots show that when the regularization strength is too small, all methods fail to mitigate the shortcuts. However, if the strength is too large, the model eliminates the shortcuts but loses its predictive ability. Additionally, the results also highlight the sensitivity of shortcut mitigation to regularization strength. For example, in the Colored-MNIST dataset, increasing the regularization strength from 10^{-3} to 10^{-1} led to an AUC improvement of nearly 0.9. Furthermore, althought the results indicate that the choice of regularization strength is critical, the optimal value for regularization strength can spread a large range. As shown in panal A of Figure 4 (varying regularization strengths vs. model performance), there is a relatively large range of regularization strengths (10^{-4} to 10^{-1}) within which shortcut mitigation remains effective. This indicates a degree of robustness in the choice of regularization strength.

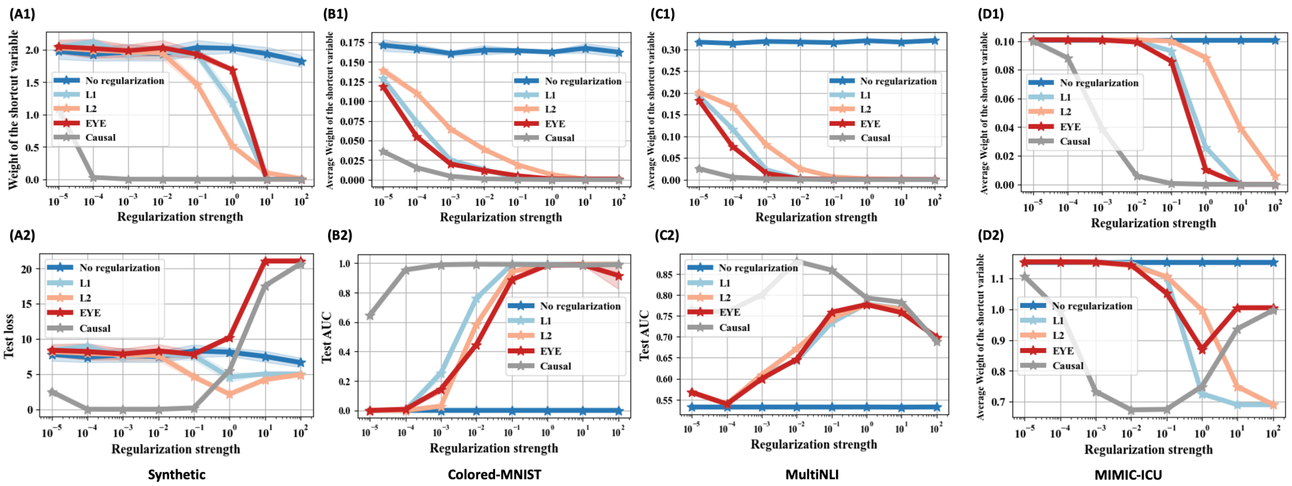

Figure 4: Comparison of weights and test loss (AUC for classification problems and MSE for regression problem) across varying regularization strengths for synthetic dataset, Colored-MNIST, MultiNLI, and MIMIC-IV dataset. Experiments were conducted ten times; the solid line represents the mean results, while the shaded area indicates the standard error of the ten experiments.

6 Practical Considerations

When applying regularization techniques such as L1, L2 and EYE to mitigate shortcut learning, several key considerations can help practitioners avoid common pitfalls and achieve better results:

Task-Specific Application. Regularization should be adapted to the complexity of the task. In high-dimensional tasks with multi-modal data, it is important to balance limiting shortcut learning while maintaining model flexibility. Over-regularization risks underfitting, while under-regularization may fail to effectively reduce shortcut reliance. Regularization should be adjusted flexibly based on the feature space.

Regularization Strength Matters. Regularization strength is critical but not a one-size-fits-all solution. Different strengths should be explored to find the right balance between reducing shortcut learning and maintaining task performance. Too little regularization may fail to address shortcut reliance, while too much could suppress important task-related information. Testing various strengths reveals the model's sensitivity to regularization. If performance is highly sensitive to this, alternative techniques may be worth considering, especially in complex tasks.

Handling Uncertainty in OOD Scenarios and Data Collection. Handling uncertainty in OOD data is a significant challenge in shortcut learning. If shortcuts are not fully mitigated, the model may still rely on them, increasing uncertainty. Regularization alone may not be enough, particularly with limited or imbalanced training data. Expanding and diversifying the dataset, especially with OOD samples, can reduce shortcut reliance and improve generalization. Testing regularization in multiple OOD scenarios helps ensure the model generalizes well beyond training data.

Refining Causal Concept Variables. Regularization techniques often treat causal variables as static and perfectly defined, but real-world tasks are more dynamic. If shortcuts continue to influence the model after regularization, it may indicate that the initial set of causal variables is insufficient. Practitioners should be open to refining the set of concepts based on performance, especially in OOD contexts, as this can reduce reliance on shortcuts. Standard regularization methods such as L1, L2, EYE are agnostic to the treatment effect and constrain learning without knowledge of this process. However, the treatment effect is crucial to ensure the underlying causal process is captured. Alternatively, one might consider iteratively refining and updating assumptions about the causal variables to improve model robustness.

7 Conclusion

Regularization methods are helpful for mitigating shortcut learning but face challenges. Our analysis shows that techniques like L1, L2 and EYE often over-regularize, suppressing both spurious and causal features and limiting their effectiveness. Theoretical exploration and experiments on synthetic and real-world datasets indicate that current methods inadequately disentangle causal features from shortcuts. Future work should aim to develop techniques that better balance shortcut reduction with the preservation of causal information, enhancing model performance.

Acknowledgements

The authors would like to thank Jenna Wiens (University of Michigan) and Amit Sharma (Microsoft Research India) for valuable discussions regarding this work, as well as the anonymous reviewers for constructive feedback.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Azulay, A. and Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25.

Bahadori, M. T., Chalupka, K., Choi, E., Chen, R., Stewart, W. F., and Sun, J. (2017). Causal regularization. *arXiv preprint arXiv:1702.02604*.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

Creager, E., Jacobsen, J.-H., and Zemel, R. (2021). Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Havasi, M., Parbhoo, S., and Doshi-Velez, F. (2022). Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397.

Hermann, K., Mobahi, H., Thomas, F., and Mozer, M. C. (2024). On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M. W., and Wiens, J. (2020). Deep learning applied to chest x-rays: exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Joshi, N. and He, H. (2021). An investigation of the (in) effectiveness of counterfactually augmented data. *arXiv preprint arXiv:2107.00753*.

Kaur, J. N., Kiciman, E., and Sharma, A. (2022). Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*.

Kaushik, D., Setlur, A., Hovy, E. H., and Lipton, Z. C. (2021). Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.

Kumar, A., Deshpande, A., and Sharma, A. (2024). Causal effect regularization: automated detection and removal of spurious correlations. *Advances in Neural Information Processing Systems*, 36.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., and D'Amour, A. (2022). Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR.

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A. C., Precup, D., and Lajoie, G. (2021). Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Subbaswamy, A., Schulam, P., and Saria, S. (2019). Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. (2021). Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208.

Vidaurre, D., Bielza, C., and Larranaga, P. (2013). A survey of l1 regression. *International Statistical Review*, 81(3):361–387.

Wang, J., Jabbour, S., Makar, M., Sjoding, M., and Wiens, J. (2022). Learning concept credible models for mitigating shortcuts. *Advances in neural information processing systems*, 35:33343–33356.

Wang, J., Oh, J., Wang, H., and Wiens, J. (2018). Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2417–2426.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Ye, W., Zheng, G., Cao, X., Ma, Y., Hu, X., and Zhang, A. (2024). Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zheng, J. and Makar, M. (2022). Causally motivated multi-shortcut identification and removal. *Advances in Neural Information Processing Systems*, 35:12800–12812.

Zwaan, L., Thijs, A., Wagner, C., van der Wal, G., and Timmermans, D. R. (2012). Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Academic Medicine*, 87(2):149–156.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes - will be made public after review process]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials: Do Regularization Methods for Shortcut Mitigation Work As Intended?

A Proofs for section

A.1 Proof for Proposition 1

Proposition 1: *Follow the problem setting* (3)*, denote the trained parameters with regularization by*

$$\hat{\boldsymbol{\beta}}^{(\lambda)} = [\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}{}^{(\lambda)} \;\; \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}{}^{(\lambda)} \;\; \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}{}^{(\lambda)}]$$

where λ is the regularization strength. When $\lambda = 0$, the model is trained without regularization.

(i) Training with L1 and EYE does not guarantee alleviation of the shortcut effects.

(ii) Training with L2, causal and causal effect regularization can help mitigate shortcut to some extents, i.e.

$$|\hat{\boldsymbol{\beta}}_s{}^{reg}| \leq |\hat{\boldsymbol{\beta}}_s{}^{without\_reg}|$$

Proof:

According to the problem setting (3), the assumed linear model is

$$\hat{\boldsymbol{Y}} = \boldsymbol{C}\hat{\boldsymbol{\beta}}_c + \boldsymbol{U}\hat{\boldsymbol{\beta}}_u + \boldsymbol{S}\hat{\boldsymbol{\beta}}_s \tag{23}$$

Loss used to train the model can be written as

$$loss = \frac{1}{n}||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||_2^2 + \lambda\mathcal{R}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}) \tag{24}$$

where n is the number of samples in training dataset, λ is the regularization strength and $\mathcal{R}(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{s}})$ is the regularization value according to different regularization methods. Further, the trained parameters are denoted as $\hat{\boldsymbol{\beta}}^{(\lambda)} = [\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}{}^{(\lambda)} \;\; \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}{}^{(\lambda)} \;\; \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}{}^{(\lambda)}]$. Some notations are listed as:

$$\boldsymbol{H} = [\boldsymbol{C} \;\; \boldsymbol{U} \;\; \boldsymbol{S}] \quad \boldsymbol{H}_{us} = [\boldsymbol{U} \;\; \boldsymbol{S}] \quad \hat{\boldsymbol{\beta}}_{us} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_u \\ \hat{\boldsymbol{\beta}}_s \end{bmatrix}$$

L1 regularization

Prove via counterexample, suppose $\hat{\boldsymbol{\beta}}_c = 0$ and $\boldsymbol{U}$ and $\boldsymbol{S}$ are perfectly correlated, then

$$\hat{\boldsymbol{Y}} = \boldsymbol{C}\hat{\boldsymbol{\beta}}_c + \boldsymbol{U}\hat{\boldsymbol{\beta}}_u + \boldsymbol{S}\hat{\boldsymbol{\beta}}_s = \boldsymbol{U}(\hat{\boldsymbol{\beta}}_u + \hat{\boldsymbol{\beta}}_s) \tag{25}$$

Then without regularization, the ordinary least square solution minimize

$$\min_{\hat{\boldsymbol{\beta}}_u, \hat{\boldsymbol{\beta}}_s} \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{U}(\hat{\boldsymbol{\beta}}_u + \hat{\boldsymbol{\beta}}_s)||_2^2 \tag{26}$$

Due to the perfect linear multicollinearity, there are infinitely many solutions for θ satisfying $\hat{\boldsymbol{\beta}}_u + \hat{\boldsymbol{\beta}}_s = \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the OLS estimate of the combined coefficient:

$$\boldsymbol{\theta} = \left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{Y} \tag{27}$$

A common choice in the unregularized case is to distribute the coefficient equally:

$$\hat{\boldsymbol{\beta}}_u^{(0)} = \hat{\boldsymbol{\beta}}_s^{(0)} = \frac{\boldsymbol{\theta}}{2} \tag{28}$$

With L1 regularization, the optimization problem becomes

$$\min_{\hat{\boldsymbol{\beta}}_u, \hat{\boldsymbol{\beta}}_s} \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{U}(\hat{\boldsymbol{\beta}}_u + \hat{\boldsymbol{\beta}}_s)||_2^2 + \lambda\left(||\hat{\boldsymbol{\beta}}_u||_1 + ||\hat{\boldsymbol{\beta}}_s||_1\right) \tag{29}$$

Since $\boldsymbol{U} = \boldsymbol{S}$, the model depends only on $\theta = \hat{\boldsymbol{\beta}}_u + \hat{\boldsymbol{\beta}}_s$, but the penalty depends on $||\hat{\boldsymbol{\beta}}_u||_1 + ||\hat{\boldsymbol{\beta}}_s||_1$. To minimize the penalty while keeping $\boldsymbol{\theta}$ constant, the optimal strategy is to set one coefficient to zero and assign all the weight to the other. According to triangle inequality,

$$||\hat{\boldsymbol{\beta}}_u||_1 + ||\hat{\boldsymbol{\beta}}_s||_1 \geq ||\hat{\boldsymbol{\beta}}_u + \hat{\boldsymbol{\beta}}_s||_1 = ||\boldsymbol{\theta}||_1 \tag{30}$$

Hence, when $\boldsymbol{\theta} > 0$:

$$\hat{\boldsymbol{\beta}}_u^{(\lambda)} = \boldsymbol{\theta},\ \ \hat{\boldsymbol{\beta}}_s^{(\lambda)} = \boldsymbol{0}\ \text{ OR }\ \hat{\boldsymbol{\beta}}_u^{(\lambda)} = \boldsymbol{0},\ \ \hat{\boldsymbol{\beta}}_s^{(\lambda)} = \boldsymbol{\theta} \tag{31}$$

Hence, when $\hat{\boldsymbol{\beta}}_u^{(\lambda)} = \boldsymbol{0},\ \ \hat{\boldsymbol{\beta}}_s^{(\lambda)} = \boldsymbol{\theta}$,

$$||\hat{\boldsymbol{\beta}}_s^{(\lambda)}||_1 = ||\boldsymbol{\theta}||_1 \geq ||\frac{\boldsymbol{\theta}}{2}||_1 = ||\hat{\boldsymbol{\beta}}_s^{(0)}||_1 \tag{32}$$

Therefore, L1 regularization does not reliably mitigate the shortcut term and, in some cases, may incorrectly assign all weights to the shortcut variables. □

EYE regularization

For EYE regularization, the loss function is

$$\begin{aligned} loss &= \frac{1}{n}||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||_2^2 + \lambda\left(||\hat{\boldsymbol{\beta_{us}}}||_1 + \sqrt{||\hat{\boldsymbol{\beta_{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}\right) && (33)\\ &= \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}||_2^2 + \lambda\left(||\hat{\boldsymbol{\beta_{us}}}||_1 + \sqrt{||\hat{\boldsymbol{\beta_{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}\right) && (34)\end{aligned}$$

Take derivative with respect to $\hat{\boldsymbol{\beta}}_c$ and $\hat{\boldsymbol{\beta}}_{us}$,

$$\frac{\partial loss}{\partial \hat{\boldsymbol{\beta}}_c} = -\frac{2}{n}\boldsymbol{C}^T\left(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) + \lambda\frac{\hat{\boldsymbol{\beta}}_c}{\sqrt{||\hat{\boldsymbol{\beta_{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}} \tag{35}$$

$$\frac{\partial loss}{\partial \hat{\boldsymbol{\beta}}_{us}} = -\frac{2}{n}\boldsymbol{H}_{us}^T\left(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) + \lambda\left(sign(\hat{\boldsymbol{\beta}}_{us}) + \frac{sign(\hat{\boldsymbol{\beta}}_{us})||\hat{\boldsymbol{\beta}}_{us}||_1}{\sqrt{||\hat{\boldsymbol{\beta_{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}}\right) \tag{36}$$

Then, at the optimum, the gradients are zero. Then,

$$\frac{2}{n}\boldsymbol{C}^T\left(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) = \lambda\frac{\hat{\boldsymbol{\beta}}_c}{\sqrt{||\hat{\boldsymbol{\beta_{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}} \tag{37}$$

$$\frac{2}{n}\boldsymbol{H}_{us}^T\left(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) = \lambda\left(sign(\hat{\boldsymbol{\beta}}_{us}) + \frac{sign(\hat{\boldsymbol{\beta}}_{us})||\hat{\boldsymbol{\beta}}_{us}||_1}{\sqrt{||\hat{\boldsymbol{\beta_{us}}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}}\right) \tag{38}$$

Suppose the inequality $|\hat{\boldsymbol{\beta}}_s^{reg}| \leq |\hat{\boldsymbol{\beta}}_s^{without\_reg}|$ always hold. Then to reduce the penalty, $\hat{\boldsymbol{\beta}}_c$ may be decreased, which leads to decreasing in $\sqrt{||\hat{\boldsymbol{\beta}_{us}}||_1^2 + ||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}$. Then $\frac{||\hat{\boldsymbol{\beta}}_{us}||_1}{\sqrt{||\hat{\boldsymbol{\beta}_{us}}||_1^2+||\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}||_2^2}}$ increases. According to the optimality condition for $\hat{\boldsymbol{\beta}}_{uc}$, $(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us})$ must also increase, which may require increasing $||\hat{\boldsymbol{\beta}}_{us}||_1$. Therefore, the inequality does not always hold. $\square$

L2, causal and causal effect regularization

The loss function with L2, causal and causal effect regularization can be formulated as

$$loss = \frac{1}{n}||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||_2^2 + \lambda||\boldsymbol{\mathcal{D}}_\lambda \hat{\boldsymbol{\beta}}_{us}||_2^2 \tag{39}$$

$$= \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}||_2^2 + \lambda||\boldsymbol{\mathcal{D}}_\lambda \hat{\boldsymbol{\beta}}_{us}||_2^2 \tag{40}$$

where $\boldsymbol{\mathcal{D}}_\lambda$ is a diagonal matrix whose diagonal entities are

- 1 for L2 regularization
- $1/P(\boldsymbol{H}_{us,i}$ is the cause of $\boldsymbol{Y})$ for causal regularization
- $1/|$Estimated treatment effect of $\boldsymbol{H}_{us,i}|$ for causal effect regularization

Hence the diagonal entities of $\boldsymbol{\mathcal{D}}_\lambda$ are all positive. Take derivative with respect to $\hat{\boldsymbol{\beta}}_c$ and $\hat{\boldsymbol{\beta}}_{us}$,

$$\frac{\partial loss}{\partial \hat{\boldsymbol{\beta}}_c} = -\frac{2}{n}\boldsymbol{C}^T\left(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) \tag{41}$$

$$= -\frac{2}{n}\left(\boldsymbol{C}^T\boldsymbol{Y} - \boldsymbol{C}^T\boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{C}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) \tag{42}$$

$$\frac{\partial loss}{\partial \hat{\boldsymbol{\beta}}_{us}} = -\frac{2}{n}\boldsymbol{H}_{us}^T\left(\boldsymbol{Y} - \boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) + 2\lambda\boldsymbol{\mathcal{D}}_\lambda^2\hat{\boldsymbol{\beta}}_{us} \tag{43}$$

$$= -\frac{2}{n}\left(\boldsymbol{H}_{us}^T\boldsymbol{Y} - \boldsymbol{H}_{us}^T\boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{H}_{us}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) + 2\lambda\boldsymbol{\mathcal{D}}_\lambda^2\hat{\boldsymbol{\beta}}_{us} \tag{44}$$

According to Karush-Kuhn-Tucker (KKT) condition, set the derivatives to 0, then

$$\frac{\partial loss}{\partial \hat{\boldsymbol{\beta}}_c} = 0 \Rightarrow \boldsymbol{C}^T\boldsymbol{Y} - \boldsymbol{C}^T\boldsymbol{C}\hat{\boldsymbol{\beta}}_c - \boldsymbol{C}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us} = 0 \tag{45}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_c = \left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\left(\boldsymbol{C}^T\boldsymbol{Y} - \boldsymbol{C}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) \tag{46}$$

$$= \left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{Y} - \left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us} \tag{47}$$

$$\frac{\partial loss}{\partial \hat{\boldsymbol{\beta}}_{us}} = 0 \Rightarrow \left(\frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{C}\hat{\boldsymbol{\beta}}_c \tag{48}$$

$$\Rightarrow \left(\frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{C}\left(\left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{Y} - \left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us}\right) \tag{49}$$

$$\Rightarrow \left(\frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{Y} + \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us} \tag{50}$$

Denote $\boldsymbol{\Pi}_c = \boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T$, then

$$\Rightarrow \left(\frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{\Pi}_c\boldsymbol{Y} + \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{\Pi}_c\boldsymbol{H}_{us}\hat{\boldsymbol{\beta}}_{us} \tag{51}$$

$$\Rightarrow \left(\frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2 - \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{\Pi}_c\boldsymbol{H}_{us}\right)\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{H}_{us}^T\boldsymbol{\Pi}_c\boldsymbol{Y} \tag{52}$$

$$\Rightarrow \left(\frac{1}{n}\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{Y} \tag{53}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{H}_{us} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)^{-1}\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{Y} \tag{54}$$

Because $\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{H}_{us}$ is symmetric and normal, it can be diagnolized as

$$\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{H}_{us} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \tag{55}$$

where $\boldsymbol{Q}$ is an orthonormal basis of $\mathbb{R}^{u+s}$, hence $\boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{Y}$ can be written as a linear combination of $\boldsymbol{Q}$, i.e.

$$\exists\ \boldsymbol{z} \text{ such that } \boldsymbol{H}_{us}^T\left(\mathbb{I}_c - \boldsymbol{\Pi}_c\right)\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{z} \tag{56}$$

Then,

$$\hat{\boldsymbol{\beta}}_{us} = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)^{-1}\boldsymbol{Q}\boldsymbol{z} \tag{57}$$

Because $\boldsymbol{Q}$ is orthonormal basis, let $\boldsymbol{z}' = \boldsymbol{Q}^T\boldsymbol{z}$, we can focus on the part

$$\frac{1}{n}\left(\frac{1}{n}\boldsymbol{\Lambda} + \lambda\boldsymbol{\mathcal{D}}_\lambda^2\right)^{-1}\boldsymbol{z}' \tag{58}$$

Then, for $i-$th entity,

$$\hat{\boldsymbol{\beta}}_{us,i} = \frac{n}{\frac{\Lambda_i}{n} + n\lambda\boldsymbol{\mathcal{D}}_{ii}^2} z_i' \tag{59}$$

Hence, increasing regularization strength λ increasing the positive denominator, which makes the absolute value of $\hat{\boldsymbol{\beta}}_{us,i}$ decrease. $\square$.

A.2 Proof for Proposition 2

Proposition 2: *Follow the problem setting, when $c = u = s = 1$, and S is correlated with Y by correlating with C and U, i.e. C and U are not co-linear, $S = \delta_c C + \delta_u U$, the regularization methods can eliminate the shortcuts under the following situations.*

(i) For L1 regularization: $\frac{\delta_c + \delta_u - 1}{\delta_c} \leq 0$

(ii) For L2 regularization: $\frac{\beta_c + \beta_c\delta_u^2 - 2\beta_u\delta_c\delta_u}{\delta_c^2 + \delta_u^2 + 1} \geq \beta_c$

(iii) For EYE regularization: $\frac{2\beta_c - 2\frac{\delta_c}{\delta_u - 1}\beta_u}{\left(\frac{\delta_c}{\delta_u - 1}\right)^2 + 1} \geq \beta_c$

(iv) For causal and causal effect regularization, denote the coefficients assigned to C, U and S as λ_c, λ_u and λ_s, the condition is $\frac{\beta_c + \frac{\lambda_u}{\lambda_s}\delta_u^2\beta_c - \frac{\lambda_u}{\lambda_s}\delta_u\delta_c\beta_u}{\frac{\lambda_c}{\lambda_s}\delta_c^2 + \frac{\lambda_u}{\lambda_s}\delta_u^2 + 1} = \beta_c$

Proof:

Given $c = u = s$, the true linear relationship between C, U and Y becomes

$$Y = \beta_c C + \beta_u U$$

Then the predicted values based on C, S and U are

$$\hat{Y} = \hat{\beta}_c C + \hat{\beta}_u U + \hat{\beta}_s S$$

The loss function becomes:

$$loss = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda\mathcal{R}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) \tag{60}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_c C_i + \beta_u U_i - \hat{\beta}_c C_i - \hat{\beta}_u U_i - \hat{\beta}_s S_i)^2 + \lambda \mathcal{R}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) \tag{61}$$

$$= \frac{1}{n}\sum_{i=1}^{n}((\beta_c - \hat{\beta}_c)C_i + (\beta_u - \hat{\beta}_u)U_i - \hat{\beta}_s(\delta_c C_i + \delta_u U_i))^2 + \lambda \mathcal{R}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) \tag{62}$$

$$= \frac{1}{n}\sum_{i=1}^{n}((\beta_c - \hat{\beta}_c)C_i + (\beta_u - \hat{\beta}_u)U_i - \hat{\beta}_s(\delta_c C_i + \delta_u U_i))^2 + \lambda \mathcal{R}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) \tag{63}$$

$$= \frac{1}{n}\sum_{i=1}^{n}((\beta_c - \hat{\beta}_c - \delta_c\hat{\beta}_s)C_i + (\beta_u - \hat{\beta}_u - \delta_u\hat{\beta}_s)U_i))^2 + \lambda \mathcal{R}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) \tag{64}$$

Select λ so that the empirical loss $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ becomes no larger after applying regularization. Because of the linear relationship between C, U and S, the empirical loss can be 0 when

$$\beta_c - \hat{\beta}_c - \delta_c\hat{\beta}_s = 0 \tag{65}$$

$$\beta_u - \hat{\beta}_u - \delta_u\hat{\beta}_s = 0 \tag{66}$$

Then

$$\hat{\beta}_s = \frac{\beta_c - \hat{\beta}_c}{\delta_c} \tag{67}$$

$$\hat{\beta}_u = \beta_u - \delta_u\frac{\beta_c - \hat{\beta}_c}{\delta_c} \tag{68}$$

Because the empirical loss can be 0, then for each regularization method, we can focus only on the regularization term $\mathcal{R}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s)$ when $\hat{\beta}_s$ and $\hat{\beta}_u$ follow equations (67) and (68).

L1 regularization

For L1 regularization, the regularization term becomes

$$\mathcal{R}_{L1}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) = |\hat{\beta}_c| + |\hat{\beta}_u| + |\hat{\beta}_s| \tag{69}$$

$$= |\hat{\beta}_c| + |\beta_u - \delta_u\frac{\beta_c - \hat{\beta}_c}{\delta_c}| + |\frac{\beta_c - \hat{\beta}_c}{\delta_c}| \tag{70}$$

Without loss of generality, assume $\hat{\beta}_c$, $\hat{\beta}_u$ and $\hat{\beta}_s$ are non-negative, then

$$\mathcal{R}_{L1}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) = |\hat{\beta}_c| + |\beta_u - \delta_u\frac{\beta_c - \hat{\beta}_c}{\delta_c}| + |\frac{\beta_c - \hat{\beta}_c}{\delta_c}| \tag{71}$$

$$= \hat{\beta}_c + \beta_u - \delta_u\frac{\beta_c - \hat{\beta}_c}{\delta_c} + \frac{\beta_c - \hat{\beta}_c}{\delta_c} \tag{72}$$

$$= \frac{\delta_c + \delta_u - 1}{\delta_c}\hat{\beta}_c + \beta_u + \frac{1 - \delta_u}{\delta_c}\beta_c \tag{73}$$

Because $\hat{\beta}_c$, $\hat{\beta}_u$ and $\hat{\beta}_s$ are non-negative, then $\hat{\beta}_c \in [\max\{0, \frac{\delta_c}{\delta_u}\beta_u\}, \beta_c]$. To make $\mathcal{R}$ achieves its minimum when $\hat{\beta}_s = 0$, equivalently $\hat{\beta}_c = \beta_c$, $\frac{\delta_c + \delta_u - 1}{\delta_c} \leq 0$. $\square$

EYE regularization

For EYE regularization, the regularization term becomes

$$\mathcal{R}_{EYE}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) = |\hat{\beta}_u| + |\hat{\beta}_s| + \sqrt{\left(|\hat{\beta}_u| + |\hat{\beta}_s|\right)^2 + \hat{\beta}_c^2} \tag{74}$$

$$\tag{75}$$

Similarly, assume $\hat{\beta}_c$, $\hat{\beta}_u$ and $\hat{\beta}_s$ are non-negative, then $\hat{\beta}_c \in [\max\{0, \frac{\delta_c}{\delta_u}\beta_u\}, \beta_c]$ and the regularization term:

$$\mathcal{R}_{EYE}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) = \hat{\beta}_u + \hat{\beta}_s + \sqrt{\left(\hat{\beta}_u + \hat{\beta}_s\right)^2 + \hat{\beta}_c^2} \tag{76}$$

$$= \beta_u - \delta_u \frac{\beta_c - \hat{\beta}_c}{\delta_c} + \frac{\beta_c - \hat{\beta}_c}{\delta_c} + \sqrt{(\beta_u - \delta_u \frac{\beta_c - \hat{\beta}_c}{\delta_c} + \frac{\beta_c - \hat{\beta}_c}{\delta_c})^2 + \hat{\beta}_c^2} \tag{77}$$

$$\frac{\partial \mathcal{R}_{EYE}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s)}{\partial \hat{\beta}_c} = \frac{\delta_u - 1}{\delta_c} + \frac{(\beta_u - \delta_u \frac{\beta_c - \hat{\beta}_c}{\delta_c} + \frac{\beta_c - \hat{\beta}_c}{\delta_c})(\frac{\delta_u - 1}{\delta_c}) + \hat{\beta}_c}{\sqrt{(\beta_u - \delta_u \frac{\beta_c - \hat{\beta}_c}{\delta_c} + \frac{\beta_c - \hat{\beta}_c}{\delta_c})^2 + \hat{\beta}_c^2}} \tag{78}$$

Then, by setting the derivative with respect to $\hat{\beta}_c$ to 0, we can get when the regularization term achieves minimum, which is

$$\hat{\beta}_c^{opt} = \frac{2\beta_c - 2\frac{\delta_c}{\delta_u - 1}\beta_u}{\left(\frac{\delta_c}{\delta_u - 1}\right)^2 + 1} \tag{79}$$

Then $\mathcal{R}_{EYE}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s)$ is decreasing within $[0, \hat{\beta}_c]$ and increasing within $[\hat{\beta}_c, \beta_c)$. Therefore, to make sure $\hat{\beta}_c$ achieves its minimum when $\hat{\beta}_s = 0$, i.e. $\hat{\beta}_c = \beta_c$, it is required that

$$\hat{\beta}_c^{opt} = \frac{2\beta_c - 2\frac{\delta_c}{\delta_u - 1}\beta_u}{\left(\frac{\delta_c}{\delta_u - 1}\right)^2 + 1} \geq \beta_c \tag{80}$$

L2, causal and causal effect regularization

For L2, causal and causal effect regularization, the regularization term becomes

$$\mathcal{R}_{L1}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s) = \lambda_c \hat{\beta}_c^2 + \lambda_u \hat{\beta}_u^2 + \lambda_s \hat{\beta}_s^2 \tag{81}$$

$$= \lambda_c \hat{\beta}_c^2 + \lambda_u \left(\beta_u - \delta_u \frac{\beta_c - \hat{\beta}_c}{\delta_c}\right)^2 + \lambda_s \left(\frac{\beta_c - \hat{\beta}_c}{\delta_c}\right)^2 \tag{82}$$

$$\frac{\partial \mathcal{R}_{L1}(\hat{\beta}_c, \hat{\beta}_u, \hat{\beta}_s)}{\partial \hat{\beta}_c} = 2\lambda_c \hat{\beta}_c + 2\lambda_u \frac{\delta_u}{\delta_c} \left(\beta_u - \delta_u \frac{\beta_c - \hat{\beta}_c}{\delta_c}\right) - 2\frac{\lambda_s}{\delta_c} \left(\frac{\beta_c - \hat{\beta}_c}{\delta_c}\right) \tag{83}$$

Hence, set the first derivative with respect to $\hat{\beta}_c$ to be 0 according to KKT condition.

$$\hat{\beta}_c = \frac{-\lambda_u \beta_u \frac{\delta_u}{\delta_c} + \lambda_u \beta_c (\frac{\delta_u}{\delta_c})^2 + \frac{\lambda_s}{\delta_c^2}\beta_c}{\lambda_c + \lambda_u (\frac{\delta_u}{\delta_c})^2 + \frac{\lambda_s}{\delta_c^2}} \tag{84}$$

$$= \frac{\lambda_s \beta_c + \lambda_u \beta_c \delta_u^2 - \lambda_u \beta_u \delta_u \delta_c}{\lambda_c \delta_c^2 + \lambda_u \delta_u^2 + \lambda_s} \tag{85}$$

$$= \frac{\beta_c + \frac{\lambda_u}{\lambda_s}\delta_u^2 \beta_c - \frac{\lambda_u}{\lambda_s}\delta_u \delta_c \beta_u}{\frac{\lambda_c}{\lambda_s}\delta_c^2 + \frac{\lambda_u}{\lambda_s}\delta_u^2 + 1} \tag{86}$$

Therefore, for L2, causal and causal effect regularization, the regularization can mitigate the shortcuts only when

$$\hat{\beta}_c = \frac{\beta_c + \frac{\lambda_u}{\lambda_s}\delta_u^2 \beta_c - \frac{\lambda_u}{\lambda_s}\delta_u \delta_c \beta_u}{\frac{\lambda_c}{\lambda_s}\delta_c^2 + \frac{\lambda_u}{\lambda_s}\delta_u^2 + 1} = \beta_c \tag{87}$$

B Experiments

Comprehensive experiments are conducted on three datasets, synthetic dataset, Colored-MNIST and MultiNLI to evaluate the effectiveness of different regularization methods. The codes will be made available during the review process.

B.1 Synthetic dataset

For synthetic dataset, we generated two known concepts C_1, C_2 and two unknown concepts U_1, U_2 and one shortcut variable S. The output label Y is generated by equation:

$$Y = 4C_1 - \frac{1}{2}C_2 + U_1 + 2U_2 \tag{88}$$

We design three types of shortcuts with different relationships to known and unknown concepts.

- S is only correlated with C_1 and C_2
$$S = 1.5C_1 - 0.5C_2$$
- S is correlated with U_1 and U_2
$$S = -0.5C_2 + U_1 + 2U_2$$
- S is correlated with Y
$$S = Y + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \text{ is some random noise}$$

In the training dataset, the shortcut variable is correlated with both known and unknown concepts. However, in the test dataset, the shortcut variable is generated by randomly sampling from a standard normal distribution. The correlations within training datasets and test datasets are shown in Figure 9. The figure illustrates that in the training datasets, the shortcut variable is designed to correlate with an increasing number of variables, while in the test dataset, it remains independent of all other variables.

Causal effect regularization. Regarding causal regularization, the weights corresponding to each variable are typically determined by treatment effect estimation. The performance of the treatment effect estimation can significantly impact the effectiveness of causal regularization. However, in our simulation experiments, we want to get rid of the effects of treatment estimation methods performance and only focus on the regularization itself. To achieve this, we manually assigned weights as the shortcut variables were known in our experimental setup. Specifically, we set the weights for known and unknown concepts to 1, while assigning 0.001 to the shortcut variables. This approach allowed us to focus solely on the behavior of the regularization.

B.1.1 Nonlinear case

If we relax the assumption of a linear relationship (Assumption 2), we can extend regularization methods to neural network models. For multi-layer neural networks, regularization can be applied to the parameters of the first layer. For example, consider our synthetic dataset where the input data contains five variables, assume the first layer of the neural network contains 10 neurons. The parameter matrix from the input data to the first layer would have a shape of $(10, 5)$. We can then apply regularization to these parameters, distinguishing between known and unknown concepts: let first two columns (assigned to known concepts) to be β_c and the second two columns (assigned to unknown concepts) to be β_u.

Since neural networks are black-box models, we cannot demonstrate whether the model mitigates shortcuts by examining the weights, as we do with linear models. Instead, we can treat neural networks as T-learners and compute the treatment effect of the shortcut variable. Ideally, if the shortcuts are mitigated, the estimated treatment effect should approach zero. The results for the synthetic datasets are presented in Figure 5.

B.1.2 Correlation with output label

To further evaluate the performance of various regularization methods, we calculate the correlations between predicted values and other variables, as shown in Figure 3.

Correlation with the shortcut variable: Consistent with the results in Figure 2, when regularization methods fail to mitigate shortcuts, the correlation between predicted values and the shortcut variable approaches 1, indicating heavy reliance on the shortcut. Moreover, as the correlation between the shortcut and other variables increases, the corresponding correlations with predicted values also rise.

Correlation with true output values: Similarly, the correlations between the predicted values and the true output values follow the same pattern. When the shortcuts are mitigated (i.e., correlation with the true concepts),

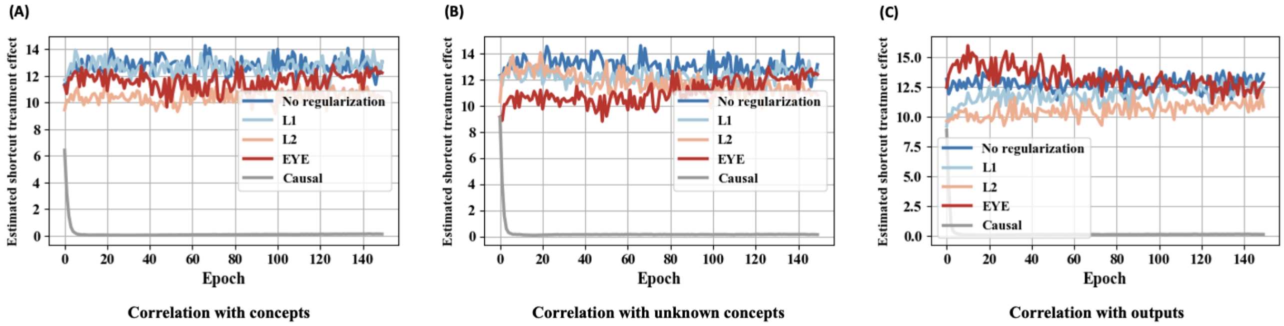

Figure 5: Treatment effect estimation of the shortcut variable along with training epoch when applying regularization terms to different shortcut types. **Neural networks might undermine the effectiveness of regularization methods.** (A) Shortcut S is only correlated with C_1 and C_2. (B) S is correlated with U_1 and U_2. (C) S is correlated with Y.

the correlations are close to 1, indicating near-perfect predictions. However, when regularization methods fail, these correlations drop significantly, showing that reliance on shortcuts leads to poorer performance on the test dataset.

B.1.3 Correlation with unknown concepts

In this experiment, the shortcut variable is generated based on $\boldsymbol{C}$ and $\boldsymbol{U}$, which makes the shortcut varibles correlated with U, i.e,

$$S = \delta_c C + \delta_u U + \epsilon$$

where ϵ is a standard noise. Changing δ_u changes the correlation between the shortcut variable and the unknown concept. For each correlation case, we apply all regulariztion methods. The results are shown in Table 1.

Correlation between U and S	No regularization	L1	L2	EYE	Causal
-0.0377 (training dataset)	0.0	0.0	1e-05	1e-05	3e-05
0 (test dataset)	0.7559	0.0	0.0001	0.1972	0.0001
0.0589 (training dataset)	0.0	0.0	0.0	1e-05	3e-05
0 (test dataset)	0.6589	0.0	0.1835	0.7066	0.0002
0.1539 (training dataset)	1e-05	1e-05	0.0	0.0	3e-05
0 (test dataset)	0.5503	0.0	0.5978	1.5746	0.0002
0.2448 (training dataset)	0.0	0.0	0.0	0.0	3e-05
0 (test dataset)	0.4373	0.0	1.019	2.7988	0.0002
0.3297 (training dataset)	0.0	0.0	0.0	0.0	3e-05
0 (test dataset)	0.3319	0.0	1.3178	4.3008	0.0002
0.4073 (training dataset)	0.0	0.0	0.0	0.0	3e-05
0 (test dataset)	0.2396	0.2395	1.48	5.8909	0.0002

Table 1: Mean square error of different correlations between unknown concepts (U) and shortcut variables (S).

From the Table 1, it is evident that as the correlation between U and S increases in the training dataset, the effectiveness of regularization methods in mitigating shortcuts diminishes significantly. This is reflected in the increasing discrepancy between the MSEs in the training and test datasets. These findings further demonstrate that regularization methods are likely to fail when shortcuts are highly correlated with unknown concepts. In addition, during this experiments, all the shortcuts and unknown concepts are not perfectly correlated (correlation less than 1).

B.1.4 Distribution shift

We evaluate the effectiveness of regularization methods in managing distribution of learned weights. The primary objective, for both synthetic and real-world datasets, is to determine whether regularization can effectively constrain weights within a specific range. For instance, we assess whether regularization methods can drive the weights associated with shortcut variables toward zero.

Over-regularization on known concepts is uncommon. Figure 6 illustrates the weight distributions for known and unknown concepts across three types of shortcut correlations. The first row shows the distributions for known concepts, while the second row focuses on unknown concepts. For known concepts (first row), when shortcuts are successfully mitigated (i.e., when the shortcut correlates only with the concepts), regularization methods broaden the range of weights assigned to the known concepts compared to no regularization, increasing their influence. However, when regularization fails to mitigate shortcuts, two scenarios arise:

1. If the shortcut correlates with unknown concepts, no adjustment is needed for the weights assigned to the known concepts because the shortcuts mainly come from the unknown concepts.
2. If the shortcut correlates with the output, only causal regularization with accurate causal effect estimations effectively broadens the weight range for the known concepts, preventing it from being concentrated near 0.

Over-regularization on unknown concepts is common. The second row in Figure 6 displays the weight distributions for unknown concepts. When the shortcut correlates with known concepts, no adjustment is needed for the weights of the unknown concepts (shown in panel A2), as they are not the source of the shortcut. However, when regularization fails to mitigate shortcuts, the shortcuts become correlated with unknown concepts in both cases. For example, in panel (B2), L1 and EYE regularization incorrectly constrain the weights of the unknown concepts to near zero, leading to over-regularization of the causal concepts. In contrast, in panel (C2), these methods do not present too much over-regularization of the unknown concepts because the correlation between the shortcut and the unknown concepts is weaker in this scenario.

B.2 Colored-MNIST & MultiNLI

B.2.1 Data generation and shortcut allocation

Colored-MNIST

Following Arjovsky et al. (2019), we first adapted the `MNIST` dataset (LeCun et al., 1998) by

1. Binarize the labels (digits 0-4 as class 0, 5-9 as class 1)
2. Set color labels s equal to y in the training set while reversing this in the test set. In training dataset, all the digits with class 0 are assigned with green colors and all the digits with class 1 are assigned with red colors. In test dataset, the assignment rule is reversed.

Thus, image color (green or red) acts as a shortcut for predicting the binary label. Additionally, we generated an unbiased dataset with random color assignments, ensuring that both colors are evenly distributed across both label classes. As a result, the model trained on this unbiased data is free from relying on shortcuts, which is further validated by its performance on the test dataset.

Two models were then trained: one for binary label prediction and another for color prediction. The last layer of the binary label model was used to extract concepts, while the last layer of the color prediction model was used to extract shortcuts. For unknown concept extraction, we utilized a pre-trained `resnet50` from `torchvision`. All concepts, unknown concepts, and shortcuts were represented with a dimension of 50. Since the last layer of the pre-trained `resnet50` has a dimension of 1000, we added a fully connected layer to reduce it from 1000 to 50, generating unknown concepts with the desired dimension. The parameters of this fully connected layer were initialized using a uniform distribution within the range $(-1, 1)$. Because this additional layer lacks an activation function, it preserves the linearity of the unknown features during further regularization training, which does not affect the regularization performance. Regarding the weights assigned to the variables in causal effect regularization, we set the weights for known and unknown concepts to 1, while assigning 0.001 to the

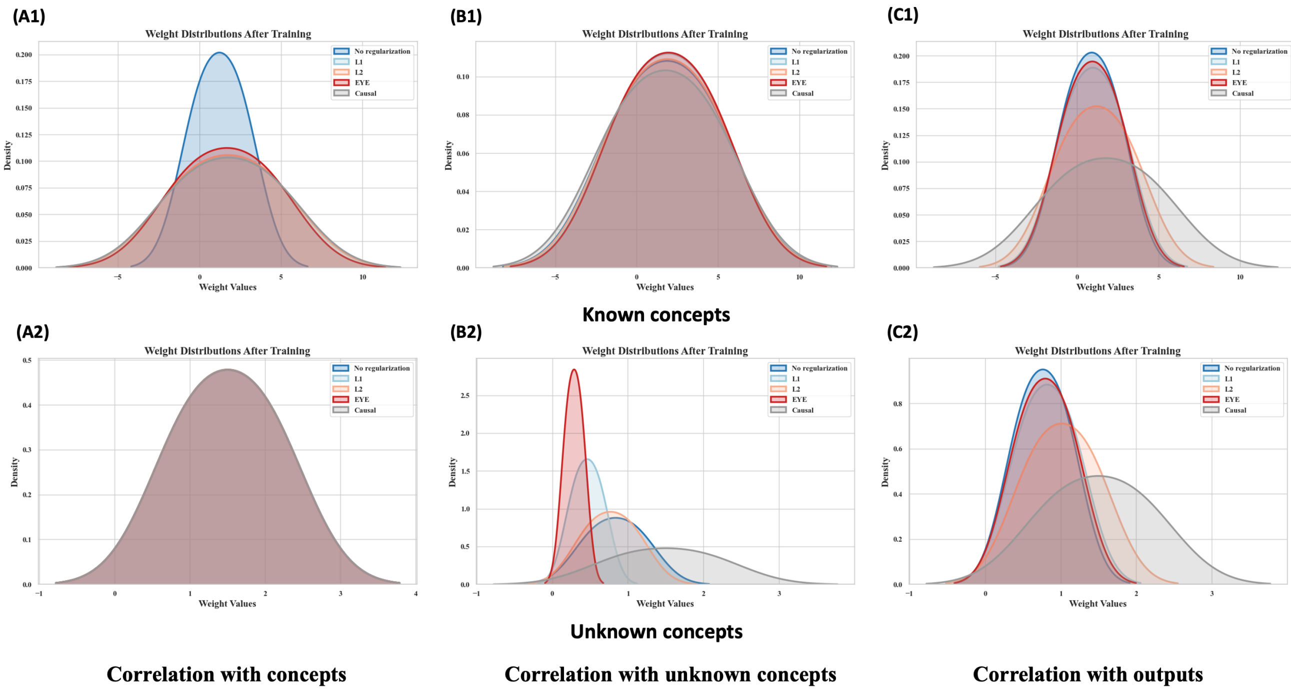

Figure 6: Distribution of learned weights across different regularization techniques for the synthetic dataset. The top row shows the distribution of weights for known concepts, while the bottom row shows the distribution for unknown concepts. Columns represent different shortcut correlations: with known concepts, unknown concepts, and the output label. **Over-regularization often happens on unknown concepts rather than known concepts**.

shortcut variables to remain consistent with the settings in the experiments of the synthetic datasets. This approach allowed us to focus solely on the behavior of the regularization.

MultiNLI

For the MultiNLI dataset (Williams et al., 2017), the task is to classify whether a hypothesis is entailed or contradicts a premise. We follow a similar approach to the shortcut allocation used in Colored-MNIST, with the following steps:

1. Exclude all records labeled as "neutral."
2. For the training dataset, select samples where the presence of negation in the second sentence corresponds to "contradiction," and for the test dataset, select samples where negation corresponds to "entailment." This reverse correlation allows us to assess the effectiveness of shortcut mitigation more clearly.

Therefore, having negation words in the second sentence acts as a shortcut for the classification. Similarly, we generated an unbiased dataset with random negation words assignments, ensuring that having or not having negation words are evenly distributed across both entailment and contradiction samples.

Two BERT-based models were trained: one for the entailment versus contradiction classification task, and another for detecting whether the second sentence contains negation words. The last layer of the first model was used to extract concepts, while the last layer of the second model was used to extract shortcuts, both with a dimension of 50. To extract unknown concepts, we employed a pre-trained `bert-base-uncased` model.

MIMIC-ICU

For the MIMIC-ICU dataset (Johnson et al., 2023) (`https://physionet.org/content/mimiciv/3.1/`), the task is to predict the length of stay (LOS) in intensive care unit (ICU). For the experiments, we identified 19 variables relevant to the prediction of LOS. These variables were categorized into known concepts and unknown concepts based on their correlations with LOS. Specifically, variables with a correlation greater than 0.1 with the output LOS were classified as known concepts, while the remaining variables were considered unknown concepts. This approach was necessary because there are no universally pre-trained models tailored for extracting information from this dataset. As a result, we identified 12 known concepts and 7 unknown concepts.

For shortcut variables, we generated 10 shortcuts derived from LOS. All shortcut variables were highly correlated with LOS in the training dataset but were randomly distributed in the test dataset to simulate spurious correlations. Consistent with the methodology of our other experiments, we applied various regularization methods to train a regression model on this dataset. Model performance was evaluated using the Mean Square Error (MSE).

Correlations within variables in the datasets are shown in Figure 7.

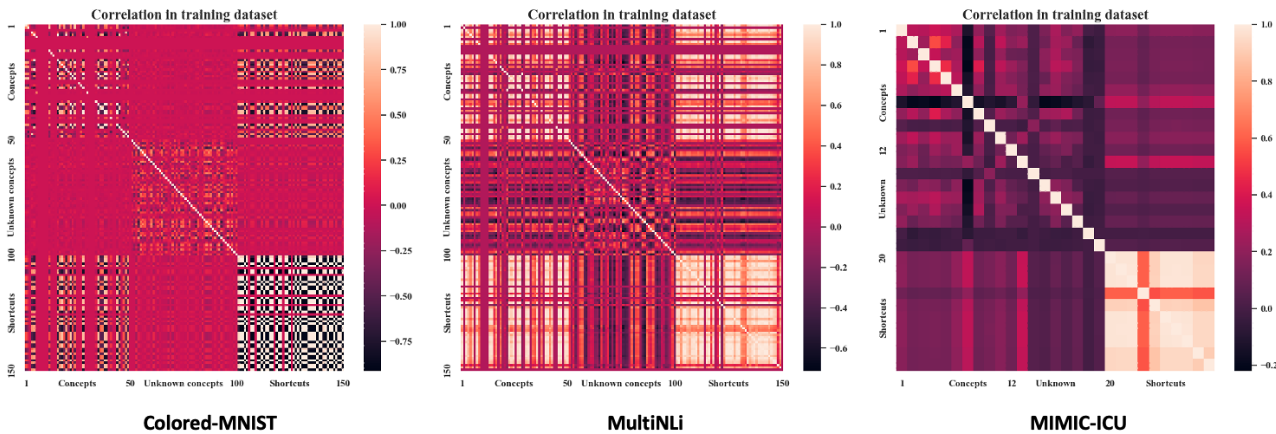

Figure 7: Correlation among the data for Colored-MNIST, MultiNLI and MIMIC-ICU datasets.

B.2.2 Distribution shift

For the real-world datasets, the distribution of learned weights for known concepts, unknown concepts, and shortcuts further shows the differences in how regularization methods mitigate shortcut learning. As illustrated in Figure 8, the weight distributions for Colored-MNIST and MultiNLI show the effectiveness of different regularization techniques in shortcut mitigation.

For both datasets, different regularization methods show varying performances in terms of learned weight distribution. For known concepts (panel A1), EYE regularization penalizes the known concepts the most, narrowing the weight distribution to near zero. In contrast, the other three regularization methods broaden the weight distribution, encouraging the model to rely more on the known concepts. However, for unknown concepts (panel A2 and B2), all regularization methods, except causal regularization, suppress their influence on the final predictions by concentrating the weights near zero. Correspondingly, these methods allow the weights of the shortcut variables to vary, indicating a reliance on shortcuts (panel A3 and B3).

Despite this, the test dataset's AUC performance for Colored-MNIST appears promising (Figure 4). This may be due to two factors:

- The known concepts provide sufficient information for accurately predicting the binary label.
- The task is simple enough that fluctuations in predicted probabilities caused by the use of shortcuts do not significantly affect the binary classification outcome.

Compared the AUC performance for MultiNLI (Figure 4), the performance is not as good as those for Colored-MNIST dataset, this might be because:

- The known and unknown concepts are highly correlated with the shortcut variables (Figure 7).
- The MultiNLI task is more complex than the binary classification in Colored-MNIST, meaning that the known concepts alone are insufficient for accurate classification, especially when the unknown concepts are also suppressed.

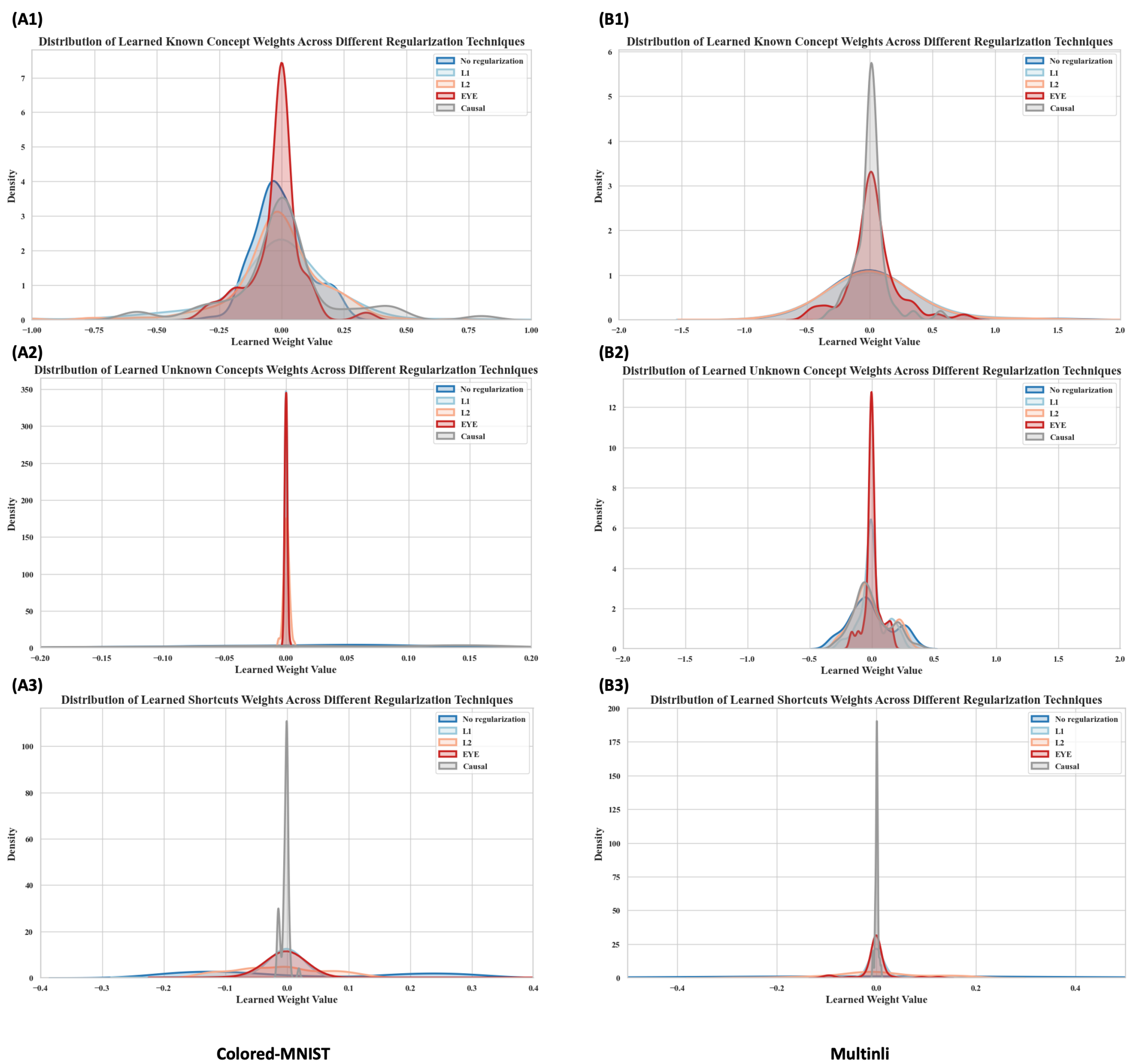

Figure 8: Distribution of learned weights across different regularization techniques for the Colored-MNIST and MultiNLI datasets. The top row shows the distribution of learned weights for known concepts, while the middle row shows the distribution for unknown concepts. The bottom row illustrates the distribution of learned weights for shortcuts. **Regularization methods tend to over regularize on unknown concepts and broaden the weight choices for the shortcut variables.**

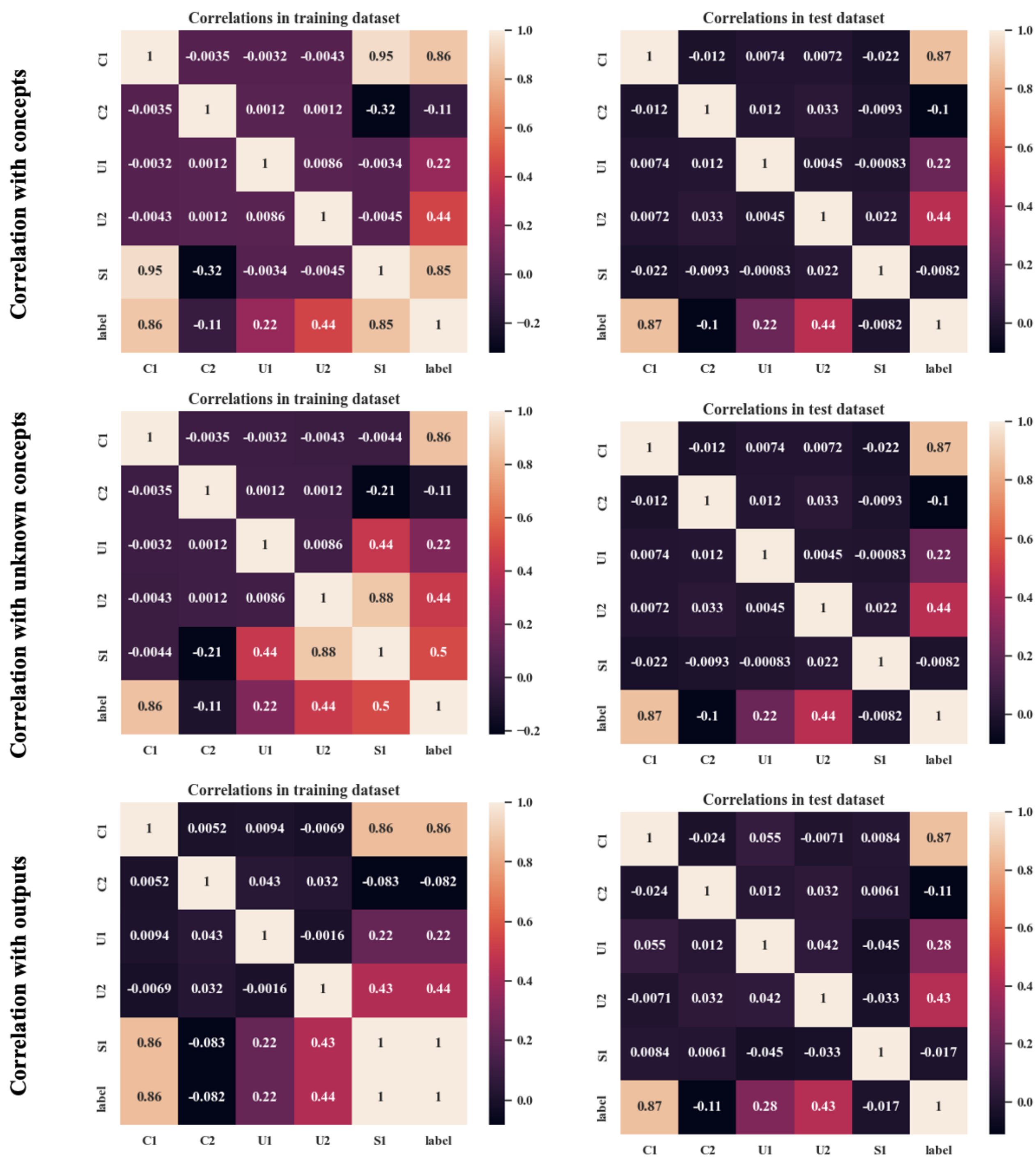

Figure 9: Correlations in the synthetic dataset for the training and test datasets. In the training dataset, the shortcut variable S_1 is designed to correlate with the output label by interacting with different variables, allowing the evaluation of regularization methods under varying conditions. In contrast, in the test dataset, the shortcut variable is independent of all other variables and the output label.