A Additional Theoretical Results and Proofs of the Covariate Shift Case

A.1 Theoretical Model Setup of the Covariate Shift Case

In this section, we will extend our theoretical model in the main text to the covariate shift setting. For the covariate shift setting, spurious features are independent of Y. Thus we can model the data generation process for environment e as

$$Y^e = \tilde{A^e}^k X_1 + n_1, \quad X_2^e = n_2 + \epsilon^e, \tag{14}$$

where the definition of n_1 and n_2 are the same as Section 3, ϵ^e represents environmental spurious features. ϵ_i^e (each dimension of ϵ^e) is a random variable that are independent for $i = 1, ..., N^e$. We assume the intra-environment expectation of the environment spurious variable is $\mathbb{E}_{\epsilon_i \sim p_e}[\epsilon_i] = \mu^e \in \mathbb{R}$ since spurious features are consistent in a certain environment. We further assume the cross-environment expectation $\mathbb{E}_e[\epsilon^e] = \mathbf{0}$ and cross-environment variance $\mathbb{E}_e[\epsilon_i^e] = \sigma^2$, $i = 1, ..., N^e$ for simplicity. This is consistent with the covariate shift case that $p(X)$ can arbitrarily change across different domains, and the support set of X may vary. Note that different from the concept shift setting, we only require $L \geq k$ to ensure the predictiveness of the network.

A.2 Theoretical Results of the Covariate Shift Case

In this subsection, we will present the failure case of VREx and IRMv1, and the success case of CIA under covariate shift (which is a different setting from the results of the concept shift case in the main text). The proofs are in Appendix E.2.

A.2.1 The Failure Case of VREx under Covariate Shift

Proposition A.1. ***(VREx will use spurious features)*** *The objective* $\min_\Theta \mathbb{V}_e[R(e)]$ *has non-unique solutions, and when part of the model parameters* $\{\theta_1^{1(l)}, \theta_1^{2(l)}, \theta_2^{1(l)}, \theta_2^{2(l)}\}$ *take the values*

$$\Theta_0 = \begin{cases} \theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-s+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-s, L-s-1, ..., 1 \\ \theta_2^{1(l)} = 0, \theta_2^{2(l)} = 1, & l = L-1, ..., 1 \end{cases}, \tag{15}$$

$0 < s < L$ *is some positive integer,* θ_1 *and* θ_2 *have four sets of solutions of the quadratic equation:*

$$\begin{cases} c_1\sigma^2(2\theta_1\theta_2 + (\theta_2)^2 - 2c_2\sigma^2\theta_2) + c_3 - \mathbb{E}_e[N^e]c_1\sigma^2\theta_1\theta_2 + \mathbb{E}_e[N^e]c_2\sigma^2\theta_2 = 0 \\ \left[c_3 - \mathbb{E}_e[N^e]c_1\sigma^2\theta_1\theta_2 + \mathbb{E}_e[N^e]c_2\sigma^2\theta_2\right] c_4 - c_5(\theta_2)^2 = 0 \end{cases}. \tag{16}$$

where $c_1 = \mathbb{E}[(\tilde{A^e}^s X_1)^\top (\tilde{A^e}^s X_1)]$, $c_2 = \mathbb{E}[(\tilde{A^e}^s X_1)^\top (\tilde{A^e}^k X_1)]$, $c_3 = \mathbb{E}_e[\epsilon^{e\top} \epsilon^e \epsilon^{e\top} (\tilde{A^e}^s X_1)]\sigma^2$, $c_4 = \mathbb{E}_e\left[(\tilde{A^e}^k X_1)^\top \mathbf{1_{N^e}}\right]\sigma^2$, $c_5 = \mathbb{E}_e\left[N^e\left(tr((\tilde{A^e}^k)^\top \tilde{A^e}^k) + N^e(1+\sigma^2)\right)\right]$.

Remark. For the covariate shift setting, $\theta_2 = 0$ is still not a solution to the VREx objective in node-level OOD tasks. Therefore it will also rely on spurious features.

A.2.2 The Failure Case of IRMv1 under Covariate Shift

Proposition A.2. ***(IRMv1 will use spurious features)*** *The objective* $\min_\Theta \mathbb{E}_e[\|\nabla_{w|w=1.0} R(e)\|^2]$ *has a solution that the invariant parameter* θ_1 *will produce inaccurate predictions,*

$$\theta_1 = \frac{\mathbb{E}_e[(\tilde{A^e}^k X_1)^\top (\tilde{A^e}^{2k} X_1)]}{\mathbb{E}_e[(\tilde{A^e}^{2k} X_1)^\top (\tilde{A^e}^{2k} X_1)]} \tag{17}$$

and there will be no constraints on the spurious parameter θ_2, when $\{\theta_1^{1(l)}, \theta_1^{2(l)}, \theta_2^{1(l)}, \theta_2^{2(l)}\}$ take the special values for some $0 < s < L$:

$$\Theta_0 = \begin{cases} \theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-s+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-s, L-s-1, ..., 1 \\ \theta_2^{1(l)} = 0, \theta_2^{2(l)} = 1, & l = L-1, ..., 1 \end{cases} . \tag{18}$$

A.2.3 THE SUCCESSFUL CASE OF CIA UNDER COVARIATE SHIFT

Proposition A.3. *Optimizing the CIA objective will lead to the optimal solution Θ^*:*

$$\begin{cases} \theta_1 = 1 \\ \theta_2 = 0 \\ \theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-k+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-k, L-k-1, ..., 1 \end{cases} . \tag{19}$$

A.3 ADDITION DISCUSSION ON THE SPECIAL FAILURE VALUE

B DETAILED EXPERIMENTAL SETUP

B.1 BASIC SETTINGS

All experimental results were averaged over three random runs. Following (Gui et al., 2022), we use an OOD validation set for model selection and use a 3-layer GCN (Kipf & Welling, 2016) as the backbone GNN, except that Mixup uses a modified GCN. The settings for learning rate, batch size, and training epochs also follow (Gui et al., 2022).

B.2 HYPERPARAMETER SETTINGS

Most hyperparameter settings are adopted from (Gui et al., 2022), except that for EERM we reduce the number of generated environments from 10 to 7 and reduce the number of adversarial steps from 5 to 1 for memory and computing complexity concerns. For each parameter of the methods, we conduct a grid search for about 3$\sim$4 values.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 PARAMETER ANALYSIS

In this section, we analyze the effect of λ and the number of adjacent hops of LoRe-CIA. From Figure 2, we can see clearly that adding LoRe-CIA regularization is beneficial for generalization since the test accuracy increases with λ. Note that most of the parameter combinations outperform the baseline methods (ERM: 55.78/60.24, IRM: 55.77/61.23, VREx: 55.97/60.69), indicating that our method leads to consistently superior performance.

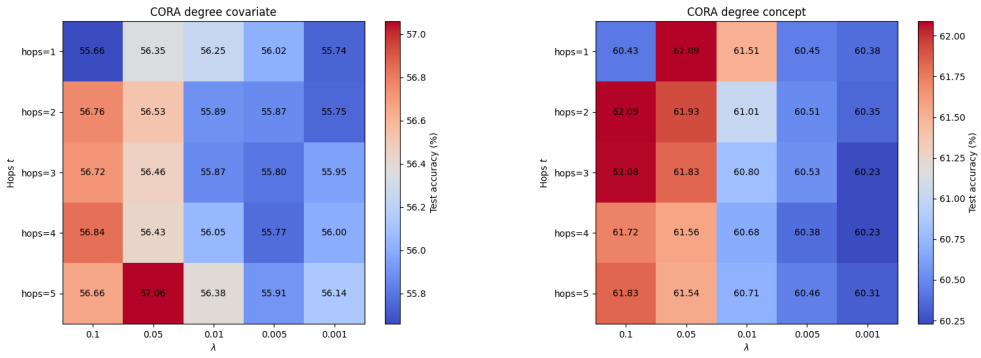
On Cora degree covariate shift, we can observe the positive effect of localized alignment: with the decrease of t, the accuracy rate increases gradually. However, the trend is not clear in the covariate shift. On covariate shift, the accuracy rate varies differently with t for different λ, indicating that there is a synergistic effect on the accuracy rate. How to better balance these two parameters is a direction worth exploring in the future.

C.2 REPRESENTATION VISUALIZATION

We visualize the representation of CIA and LoRe-CIA in Figure 3 to show that LoRe-CIA can alleviate feature collapse caused by overalignment.

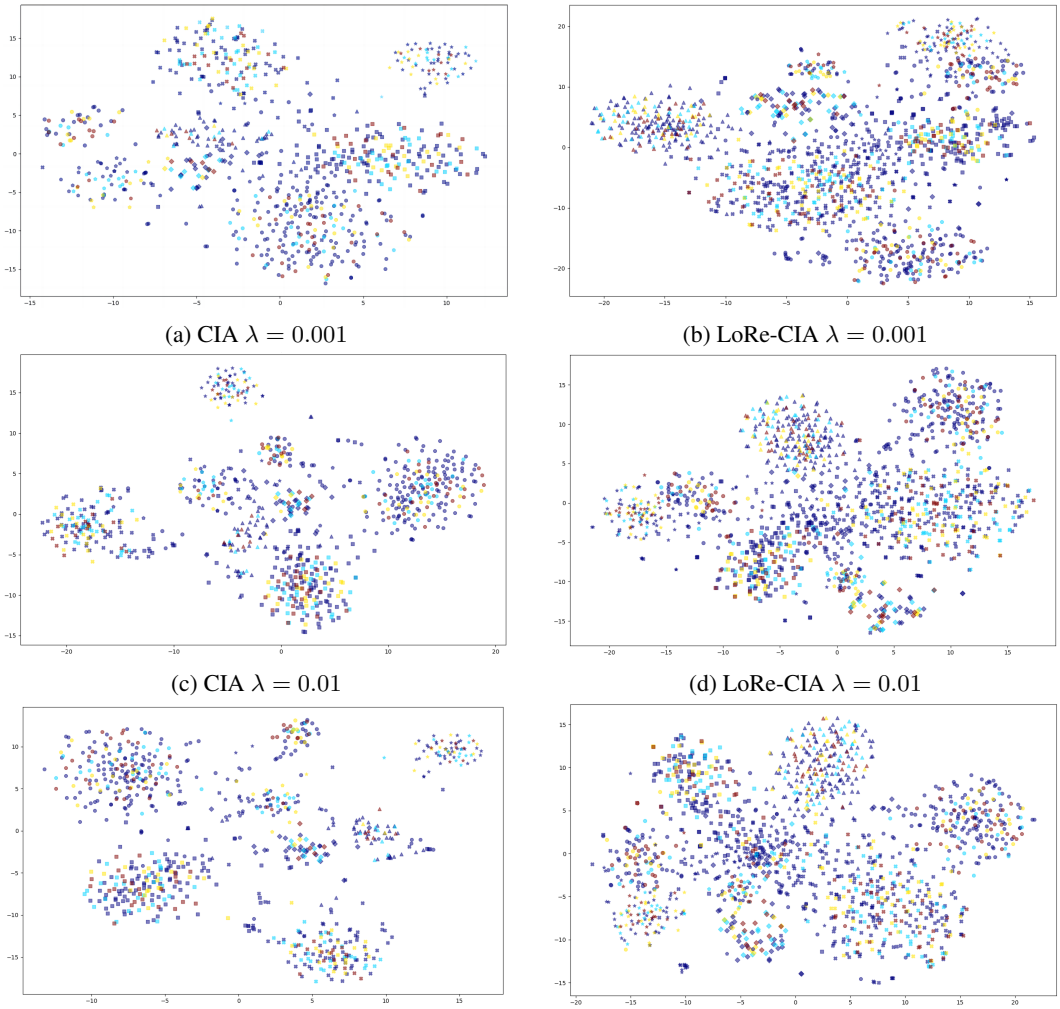
Table 4: Hyperparameter setting of the experiments.

Algorithm	Search Space
IRM	0.1, 1, 10, 100
VREx	1, 10, 100, 1000
GroupDRO	0.001, 0.01, 0.1
DANN	0.001, 0.01, 0.1
Deep Coral	0.01, 0.1, 1
Mixup	0.4, 1.0, 2.0
EERM	β=0.5, 1, 3 number of generated environments k=7 adversarial training steps t=1 numbers of nodes for each node should be modified the link with s=5 subgraph generator learning rate r=0.0001, 0.001, 0.005, 0.01
SRGNN	0.000001, 0.00001, 0.0001
CIA	λ=0.0001, 0.001, 0.005, 0.01, 0.05, 0.1
LoRe-CIA	λ= 0.001, 0.005, 0.01, 0.05, 0.1 hops t=2, 3, 4, 5



(a) Cora degree concept (b) Cora degree covariate

Figure 2: Effect of λ and the number of hops on OOD test accuracy (%).



(a) CIA $\lambda = 0.001$

(b) LoRe-CIA $\lambda = 0.001$

(c) CIA $\lambda = 0.01$

(d) LoRe-CIA $\lambda = 0.01$

(e) CIA $\lambda = 0.1$

(f) LoRe-CIA $\lambda = 0.1$

Figure 3: Visualization of the learned representations of nodes. LoRe-CIA can prevent the features of each class from being too concentrated.

C.3 VALIDATION OF THE TRUE FEATURE GENERATION DEPTH

For the theoretical model in section 3, we assume $L \geq k$. To empirically find out how large k really is, we use GCN with different layer to predict the ground-truth label Y on Cora and Arxiv dataset respectively (results are in Table 5 and 6). As mentioned above, since a GCN with layer l will aggregate features from l-hop neighbors for prediction, if the depth of the GCN is equal to the true generation depth, then the performance should be close to optimal. Suppose the empirical optimal layer number is L^* for prediction, we have: $L^* = k$ **We find that the $L_s^* \leq 4$ in most cases (even on large-scale graphs in Arxiv).** This indicates that our assumptions holds easily.

Table 5: OOD accuracy (%) of GCN with different numbers of layers on Cora.

Dataset	Shift	$L = 1$	$l = 2$	$L = 3$	$L = 4$
Cora (degree)	covariate	**59.04(0.15)**	58.44(0.44)	55.78(0.52)	55.15(0.24)
	concept	**62.88(0.34)**	61.53(0.48)	60.24(0.40)	60.51(0.17)
Cora (word)	covariate	64.05(0.18)	**65.81(0.12)**	65.07(0.52)	64.58(0.10)
	concept	64.76(0.91)	**64.85(0.10)**	64.61(0.11)	64.16(0.23)

Table 6: OOD accuracy on causal prediction (%) of GCN with different numbers of layers on Arxiv.

Dataset	Shift	$l = 2$	$L = 3$	$L = 4$	L=5
Arxiv (degree)	covariate	57.28(0.09)	58.92(0.14)	**60.18(0.41)**	60.17(0.12)
	concept	63.32(0.19)	62.92(0.21)	**65.41(0.13)**	63.93(0.58)
Arxiv (time)	covariate	71.17(0.21)	70.98(0.20)	**71.71(0.21)**	70.84(0.11)
	concept	65.14(0.12)	67.36(0.07)	65.20(0.26)	**67.49(0.05)**

C.4 DISCUSSION AND VALIDATION OF THE ASSUMPTION ON THE RATE OF CHANGE OF CAUSAL AND SPURIOUS FEATURES W.R.T SPATIAL POSITION

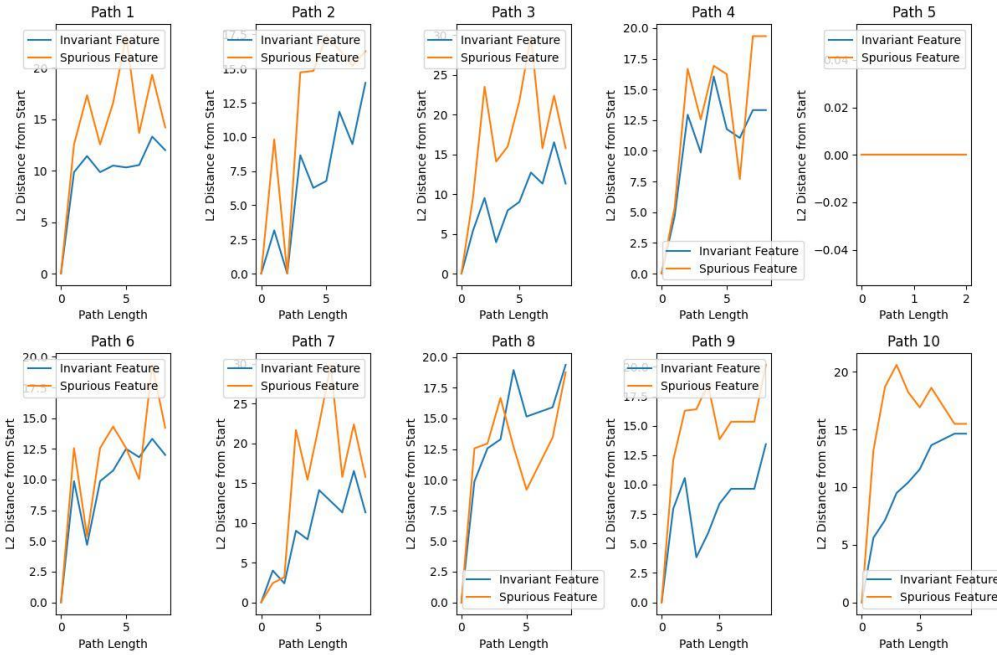
To verify the intuition used in Section 4.2 that the change rate of node's spurious features w.r.t spatial location is faster than that of the causal/invariant features within a certain range of hops, we conduct experiments on GOOD-Arxiv and GOOD-Cora, both are real-world citation networks. To extract invariant features, we use a pretrained VREx model and take the output of the last layer as invariant features[3]. To obtain spurious features, we train a ERM model to predict the environment label and also take the output of the last layer as spurious features. For each class, we randomly sample 10 nodes and generate corresponding 10 paths using Breadth-First Search (BFS). We extract invariant and spurious features of the nodes on each paths, and plot the distances between the node representations on the paths and the starting node. The results of Cora are in Figure 4 and 5, and the results of Arxiv are in Figure 6 and 7. (we choose some of the classes to avoid excessive paper length, the results for the other classes are similar).

We can see that within about 5$\sim$10 hops, the changes of spurious features grow more rapidly than invariant ones. Hence we propose to align the representations of adjacent nodes to better eliminate spurious features and avoid the collapse of the invariant features. And this explains we add a weighting term $d(i, j)$ in our loss function to assign smaller weight node pairs farther apart.

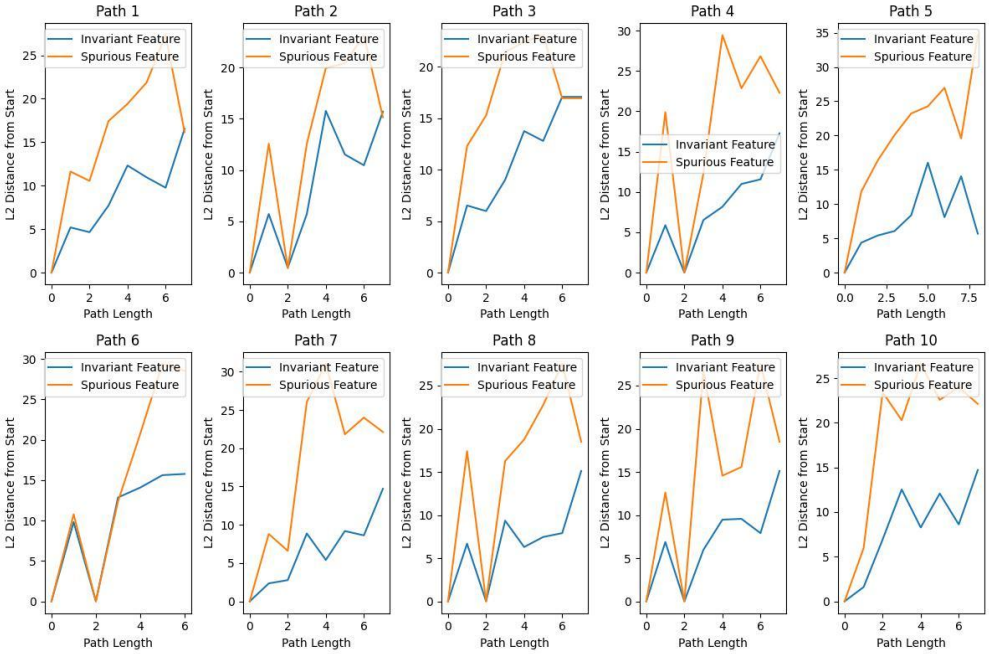
This assumption is similar to the ones adopted by a series of previous works on causality and invariant learning (Chen et al., 2022; Burshtein et al., 1992; Schölkopf, 2022; Schölkopf et al., 2021).

[3]though we reveal in our theory that VREx could rely on spurious features, we still use VREx here to approximately extract invariant features as many previous graph OOD works did since VREx already gains some advantages.

They assume causal features are more well-clustered than spurious features. In node-level graph OOD scenario, we observe this phenomenon only within local parts of a graph. In some cases, when two nodes are too far away from each other, their causal features can also vary more than the spurious features, as can been seen in Figure 7 (a) path 1,2,4,6,9 and 10. Therefore, choosing to match the representations in a local region can help to alleviate the feature collapse problem.

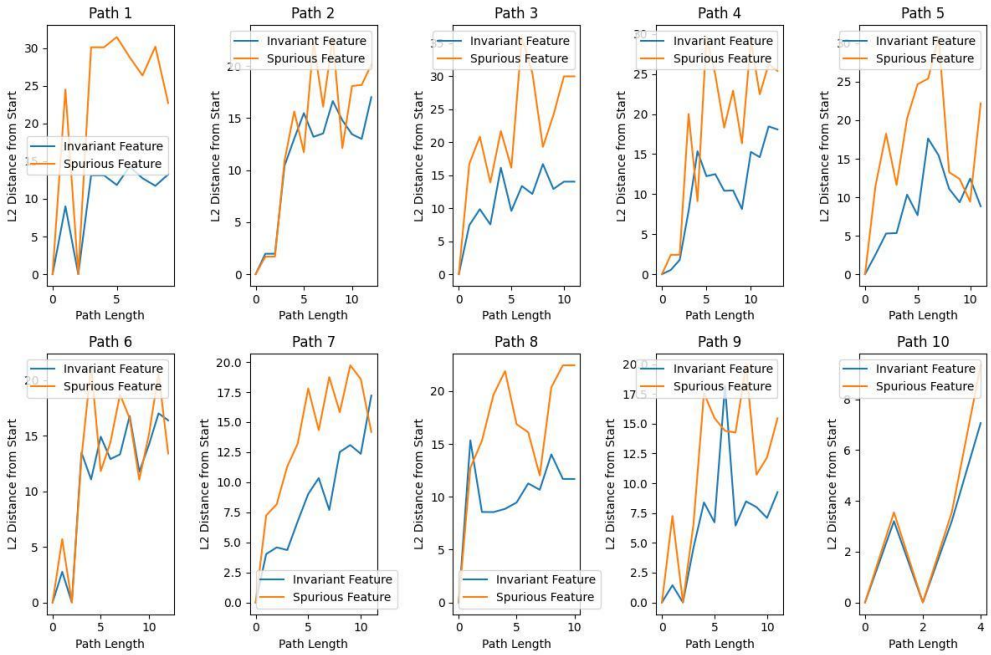


(a) class 16 of Cora

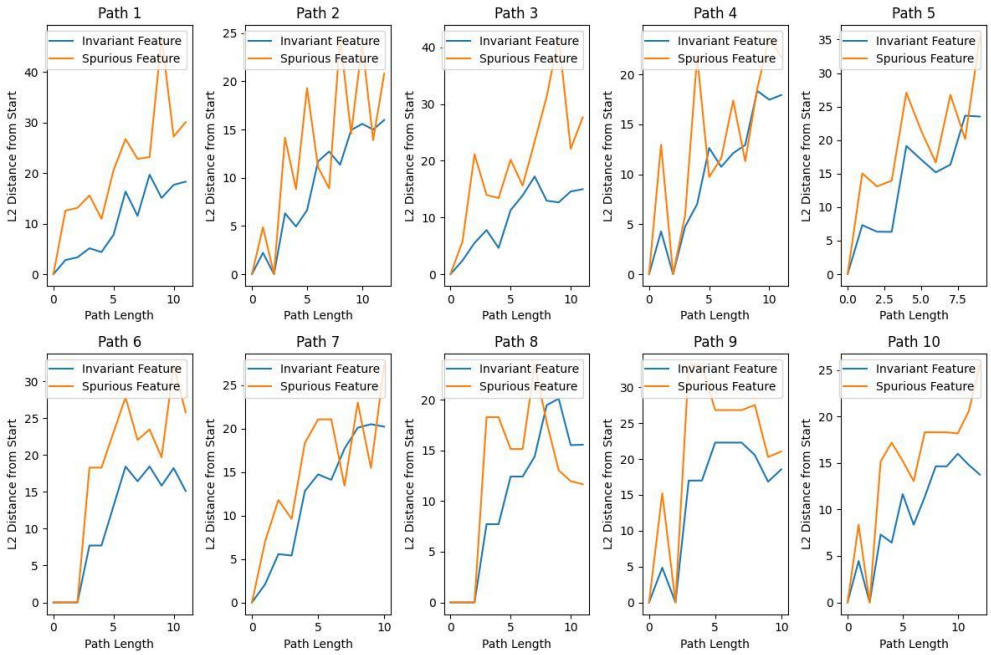


(b) class 17 of Cora

Figure 4: Visualization of the rate of change of invariant features and spurious features on Cora (part 1).

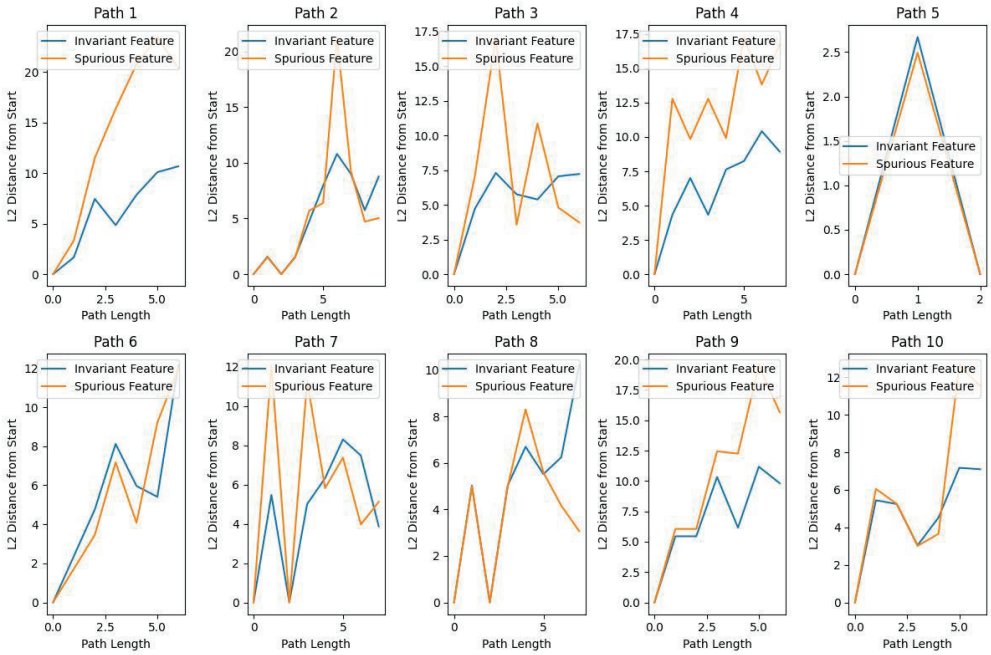


(a) class 39 of Cora

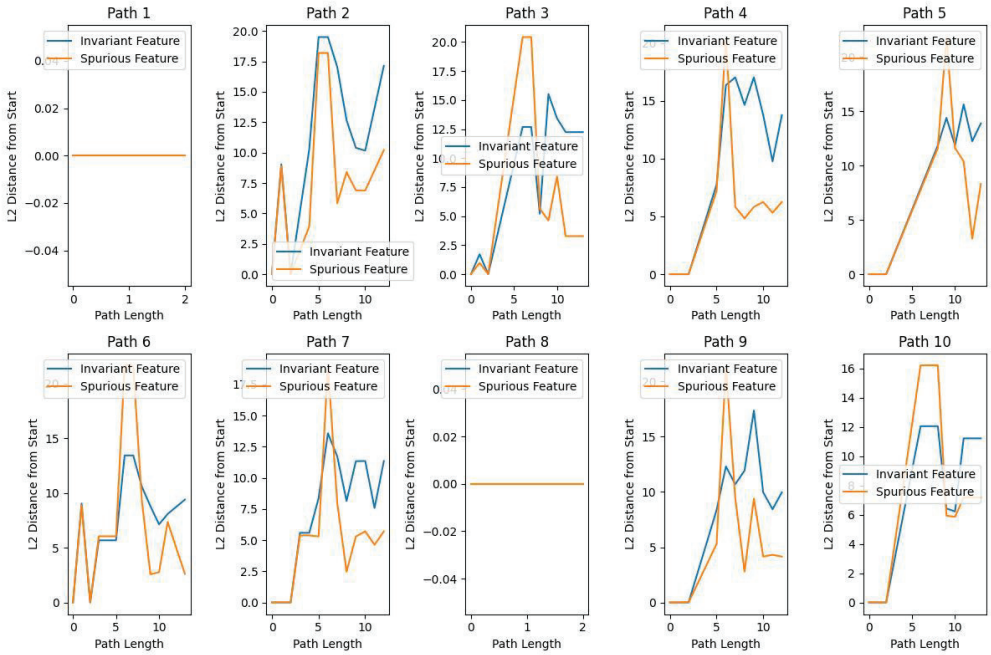


(b) class 41 of Cora

Figure 5: Visualization of the rate of change of invariant features and spurious features on Cora (part 2).

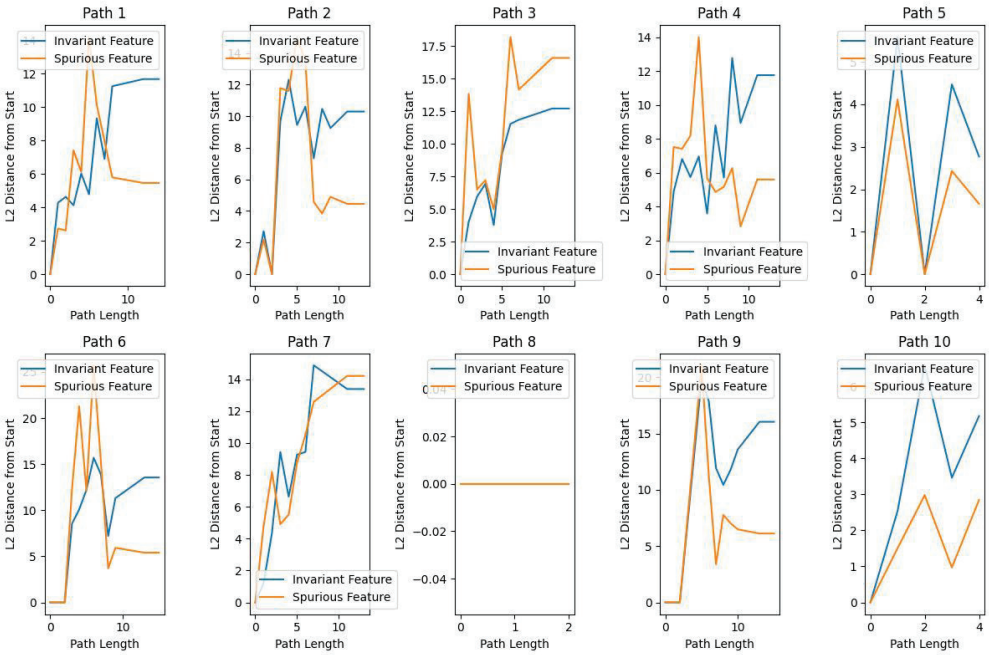


(a) class 25 of Arxiv

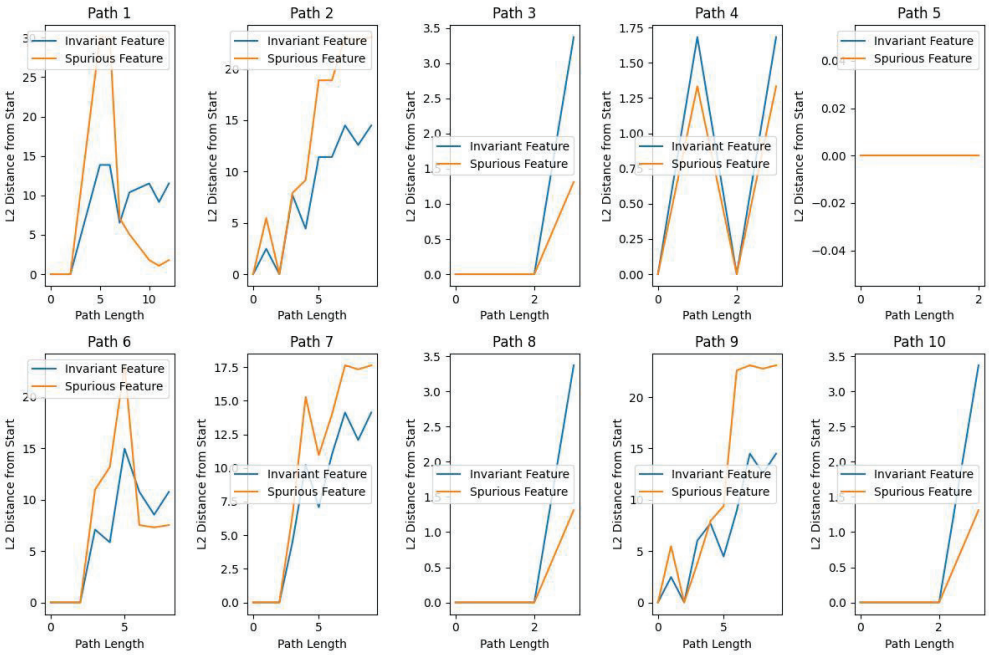


(b) class 29 of Arxiv

Figure 6: Visualization of the rate of change of invariant features and spurious features on Arxiv (part 1).



(a) class 13 of Arxiv



(b) class 17 of Arxiv

Figure 7: Visualization of the rate of change of invariant features and spurious features on Arxiv (part 2).

C.5 Discussion and Validation of the Assumption on the Feature Distance and Neighboring Label Distribution Discrepancy

C.5.1 Class-different Neighboring Labels Reflect Spurious Feature Distribution

In this section, we will empirically validate the key intuition of LoRe-CIA: the label distribution of the neighbors from different classes (which we call *Heterophilous Neighboring Label Distribution (HNLD)* in the following contents) reflects the spurious feature of the centered node. This idea is similar to the observation in Song & Wang (2022): heterophily is a main source of node distribution shift. Moreover, as recommended by Ye et al. (2022), we will further investigate the impact of HNLD on spurious feature distribution under two types of OOD shift: *concept shift* (or correlation shift in (Ye et al., 2022)), where $p(Y|X)$ varies across environments, and *covariate shift* (or diversity shift in (Ye et al., 2022)), where $p(X)$ changes with environments, respectively. We will show that HNLD affect the spurious features of the centered node in different manners under concept shift and covariate shift.

Spurious features represent features that have no predictive power for labels, and spurious features of a node come from two sources: (1) the environmental spurious feature, i.e. features determined by environments that contain no invariant and predictive information about labels, (2) class-different (heterophilous) neighboring features. The first source of spurious features is mentioned all the time in OOD and Domain Generalization (DG) topics, and many recent works have revealed that heterophilous neighbors harm node classification performance (Ma et al., 2021; Huang et al., 2023). In the follow part, we will first point out how to approximately measure spurious features for covariate and concept shift, and empirically validate our intuition.

Covariate shift. For covariate shifts on graphs, since spurious features are not necessarily correlated with labels, the environmental spurious features cannot be reflected by HNLD. However, we can still measure the distribution of the spurious features caused by heterophilous neighbors. To extract spurious features induced by class-different labels, we train a 1-layer GCN that aggregates neighboring features and discards the features of the centered node. The reason why we use features from all neighbors rather than only heterophilous neighbors is we want to simulate message-passing as authentically as possible, that is, we hope to observe whether the gap of HNLD accurately reflects the distance of heterophilous neighboring feature in the presence of both homophilous and heterophilous neighbors. To ensure that the discrepancy in the aggregated neighboring feature is caused solely by heterophilous neighbors, we only use point pairs with the same number of homophilous neighbors. Specifically, we compute the L2 distance between the neighbor representations of two nodes with the same number of class-same neighbors, and plot its trend w.r.t. the distance of HNLD (according to the definition of $Q_{i,j}^{\text{diff}}$ in Equation 13). We run experiments on Cora to verify this. We evaluate on both *word* shifts (node feature shifts) and *degree* (graph structure shifts) for a comprehensive understanding. We show the results of first 30 classes of Cora. **The results in Figure 8 and 9 show a clear positive correlation between the spurious feature distance and HNLD discrepancy under covariate shifts.**

Concept shift. As for concept shift, spurious features are correlated with labels, thus the label of a node contains information about spurious features correlated with this class. Moreover, due to the massage-passing mechanism of GNNs, the spurious features of a centered node are also affected by neighboring nodes. Assuming that most adjacent nodes are from the same environment, the spurious features of same-class neighbors will not change that of the centered node since the spurious distribution is fixed given the class and the environment (Yi et al., 2022). Hence, by observing HNLD, we can measure the distribution of the spurious feature. For concept shift, we train a GNN to predict environment labels to obatin spurious representations. **Table 10 and 11 also show a clear positive correlation between spurious featured distance and HNLD discrepancy on concept shift.**

C.5.2 Class-same Neighboring Labels Reflect Invariant Feature Distribution

Now will validate that the label distribution of the neighbors from the same class as the centered node reflects the invariant feature of the centered node. We use VREx to approximate invariant features, and compute the their distance w.r.t. the discrepancies of the neighboring label distribution of the same class. We evaluate on 4 splits of Cora: *word*+**covariate**, *word*+**concept**, *degree*+**covariate**

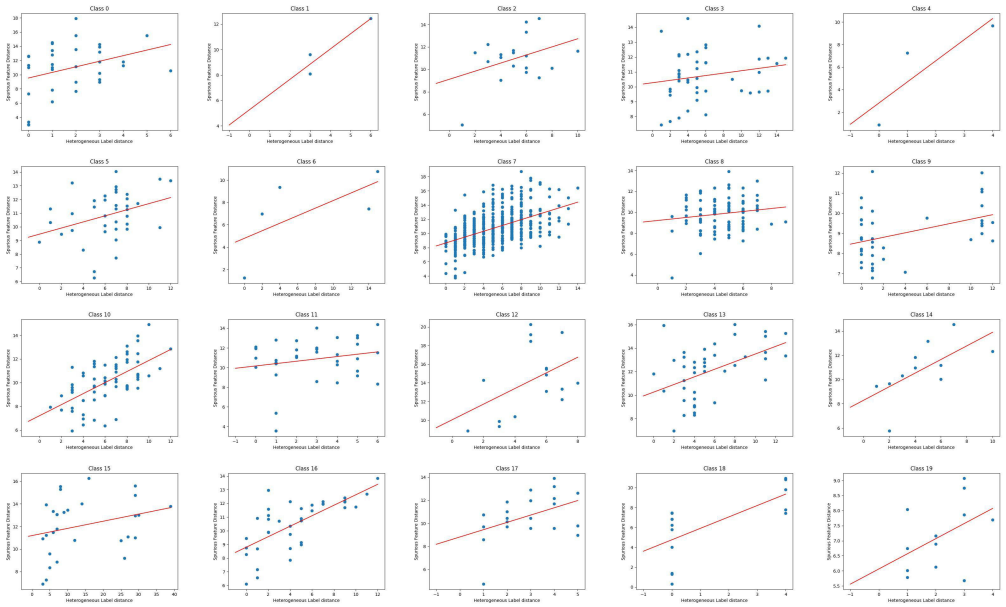


Figure 8: The relationship between the distance of spurious features induced by class-different neighbors and distance of HNLD on Cora *word* domain, **covariate shift**. Each sub-figure is a class, and each dot in the figure represents a node pair in the graph. The red line is obtained by linear regression. The positive correlation is clear.

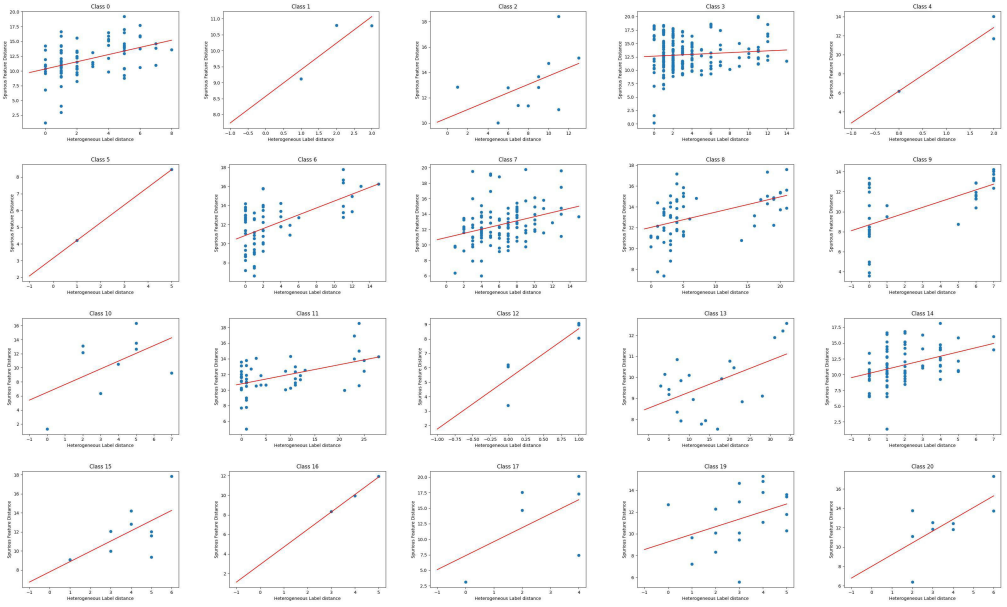


Figure 9: The relationship between the distance of spurious features induced by class-different neighbors and distance of HNLD on Cora *degree* domain, **covariate shift**. The positive correlation is clear.

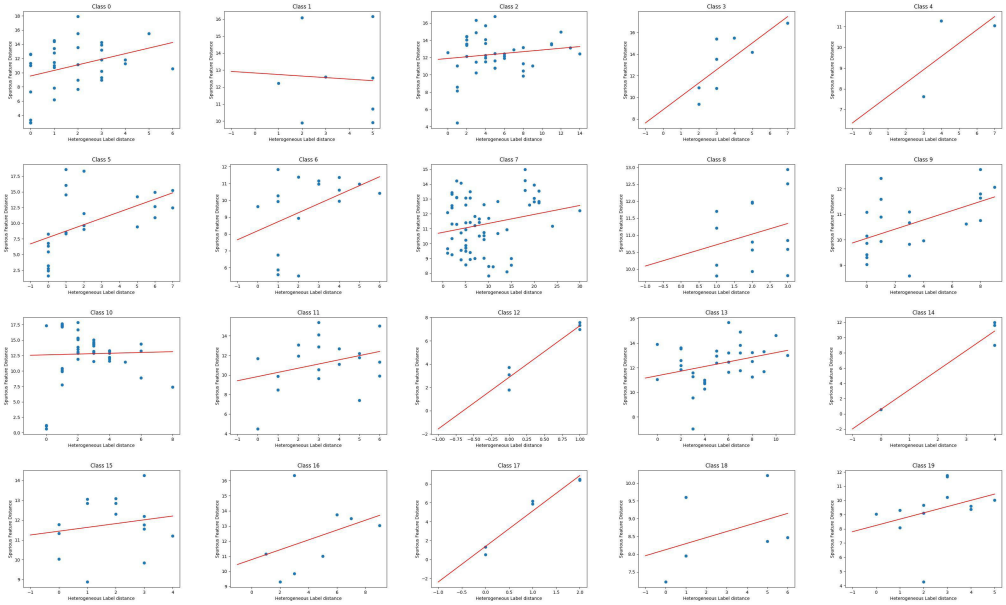


Figure 10: The relationship between the distance of environmental spurious features and distance of HNLD on Cora *word*, **concept shift**. The positive correlation is clear.

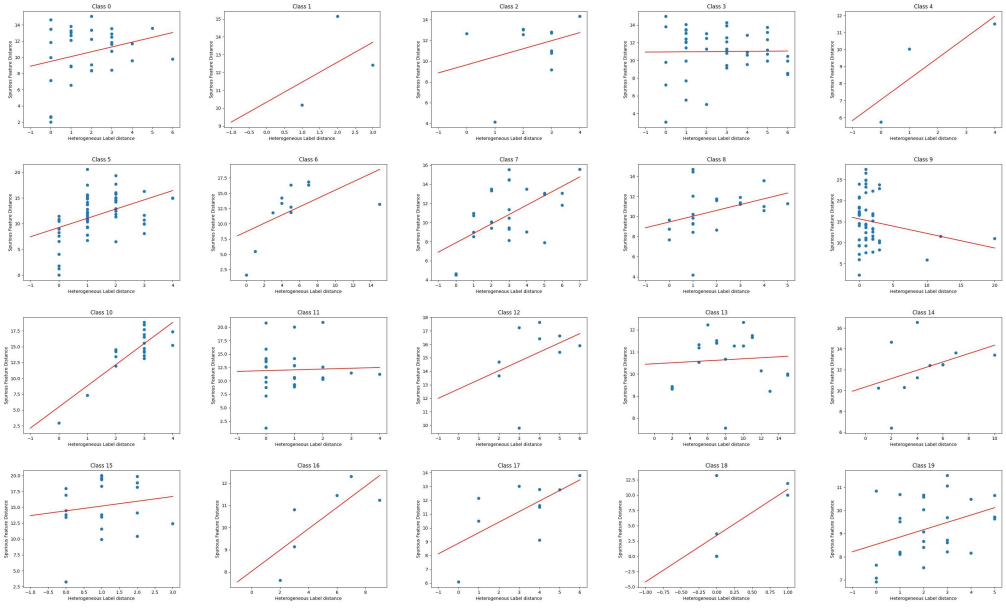


Figure 11: The relationship between the distance of environmental spurious features and distance of HNLD on Cora *degree*, **concept shift**. The positive correlation is clear.

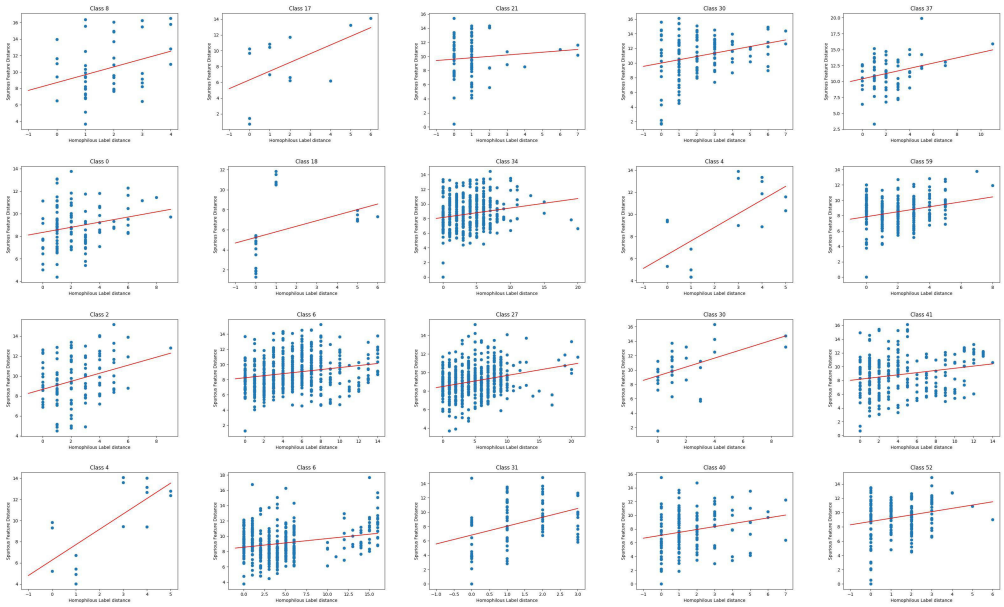


Figure 12: The relationship between the distance of invariant features and discrepancy of class-same neighboring labels on Cora *degree*, **concept shift**. Line 1 to 4 are results of Cora *word*+**covariate**, *word*+**concept**, *degree*+**covariate** and *degree*+**concept**, respectively. There is a positive correlation between the invariant feature distance and difference in neighboring labels of the same class as the centered node.

and *degree*+**concept**. For each data split, we randomly choose 5 classes that have node pairs with difference of larger than 5 in class-same neighboring labels. **The results in Table 12 also show a positive correlation trend.**

D Detailed Training Process

Table 1 show the detailed training process of LoRe-CIA.

E Proofs

E.1 Proofs of the Concept Shift Case Presented in the Main Text

In this section, we give proof of the propositions of the concept shift model presented in the main text.

E.1.1 Proof of the Failure Case of VREx under Concept Shift

Proposition E.1. ***(VREx will use spurious features)*** *The objective* $\min_\Theta \mathbb{V}_e[R(e)]$ *has non-unique solutions, and when part of the model parameters* $\{\theta_1^{1(l)}, \theta_1^{2(l)}, \theta_2^{1(l)}, \theta_2^{2(l)}\}$ *take the values*

$$\Theta_0 = \begin{cases} \theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-s+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-s, L-s-1, ..., 1 \\ \theta_2^{1(l)} = 0, \theta_2^{2(l)} = 1, & l = L-1, ..., 1 \end{cases}, \tag{20}$$

for some $0 < s < L$, θ_1 *and* θ_2 *have four sets of solutions of the cubic equation:*

$$\begin{cases} (3c_1\theta_1\theta_2 + c_1(\theta_2)^2 - 2c_6\theta_2)\sigma^2 - \mathbb{E}_e[N^e(2c_1(\theta_1+\theta_2) - c_6)]\sigma^2\theta_2 + c_7 = 0 \\ (\mathbb{E}_e[N^e(2c_1(\theta_1+\theta_2) - c_6)]\sigma^2\theta_2 - c_7)(c_3\theta_2 - c_4) - [c2(\theta_1+\theta_2) - c_5](\theta_2)^2 = 0 \end{cases}. \tag{21}$$

Algorithm 1 Detailed Training Procedure of LoRe-CIA

Require:
A labeled training graph $\mathcal{G} = (A, X, Y)$.
The number of hops t, LoRe-CIA weight λ, the number of classes C, total iterations T, model learning rate r.
Ensure:
Updated model f_Θ with parameter Θ.
for iterations in $1, 2, ..., T$ **do**
  Initialize $\mathcal{L}_{\text{LoRe}} = 0$
  for c in $1, 2, ..., C$ **do**
    Calculate the node representations $\phi(A, X)$
    Calculate A_c^t, where the (i, j)-th element of A_c^t equals the length of the shortest path from node i to j if the length is less than t else infinity.
    Use A_c^t to screen for pairs of nodes not exceeding a distance of t hops $\Omega_c(t)$.
    Compute LoRe-CIA loss of class c: $\mathcal{L}_{\text{LoRe}}^c$ according to Equation (13) using $\Omega_c(t)$, A_c^t and $\phi(A, X)$.
    $\mathcal{L}_{\text{LoRe}} = \mathcal{L}_{\text{LoRe}} + \mathcal{L}_{\text{LoRe}}^c$
  end for
  Compute final loss $\mathcal{L} = \mathcal{L}_{\text{ce}}(f_\Theta(A, X), Y) + \lambda\mathcal{L}_{\text{LoRe}}$, $\mathcal{L}_{\text{ce}}$ is the cross entropy loss.
  Update model parameters $\Theta = \Theta - r\nabla_\Theta\mathcal{L}$
end for

where $c_1 = \mathbb{E}_e[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)]$, $c_2 = \mathbb{E}_e[N_e(\tilde{A^e}^s X_1)^\top \tilde{A^e}^s X_1]$, $c_3 = \mathbb{E}_e[(\tilde{A^e}^s X_1)^\top \mathbf{1}]$, $c_4 = \mathbb{E}_e[((\tilde{A^e}^k X_1)^\top \mathbf{1}]$, $c_5 = \mathbb{E}_e[N_e((\tilde{A^e}^k X_1)^\top \tilde{A^e}^s X_1 + tr((\tilde{A^e}^k)^\top \tilde{A^e}^k) + N^e(1+\sigma^2))]$, $c_6 = \mathbb{E}_e[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)]$, $c_7 = \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]$.

Proof. We will use some symbols to simplify the expression of the toy GNN. Denote $\tilde{A}^m n_1 + n_2 + \epsilon^e$ as η. Use the following notations to represent the components of the L-layer GNN model:

$$
\begin{aligned}
f_\Theta(A, X) &= H_1^{(L)}\theta_1 + H_2^{(L)}\theta_2 \\
&= \underbrace{\left[\theta_1^{1^{(L-1)}}\bar{A}\left(...\theta_1^{1^{(3)}}\left(\theta_1^{1^{(2)}}\bar{A}(\theta_1^{1^{(1)}}\bar{A} + \theta_1^{2^{(1)}}\bar{I})X_1 + \theta_1^{2^{(2)}}(\theta_1^{1^{(1)}}\bar{A} + \theta_1^{2^{(1)}}\bar{I})X_1\right) + ...\right)\right]}_{C_1}\theta_1 \\
&+ \underbrace{\left[\theta_2^{1^{(L-1)}}\bar{A}\left(...\theta_2^{1^{(3)}}\left(\theta_2^{1^{(2)}}\bar{A}(\theta_2^{1^{(1)}}\bar{A} + \theta_2^{2^{(1)}}\bar{I})\tilde{A}^s X_1 + \theta_2^{2^{(2)}}(\theta_2^{1^{(1)}}\bar{A} + \theta_2^{2^{(1)}}\bar{I})\tilde{A}^s X_1\right) + ...\right)\right]}_{C_2}\theta_2 \\
&+ \underbrace{\left[\theta_2^{1^{(L-1)}}\bar{A}\left(...\theta_2^{1^{(3)}}\left(\theta_2^{1^{(2)}}\bar{A}(\theta_2^{1^{(1)}}\bar{A} + \theta_2^{2^{(1)}}\bar{I})\eta + \theta_2^{2^{(2)}}(\theta_2^{1^{(1)}}\bar{A} + \theta_2^{2^{(1)}}\bar{I})\eta\right) + ...\right)\right]}_{Z}\theta_2 \\
&= C_1\theta_1 + (C_2 + Z)\theta_2.
\end{aligned}
\tag{22}
$$

$C_1, C_2, Z \in \mathbb{R}^{N\times 1}$. We use C_1^e, C_2^e, and Z^e to denote the variables from the corresponding environment e. We further denote $C_2^e = C_2^{e\prime}\tilde{A^e}^s X_1$, $Z^e = C_2^{e\prime}\eta$.

Using these notations, the loss of environment e is

$$
\begin{aligned}
R(e) &= \mathbb{E}_{n_1,n_2}\left[\|f_\Theta(A^e, X^e) - Y^e\|_2^2\right] \\
&= \mathbb{E}_{n_1,n_2}\left[\left\|C_1^e\theta_1 + (C_2^e + Z^e)\theta_2 - \tilde{A^e}^k X_1 - n_1\right\|_2^2\right].
\end{aligned}
\tag{23}
$$

Denote the inner term $C_1^e\theta_1 + (C_2^e + Z^e)\theta_2 - \tilde{A^e}^k X_1 - n_1$ as l_e.

The variance of loss across environments is:

$$\begin{aligned}\mathbb{V}_e[R(e)] &= \mathbb{E}_e[R^2(e)] - \mathbb{E}_e^2[R(e)] \\ &= \mathbb{E}_e\left[\left(\mathbb{E}_{n_1,n_2}\left\|C_1^e\theta_1 + (C_2^e + Z^e)\theta_2 - \tilde{A^e}^k X_1 - n_1\right\|_2^2\right)^2\right] \\ &- \mathbb{E}_e^2\left[\mathbb{E}_{n_1,n_2}\left\|C_1^e\theta_1 + (C_2^e + Z^e)\theta_2 - \tilde{A^e}^k X_1 - n_1\right\|_2^2\right]. \\ &= \mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[(l_e^\top l_e)^2\right]\right] - \mathbb{E}_e^2\left[\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\right].\end{aligned} \tag{24}$$

Take the derivative of $\mathbb{V}_e[R(e)]$ with respect to θ_1:

$$\begin{aligned}\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_1} &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\mathbb{E}_{n_1,n_2}\left[2l_e^\top C_1^e\right]\right] \\ &\quad - 2\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\right]\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[2l_e^\top C_1^e\right]\right]\end{aligned} \tag{25}$$

Calculate the derivative by terms:

$$\begin{aligned}\mathbb{E}_{n_1,n_2}[l_e^\top l_e] = \mathbb{E}_{n_1,n_2}[&C_1^{e\top}C_1^e(\theta_1)^2 + C_1^{e\top}C_2^e\theta_1\theta_2 + C_1^{e\top}Z^e\theta_1\theta_2 - C_1^{e\top}\tilde{A^e}^k X_1\theta_1 - C_1^{e\top}n_1\theta_1 \\ &+ C_2^{e\top}C_1^e\theta_1\theta_2 + C_2^{e\top}C_2^e(\theta_2)^2 + C_2^{e\top}Z^e\theta_1\theta_2 - C_2^{e\top}\tilde{A^e}^k X_1\theta_2 - C_2^{e\top}n_1\theta_2 \\ &+ Z^{e\top}C_1^e\theta_1\theta_2 + Z^{e\top}C_2^e(\theta_2)^2 + Z^{e\top}Z^e(\theta_2)^2 - Z^{e\top}\tilde{A^e}^k X_1\theta_2 - Z^{e\top}n_1\theta_2 \\ &- (\tilde{A^e}^k X_1)^\top(C_1^e\theta_1 + C_2^e\theta_2) - (\tilde{A^e}^k X_1)^\top Z^e\theta_2 + (\tilde{A^e}^k X_1)^\top\tilde{A^e}^k X_1 \\ &+ (\tilde{A^e}^k X_1)^\top n_1 - n_1^\top(C_1^e\theta_1 + C_2^e\theta_2) - n_1^\top Z^e\theta_2 + n_1^\top\tilde{A^e}^k X_1 + n_1^\top n_1]\end{aligned} \tag{26}$$

Since n_1 and n_2 are independent standard Gaussian noise, we have $\mathbb{E}_{n_1,n_2}[n_1] = \mathbb{E}_{n_1,n_2}[n_2] = \mathbf{0}$, $\mathbb{E}_{n_1,n_2}[n_1^\top n_2] = \mathbb{E}_{n_1,n_2}[n_2^\top n_1] = 0$ and $\mathbb{E}_{n_1,n_2}[n_1^\top n_1] = \mathbb{E}_{n_1,n_2}[n_2^\top n_2] = N^e$ if it is the noise from e. Also, since ϵ^e and n_1, n_2 are independent, we have $\mathbb{E}_{n_1,n_2}[n_1^\top \epsilon^e] = \mathbb{E}_{n_1,n_2}[n_2^\top \epsilon^e] = 0$.

When

$$\begin{cases}\theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-s+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-s, L-s-1, ..., 1 \\ \theta_2^{1(l)} = 0, \theta_2^{2(l)} = 1, & l = L-1, ..., 1\end{cases}, \tag{27}$$

we have $C_2^{e\prime} = I_{N^e} \in \mathbb{R}^{N^e \times N^e}$ and $C_1^e = \tilde{A^e}^s X_1$. Consequently, we get $\mathbb{E}_{n_1,n_2}[Z^{e\top}n_1] = \text{tr}(C_2^{e\prime}\tilde{A^e}^k) = \text{tr}(\tilde{A^e}^k)$, $\mathbb{E}_{n_1,n_2}[Z^{e\top}Z^e] = \text{tr}\left((\tilde{A^e}^k)^\top(\tilde{A^e}^k)\right) + N^e + \epsilon^{e\top}\epsilon^e$.

Use the above conclusions and rewrite Equation (26) as (here we only plug in the value of $C_2^{e\prime}$):

$$\begin{aligned}&\mathbb{E}_{n_1,n_2}[l_e^\top l_e] = \\ &\left.\begin{aligned}&C_1^{e\top}C_1^e(\theta_1)^2 + C_1^{e\top}C_2^e\theta_1\theta_2 - C_1^{e\top}\tilde{A^e}^k X_1\theta_1 + C_2^{e\top}C_1^e\theta_1\theta_2 + C_2^{e\top}C_2^e(\theta_2)^2 - C_2^{e\top}\tilde{A^e}^k X_1\theta_2 \\ &+\text{tr}\left((\tilde{A^e}^k)^\top(\tilde{A^e}^k)\right)(\theta_2)^2 - (\tilde{A^e}^k X_1)^\top(C_1^e\theta_1 + C_2^e\theta_2) + (\tilde{A^e}^k X_1)^\top\tilde{A^e}^k X_1 + N^e\left(1 + (\theta_2)^2\right) \\ &-2\text{tr}(\tilde{A^e}^k)\end{aligned}\right\}(*) \\ &\left.+[C_1^{e\top}\epsilon^e + C_2^{e\top}\epsilon^e + \epsilon^{e\top}C_1^e]\theta_1\theta_2 + \epsilon^{e\top}\epsilon^e(\theta_2)^2 - 2(\tilde{A^e}^k X_1)^\top\epsilon^e\theta_2\right\}(**),\end{aligned} \tag{28}$$

$(*)$ and $(**)$ represent terms that are independent and associated with ϵ^e, respectively. Additionally,

$$\begin{aligned}\mathbb{E}_{n_1,n_2}[2l_e^\top C_1^e] &= 2\left[C_1^{e\top}C_1^e\theta_1 + C_2^{e\top}C_1^e\theta_2 + (C_2^{e\prime}\epsilon^e)^\top C_1^e\theta_2 - (\tilde{A^e}^k X_1)^\top C_1^e\right] \\ &= 2\left[C_1^{e\top}C_1^e\theta_1 + C_2^{e\top}C_1^e\theta_2 + \epsilon^{e\top}C_1^e\theta_2 - (\tilde{A^e}^k X_1)^\top C_1^e\right].\end{aligned} \tag{29}$$

Multiplying Equation (28) and (29) and take the expectation on e, using the assumption that $\mathbb{E}_e[(\epsilon^e{}_i)^2] = \sigma^2$ ($\epsilon^e{}_i$ is the i-th element of ϵ^e):

$$\begin{aligned}\mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}[l_e^\top l_e]\mathbb{E}_{n_1,n_2}\left[2l_e^\top C_1\right]\right] &= 4\mathbb{E}_e\left[(*)\left(C_1^{e\top}C_1^e\theta_1 + C_2^{e\top}C_2^{e\prime}\theta_2 - (\tilde{A^e}^k X_1)^\top C_1^e\right)\right] \\ &+ 4\mathbb{E}_e\left[(\tilde{A^e}^s X_1)^\top \tilde{A^e}^s X_1(3\theta_1\theta_2 + (\theta_2)^2) - 2(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)\theta_2\right]\theta_2\sigma^2 \\ &+ 4\mathbb{E}_e[N^e\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]\theta_2.\end{aligned} \tag{30}$$

Next target is to compute $2\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[l_e^\top l_e]]$ and $\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[2l_e^\top C_1]]$ Since ϵ^e has zero mean, we have:

$$2\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[l_e^\top l_e]] = \mathbb{E}_e[(*)] + 2\mathbb{E}_e[N^e](\theta_2)^2\sigma^2 \tag{31}$$

and

$$\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[2l_e^\top C_1^e]] = 2\mathbb{E}_e\left[C_1^{e\top}C_1^e\theta_1 + C_2^{e\top}C_1^e\theta_2 - (\tilde{A^e}^k X_1)^\top C_1^e\right]. \tag{32}$$

Use Equation (30) (31) and (32) and let $\frac{\partial \mathbb{V}_e[R(e)]}{\partial\theta_1} = 0$, we have:

$$\begin{aligned}&\mathbb{E}_e\left[3(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)(\theta_1\theta_2 + \frac{1}{3}(\theta_2)^2) - 2(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)\theta_2\right]\sigma^2 + \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)] \\ &-\mathbb{E}_e[N^e]\mathbb{E}_e\left[2(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)(\theta_1+\theta_2) - (\tilde{A^e}^k X_1)^\top C_1^e\right]\theta_2\sigma^2 = 0.\end{aligned} \tag{33}$$

Now we start calculating the expression of $\frac{\partial \mathbb{V}_e[R(e)]}{\partial\theta_2}$:

$$\begin{aligned}\frac{\partial \mathbb{V}_e[R(e)]}{\partial\theta_2} &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\mathbb{E}_{n_1,n_2}\left[2l_e^\top(C_2 + Z^e)\right]\right] \\ &- 2\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\right]\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[2l_e^\top(C_2^e + Z^e)\right]\right].\end{aligned} \tag{34}$$

Let $\frac{\partial \mathbb{V}_e[R(e)]}{\partial\theta_2} = 0$:

$$\begin{aligned}&\mathbb{E}_e\left[(C_1^{e\top}C_2^{e\prime} + C_2^{e\top}C_2^{e\prime} + C_2^{e\prime\top}C_1^{e\top})\theta_1\theta_2 + (C_2^{e\prime})^\top C_2^{e\prime}(\theta_2)^2 - 2(\tilde{A^e}^k X_1)^\top C_2^{e\prime}\theta_2\right] \\ &\mathbb{E}_e\left[(C_2^{e\prime\top}C_2^e\theta_2 - (\tilde{A^e}^k X_1)^\top C_2^{e\prime})\right]\sigma^2 \\ &- \mathbb{E}_e\left[N^e\sigma^2\left(C_1^{e\top}C_2^e\theta_1 + C_2^{e\top}C_2^e\theta_2 - (\tilde{A^e}^k X_1)^\top C_2^e + \text{tr}((\tilde{A^e}^k)^\top\tilde{A^e}^k) + N^e + C_2^{e\prime\top}C_2^{e\prime}\sigma^2\right)(\theta_2)^2\right] \\ &= 0.\end{aligned} \tag{35}$$

Plug Equation (33) in (35), we reach:

$$\begin{aligned}&\left[\mathbb{E}_e\left[N^e(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)(\theta_1+\theta_2) - (\tilde{A^e}^k X_1)^\top C_1^e\right]\theta_2\sigma^2 - \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]\right] \\ &\mathbb{E}_e\left((\tilde{A^e}^s X_1)^\top\mathbf{1}_{\mathbf{N^e}}\theta_2 - (\tilde{A^e}^k X_1)^\top\mathbf{1}_{\mathbf{N^e}}\right) \\ &- \mathbb{E}_e\left[N^e\left((\tilde{A^e}^s X_1)^\top\tilde{A^e}^s X_1(\theta_1+\theta_2) - (\tilde{A^e}^k X_1)^\top\tilde{A^e}^s X_1 + \text{tr}((\tilde{A^e}^k)^\top\tilde{A^e}^k) + N^e(1+\sigma^2)\right)\right](\theta_2)^2 \\ &= 0.\end{aligned} \tag{36}$$

Let $c_1 = \mathbb{E}_e[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)]$, $c_2 = \mathbb{E}_e[N_e(\tilde{A^e}^s X_1)^\top\tilde{A^e}^s X_1]$, $c_3 = \mathbb{E}_e[(\tilde{A^e}^s X_1)^\top\mathbf{1}]$, $c_4 = \mathbb{E}_e[((\tilde{A^e}^k X_1)^\top\mathbf{1}]$, $c_5 = \mathbb{E}_e[N_e((\tilde{A^e}^k X_1)^\top\tilde{A^e}^s X_1 + \text{tr}((\tilde{A^e}^k)^\top\tilde{A^e}^k) + N^e(1+\sigma^2))]$, $c_6 = \mathbb{E}_e[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)]$, $c_7 = \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]$,

we conclude that

$$\begin{cases} (3c_1\theta_1\theta_2 + c_1(\theta_2)^2 - 2c_6\theta_2)\sigma^2 - \mathbb{E}_e[N^e(2c_1(\theta_1+\theta_2)-c_6)]\sigma^2\theta_2 + c_7 = 0 \\ (\mathbb{E}_e[N^e(2c_1(\theta_1+\theta_2)-c_6)]\sigma^2\theta_2 - c_7)(c_3\theta_2 - c_4) - [c2(\theta_1+\theta_2)-c_5](\theta_2)^2 = 0 \end{cases} . \tag{37}$$

As for the derivative respect to ${\theta_1^1}^{(l)}$, ${\theta_1^2}^{(l)}$, ${\theta_2^1}^{(l)}$,${\theta_2^2}^{(l)}$, when they take the special value in (27), we have $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_1} \Rightarrow {\theta_1^1}^{(l)} = {\theta_1^2}^{(l)} = 0$ and $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_2} \Rightarrow {\theta_2^1}^{(l)} = {\theta_2^2}^{(l)} = 0, l = 1, ..., L$ So we conclude the solution induced by 37 is the solution of the objective.

□

E.1.2 Proof of the Failure Case of IRMv1 under Concept Shift

Proposition E.2. ***(IRMv1 will use spurious features)*** *The objective* $\min_\Theta \mathbb{E}_e[\|\nabla_{w|w=1.0} R(e)\|^2]$ *has a solution that uses spurious features:*

$$\begin{cases} \theta_1 = \frac{\mathbb{E}_e\left\{(\tilde{A^e}^s X_1)^\top (\tilde{A^e}^k \mathbf{1})\left[\mathbf{1}^\top \tilde{A^e}^s X_1 + (\tilde{A^e}^k X_1)^\top (\tilde{A^e}^k \mathbf{1})\right] + (1+\sigma^2)(\tilde{A^e}^s X_1)^\top \mathbf{1} (\tilde{A^e}^k X_1)^\top \mathbf{1}\right\}}{(2+\sigma^2)(\mathbb{E}_e[\tilde{A^e}^s]X_1)^\top \mathbf{1}} \\ \theta_2 = \frac{\mathbb{E}_e\left\{(\tilde{A^e}^s X_1)^\top (\tilde{A^e}^s X_1)[\mathbf{1}^\top (\tilde{A^e}^k \mathbf{1})] + (\tilde{A^e}^s X_1)^\top (\tilde{A^e}^k \mathbf{1})(\mathbf{1}^\top \tilde{A^e}^s X_1)\right\}}{(2+\sigma^2)(\mathbb{E}_e[\tilde{A^e}^s]X_1)^\top \mathbf{1}} \end{cases} . \tag{38}$$

when $\{{\theta_1^1}^{(l)}, {\theta_1^2}^{(l)}, {\theta_2^1}^{(l)}, {\theta_2^2}^{(l)}\}$ *take the special values for some* $0 < s < L$*:*

$$\Theta_0 = \begin{cases} {\theta_1^1}^{(l)} = 1, {\theta_1^2}^{(l)} = 1, & l = L-1, ..., L-s+1 \\ {\theta_1^1}^{(l)} = 0, {\theta_1^2}^{(l)} = 1, & l = L-s, L-s-1, ..., 1 \\ {\theta_2^1}^{(l)} = 0, {\theta_2^2}^{(l)} = 1, & l = L-1, ..., 1 \end{cases} . \tag{39}$$

Proof.

$$\begin{aligned} \mathcal{L}_{\text{IRMv1}} &= \mathbb{E}_e\left[\|\nabla_{w|w=1.0}\mathbb{E}_{n_1,n_2}[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]\|_2^2\right] \\ &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}[(\hat{Y}^e - Y^e)^\top \phi(A^e, X^e)]\right] \\ &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[\left(C_1^e\theta_1 + (C_2^e + Z^e)\theta_2 - \tilde{A^e}^k X_1 - n_1\right)^\top (C_1^e\theta_1 + (C_2^e + Z^e)\theta_2)\right]\right] \\ &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[{C_1^e}^\top C_1^e(\theta_1)^2 + 2{C_1^e}^\top (C_2^e + Z^e)\theta_1\theta_2 + {C_2^e}^\top C_2^e(\theta_2)^2\right]\right] \\ &\quad - \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[(\tilde{A^e}^k X_1 + n_1)^\top C_1^e\theta_1 - ((\tilde{A^e}^k X_1) + n_1)^\top (C_2^e + Z^e)\theta_2\right]\right]. \end{aligned} \tag{40}$$

Take the derivative of $\mathcal{L}_{\text{IRMv1}}$ w.r.t. θ_1 and θ_2:

$$\begin{cases} \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_1} = \mathbb{E}_e[2\mathbb{E}_{n_1,n_2}[{C_1^e}^\top C_1^e\theta_1 + 2{C_1^e}^\top (C_2^e + Z^e)\theta_2 - ((\tilde{A^e}^k X_1) + n_1)^\top C_1^e]] \\ \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_2} = \mathbb{E}_e[2\mathbb{E}_{n_1,n_2}[{C_2^e}^\top C_2^e\theta_2 + 2{C_1^e}^\top (C_2^e + Z^e)\theta_1 - ((\tilde{A^e}^k X_1) + n_1)^\top (C_2^e + Z^e)^e]] \end{cases} . \tag{41}$$

For brevity, let $a = {C_1^e}^\top C_1^e\theta_1$, $b = {C_1^e}^\top (C_2^e + Z^e)$, $c = ((\tilde{A^e}^k X_1) + n_1)^\top C_1^e$, $d = {C_2^e}^\top C_2^e$, $e = ((\tilde{A^e}^k X_1) + n_1)^\top (C_2^e + Z^e)^e$. By letting $\{{\theta_1^1}^{(l)}, {\theta_1^2}^{(l)}, {\theta_2^1}^{(l)}, {\theta_2^2}^{(l)}\}$ take the values of Θ_0, let the derivative w.r.t. θ_1 and θ_2 to be zero, we have

$$\begin{cases} \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_1} = \frac{ac - be}{2(a^2 - b^2)} \\ \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_2} = \frac{ae - bc}{2(a^2 - b^2)} \end{cases} . \tag{42}$$

Also, when $\{{\theta_1^1}^{(l)}, {\theta_1^2}^{(l)}, {\theta_2^1}^{(l)}, {\theta_2^2}^{(l)}\}$ take the values of Θ_0, we have

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_1} = \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_1^1}^{(l)}} = \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_1^2}^{(l)}} = 0 \\ \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_2} = \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_2^1}^{(l)}} = \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_2^2}^{(l)}} = 0 \end{aligned} \tag{43}$$

□

E.1.3 PROOF OF THE SUCCESSFUL CASE OF CIA UNDER CONCEPT SHIFT

Proposition E.3. *Optimizing the CIA objective will lead to the optimal solution Θ^*:*

$$\begin{cases} \theta_1 = 1 \\ \theta_2 = 0 \quad or \quad \exists l \in \{1, ..., L-1\} \text{ s.t. } {\theta_2^1}^{(l)} = {\theta_2^2}^{(l)} = 0 \\ {\theta_1^1}^{(l)} = 1, {\theta_1^2}^{(l)} = 1, \quad l = L-1, ..., L-k+1 \\ {\theta_1^1}^{(l)} = 0, {\theta_1^2}^{(l)} = 1, \quad l = L-k, L-k-1, ..., 1 \end{cases}. \tag{44}$$

Proof. For brevity, denote a node representation of ${C_1^e}_c$ as C_1^i and the one of ${C_1^{e'}}_c$ as C_1^j. The same is true for C_2^i and C_2^j. In this toy model, we need to consider the expectation of the noise, while in real cases such noise is included in the node features so taking expectation on e will handle this. Therefore, we add $\mathbb{E}_{n_1,n_2}$ in this proof, and this expectation is excluded in the formal description of the objective in the main paper.

$$\begin{aligned} \mathcal{L}_{\text{CIA}} &= \mathbb{E}_{\substack{e,e' \\ e \neq e'}} \mathbb{E}_{n_1,n_2} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j) \in \Omega^{e,e'}}} \left[\mathcal{D}(\phi_\Theta(A^e, X^e)_{[c][v_i]}, \phi_\Theta(A^e, X^{e'})_{[c][v_j]}) \right] \\ &= \mathbb{E}_{\substack{e,e' \\ e \neq e'}} \mathbb{E}_{n_1,n_2} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j) \in \Omega^{e,e'}}} \| C_1^i \theta_1 + (C_2^i + Z^e)\theta_2 - C_1^j \theta_1 - (C_2^j + Z^{e'})\theta_2 \|_2^2 \end{aligned} \tag{45}$$

$$\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1} = \mathbb{E}_{\substack{e,e' \\ e \neq e'}} \mathbb{E}_{n_1,n_2} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j) \in \Omega^{e,e'}}} \left[C_1^i \theta_1 + (C_2^i + Z^e)\theta_2 - C_1^j \theta_1 - (C_2^j + Z^{e'})\theta_2 \right]^\top (C_1^i - C_1^j) \tag{46}$$

Let $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1} = 0$, we have:

$$\mathbb{E}_{\substack{e,e' \\ e \neq e'}} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j) \in \Omega^{e,e'}}} \left[(C_1^i - C_1^j)^\top (C_1^i - C_1^j)\theta_1 + (C_2^i - C_2^k)^\top (C_1^i - C_1^j)\theta_2 \right] = 0 \tag{47}$$

Also, we have:

$$\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2} = \mathbb{E}_{\substack{e,e' \\ e \neq e'}} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j) \in \Omega^{e,e'}}} \left[(C_1^i - C_1^j)^\top (C_2^i - C_2^j)\theta_1 + \left[(C_2^i - C_2^k)^\top (C_2^i - C_2^j) + (Z^e - Z^{e'})^\top (Z^e - Z^{e'}) \right] \theta_2 \right] \tag{48}$$

Further let $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2} = 0$, combining Equation (47), we get

$$\begin{cases} \theta_1 = 0 \quad \text{or} \quad \exists l \in \{1, ..., L-1\} \text{ s.t. } {\theta_1^1}^{(l)} = {\theta_1^2}^{(l)} = 0 \\ \theta_2 = 0 \end{cases}. \tag{49}$$

or, if $\theta_1 \neq 0$ and $\forall l \in \{1, ..., L-1\}$, the parameters of that layer l of the invariant branch of the GNN are not all zero: ${\theta_1^1}^{(l)} \neq 0$ or ${\theta_1^2}^{(l)} \neq 0$, then we get

$$\underbrace{\theta_2 \, \mathbb{E}_{\substack{e,e' \\ e \neq e'}} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j) \in \Omega^{e,e'}}} \left[-\frac{[(C_1^i - C_1^j)^\top (C_2^i - C_2^j)]^2}{(C_1^i - C_1^j)^\top (C_1^i - C_1^j)} + (C_2^i - C_2^j)^\top (C_2^i - C_2^j) + (Z^e - Z^{e'})^\top (Z^e - Z^{e'}) \right]}_{F} = 0 \tag{50}$$

Due to the property of the inner product, $F > 0$ unless $\exists l \in \{1, ..., L-1\}$ s.t. ${\theta_2^1}^{(l)} = {\theta_2^2}^{(l)} = 0$. To ensure $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2}$, we conclude that $\theta_2 = 0$ or $\exists l \in \{1, ..., L-1\}$ s.t. ${\theta_2^1}^{(l)} = {\theta_2^2}^{(l)} = 0$.

In conclusion, to satisfy the constraint of CIA, no matter whether the invariant branch has zero output, the spurious branch must have zero parameters, i.e.,

$$\theta_2 = 0 \quad \text{or} \quad \exists l \in \{1, ..., L-1\} \text{ s.t. } {\theta_2^1}^{(l)} = {\theta_2^2}^{(l)} = 0 \tag{51}$$

Thus, CIA will remove spurious features.

Now we show that when CIA objective has been reached (the spurious branch has zero outputs), the objective of $\min_\Theta \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]$ will help to learn predictive paramters of the invariant branch θ_1, $\theta_1^{1}{}^{(}l)$ and $\theta_1^{2}{}^{(}l)$. When Equation (51) holds,

$$\begin{aligned}\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1} &= 2\mathbb{E}_e \mathbb{E}_{n_1,n_2}\left[\left(C_1^e\theta_1 - \tilde{A^e}^k X_1 - n_1\right)^\top C_1^e\right] \\ &= 2\mathbb{E}_e\left[\left(C_1^e\theta_1 - \tilde{A^e}^k X_1\right)^\top C_1^e\right]\end{aligned} \tag{52}$$

Let $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1} = 0$, we get the predictive parameters

$$\begin{cases} \theta_1 = 1 \\ \theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-k+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-k, L-k-1, ..., 1 \end{cases}. \tag{53}$$

Plug the final solution back in $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1^{1(l)}}$, $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1^{2(l)}}$, $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2^{1(l)}}$, $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2^{2(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1^{1(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1^{2(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_2^{1(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_2^{2(l)}}$, we can verify that these terms are all 0. □

E.2 Proof of the Covariate Shift Case

E.2.1 Proof of the Failure Case of VREx under Covariate Shift

Proof. We will use some symbols to simplify the expression of the toy GNN. Denote $n_2 + \epsilon^e$ as η. Use the following notations to represent the components of the L-layer GNN model:

$$\begin{aligned} f_\Theta(A, X) &= H_1^{(L)}\theta_1 + H_2^{(L)}\theta_2 \\ &= \underbrace{\left[\theta_1^{1(L-1)}\bar{A}\left(...\theta_1^{1(3)}\left(\theta_1^{1(2)}\bar{A}(\theta_1^{1(1)}\bar{A} + \theta_1^{2(1)}\bar{I})X_1 + \theta_1^{2(2)}(\theta_1^{1(1)}\bar{A} + \theta_1^{2(1)}\bar{I})X_1\right) + ...\right)\right]}_{C_1}\theta_1 \\ &+ \underbrace{\left[\theta_2^{1(L-1)}\bar{A}\left(...\theta_2^{1(3)}\left(\theta_2^{1(2)}\bar{A}(\theta_2^{1(1)}\bar{A} + \theta_2^{2(1)}\bar{I})\eta + \theta_2^{2(2)}(\theta_2^{1(1)}\bar{A} + \theta_2^{2(1)}\bar{I})\eta\right) + ...\right)\right]}_{Z}\theta_2 \\ &= C_1\theta_1 + Z\theta_2. \end{aligned} \tag{54}$$

$C_1, Z \in \mathbb{R}^{N\times 1}$. We use C_1^e and Z^e to denote the variables from the corresponding environment e. We further denote $Z^e = C_2^{e\prime}\eta$.

Using these notations, the loss of environment e is

$$\begin{aligned} R(e) &= \mathbb{E}_{n_1,n_2}\left[\|f_\Theta(A^e, X^e) - Y^e\|_2^2\right] \\ &= \mathbb{E}_{n_1,n_2}\left[\left\|C_1^e\theta_1 + Z^e\theta_2 - \tilde{A^e}^k X_1 - n_1\right\|_2^2\right]. \end{aligned} \tag{55}$$

Denote the inner term $C_1^e\theta_1 + Z^e\theta_2 - \tilde{A^e}^k X_1 - n_1$ as l_e.

The variance of loss across environments is:

$$\begin{aligned} \mathbb{V}_e[R(e)] &= \mathbb{E}_e[R^2(e)] - \mathbb{E}_e^2[R(e)] \\ &= \mathbb{E}_e\left[\left(\mathbb{E}_{n_1,n_2}\left\|C_1^e\theta_1 + Z^e\theta_2 - \tilde{A^e}^k X_1 - n_1\right\|_2^2\right)^2\right] \\ &- \mathbb{E}_e^2\left[\mathbb{E}_{n_1,n_2}\left\|C_1^e\theta_1 + Z^e\theta_2 - \tilde{A^e}^k X_1 - n_1\right\|_2^2\right]. \\ &= \mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[(l_e^\top l_e)^2\right]\right] - \mathbb{E}_e^2\left[\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\right]. \end{aligned} \tag{56}$$

Take the derivative of $\mathbb{V}_e[R(e)]$ with respect to θ_1:

$$\begin{aligned}\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_1} &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\mathbb{E}_{n_1,n_2}\left[2l_e^\top C_1^e\right]\right] \\ &\quad - 2\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\right]\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[2l_e^\top C_1^e\right]\right]\end{aligned} \tag{57}$$

Calculate the derivative by terms:

$$\begin{aligned}\mathbb{E}_{n_1,n_2}[l_e^\top l_e] = \mathbb{E}_{n_1,n_2}[&C_1^{e\top} C_1^e(\theta_1)^2 + C_1^{e\top} Z^e\theta_1\theta_2 - C_1^{e\top}\tilde{A^e}^k X_1\theta_1 - C_1^{e\top} n_1\theta_1 \\ &+ Z^{e\top} C_1^e\theta_1\theta_2 + Z^{e\top} Z^e(\theta_2)^2 - Z^{e\top}\tilde{A^e}^k X_1\theta_2 - Z^{e\top} n_1\theta_2 \\ &- (\tilde{A^e}^k X_1)^\top C_1^e\theta_1 - (\tilde{A^e}^k X_1)^\top Z^e\theta_2 + (\tilde{A^e}^k X_1)^\top\tilde{A^e}^k X_1 \\ &+ (\tilde{A^e}^k X_1)^\top n_1 - n_1^\top C_1^e\theta_1 - n_1^\top Z^e\theta_2 + n_1^\top\tilde{A^e}^k X_1 + n_1^\top n_1]\end{aligned} \tag{58}$$

Since n_1 and n_2 are independent standard Gaussian noise, we have $\mathbb{E}_{n_1,n_2}[n_1] = \mathbb{E}_{n_1,n_2}[n_2] = \mathbf{0}$, $\mathbb{E}_{n_1,n_2}[n_1^\top n_2] = \mathbb{E}_{n_1,n_2}[n_2^\top n_1] = 0$ and $\mathbb{E}_{n_1,n_2}[n_1^\top n_1] = \mathbb{E}_{n_1,n_2}[n_2^\top n_2] = N^e$ if it is the noise from e. Also, since ϵ^e and n_1, n_2 are independent, we have $\mathbb{E}_{n_1,n_2}[n_1^\top\epsilon^e] = \mathbb{E}_{n_1,n_2}[n_2^\top\epsilon^e] = 0$.

When

$$\begin{cases}\theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-s+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-s, L-s-1, ..., 1 \\ \theta_2^{1(l)} = 0, \theta_2^{2(l)} = 1, & l = L-1, ..., 1\end{cases}, \tag{59}$$

we have $C_2^{e\prime} = I_{N^e} \in \mathbb{R}^{N^e\times N^e}$ and $C_1^e = \tilde{A^e}^s X_1$. Consequently, we get $\mathbb{E}_{n_1,n_2}[Z^{e\top} n_1] = 0$, $\mathbb{E}_{n_1,n_2}[Z^{e\top} Z^e] = N^e + \epsilon^{e\top}\epsilon^e$.

Use the above conclusions and rewrite Equation (58) as (here we only plug in the value of $C_2^{e\prime}$):

$$\begin{aligned}&\mathbb{E}_{n_1,n_2}[l_e^\top l_e] = \\ &\left.\begin{aligned}&C_1^{e\top} C_1^e(\theta_1)^2 - C_1^{e\top}\tilde{A^e}^k X_1\theta_1 \\ &-(\tilde{A^e}^k X_1)^\top C_1^e\theta_1 + (\tilde{A^e}^k X_1)^\top\tilde{A^e}^k X_1 + N^e\left(1 + (\theta_2)^2\right)\end{aligned}\right\}(*) \\ &\left.+[C_1^{e\top}\epsilon^e + \epsilon^{e\top} C_1^e]\theta_1\theta_2 + \epsilon^{e\top}\epsilon^e(\theta_2)^2 - 2(\tilde{A^e}^k X_1)^\top\epsilon^e\theta_2\right\}(**),\end{aligned} \tag{60}$$

$(*)$ and $(**)$ represent terms that are independent and associated with ϵ^e, respectively.

Additionally,

$$\mathbb{E}_{n_1,n_2}[2l_e^\top C_1^e] = 2\left[C_1^{e\top} C_1^e\theta_1 + \epsilon^{e\top} C_1^e\theta_2 - (\tilde{A^e}^k X_1)^\top C_1^e\right] \tag{61}$$

.

Multiplying Equation (60) and (61) and take the expectation on e, using the assumption that $\mathbb{E}_e[(\epsilon^e{}_i)^2] = \sigma^2$ ($\epsilon^e{}_i$ is the i-th element of ϵ^e):

$$\begin{aligned}\mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}[l_e^\top l_e]\mathbb{E}_{n_1,n_2}\left[2l_e^\top C_1\right]\right] &= 4\mathbb{E}_e\left[(*)\left(C_1^{e\top} C_1^e\theta_1 - (\tilde{A^e}^k X_1)^\top C_1^e\right)\right] \\ &\quad + 4\mathbb{E}_e\left[(\tilde{A^e}^s X_1)^\top\tilde{A^e}^s X_1(2\theta_1\theta_2 + (\theta_2)^2) - 2(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)\theta_2\right]\theta_2\sigma^2 \\ &\quad + 4\mathbb{E}_e[N^e\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]\theta_2.\end{aligned} \tag{62}$$

Next target is to compute $2\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[l_e^\top l_e]]$ and $\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[2l_e^\top C_1]]$ Since ϵ^e has zero mean, we have:

$$2\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[l_e^\top l_e]] = \mathbb{E}[(*)] + \mathbb{E}[2N^e]\sigma^2(\theta_2)^2 \tag{63}$$

and

$$\mathbb{E}_e[\mathbb{E}_{n_1,n_2}[2l_e^\top C_1^e]] = 2\mathbb{E}_e\left[C_1^{e\top} C_1^e\theta_1 - (\tilde{A^e}^k X_1)^\top C_1^e\right]. \tag{64}$$

Use Equation (62) (63) and (64) and let $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_1} = 0$, we have:

$$\mathbb{E}_e\left[2(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)(\theta_1\theta_2 + \frac{1}{2}(\theta_2)^2) - 2(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)\theta_2\right]\sigma^2 + \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]$$
$$-\mathbb{E}_e[N^e]\mathbb{E}_e\left[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)\theta_1 - (\tilde{A^e}^k X_1)^\top C_1^e\right]\theta_2\sigma^2 = 0. \tag{65}$$

Now we start calculating the expression of $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_2}$:

$$\begin{aligned}\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_2} &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\mathbb{E}_{n_1,n_2}\left[2l_e^\top Z^e\right]\right] \\ &\quad - 2\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[l_e^\top l_e\right]\right]\mathbb{E}_e\left[\mathbb{E}_{n_1,n_2}\left[2l_e^\top Z^e\right]\right].\end{aligned} \tag{66}$$

Let $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_2} = 0$:

$$\mathbb{E}_e\left[(C_1^{e\top}C_2^{e\prime} + C_2^{e\prime\top}C_1^{e\top})\theta_1\theta_2 + (C_2^{e\prime})^\top C_2^{e\prime}(\theta_2)^2 - 2(\tilde{A^e}^k X_1)^\top C_2^{e\prime}\theta_2\right]\mathbb{E}_e\left[(-(\tilde{A^e}^k X_1)^\top C_2^{e\prime})\right]\sigma^2$$
$$-\mathbb{E}_e\left[N^e\sigma^2\left(\text{tr}((\tilde{A^e}^k)^\top\tilde{A^e}^k) + N^e + C_2^{e\prime\top}C_2^{e\prime}\sigma^2\right)(\theta_2)^2\right]$$
$$= 0. \tag{67}$$

Plug Equation (65) in (67), we reach:

$$\left[\mathbb{E}_e\left[N^e(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)\theta_1 - N^e(\tilde{A^e}^k X_1)^\top C_1^e\right]\theta_2\sigma^2 - \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]\right]\mathbb{E}_e\left(-(\tilde{A^e}^k X_1)^\top \mathbf{1_{N^e}}\right)$$
$$-\mathbb{E}_e\left[N^e\left(\text{tr}((\tilde{A^e}^k)^\top\tilde{A^e}^k) + N^e(1+\sigma^2)\right)\right](\theta_2)^2$$
$$= 0. \tag{68}$$

Let $c_1 = \mathbb{E}[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^s X_1)]$, $c_2 = \mathbb{E}[(\tilde{A^e}^s X_1)^\top(\tilde{A^e}^k X_1)]$, $c_3 = \mathbb{E}_e[\epsilon^{e\top}\epsilon^e\epsilon^{e\top}(\tilde{A^e}^s X_1)]\sigma^2$, $c_4 = \mathbb{E}_e\left[(\tilde{A^e}^k X_1)^\top \mathbf{1_{N^e}}\right]\sigma^2$, $c_5 = \mathbb{E}_e\left[N^e\left(\text{tr}((\tilde{A^e}^k)^\top\tilde{A^e}^k) + N^e(1+\sigma^2)\right)\right]$,

we conclude that

$$\begin{cases} c_1\sigma^2(2\theta_1\theta_2 + (\theta_2)^2 - 2c_2\sigma^2\theta_2) + c_3 - \mathbb{E}_e[N^e]c_1\sigma^2\theta_1\theta_2 + \mathbb{E}_e[N^e]c_2\sigma^2\theta_2 = 0 \\ \left[c_3 - \mathbb{E}_e[N^e]c_1\sigma^2\theta_1\theta_2 + \mathbb{E}_e[N^e]c_2\sigma^2\theta_2\right]c_4 - c_5(\theta_2)^2 \end{cases}. \tag{69}$$

As for the derivative respect to ${\theta_1^1}^{(l)}$, ${\theta_1^2}^{(l)}$, ${\theta_2^1}^{(l)}$,${\theta_2^2}^{(l)}$, when they take the special value in (59), we have $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_1} \Rightarrow {\theta_1^1}^{(l)} = {\theta_1^2}^{(l)} = 0$ and $\frac{\partial \mathbb{V}_e[R(e)]}{\partial \theta_2} \Rightarrow {\theta_2^1}^{(l)} = {\theta_2^2}^{(l)} = 0, l = 1, ..., L$ So we conclude the solution induced by 69 is the solution of the objective. □

E.2.2 Proof of the Failure Case of IRMv1 under Covariate Shift

Proof.

$$\begin{aligned}\mathcal{L}_{\text{IRMv1}} &= \mathbb{E}_e\left[\|\nabla_{w|w=1.0}\mathbb{E}_{n_1,n_2}[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]\|_2^2\right] \\ &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}[(\hat{Y}^e - Y^e)^\top\phi(A^e, X^e)]\right] \\ &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[\left(C_1^e\theta_1 + Z^e\theta_2 - \tilde{A^e}^k X_1 - n_1\right)^\top(C_1^e\theta_1 + Z^e\theta_2)\right]\right] \\ &= \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[C_1^{e\top}C_1^e(\theta_1)^2 + 2C_1^{e\top}Z^e\theta_1\theta_2\right]\right] \\ &\quad - \mathbb{E}_e\left[2\mathbb{E}_{n_1,n_2}\left[(\tilde{A^e}^k X_1 + n_1)^\top C_1^e\theta_1 - (\tilde{A^e}^k X_1 + n_1)^\top Z^e\theta_2\right]\right].\end{aligned} \tag{70}$$

Take the derivative of $\mathcal{L}_{\text{IRMv1}}$ w.r.t. θ_1:

$$\frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_1} = \mathbb{E}_e[2\mathbb{E}_{n_1,n_2}[{C_1^e}^\top C_1^e \theta_1 + 2{C_1^e}^\top Z^e \theta_2 - ((\tilde{A^e}^k X_1) + n_1)^\top C_1^e]], \tag{71}$$

Let it be zero, we get $\theta_1 = \frac{\mathbb{E}_e[(\tilde{A^e}^k X_1)^\top (\tilde{A^e}^{2k} X_1)]}{\mathbb{E}_e[(\tilde{A^e}^{2k} X_1)^\top (\tilde{A^e}^{2k} X_1)]}$

Take the derivative of $\mathcal{L}_{\text{IRMv1}}$ w.r.t. θ_2:

$$\frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_2} = \mathbb{E}_e[2\mathbb{E}_{n_1,n_2}[2{C_1^e}^\top Z^e \theta_1 - ((\tilde{A^e}^k X_1) + n_1)^\top Z^e]] \equiv 0 \tag{72}$$

Also, when $\{{\theta_1^1}^{(l)}, {\theta_1^2}^{(l)}, {\theta_2^1}^{(l)}, {\theta_2^2}^{(l)}\}$ take the values of Θ_0, we have

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_1} &= \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_1^1}^{(l)}} = \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_1^2}^{(l)}} = 0 \\ \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial \theta_2} &= \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_2^1}^{(l)}} = \frac{\partial \mathcal{L}_{\text{IRMv1}}}{\partial {\theta_2^2}^{(l)}} = 0 \end{aligned} \tag{73}$$

□

E.3 Proof of the Successful Case of CIA under Covariate Shift

Proof. For brevity, denote a node representation of ${C_1^e}_c$ as C_1^i and the one of ${C_1^{e'}}_c$ as C_1^j. The same is true for C_2^i and C_2^j. In this toy model, we need to consider the expectation of the noise, while in real cases such noise is included in the node features so taking expectation on e will handle this. Therefore, we add $\mathbb{E}_{n_1,n_2}$ in this proof, and this expectation is excluded in the formal description of the objective in the main paper.

$$\begin{aligned} \mathcal{L}_{\text{CIA}} &= \mathop{\mathbb{E}}_{\substack{e,e' \\ e\neq e'}} \mathbb{E}_{n_1,n_2} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j)\in\Omega^{e,e'}}} \left[\mathcal{D}(\phi_\Theta(A^e, X^e)_{[c][v_i]}, \phi_\Theta(A^e, X^{e'})_{[c][v_j]})\right] \\ &= \mathop{\mathbb{E}}_{\substack{e,e' \\ e\neq e'}} \mathbb{E}_{n_1,n_2} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j)\in\Omega^{e,e'}}} \|C_1^i + Z^e - C_1^j - Z^{e'}\|_2^2 \end{aligned} \tag{74}$$

$$\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1} = \mathop{\mathbb{E}}_{\substack{e,e' \\ e\neq e'}} \mathbb{E}_{n_1,n_2} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j)\in\Omega^{e,e'}}} \left[C_1^i\theta_1 + Z^e\theta_2 - C_1^j\theta_1 - Z^{e'}\theta_2\right]^\top (C_1^i - C_1^j) \tag{75}$$

Let $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1} = 0$, we have:

$$\mathop{\mathbb{E}}_{\substack{e,e' \\ e\neq e'}} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j)\in\Omega^{e,e'}}} \left[(C_1^i - C_1^j)^\top (C_1^i - C_1^j)\theta_1\right] = 0 \tag{76}$$

Thus, we get two possible solutions of the invariant branch. The first valid solution is the optimal one:

$$\begin{cases} \theta_1 = 1 \\ {\theta_1^1}^{(l)} = 1, {\theta_1^2}^{(l)} = 1, \quad l = L-1, ..., L-k+1 \\ {\theta_1^1}^{(l)} = 0, {\theta_1^2}^{(l)} = 1, \quad l = L-k, L-k-1, ..., 1 \end{cases}. \tag{77}$$

The second valid solution is a trivial one:

$$\left\{ \theta_1 = 0 \quad \text{or} \quad \exists l \in \{1, ..., L-1\} \text{ s.t. } {\theta_1^1}^{(l)} = {\theta_1^2}^{(l)} = 0 \right. . \tag{78}$$

Take the derivative of the objective w.r.t. θ_2:

$$\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2} = \mathop{\mathbb{E}}_{\substack{e,e' \\ e\neq e'}} \mathbb{E}_c \mathbb{E}_{\substack{i,j \\ (i,j)\in\Omega^{e,e'}}} \left[\left[(Z^e - Z^{e'})^\top (Z^e - Z^{e'})\right]\theta_2\right] = 2(1+\sigma^2)\theta_2 \tag{79}$$

Let $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2} = 0$, we get $\theta_2 = 0$. Thus, CIA will remove spurious features.

Now we show that when CIA objective has been reached (the spurious branch has zero outputs), the objective of $\min_{\Theta} \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]$ will help to learn predictive parameters of the invariant branch θ_1, $\theta_1^{1(}l)$ and $\theta_1^{2(}l)$. When $\theta_2 = 0$:

$$\begin{aligned}\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1} &= 2\mathbb{E}_e \mathbb{E}_{n_1, n_2}\left[\left(C_1^e \theta_1 - \tilde{A^e}^k X_1 - n_1\right)^\top C_1^e\right] \\ &= 2\mathbb{E}_e\left[\left(C_1^e \theta_1 - \tilde{A^e}^k X_1\right)^\top C_1^e\right]\end{aligned} \tag{80}$$

Let $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1} = 0$, we get the predictive parameters

$$\begin{cases} \theta_1 = 1 \\ \theta_1^{1(l)} = 1, \theta_1^{2(l)} = 1, & l = L-1, ..., L-k+1 \\ \theta_1^{1(l)} = 0, \theta_1^{2(l)} = 1, & l = L-k, L-k-1, ..., 1 \end{cases}. \tag{81}$$

Plug the final solution back in $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1^{1(l)}}$, $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_1^{2(l)}}$, $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2^{1(l)}}$, $\frac{\partial \mathcal{L}_{\text{CIA}}}{\partial \theta_2^{2(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1^{1(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_1^{2(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_2^{1(l)}}$, $\frac{\partial \mathbb{E}_e[\mathcal{L}(f_\Theta(A^e, X^e), Y^e)]}{\partial \theta_2^{2(l)}}$, we can verify that these terms are all 0. □