ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets

Appendix

A Reviewers' FAQ

For transparency and to facilitate the reviewing process, we summarize questions and answers that arose during the review of an earlier version of this paper.

The most significant update to this version is the inclusion of **additional datasets**, and **experiments without a diversity-inducing method** (Section 5 and Appendix C). Some reviewers previously commented that our findings are not surprising, yet we continually see misunderstandings of their implications in the literature. We added a **discussion of such a recent case** [20] (April 2023) in Section 8. Other improvements include a discussion of **occurrences the existing literature** (Section 5.1) and various clarifications throughout the text.

Q: Isn't the message of the paper unsurprising? Few would disagree that "focusing on ID performance alone may not lead to optimal OOD performance".

A: We also would like this to be self-evident. Yet, multiple papers precisely conclude that focusing exclusively on ID performance is a fine strategy:

- "*If practitioners want to make the model more robust on OOD data, the main focus should be to improve the ID classification error*" [48]
- "*We see the following potential prescriptive outcomes: (...) the correlation between OOD and ID performance can simplify model development since we can focus on a single metric*" [24]

Many readers will clearly benefit from our exposition of the necessary caveats to such statements.

Q: Past work on positive correlations is not a "for all" claim. Isn't it merely pointing out that this correlation is surprisingly strong in many benchmarks?

A: This is indeed what the experiments show in "accuracy on the line" for example. But the takeaways (in this and other papers) are over-generalized and much overblown (see the citations in the previous answer above). The issue is clear when subsequent works apply this message uncritically and use the phenomenon ("accuracy on the line") as an unverified assumption on other datasets.

When this assumption serves to justify an experimental design (e.g. model selection) there is often no more chance to verify its validity later on, nor to recover from this faulty choice of assumptions (because they serve as premises for the whole analysis). The risk of such methodological mistakes in future research is why this paper is important.

Q: Why the focus on the Camelyon17 dataset?

A: We have now added experiments on five other datasets. They show that the phenomenon is not an isolated case.

Large-scale support for our message has also appeared after the release of a first version of this paper. Naganuma et al. [25] show that a re-evaluation of OOD benchmarks with a wider range of hyperparameters than previous studies leads to more diverse types of ID/OOD relations than the "linear trend".

Regarding the value of a large-scale evaluation of many datasets, note that this paper is in a very different situation from past studies claiming that positive correlations were widespread (cf. proof of existence vs. absence). The point of this paper is that "worst-case scenarios" (inverse correlation) are a possibility, hence the "best case" (positive correlation) cannot serve as an unverified assumption.

Q: Why is the theoretical analysis restricted to linear models?

A: Its value is precisely in demonstrating the phenomenon in a hypothesis space as simple as linear models. High-capacity models would be less surprising in displaying complex performance patterns.

Q: Why aren't domain generalization (DG) methods investigated?

A: Because the existence of predictors showing an inverse correlation is mostly relevant to the data and hypothesis space rather than learning methods. Using DG methods could have been a way to obtain a variety of models in the hypothesis space. We chose a less ad hoc solution with a general-purpose diversity-inducing method. This allows covering more of the ID/OOD spectrum than specific DG methods (i.e. we avoid the analysis to focus on the specific inductive biases of an arbitrary set of methods).

B Additional results on Camelyon17

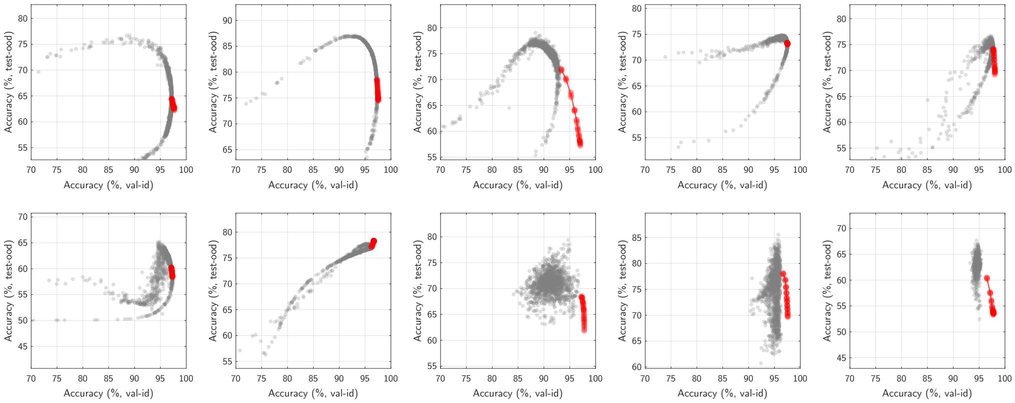

Figure 8: As in Figure 3, we show that higher OOD accuracy can be sometimes be traded off for a lower ID accuracy. Each panel shows results from a different pretrained model (i.e. pretrained with a different random seed). Each dot represents a linear classifier re-trained on features from this pretrained model with standard ERM (red dots ●) or with a diversity-inducing method [45] (gray dots ●). The latter set includes models with higher OOD / lower ID accuracies.

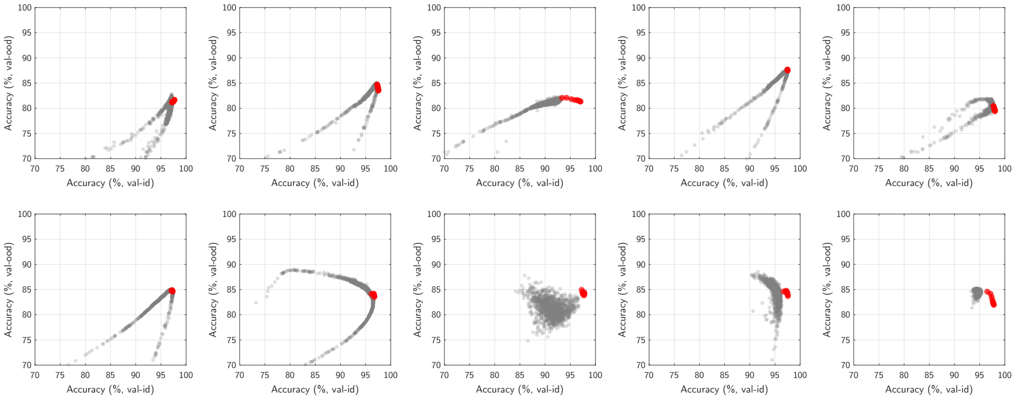

Figure 9: Same as in Figure 8, but using `val-ood` (instead of `test-ood`) as the OOD evaluation set.

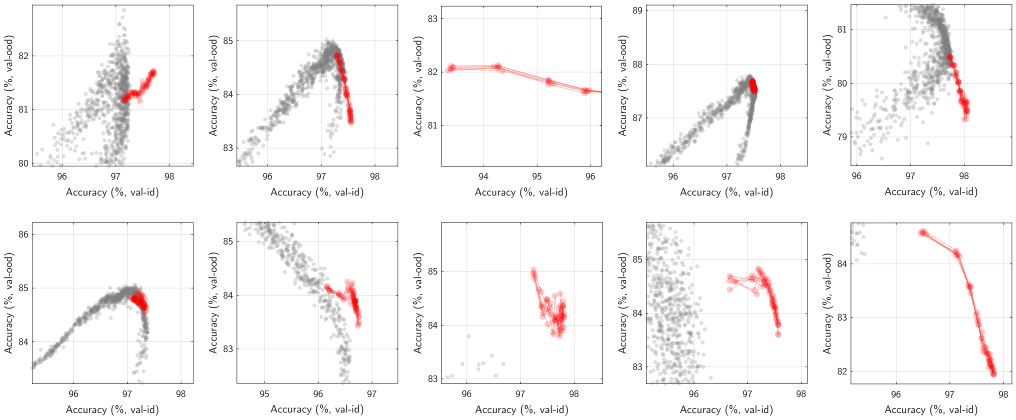

Figure 10: Same as in Figure 9, zoomed-in on ERM models (red dots ●).

C Results on other datasets

In addition to Camelyon17, we performed experiments on five other datasets. We selected these datasets from the current literature on OOD generalization with no a priori knowledge of particular patterns of ID vs. OOD performance. We find inverse patterns to different extent on four out of five.

C.1 WildTime-arXiv

Data. The WildTime-arXiv [53] dataset contains text abstracts from arXiv preprints. The task is to predict each paper's category among 172 classes. The ID and OOD splits are made of data from different time periods.

Methods. We fine-tune a standard BERT-tiny model with a new linear head, using any of these well-known methods: standard ERM, simple class balancing [12], mixup [58], selective mixup [54], and post hoc adjustment for label shift [19] (we did not use the diversification method from Section 3). We repeat every experiment with 10 seeds and record the ID and OOD accuracy at every training epoch. We then plot each of these points in Figure 11 and highlight the epoch of highest ID or OOD accuracy per run (method/seed combination).

Results. As discussed in Section 5, there is a clear trade-off both within methods (i.e. across seeds and epochs) and across methods.

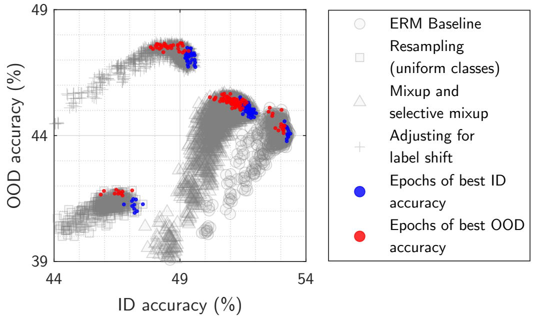

Figure 11: Results on WildTime-arXiv.

C.2 Waterbirds

Data. The **waterbirds** dataset [36] is a synthetic dataset widely used for evaluating OOD generalization. The task is to classify images of birds into 2 classes. The image backgrounds are also of two types, and the correlation between birds and background is reversed across the training and test splits. The standard metric is the worst-group accuracy, where each group is any of the 4 combinations of bird/background.

Methods. We follow the same procedure as described above. We experiment two classes of architectures: ResNet-50 models pretrained on ImageNet and fine-tuned on waterbirds, and linear classifiers trained of features from the same frozen (non-fine-tuned) ResNet-50.

Results. We observe in Figure 12 patterns of inverse correlations in both cases.

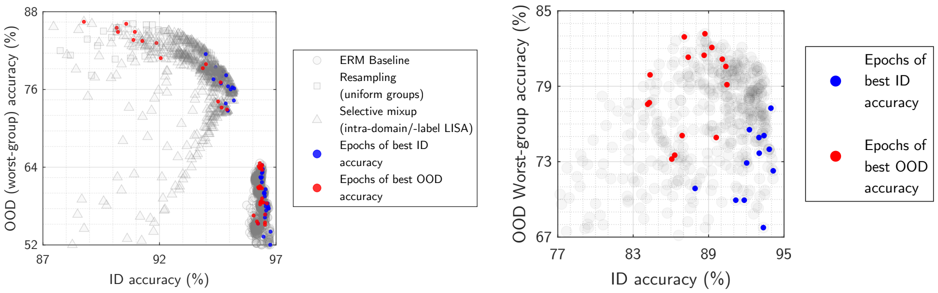

Figure 12: Results on waterbirds with linear probing (left) and fine-tuned ResNet-50 models.

C.3 CivilComments

Data. The CivilComments dataset [15] is another widely-used dataset in OOD research. It contains text comments from internet forums to be classified as toxic or not. Each example is labeled with a topical attribute (e.g. Christian, male, LGBT, etc.) that is spuriously associated with ground truth labels in the training data. The target metric is again worst-group accuracy, where a group is any label/attribute combination.

Methods. We follow the same procedure as described above. We experiment two classes of architectures: pretrained BERT-tiny fine-tuned on CivilComments, and linear classifiers trained of features from the same frozen BERT-tiny models (a.k.a. linear probing).

Results. We observe in Figures 13–14 different patterns with the two classes of architectures. With linear probing, the ID vs. OOD trade-off is minimal, and the model of highest ID performance within a run as well as across methods is very similar to the model of highest OOD performance. With fine-tuning however, the trade-off is more pronounced. The ID and OOD performance usually peak then diminish at different epochs during the fine-tuning. This agrees with previous reports [2] of OOD robustness being progressively lost during fine-tuning.

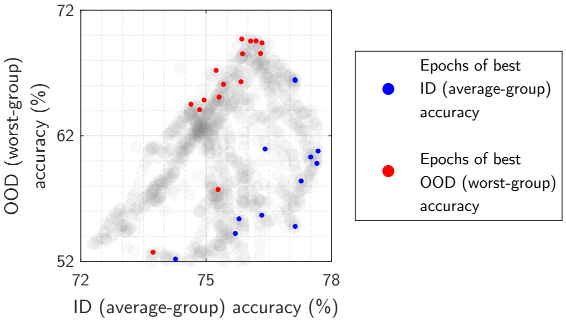

Figure 13: Results on CivilComments with fine-tuned BERT.

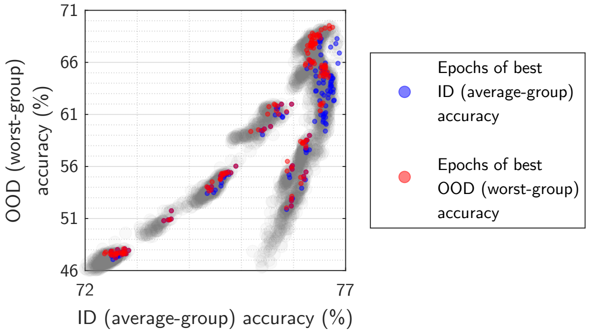

Figure 14: Results on CivilComments with linear probing on frozen BERT embeddings.

C.4 WildTime-MIMIC-Readmission

Data. The WildTime-MIMIC-Readmission [53] dataset contains hospital records (sequences of codes representing diagnoses and treatments) to be classified into two classes, corresponding to the readmission of the patient within a short time. ID and OOD splits contain records from different time periods.

Methods. We follow the same procedure as described above. We train a standard bag-of-embeddings architecture, which associate each diagnosis/treatment with a learned embedding, then summed and fed to a linear classifier. We train this model with standard ERM, and with a resampling to balance the classes in the training data [12], which is a standard approach for imbalanced datasets. We also train models with a “mild balancing”, where classes are sampled according to a distribution half-way between the original one of the training data, and a uniform (50%–50%) one.

Results. In Figure 15 we observe that ID and OOD performance are mostly positively correlated across methods. The best models are obtained with uniform balancing of classes, in which case model selection based on OOD performance could give a small advantage, but it is marginal compared to the improvement over the ERM baseline, which can be clearly detected on both the ID and OOD performance.

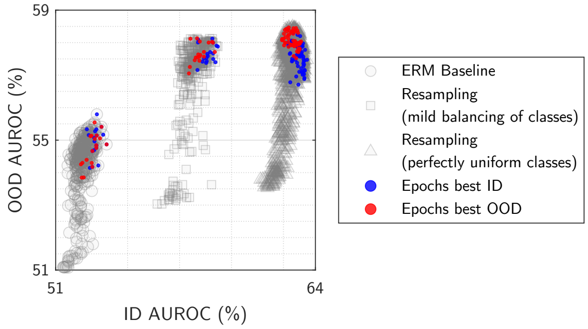

Figure 15: Results on WildTime-MIMIC-Readmission.

C.5 WildTime-Yearbook

Data. The WildTime-Yearbook [53] dataset contains yearbook portraits. Each image is to be classified as male or female, and the ID and OOD splits contain images from different time periods.

Methods. We follow the same procedure as described above. We train the simple CNN architecture described in [53]. We report in Figure 16 both the "average-group" accuracy (over the entire OOD test set) and the "worst-group" accuracy (where a group is any 5-year period within the OOD test period.

Results. The patterns are slightly different in the two cases but similar conclusions can be drawn from both. There is a mostly-positive correlation, but at the highest accuracies (upper-right quadrant), some small trade-off exists. This suggests that fine-grained differences exist that are useful for either ID or OOD generalization, but not both. Although differences are small, this "pointy end" of the spectrum is where the state-of-the-art models compete, hence the relevance of this observation.

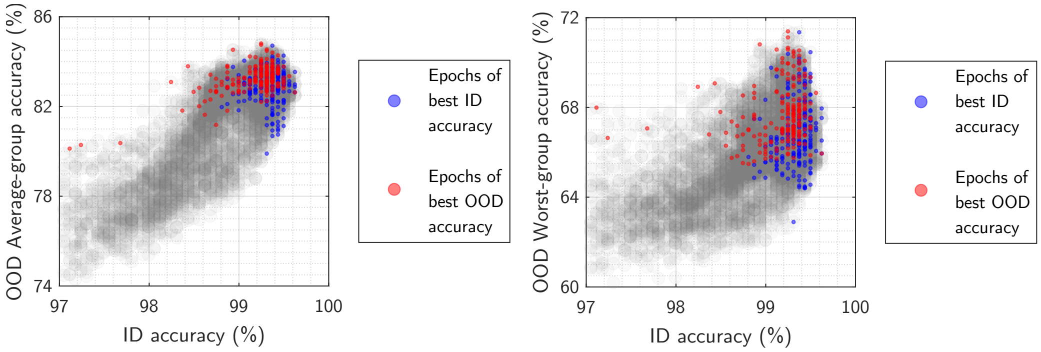

Figure 16: Results on WildTime-Yearbook (left: average-group accuracy, right: worst-group accuracy).

D Proof of Theorem 1

Theorem 1. Including an additional spurious feature leads to the following change in the risks:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}+1}) \;-\; \mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}}) \;&=\; \mathbb{E}^{\mathrm{ID}}[y - \Phi_{\hat{d}}(\boldsymbol{x})^\top \boldsymbol{\beta}_{\hat{d}}^{\mathrm{ID}}]^2 \;-\; \mathbb{E}^{\mathrm{ID}}[y - \Phi_{\hat{d}+1}(\boldsymbol{x})^\top \boldsymbol{\beta}_{\hat{d}+1}^{\mathrm{ID}}]^2 \;<\; 0 \\
\mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}+1}) \;-\; \mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}}) \;&=\; \mathbb{E}^{\mathrm{OOD}}[y - \Phi_{\hat{d}}(\boldsymbol{x})^\top \boldsymbol{\beta}_{\hat{d}}^{\mathrm{OOD}}]^2 \;-\; \mathbb{E}^{\mathrm{OOD}}[y - \Phi_{\hat{d}+1}(\boldsymbol{x})^\top \boldsymbol{\beta}_{\hat{d}+1}^{\mathrm{OOD}}]^2 \;>\; 0 \\
\mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}+1}) \;-\; \mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}}) \;&=\; Q_1 + Q_2 + Q_3
\end{aligned}
$$

with Q_1, Q_2, Q_3 defined as:

$$
\begin{aligned}
Q_1 \;&=\; \mathbb{E}^{\mathrm{OOD}}[y - \Phi_{\hat{d}}(\boldsymbol{x})^\top \beta_{\hat{d}}^{\mathrm{OOD}}]^2 \;-\; \mathbb{E}[y - \Phi_{\hat{d}+1}(\boldsymbol{x})^\top \beta_{\hat{d}+1}^{\mathrm{OOD}}]^2 \\
Q_2 \;&=\; \sum_{i=1}^{\hat{d}} \Bigg[\big(\mathbb{E}^{\mathrm{OOD}}[\Phi_{\hat{d}}(\boldsymbol{x})y]^\top \boldsymbol{v}_i^{\mathrm{OOD},\hat{d}}\big)^2 \, \big(\lambda_i^{\mathrm{OOD},\hat{d}}\big) \, \Big(\frac{1}{\lambda_i^{\mathrm{ID},\hat{d}}} - \frac{1}{\lambda_i^{\mathrm{OOD},\hat{d}}}\Big)^2 \\
&\qquad - \big(\mathbb{E}^{\mathrm{OOD}}[\Phi_{\hat{d}}(\boldsymbol{x})y]^\top \boldsymbol{v}_i^{\mathrm{OOD},\hat{d}+1}\big)^2 \, \big(\lambda_i^{\mathrm{OOD},\hat{d}+1}\big) \, \Big(\frac{1}{\lambda_i^{\mathrm{ID},\hat{d}+1}} - \frac{1}{\lambda_i^{\mathrm{OOD},\hat{d}+1}}\Big)^2 \Bigg] \\
Q_3 \;&=\; \big(\mathbb{E}^{\mathrm{OOD}}[\Phi_{\hat{d}+1}(\boldsymbol{x})y]^\top \boldsymbol{v}_{\hat{d}+1}^{\mathrm{OOD},\hat{d}+1}\big)^2 \, \frac{((\alpha_{\hat{d}+1}^{\mathrm{ID}})^2 \;-\; (\alpha_{\hat{d}+1}^{\mathrm{OOD}})^2)^2}{(\lambda_{\hat{d}+1}^{\mathrm{ID},\hat{d}+1})^2 \, \lambda_{\hat{d}+1}^{\mathrm{OOD},\hat{d}+1}} \quad > \quad 0.
\end{aligned}
$$

Further, if the new feature is sufficiently unstable in the test domain, i.e. if $((\alpha_{\hat{d}+1}^{\mathrm{ID}})^2 - (\alpha_{\hat{d}+1}^{\mathrm{OOD}})^2)^2$ is sufficiently large such that:

$$
|(\alpha_{\hat{d}+1}^{\mathrm{ID}})^2 - (\alpha_{\hat{d}+1}^{\mathrm{OOD}})^2| \quad > \quad \sqrt{\frac{(\lambda_{\hat{d}+1}^{\mathrm{ID},\hat{d}+1})^2 \lambda_{\hat{d}+1}^{\mathrm{OOD},\hat{d}+1}}{\big(\mathbb{E}^{\mathrm{OOD}}[\Phi(\boldsymbol{x})y]^\top \boldsymbol{v}_{\hat{d}+1}^{\mathrm{OOD},\hat{d}+1}\big)^2} \;\; |Q_1 + Q_2|} \;,
$$

then we have $Q_3 > |Q_1 + Q_2|$ and therefore $\mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}+1}) - \mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}}) \;>\; 0\,.$
Let $\boldsymbol{x}_{\hat{d}} := \Phi_{\hat{d}}(\boldsymbol{m}x)[\boldsymbol{x}_{\mathrm{inv},1},\, ...,\, \boldsymbol{x}_{\mathrm{inv},\hat{d}_{\mathrm{inv}}},\, \boldsymbol{x}_{\mathrm{spu},1},\, ...,\, \boldsymbol{x}_{\mathrm{spu},\hat{d}_{\mathrm{spu}}}]$ be the $\hat{d}$ features already selected,

$\boldsymbol{x}_{\hat{d}+1} := \Phi_{\hat{d}+1}(\boldsymbol{m}x)$ the features after adding a new spurious feature $\boldsymbol{x}_{\mathrm{spu},\hat{d}_{\mathrm{spu}}+1}$ to $\boldsymbol{x}_{\hat{d}}$,
$[\lambda_1^{\hat{d}}, \lambda_2^{\hat{d}}, ..., \lambda_{\hat{d}}^{\hat{d}}]$ the eigenvalues of $\mathbb{E}[\boldsymbol{x}_{\hat{d}}^\top \boldsymbol{x}_{\hat{d}}]$ and $[\boldsymbol{v}_1^{\hat{d}}, \boldsymbol{v}_2^{\hat{d}}, ..., \boldsymbol{v}_{\hat{d}}^{\hat{d}}]$ the corresponding eigenvectors.

Assumption 1. The projection of $\mathbb{E}[\boldsymbol{x}_{\hat{d}}^{\intercal}\, \boldsymbol{v}_i^{\hat{d}}]$ on each basis corresponding to feature is non zero, i.e.

$$
\big|\, \mathbb{E}^e[\boldsymbol{x}_{\hat{d}}^{\intercal}\, \boldsymbol{v}_i^{\hat{d}}] \,\big| > 0, \;\; \forall\, e \in \{e_{\mathrm{ID}},\, e_{\mathrm{OOD}}\},\; i \in [d].
$$

This ensures that coefficients of a feature can not be always 0, otherwise we can simply remove it.

Proof. Let β^{ID} and β^{OOD} denote the solution of linear regression in the ID and OOD domains, i.e.,

$$
\beta_{\hat{d}}^{\mathrm{ID}} = \arg\min_{\beta} \mathbb{E}^{\mathrm{ID}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta)^2 \tag{2}
$$

$$
\beta_{\hat{d}}^{\mathrm{OOD}} = \arg\min_{\beta} \mathbb{E}^{\mathrm{OOD}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta)^2 \tag{3}
$$

Now let us compare the OOD loss after we include $\boldsymbol{x}_{\mathrm{spu},\hat{d}_{\mathrm{spu}}+1}$. In practice, we can only obtain β^{ID} and then apply it on both the ID and OOD domains, which elicits the following errors:

$$
\mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}}) = \mathbb{E}^{\mathrm{ID}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta^{\mathrm{ID}}) \tag{4}
$$

$$
\begin{aligned}
\mathcal{L}_{\mathrm{OOD}}(\Phi_{\hat{d}}) &= \mathbb{E}^{\mathrm{OOD}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta^{\mathrm{ID}}) \\
&= \underbrace{\mathbb{E}^{\mathrm{OOD}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta^{\mathrm{ID}}) - \mathbb{E}^{\mathrm{OOD}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta^{\mathrm{OOD}})}_{\xi_1^{\hat{d}}} + \underbrace{\mathbb{E}^{\mathrm{OOD}}(y - \boldsymbol{x}_{\hat{d}}^\top \beta^{\mathrm{OOD}})}_{\xi_2^{\hat{d}}}
\end{aligned} \tag{5}
$$

It is well known that the residual of the linear fitting y by $\boldsymbol{x}_{\hat{d}}$ on the ID domain is

$$
\mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}}) = \mathbb{E}^{\mathrm{ID}}\big[y - \boldsymbol{x}_{\hat{d}} \mathbb{E}^{\mathrm{ID}}[\boldsymbol{x}_{\hat{d}}^\top \boldsymbol{x}_{\hat{d}}]^{-1} \mathbb{E}^{\mathrm{ID}}[\boldsymbol{x}_{\hat{d}} y]\big]^2 = \mathbb{E}^{\mathrm{ID}}[y - \Phi_{\hat{d}}(\boldsymbol{x})^\top \beta_{\hat{d}}^{\mathrm{ID}}]^2, \tag{6}
$$

Similarly, we have

$$\mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}+1}) = \mathbb{E}^{\mathrm{ID}}[y - \boldsymbol{x}_{\hat{d}+1}^\top \beta_{\hat{d}+1}^{\mathrm{ID}}]^2. \tag{7}$$

Since $\boldsymbol{x}_{\mathrm{spu},\hat{d}_{\mathrm{spu}+1}}$ does not lies in the space spaned by $\boldsymbol{x}_{\hat{d}}$, so the space spanned by $\boldsymbol{x}_{\hat{d}+1}$ is strictly larger than $\boldsymbol{x}_{\hat{d}}$.

Together with Assumption 1, we have

$$\mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}}) - \mathcal{L}_{\mathrm{ID}}(\Phi_{\hat{d}+1}) = \mathbb{E}^{\mathrm{ID}}[y - \boldsymbol{x}_{\hat{d}}^\top \beta_{\hat{d}}^{\mathrm{ID}}]^2 - \mathbb{E}^{\mathrm{ID}}[y - \boldsymbol{x}_{\hat{d}+1}^\top \beta_{\hat{d}+1}^{\mathrm{ID}}]^2 > 0, \tag{8}$$

and also

$$\xi_2^{\hat{d}} - \xi_2^{\hat{d}+1} = \mathbb{E}^{\mathrm{OOD}}[y - \boldsymbol{x}_{\hat{d}}^\top \beta_{\hat{d}}^{\mathrm{OOD}}]^2 - \mathbb{E}^{\mathrm{OOD}}[y - \boldsymbol{x}_{\hat{d}+1}^\top \beta_{\hat{d}+1}^{\mathrm{OOD}}]^2 > 0. \tag{9}$$

By the proof in Appendix B.6.3 (above Eq. 29) in [59], we have

$$\xi_1^{\hat{d}} = \sum_i^{\hat{d}} (\mathbb{E}^{\mathrm{OOD}}[\boldsymbol{x}_{\hat{d}}^\top y]^\intercal \boldsymbol{v}_i^{\mathrm{OOD},\hat{d}})^2 \lambda_i^{\mathrm{OOD}} (\frac{1}{\lambda_i^{\mathrm{IID},\hat{d}}} - \frac{1}{\lambda_i^{\mathrm{OOD},\hat{\mathrm{d}}}})^2. \tag{10}$$

By Eq. (20) in [59], we have

$$\lambda_i^{\mathrm{IID},\hat{d}} - \lambda_i^{\mathrm{OOD},\hat{d}} = (\alpha_i^{\mathrm{IID}})^2 - (\alpha_i^{\mathrm{OOD}})^2. \tag{11}$$

So we have:

$$\begin{aligned}
\xi_1^{\hat{d}+1} - \xi_1^{\hat{d}+1} = & \sum_{i=1}^{\hat{d}} \Bigg[\left(\mathbb{E}^{\mathrm{OOD}}[\boldsymbol{x}_{\hat{d}} y]^\top \boldsymbol{v}_i^{\mathrm{OOD},\hat{d}}\right)^2 \left(\lambda_i^{\mathrm{OOD},\hat{d}}\right) \left(\frac{1}{\lambda_i^{\mathrm{ID},\hat{d}}} - \frac{1}{\lambda_i^{\mathrm{OOD},\hat{d}}}\right)^2 \\
& - \left(\mathbb{E}^{\mathrm{OOD}}[\boldsymbol{x}_{\hat{d}+1} y]^\top \boldsymbol{v}_i^{\mathrm{OOD},\hat{d}+1}\right)^2 \left(\lambda_i^{\mathrm{OOD},\hat{d}+1}\right) \left(\frac{1}{\lambda_i^{\mathrm{ID},\hat{d}+1}} - \frac{1}{\lambda_i^{\mathrm{OOD},\hat{d}+1}}\right)^2 \Bigg] \\
& + \left(\mathbb{E}^{\mathrm{OOD}}[\boldsymbol{x}_{\hat{d}+1} y]^\top \boldsymbol{v}_{\hat{d}+1}^{\mathrm{OOD},\hat{d}+1}\right)^2 \frac{((\alpha_{\hat{d}+1}^{\mathrm{ID}})^2 - (\alpha_{\hat{d}+1}^{\mathrm{OOD}})^2)^2}{(\lambda_{\hat{d}+1}^{\mathrm{ID},\hat{d}+1})^2 \, \lambda_{\hat{d}+1}^{\mathrm{OOD},\hat{d}+1}}.
\end{aligned} \tag{12}$$

From Eq. 10 and 12, we have the desired result. □