

Probing Classifiers are Unreliable for Concept Removal and Detection

Abhinav Kumar¹, Chenhao Tan², Amit Sharma¹

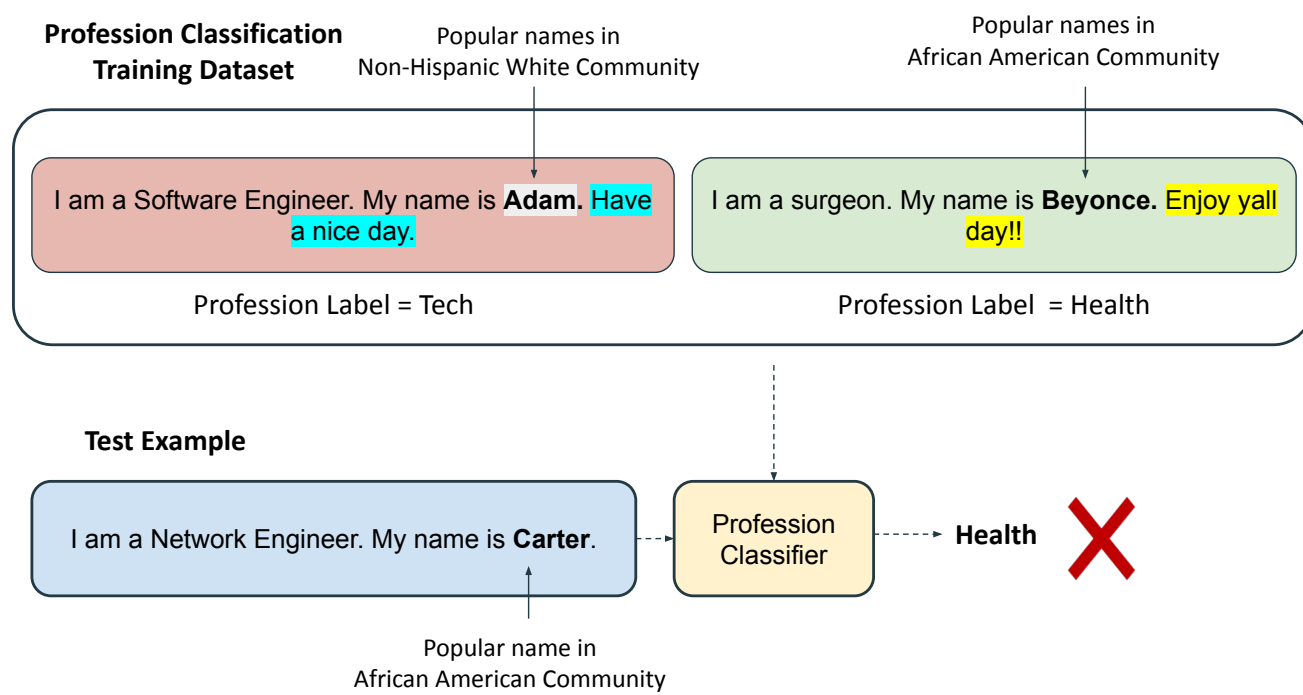
¹Microsoft Research India, ²University of Chicago

¹{t-abkumar, amshar}@microsoft.com, ²chenhao@uchicago.edu



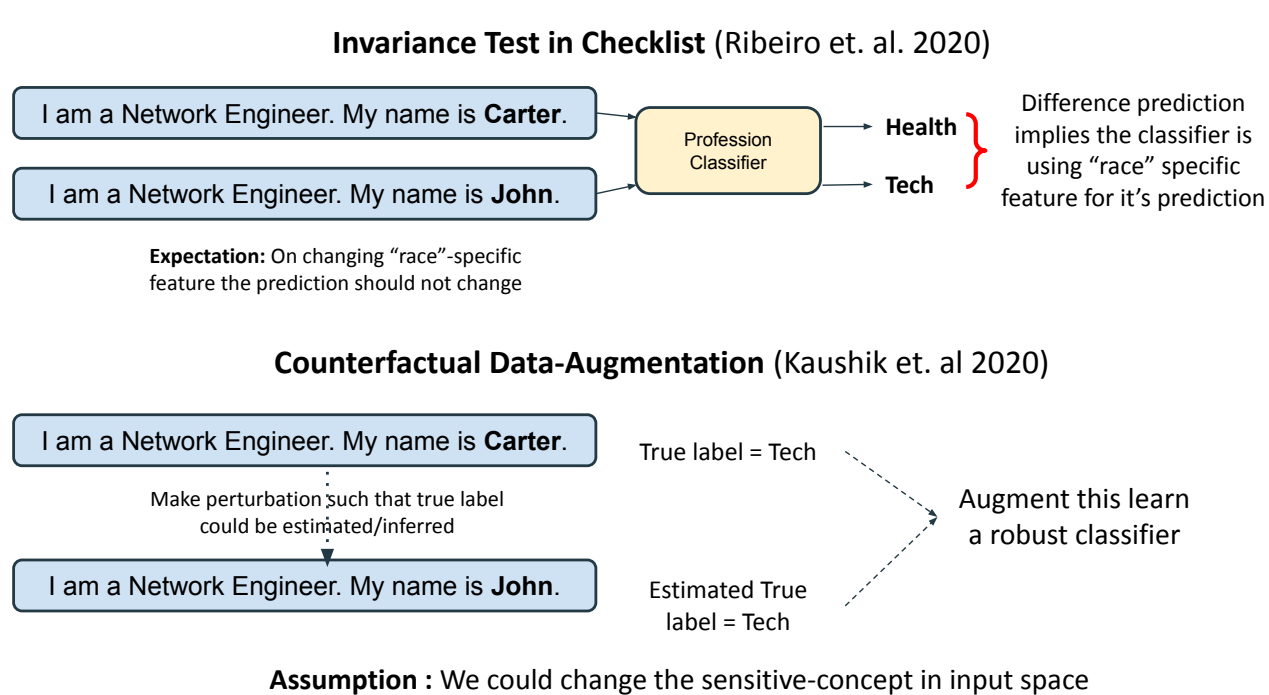
arXiv: 2207.04153

Background



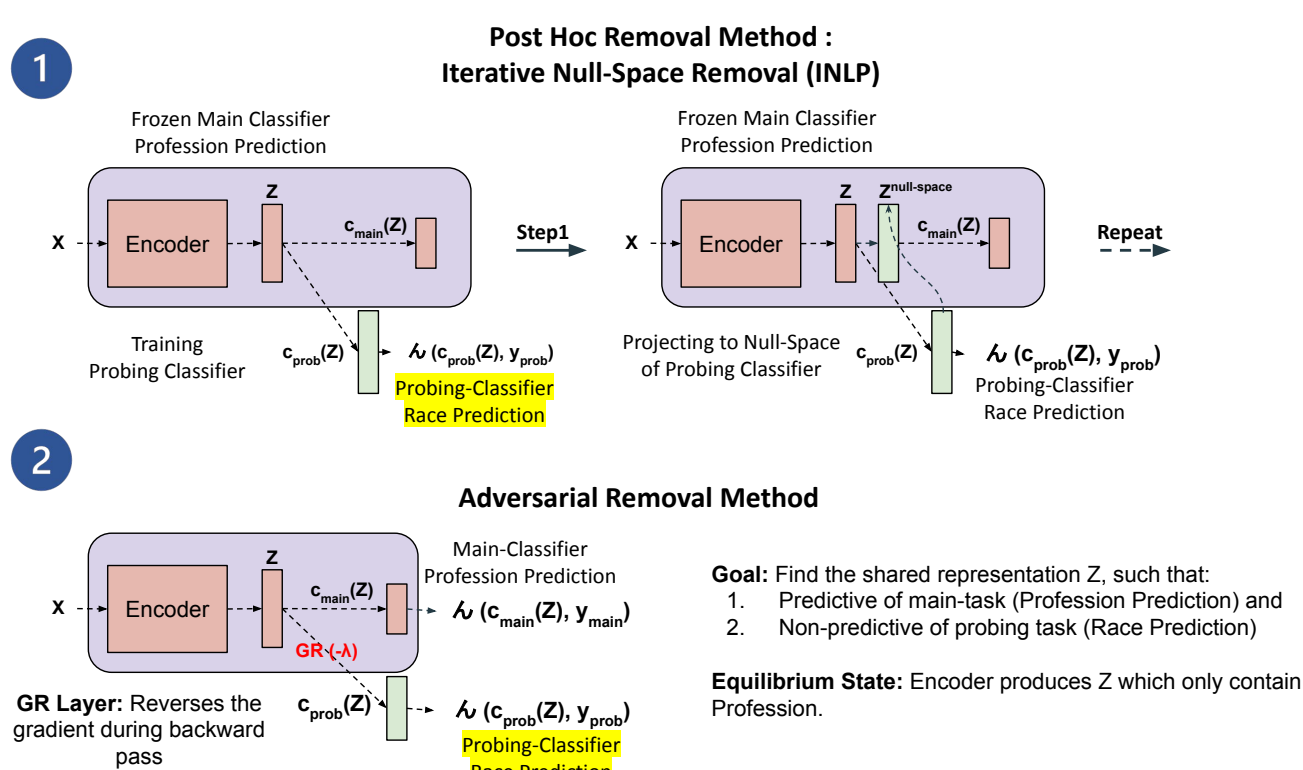
- ML models tend to rely of spurious signal for prediction
- Eg. Race specific names correlated to Profession feature.
- At test time if correlation breaks → wrong prediction.

Input Space Based Removal Methods



Latent Space Based Removal Methods

- Making perturbation in input is not always possible.
- Make perturbation or changes in latent space.
- **Null Space Removal (INLP)**: Removes spurious features by projecting latent space to null-space of spurious feature classifier.
- **Adversarial Removal (ADV)**: Jointly trains main-task and spurious feature classifier adversarially.



Our Contribution

- 1 Any method using Probing classifier will fail.
- 2 Theoretical and Empirical Results showing failure of INLP and Adversarial Removal.

Dataset Description

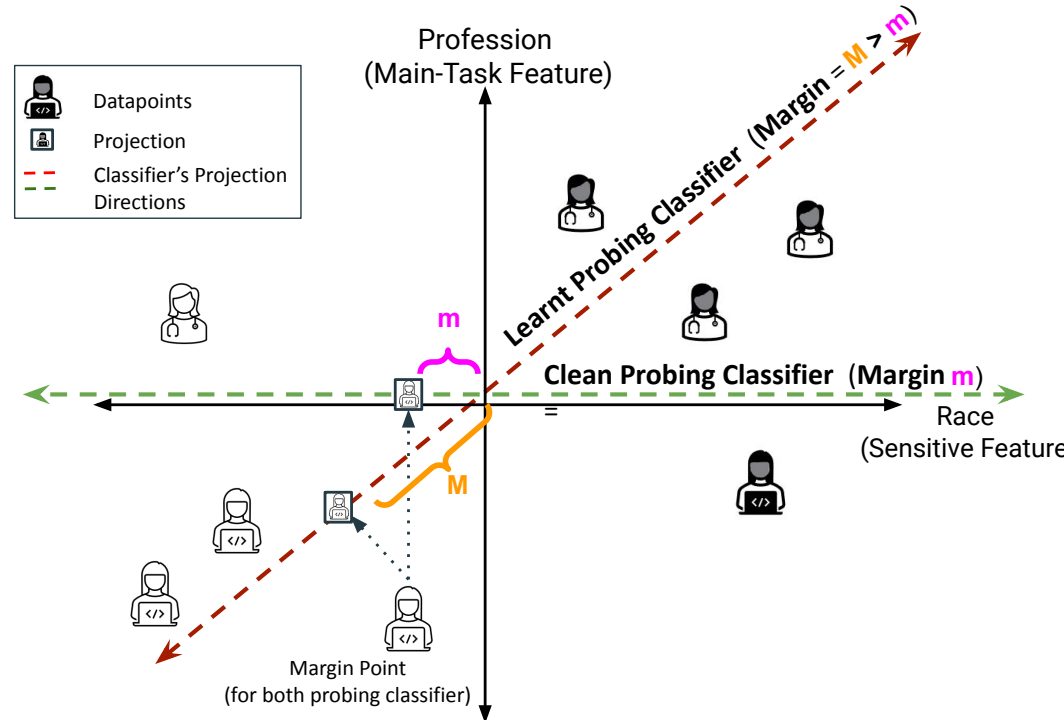
Dataset	Main Task	Spurious Feature
Multi-NLI	Contradiction Prediction	Negation Words
Twitter AAE	Sentiment Prediction	Race
Synthetic-Text	Presence of Numbered Word	Length of Text

Future Direction

- 1 Extending input space based removal method like Counterfactual augmentation and Checklist to non-trivial concepts.
- 2 Focus on debiasing classifier's prediction than representation e.g. in algorithmic fairness literature.

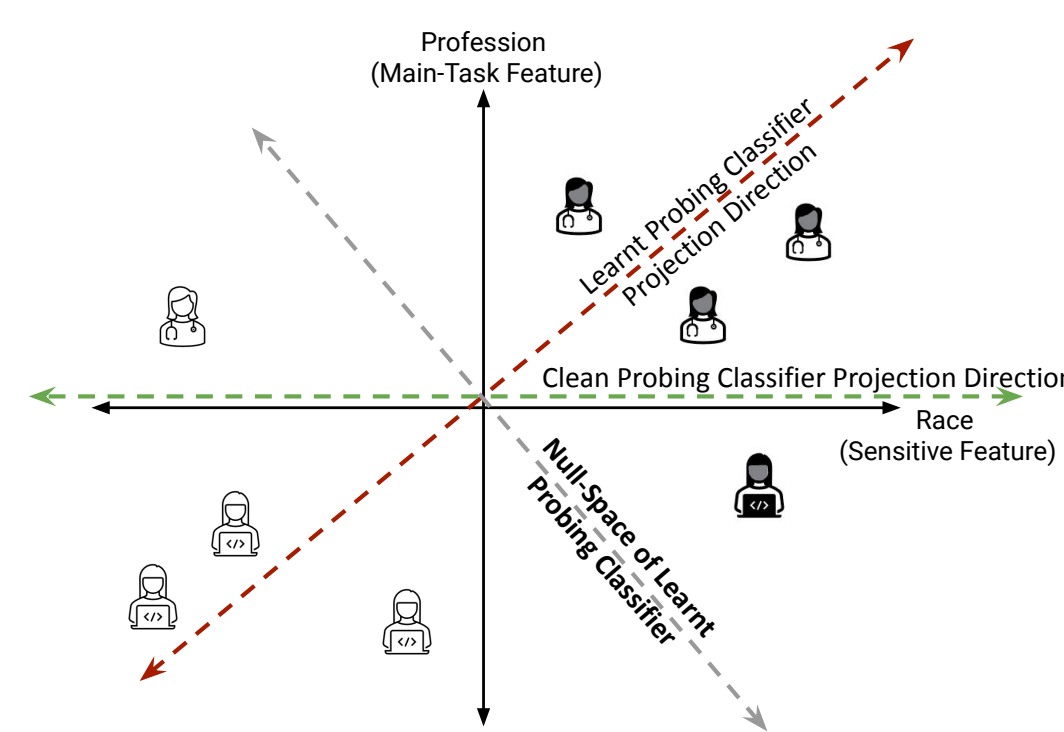
Example: Probing Failure

Goal: Learn a *clean* probing (Race) classifier which just uses probing (Race) feature (green dashed line).



Observation: Slanted *unclean* classifier (red) is better than *clean* probing classifier (green) when trained with max-margin objective.

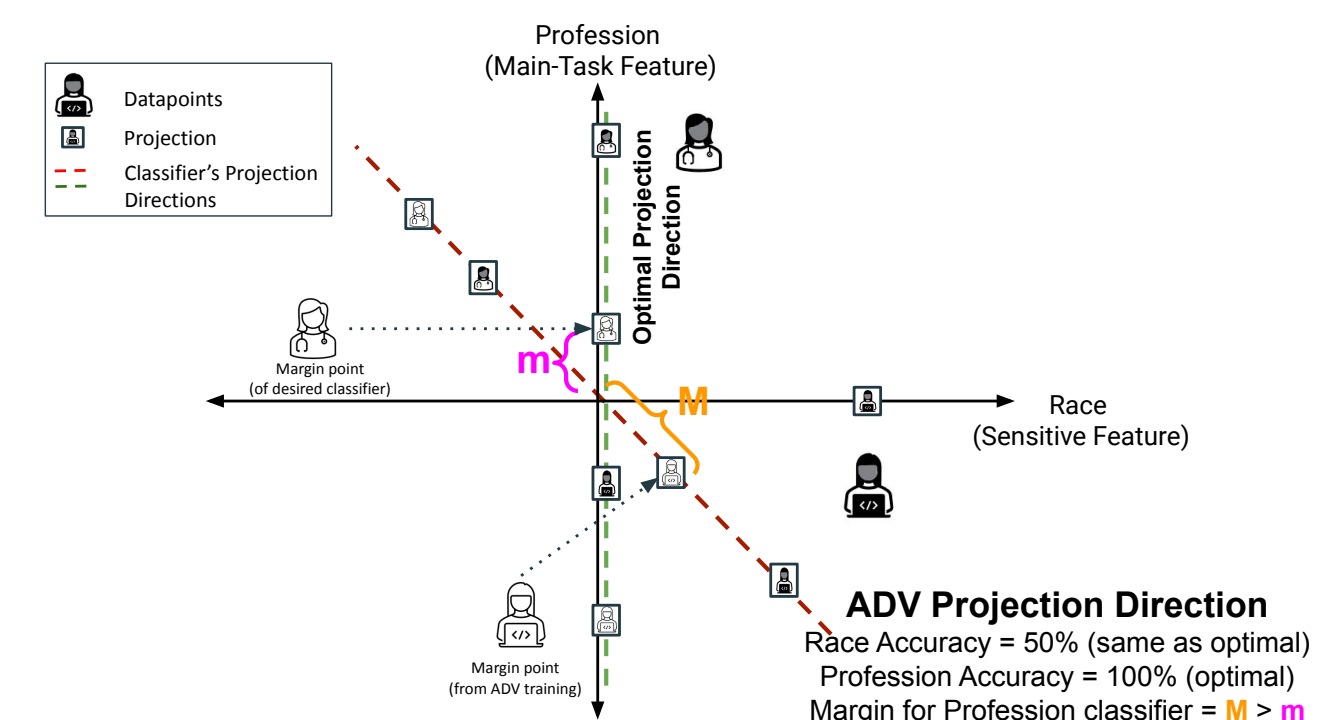
Example: INLP Probing Classifier



- From **Lemma 2.1** → Probing Classifier is slanted
- Null-Space of Probing Classifier is also slanted (wrong).

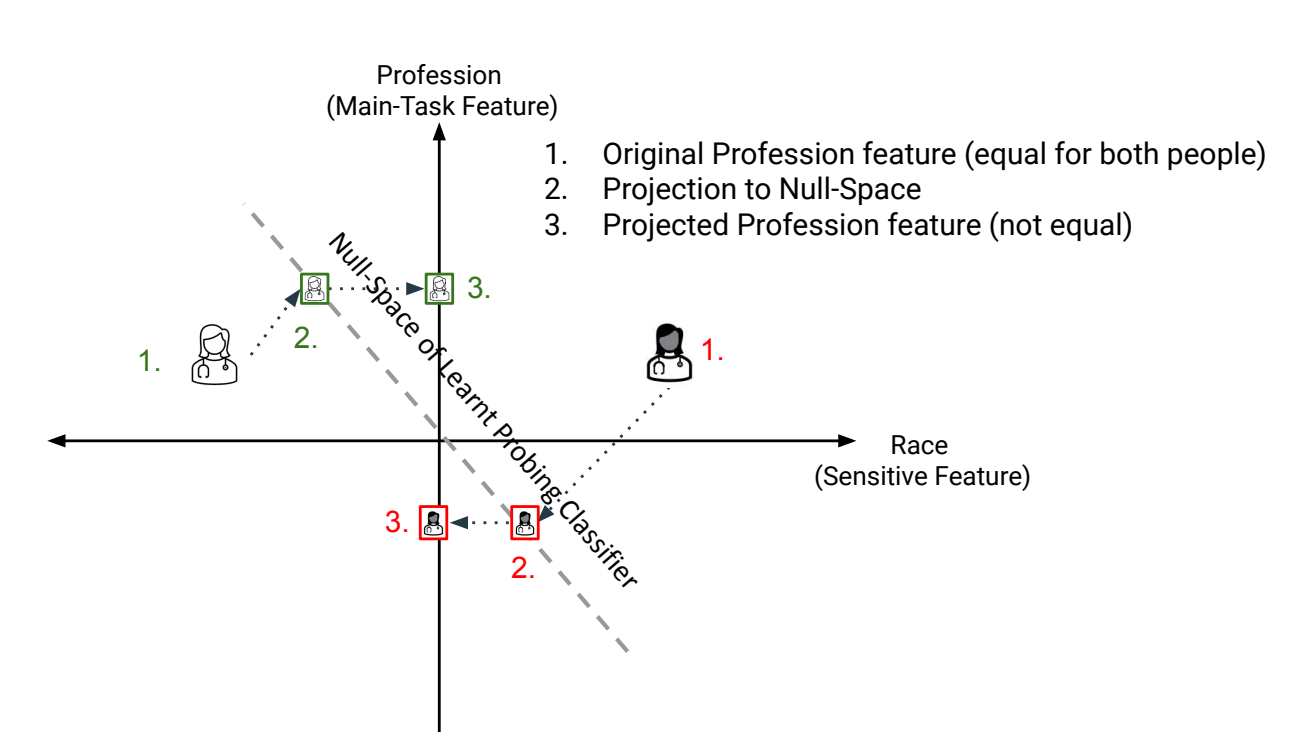
Example: ADV Removal Failure

Goal: Learn shared 1D latent representation s.t. only main-task feature (Profession) is present.



Observation: Slanted Projection Direction (red) better in Main-task objective and equivalent in Adversarial Objective than desired direction (green).

Example: INLP Feature Corruption



- Two individual with same profession but different race.
- Projecting to *wrong* Null Space → inverts Profession.

INLP Empirical Result

Expectation: Clean Main-Task classifier is given as input to INLP, so it should have no effect on main-task classifier.

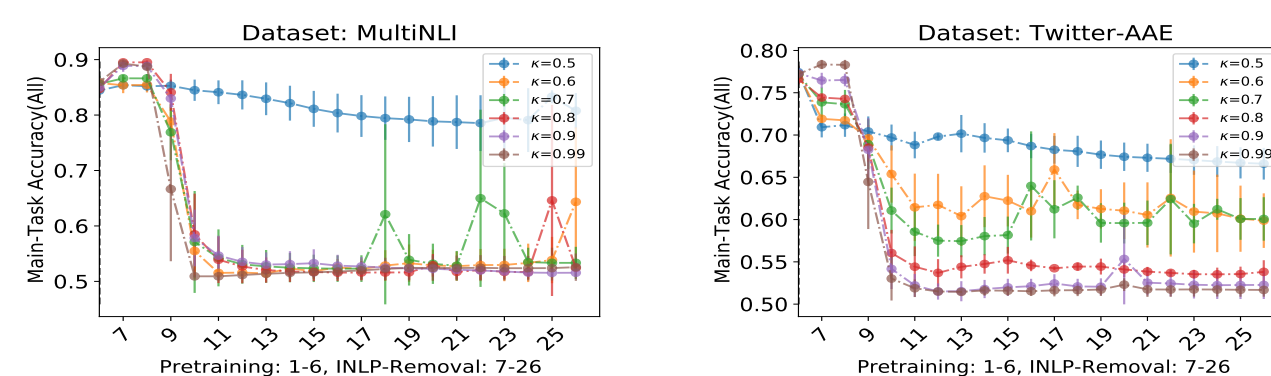


Figure 1: Variation of Main Task accuracy with INLP steps. Main-Task Accuracy goes to random guess as INLP proceeds.

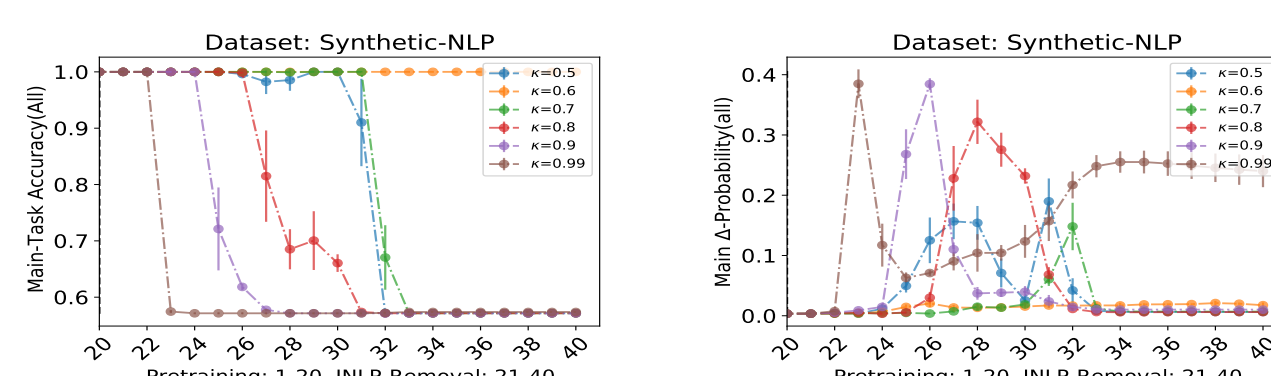


Figure 2: Measuring ΔProb i.e Change in prediction probability of classifier by changing the spurious feature

Early Stopping Doesn't Help: ΔProb increases in the initial phase of INLP. Stopping early could lead to relatively more *unclean* classifier.

ADV Empirical Result

Metric: Post Adversarial Training accuracy on subset of data where spurious correlation breaks (minority group).

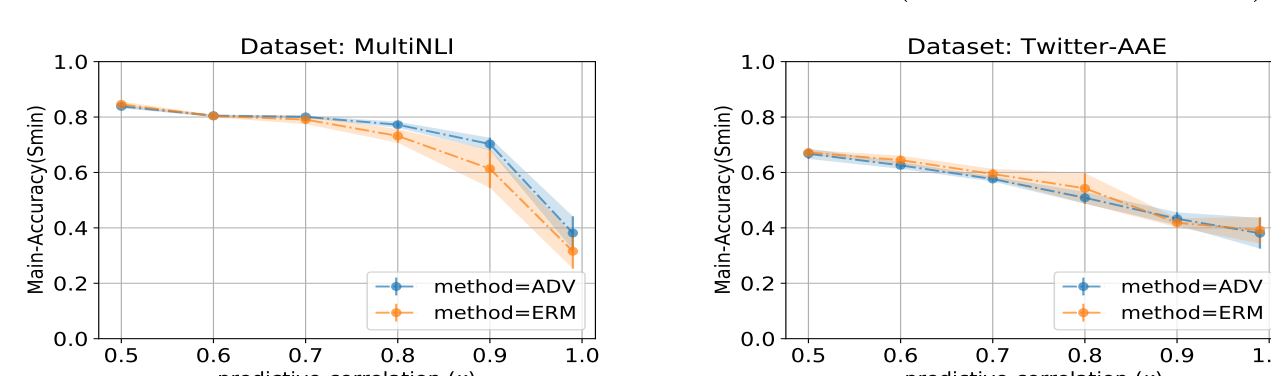


Figure 3: Variation of Main-Task classifier's accuracy on minority group as we vary the degree of correlation between main and spurious feature. Little to no improvement in minority group accuracy across dataset and degree of correlation.

Assumptions

- 1 Latent Space is disentangled and frozen.
- 2 Probing feature is fully (100%) predictive.
- 3 Main-Task feature is linearly separable w.r.t. probing label for the margin point of *clean* probing classifier.

Theory: Probing Failure

Lemma 2.1 (Informal) Given Assm 1,2,3 is satisfied:

- $c_{\text{prob}}(\mathbf{z}) = \mathbf{w}_{\text{prob}} \cdot \mathbf{z}_{\text{prob}} + \mathbf{w}_{\text{main}} \cdot \mathbf{z}_{\text{main}}$ where $\mathbf{w}_{\text{main}} \neq \mathbf{0}$.
- Generalized version for any classifier in paper.

Theory: ADV Failure

Theorem 2.3 (Informal) Given adversarial removal methods is just training the last layer. Then there exist an *unclean* shared latent representation (\mathbf{Z}) s.t.:

- $\text{Margin}^{\text{main}}(\text{unclean } \mathbf{Z}) = \text{Margin}^{\text{main}}(\text{clean } \mathbf{Z})$.
- $\text{Acc}^{\text{adv}}(\text{unclean } \mathbf{Z}) = \text{Acc}^{\text{adv}}(\text{clean } \mathbf{Z})$.
- $\mathcal{L}^{\text{adv}}(\text{unclean } \mathbf{Z}) > \mathcal{L}^{\text{adv}}(\text{clean } \mathbf{Z})$, when main-task and probing labels are correlated and probing feature is more *useful* for main-task than probing task for the margin point of *clean* main-task classifier.

Theory: INLP Failure

Theorem 2.2 (Informal) Given the probing classifier used by INLP is trained using max-margin objective (Lemma 2.1), following happens:

Mixing or Damage: After first step of INLP we have:

- if $\mathbf{w}_{\text{prob}} = \mathbf{0}$, $\mathbf{z}_{\text{main}}^{\text{after}} \neq \mathbf{z}_{\text{main}}^{\text{before}}$ and $\mathbf{z}_{\text{prob}}^{\text{after}} = \mathbf{z}_{\text{prob}}^{\text{before}}$; else,
- $\mathbf{z}_{\text{main}}^{\text{after}} = \psi(\mathbf{z}_{\text{main}}^{\text{before}}, \mathbf{z}_{\text{prob}}^{\text{before}})$ and $\mathbf{z}_{\text{prob}}^{\text{after}} = \phi(\mathbf{z}_{\text{main}}^{\text{before}}, \mathbf{z}_{\text{prob}}^{\text{before}})$
- Mixing is non-invertible in subsequent step of INLP.

Destruction: In long term:

- $\|\mathbf{Z}\|$ decreases after every step.
- Continued removal leads to complete destruction of \mathbf{Z} .