Linear Adversarial Concept Erasure

Shauli Ravfogel, Michael Twiton, Yoav Goldberg and Ryan Cotterell

Motivation

Neural models learn rich representations

Motivation

Neural models learn rich representations

But can we control their content?

Controlled Representation Learning

- Often, we want to make sure some concept is *not* encoded.

Controlled Representation Learning

- Often, we want to make sure some concept is *not* encoded.
  - Word embeddings without tense distinctions

Controlled Representation Learning

- Often, we want to make sure some concept is *not* encoded.
  - Word embeddings without tense distinctions
  - Sensitivity to content, but not to style

Controlled Representation Learning

- Often, we want to make sure some concept is *not* encoded.
  - Word embeddings without tense distinctions
  - Sensitivity to content, but not to style
  - **Representations that do not leak protected attributes**

The linear concept subspace hypothesis

Useful use case: the concept lives in low-dimensional **subspace** within the representation space.

The linear concept subspace hypothesis

Useful use case: the concept lives in low-dimensional **subspace** within the representation space.

How can we identify the concept subspace?

Formulation

We formulate an **adversarial game** between a projection matrix $P \in \mathcal{P}_k$ that tries to remove the information, and a predictor $\boldsymbol{\theta} \in \Theta$ that tries to recover it.

Formulation

We formulate an **adversarial game** between a projection matrix $P \in \mathcal{P}_k$ that tries to remove the information, and a predictor $\boldsymbol{\theta} \in \Theta$ that tries to recover it.

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{P \in \mathcal{P}_k} \sum_{n=1}^{N} \ell\Big(y_n,\, g^{-1}\Big(\boldsymbol{\theta}^\top P \boldsymbol{x}_n\Big)\Big)$$

Formulation

We formulate an **adversarial game** between a projection matrix $P \in \mathcal{P}_k$ that tries to remove the information, and a predictor $\boldsymbol{\theta} \in \Theta$ that tries to recover it.

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{P \in \mathcal{P}_k} \sum_{n=1}^{N} \ell\left(y_n,\, g^{-1}\left(\boldsymbol{\theta}^\top P \boldsymbol{x}_n\right)\right)$$

Projection matrix that tries to maximize the loss

Formulation

We formulate an **adversarial game** between a projection matrix $P \in \mathcal{P}_k$ that tries to remove the information, and a predictor $\boldsymbol{\theta} \in \Theta$ that tries to recover it.

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{P \in \mathcal{P}_k} \sum_{n=1}^{N} \ell\left(y_n,\, g^{-1}\left(\boldsymbol{\theta}^\top P \boldsymbol{x}_n\right)\right)$$

Predictor that tries to minimize it

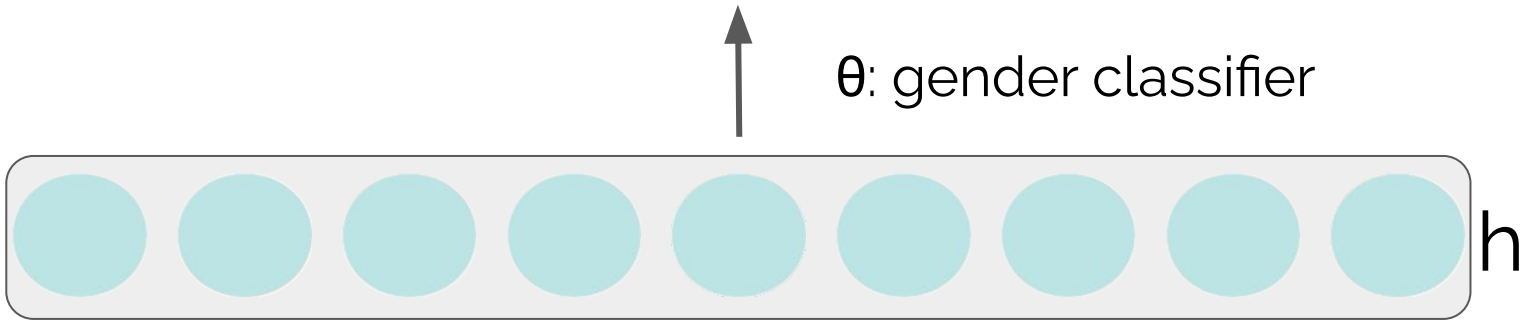
Projection matrix that tries to maximize the loss

Side trip: Iterative Nullspace Projection (INLP)

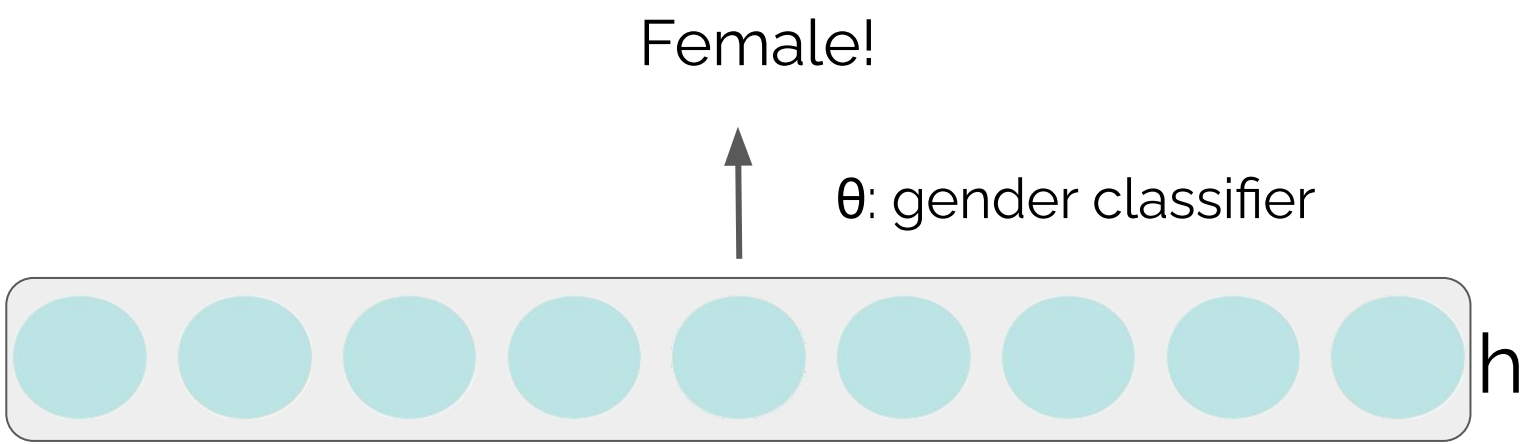
Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel[1,2] **Yanai Elazar**[1,2] **Hila Gonen**[1] **Michael Twiton**[3] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Independent researcher
{shauli.ravfogel, yanaiela, hilagnn, mtwito101, yoav.goldberg1@gmail.com

Nullspace projections

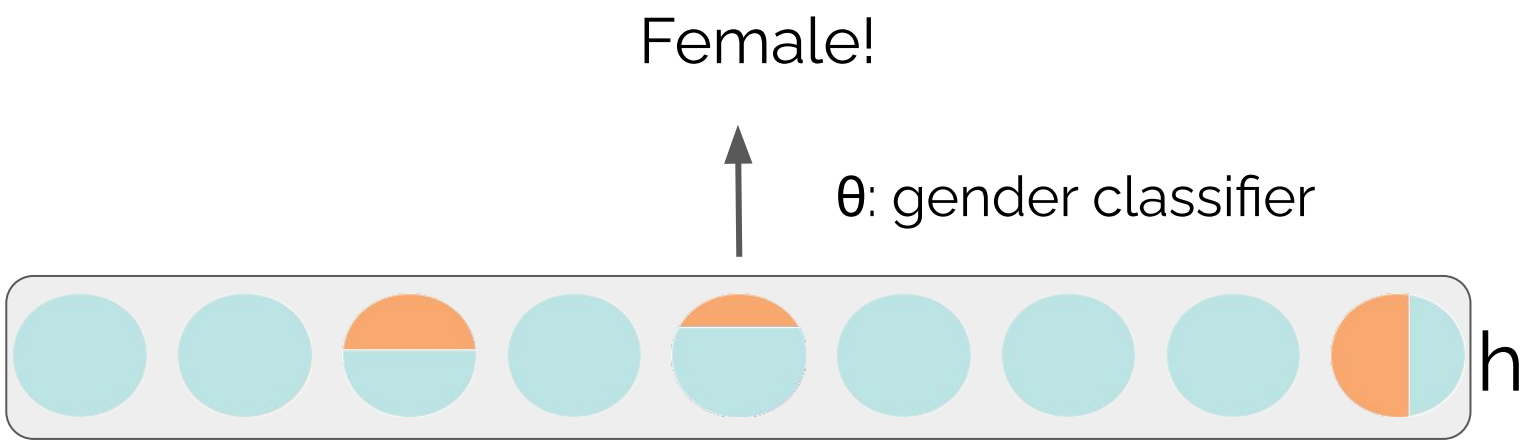


Nullspace projections



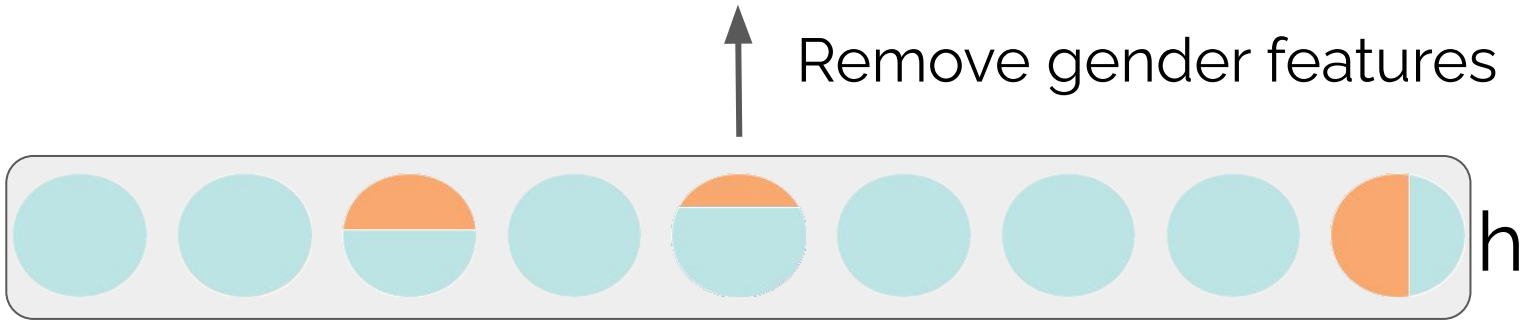
Nullspace projections

Features that θ finds indicative of gender



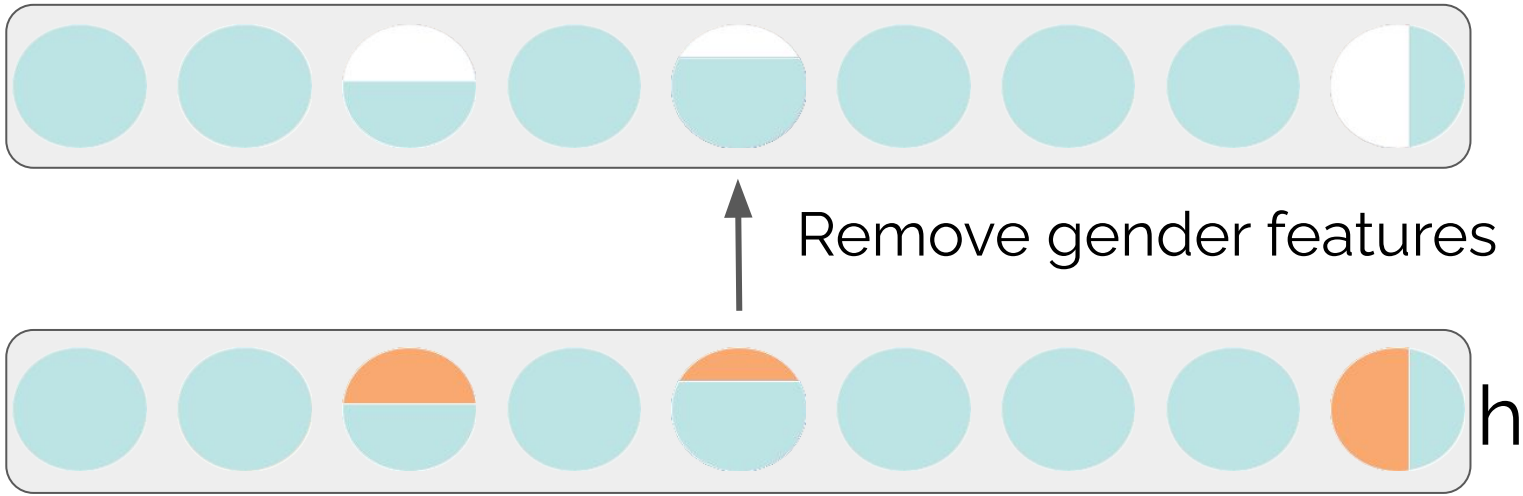
Nullspace projections

Features that θ finds indicative of gender



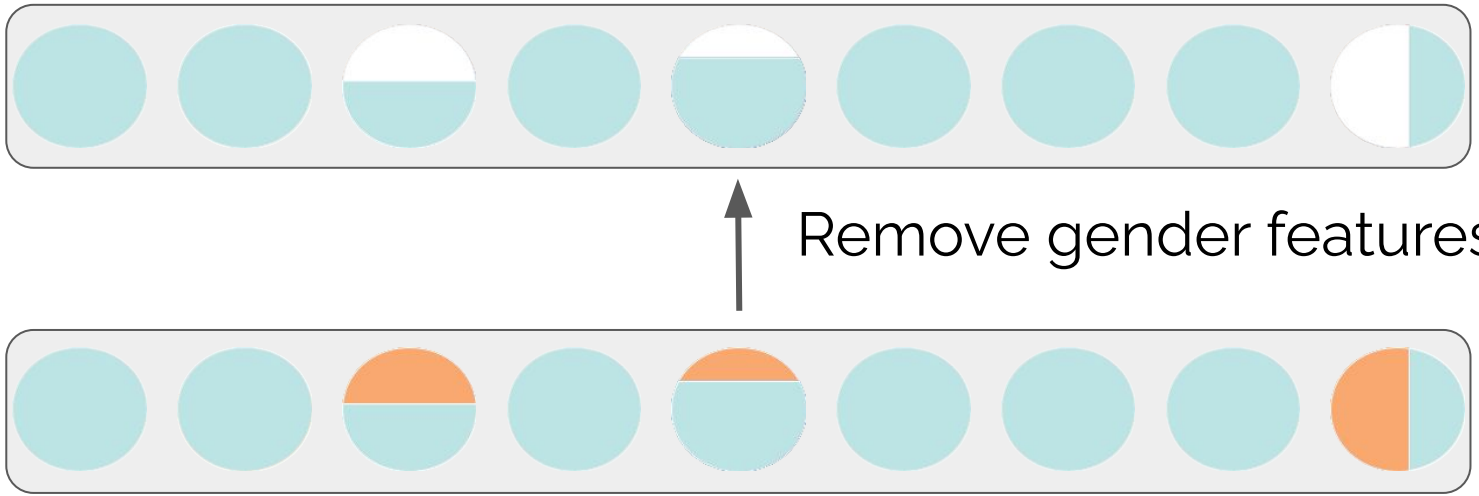
Nullspace projections

Features that θ finds indicative of gender



Nullspace projections

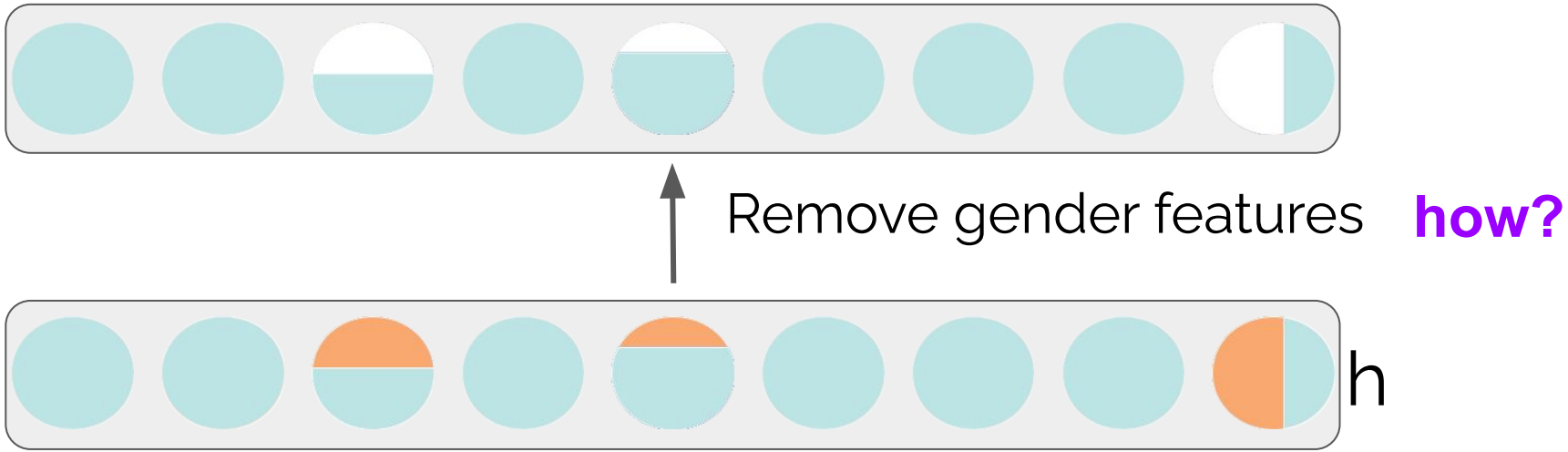
Features that θ finds indicative of gender



Nullspace projections

Features that θ finds indicative of gender

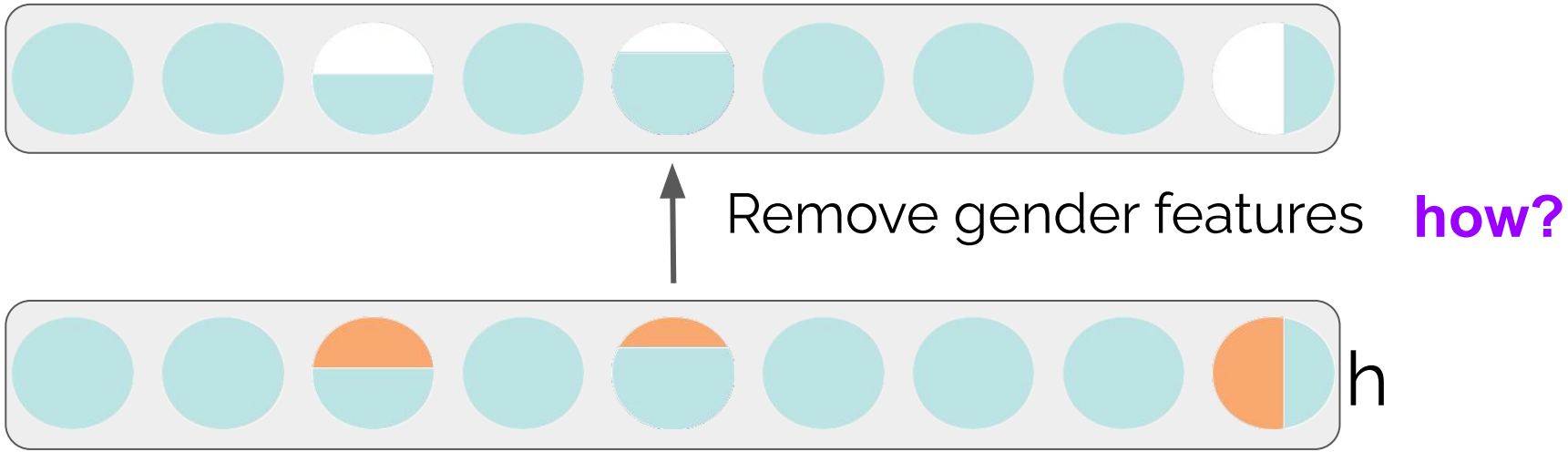
Project h to the orthogonal complement of θ



Nullspace projections

Features that θ finds indicative of gender

Project h to the orthogonal complement of θ **Iteratively**



Special Cases

We treat several important special cases of this objective:

Special Cases

We treat several important special cases of this objective:

1. Linear regression

Special Cases

We treat several important special cases of this objective:

1. Linear regression
2. Rayleigh quotient losses

Special Cases

We treat several important special cases of this objective:

1. Linear regression
2. Rayleigh quotient losses
3. **Classification**

Special Cases

We treat several important special cases of this objective:

1. Linear regression
2. Rayleigh quotient losses
3. **Classification**

Closed-form solution (details in the paper)

Special Cases

We treat several important special cases of this objective:

1. Linear regression
2. Rayleigh quotient losses

Closed-form solution (details in the paper)

3. **Classification**

Gradient-based optimization

Classification Case

The loss is an arbitrary classification loss (hinge, logistic, etc).

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \max_{P \in \mathcal{P}_k} \sum_{n=1}^{N} y_n \log \frac{\exp \boldsymbol{\theta}^\top P \boldsymbol{x}_n}{1 + \exp \boldsymbol{\theta}^\top P \boldsymbol{x}_n}$$

Convex Relaxation (RLACE)

The problem is nonconvex in P.

Convex Relaxation (RLACE)

The problem is nonconvex in P.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \max_{\cancel{P \in \mathcal{P}_k} \; P \in \mathcal{F}_k} \sum_{n=1}^{N} y_n \log \frac{\exp \boldsymbol{\theta}^\top P \boldsymbol{x}_n}{1 + \exp \boldsymbol{\theta}^\top P \boldsymbol{x}_n}$$

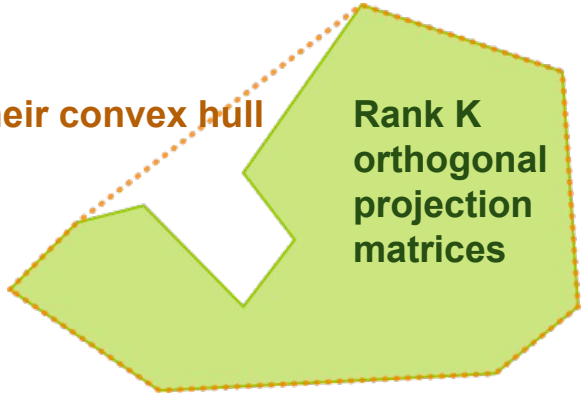
Convex Relaxation (RLACE)

The problem is nonconvex in P.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \max_{\cancel{P \in \mathcal{P}_k} \; P \in \mathcal{F}_k} \sum_{n=1}^{N} y_n \log \frac{\exp \boldsymbol{\theta}^\top P \boldsymbol{x}_n}{1 + \exp \boldsymbol{\theta}^\top P \boldsymbol{x}_n}$$

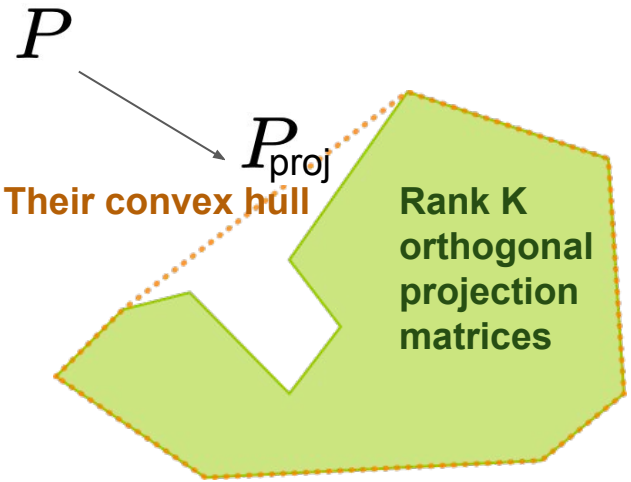
Where we define:

$$\mathcal{F}_k = \text{conv}\left(\mathcal{P}_k\right)$$



Convex Relaxation (RLACE)

In training, we optimize over an arbitrary matrix P, and we project it to the convex hull of orthogonal projection matrices at each step.



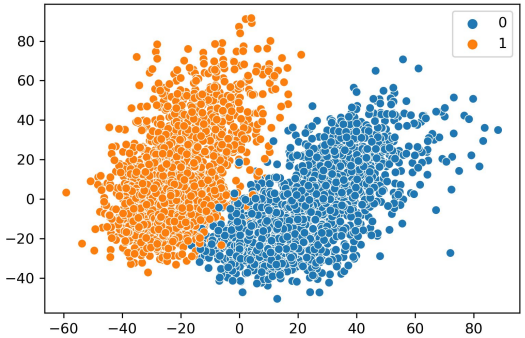
Experimental Evaluation

We conduct experiments on GloVe embeddings and on contextualized representations of short biographies (annotated for both gender and profession).

Results

PCA: less clustering of representations by gender

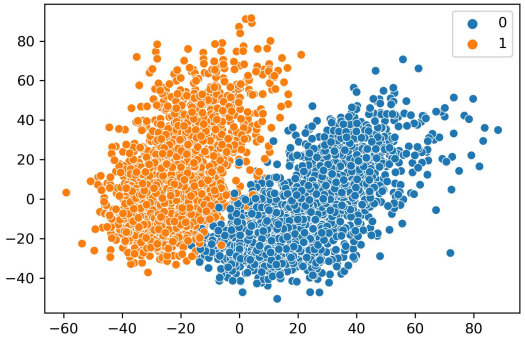
Original



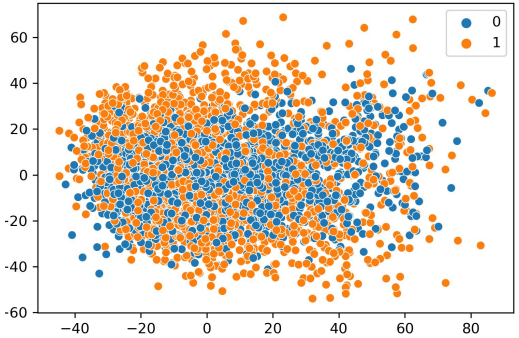
Results

PCA: less clustering of representations by gender

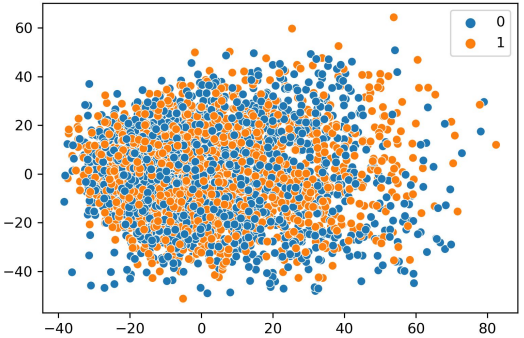
Original



Rank 1 - ours



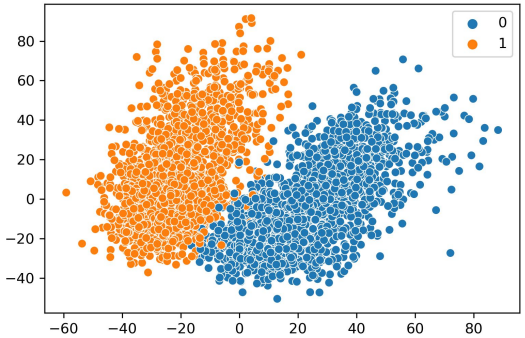
Rank 4 - ours



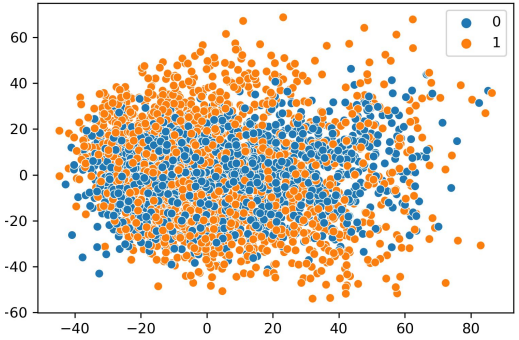
Results

PCA: less clustering of representations by gender

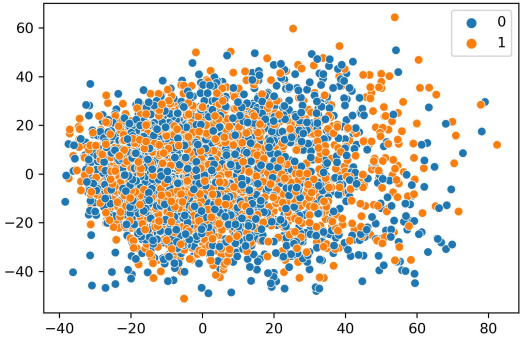
Original



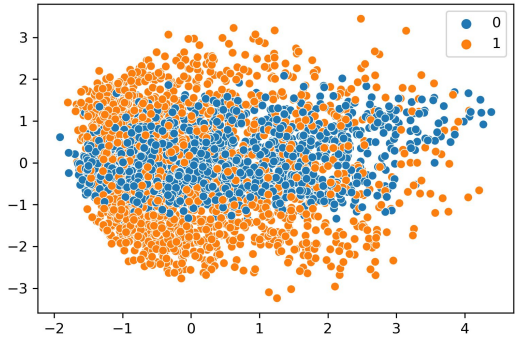
Rank 1 - ours



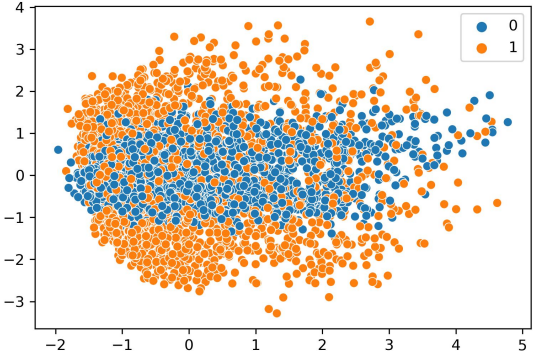
Rank 4 - ours



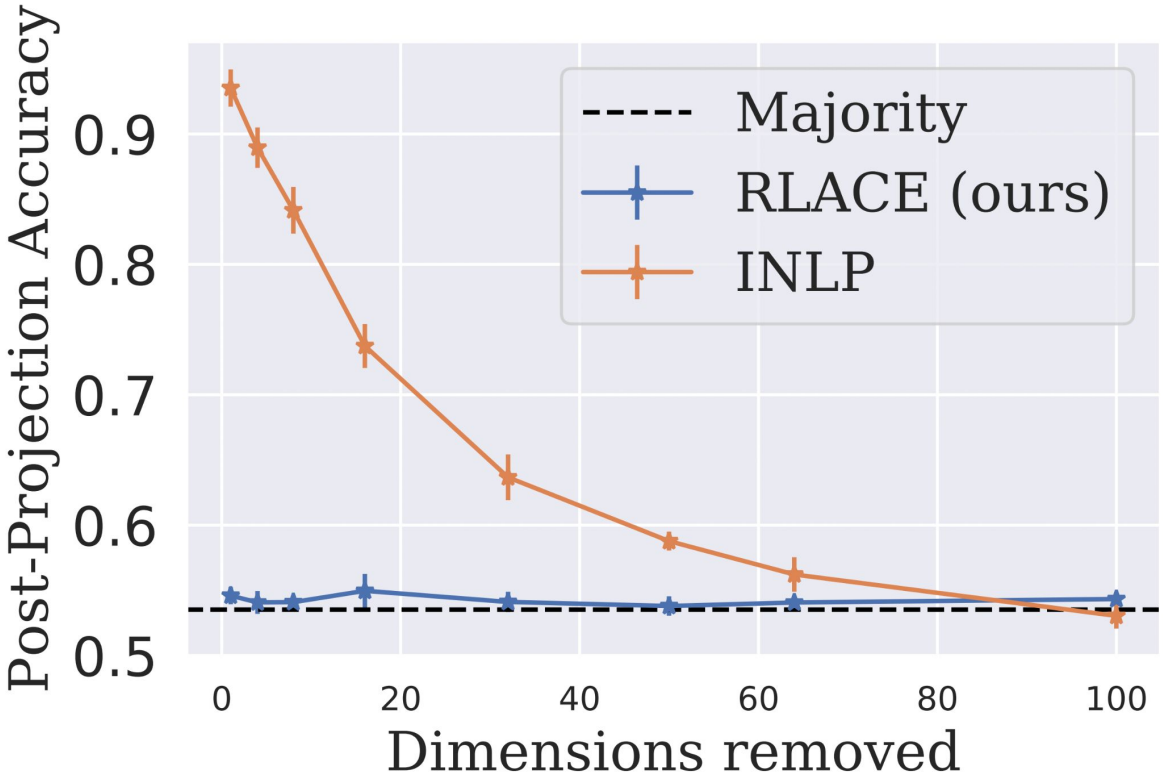
Rank 1 - INLP



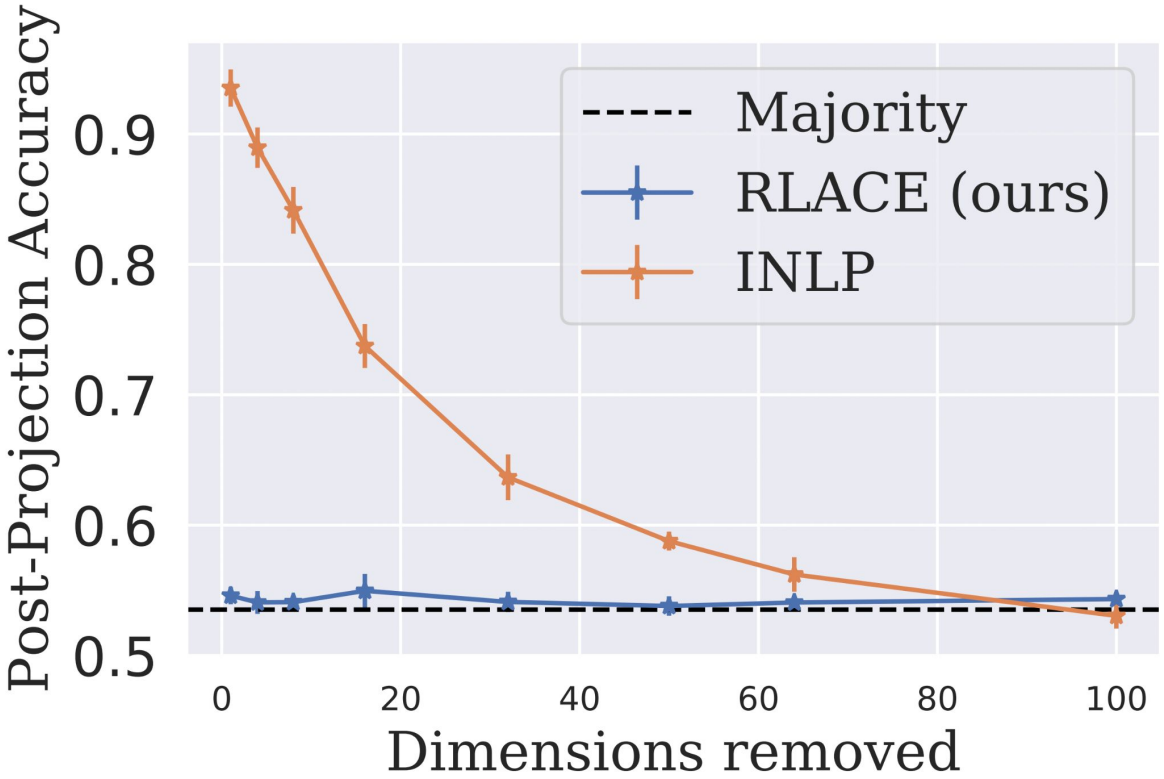
Rank 4 - INLP



Comparison with INLP - BERT (bios data)



Comparison with INLP - BERT (bios data)



Additional results in the paper.

Application on Images






Original

Application on Images

Application on Images

Conclusions

- We have formally defined the problem of removing linear “concept subspaces”

Conclusions

- We have formally defined the problem of removing linear “concept subspaces”
- We present analytical solutions in some cases, and provide a relaxation which works well for others.

Conclusions

- We have formally defined the problem of removing linear “concept subspaces”
- We present analytical solutions in some cases, and provide a relaxation which works well for others.
- High level conclusion: it’s sometimes valuable to constrain our model (in contrast to the “more parameters is better” trend)

Thanks!