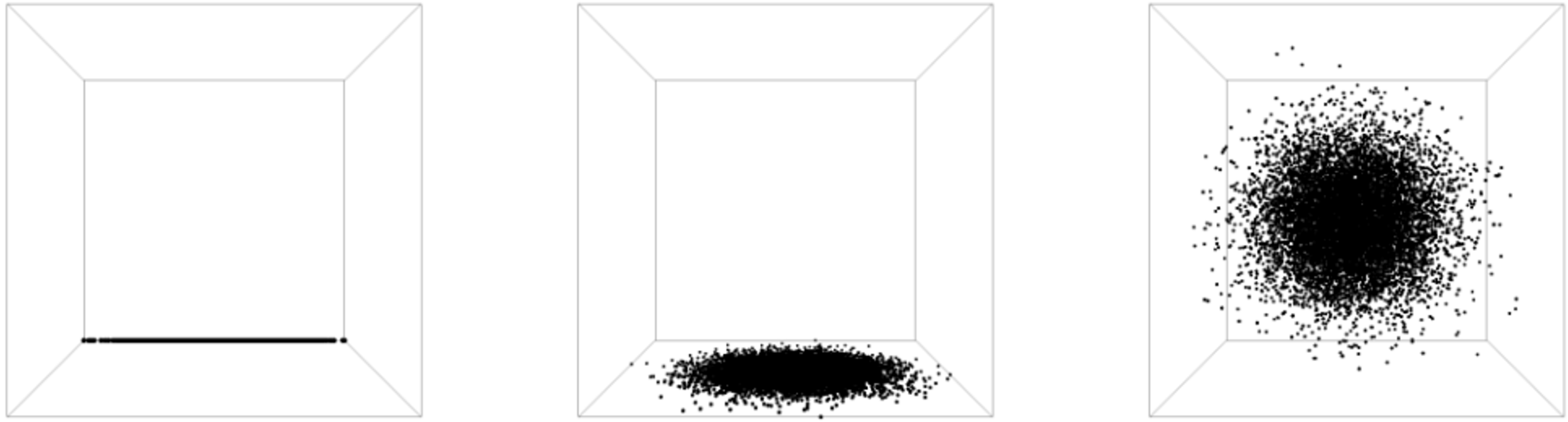
Stable Anisotropic Regularization

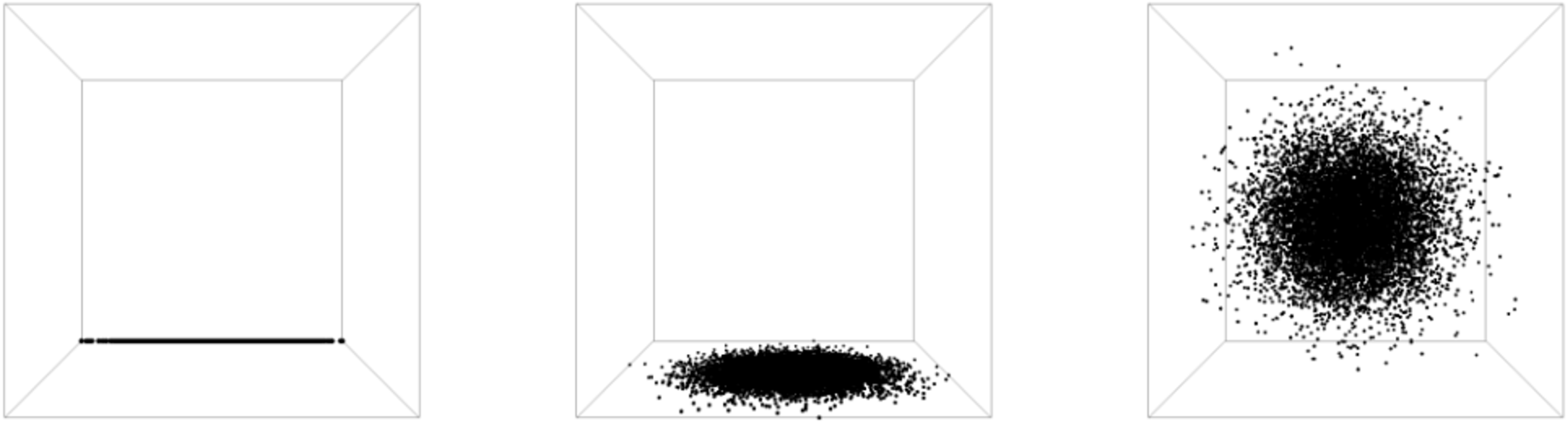
William Rudman[1] & Carsten Eickhoff[2]

Department of Computer Science, Brown University[1]
School of Medicine, University of Tübingen[2]



Defining Isotropy

Isotropy: a distribution is *isotropic* if the variance of the data is uniformly distributed (i.e. the covariance matrix is proportional to the identity matrix).



Left: line embedded in 3D space. **Middle**: circle embedded in 3D space. **Right**: Sphere in 3D space.

Geometry of LLM Embeddings

- **<u>Narrow Cone Hypothesis</u>**:
  - Average random cosine similarity approaches 1 (Ethayarajh 2019).
  - Limits expressiveness of textual representations (Zhang et al. 2020; Timkey & Van Schijndel et al. 2021).

- **<u>Limits Downstream Performance:</u>**
  - Making embeddings more uniformly distributed improves performance (Rajaee et al. 2021; Zhou et al. 2021).

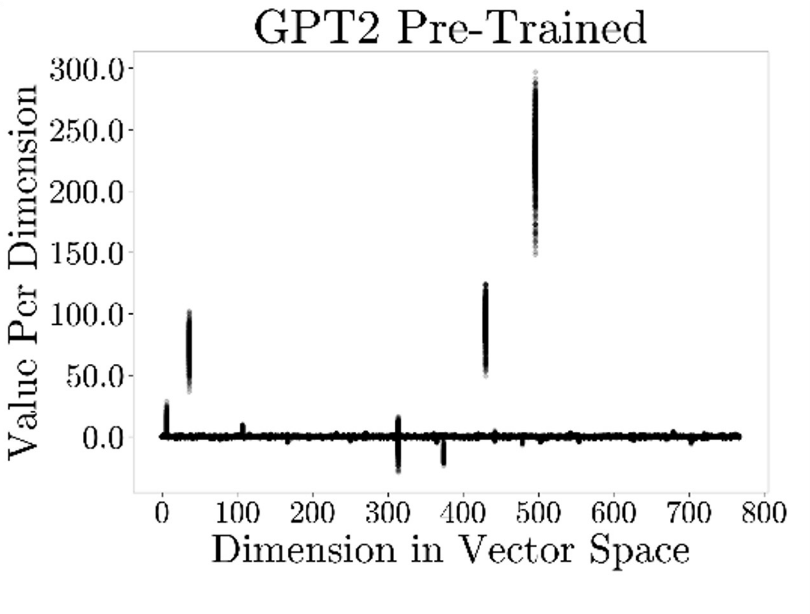


Figure 1: Average activation diagrams of sentence embeddings on the SST-2 validation dataset. The x-axis represents the index of the dimension, and the y-axis is the magnitude in that given dimension.

Methods for Improving Isotropy

- **<u>All-But-The-Top:</u>**
  - Post-processing algorithm that removes the *top* principal components from embeddings.
- **<u>Cosine Similarity Regularization:</u>**
  - CosReg adds an cosine similarity penalty to the standard CrossEntropy loss.
- **<u>Can we regularize using IsoScore?:</u>**
  - IsoScore is a more recent tool design for measuring isotropy.
  - *Not stable on mini-batch computations.*

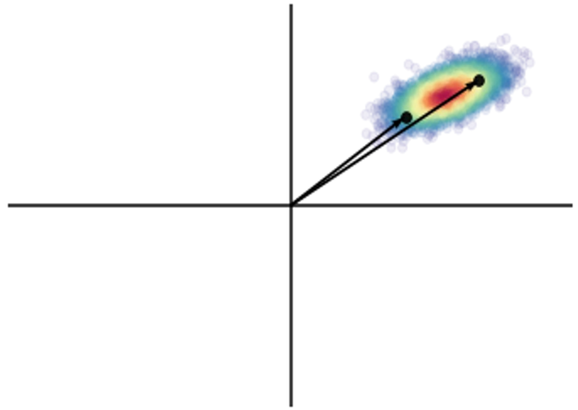


Figure 2: Average cosine similarity of **0.978**.

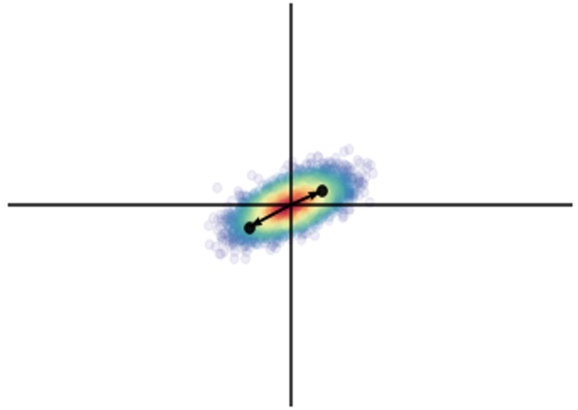
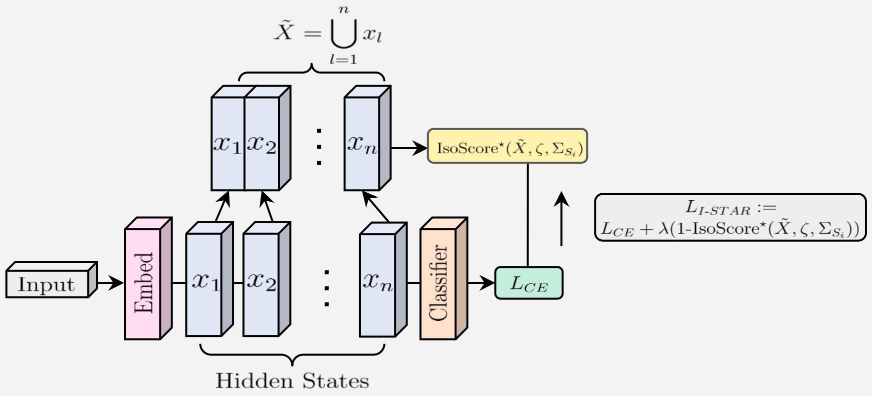
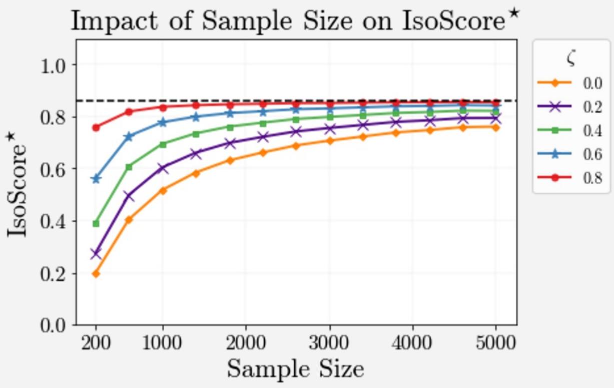


Figure 3: Average cosine similarity of **0.005**.

AIM 1

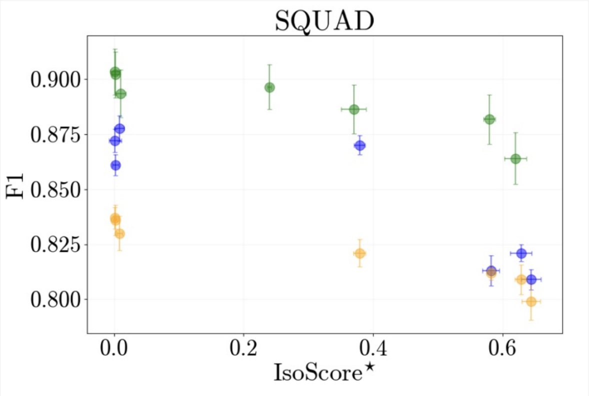
Develop a stable method of measuring isotropy.

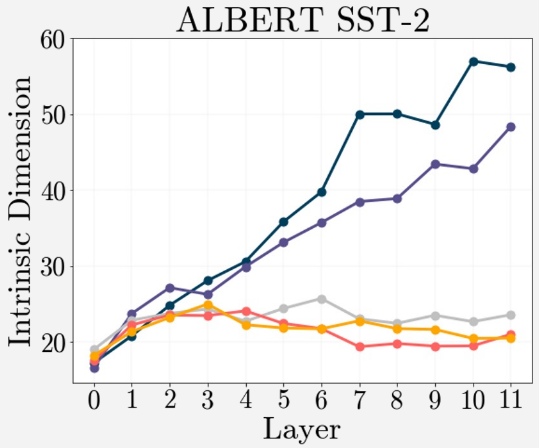




AIM 2

Understand how isotropy correlates to model performance.





IsoScore*: A Stable and Differentiable Metric of Isotropy

- **<u>IsoScore*:</u>**
  - Stable on mini-batch computations thanks to Regularized Discriminant Analysis (Friedman 1989).
  - IsoScore*(X) = 1 implies that all principal components are equal.
  - IsoScore*(X) = 0 implies that exactly principal component is non-zero.

- **<u>IsoScore vs. IsoScore*:</u>**
  - Approaches IsoScore when the number of samples is large.
  - Equivalent IsoScore when no RDA-shrinkage is performed.

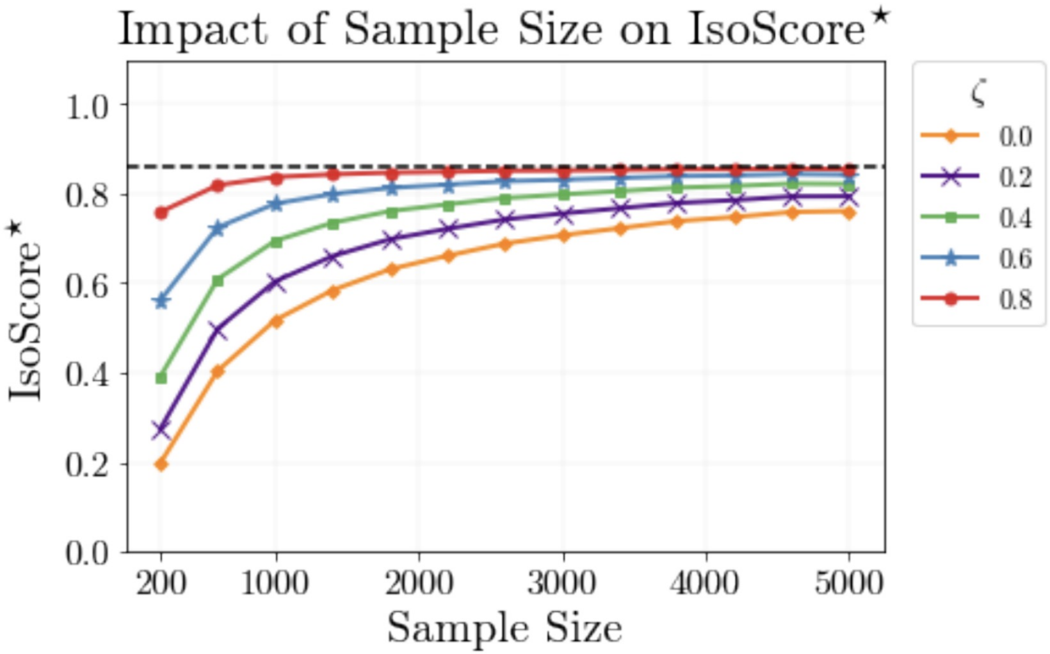


Figure 4: Impact of varying the RDA-shrinkage parameter on IsoScore*. Note that when ζ = 0, IsoScore = IsoScore*. Dashed line is the true isotropy score of the point cloud.

IsoScore*

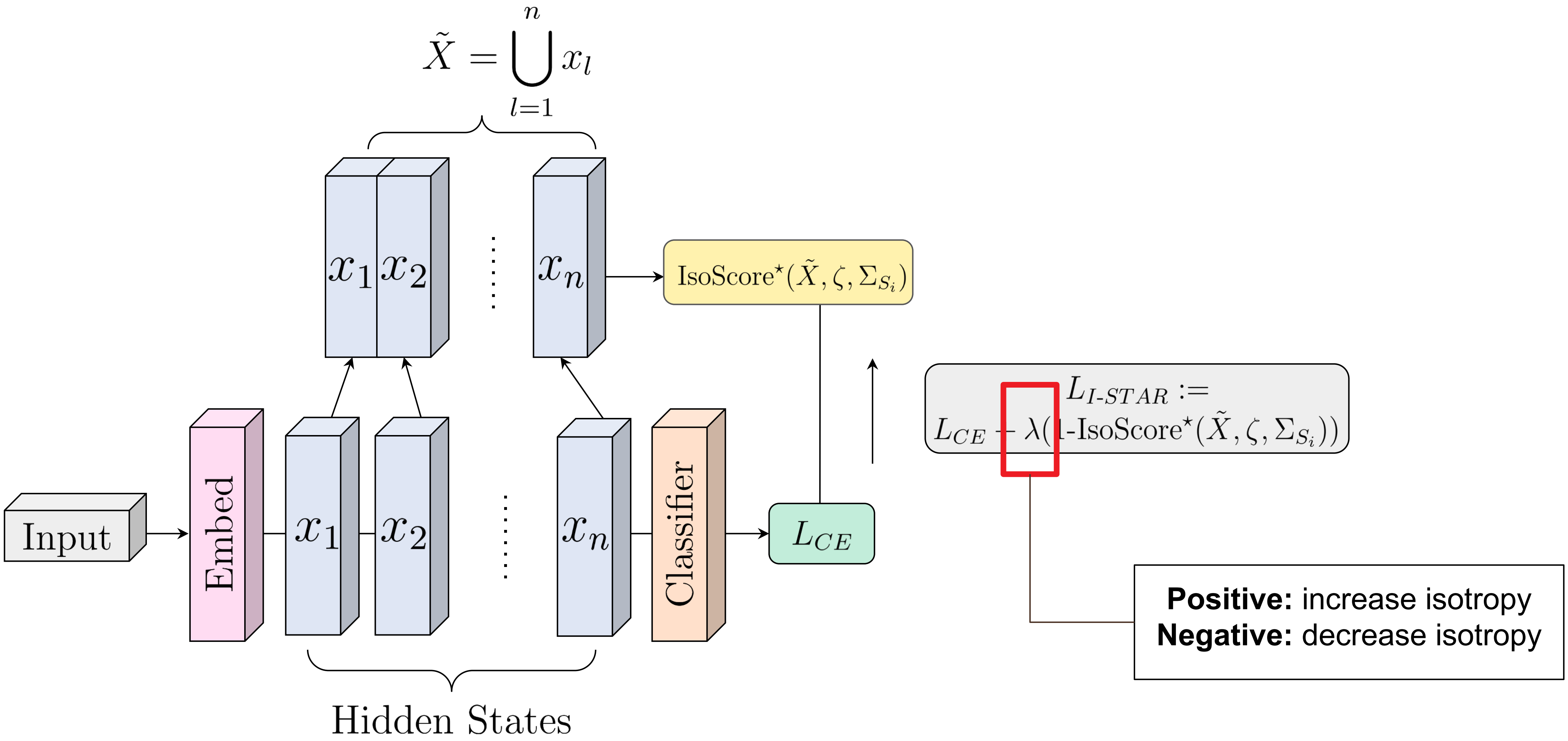
Algorithm 1 IsoScore* Forward Pass

Input: $X \subset \mathbb{R}^d$ point cloud, $\Sigma_S \in \mathbb{R}^{d\times d}$ shrinkage covariance matrix, $\zeta \in (0, 1)$.
Outputs: I-STAR penalty of X.
calculate covariance matrix: Σ_X of X
calculate shrinkage matrix: $\Sigma_\zeta := (1-\zeta)\cdot\Sigma_X + \zeta\cdot\Sigma_S$
calculate eigenvalues: $\Lambda := \{\lambda_1, .., \lambda_d\}$ of Σ_ζ
normalize eigenvalues: $\hat{\Lambda} := \sqrt{d}\cdot\Lambda/||\Lambda||_2$ such that $||\hat{\Lambda}|| = \sqrt{d}$
calculate the isotropy defect:

$$\delta(\hat{\Lambda}) := ||\hat{\Lambda} - \mathbf{1}||/\sqrt{2(d-\sqrt{d})}$$

where $\mathbf{1} = (1, ..., 1)^\top \in \mathbb{R}^d$
calculate: $\phi(\hat{\Lambda}) := (d - \delta(\hat{\Lambda})^2(d-\sqrt{d}))^2/d^2$
calculate: $\iota(\hat{\Lambda}) := (d\cdot\phi(\hat{\Lambda}) - 1)/(d-1)$.

I-STAR Loss


$\tilde{X} = \bigcup_{l=1}^{n} x_l$
x_1 x_2 x_n
$\text{IsoScore}^{\star}(\tilde{X}, \zeta, \Sigma_{S_i})$
$L_{I\text{-}STAR} :=$
$L_{CE} + \lambda(1\text{-IsoScore}^{\star}(\tilde{X}, \zeta, \Sigma_{S_i}))$
Input
Embed
x_1 x_2 x_n
Classifier
L_{CE}
Hidden States
Positive: increase isotropy
Negative: decrease isotropy

Isotropy Negatively Correlates with Performance

BROWN

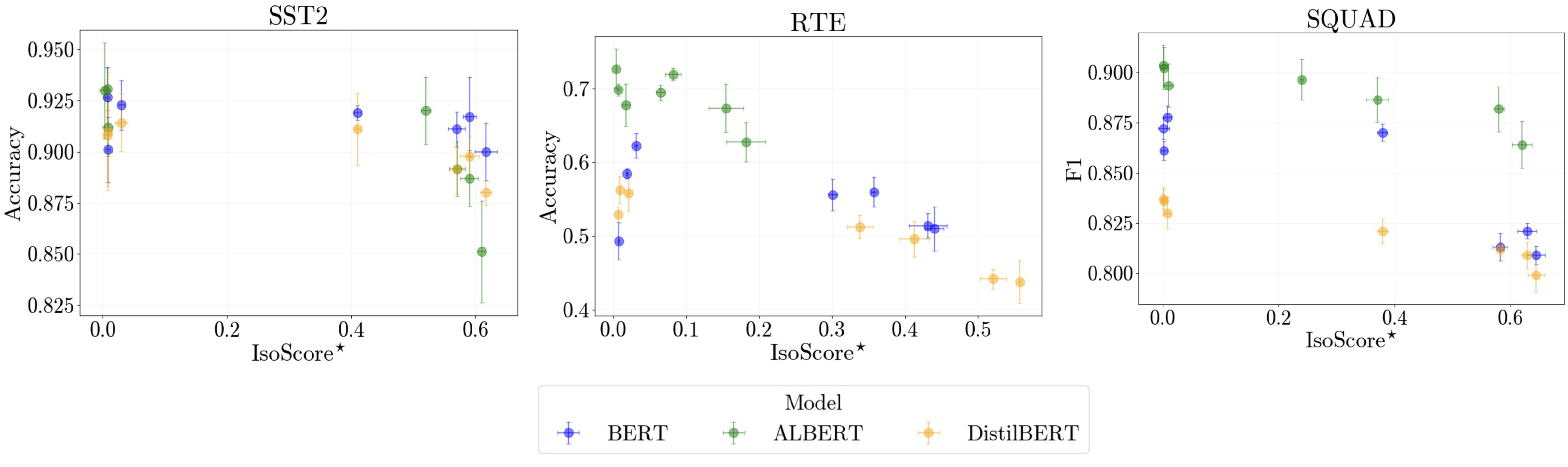


Figure 5: Relationship between IsoScore* (x-axis) and model performance (y-axis). We fine-tune each model with I-STAR using the tuning parameters, λ, in {-5, -3, -1, 0.50, 1, 3, 5}. We train each model over five random seeds and report the standard deviation of both performance and IsoScore*

Further Decreasing Isotropy with I-STAR Improves Performance

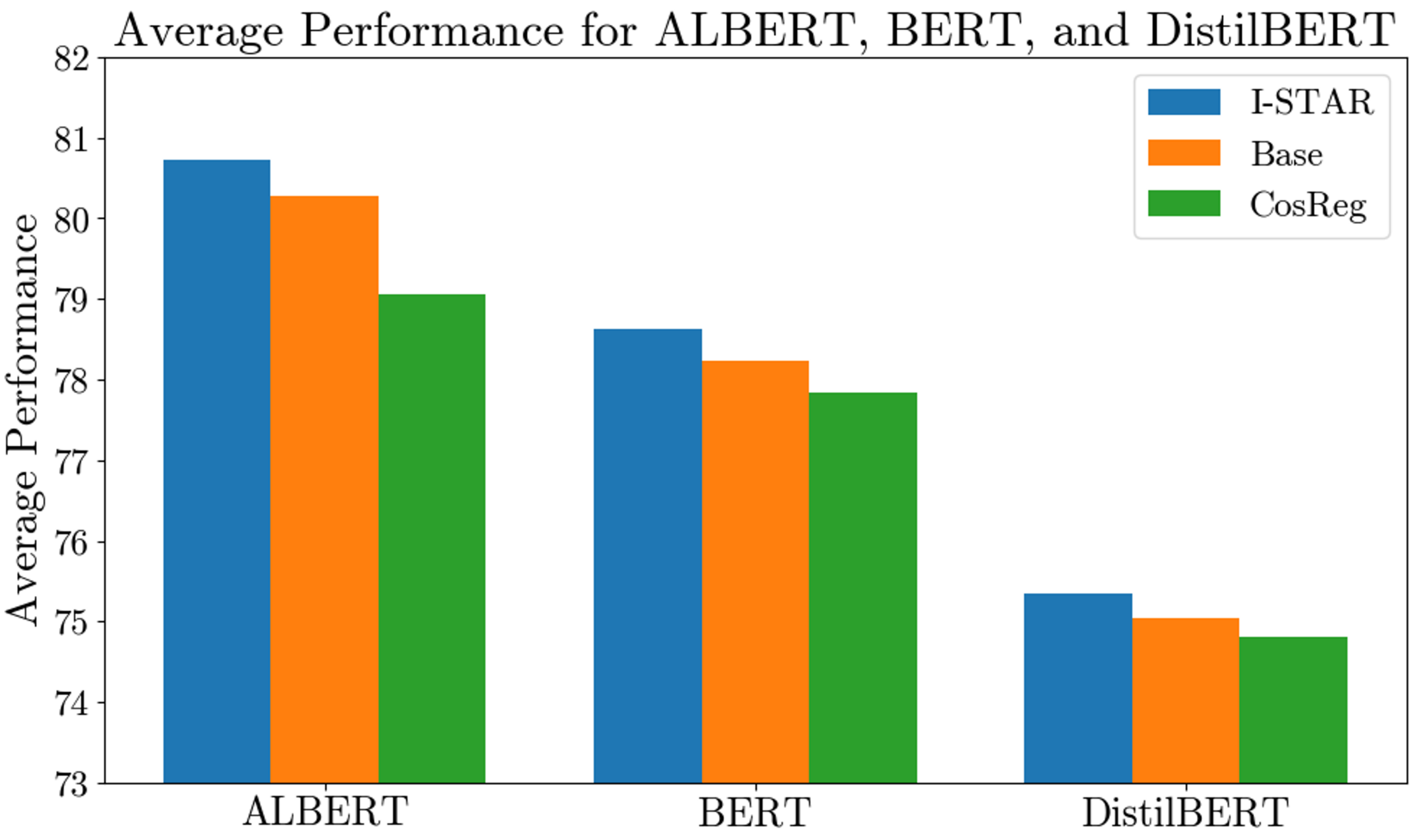


Figure 6: Average performance of ALBERT, BERT and DistilBERT fine-tuned using I-STAR (negative tuning params only), no regularization and Cosine Similarity regularization. Average is computed across 9 common NLP Benchmarks. GLUE (7 tasks), SST-5 and SQUAD.

Why do models prefer anisotropy?

- **<u>Benefits of Anisotropic Noise:</u>**
  - Helps models escape local minima in the loss landscape (Zhu et al. 2018).

- **<u>I-STAR Alters Intrinsic Dimension:</u>**
  - Low intrinsic dimension correlates with improved classification performance (Ansuini et al. 2019).

- **<u>Minimizing IsoScore Maximizes Silhouette Scores:</u>**
  - Isotropy objectives are incompatible with clustering objectives (Mickus et al. 2024).

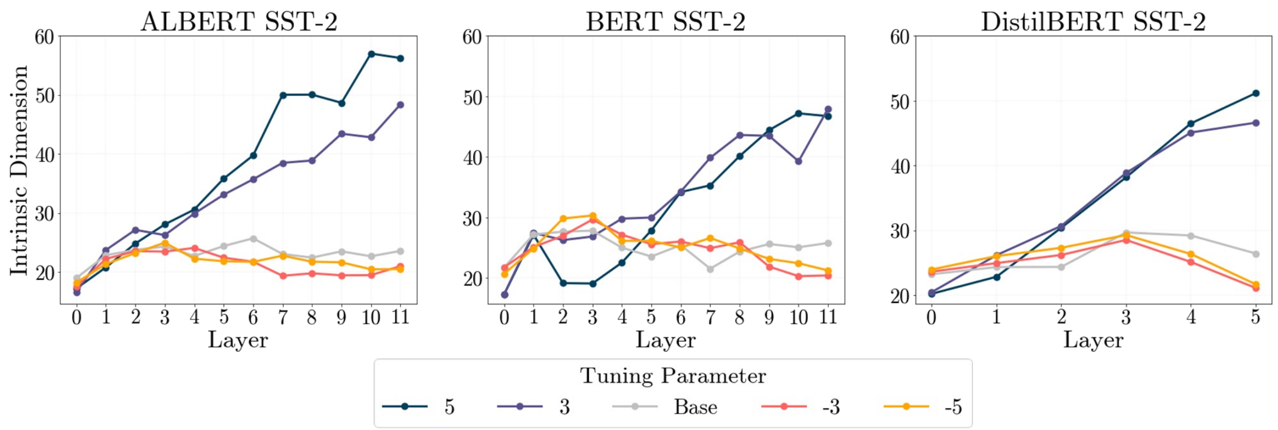
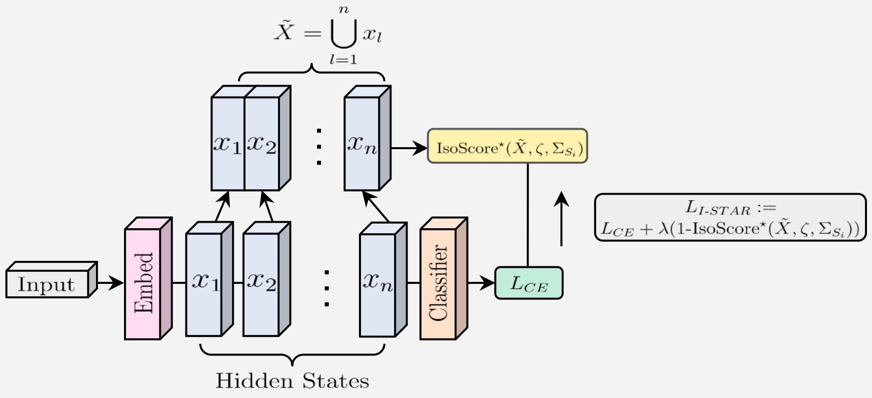
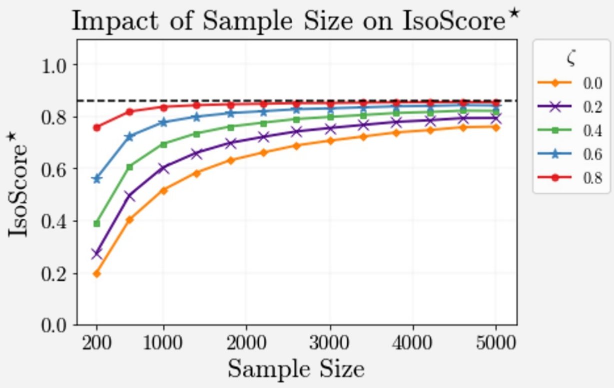


Figure 7: TwoNN Intrinsic Dimensionality estimate of ALBERT, BERT, and DistilBERT sentence embeddings obtained from the SST-2 validation data for models fine-tuned on the SST-2 using I-STAR with tuning-parameters. Base' represents the case where no regularization is used.

AIM 1

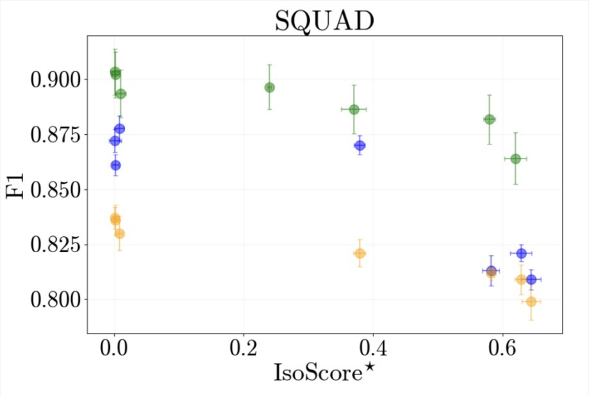
Develop a stable method of measuring isotropy.

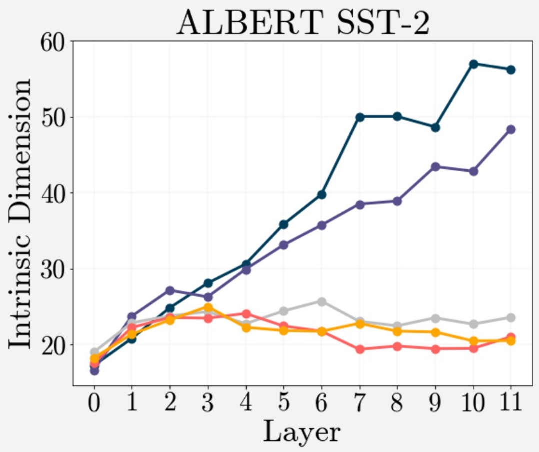




AIM 2

Understand how isotropy correlates to model performance.





Thank you!

- **GitHub:** https://github.com/bcbi-edu/p_eickhoff_isoscore.git
- **Pip installation:**

  IsoScore 2.0.1

  ```
  pip install IsoScore
  ```

- **Paper:** https://arxiv.org/pdf/2305.19358.pdf
- **Email:** william_rudman@brown.edu