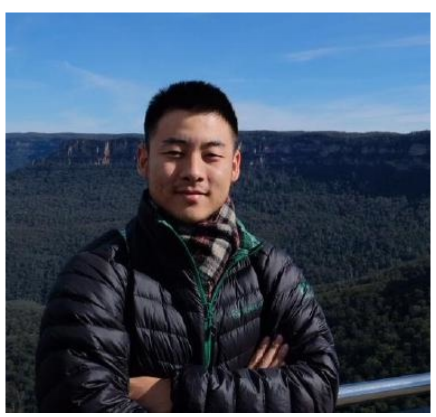
Predicting Out-of-distribution Error with the Projection Norm

Speaker: Zitong Yang*


*Yaodong Yu**


Alex Wei


Yi Ma


Jacob Steinhardt

*equal contribution

Paper link: https://arxiv.org/pdf/2103.04554.pdf

Problem: predicting OOD error

Problem: predicting OOD error

Given

- A prediction model $\hat{\boldsymbol{\theta}}$ fitted on a training set

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n) \; ;$$

- Test covariates $\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_m$,

Problem: predicting OOD error

Given

- A prediction model $\hat{\boldsymbol{\theta}}$ fitted on a training set

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n) \;;$$

- Test covariates $\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_m$,

predict the prediction error on the test set

$$(\widetilde{\boldsymbol{x}}_1, \widetilde{y}_1), (\widetilde{\boldsymbol{x}}_2, \widetilde{y}_2), \ldots, (\widetilde{\boldsymbol{x}}_m, \widetilde{y}_m)$$

without having access to labels $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_m$.

Problem: predicting OOD error

Given

- A prediction model $\hat{\boldsymbol{\theta}}$ fitted on a training set

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n) \; ;$$

- Test covariates $\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_m$,

predict the prediction error on the test set

$$(\widetilde{\boldsymbol{x}}_1, \widetilde{y}_1), (\widetilde{\boldsymbol{x}}_2, \widetilde{y}_2), \ldots, (\widetilde{\boldsymbol{x}}_m, \widetilde{y}_m)$$

without having access to labels $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_m$.

Golden machine learning wisdom:

- Holdout validation set
- Cross validation
- …

Problem: predicting OOD error

Given

- A prediction model $\hat{\theta}$ fitted on a training set

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n) \;;$$

- Test covariates $\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_m$,

predict the prediction error on the test set

$$(\widetilde{\boldsymbol{x}}_1, \widetilde{y}_1), (\widetilde{\boldsymbol{x}}_2, \widetilde{y}_2), \ldots, (\widetilde{\boldsymbol{x}}_m, \widetilde{y}_m)$$

without having access to labels $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_m$.

Golde chine l ing wisdom:

- Holdou ion set
- Cros da
- …

Test time distribution shift

Solution: Projection Norm

We propose a quantity named Projection Norm that help predict test error.

Solution: Projection Norm

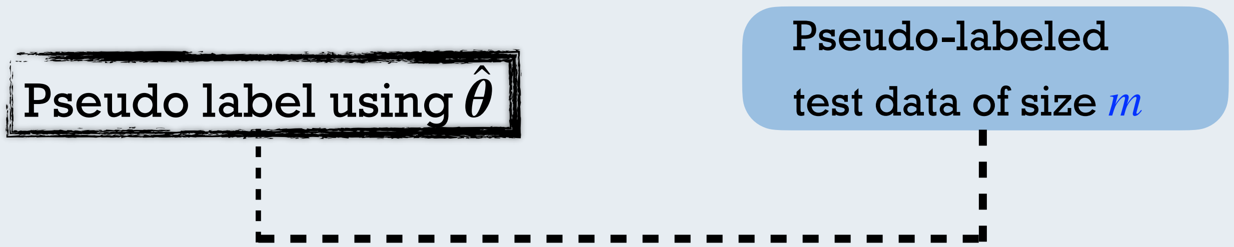
We propose a quantity named Projection Norm that help predict test error.

Projection Norm for neural network.

Solution: Projection Norm

We propose a quantity named Projection Norm that help predict test error.

Projection Norm for neural network.

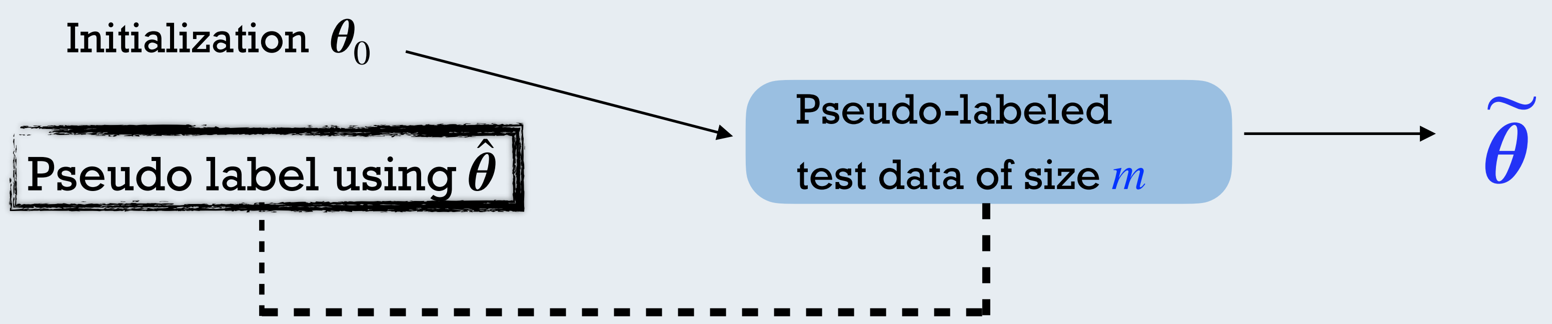


- Step 1: Use $\hat{\theta}$ (the model whose test accuracy we care about) to pseudo label the test covariates of size m.

Solution: Projection Norm

We propose a quantity named Projection Norm that help predict test error.

Projection Norm for neural network.

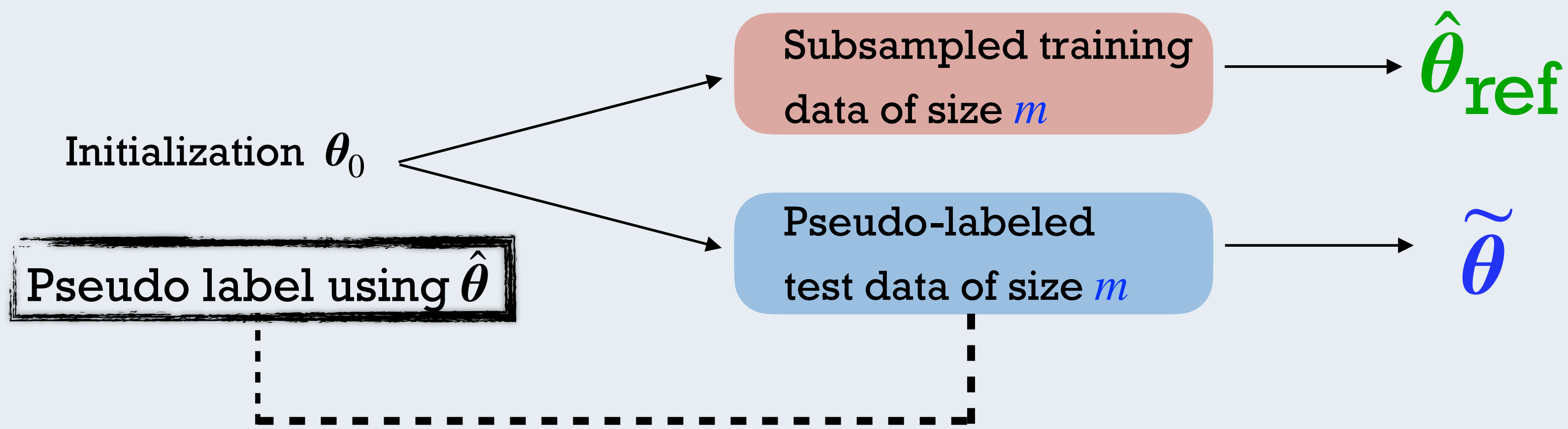


- Step 1: Use $\hat{\theta}$ (the model whose test accuracy we care about) to pseudo label the test covariates of size m.
- Step 2: Starting from initialization θ_0 (e.g. pertained ResNet), train a new model $\widetilde{\theta}$ on the pseudo labeled test set from Step 1.

Solution: Projection Norm

We propose a quantity named Projection Norm that help predict test error.

Projection Norm for neural network.

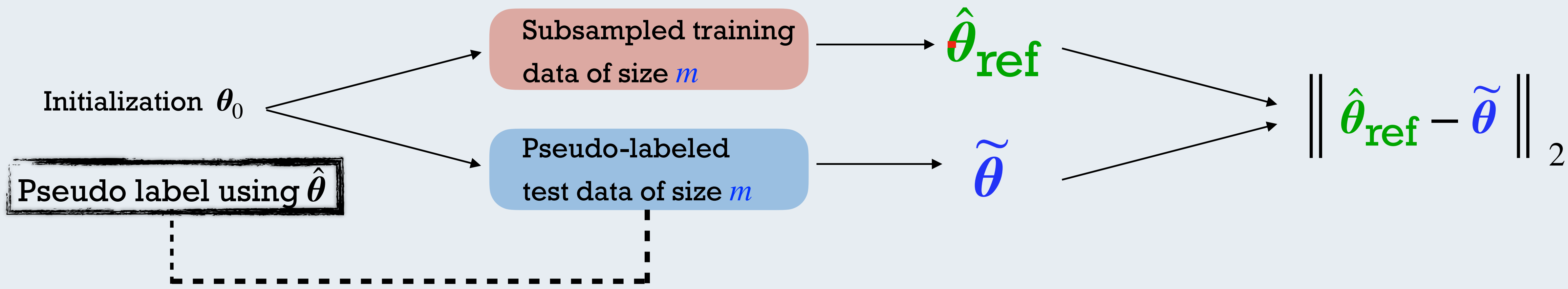


- Step 1: Use $\hat{\theta}$ (the model whose test accuracy we care about) to pseudo label the test covariates of size m.
- Step 2: Starting from initialization θ_0 (e.g. pertained ResNet), train a new model $\tilde{\theta}$ on the pseudo labeled test set from Step 1.
- Step 3: Subsample m samples from the training set (original size = n), and train a reference model $\hat{\theta}_{\text{ref}}$. Compute the norm of difference.

Solution: Projection Norm

We propose a quantity named Projection Norm that help predict test error.

Projection Norm for neural network.

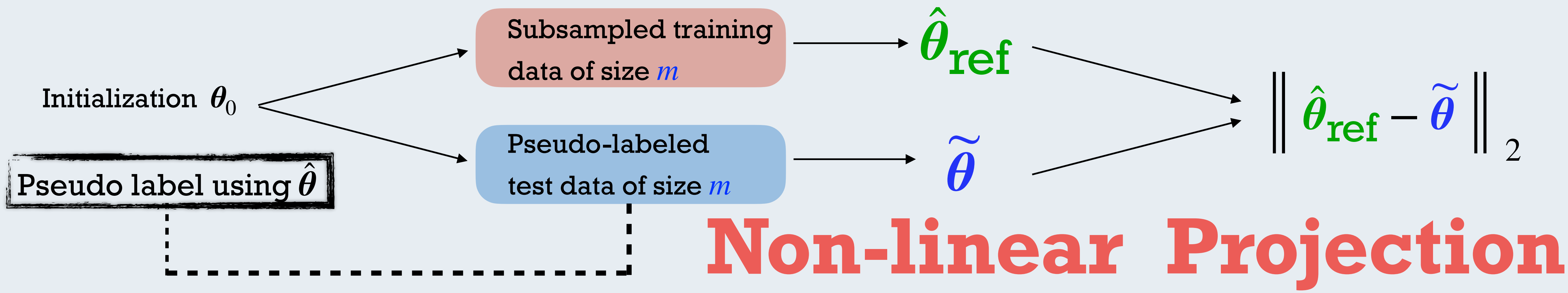


- Step 1: Use $\hat{\theta}$ (the model whose test accuracy we care about) to pseudo label the test covariates of size m.
- Step 2: Starting from initialization θ_0 (e.g. pertained ResNet), train a new model $\widetilde{\theta}$ on the pseudo labeled test set from Step 1.
- Step 3: Subsample m samples from the training set (original size = n), and train a reference model $\hat{\theta}_{\text{ref}}$. Compute the norm of difference.

Solution: Projection Norm

We propose a quantity named Projection Norm that help predict test error.

Projection Norm for neural network.



- Step 1: Use $\hat{\theta}$ (the model whose test accuracy we care about) to pseudo label the test covariates of size m.
- Step 2: Starting from initialization θ_0 (e.g. pertained ResNet), train a new model $\widetilde{\theta}$ on the pseudo labeled test set from Step 1.
- Step 3: Subsample m samples from the training set (original size = n), and train a reference model $\hat{\theta}_{\text{ref}}$. Compute the norm of difference.

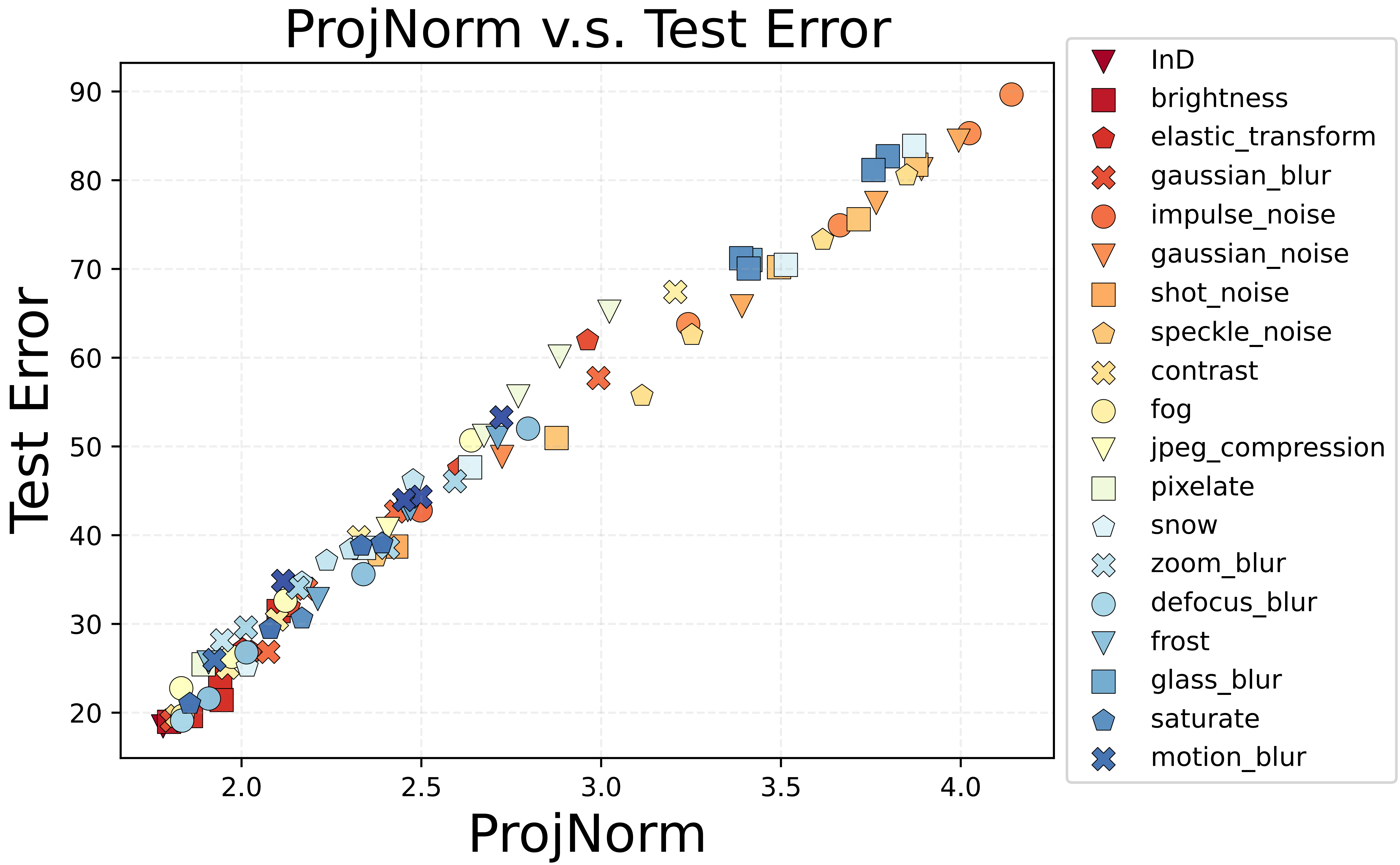
Experiments

Experiments

Mainline experiment: CIFAR100C ($16 \times 5 = 80$ corruptions) with ResNet50 pertained on ImageNet.

Experiments

Mainline experiment: CIFAR100C ($16 \times 5 = 80$ corruptions) with ResNet50 pertained on ImageNet.



Experiments: qualitative comparison

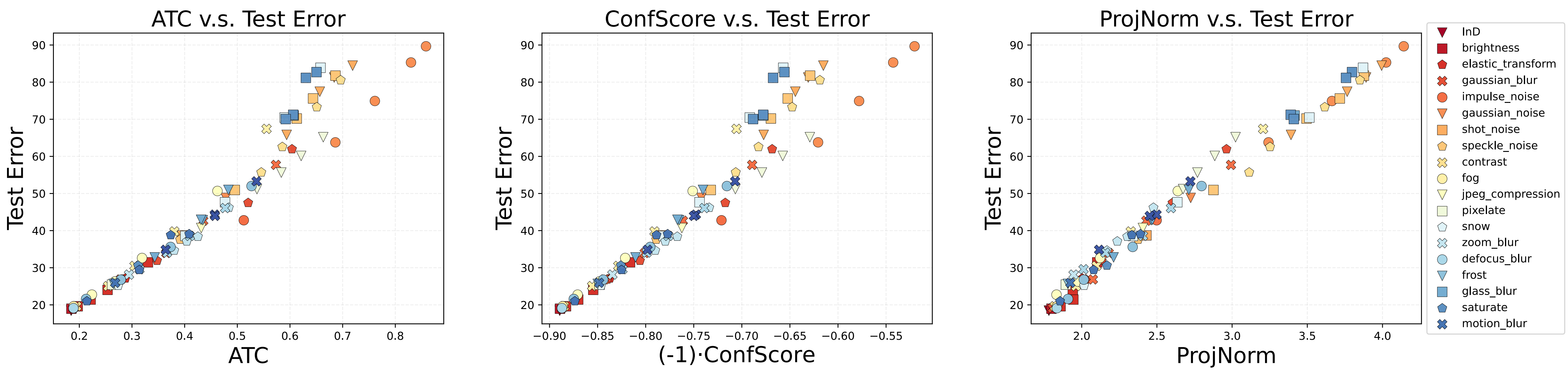
We consider several baselines:

- Confidence score (Hendrycks & Gimpel, 2016), ATC (Grag el al., 2022)
- Agreement score (Madani et al., 2004)
- Rotation prediction (Deng et al., 2021)

Experiments: qualitative comparison

We consider several baselines:

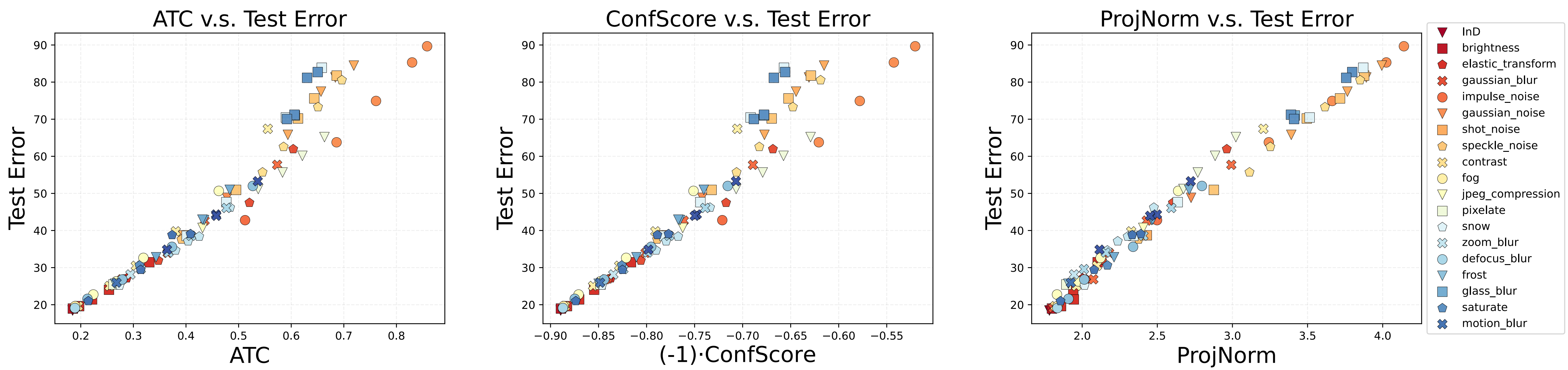
- Confidence score (Hendrycks & Gimpel, 2016), ATC (Grag el al., 2022)
- Agreement score (Madani et al., 2004)
- Rotation prediction (Deng et al., 2021)



Experiments: qualitative comparison

We consider several baselines:

- Confidence score (Hendrycks & Gimpel, 2016), ATC (Grag el al., 2022)
- Agreement score (Madani et al., 2004)
- Rotation prediction (Deng et al., 2021)



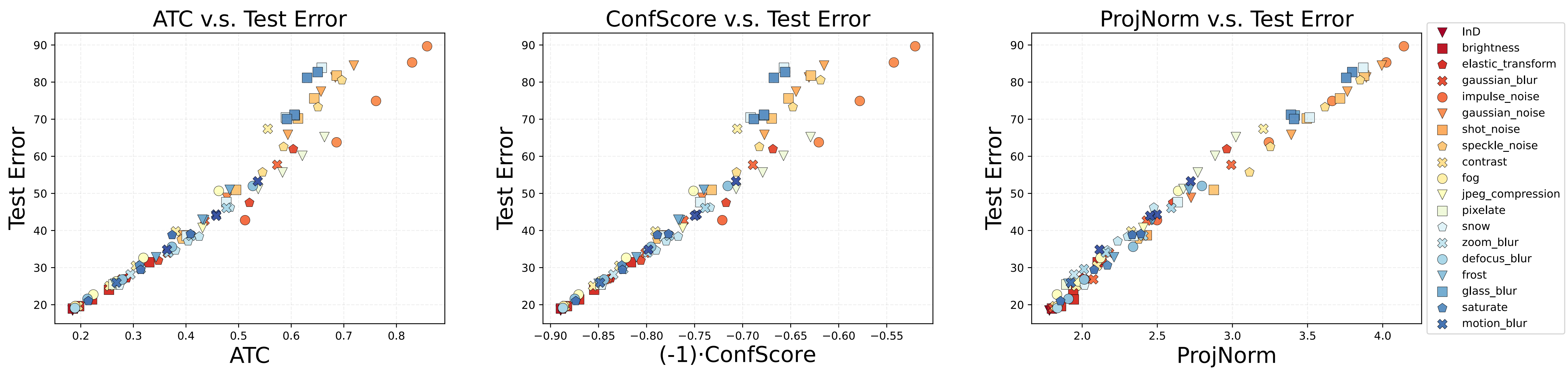
Some observations:

- All methods tends to behave well when test error is small.
- Projection Norm outperforms the other methods when test error is large.

Experiments: qualitative comparison

We consider several baselines:

- Confidence score (Hendrycks & Gimpel, 2016), ATC (Grag el al., 2022)
- Agreement score (Madani et al., 2004)
- Rotation prediction (Deng et al., 2021)



Some observations:

- All methods tends to behave well when test error is small.
- Projection Norm outperforms the other methods when test error is large.

Conclusion: Superiority comes from its ability to handle "hard" distribution shifts. We will latter present a synthetic example that further illustrates this empirical conclusion.

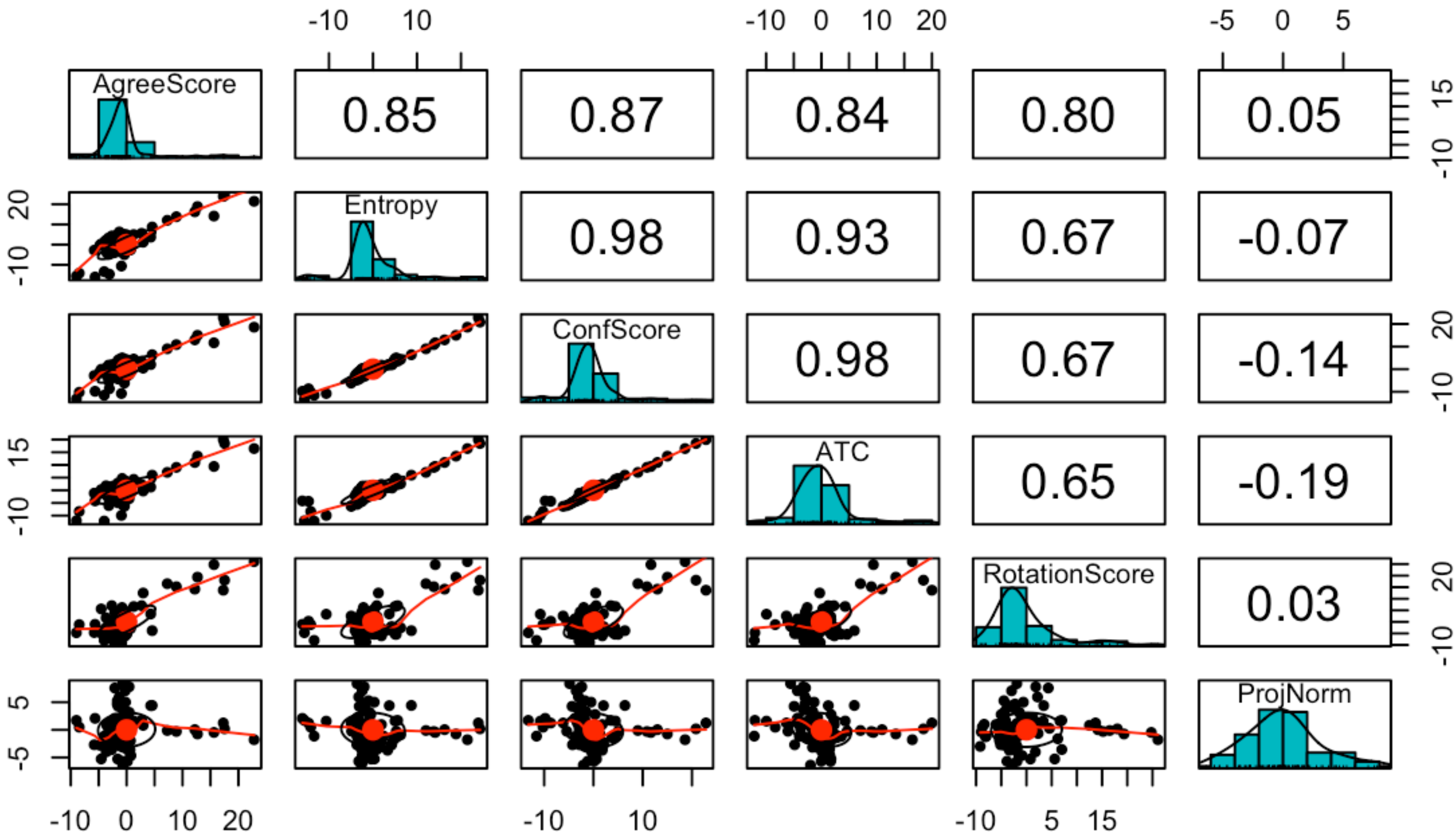
Experiments: quantitative comparison

Dataset	Network	Rotation		ConfScore		Entropy		AgreeScore		ATC		ProjNorm	
		R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
CIFAR10	ResNet18	0.839	0.953	0.847	0.981	0.872	0.983	0.556	0.871	0.860	0.983	**0.962**	**0.992**
	ResNet50	0.784	0.950	0.935	0.993	0.946	**0.994**	0.739	0.961	0.949	**0.994**	**0.951**	0.991
	VGG11	0.826	0.876	0.929	0.988	0.927	0.989	0.907	0.989	**0.931**	0.989	0.891	**0.991**
	Average	0.816	0.926	0.904	0.987	0.915	0.989	0.734	0.940	0.913	0.989	**0.935**	**0.991**
CIFAR100	ResNet18	0.903	0.955	0.917	0.958	0.879	0.938	0.939	0.969	0.934	0.966	**0.978**	**0.989**
	ResNet50	0.916	0.963	0.932	0.986	0.905	0.980	0.927	0.985	0.947	0.989	**0.984**	**0.993**
	VGG11	0.780	0.945	0.899	0.981	0.880	0.979	0.919	0.988	0.935	0.986	**0.953**	**0.993**
	Average	0.866	0.954	0.916	0.975	0.888	0.966	0.928	0.981	0.939	0.980	**0.972**	**0.992**
MNLI	BERT	-	-	0.516	0.671	0.533	**0.734**	0.318	0.524	0.524	0.699	**0.585**	0.664
	RoBERTa	-	-	0.493	0.727	0.498	0.734	0.499	0.762	0.519	0.734	**0.621**	**0.790**
	Average	-	-	0.505	0.699	0.516	**0.734**	0.409	0.643	0.522	0.717	**0.603**	0.727

Two metrics considered:

- R^2: perform a simple linear regression using the 80 samples and compute the R^2 statistics.
- ρ: Rank correlation between the vector of OOD test errors and the vector of Projection Norm.

Experiments: a statistical analysis



Synthetic linear regression: setup

Let’s consider a simplified overparameterized liner regression model to see the intuition behind the Projection Norm…

Synthetic linear regression: setup

Let's consider a simplified overparameterized liner regression model to see the intuition behind the Projection Norm…

We denote the training set by the matrix-vector pair $(\boldsymbol{X}, \boldsymbol{y})$ where $\boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{y} \in \mathbb{R}^{n}$. Similarly for the test set with $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m \times d}, \widetilde{\boldsymbol{y}} \in \mathbb{R}^{m}$.

Synthetic linear regression: setup

Let's consider a simplified overparameterized liner regression model to see the intuition behind the Projection Norm…

We denote the training set by the matrix-vector pair $(\boldsymbol{X}, \boldsymbol{y})$ where $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{y} \in \mathbb{R}^{n}$. Similarly for the test set with $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m\times d}, \widetilde{\boldsymbol{y}} \in \mathbb{R}^{m}$.

Assumptions: covaraite shift

We assume $d > n$, and that there exists a ground truth $\boldsymbol{\theta}_\star$ that defines the relation $y \mid \boldsymbol{x}$ in a noiseless fashion:

$$\boldsymbol{X}\boldsymbol{\theta}_\star = \boldsymbol{y}, \quad \widetilde{\boldsymbol{X}}\boldsymbol{\theta}_\star = \widetilde{\boldsymbol{y}}.$$

Synthetic linear regression: setup

Let's consider a simplified overparameterized liner regression model to see the intuition behind the Projection Norm…

We denote the training set by the matrix-vector pair $(\boldsymbol{X}, \boldsymbol{y})$ where $\boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{y} \in \mathbb{R}^{n}$. Similarly for the test set with $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m \times d}, \widetilde{\boldsymbol{y}} \in \mathbb{R}^{m}$.

Assumptions: covaraite shift

We assume $d > n$, and that there exists a ground truth $\boldsymbol{\theta}_\star$ that defines the relation $y \,|\, \boldsymbol{x}$ in a noiseless fashion:

$$\boldsymbol{X}\boldsymbol{\theta}_\star = \boldsymbol{y}, \quad \widetilde{\boldsymbol{X}}\boldsymbol{\theta}_\star = \widetilde{\boldsymbol{y}}.$$

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{\boldsymbol{y}}$, the test loss $(1/m)\|\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{y}}\|_2^2$ of the min-norm solution

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{X}\boldsymbol{\theta}=y} \|\boldsymbol{\theta}\|_2 = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\, \boldsymbol{\theta}_\star = \boldsymbol{P}\boldsymbol{\theta}_\star\,.$$

Synthetic linear regression: setup

Let's consider a simplified overparameterized liner regression model to see the intuition behind the Projection Norm…

We denote the training set by the matrix-vector pair $(\boldsymbol{X}, \boldsymbol{y})$ where $\boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{y} \in \mathbb{R}^{n}$. Similarly for the test set with $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m \times d}, \widetilde{\boldsymbol{y}} \in \mathbb{R}^{m}$.

Assumptions: covaraite shift

We assume $d > n$, and that there exists a ground truth $\boldsymbol{\theta}_\star$ that defines the relation $y \,|\, \boldsymbol{x}$ in a noiseless fashion:

$$\boldsymbol{X}\boldsymbol{\theta}_\star = \boldsymbol{y}, \quad \widetilde{\boldsymbol{X}}\boldsymbol{\theta}_\star = \widetilde{\boldsymbol{y}}.$$

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{\boldsymbol{y}}$, the test loss $(1/m)\|\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{y}}\|_2^2$ of the min-norm solution

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{X}\boldsymbol{\theta}=y} \|\boldsymbol{\theta}\|_2 = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \boxed{\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}}\,\boldsymbol{\theta}_\star = \boldsymbol{P}\boldsymbol{\theta}_\star\,.$$

Synthetic linear regression: intuition

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{y}$, the test loss

$(1/m)\|\widetilde{X}\hat{\theta} - \widetilde{y}\|_2^2$ of the min-norm solution

$$\hat{\boldsymbol{\theta}} = \min_{X\theta = y} \|\boldsymbol{\theta}\|_2 = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \boxed{\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}}\,\boldsymbol{\theta}_\star = \boldsymbol{P}\boldsymbol{\theta}_\star \,.$$

Synthetic linear regression: intuition

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{y}$, the test loss

$(1/m)\|\widetilde{X}\hat{\theta} - \widetilde{y}\|_2^2$ of the min-norm solution

$$\hat{\theta} = \min_{X\theta=y} \|\theta\|_2 = X^\top(XX^\top)^{-1}y = \boxed{X^\top(XX^\top)^{-1}X}\,\theta_\star = P\theta_\star\,.$$

Two observations:

- From the training set, we only learned the portion of $\theta_\star$ that is in the span of X.

Synthetic linear regression: intuition

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{\boldsymbol{y}}$, the test loss

$(1/m)\|\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{y}}\|_2^2$ of the min-norm solution

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{X\theta}=y} \|\boldsymbol{\theta}\|_2 = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \boxed{\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}}\,\boldsymbol{\theta}_\star = \boldsymbol{P}\boldsymbol{\theta}_\star\,.$$

Two observations:

- From the training set, we only learned the portion of $\boldsymbol{\theta}_\star$ that is in the span of X.
- If the training and test set are perfect aligned, i.e. $\text{row}(X) = \text{row}(\widetilde{X})$, the test loss would be just 0.

Synthetic linear regression: intuition

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{y}$, the test loss

$(1/m)\|\widetilde{X}\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{y}}\|_2^2$ of the min-norm solution

$$\hat{\boldsymbol{\theta}} = \min_{X\boldsymbol{\theta}=y} \|\boldsymbol{\theta}\|_2 = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \boxed{\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}}\,\boldsymbol{\theta}_\star = \boldsymbol{P}\boldsymbol{\theta}_\star\,.$$

Two observations:

- From the training set, we only learned the portion of $\boldsymbol{\theta}_\star$ that is in the span of X.
- If the training and test set are perfect aligned, i.e. $\text{row}(X) = \text{row}(\widetilde{X})$, the test loss would be just 0.

Projection norm for linear regression

Therefore the non-zero test error stems from the portion of $\boldsymbol{\theta}_\star$ that is in $\text{row}(\widetilde{X})$ but not in $\text{row}(X)$. This quantity is intuitively measured by

$$\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$$

Synthetic linear regression: intuition

Problem formulation in the linear setting

The problem is to estimate, without access to $\widetilde{y}$, the test loss

$(1/m)\|\widetilde{X}\hat{\theta} - \widetilde{y}\|_2^2$ of the min-norm solution

$$\hat{\theta} = \min_{X\theta=y} \|\theta\|_2 = X^\top(XX^\top)^{-1}y = \boxed{X^\top(XX^\top)^{-1}X}\,\theta_\star = P\theta_\star\,.$$

Two observations:

- From the training set, we only learned the portion of $\theta_\star$ that is in the span of X.
- If the training and test set are perfect aligned, i.e. $\text{row}(X) = \text{row}(\widetilde{X})$, the test loss would be just 0.

Projection norm for linear regression

Therefore the non-zero test error stems from the portion of $\theta_\star$ that is in $\text{row}(\widetilde{X})$ but not in $\text{row}(X)$. This quantity is intuitively measured by

$$\|(I - \widetilde{P})P\theta_\star\|_2 \qquad \boxed{\widetilde{X}^\top(\widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}}$$

Synthetic linear regression: intuition

Question: How does $\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

Synthetic linear regression: intuition

Question: How does $\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

$$\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2 = \|\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}\|_2$$

Synthetic linear regression: intuition

Question: How does $\|(I - \widetilde{P})P\theta_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

$$\|(I - \widetilde{P})P\theta_\star\|_2 = \|\hat{\theta} - \widetilde{P}\hat{\theta}\|_2$$

A nonlinear projection

Synthetic linear regression: intuition

Question: How does $\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

$$\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2 = \|\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}\|_2$$

A nonlinear projection

The quantity $\widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}$ can be regarded as the minimum norm solution to

$$\min_{\boldsymbol{\theta}} \|\widetilde{\boldsymbol{X}}\boldsymbol{\theta} - \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}}\|_2.$$

Synthetic linear regression: intuition

Question: How does $\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

$$\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2 = \|\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}\|_2$$

A nonlinear projection

The quantity $\widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}$ can be regarded as the minimum norm solution to

$$\min_{\boldsymbol{\theta}} \|\widetilde{\boldsymbol{X}}\boldsymbol{\theta} - \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}}\|_2.$$

Writing the optimization problem differently with $f(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$:

$$\min_{\boldsymbol{\theta}} \sum_{j=1}^{m} \left[f(\widetilde{\boldsymbol{x}}_j, \boldsymbol{\theta}) - f(\widetilde{\boldsymbol{x}}_j, \hat{\boldsymbol{\theta}}) \right],$$

Synthetic linear regression: intuition

Question: How does $\|(\boldsymbol{I}-\widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

$$\|(\boldsymbol{I}-\widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2 = \|\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}\|_2$$

A nonlinear projection

The quantity $\widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}$ can be regarded as the minimum norm solution to

$$\min_{\boldsymbol{\theta}} \|\widetilde{\boldsymbol{X}}\boldsymbol{\theta} - \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}}\|_2.$$

Writing the optimization problem differently with $f(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$:

$$\min_{\boldsymbol{\theta}} \sum_{j=1}^{m} \left[f(\widetilde{\boldsymbol{x}}_j, \boldsymbol{\theta}) - f(\widetilde{\boldsymbol{x}}_j, \hat{\boldsymbol{\theta}}) \right],$$

where $f(\widetilde{\boldsymbol{x}}_j, \hat{\boldsymbol{\theta}})$ is exactly the pseudo-labels mentioned earlier.

Synthetic linear regression: intuition

Question: How does $\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2$ relates to the Projection Norm for neural network that is introduced earlier?

$$\|(\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{P}\boldsymbol{\theta}_\star\|_2 = \|\hat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}\|_2$$

A nonlinear projection

The quantity $\widetilde{\boldsymbol{P}}\hat{\boldsymbol{\theta}}$ can be regarded as the minimum norm solution to

$$\min_{\boldsymbol{\theta}} \|\widetilde{\boldsymbol{X}}\boldsymbol{\theta} - \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\theta}}\|_2.$$

Writing the optimization problem differently with $f(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$:

$$\min_{\boldsymbol{\theta}} \sum_{j=1}^{m} \left[f(\widetilde{\boldsymbol{x}}_j, \boldsymbol{\theta}) - f(\widetilde{\boldsymbol{x}}_j, \hat{\boldsymbol{\theta}}) \right],$$

where $f(\widetilde{\boldsymbol{x}}_j, \hat{\boldsymbol{\theta}})$ is exactly the pseudo-labels mentioned earlier.

In this case, we start from theoretical analysis on a toy model and end up with an algorithm works well on real architectures!

Synthetic linear regression: a special case

To see why Projection Norm handles “hard” distribution shifts, consider the example:

Synthetic linear regression: a special case

To see why Projection Norm handles "hard" distribution shifts, consider the example:

Training samples: $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)$; Test samples: $\widetilde{\boldsymbol{x}}_j \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \boldsymbol{I}_{d_2} \end{bmatrix}\right)$.

Synthetic linear regression: a special case

To see why Projection Norm handles "hard" distribution shifts, consider the example:

$$\text{Training samples: } \boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right); \quad \text{Test samples: } \widetilde{\boldsymbol{x}}_j \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \boldsymbol{I}_{d_2} \end{bmatrix}\right).$$

Even with enough samples, $\hat{\boldsymbol{\theta}}$ is just the projection of $\boldsymbol{\theta}_\star$ to its first d_1 coordinates. Therefore, methods like the confidence score that only depends on the neural network output $f(\widetilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) = \langle \widetilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}} \rangle$ can't capture any information about σ.

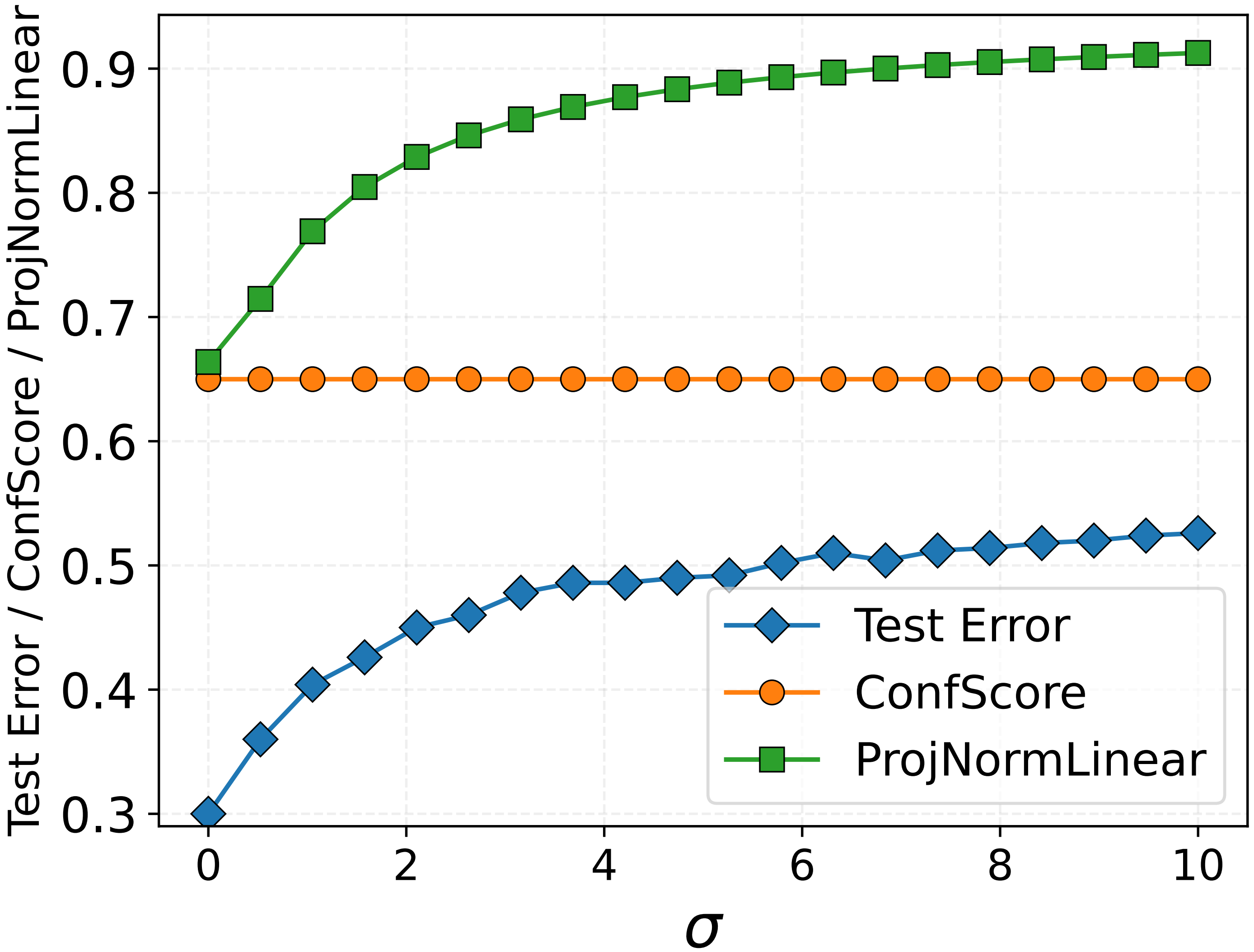
Synthetic linear regression: a special case

To see why Projection Norm handles "hard" distribution shifts, consider the example:

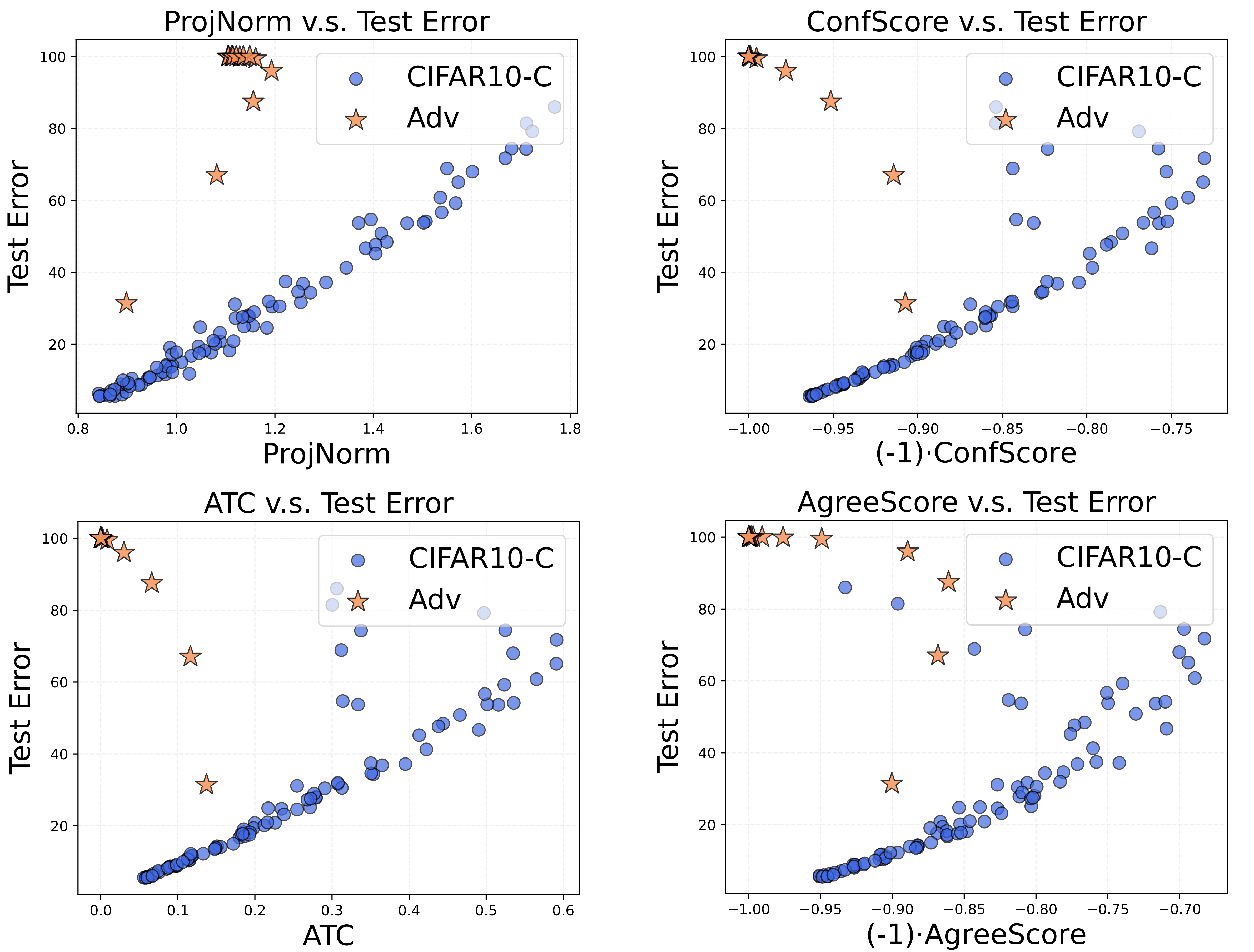
$$\text{Training samples: } \boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right); \quad \text{Test samples: } \widetilde{\boldsymbol{x}}_j \overset{i.i.d.}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \boldsymbol{I}_{d_2} \end{bmatrix}\right).$$

Even with enough samples, $\hat{\boldsymbol{\theta}}$ is just the projection of $\boldsymbol{\theta}_\star$ to its first d_1 coordinates.

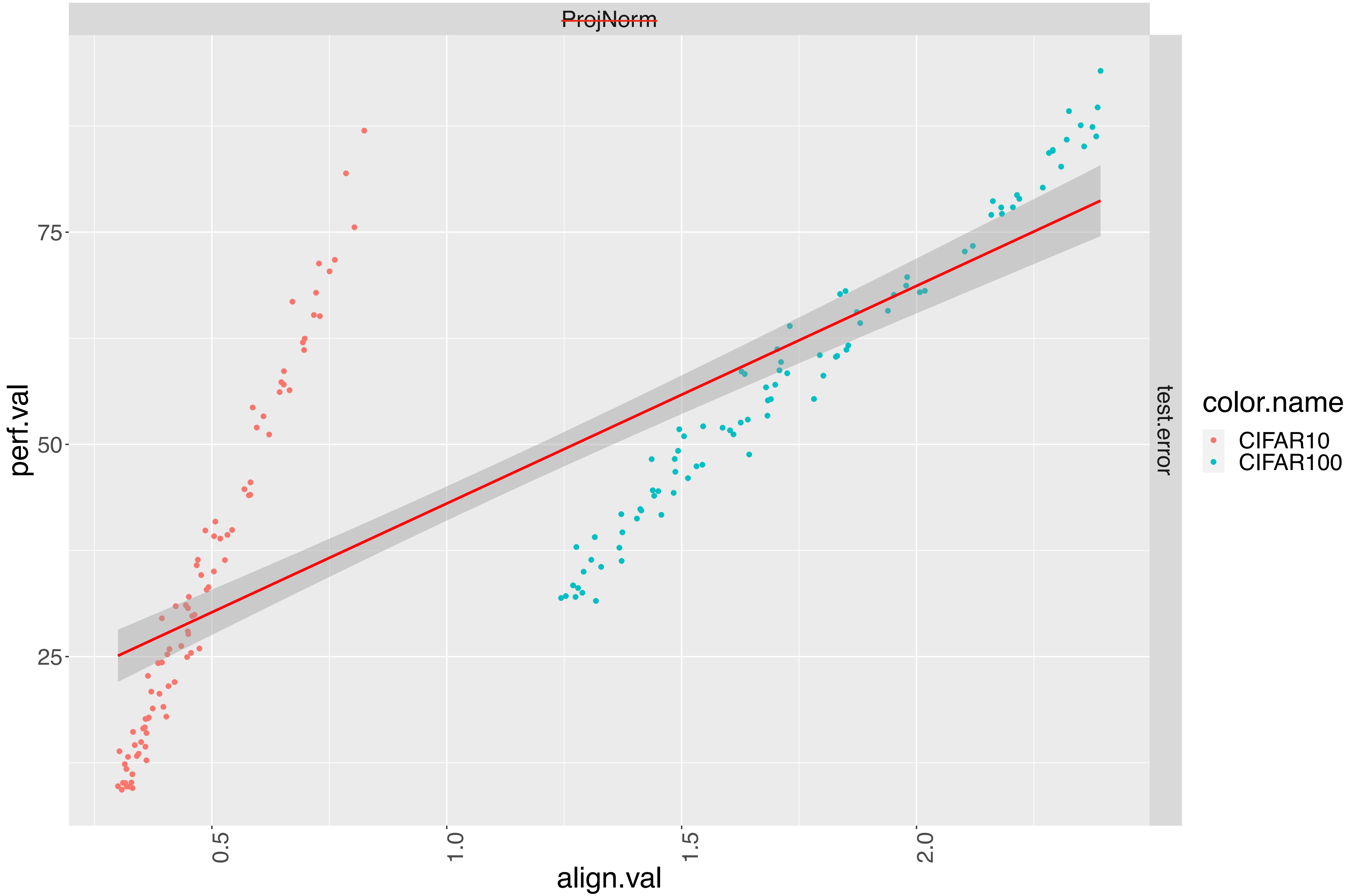
Therefore, methods like the confidence score that only depends on the neural network output $f(\widetilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) = \langle \widetilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}} \rangle$ can't capture any information about σ.



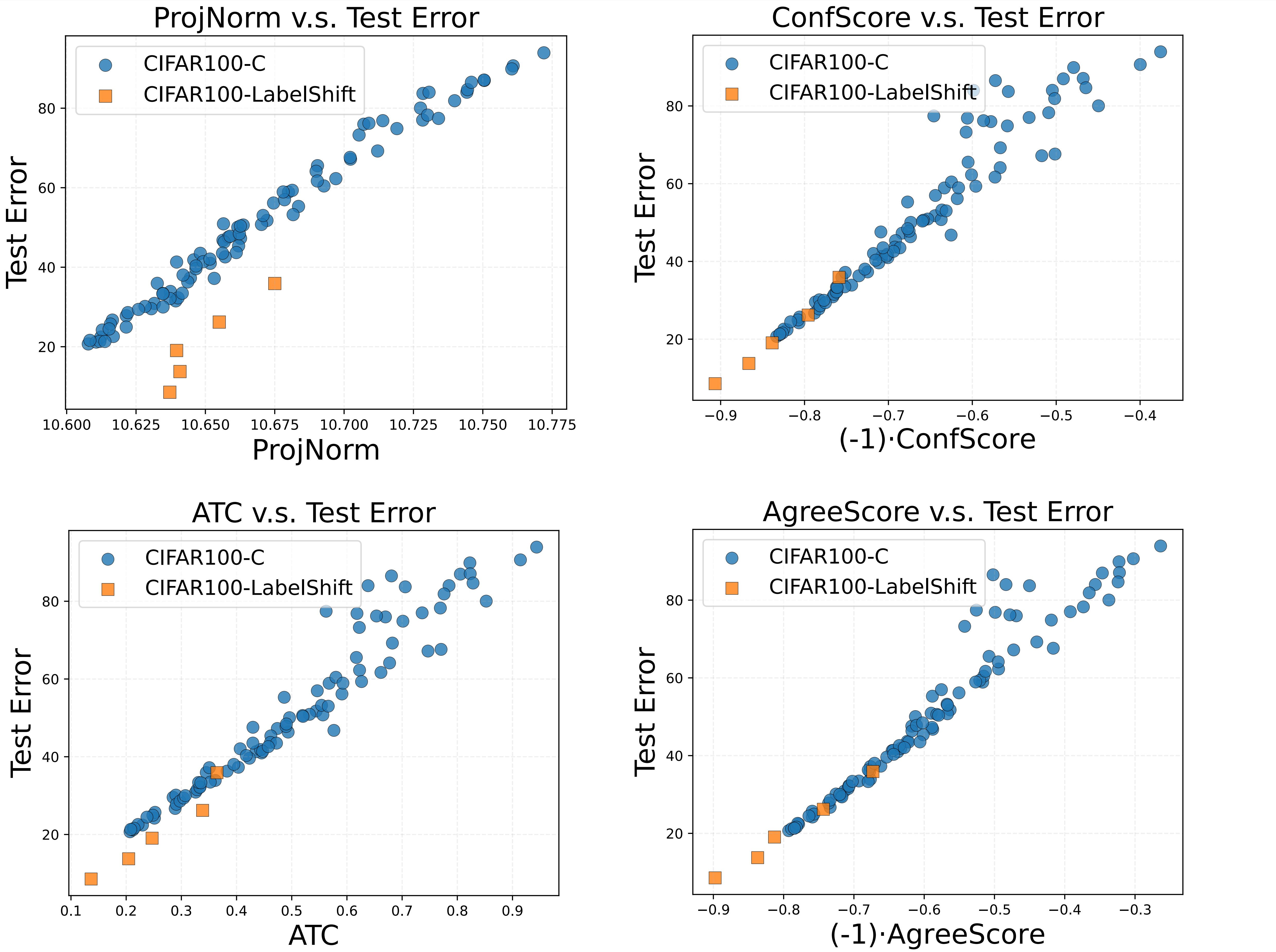
Stress test: adversarial example



Limitations: across dataset prediction



Limitations: easy distribution shift



Conclusion

We propose a quantity named Projection Norm that help predict test error that is **almost better than every existing procedures**!

Conclusion

We propose a quantity named Projection Norm that help predict test error that is **almost better than every existing procedures**!

It has three limitations that opens room for improvement:

- Projection Norm **requires the test set to be large enough** to allow meaningful fine tuning, whereas methods such as confidence score only require one test sample

Conclusion

We propose a quantity named Projection Norm that help predict test error that is **almost better than every existing procedures**!

It has three limitations that opens room for improvement:

- Projection Norm **requires the test set to be large enough** to allow meaningful fine tuning, whereas methods such as confidence score only require one test sample

- Projection Norm can't handle **easy distribution shift** very well.

Conclusion

We propose a quantity named Projection Norm that help predict test error that is **almost better than every existing procedures**!

It has three limitations that opens room for improvement:

- Projection Norm **requires the test set to be large enough** to allow meaningful fine tuning, whereas methods such as confidence score only require one test sample
- Projection Norm can't handle **easy distribution shift** very well.
- Projection Norm can't handle **prediction across dataset**.

Thanks!