# Data Continuity Matters: Improving Sequence Modeling with Lipschitz Regularizer

ICLR

Selected as Spotlight

Eric Qu [1,2,3] Xufang Luo [1] Dongsheng Li [1]

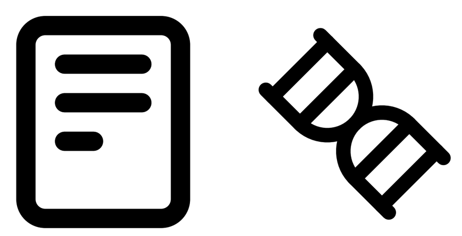
[1]Microsoft Research Asia [2]UC Berkeley [3]Duke Kunshan University

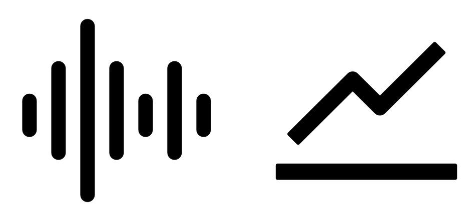
## Motivation

Sequence Models Works Well On Specific Tasks

Transformers | State Space Models

Text Gene | Audio Time-series

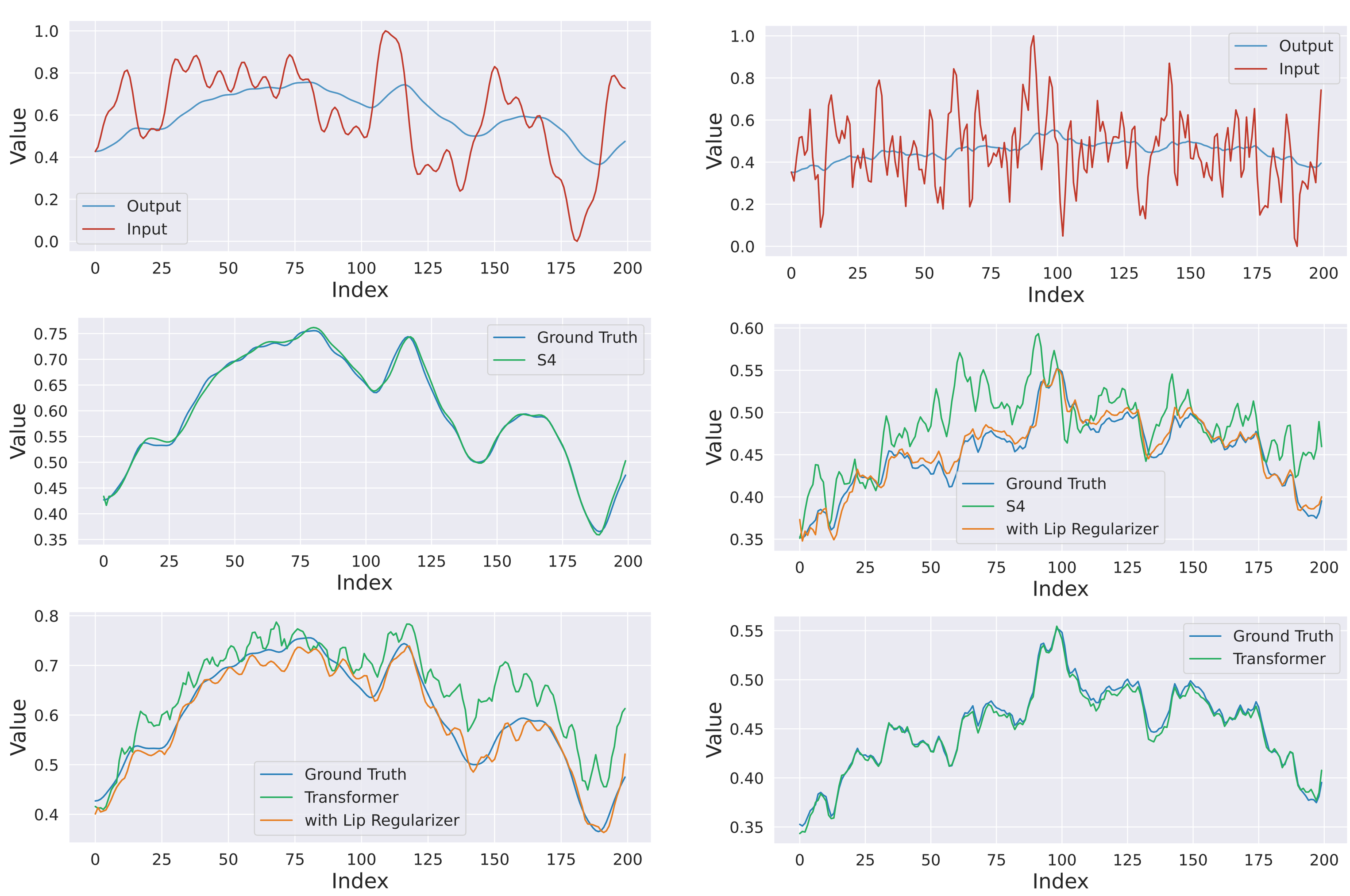




Sequence models have preferences in **Data Continuity**

Transformers ⇔ Discrete Data

State Space Models ⇔ Continuous Data




(a) High Continuity (b) Low Continuity

Sequence Models + Unpreferred Data Continuity

⇓

Deteriorated Performance

Solution

A Regularizer That Alters Input Data Continuity!

Apply The Regularizer to The Input Embedding

## Lipschitz Regularizer

Lipschitz Constant

$$L_f = \max_{i,j\in\{0,1,\ldots,n\}} \frac{|x_i - x_j|}{|i-j|} = \max_{k\in\{0,1,\ldots,n-1\}} |x_{k+1} - x_k|$$

Max → Mean ⇓ L1 → L2 norm

Lipschitz Regularizer

$$\mathcal{L}_{\text{Lip}} = \frac{1}{n}\sum_{i=0}^{n-1}(x_{i+1} - x_i)^2$$

## Experiments

State Space Models prefer Continuous Input

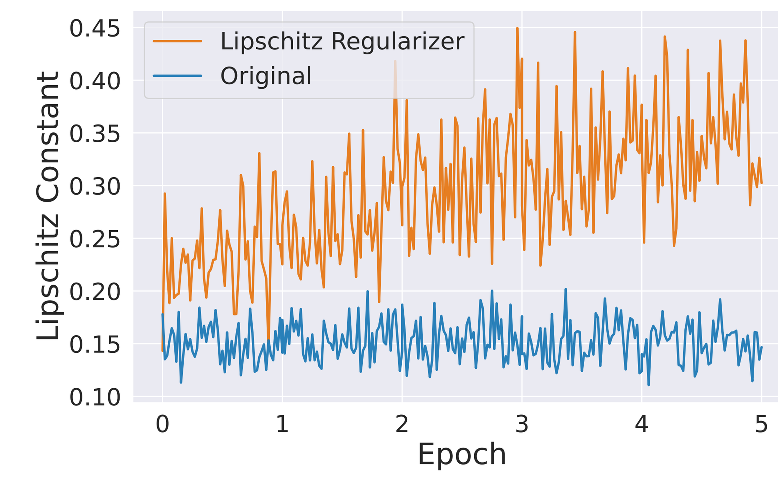
$$\mathcal{L}(y, \hat{y}, \hat{l}) = \mathcal{L}_{\text{S4}}(y, \hat{y}) + \lambda \mathcal{L}_{\text{Lip}}(\hat{l})$$

| | ListOps | Text | Retrieval | Image | Image-c | Path | Path-c | PathX | PathX-c |
|---|---|---|---|---|---|---|---|---|---|
| S4 | 59.53 | 86.51 | 91.07 | 88.54 | 84.27 | **94.02** | 89.11 | **96.03** | 92.41 |
| S4 + Emb | 58.94 | 87.12 | 90.28 | 87.25 | 85.13 | 92.37 | 90.32 | 93.87 | 92.81 |
| **S4 + Emb + Lip** | **61.37** | **89.74** | **93.83** | **89.19** | **88.43** | 93.52 | **91.39** | 95.72 | **94.36** |

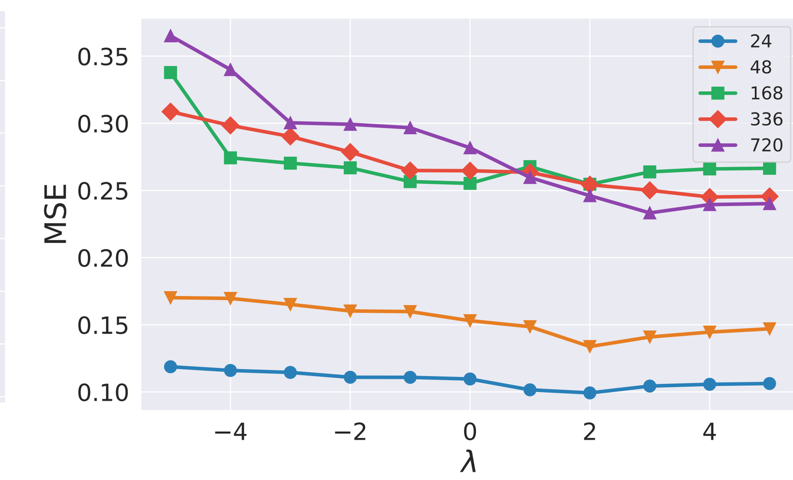
Transformers prefer Discrete Input

$$\mathcal{L}(y, \hat{y}, \hat{l}) = \mathcal{L}_{\text{Transformer}}(y, \hat{y}) - \lambda \mathcal{L}_{\text{Lip}}(\hat{l})$$

| Methods | | Transformer | | Transformer + Lip | | Informer | | Informer + Lip | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| $ETTh_1$ | 24 | 0.07047 | 0.20586 | **0.07019** | **0.20570** | 0.09842 | 0.24747 | **0.08882** | **0.23674** |
| | 48 | 0.18902 | 0.37046 | **0.16716** | **0.34974** | 0.15845 | 0.31907 | **0.12615** | **0.28333** |
| | 168 | 0.39773 | 0.55569 | **0.30811** | **0.48183** | 0.18314 | 0.34619 | **0.10579** | **0.25552** |
| | 336 | 0.41523 | 0.56902 | **0.41324** | **0.56402** | 0.22164 | 0.38720 | **0.11810** | **0.26959** |
| | 720 | 0.65586 | 0.75324 | **0.62233** | **0.73160** | 0.26883 | 0.43506 | **0.13131** | **0.28731** |
| $ETTh_2$ | 24 | 0.09449 | 0.24259 | **0.07560** | **0.20989** | 0.09309 | 0.24015 | **0.08626** | **0.22559** |
| | 48 | 0.15016 | 0.30996 | **0.13229** | **0.29278** | 0.15483 | 0.31445 | **0.13684** | **0.28936** |
| | 168 | 0.25197 | 0.41087 | **0.21046** | **0.37453** | **0.23193** | **0.38947** | 0.30071 | 0.43671 |
| | 336 | 0.22258 | 0.38170 | **0.20867** | **0.37298** | 0.26321 | 0.41659 | **0.24875** | **0.40827** |
| | 720 | 0.21932 | 0.38844 | **0.18445** | **0.35793** | 0.27722 | 0.43063 | **0.23646** | **0.39648** |
| $ETTm_1$ | 24 | 0.01279 | 0.08410 | **0.01210** | **0.08312** | 0.03016 | 0.13717 | **0.01815** | **0.09147** |
| | 48 | 0.08974 | 0.25869 | **0.02872** | **0.12820** | 0.06944 | 0.20255 | **0.05848** | **0.19686** |
| | 96 | 0.05341 | 0.17696 | **0.05182** | **0.15017** | 0.19414 | 0.37236 | **0.13336** | **0.30091** |
| | 288 | 0.22354 | 0.40455 | **0.13780** | **0.29825** | 0.40140 | 0.55355 | **0.30266** | **0.46864** |
| | 672 | **0.40726** | **0.55824** | **0.40726** | 0.55826 | 0.51164 | 0.64390 | **0.27543** | **0.45377** |
| Weather | 24 | 0.00223 | 0.03468 | **0.00154** | **0.02497** | 0.11676 | 0.25142 | **0.11256** | **0.23844** |
| | 48 | 0.00422 | 0.04106 | **0.00292** | **0.03026** | **0.17822** | **0.31846** | 0.19134 | 0.32408 |
| | 168 | 0.00537 | 0.05975 | **0.00319** | **0.04464** | 0.26585 | 0.39764 | **0.25138** | **0.37400** |
| | 336 | 0.00524 | 0.05772 | **0.00417** | **0.03673** | 0.29713 | 0.41571 | **0.24748** | **0.37725** |
| | 720 | 0.00933 | 0.07630 | **0.00272** | **0.03823** | 0.35875 | 0.46647 | **0.26479** | **0.39214** |
| Count | | 2 | | 49 | | 4 | | 46 | |




Change of $L_f$ during training




MSE with different $\lambda$

## Frequency Domain

In the Frequency Domain

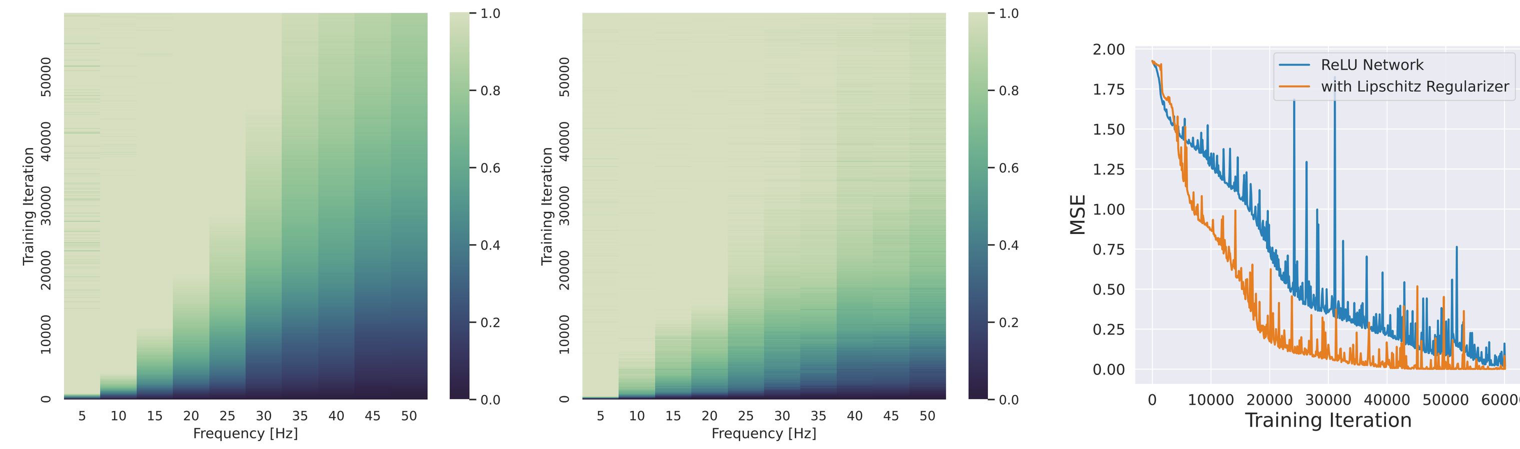
$$\sum_{i=0}^{n-1}(x_{i+1} - x_i)^2 \approx 4\pi^2 C \mathbb{E}_{p(\xi)}[\xi^2]$$

LipReg ⇔ Expectation Over the Frequency

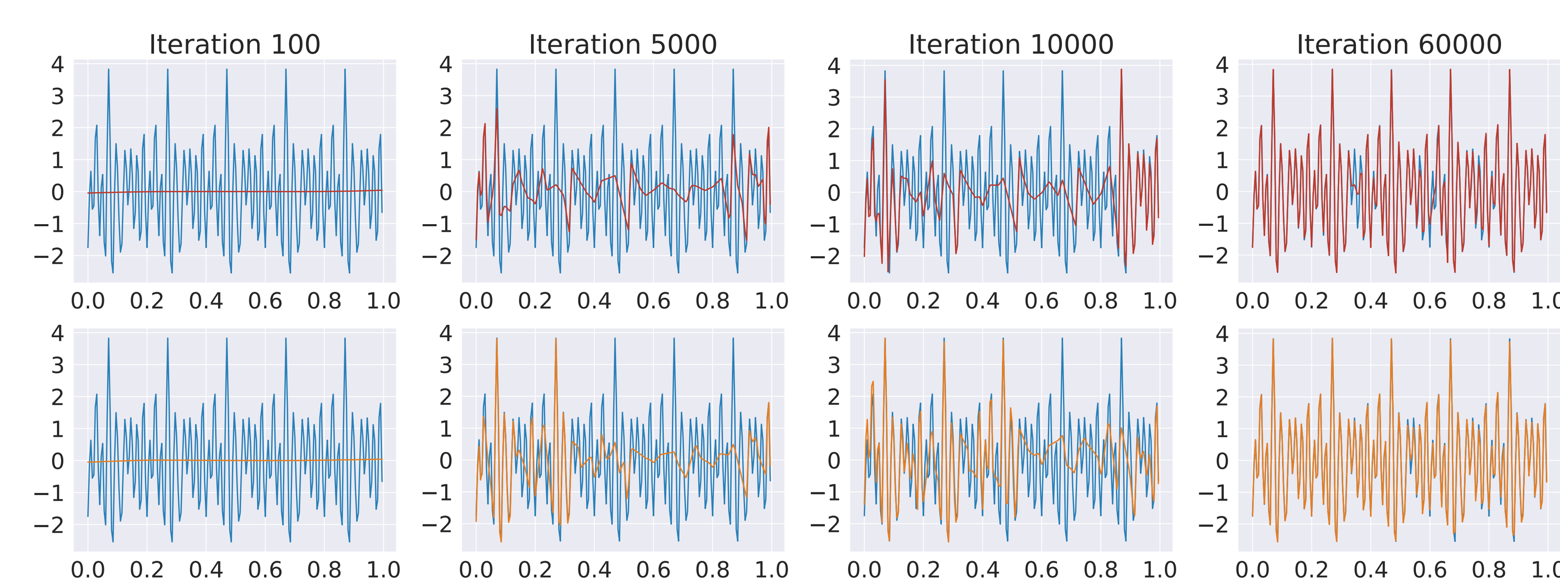
Spectral Bias: Low-frequency Part is Learned First

Use LipReg to Penalize the Low-frequency Part of NN

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_{\text{MSE}}(y, \hat{y}) - \lambda e^{-\epsilon t}\mathcal{L}_{\text{Lip}}(\hat{y})$$




(a) Without LipReg (b) With LipReg (c) Training Loss




Top: without LipReg; Bottom: with LipReg

Link to Paper

Link to Code