Understanding PFNs

Insights and issues

universität freiburg

Machine Learning Lab

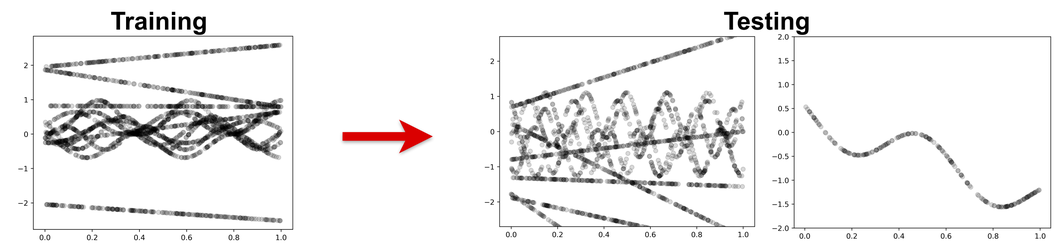
Marek Schuster, Pablo Marhoff

TL;DR

- **building custom PFNs requires in-depth knowledge of the model hyperparameters and high effort to design well performing priors**
- **Toy datasets offer the possibility to control the prior data distribution, which can provide insights into PFNs**
- **We propose a new mixed bucket distribution - especially for unevenly distributed data - with just one additional HP**

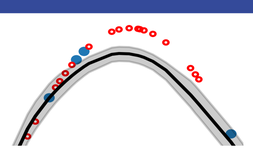
Toy Dataset

- **Desired properties:** visually interpretable, modular and parameterizable
- **Multiple priors:** lines, sines and sloped sines (convex combination)
- **Our approach to detecting ICL:** testing a trained PFN on deviating prior distributions (measuring OOD performance)

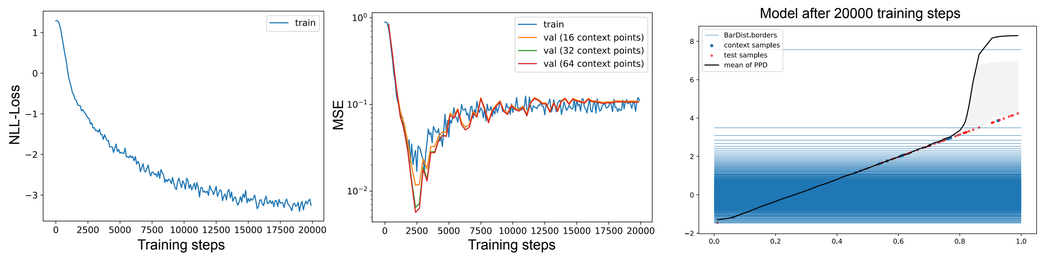


Importance of Performance Metrics

- NLL tends to underestimate the **regression performance on OOD data** (e.g., higher Amplitude):



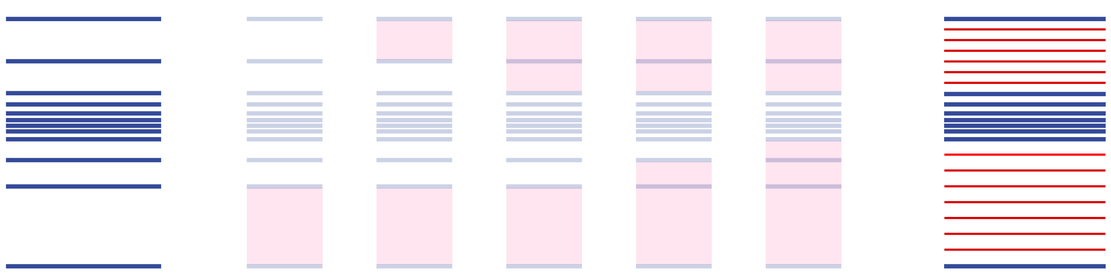
- Additional tracking of the **MSE revealed poor parameterization:**
  - ➡ Trained on NLL-Loss - “normal” convergence
  - ➡ However, MSE increased again after a small number of training steps. This could not be fixed by HPO
  - ➡ The increase could be traced back to an unfavorable bucket distribution

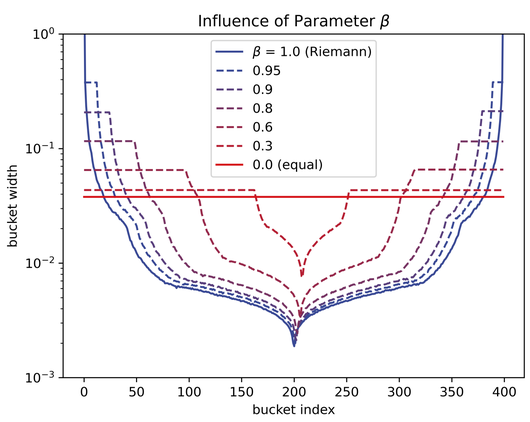


New Bucket Distribution

- Output discretization is essential for calculating the PPD, but the **Riemann distribution might lead to undersampled regions**
- **Tradeoffs:** having high density of buckets in the center vs. reducing sparsity at outer buckets **and** higher accuracy vs. faster training
- Hyperparameter defining the initial ratio $\beta = \frac{\textbf{Equal distributed buckets}}{\textbf{Riemann distributed buckets}}$

Riemann Distribution → Gathering the widest buckets until an optimal distribution of the evenly distributed buckets with regard to β is achieved. → **Mixed Distribution**





ECG Prior

Riemann

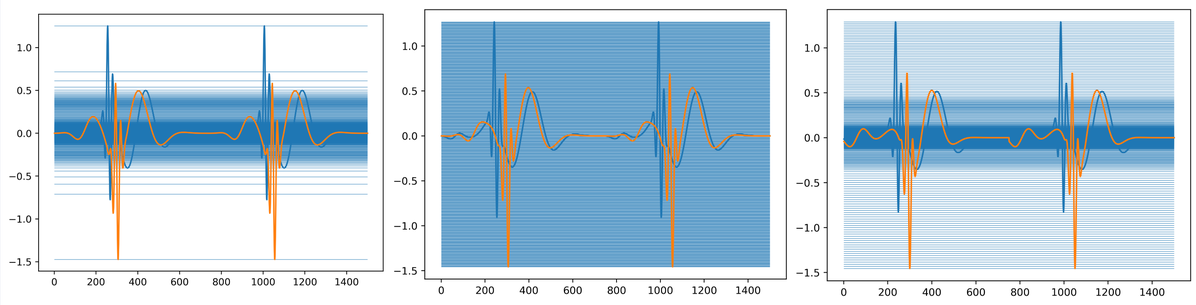
Too sparse on spikes

Equal

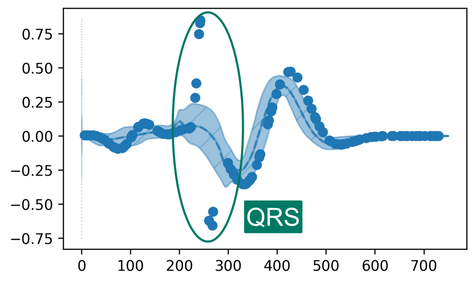
Easy fix - loses details

Mixed
(Hypothesis)

Precise in crucial area

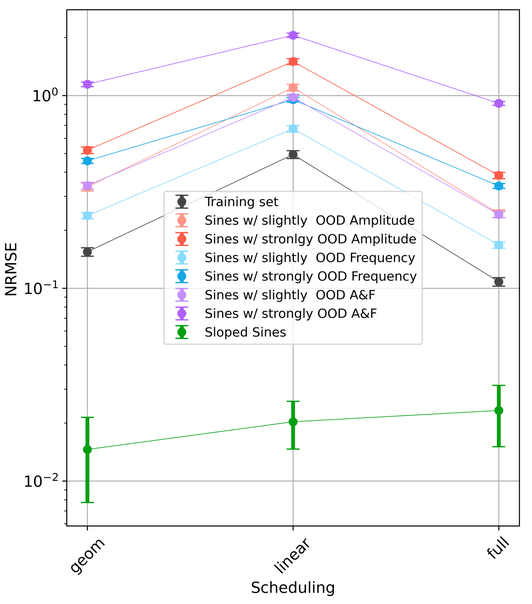


- Training on these ECG Waves did not yield results proving our hypothesis
- Examples show why: Model doesn't try to fit the fluctuating QRS-Section
- Problem likely stems from **our** implementation of the prior - small number of y-values make QRS *irrelevant* for the NLL-loss



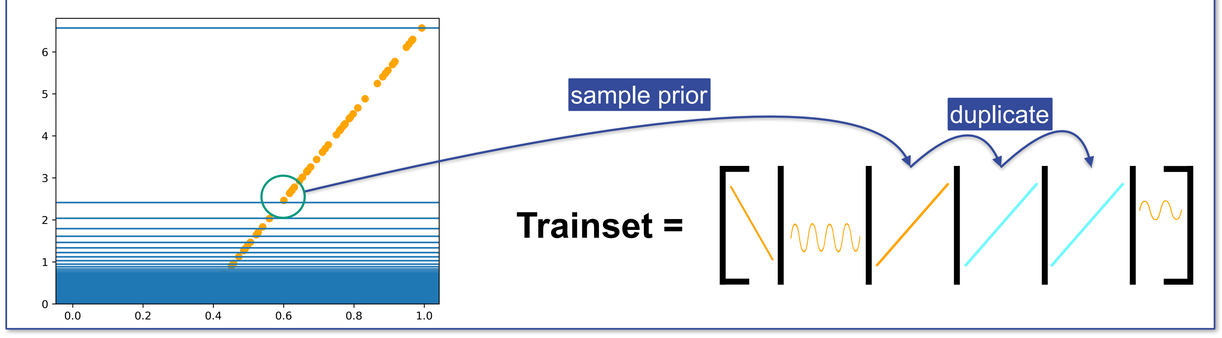
Curriculum Learning

- **Hypothesis:** Scaling difficulty during training improves Models ability to solve unseen problems
- **Experiment**: Increase parameter range of train-set prior linearly or geometrically. Also train w/o scheduler for comparison
- **Result:** Except sloped sines, training w/o scheduler had best overall performance on our test-sets
- **Conclusion:** CL seems irrelevant for **our** PFNs performance



Burstiness

- **Chan et al. (2022)** proposed great LLM performance stems from language being ***bursty*** **- e.g. outliers appear in clusters**
- **Our "outlier classes" are the outermost buckets**
- **Idea:** Duplicate all priors that touch them
- **Result:** For our case this did **not** yield **a noticeable performance increase**



Future Directions

- **Robustness regarding data shift**
- **Effects of a disproportionate number of buckets**
- **Influence of Noise on OOD-performance**
- **Fixing the number of train-samples**
- **Loss smoothing**
- **Other definitions of burstiness**