How Noisy Is Too Noisy? – Robust Extrapolation of Learning Curves with LC-PFN

Author: Razvan Gherasa(rgherasa@tudelft.nl) Supervisors: Sayak Mukherjee (s.mukherjee-3@tudelft.nl), Cheng Yan (c.yan-1@tudelft.nl)
Responsible Professor: Tom Viering (T.J.Viering@tudelft.nl)

TU Delft Delft University of Technology

1. INTRODUCTION

What are learning curves?
Learning curves (LCs) describe how a model's generalization error decreases as the size of the training dataset increases. They are used in machine learning to forecast future performance, decide when to stop training, and reduce computation or labeling costs.

Why study their shape?
In practice, learning curves are not always smooth. They can be non-monotonic, noisy, or irregular, making them hard to extrapolate reliably. This is especially true for real-world model training.

Why study noise?
Noise from label errors or metric instability often distorts learning curves, yet its effect on extrapolation methods like LC-PFN is still poorly understood. Studying this is key for reliable forecasting in AutoML.

2. RESEARCH QUESTION

How robust is LC-PFN to increasing levels of Gaussian noise in learning curves?

3. SUBQESTIONS

Research Question 1: How does noise in the observed portion of a learning curve affect LC-PFN's extrapolation performance?

Research Question 2: What mitigation strategies can improve robustness to noisy inputs?

Research Question 3: What mitigation strategies can improve robustness to noisy inputs?

4. Methodology

Dataset & Setup:

- LCDB 1.1 used — (265 datasets, 24 learners, 5 seeds, 5 resamples, 137 steps)
- Curves represent validation accuracy across 137 anchor points.
- 80/20 train/test split on all curves of length 80

Baseline:

- LC-PFN trained on clean curves (no noise), with data augmentation via random linear rescaling between anchor points.
- Evaluate on test curves with added Gaussian noise, for $\sigma \in \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$

Mitigation Strategies

1. Noisy Training
Goal: expose model to noise during training for improved robustness.
Train LC-PFN on noisy curves (σ = 0.05, 0.10, 0.30):

- Fixed: Constant noise across each epoch.
- Ramp: Noise increases gradually each epoch.

→ Helps the model generalize better under noisy conditions.

2. EMA Smoothing
Goal: reduce input volatility for more stable extrapolation.
Apply Exponential Moving Average at test time:
smoothed[t] = α * y[t] + (1 - α) * smoothed[t-1], with α = 0.2
→ Chosen α = 0.2 balances reactivity to change and smoothing of high-frequency noise.

3. MC(Monte Carlo) Dropout
Goal: estimate prediction uncertainty and reduce overfitting effects.
At test time, perform 20 stochastic forward passes with dropout active (instead of disabled as usual).
→ This generates diverse predictions for the same input.
→ The final prediction is the average across these samples.
→ Helps reduce variance from individual predictions and mitigate overconfidence on noisy inputs.

5. CONCLUSIONS

- RQ1: Clean-trained LC-PFNs degrade sharply under noise — MAE nearly doubles at σ = 0.10 and increases 5.6× at σ = 0.30, highlighting their sensitivity to even mild input corruption.
- RQ2: Noise-aware training, especially with a gradual ramp (Ramp-0→0.05 or 0→0.10), significantly improves robustness across noise levels, cutting MAE by over 60% in noisy conditions.
- RQ3: Test-time Exponential Moving Average (EMA) smoothing greatly boosts robustness with minimal cost, and when combined with Ramp-0→0.05 training, delivers the best overall performance—achieving up to 75% MAE reduction without harming clean-data accuracy.

Further work

- *Broader data coverage: Extend training to include all curves with ≥40 valid points from LCDB 1.1, improving applicability to real-world, incomplete logs.*
- *Expanded evaluation: Explore alternative cutoff lengths and additional metrics (e.g., CRPS, ECE) to assess extrapolation quality and uncertainty calibration more comprehensively.*
- *Optimized configurations: Systematically tune EMA smoothing parameters and MC-Dropout rates, and experiment with longer training or diverse architectures to enhance robustness further.*

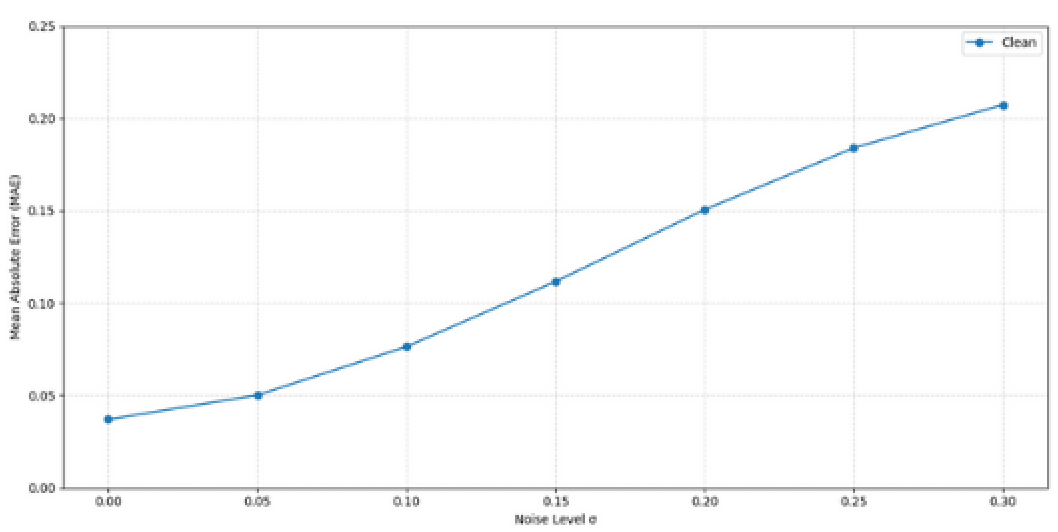
References

- [1] Viering, T., & Loog, M. (2022). The Shape of Learning Curves: A Review. IEEE TPAMI, 45(6), 7799–7819.
- [2] Viering, T., et al. (2024). From Epoch to Sample Size: Developing New Data-driven Priors for Learning Curve Prior-Fitted Networks. AutoML Conference (Workshop Track).
- [3] Yan, C., Mohr, F., & Viering, T. (2025). LCDB 1.1: A Database Illustrating Learning Curves Are More Ill-Behaved Than Previously Thought. arXiv:2505.15657.

Results RQ1

LC-PFN is highly sensitive to noise: Even a small level of noise (σ = 0.05) significantly increases the error (up to 35% more MAE).

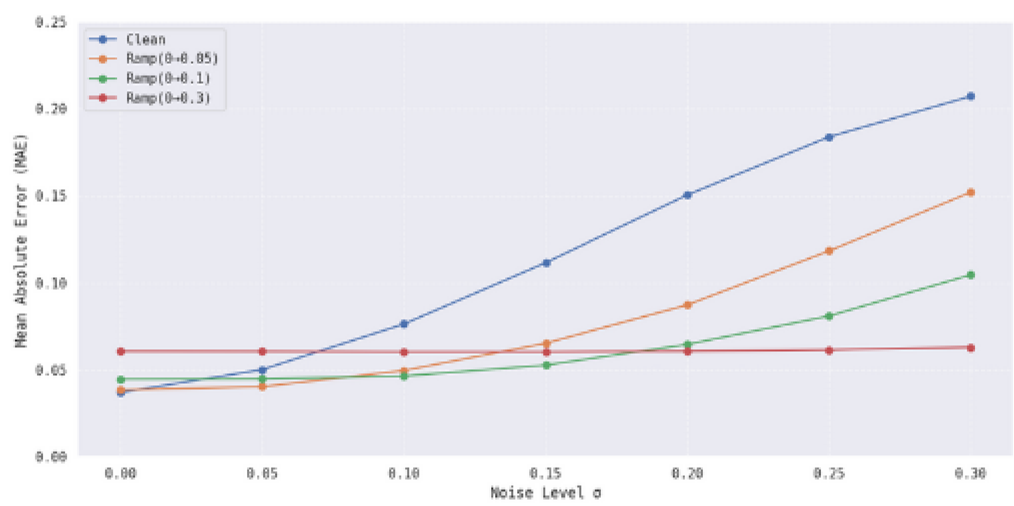
Error increases predictably: As noise increases, prediction accuracy degrades linearly, with the error more than 5 times higher at σ = 0.30 compared to the clean baseline.



Results RQ2

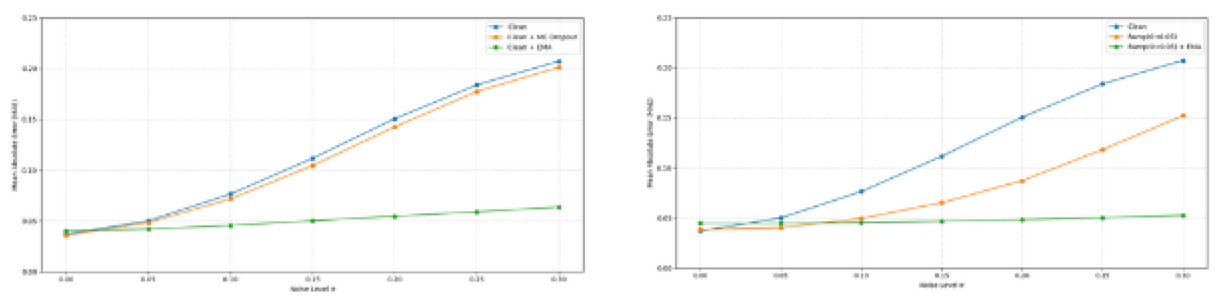
Ramp-based training improves robustness: Gradual noise exposure during training significantly enhances performance under noisy conditions.

Each ramp excels in its targeted noise range:
Ramp-0→0.05 performs best with low noise (σ = 0.05).
Ramp-0→0.10 excels at moderate noise (σ = 0.10–0.15).
Ramp-0→0.30 shows superior performance at high noise levels ($\sigma \geq$ 0.20).



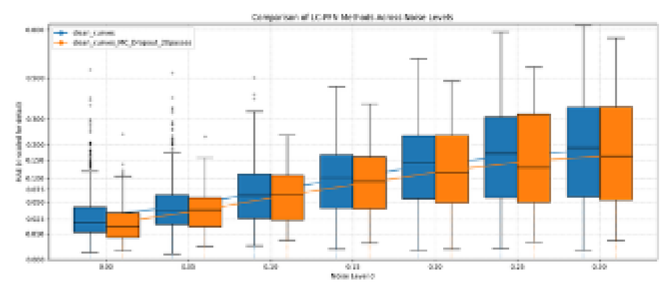
Results RQ3

EMA Smoothened Input



Best configuration: The Ramp-0→0.05 + EMA combination achieves state-of-the-art performance, reducing error by up to 75% across different noise levels, while maintaining efficiency and robustness.

MC Dropout



20x higher test-time computation cost

Findings:

- 7.2% lower mean MAE
- 5.8% lower mean SD