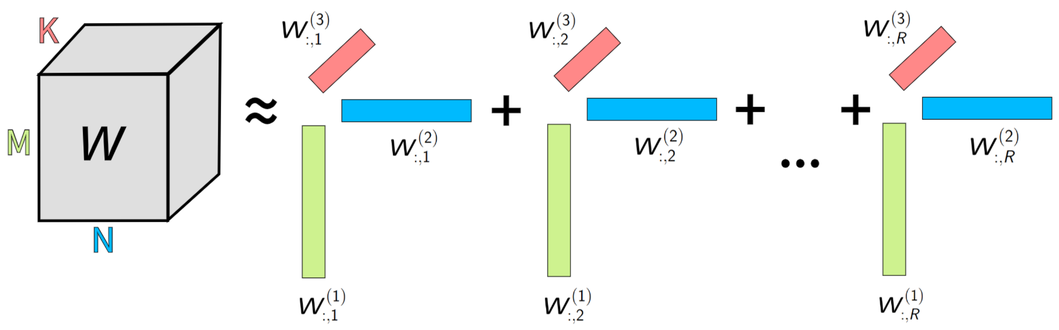
# Tensor Network Based Feature Learning Model

Albert Saiapin[1] Kim Batselier[1]

[1] Delft Center for Systems and Control, Delft University of Technology

## 1 Introduction

In prior works [1] [2], the authors explored a kernel-based model parameterized by a CP tensor network [3]:




$$y = \phi_\theta(x)^\top w = \left(\psi_\theta^{(D)}(x_D) \otimes \cdots \otimes \psi_\theta^{(1)}(x_1)\right)^\top w,$$

$$w = \sum_{r=1}^{R} W_{:,r}^{(D)} \otimes \cdots \otimes W_{:,r}^{(1)}.$$

**Why is this CP network-based kernel model interesting?**

- Leverages deterministic features to achieve better convergence compared to random Fourier features;
- Allows to work with large-scale datasets compared to Kernel methods (GP, Kernel Ridge regression);
- Mitigates the curse of dimensionality associated with a direct solution approach: $\mathcal{O}\left(e^D\right) \rightarrow \mathcal{O}\left(D\right)$.

**What is the problem?**

How to choose hyperparameters $\theta$ of the feature map $\phi_\theta(x)$?

## 2 Feature Learning Model

**Problem Solution - FL model**

$$f(x) = \left[\sum_{p=1}^{P} \lambda_p \psi_{\theta_p}^{(D)}(x_D) \otimes \cdots \otimes \psi_{\theta_p}^{(1)}(x_1)\right]^\top w$$

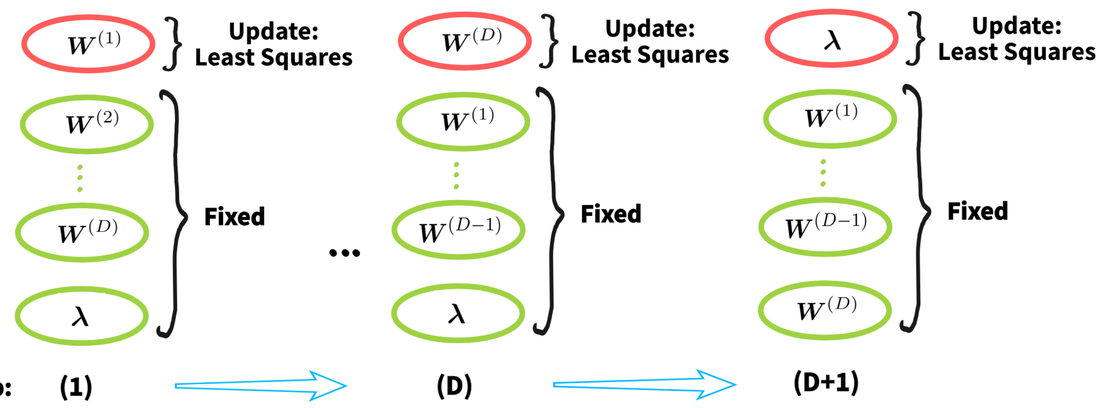
Optimal model parameters - $[\lambda_p]_{p=1}^P$ and $\left[W^{(k)}\right]_{k=1}^D$ can be learned from the data by solving non-convex non-linear optimization problem:

$$\frac{1}{2}\|y - \sum_{p=1}^{P} \lambda_p \Phi_p w\|_2^2 + \frac{\alpha}{2}\|w\|_2^2 + \frac{\beta}{2}\|\lambda\|_2^2,$$

$$\text{s.t. } w = \sum_{r=1}^{R} W_{:,r}^{(D)} \otimes \cdots \otimes W_{:,r}^{(1)}.$$

**What are the advantages of the FL model?**

- Replaces discrete hyperparameters search with continuous optimization;
- Enables combining various feature maps (Fourier, polynomial, etc.);
- Uses effective Alternating Least Squares (ALS) optimization:




## 3 Results

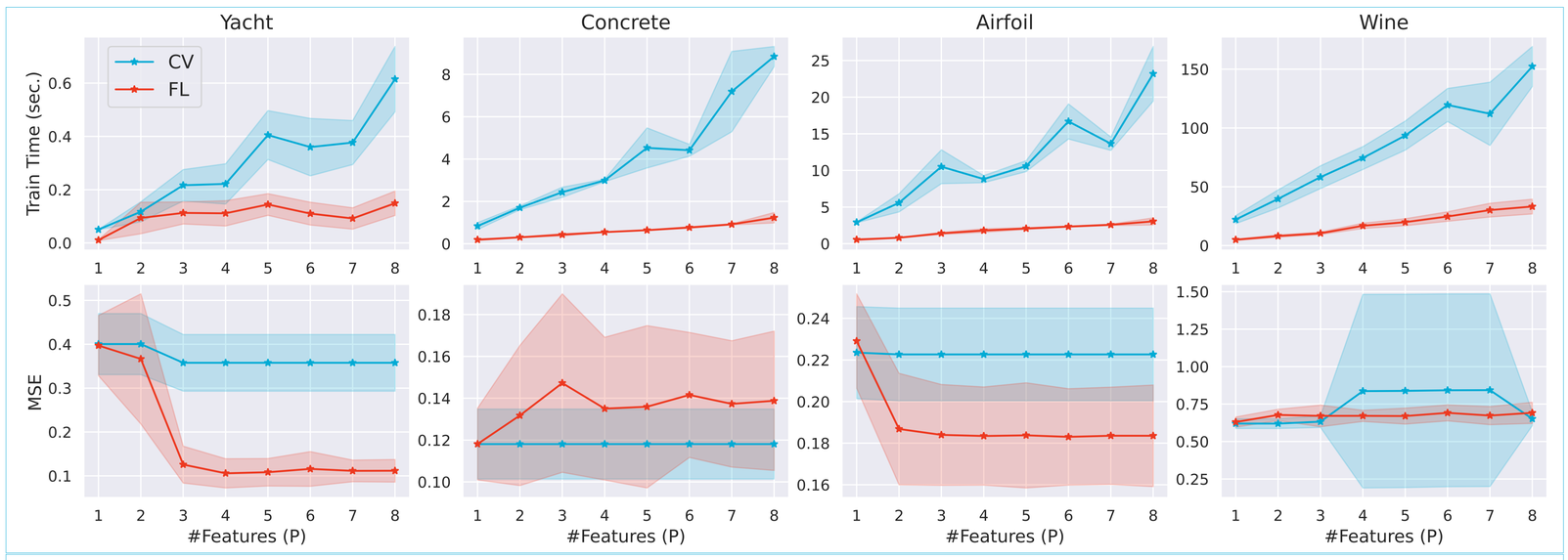



Figure: Plots of the training time (first row) and test MSE (second row) of FL and CV models (red and blue curves respectively) as a function of the number of features $P$ for different real-life datasets (column-wise). Solid lines represent mean metric calculations and shaded regions depict $\pm 1$ standard deviation around the mean across 10 restarts.

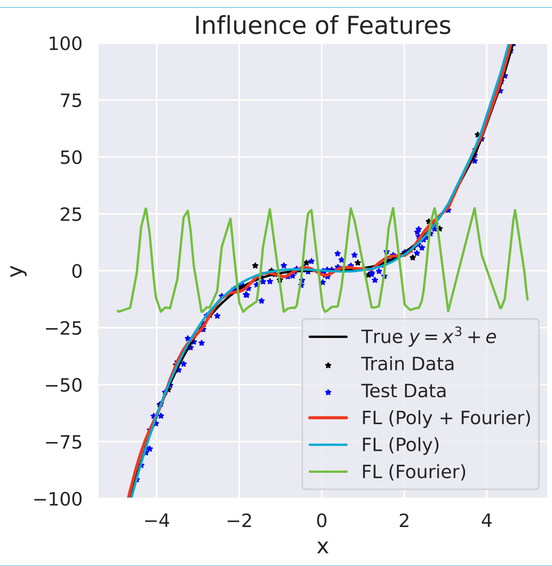



Figure: Plot illustrating the approximation quality of different feature maps. 'Poly' denotes polynomial features (powers), while 'Fourier' represents Fourier features.

**Main Outcome**

The proposed FL model requires consistently less time to train compared to the conventional cross-validation. Likewise the prediction error of the FL model is either similar to CV (shaded regions intersect) or significantly lower (Yacht data) that demonstrates the superiority of the FL model.

## 4 Future Work

- Develop a parallel FL model by utilizing the model's CP structure;
- Explore a probabilistic formulation of the FL model, enabling uncertainty estimation for classification and regression tasks;
- Incorporate a sparsity-inducing prior over factors $W_{(d)}$, resulting in automatic rank determination.

## References

[1] A. Haliassos, K. Konstantinidis, and D. Mandic. *Supervised Learning for Nonsequential Data: A Canonical Polyadic Decomposition Approach*. 2021.

[2] F. Wesel and K. Batselier. *Quantized Fourier and Polynomial Features for more Expressive Tensor Network Models*. 2024.

[3] A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, and D. P. Mandic. *Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 1 Low-Rank Tensor Decompositions*. 2016.

**Acknowledgements:** This publication is part of the project Sustainable learning for Artificial Intelligence from noisy large-scale data (with project number VI.Vidi.213.017) which is financed by the Dutch Research Council (NWO).