Stochastic Machine Learning
Chapter 04 - Prior Fitted Networks

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

SS 2024

PFNs

- Now we want to study also related approaches like **mullertransformers**.
- The authors consider a supervised training approach. In our sense this means for example we observe $Y_1, \dots, Y_n$ and estimate Y_{n+1} such that

$$(\xi^i, \eta^i) = ((Y_1^i, \dots, Y_n^i), Y_{n+1}^i), \qquad i = 1, \dots, N$$

  would be our training data. Note that (X_n) is typically not observed - this is our filtering problem.
- The Bayesian predictor is achieved via some latent state (in our case $X = (X_1, \dots, X_n)$ such that

$$\hat{\eta} = \hat{\eta}(\xi) = \int_{\mathcal{X}} \hat{F}(x, \xi)\, df(x|\xi)$$

- The function $\hat{F}(x, \xi)$ is the best predictor of η given x, ξ, in our case

$$\hat{F}(x, \xi) = E[Y_{n+1}|X_1, \dots, X_n, Y_1, \dots, Y_n],$$

  which is often easy to compute. The filtering problem is of course to compute

$$f(x|\xi) = f(x_1, \dots, x_n|y_1, \dots, y_n)$$

  (when written in densities).

We look into the paper and compare it to the Kalman filter approach