# Uncertainty Estimation with Recursive Feature Machines

**Daniel Gedon**[1], Amirhesam Abedsoltan[2], Thomas B. Schön[1], Mikhail Belkin[2]
[1]Uppsala University (Sweden), [2]UC San Diego (USA)

## Motivation

**Setup:** Regression with tabular data

**Point estimate** + **Uncertainty quantification**

RMSE, ... NLL, coverage, ...

**Boosting** vs. **Gaussian Process**

**Can we include feature learning in GPs?**

## Method

$x_* \rightarrow$ **Recursive Feature Machine** $\xrightarrow{f_M}$ / $\xrightarrow{K_M}$ **GP** (post. mean $\rightarrow f_M$; post. cov. $\rightarrow \mathbb{V}[f_M]$)

**GP-RFM**

Feature learning Uncertainty quant.

Restrict to diag. features: **GP-RFM-diag**

Penalize off-diagonal elements:

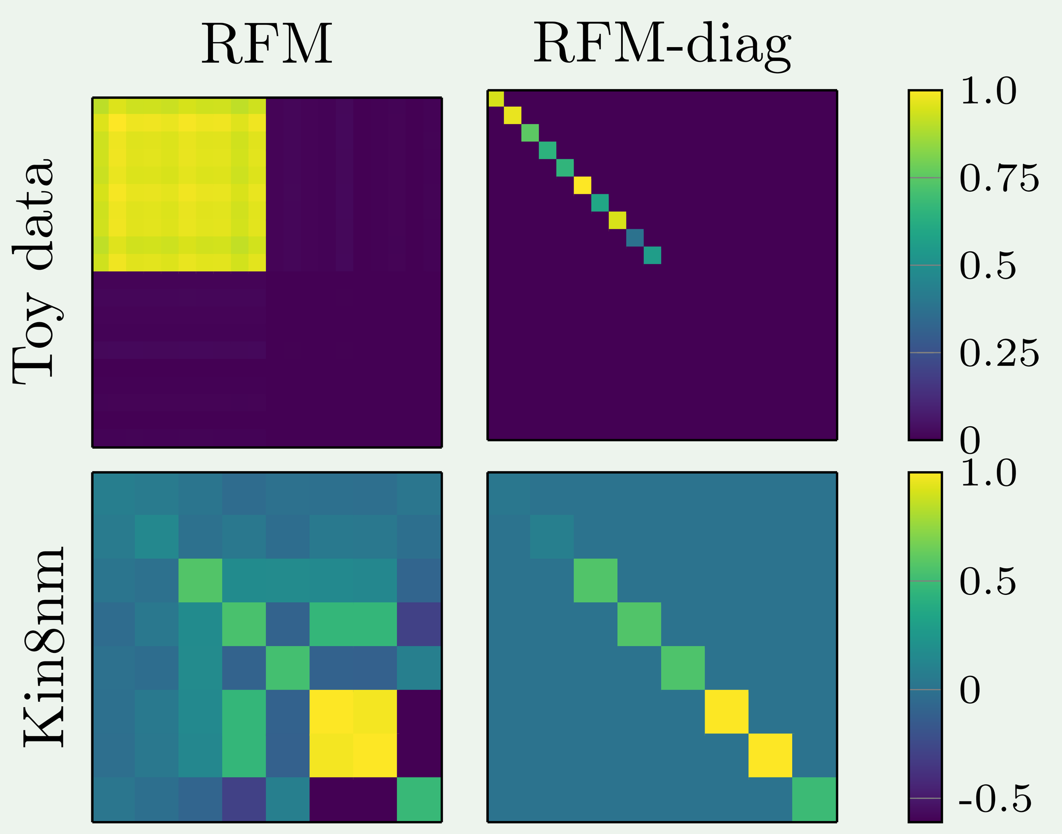
$$M = \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_M(x) \nabla_x f_M(x)^\top + \lambda_M I_d$$

## Visualising feature matrix $M$

**Data:** $x \sim \mathcal{U}(0_d, 1_d)$, $y = (\sum_{j=1}^{10} x_{[j]})^2$
$\rightarrow$ introduce correlation.

**Interpretation:**

- RFM features capture correlation.
- Correlation in real-world data sets.




## Gaussian Process (GP) | Recursive Feature Machine (RFM) [1]

$x_i \rightarrow$ **Gaussian Process** $\mathcal{N}(\mu, \Sigma)$, $\mu = f(x)$, $\Sigma = \mathbb{V}[f(x)]$ $\rightarrow y_i$

$x_i \rightarrow$ **Kernel Machine** $f_M(x)$ $\rightarrow y_i$ ⇄ **Feature Matrix** $M = \mathrm{AGOP}(f_M)$

**Model parameterization**

pred. function $f(x) = k(x, X)\alpha$

RBF (or Laplace) kernel
$K_M(x, z) = \exp(-\gamma \|x - z\|_M^2)$

with Auto. Relevance Det. (ARD)
$M^{-1} = \mathrm{diag}([\ell_1^2, \dots, \ell_d^2])$

Laplace kernel
$K_M(x, z) = \exp(-\gamma \|x - z\|_M)$

with Mahalanobis distance
$\|x - z\|_M = \sqrt{(x - z)^\top M (x - z)}$

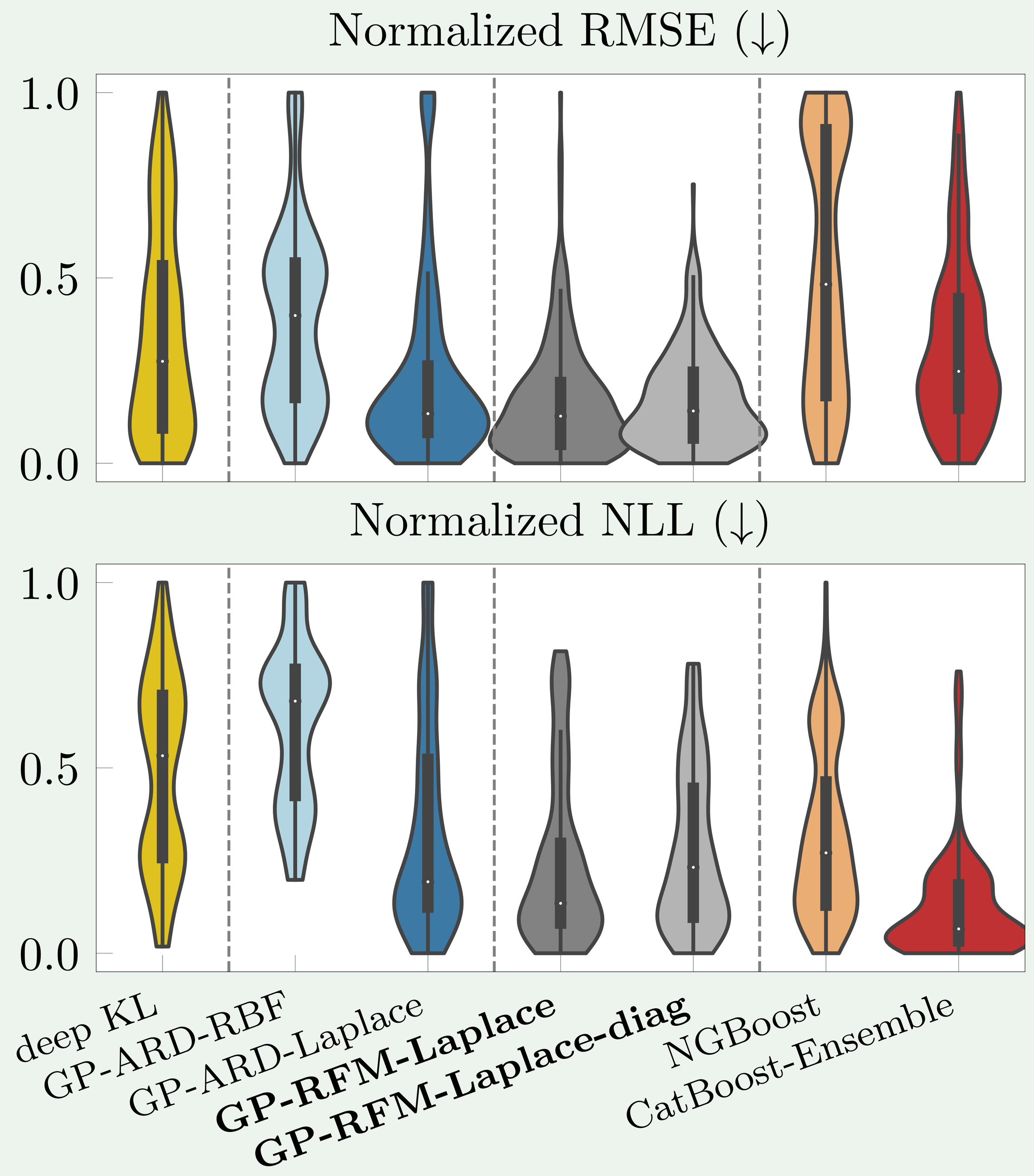
**Training procedure**

Maximum Likelihood Estimation
$\arg\min_\theta - \log p(y \mid X, \theta)$

with $\theta = \{\ell_1, \dots, \ell_d\}$

For $t$ in $T$:
1. Solve for kernel weights
$$\alpha = (k_M(x, X) + \lambda_\alpha I_n)^{-1} y$$
2. Average gradient outer product (AGOP)
$$M = \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_M(x_i) \nabla_x f_M(x_i)^\top$$

## Main results




Negative log-likelihood on tabular data.

| Dataset | Gauss. Process ARD-Lap. | Gauss. Process RFM | Boosting NG | Boosting Cat |
|---|---|---|---|---|
| cpu-act | 2.30 | 2.21 | 2.33 | **2.17** |
| pol | 2.84 | 2.73 | 3.55 | **2.09** |
| elevators | -4.75 | **-4.86** | -4.48 | -4.73 |
| isolet | 3.43 | **2.34** | 2.71 | 2.52 |
| wine | **0.95** | **0.95** | 1.04 | 1.03 |
| Ailerons | -7.33 | -7.37 | **-7.42** | -7.41 |
| houses | -0.10 | -0.07 | 0.07 | **-0.12** |
| houses-16H | 0.72 | 0.69 | 0.57 | **0.51** |
| Bra-houses | -1.82 | -2.11 | -2.18 | **-2.66** |
| bike | 6.03 | 6.04 | 5.62 | **5.58** |
| house-sales | -0.30 | **-0.32** | -0.27 | -0.31 |

**best**, 2nd best

**OpenML datasets [2].**

- 16 tabular datasets
- 5 - 613 features
- 6,497 - 22,784 samples

**Setup.**

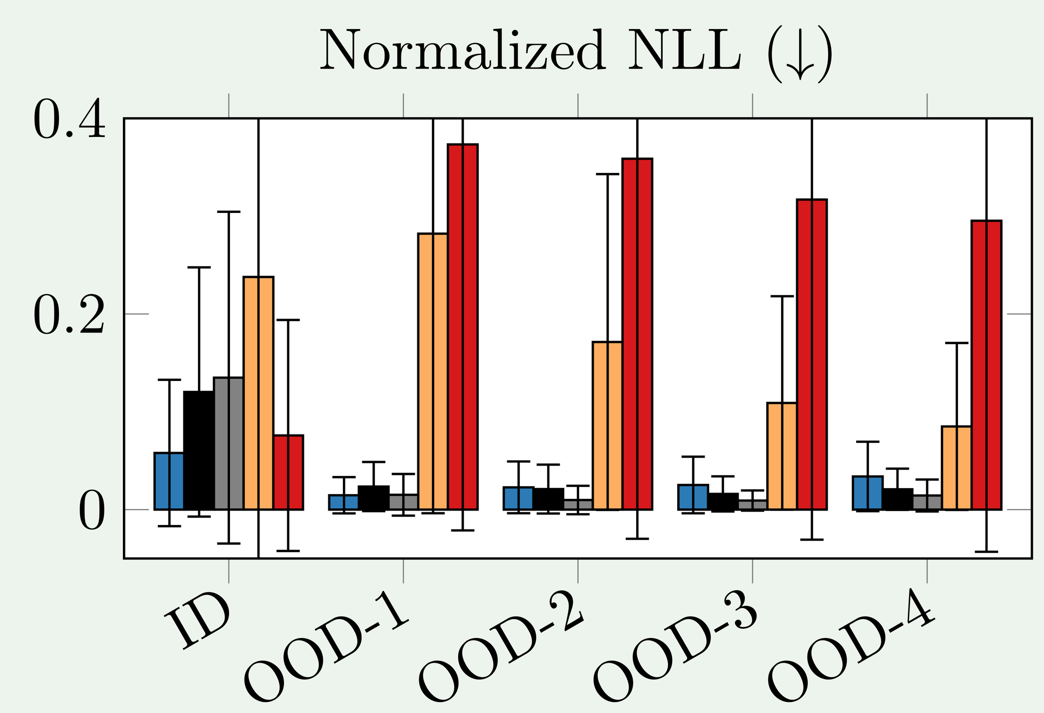
- 20 seeds
- normalised metrics for each dataset

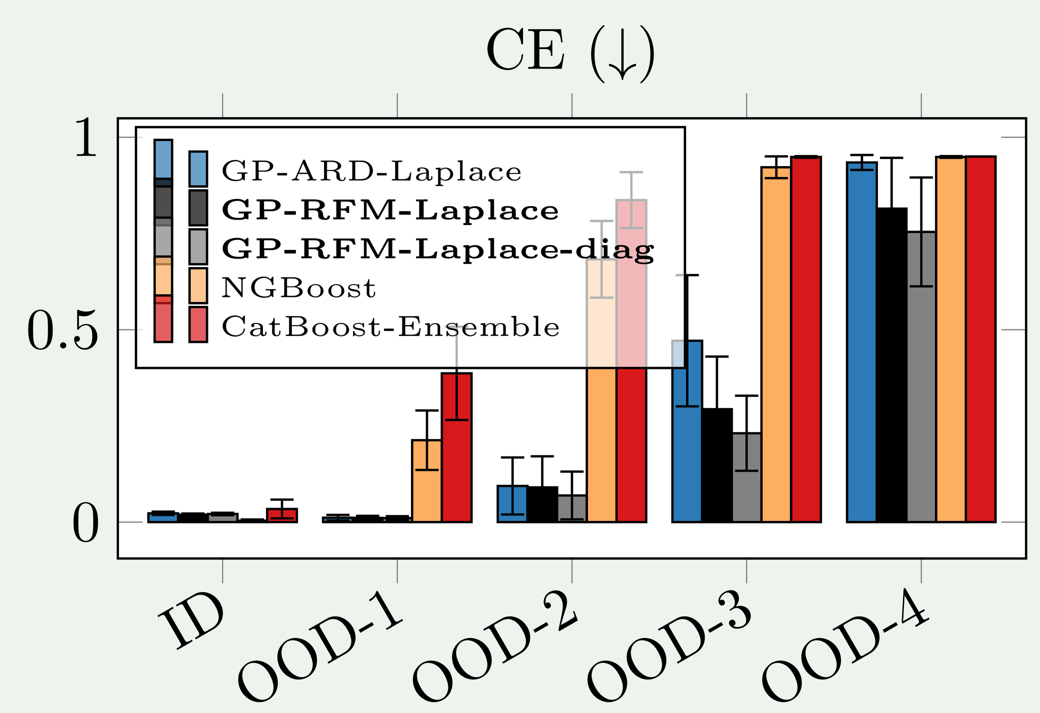
**Interpretation.**

- GP-RFM best in RMSE
- GP-RFM overall competitive

## Out-of-distribution data

**Data.** Housing data with increasing target (price) OOD shift.
**Interpretation.** GP-RFM most reliable method under OOD shift.




## Conclusion

**GP + RFM → competitive results**

**Missing in GP literature.**
Full feature matrix $M$ often outperforms diagonal approach

## References

[1] Radhakrishnan A, Beaglehole D, Pandit P, Belkin M. "*Mechanism for feature learning in neural networks and backpropagation-free machine learning models*". Science, 2024.
[2] Vanschoren J, Van Rijn JN, Bischl B, Torgo L. "*OpenML: networked science in machine learning.*" ACM SIGKDD Explorations Newsletter, 2014.