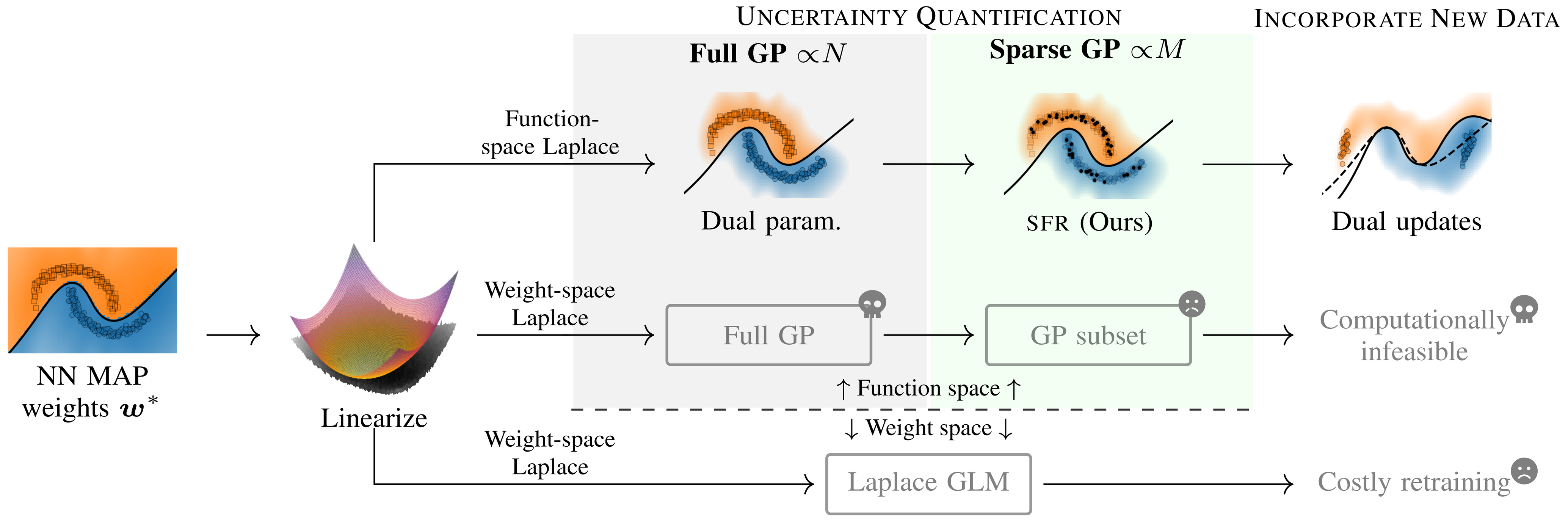
Function-space Parameterization of Neural Networks for Sequential Learning

Aidan Scannell*[1,2] Riccardo Mereu*[1] Paul Chang[1] Ella Tamir[1] Joni Pajarinen[1] Arno Solin[1]

[1]Aalto University [2]Finnish Center for Artificial Intelligence *Equal Contribution

Aalto University



TL;DR

- Neural networks (NNs) have limitations: *estimating uncertainty*, *incorporating new data*, and *avoiding catastrophic forgetting*.
- Our method, Sparse Function-space Representation (SFR):
  - converts NN to sparse Gaussian process (GP) via dual parameters,
  - gives good uncertainty estimates,
  - can incorporate new data without retraining,
  - can maintain a functional representation for continual learning,
  - can be used for uncertainty-guided exploration in model-based RL.

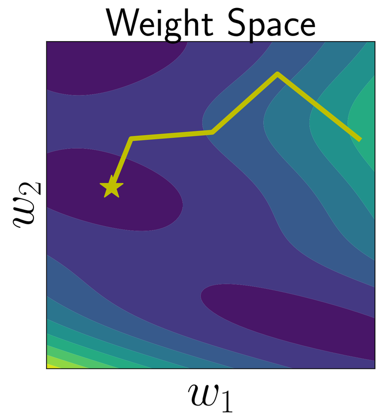
Motivation

	SFR (Ours)	GP	NN
Uncertainty estimates	✅	✅	❌
Image inputs	✅	❌	✅
Large data	✅	❌	✅
Incorporate new data	✅	✅	❌

1. Train Neural Network

Inputs: NN $f_{\mathbf{w}}(\cdot)$, data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$
Outputs: Maximum A-Posteriori (MAP) weights

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathcal{D}, \mathbf{w}) = \sum_{i=1}^{N} \underbrace{\ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}_{-\log p(y_i \mid f_{\mathbf{w}}(\mathbf{x}_i))} + \underbrace{\mathcal{R}(\mathbf{w})}_{-\log p(\mathbf{w})}$$



2. From NN to Function-space Laplace

(1) Linerised NN $f_{\mathbf{w}^*}(\mathbf{x}) \approx \nabla_{\mathbf{w}} f_{\mathbf{w}^*}(\mathbf{x})^\top \mathbf{w} \rightarrow$ function space formulation:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mathbf{0}, \kappa(\boldsymbol{X}, \boldsymbol{X})) \quad \text{with} \quad \kappa(\mathbf{x}, \mathbf{x}') = \frac{1}{\delta} \nabla_{\mathbf{w}} f_{\mathbf{w}^*}(\mathbf{x})^\top \nabla_{\mathbf{w}} f_{\mathbf{w}^*}(\mathbf{x})$$

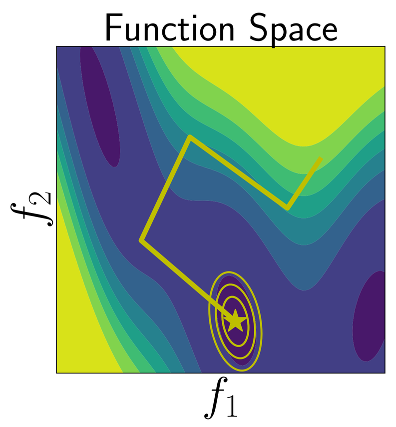
(2) Convert training objective to function space,

$$\mathcal{L}(\mathcal{D}, \mathbf{w}) = -\textstyle\sum_{i=1}^N \log p(y_i \mid f_i) - \log p(\mathbf{f}).$$

(3) Function-space Laplace approximation:

$$p(\mathbf{f} \mid \mathcal{D}) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mathbf{m}_{\mathbf{f}}, \boldsymbol{S}_{\mathbf{f}})$$

⚠ GP predictive posterior is computational expensive.



Sparse Function-space Representation (SFR)

- Sample inducing inputs $\boldsymbol{Z} \subseteq \boldsymbol{X}$ from training inputs $\boldsymbol{X}$.
- SFR predictive posterior:

$$\mathbb{E}_{q(f_i)}[f_i] \approx \mathbf{k}_{\mathbf{z}i}^\top \boldsymbol{K}_{\mathbf{zz}}^{-1} \boldsymbol{\alpha}_{\mathbf{u}} \quad \text{and}$$
$$\mathrm{Var}_{q(f_i)}[f_i] \approx k_{ii} - \mathbf{k}_{\mathbf{z}i}^\top [\boldsymbol{K}_{\mathbf{zz}}^{-1} - (\boldsymbol{K}_{\mathbf{zz}} + \boldsymbol{B}_{\mathbf{u}})^{-1}] \mathbf{k}_{\mathbf{z}i}$$

with sparse dual parameters,

$$\boldsymbol{\alpha}_{\mathbf{u}} = \textstyle\sum_{i=1}^N \mathbf{k}_{\mathbf{z}i} \hat{\alpha}_i \quad \text{and} \quad \boldsymbol{B}_{\mathbf{u}} = \sum_{i=1}^N \mathbf{k}_{\mathbf{z}i} \hat{\beta}_i \mathbf{k}_{\mathbf{z}i}^\top$$

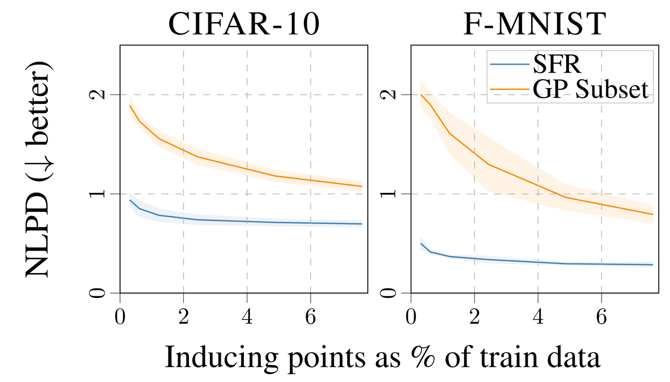
$$\hat{\alpha}_i := \nabla_f \log p(y_i \mid f)|_{f=f_i} \quad \text{and} \quad \hat{\beta}_i := -\nabla_{ff}^2 \log p(y_i \mid f_i)|_{f=f_i}$$

- Incorporating new data $\mathcal{D}^{\text{new}}$ with dual updates is easy,

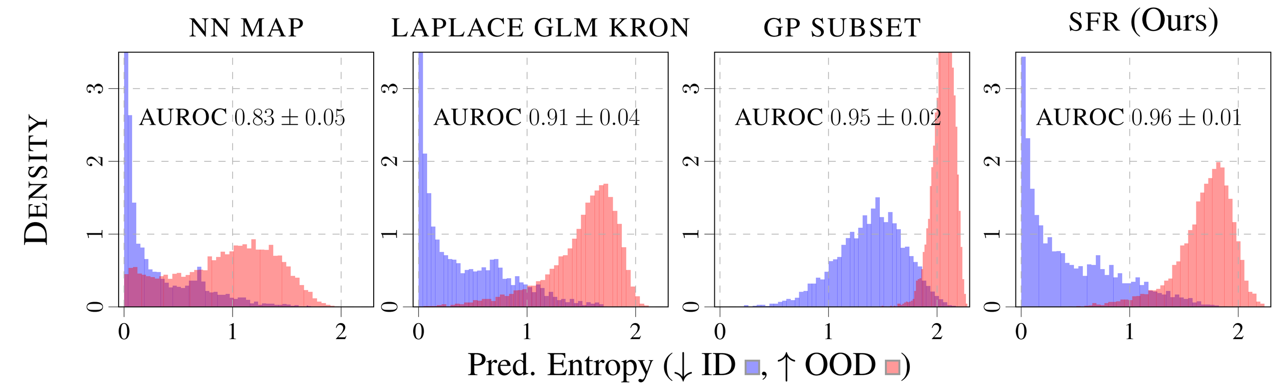
$$\boldsymbol{\alpha}_{\mathbf{u}} \leftarrow \boldsymbol{\alpha}_{\mathbf{u}} + \underbrace{\textstyle\sum_{\mathbf{x}_i, y_i \in \mathcal{D}^{\text{new}}} \mathbf{k}_{\mathbf{z}i} \hat{\alpha}_i}_{\text{update}} \quad \text{and} \quad \boldsymbol{B}_{\mathbf{u}} \leftarrow \boldsymbol{B}_{\mathbf{u}} + \underbrace{\textstyle\sum_{\mathbf{x}_i, y_i \in \mathcal{D}^{\text{new}}} \mathbf{k}_{\mathbf{z}i} \hat{\beta}_i \mathbf{k}_{\mathbf{z}i}^\top}_{\text{update}}$$

Sparsification in Image Classification

- SFR (—) requires fewer inducing points than a GP subset (—) to achieve good (low) NLPD.



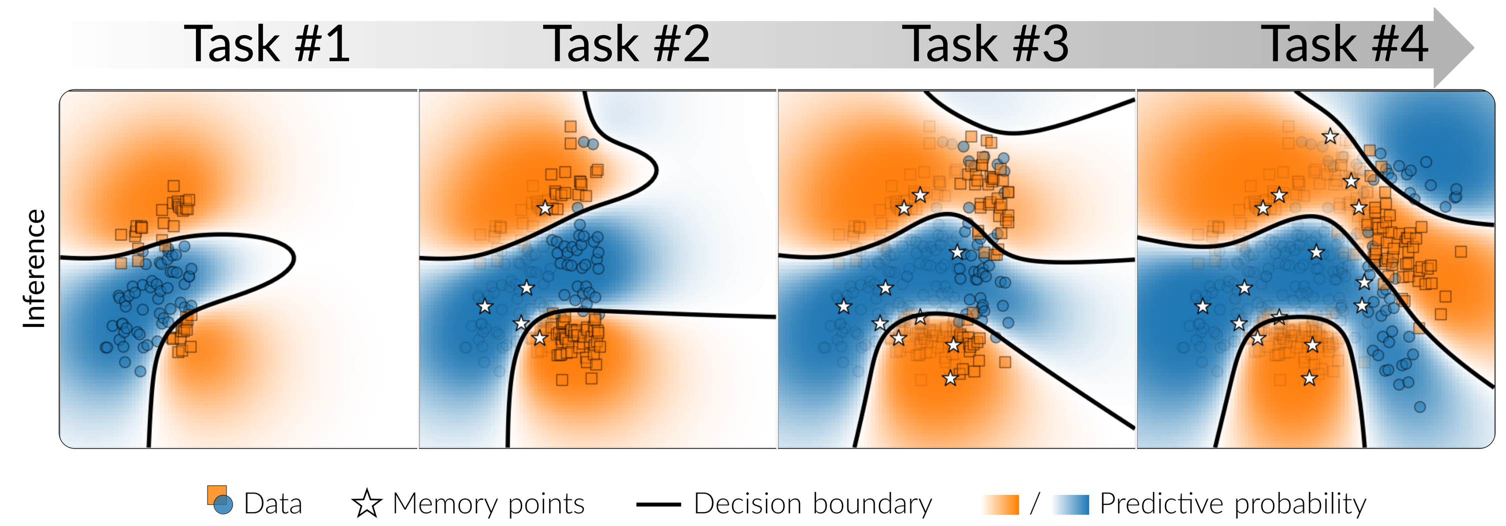
OOD Detection with CNNs



- SFR demonstrates good out-of-distribution (OOD) detection as it has low predictive entropy for in-distribution data (FMNIST, ■) and high predictive entropy for out-of-distribution data (MNIST, ■).

Continual Learning

- SFR is effective for function-space regularization in CL.

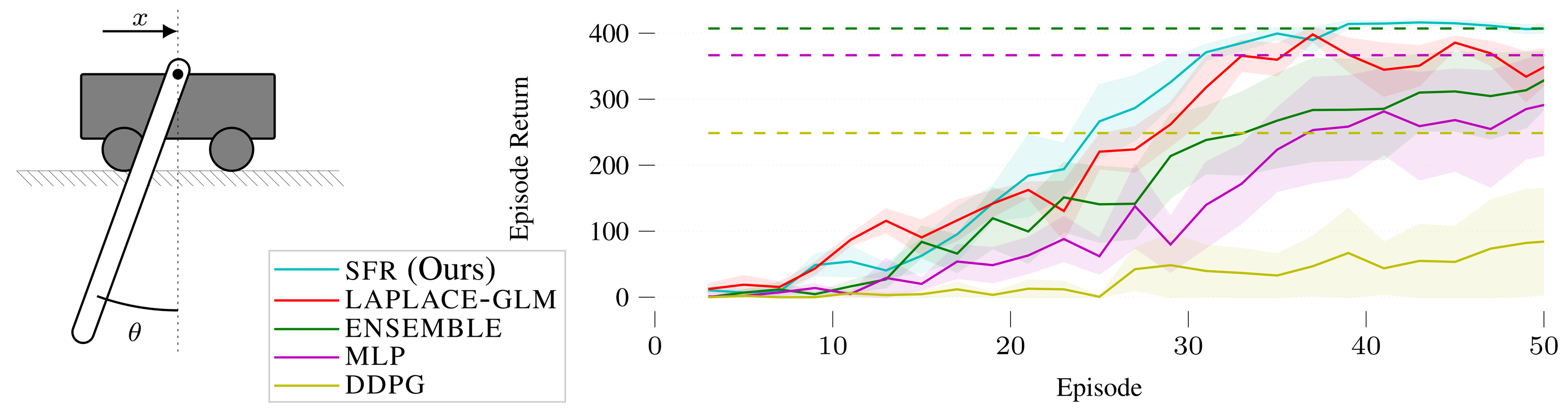


Method	S-MNIST (SH) 40 pts./task	S-MNIST (SH) 200 pts./task	S-FMNIST (SH) 200 pts./task	P-MNIST (SH) 200 pts./task
DER	85.26±0.54	92.13±0.45	**82.03**±0.57	93.08±0.11
FROMP	75.21±2.05	89.54±0.72	78.83±0.46	94.90±0.04
S-FSVI	84.51±1.30	92.87±0.14	77.54±0.40	**95.76**±0.02
SFR (Ours)	**89.22**±0.76	**94.19**±0.26	**81.96**±0.24	95.58±0.08

Model-based Reinforcement Learning

Strategy Policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ based on posterior sampling:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\boldsymbol{\epsilon}_{0:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \mid \mathbf{s}_{t+1} = \tilde{f}(\mathbf{s}_t, \mathbf{a}_t) \right] \quad \text{s.t. } \tilde{f} \sim q_{\mathbf{u}}(f \mid \mathcal{D}),$$



- SFR's uncertainty estimates can improve sample efficiency in model-based RL by guiding exploration.

aidan.scannel@aalto.fi riccardo.mereu@aalto.fi