Sparse Function-space Representation of Neural Networks

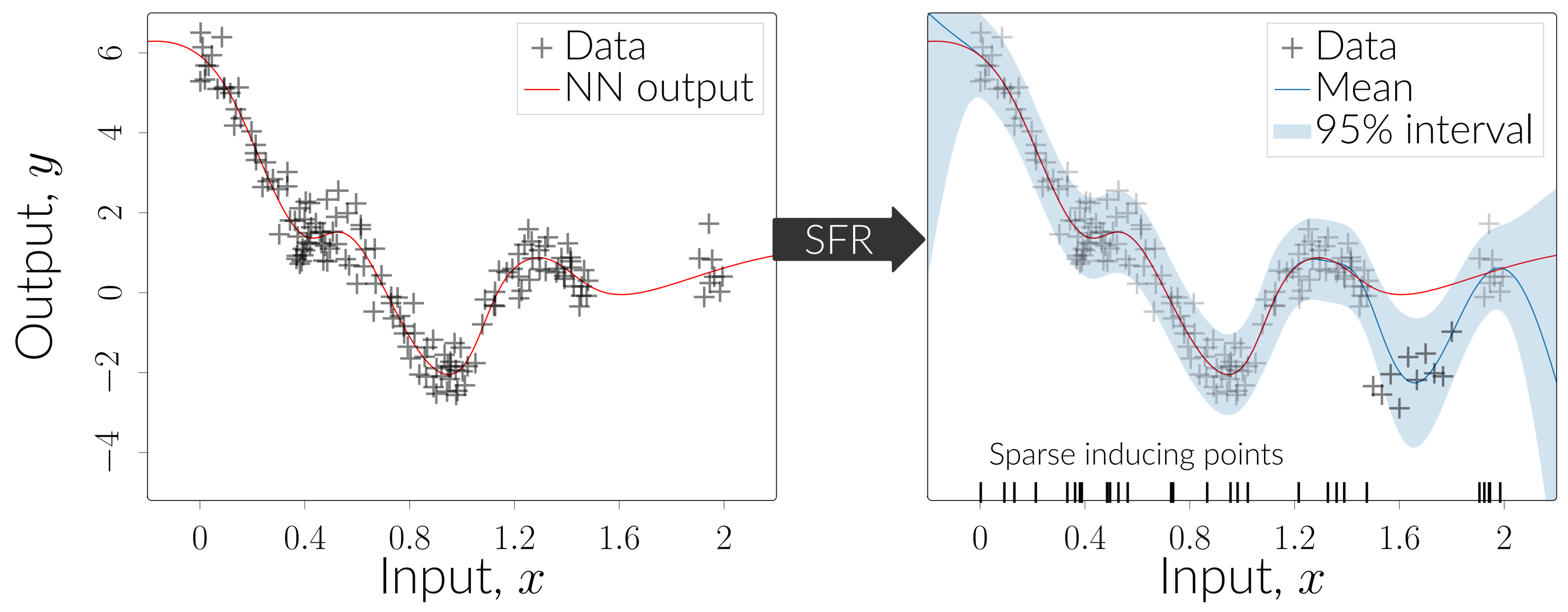
Aidan Scannell⋆ [1,2] Riccardo Mereu⋆ [1] Paul Chang [1]
Ella Tamir [1] Joni Pajarinen [1] Arno Solin [1]

[1]Aalto University [2]Finnish Center for Artificial Intelligence

Aalto University

Summary

Deep neural networks have limitations in *estimating uncertainty*, *incorporating new data*, and *avoiding catastrophic forgetting*. To overcome these issues, we introduce a method that converts neural networks from weight-space to a low-rank function-space representation using dual parameters. Unlike previous methods, our approach, named Sparse Function Representation (SFR), captures the full joint distribution of the entire data set, not just a subset. This allows for a concise and reliable way of capturing uncertainty and facilitates the inclusion of new data without the need for retraining. We provide a proof-of-concept quantifying uncertainty for supervised learning tasks on UCI benchmark data sets.



Regression w/ 2-layer MLP. Prediction from a trained neural network *(left)* and from our approach using inducing points to summarize the training data *(right)*. SFR captures the predictive mean and uncertainty, and can incorporate new data without retraining the model.

NN Function-space Representation

Inputs in a *supervised setting* for NNs $f_{\mathbf{w}} : \mathbb{R}^D \to \mathbb{R}^C$:

- $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, a data set w/ input $\mathbf{x}_i \in \mathbb{R}^D$ and output $\mathbf{y}_i \in \mathbb{R}^C$;
- $\mathbf{w} \in \mathbb{R}^P$, the initial weights of the neural network.

Goal: minimize the empirical (regularized) risk loss function:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathcal{D}, \mathbf{w}) = \arg\min_{\mathbf{w}} \textstyle\sum_{i=1}^N \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \delta \mathcal{R}(\mathbf{w}).$$

Output: $\mathbf{w}^*$, the Maximum A-Posteriori (MAP) weights of the NN.

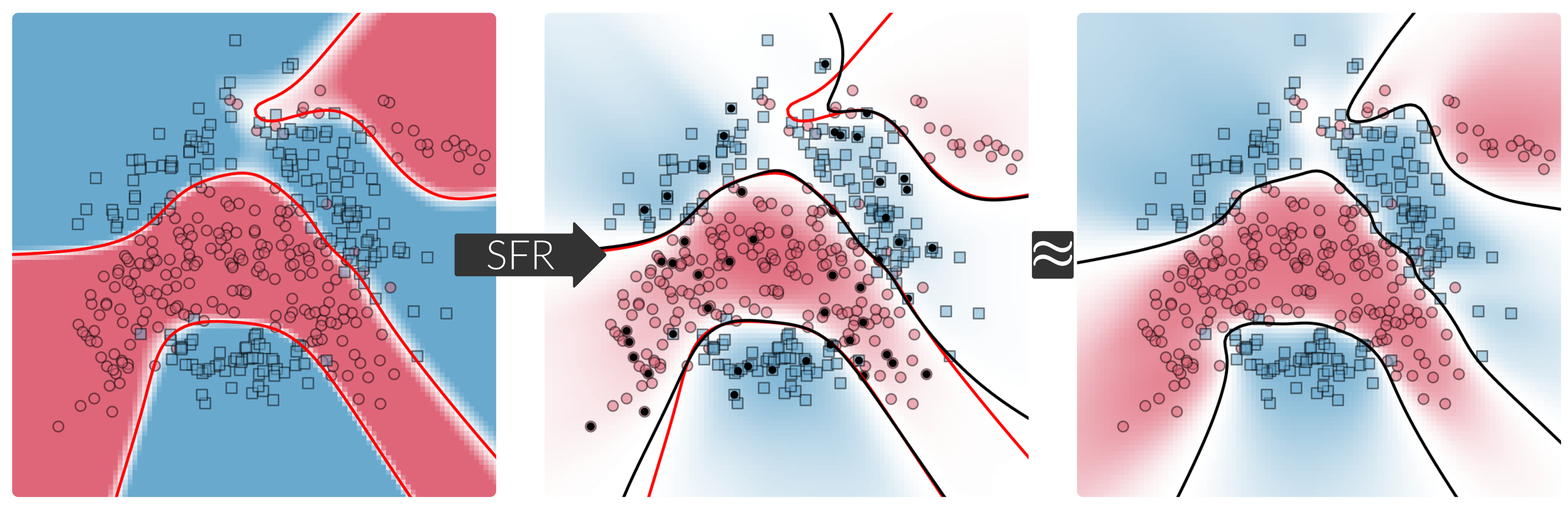
How to capture distribution over NN model functions?
Use their first two moments, obtaining a Gaussian process with a mean function $\mu(\cdot)$ and a covariance function $\kappa(\cdot,\cdot)$ (or kernel).
For GPs, linear approximations in weight space lead to function-space equivalent approximations:

$$f_{\mathbf{w}}(\mathbf{x}) \approx \phi^\top(\mathbf{x})\,\mathbf{w} \quad \Longrightarrow \quad \mu(\mathbf{x}) = 0 \quad \text{and} \quad \kappa(\mathbf{x}, \mathbf{x}') = \frac{1}{\delta}\phi^\top(\mathbf{x})\,\phi(\mathbf{x}')$$

For NNs, we can use the Laplace-GGN approximation to get a linear model of the neural network at the MAP as:

$$f_{\mathbf{w}^*}(\mathbf{x}) \approx \mathcal{J}_{\mathbf{w}_*}(\mathbf{x})\,\mathbf{w} \implies \mu(\mathbf{x}) = 0 \quad \text{and} \quad \kappa(\mathbf{x}, \mathbf{x}') = \frac{1}{\delta}\mathcal{J}_{\mathbf{w}^*}(\mathbf{x})\,\mathcal{J}_{\mathbf{w}^*}^\top(\mathbf{x}'),$$

where $\mathcal{J}_{\mathbf{w}}(\mathbf{x}) \coloneqq [\nabla_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{x})]^\top \in \mathbb{R}^{C \times P}$ is the Jacobian at $\mathbf{w}^*$.



Neural network prediction | Sparse function-space representation | HMC result as baseline

Uncertainty quantification for classification (■ vs. ●). We convert the trained neural network *(left)* to a sparse GP model with a set of inducing points ● *(middle)*. Results show a similar behaviour as running full Hamiltonian Monte Carlo (HMC) on the original NN model weights *(right)*. Marginal uncertainty depicted by colour intensity.

SFR: Sparse Function Representation

GP predictive posterior:

$$\mathbb{E}_{p(f_i \mid \mathbf{y})}[f_i] = \mathbf{k}_{\mathbf{x}i}^\top \boldsymbol{\alpha} \quad \text{and} \tag{1}$$

$$\mathrm{Var}_{p(f_i \mid \mathbf{y})}[f_i] = k_{ii} - \mathbf{k}_{\mathbf{x}i}^\top(\mathbf{K}_{\mathbf{xx}} + \mathrm{diag}(\boldsymbol{\beta})^{-1})^{-1}\mathbf{k}_{\mathbf{x}i} \tag{2}$$

with dual parameters $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^N, \boldsymbol{\beta} = \{\beta_i\}_{i=1}^N$:

$$\alpha_i \coloneqq \mathbb{E}_{p(\mathbf{w} \mid \mathbf{y})}[\nabla_f \log p(y_i \mid f)|_{f=f_i}] \quad \text{and} \tag{3}$$

$$\beta_i \coloneqq -\mathbb{E}_{p(\mathbf{w} \mid \mathbf{y})}[\nabla^2_{ff} \log p(y_i \mid f_i)|_{f=f_i}] \tag{4}$$

We consider the MAP of $p(\mathbf{w} \mid \mathbf{y})$, and get $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}} \in \mathbb{R}^N$:

$$\alpha_i \approx \hat{\alpha}_i \coloneqq \nabla_f \log p(y_i \mid f)|_{f=f_i} \quad \text{and} \tag{5}$$

$$\beta_i \approx \hat{\beta}_i \coloneqq -\nabla^2_{ff} \log p(y_i \mid f)|_{f=f_i} \tag{6}$$

Cons: requires access to all data $\to \mathcal{O}(N^3)$

Sparse GP (SFR) predictive posterior:

$$\mathbb{E}_{p(f_i \mid \mathbf{y})}[f_i] \approx \mathbb{E}_{q_{\mathbf{u}}(\mathbf{f})}[f_i] = \mathbf{k}_{\mathbf{z}i}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \boldsymbol{\alpha}_{\mathbf{u}} \quad \text{and} \tag{7}$$

$$\mathrm{Var}_{p(f_i \mid \mathbf{y})}[f_i] \approx \mathrm{Var}_{q_{\mathbf{u}}(\mathbf{f})}[f_i] = k_{ii} - \mathbf{k}_{\mathbf{z}i}^\top[\mathbf{K}_{\mathbf{zz}}^{-1} - (\mathbf{K}_{\mathbf{zz}} + \boldsymbol{B}_{\mathbf{u}})^{-1}]\mathbf{k}_{\mathbf{z}i} \tag{8}$$

given inducing points $\mathbf{u}_j = f_{\mathbf{w}^*}(\mathbf{z}_j)$ with $\{\mathbf{z}_j\}_{j=1}^M \to \mathcal{O}(M^3)$ with $(M \ll N)$

with SFR dual parameters $\boldsymbol{\alpha}_{\mathbf{u}} \in \mathbb{R}^M, \boldsymbol{B}_{\mathbf{u}} \in \mathbb{R}^{M \times M}$:

$$\boldsymbol{\alpha}_{\mathbf{u}} = \sum_{i=1}^N \mathbf{k}_{\mathbf{z}i}\,\hat{\alpha}_i \quad \text{and} \quad \boldsymbol{B}_{\mathbf{u}} = \sum_{i=1}^N \mathbf{k}_{\mathbf{z}i}\,\hat{\beta}_i\,\mathbf{k}_{\mathbf{z}i}^\top \tag{9}$$

Results on UCI datasets

	NN MAP	BNN	GLM	SFR (GP)	Full GP	Ablations (M=32): GP Subset (GP)	GP Subset (NN)	SFR (GP)	SFR (NN)
AUSTRALIAN	**0.31**±.01	0.42±.00	**0.32**±.02	**0.32**±.03	**0.32**±.03	0.51±.01	**0.33**±.02	**0.33**±.03	**0.32**±.03
CANCER	**0.11**±.02	0.19±.00	**0.10**±.01	**0.11**±.03	**0.11**±.03	0.41±.02	**0.11**±.03	**0.11**±.03	**0.10**±.04
IONOSPHERE	0.35±.02	0.50±.00	**0.29**±.01	0.34±.04	**0.34**±.04	0.54±.02	**0.34**±.05	0.34±.04	**0.30**±.06
GLASS	0.95±.03	1.41±.00	**0.86**±.01	0.93±.08	**0.93**±.08	1.15±.05	0.99±.07	0.95±.08	**0.87**±.07
VEHICLE	**0.42**±.01	0.89±.00	0.43±.01	0.48±.03	**0.48**±.03	1.02±.03	0.58±.02	0.54±.02	**0.49**±.02
WAVEFORM	**0.34**±.00	0.52±.00	0.34±.00	**0.35**±.02	**0.35**±.02	0.57±.02	**0.36**±.02	**0.35**±.02	**0.34**±.03
DIGITS	**0.09**±.00	0.88±.00	0.25±.00	0.37±.02	**0.08**±.03	1.65±.04	0.43±.01	0.43±.01	0.32±.02
SATELLITE	**0.23**±.00	0.48±.00	0.24±.00	0.29±.01	**0.27**±.01	1.12±.04	0.36±.01	0.33±.01	**0.28**±.01

References

[1] V. Adam, P. Chang, M. E. E. Khan, and A. Solin, "Dual parameterization of sparse variational Gaussian processes," in *Advances in Neural Information Processing Systems 34 (NeurIPS)*.

[2] A. Immer, M. Korzepa, and M. Bauer, "Improving predictions of Bayesian neural nets via local linearization," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

aidan.scannell@aalto.fi