MARS: Meta-learning as score matching in the function space

Krunoslav Lehman Pavasovic*, **Jonas Rothfuss***, **Andreas Krause**
ETH Zürich, Switzerland
{klehman, rojonas, krausea}@ethz.ch

Abstract

We approach meta-learning through the lens of functional Bayesian neural network inference, which views the prior as a stochastic process and performs inference in the function space. Specifically, we view the meta-training tasks as samples from the data-generating process and formalize meta-learning as empirically estimating the law of this stochastic process. Our approach can seamlessly acquire and represent complex prior knowledge by meta-learning the score function of the data-generating process marginals. In a comprehensive benchmark, we demonstrate that our method achieves state-of-the-art performance in terms of predictive accuracy and substantial improvements in the quality of uncertainty estimates.

1 Introduction

Meta-learning attempts to extract prior knowledge about the unknown data generation process from related tasks and embed it into the learner (Thrun & Pratt, 1998).We take a new approach to formulating the meta-learning problem and represent prior knowledge in a novel way by building on recent advances in functional BNNs (Wang et al., 2018; Sun et al., 2019). When viewing the BNN prior and posterior as stochastic processes, the perfect Bayesian prior is the (true) data-generating stochastic process itself. Hence, we view the meta-training datasets as samples from the meta data-generating process and interpret meta-learning as empirically estimating the law of this *stochastic process*. More specifically, we *meta-learn* the *score function* of its *marginal distributions*, which can then directly be used as a source of prior knowledge when performing approximate functional BNN inference on a new target task. This allows us to use flexible neural network models for learning the score and overcome the issues of meta-learning BNN priors in the parameter space.

In our experiments, we demonstrate that our proposed approach, *Meta-learning via Attention-based Regularised Score estimation (MARS)*, consistently outperforms previous meta-learners in predictive accuracy and yields significant improvements in the quality of uncertainty estimates. This promises fruitful future applications to domains like molecular biology or medicine, where meta-training data is scarce, and reasoning about epistemic uncertainty is crucial.

2 Background

Bayesian Neural Networks. Consider a regression task with data $\mathcal{D} = \left(\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}\right)$ that consists of m i.i.d. noisy function evaluations $y_j = f\left(\mathbf{x}_j\right) + \epsilon_j$ of an unknown function $f : \mathcal{X} \mapsto \mathcal{Y}$. Here, $\mathbf{X}^{\mathcal{D}} = \{\mathbf{x}_j\}_{j=1}^m \in \mathcal{X}^m$ denotes training inputs and $\mathbf{y}^{\mathcal{D}} = \{y_j\}_{j=1}^m \in \mathcal{Y}^m$ the corresponding noisy function values. Let $h_\theta : \mathcal{X} \to \mathcal{Y}$ be a regression function parametrized by a NN with weights $\theta \in \Theta$. Given a prior distribution $p(\theta)$ over the model parameters θ and a likelihood $p(y|\mathbf{x}, \theta)$, Bayes' theorem yields a posterior distribution $p(\theta|\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) \propto p(\mathbf{y}^{\mathcal{D}}|\mathbf{X}^{\mathcal{D}}, \theta)p(\theta)$.

BNN inference in the function space. BNN inference is difficult due to the high-dimensional parameter space Θ and the over-parameterized nature of the NN mapping $h_\theta(x)$. An alternative

*Equal contribution.

6th Workshop on Meta-Learning at NeurIPS 2022, New Orleans.

approach views BNN inference in the function space, i.e., the space of functions $h : \mathcal{X} \mapsto \mathcal{Y}$, yielding the Bayes rule[2] $p(h|\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) \propto p(\mathbf{y}^{\mathcal{D}}|\mathbf{X}^{\mathcal{D}}, h)p(h)$ (Wang et al., 2019; Sun et al., 2019). Given finite measurement sets $\mathbf{X} := [\mathbf{x}_1, ..., \mathbf{x}_k] \in \mathcal{X}^k, k \in \mathbb{N}$, we can characterize the stochastic process by its marginal distributions of function values $\rho(\mathbf{h}^{\mathbf{X}}) := \rho(h(\mathbf{x}_1), ...h(\mathbf{x}_k))$ (Oksendal, 2013).

3 Meta-Learning as Score Estimation in the Function Space

3.1 Problem Statement: Meta-Learning

The meta-learner is given n datasets $\mathcal{D}_1, ..., \mathcal{D}_n$, where each dataset $\mathcal{D}_i = (\mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})$ consists of m_i noisy function evaluations $y_{i,t} = f_i(\mathbf{x}_{i,t}) + \epsilon$ corresponding to a function $f_i : \mathcal{X} \mapsto \mathcal{Y} \subseteq \mathbb{R}$ and additive σ sub-Gaussian noise ϵ. In short, we write $\mathbf{X}_i^{\mathcal{D}} = (\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,m_i})^\top$ for the matrix of function inputs and $\mathbf{y}_i^{\mathcal{D}} = (y_{i,1}, ..., y_{i,m_i})^\top$ for the vector of corresponding observations. We assume that the functions $f_i \sim \mathcal{T}$ are sampled i.i.d. from a task distribution $\mathcal{T}$, which can be thought of as a stochastic process $p(f)$ that governs the random function $f : \mathcal{X} \mapsto \mathcal{Y}$. Our goal is to *extract knowledge from the observed datasets, which can then be used as a form of prior for learning a new, unknown target function* $f^* \sim \mathcal{T}$ from a corresponding dataset $\mathcal{D}^*$. By performing Bayesian inference with a meta-learned prior that is attuned to the task distribution $\mathcal{T}$, we hope to obtain posterior predictions that generalize better.

3.2 Meta-Learning as Score Estimation on the Data-Generating Process

We view acquiring prior knowledge in a data-driven manner from a new perspective. Our novel approach to meta-learning hinges upon three key ideas:

First, we view BNN inference in the function space (see Sec. 2), i.e., as posterior inference $p(h|\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) \propto p(\mathbf{y}^{\mathcal{D}}|\mathbf{X}^{\mathcal{D}}, h)p(h)$ over neural network mappings $h_\theta : \mathcal{X} \mapsto \mathcal{Y}$ instead of parameters θ. From this viewpoint, the prior, which is the target of our meta-learning problem, is $p(h)$ a stochastic process. Second, from a Bayesian perspective, the best possible prior is the stochastic process of the task generating distribution $\mathcal{T}$ itself, i.e., $p(h) = p(f)$. Hence, we aim to meta-learn a prior that matches $p(f)$ as closely as possible. We treat the meta-training datasets $\mathcal{D}_1, ..., \mathcal{D}_n$ as noisy observations from $p(f)$ and use them to estimate the stochastic process marginals $p(\mathbf{f}^{\mathbf{X}})$. Third, we estimate the score of the marginals rather than the marginal distribution itself. Since the score does not have to integrate to 1, it allows for much more flexible neural network representations.

In summary, our high-level approach is to *meta-learn / estimate the prior score* $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$ *that matches the data-generating stochastic process of the meta-training data* $\mathcal{D}_1, ..., \mathcal{D}_n$. At meta-test time, the meta-learned prior marginals are used in the approximate functional BNN inference on a target dataset $\mathcal{D}^*$, to infuse the acquired prior knowledge into the posterior predictions.

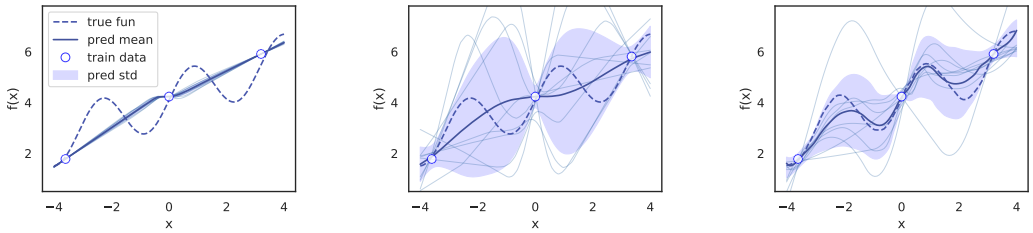
4 The MARS meta-learning algorithm

In the following, we discuss how to solve various challenges towards implementing our approach from Sec.3 and develop a practical algorithm for estimating the stochastic prior scores.

Parametric Score Matching for Stochastic Process Marginals. Performing functional approximate posterior inference with fSVGD updates as in (4) requires estimation of the prior marginal scores $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$ for arbitrary measurement sets $\mathbf{X}$. As the parametric model of the prior marginal scores, we use a transformer encoder architecture (Vaswani et al., 2017), takes as input a concatenation of measurement set $\mathbf{X} \in \mathbb{R}^{\dim(\mathcal{X}) \times k}$ of k points and corresponding query function values $\mathbf{h}^{\mathbf{X}} \in \mathbb{R}^{\dim(\mathcal{Y}) \times k}$. The network performs attention over the second dimension, i.e., the k columns corresponding to measurement points. Our score network model is denoted as $\mathbf{s}_\phi(\mathbf{h}^{\mathbf{X}}, \mathbf{X})$ with trainable parameters ϕ. The model is permutation equivariant w.r.t. the columns that correspond to measurement points and, thus, reflects a key property of stochastic process marginal scores.

We train the score network with a modified version of the score matching loss (Hyvärinen & Dayan, 2005) which is based on samples $\mathbf{f}_1^{\mathbf{X}}, ..., \mathbf{f}_n^{\mathbf{X}}$ from the data-generating process marginals. In contrast to the normal score matching loss, our modified loss takes the dependence of marginals on their

[2]Here, $p(h)$ is a stochastic process prior.



(a) Vanilla BNN (b) fBNN with GP prior (c) MARS (ours)

Figure 1: BNN posterior predictions with the corresponding prior distributions/processes.

measurements sets into account and uses an expectation over random measurement sets:

$$\mathcal{L}(\phi) := \mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\text{tr}(\nabla_{\mathbf{f}_i^{\mathbf{X}}}\mathbf{s}_\phi(\mathbf{f}_i^{\mathbf{X}},\mathbf{X})) + \frac{1}{2}||\mathbf{s}_\phi(\mathbf{f}_i^{\mathbf{X}},\mathbf{X})||_2^2\right)\right] \tag{1}$$

Interpolating the Datasets across $\mathcal{X}$. In our meta-learning setup (see Section 3.1), we only receive noisy function evaluations $\mathbf{y}_i^{\mathcal{D}}$ in m_i locations $\mathbf{X}_i^{\mathcal{D}}$ for each meta-training task. However, to minimize the score matching loss in (1), we need the corresponding function values in arbitrary measurement locations of the domain $\mathcal{X}$. To solve this issue, we fit a Gaussian Process (GP) on each dataset $\mathcal{D}_i$ which gives us a GP posterior $p(\mathbf{f}_i^{\mathbf{X}}|\mathbf{X},\mathbf{X}_i^{\mathcal{D}},\mathbf{y}_i^{\mathcal{D}})$ over function values $\mathbf{f}_i^{\mathbf{X}}$ for arbitrary measurement sets $\mathbf{X}$. To account for the epistemic uncertainty of the GP interpolation, for each task, we sample function values $\tilde{\mathbf{f}}_i^{\mathbf{X}} \sim p(\mathbf{f}_i^{\mathbf{X}}|\mathbf{X},\mathbf{X}_i^{\mathcal{D}},\mathbf{y}_i^{\mathcal{D}})$ from the corresponding GP posterior and use them as an input to the score matching loss in (1). Correspondingly, the score matching loss modifies to

$$\tilde{\mathcal{L}}(\phi) := \mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(\tilde{\mathbf{f}}_i^{\mathbf{X}}|\mathbf{X},\mathbf{X}_i^{\mathcal{D}},\mathbf{y}_i^{\mathcal{D}})}\left[\text{tr}(\nabla_{\tilde{\mathbf{f}}_i^{\mathbf{X}}}\mathbf{s}_\phi(\tilde{\mathbf{f}}_i^{\mathbf{X}},\mathbf{X})) + \frac{1}{2}||\mathbf{s}_\phi(\tilde{\mathbf{f}}_i^{\mathbf{X}},\mathbf{X})||_2^2\right]\right] . \tag{2}$$

Preventing Meta-Overfitting of the Prior Score Network. We only have meta-training data $\{\mathcal{D}_i\}_{i=1}^n$, corresponding to n functions f_i drawn from the data-generating stochastic process, i.e., we only have n samples for estimation of marginal scores $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$. This makes us prone to overfitting of the prior score network. To counteract the tendency to overfit, we regularize the score network via spectral normalization (Miyato et al., 2018) of the linear layers in the transformer encoder blocks. This biases the score estimates towards higher entropy.

The full MARS algorithm. Overall, the MARS meta-learning algorithm consists of two stages:

Stage 1: Meta-Learning the Prior Score Network. We fit a GP to each of the n meta-training datasets $\mathcal{D}_i$. Then, we train the score network by stochastic gradient descent on the modified score matching loss $\tilde{\mathcal{L}}(\phi)$ in (2). In each iteration, we first sample a measurement set $\mathbf{X}$ i.i.d. from the measurement distribution $\nu = \mathcal{U}(\tilde{\mathcal{X}})$, with $\tilde{\mathcal{X}} \subseteq \mathcal{X}$. For the measurement set, we sample a vector of functions values $\tilde{\mathbf{f}}_i^{\mathbf{X}}$ from the corresponding GP posterior marginals $p(\mathbf{f}_i^{\mathbf{X}}|\mathbf{X},\mathbf{X}_i^{\mathcal{D}},\mathbf{y}_i^{\mathcal{D}}), i = 1,...,n$. Based on these samples, we form an empirical estimate $\hat{\mathcal{L}}(\phi)$ of the score matching loss and perform a gradient update step on ϕ. This is repeated till convergence and summarized in Alg. 1.

Stage 2: Functional BNN Inference with the Prior Score Network. When concerned with a target learning task with training dataset $\mathcal{D}^* = (\mathbf{X}_*^{\mathcal{D}},\mathbf{y}_*^{\mathcal{D}})$, we infuse the meta-learned inductive bias into the functional BNN inference by using the score network $\mathbf{s}_\phi(\mathbf{h}^{\mathbf{X}},\mathbf{X})$ predictions as a swap-in replacement for the marginal scores $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$. In our experiments, we use fSVGD (see Section 2) as a functional approximate inference method (see Alg. 2 in Appx. A).

5 Experiments

We provide a detailed benchmark comparison, demonstrating that MARS: *(i)* achieves state-of-the-art performance in terms of predictive accuracy, *(ii)* yields well-calibrated uncertainty estimates.

5.1 Experiment Setup

Baselines. We consider five realistic meta-learning *regression* environments with detailed descriptions provided in Appendix B.2. As non-meta-learning baselines, we use a *Vanilla GP* with RBF

	SwissFEL	Physionet-GCS	Physionet-HCT	Berkeley-Sensor	Argus-Control
Vanilla GP	0.876 ± 0.000	2.240 ± 0.000	2.768 ± 0.000	0.258 ± 0.000	0.026 ± 0.000
Vanilla BNN	0.529 ± 0.022	2.664 ± 0.274	3.938 ± 0.869	0.151 ± 0.018	0.016 ± 0.002
MAML	0.730 ± 0.057	1.895 ± 0.141	$\mathbf{2.413 \pm 0.113}$	$\mathbf{0.121 \pm 0.027}$	0.017 ± 0.001
BMAML	0.577 ± 0.044	1.894 ± 0.062	2.500 ± 0.002	0.222 ± 0.032	0.037 ± 0.003
NP	0.471 ± 0.053	2.056 ± 0.209	2.594 ± 0.107	0.173 ± 0.018	0.020 ± 0.001
MLAP	0.486 ± 0.026	2.009 ± 0.248	2.470 ± 0.039	0.348 ± 0.034	0.030 ± 0.003
PACOH-NN	0.437 ± 0.021	1.623 ± 0.057	2.405 ± 0.017	0.160 ± 0.070	0.018 ± 0.002
MARS (ours)	$\mathbf{0.391 \pm 0.011}$	$\mathbf{1.471 \pm 0.083}$	$\mathbf{2.309 \pm 0.041}$	$\mathbf{0.116 \pm 0.024}$	$\mathbf{0.013 \pm 0.001}$

Table 1: Meta-Learning benchmark results in terms of the test RMSE.

	SwissFEL	Physionet-GCS	Physionet-HCT	Berkeley-Sensor	Argus-Control
Vanilla GP	0.135 ± 0.000	0.268 ± 0.000	0.277 ± 0.000	0.119 ± 0.000	0.090 ± 0.000
Vanilla BNN	0.085 ± 0.008	0.277 ± 0.013	0.307 ± 0.009	0.206 ± 0.025	0.104 ± 0.005
BMAML	0.115 ± 0.036	0.279 ± 0.010	0.423 ± 0.106	0.154 ± 0.021	0.068 ± 0.005
NP	0.131 ± 0.056	0.299 ± 0.012	0.319 ± 0.004	0.140 ± 0.035	0.094 ± 0.015
MLAP	0.090 ± 0.021	0.343 ± 0.017	0.344 ± 0.016	0.183 ± 0.017	0.168 ± 0.009
PACOH-NN	$\mathbf{0.037 \pm 0.005}$	$\mathbf{0.267 \pm 0.005}$	0.302 ± 0.003	0.223 ± 0.012	0.119 ± 0.005
MARS (ours)	$\mathbf{0.035 \pm 0.002}$	$\mathbf{0.263 \pm 0.001}$	$\mathbf{0.136 \pm 0.007}$	$\mathbf{0.080 \pm 0.005}$	$\mathbf{0.055 \pm 0.002}$

Table 2: Meta-learning benchmark results in terms of the calibration error.

kernel and a *Vanilla BNN* with a zero-centered, spherical Gaussian prior and SVGD posterior inference (Liu & Wang, 2016). For meta-learning baselines, we compare to model agnostic meta-learning (*MAML*) (Finn et al., 2017), Bayesian MAML (*BMAML*) (Yoon et al., 2018), neural processes (*NPs*) (Garnelo et al., 2018), *MLAP* (Amit & Meir, 2018) and *PACOH-NN* (Rothfuss et al., 2021a).

5.2 Empirical Benchmark Study

Qualitative illustration. Fig. 1 illustrates the superiority of MARS in (c) to posterior predictions of a) a BNN with standard Gaussian Prior in the parameter space b) a functional BNN with GP prior (Wang et al., 2019). For the meta-training, we use $n = 20$ tasks with each $m = 8$ data points, generated from a student-t process sinusoidal mean function $2x + 5\sin(2x)$.

MARS provides accurate predictions. We perform a comprehensive benchmark study with the baselines introduced in Sec. 5.1. Table 1 shows that MARS yields substantial RMSE improvements on the unseen meta-data when compared to the non-meta-learning baselines and significantly outperforms all other meta-learning baselines in the majority of the environments.

MARS yields well-calibrated uncertainty estimates. In order to compare the quality of uncertainty estimates of MARS to other meta-learning methods, we compute the *calibration error*, which measures how much the predicted confidence regions deviate from the actual frequencies of test data in the respective regions. The results are listed in Table 2, and the metric is detailed Appendix B.4. We can see that MARS consistently yields the best-calibrated uncertainty estimates.

6 Conclusion

We have introduced a novel meta-learning approach in the function space that estimates the score of the data-generating process marginals from a set of related datasets. When we face a new learning task, the meta-learned score network is used as a source of prior for functional approximate BNN inference. By representing inductive bias as the score of the stochastic process, our approach is versatile and can seamlessly acquire/represent complex prior knowledge. Empirically, this translates into strong performance when compared to previous meta-learning methods. The substantial improvements of MARS in terms of the quality of uncertainty estimates open up many potential extensions towards interactive machine learning where exploration based on epistemic uncertainty is key.

7 Acknowledgements

This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 815943 and the Swiss National

Science Foundation under NCCR Automation, grant agreement 51NF40 180545. Jonas Rothfuss was supported by an Apple Scholars in AI/ML fellowship. We thank Alex Hägele, Parnian Kassraie, Lars Lorch and Danica J. Sutherland for their valuable feedback.

References

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pp. 205–214. PMLR, 2018.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 2000.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69 (7-9):714–720, 2006.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2016.

Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. The born supremacy: quantum advantage and training of an ising born machine. *npj Quantum Information*, 6(1):1–11, 2020.

Charles-Alban Deledalle, Samuel Vaiter, Jalal M. Fadili, and Gabriel Peyré. Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection, 2014. URL `https://arxiv.org/abs/1405.1164`.

Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. Tensorflow distributions, 2017. URL `https://arxiv.org/abs/1711.10604`.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Vincent Fortuin and Gunnar Rätsch. Deep mean functions for meta-learning in gaussian processes. *arXiv preprint arXiv:1901.08098*, 2019.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020. URL `http://github.com/deepmind/dm-haiku`.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Johannes Kirschner, Mojmir Mutný, Nicole Hiller, Rasmus Ischebeck, and Andreas Krause. Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional Subspaces. In *International Conference on Machine Learning*, 2019a.

Johannes Kirschner, Manuel Nonnenmacher, Mojmir Mutnỳ, Andreas Krause, Nicole Hiller, Rasmus Ischebeck, and Andreas Adelmann. Bayesian optimisation for fast and safe parameter tuning of swissfel. In *FEL2019, Proceedings of the 39th International Free-Electron Laser Conference*, pp. 707–710. JACoW Publishing, 2019b.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*. PMLR, 2016.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. *arXiv preprint arXiv:2205.12934*, 2022.

Samuel Madden. Intel lab data. `http://db.csail.mit.edu/labdata/labdata.html`, 2004. Accessed: Sep 8, 2020.

Christopher J Milne, Thomas Schietinger, Masamitsu Aiba, Arturo Alarcon, Jürgen Alex, Alexander Anghel, Vladimir Arsov, Carl Beard, Paul Beaud, Simona Bettoni, et al. Swissfel: the swiss x-ray free electron laser. *Applied Sciences*, 7(7):720, 2017.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A Simple Neural Attentive Meta-Learner. In *International Conference on Learning Representations*, 7 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Tianyu Pang, Kun Xu, Chongxuan Li, Yang Song, Stefano Ermon, and Jun Zhu. Efficient learning of generative models via finite-difference score matching. *Advances in Neural Information Processing Systems*, 33:19175–19188, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

Thomas Pinder and Daniel Dodd. Gpjax: A gaussian process framework in jax. *Journal of Open Source misc*, 7(75):4455, 2022. doi: 10.21105/joss.04455. URL `https://doi.org/10.21105/joss.04455`.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Sachin Ravi and Alex Beatson. Amortized bayesian meta-learning. In *International Conference on Learning Representations*, 2018.

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. ProMP: Proximal Meta-Policy Search. In *International Conference on Learning Representations*, 2019.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pp. 9116–9126. PMLR, 2021a.

Jonas Rothfuss, Dominique Heyn, Jinfan Chen, and Andreas Krause. Meta-learning reliable priors in the function space. *Advances in Neural Information Processing Systems*, 34, 2021b.

Jonas Rothfuss, Christopher Koenig, Alisa Rupenyan, and Andreas Krause. Meta-learning priors for safe bayesian optimization. In *Conference on Robot Learning*, 2022.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, pp. 1842–1850, 2016.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. In *Artificial intelligence and statistics*, pp. 877–885. PMLR, 2014.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, pp. 4644–4653. PMLR, 2018.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology*, 2012.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020.

Ingo Steinwart. A sober look at neural network initializations. *arXiv preprint arXiv:1903.11482*, 2019.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger B. Grosse. Functional variational bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Danica J Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pp. 652–660. PMLR, 2018.

Sebastian Thrun and Lorien Pratt (eds.). *Learning to Learn*. Kluwer Academic Publishers, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

Dilin Wang, Zhe Zeng, and Qiang Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning*, pp. 5219–5227. PMLR, 2018.

Ziyu Wang, Tongzheng Ren, Jun Zhu, and Bo Zhang. Function space particle optimization for bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.

Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *International Conference on Machine Learning*, pp. 11513–11522. PMLR, 2020.

A Related Work

Meta-Learning. Common approaches in meta-learning amortize the entire inference process (Santoro et al., 2016; Mishra et al., 2018; Ravi & Beatson, 2018; Garnelo et al., 2018), learn a good neural network initialization (Finn et al., 2017; Rothfuss et al., 2019; Nichol et al., 2018; Kim et al., 2018) or a shared embedding space (Baxter, 2000; Vinyals et al., 2016; Snell et al., 2017). Although these approaches can meta-learn complex inference patterns, they require a large amount of meta-training data and often perform poorly in settings where data is scarce. Another line of work uses a hierarchical Bayesian approach to meta-learn priors over the NN parameters (Pentina & Lampert, 2014; Amit & Meir, 2018; Rothfuss et al., 2021a). Such methods perform much better on small data. However, they suffer from the lack of expressive families of priors for the high-dimensional and complex parameter space of NNs, making too restrictive assumptions to represent complex inductive biases. Our approach overcomes these issues by viewing the problem in the function space and directly learning the score, which can easily be represented by a NN instead of a prior distribution. Also related to our stochastic process approach are methods that meta-learn Gaussian Process (GP) priors (Fortuin & Rätsch, 2019; Rothfuss et al., 2021b, 2022). However, the GP assumption is quite limiting, while MARS can, in principle, match the marginals of any data-generating process.

Score estimation. We use score estimation as the central element of our meta-learning method. In particular, we use a parametric approach to score matching and employ an extended version of the score matching objective of Hyvärinen & Dayan (2005). For high-dimensional problems, Song et al. (2020); Pang et al. (2020) propose randomly sliced variations of the score matching loss. Alternatively, there is a body of work on nonparametric score estimation (Canu & Smola, 2006; Liu et al., 2016; Shi et al., 2018; Engl et al., 1996; Zhou et al., 2020). Among those, the Spectral Stein Gradient Estimator (Shi et al., 2018) has been used for estimating the stochastic process marginals for functional BNN inference in a setting where the stochastic prior is user-defined and allows for generating arbitrarily many samples (Sun et al., 2019). However, compared to our score network, it is harder to prevent meta-overfitting via regularization and to add an explicit dependence on the measurement sets in such nonparametric estimators, making them less suited to our problem.

This section focuses on a detailed explanation of the proposed framework. We first elaborate on the fSVGD inference process and then provide information regarding the network architecture and the hyperparameter configuration.

A.1 fSVGD inference

To perform approximate BNN inference with the meta-learned score network, we use *functional Stein Variational Gradient Descent (fSVGD)* (Wang et al., 2018). In the following, we explain our fSVGD implementation and how it interplays with the score network in our MARS framework.

In our context, the goal of fSVGD is to approximate the posterior $p(h_\theta|\mathbf{X}^\mathcal{D}, \mathbf{y}^\mathcal{D}) \propto p(\mathbf{y}^\mathcal{D}|\mathbf{X}^\mathcal{D}, h_\theta)p(h_\theta)$ over neural network mappings h_θ. Recall from Section 2, that $\mathbf{X}^\mathcal{D} = \{\mathbf{x}_j\}_{j=1}^m \in \mathcal{X}^m$ are the training inputs, $\mathbf{y}^\mathcal{D} = \{y_j\}_{j=1}^m \in \mathcal{Y}^m$ the corresponding noisy function values for a new, unseen learning tasks. The fSVGD algorithm approximates the posterior using a set of L NN parameter particles $\{\theta_1, ..., \theta_L\}$ where each θ_l corresponds to the weights and biases of a neural network. The particles (i.e., weights and biases) are initialized based on the scheme of Steinwart (2019) (see Appendix A.2 for details).

To make the BNN inference in the function space tractable, in each iteration, fSVGD samples a measurement set $\mathbf{X}$ from a measurement distribution ν and performs its particle updates based on the posterior marginals $p(\mathbf{h}^\mathbf{X}|\mathbf{X}, \mathbf{X}^\mathcal{D}, \mathbf{y}^\mathcal{D}) \propto p(\mathbf{y}^\mathcal{D}|\mathbf{h}^{\mathbf{X}^\mathcal{D}})p(\mathbf{h}^\mathbf{X})$ corresponding to $\mathbf{X}$. We use a uniform distribution over the bounded domain $\mathcal{X}$ as measurement distribution ν and sample the L measurement points i.i.d. from it.

For each NN particle, we compute the NN function values in the measurement points, i.e. $\mathbf{h}_{\theta^l}^\mathbf{X} = (h_\theta^l(\mathbf{x}_1), ..., h_\theta^l(\mathbf{x}_k)),\ l = 1, ..., L$ and the corresponding posterior marginal scores:

$$\nabla_{\mathbf{h}_{\theta^l}^\mathbf{X}} \ln p(\mathbf{h}_{\theta^l}^\mathbf{X}|\mathbf{X}, \mathbf{y}^\mathcal{D}) \leftarrow \underbrace{\nabla_{\mathbf{h}_{\theta^l}^\mathbf{X}} \ln p(\mathbf{y}^\mathcal{D}|\mathbf{h}_{\theta^l}^{\mathbf{X}^\mathcal{D}})}_{\text{likelihood score}} + \underbrace{\nabla_{\mathbf{h}_{\theta^l}^\mathbf{X}} \ln \hat{p}(\mathbf{h}_{\theta^l}^\mathbf{X})}_{:= \mathbf{s}_\phi(\mathbf{h}_{\theta^l}^\mathbf{X}, \mathbf{X})} \tag{3}$$

Algorithm 1 MARS: Meta-Learning the Data-Generating Process Score

Input: datasets $\mathcal{D}_1, ..., \mathcal{D}_n$, measurement point distribution ν, step size η
Initialize score network parameters ϕ
for $n = 1, ..., n$ **do**
  fit GP on $\mathcal{D}_i$, obtain GP posterior $p(f_i|\mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})$
while not converged **do**
  $\mathbf{X} \overset{iid}{\sim} \nu$ // sample measurement set
  for $i = 1, ..., n$ **do**
    $\tilde{\mathbf{f}}_i^{\mathbf{X}} \sim p(\mathbf{f}_i^{\mathbf{X}}|\mathbf{X}, \mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})$ // sample function values from posterior marginal
  $\hat{\mathcal{L}}(\phi) \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\text{tr}(\nabla_{\tilde{\mathbf{f}}_i^{\mathbf{X}}}\mathbf{s}_\phi(\tilde{\mathbf{f}}_i^{\mathbf{X}}, \mathbf{X})) + \frac{1}{2}||\mathbf{s}_\phi(\tilde{\mathbf{f}}_i^{\mathbf{X}}, \mathbf{X})||_2^2\right)$ // score matching loss
  $\phi \leftarrow \phi + \eta\nabla_\phi\hat{\mathcal{L}}(\phi)$ // score network gradient update
Output: trained score network $\mathbf{s}_\phi$

Here, we use a Gaussian likelihood $p(\mathbf{y}^{\mathcal{D}}|\mathbf{h}^{\mathbf{X}^{\mathcal{D}}}) = \prod_{j=1}^{k}\mathcal{N}(y_i^{\mathcal{D}}; h_\theta^l(x_j), \sigma^2)$ where σ^2 is the likelihood variance. Unlike in Wang et al. (2018) where the stochastic process prior is exogenously given, and its marginals $p(\mathbf{h}^{\mathbf{X}})$ are approximated as multivariate Gaussians, we use our meta-learned marginal prior scores. In particular, we use the score networks $\mathbf{s}_\phi(\mathbf{h}_{\theta^l}^{\mathbf{X}}, \mathbf{X})$ output as a swap-in for the prior score $\nabla_{\mathbf{h}_{\theta^l}^{\mathbf{X}}}\ln\hat{p}(\mathbf{h}_{\theta^l}^{\mathbf{X}})$. Finally, based on the score in (3) and the function values $\mathbf{h}_{\theta^l}^{\mathbf{X}}$, we can compute the SVGD updates in the function space, and project them back into the parameter space via the NN Jacobian $\nabla_{\theta^l}\mathbf{h}_{\theta^l}^{\mathbf{X}}$:

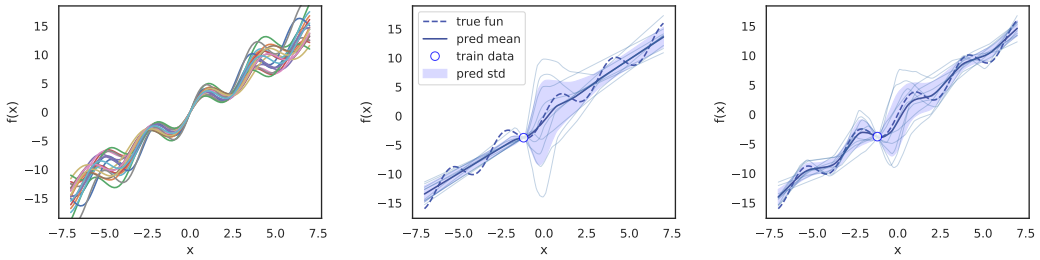
$$\theta^l \leftarrow \theta^l - \gamma\underbrace{\left(\nabla_{\theta^l}\mathbf{h}_{\theta^l}^{\mathbf{X}}\right)^\top}_{\text{NN Jacobian}}\underbrace{\left(\frac{1}{L}\sum_{i=1}^{L}\mathbf{K}_{li}\nabla_{\mathbf{h}_{\theta^i}^{\mathbf{X}}}\ln p(\mathbf{h}_{\theta^l}^{\mathbf{X}}|\mathbf{X}, \mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) + \nabla_{\mathbf{h}_{\theta^l}^{\mathbf{X}}}\mathbf{K}_{li}\right)}_{\text{SVGD update in the function space}}. \tag{4}$$

Here, γ is the SVGD step size and $\mathbf{K} = [k(\mathbf{h}_{\theta^l}^{\mathbf{X}}, \mathbf{h}_{\theta^i}^{\mathbf{X}})]_{li}$ is the kernel matrix between the function values in the measurement points based on a kernel function $k(\cdot,\cdot) : \mathcal{Y}^k \times \mathcal{Y}^k \mapsto \mathbb{R}$. We use the RBF kernel $k(\mathbf{h}, \mathbf{h}') = \exp\left((-||\mathbf{h} - \mathbf{h}'||^2)/(2\ell_k)\right)$ where ℓ_k is the bandwidth hyper-parameter. The particle update in (4) completes one iteration of fSVGD. We list the full procedure in Algorithm 2.

Algorithm 2 Approximate BNN Inference with fSVGD (Wang et al., 2019)

Input: SVGD kernel function $k(\cdot, \cdot\,; \ell_k)$, bandwidth ℓ_k, step size γ
Input: dataset $\mathcal{D}^* = (\mathbf{X}_*^{\mathcal{D}}, \mathbf{y}_*^{\mathcal{D}})$ for target task, trained score network $\mathbf{s}_\phi(\cdot,\cdot)$
Initialize set of BNN particles $\{\theta_1, ..., \theta_L\}$
while not converged **do**
  $\mathbf{X} \overset{iid}{\sim} \nu$ // sample measurement set
  for $l = 1, ..., L$ **do**
    $\mathbf{h}_{\theta^l}^{\mathbf{X}} \leftarrow (h_\theta^l(\mathbf{x}_1), ..., h_\theta^l(\mathbf{x}_k))$ where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_k)$ // compute NN function values in $\mathbf{X}$
    $\nabla_{\mathbf{h}_{\theta^l}^{\mathbf{X}}}\ln p(\mathbf{h}_{\theta^l}^{\mathbf{X}}|\mathbf{X}, \mathbf{y}^{\mathcal{D}}) \leftarrow \nabla_{\mathbf{h}_{\theta^l}^{\mathbf{X}}}\ln p(\mathbf{y}^{\mathcal{D}}|\mathbf{h}_{\theta^l}^{\mathbf{X}^{\mathcal{D}}}) + \mathbf{s}_\phi(\mathbf{h}_{\theta^l}^{\mathbf{X}}, \mathbf{X})$ // posterior marginal score
    $\theta^l \leftarrow \theta^l - \gamma\left(\nabla_{\theta^l}\mathbf{h}_{\theta^l}^{\mathbf{X}}\right)^\top\left(\frac{1}{L}\sum_{i=1}^{L}\mathbf{K}_{li}\nabla_{\mathbf{h}_{\theta^i}^{\mathbf{X}}}\ln p(\mathbf{h}_{\theta^l}^{\mathbf{X}}|\mathbf{X}, \mathbf{y}^{\mathcal{D}}) + \nabla_{\mathbf{h}_{\theta^l}^{\mathbf{X}}}\mathbf{K}_{li}\right)$ // fSVGD update
Output: Set of BNN particles $\{\theta_1, ..., \theta_L\}$ that approximate the BNN posterior process

fSVGD hyperparameter selection. For the fBNN training using fSVGD, among other parameters, we need to choose the step size γ, kernel bandwidth ℓ_k, and likelihood standard deviation σ, as we found these three to have the most substantial impact on training dynamics. We fix the number of particles to $L = 10$ and perform 10000 fSVGD update steps. We always standardize both the input and output data based on the meta-training data's empirical mean and standard deviation. For the NNs, we use three hidden layers, each of size 32 and with leaky ReLU activations. To initialize the NN weights, we use He initialization with a uniform distribution (He et al., 2015) and the bias initializer of Steinwart (2019) (see Section A.2 for details). Generally, we choose the kernel bandwidth ℓ_k via a random hyper-parameter search over the values range of $[0.1, 10]$. When



(a) Data generation process (b) Constant bias initializer (c) Steinwart bias initializer

Figure 2: a) The underlying data generating process and corresponding fBNN prediction after training using the fSVGD algorithm for 2000 iterations with b) constant bias initializer, and c) Steinwart bias initializer. With the Steinwart bias initializer, we get the desirable non-linear behavior of the fSVGD BNN much faster.

comparing to the original fSVGD implementation with SSGE, we fix the SSGE lengthscale to 0.2 for comparability. Note that we also experimented with using the median heuristic (as proposed by Shi et al. (2018)); however, we observed this to yield inferior performance.

A.2 Bias initialization

When initializing the biases to zeros or small positive constants, we find that the learned neural network maps behave like linear functions further away from zero and that it takes many SVGD iterations for them to assume non-linear behavior at the boundaries of the domain.

To address this issue, we use the bias initialization scheme of Steinwart (2019), which initializes the biases in such a way that the kinks of the leaky ReLU functions are more evenly distributed across the domain and less concentrated around zero. More specifically, we initialize each bias b_i as

$$b_i := - \langle w_i, x_i^\star \rangle \,,$$

where w_i is uniformly sampled from a sphere by taking $w_i = \frac{a_i}{||a_i||}$, $a_i \sim U(0,1)$. x_i^*'s are sampled uniformly: $x_i^* \sim U(min_x, max_x)$, where min_x and max_x are the minimum and maximum points in the input domain respectively.

By using *Steinwart* initialization, the neural networks show much more non-linear behavior after initialization. Compared to constant bias initialization, learning non-linear functional relationships happens much more quickly. We showcase this in Figure 2: Figure 2a displays a simple data generation process of sinusoids of varying amplitude, frequency, phase shift, and slope. Figure 2b displays corresponding BNN predictions after 2000 iterations with a constant bias initializer (where the constant equals 0.01), and Figure 2c shows the result of the same training dynamics, using *Steinwart's* bias initializer instead of the constant initializer. The BNNs with the *Steinwart* initialization assume desirable non-linear behavior much earlier during training, thus, significantly speeding up training. However, if trained for a large enough number of iterations, the performance of the two networks with different bias initializations becomes much less distinct. For further details on the implications of the *Steinwart* initializer on the training dynamics, we refer to Steinwart (2019).

A.3 Score estimation network

We now give an overview of the score estimation network. We start by providing motivation for the required architecture, detailing the permutation equivariance properties of the network. We construct the proposed architecture step-by-step, giving the architectural details and mentioning additional architectural designs we experimented with. Finally, we comment on the optimization method and acknowledge the libraries we used in our implementations.

Incorporating task invariances. The proposed network is permutation equivariant across the k dimension: reordering the measurement inputs would result in reordering the network's prediction in the same manner: Formally, for any permutation π of the measurement set indices we have

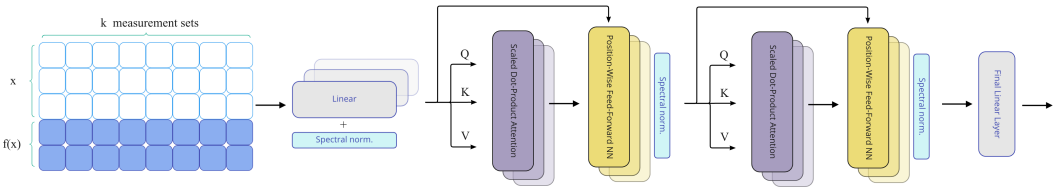


Figure 3: Architecture of MARS score estimation network. From left to right: k measurement sets consisting of input-output concatenations with $x_i \in \mathbb{R}^3, f(x_i) \in \mathbb{R}^2, i = 1, ..., k$; inputs/outputs embeddings using spectrally normalized linear layers; two identical blocks of scaled dot product attention, residual layers and feed-forward position-wise NN with spectrally normalized linear layers; the final linear layer, *not* spectrally normalized.

that $\mathbf{s}_\phi(\mathbf{h}^{\mathbf{X}_{\pi(1:d)}}, \mathbf{X}_{\pi(1:d)})_{i,j} \mapsto \tilde{\nabla}_{\mathbf{h}^{\mathbf{X}_{1:d}}} p(\mathbf{h}^{\mathbf{X}_{1:d}})_{\pi(i),\pi(j)}$. To impose permutation equivariance, we use the self-attention mechanism (Vaswani et al., 2017). Permutation equivariance is obtained by concatenating inputs (measurement points) and the corresponding functional evaluations (i.e., the concatenation constructs an object corresponding to Transformer *tokens*), which are then embedded and inputted to the attention mechanism. The architecture is displayed in Figure 3.

Constructing the network architecture. The core of our model is composed of *two* identical blocks, each consisting of *two* residual layers, the first one applying multi-head self-attention and the second one position-wise feed-forward neural network, similar to the vanilla Transformer encoder (Vaswani et al., 2017). Since the multi-head attention is permutation equivariant over the measurement point (i.e., token) dimension, the representation is permutation equivariant at all times (Lee et al., 2019). Finally, to minimize training time, we select attention embedding dimensions proportional to the data dimension of the environments; higher-dimensional environments (e.g., SwissFEL, detailed in Appendix B) correspond to the higher number of attention parameters/embedding dimensions and lower environments (e.g., the Sinusoid and Berkeley environment) to lower embeddings. Furthermore, we tune the step size and report the chosen configurations under `https://tinyurl.com/376wp8xe`. We describe the corresponding implementation details in the following subsection.

Constructing the score network. We train the score network on the input/output pair concatenations, which are then embedded onto a higher-dimensional space. After this, we perform self-attention. Specifically, we fix the initialization scale of self-attention weights to 2.0. As mentioned, the size of the model and embeddings varies across meta-learning environments. Next follows a position-wise feed-forward neural network, for which we use Exponential Linear-Unit (ELU) as the activation, as implemented in Haiku's vanilla Transformer encoder (Hennigan et al., 2020). Afterward, we apply the self-attention mechanism again, following a position-wise feed-forward neural network, after which we apply the final linear layer.

Variants of the attention mechanism. We experimented with multiple variations of the attention mechanism, all being permutation equivariant in the k dimension, similar to the work of Lorch et al. (2022). We mention two other architectural designs: in the first design, rather than using the embeddings of the concatenation of input-output points as keys, queries, and values (as performed in the vanilla Transformer encoder), we experimented with using the embeddings of the inputs as the keys and queries and the embeddings of functional outputs. However, this design yields varying performance: on several low-dimensional tasks, the performance was slightly better, whereas performance on tasks with higher-dimensional inputs was substantially worse. In another attempt, we experimented with learning different embeddings for inputs and functional outputs, which we then concatenated and used as keys, queries, and values. The difference in performance in this method was marginal, and we decided against it for simplicity.

Optimizer setup and employed libraries. Finally, across all experiments, we use gradient clipping of the prior score together with the ADAM optimizer (Kingma & Ba, 2014), with the default values set in Jax (Bradbury et al., 2018) and Haiku (Hennigan et al., 2020). For the gradient clipping, we use values of 1., 10., or 100., depending on the task and the underlying properties of the

	Sinusoid	SwissFEL	Physionet	Berkeley	Berkeley*	Argus-Control
n	20	5	100	10	36	20
m_i	8	200	4 - 24	30	288	500

Table 3: Number of tasks n and samples per task m_i for the different meta-learning environments.

data generating process. For example, for the first experiment in Appendix D, larger clipping values (50 or 100) performed better for the heavier-tailed Student't-*t* process, whereas clipping at 10. resulted in a good performance of the GP task. We train the score network for 20000 iterations. For the GP and Student-*t* process implementations, we use GP-Jax (Pinder & Dodd, 2022), scikit-learn (Pedregosa et al., 2011) and TensorFlow Distributions packages (Dillon et al., 2017).

A.4 Interpolating the datasets across $\mathcal{X}$ using GPs

In GP regression, each data point corresponds to a feature-target tuple $z_{i,j} = (\mathbf{x}_{i,j}, y_{i,j}) \in \mathbb{R}^d \times \mathbb{R}$. For the i-th dataset, we write $\mathcal{D}_i = (\mathbf{X}_i, \mathbf{y}_i)$, where $\mathbf{X}_i = (x_{i,1}, \ldots, x_{i,m_i})^\top$ and $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m_i})^\top$. GPs are a Bayesian method in which the prior $\mathcal{P}(h) = \mathcal{GP}\left(h \mid m(x), k\left(x, x'\right)\right)$ is specified by a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a mean function $m : \mathcal{X} \to \mathbb{R}$. In this section, we assume a zero mean GP and omit writing the dependence on $m(\cdot)$. As the GP kernel, we use the Matérn covariance function

$$k(\mathbf{x}, \mathbf{x}'; \ell, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \left\| \frac{\mathbf{x} - \mathbf{x}'}{\ell} \right\|_2 \right)^{\nu} \mathrm{K}_\nu \left(\sqrt{2\nu} \left\| \frac{\mathbf{x} - \mathbf{x}'}{\ell} \right\|_2 \right), \tag{5}$$

with degree ν, lengthscale ℓ, and $\Gamma(\cdot)$ representing the gamma function, and K_v the modified Bessel function of the second kind. We use fix the degree of the Matérn kernel to $\nu = 5/2$ and choose the lengthscale ℓ via 4-fold cross-validation (CV) from 10 log-uniformly spaced points in $[0.001, 10]$. We select the lengthscale that maximizes the 4-fold CV log marginal likelihood, averaged across the n tasks. In Alg. 3, we give the full procedure of selecting the lengthscale and fitting GPs to the meta-training tasks.

Algorithm 3 Fitting the GPs to the meta tasks

Input: n datasets $\{\mathcal{D}_i\}_{i=1}^n$, where $\mathcal{D}_i = \{(\mathbf{x}_{ij}, y_{ij})\}_{j=1}^{m_i}$
Input: set of p Matérn lengthscale value candidates $\mathcal{L} = \{\ell_j\}_{j=1}^p$
Input: zero-mean GP prior $\mathcal{P}(f) = \mathcal{GP}\left(f \mid k_\ell\left(\mathbf{x}, \mathbf{x}'\right)\right)$, specified by the Matérn kernel $k_\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
for $i = 1, ..., n$ **do**
  $\{\mathcal{GP}(f \mid k_{\ell_j})_i\}_{j=1}^p, \{\text{score}_{i,j}\}_{j=1}^p = CV_{\text{4-fold}}(\mathcal{D}_i, \mathcal{P})$ // cross-validation on $\mathcal{D}_i$
$j^* = \arg\max_{j=1,\ldots,p} \frac{1}{n} \sum_{i=1}^n \text{score}_{i,j}$ // selecting the optimal kernel parameters
Output: n GP posteriors $\{\mathcal{GP}(f \mid k_{(\ell,\nu)_{j^*}})\}_{i=1}^n$

B Meta-Learning Environments

In the following, we provide details on the meta-learning environments used in our experiments in Section 5. We list the number of tasks and samples in each environment in Table 3.

B.1 Sinusoids (Synthetic Environment)

The sinusoid environment corresponds to a simple 1-dimensional regression problem with a sinusoidal structure. It is used for visualization purposes in Figure 1 and Figure 2a. Each task of the sinusoid environment corresponds to a parametric function

$$f_{a,b,c,\beta}(x) = \beta * x + a * \sin(1.5 * (x - b)) + c\,, \tag{6}$$

yielding a sum of affine and a sinusoid function. Tasks differ in the function parameters (a, b, c, β) that are sampled from the task environment $\mathcal{T}$ as follows:

$$a \sim \mathcal{U}(0.7, 1.3), \quad b \sim \mathcal{N}(0, 0.1^2), \quad c \sim \mathcal{N}(5.0, 0.1^2), \quad \beta \sim \mathcal{N}(0.5, 0.2^2)\,. \tag{7}$$

Figure 2a displays functions $f_{a,b,c,\beta}$ with parameters sampled according to (7). To draw training samples from each task, we uniformly sample x from $\mathcal{U}(-5, 5)$ and add Gaussian noise with standard deviation 0.1 to the function values $f(x)$:

$$x \sim \mathcal{U}(-5, 5) \,, \qquad y \sim \mathcal{N}(f_{a,b,c,\beta}(x), 0.1^2) \,. \tag{8}$$

B.2 Real-World Data

B.2.1 SwissFEL

Free-electron lasers (FELs) accelerate electrons to a very high speed to generate shortly pulsed laser beams with wavelengths in the X-ray spectrum. The X-ray pulses from the accelerator can map nanometer-scale structures, thus facilitating molecular biology and material science experiments. The accelerator and the electron beam line of an FEL consist of multiple magnets and other adjustable components, which have several parameters that experts adjust in order to maximize the pulse energy (Kirschner et al., 2019a). Due to different operational modes, parameter drift, and changing (latent) conditions, the laser's pulse energy function, in response to its parameters, changes across time. As a result, optimizing the laser's parameters is a recurrent task.

Meta-learning setup. The meta-learning environment represents different parameter optimization runs (i.e., tasks) on SwissFEL, an 800-meter-long free electron laser located in Switzerland (Milne et al., 2017). The input space is 12-dimensional and corresponds to the laser parameters, whereas the regression target corresponds to the one-dimensional pulse energy. We refer to Kirschner et al. (2019b) for details on the individual parameters. Each optimization run consists of roughly 2000 data points generated with online optimization methods, yielding non-i.i.d. data, which becomes successively less diverse throughout the optimization. To avoid issues with highly dependent data points, we take the first 400 data points per run and split them into training and test subsets of size 200. As we have a total of 9 runs (tasks) available, we use 5 of them for meta-training and the remaining 4 for meta-testing.

B.2.2 PhysioNet

In the context of the Physionet competition 2012, Silva et al. (2012) have published an open-access dataset of patient stays in the intensive care unit (ICU). The dataset consists of measurements taken during the patient stays, where up to 37 clinical variables are measured over the span of 48 hours, yielding a time series of measurements. The intended task for the competition was the binary classification of patient mortality. However, the dataset is also often used for time series prediction methods due to a large number of missing values (around 80 % across all features).

Meta-learning setup. To set up the meta-learning environment, we treat each patient as a separate task and the different clinical variables as different environments. Out of the 37 variables, we picked the Glasgow coma scale (GCS) and hematocrit value (HCT) as environments for our study since they are among the most frequently measured variables in this dataset. From the dataset, we remove all patients where less than four measurements of CGS (and HCT, respectively) are available. From the remaining patients, we used 100 patients for meta-training and 500 patients for meta-validation and meta-testing. Since the number of available measurements differs across patients, the number of training points m_i ranges between 4 and 24.

B.2.3 Berkeley-Sensor

The Berkeley dataset contains data from 46 temperature sensors deployed in different locations at the Intel Research lab in Berkeley (Madden, 2004). The temperature measurements are taken over four days and sampled at 10-minute intervals. Each task corresponds to one of the 46 sensors and requires auto-regressive prediction, particularly predicting the subsequent temperature measurement given the last ten measurements.

Meta-learning setup. The Berkeley environment, as used in Rothfuss et al. (2021a), uses 36 sensors (tasks) with data for the first two days for meta-training and the last 10 for meta-testing. The meta-training and meta-testing are separated temporally and spatially since the data is non-i.i.d. Data are abundant, and the measurements are taken at very close intervals. Thus, the features and

	RMSE		Calib. error	
	Full dataset	Partial dataset	Full dataset	Partial dataset
Vanilla GP	0.276 ± 0.000	0.258 ± 0.000	$\mathbf{0.109 \pm 0.000}$	0.119 ± 0.000
Vanilla BNN	0.109 ± 0.004	0.151 ± 0.018	0.179 ± 0.002	0.206 ± 0.025
MAML	$\mathbf{0.045 \pm 0.003}$	$\mathbf{0.121 \pm 0.027}$	/	/
BMAML	0.073 ± 0.014	0.222 ± 0.032	0.161 ± 0.013	0.154 ± 0.021
NP	0.079 ± 0.014	0.173 ± 0.018	0.210 ± 0.000	0.140 ± 0.035
MLAP	0.050 ± 0.034	0.348 ± 0.034	$\mathbf{0.108 \pm 0.024}$	0.183 ± 0.017
PACOH-NN	0.130 ± 0.009	0.160 ± 0.070	0.167 ± 0.005	0.223 ± 0.012
MARS	0.093 ± 0.002	$\mathbf{0.116 \pm 0.024}$	0.140 ± 0.002	$\mathbf{0.080 \pm 0.005}$

Table 4: Prediction accuracy and uncertainty calibration on full and partial Berkeley-Sensor dataset. On the full dataset, MARS performance is less competitive due to the strong auto-correlation of the data which is not taken into account in the BNN likelihood. On the partial dataset, which has less dependency among the data points, MARS outperforms all other meta-learning methods.

the train/context data points are strongly correlated, violating the i.i.d. assumption that underlies our factorized Gaussian likelihood and Bayes rule in Section 2, causing the BNN to over-weight the empirical evidence and making over-confident predictions. To alleviate this problem, we subsample the data. In particular, we randomly select 10 out of the 36 training tasks, and instead of using all measurements, we randomly sample 30 of them. This has two effects: First, it makes the data less dependent/correlated and thus more compatible with our Bayesian formulation. Second, we increase the epistemic uncertainty by using less data, making the calibration metrics more meaningful. The results reported in Section 5 correspond to the sub-sampled data.

For completeness, we also report the results for the full dataset as in Rothfuss et al. (2021a) in Table 4. Other meta-learning baselines, such as MAML or MLAP, perform better than MARS on the full dataset since they do not explicitly use the Bayes rule with i.i.d. assumption or weight the empirical evidence less. Note that MARS performs worse due to the Bayesian inference at meta-test time rather than our meta-learning approach. Accounting for the strong auto-correlation of the data in the likelihood would most likely resolve the issue. On the sub-sampled environment, MARS again performs best.

B.2.4 Argus-Control

The final environment we use in our experiments is a robot case study. In particular, it aims at tuning the controller of the Argus linear motion system by Schneeberger Linear Technology. The goal is to choose the controller parameters so that the position error is minimal. In our setup, each task is a regression problem where the goal is to predict the total variation (TV) of the robot's position error signal when controlled by a PID controller in simulation. The regression features are the three PID controller gain parameter parameters.

Meta-learning setup. Overall, the environment consists of 24 tasks, of which 20 are used for meta-training and the remaining 4 for meta-testing. Each task corresponds to a different step size for the robot to move, ranging from $10\mu m$ to $10mm$. At different scales, the robot behaves differently in response to the controller parameters, resulting in different target functions. This presents a good environment for transferring similarities across different scales while leaving enough flexibility in the prior to adjust to the target function at a step size.

In the following, we describe our experimental methodology used in Section 5.

B.3 Overview of the meta-training and meta-testing phases

Evaluating the performance of a meta-learner consists of two phases, *meta-training* and *meta-testing*. The latter phase, *meta-testing*, can be further split into *target training* and *target testing*. In particular, for MARS the phases consist of the following:

- *Meta-training:* The meta-training datasets $\mathcal{D}_{i=1}^{n}$ are used to train the score estimator network (see Algorithm 1).

- *Target training:* Equipped with knowledge about the underlying data-generation process, i.e., the score network, we perform BNN inference on a new target task with a corresponding context dataset $\mathcal{D}^*$. In particular, we run fSVGD with the score network as a swap-in for the marginal scores of the stochastic process prior (see Algorithm 2). As a result, we obtain a set of NN particles that approximates the BNN posterior in the function space.
- *Target testing:* Finally, we evaluate the approximate posterior predictions on a test set $\mathcal{D}^\dagger$ corresponding to the same target task. In particular, we compute the residual mean squared error (RMSE) and the calibration error as performance metrics. We describe the evaluation metrics in more detail in Section B.4.

The target training and testing are performed independently with the meta-learned score network for each test task. The metrics are reported as averages over the test tasks.

The entire meta-training and meta-testing procedure are repeated for five random seeds that influence the score network initialization, the sampling-based estimates in Algorithm 1 as well as the initialization of the BNN particles for target training. The reported averages and standard deviations are based on the results obtained for different seeds.

B.4 Evaluation metrics

During *target-testing*, we evaluate the posterior predictions on a held-out test dataset $\mathcal{D}^\dagger$. Among the methods employed in Section 5, MARS, PACOH-NN, NPs, MLAP, Vanilla BNN and Vanilla GP yield probabilistic predictions $\hat{p}(y^\dagger|x^\dagger, \mathcal{D}^*)$ for the test points $x^\dagger \in \mathcal{D}^\dagger$. For instance, in the case of MARS, PACOH-NN, and Vanilla BNN where the posterior is approximated by a set of NN particles $\{\theta_1, ..., \theta_L\}$ and we use a Gaussian likelihood, the predictive distribution is an equally weighted mixture of Gaussians:

$$\hat{p}(y^\dagger|x^\dagger, \mathcal{D}^*) = \frac{1}{L}\sum_{l=1}^{L} \mathcal{N}(y^\dagger|h_{\theta_l}(x^\dagger), \sigma^2) \; . \tag{9}$$

The respective mean prediction corresponds to the expectation of $\hat{p}$, that is $\hat{y} = \hat{\mathbb{E}}(y^*|x^*, \mathcal{D}^*)$. In the case of MAML, only the mean prediction is available.

Evaluating prediction accuracy (RMSE) Based on the mean predictions, we compute the *root-mean-squared error (RMSE)*

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{D}^\dagger|}\sum_{(x^\dagger, y^\dagger)\in\mathcal{D}^\dagger} (y^\dagger - \hat{\mathbb{E}}(y^\dagger|x^\dagger, \mathcal{D}^*))^2} \tag{10}$$

which quantifies how accurate the mean predictions are.

Evaluating uncertainty calibration (Calibration error) In addition to the prediction accuracy, we also assess the quality of the uncertainty estimates. For this purpose, we use the concept of calibration, i.e., a probabilistic predictor is calibrated if the predicted probabilities are consistent with the observed frequencies on unseen test data. We use a regression calibration error similar to Kuleshov et al. (2018) in order to quantify how much the predicted probabilities deviate from the empirical frequencies.

Let us denote the predictor's cumulative density function (CDF) as $\hat{F}\left(y \mid \mathbf{x}, \mathcal{D}^*\right) = \int_{-\infty}^{y} \hat{p}(\tilde{y}|\mathbf{x}, \mathcal{D}^*)d\tilde{y}$, where $\hat{p}(\tilde{y}|\mathbf{x}, \mathcal{D}^*)d\tilde{y}$ is the predictive posterior distribution. For confidence levels $0 \leq q_h < \ldots < q_H \leq 1$, we compute the corresponding empirical frequency

$$\hat{q}_h = \frac{\left|\left\{y^\dagger \mid \hat{F}\left(y^\dagger \mid \mathbf{x}, \mathcal{D}^*\right) < q_h,\ (\mathbf{x}^\dagger, y^\dagger) \in \mathcal{D}^\dagger\right\}\right|}{|\mathcal{D}^\dagger|},$$

based on some test dataset $\mathcal{D}^\dagger$. If the predictions are calibrated, we expect that $\hat{q}_h \to q_h$ as $m \to \infty$. Following Kuleshov et al. (2018), we define the calibration error metric as a function of the residuals $\hat{q}_h - q_h$:

$$\text{calib-err} = \frac{1}{H}\sum_{h=1}^{H} |\hat{q}_h - q_h| \, .$$

Estimator	SwissFEL	Physionet-GCS	Physionet-HCT	Berkeley-Sensor	Argus-Control
MARS	$\mathbf{0.391 \pm 0.011}$	$\mathbf{1.471 \pm 0.083}$	$\mathbf{2.309 \pm 0.041}$	$\mathbf{0.116 \pm 0.024}$	$\mathbf{0.013 \pm 0.001}$
SSGE score estimates	0.449 ± 0.027	3.292 ± 0.562	2.784 ± 0.257	1.105 ± 0.562	0.030 ± 0.003
No spectral reg.	0.420 ± 0.060	2.208 ± 0.338	$\mathbf{2.560 \pm 0.341}$	1.734 ± 0.169	$\mathbf{0.014 \pm 0.001}$
No GP sampling	0.471 ± 0.059	2.994 ± 0.363	5.995 ± 1.108	1.253 ± 0.112	0.073 ± 0.003

Table 5: Ablation study results for MARS components in terms of the RMSE.

Estimator	SwissFEL	Physionet-GCS	Physionet-HCT	Berkeley-Sensor	Argus-Control
MARS	$\mathbf{0.035 \pm 0.002}$	0.263 ± 0.001	$\mathbf{0.136 \pm 0.007}$	$\mathbf{0.080 \pm 0.005}$	$\mathbf{0.055 \pm 0.002}$
SSGE score estimates	0.151 ± 0.001	0.249 ± 0.002	0.246 ± 0.007	0.232 ± 0.010	0.210 ± 0.011
No spectral reg.	0.233 ± 0.041	0.265 ± 0.012	0.244 ± 0.009	0.244 ± 0.009	0.187 ± 0.028
No GP sampling	0.204 ± 0.013	$\mathbf{0.225 \pm 0.021}$	0.237 ± 0.018	0.141 ± 0.029	0.216 ± 0.066

Table 6: Ablation study results for MARS components in terms of the calibration error.

Note that we report the average of absolute residuals $|\hat{q}_h - q_h|$, rather than reporting the average of squared residuals $|\hat{q}_h - q_h|^2$, as done by Kuleshov et al. (2018). This is done to preserve the units and keep the calibration error easier to interpret. In our experiments, we compute the empirical frequency using $M = 20$ equally spaced confidence levels between 0 and 1.

B.5 Hyper-Parameter Selection

For each of the meta-environments and algorithms, we ran a separate hyper-parameter search to select the hyper-parameters. In particular, we use 30 randomly sampled hyperparameter configurations across five randomly selected seeds and select the best-performing one in terms of either RMSE or calibration error. For the reported results, we provide the chosen hyperparameters and detailed evaluation results in `https://tinyurl.com/376wp8xe`.

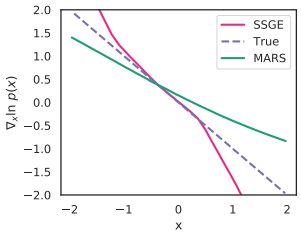
C Ablation Study

In this section, we empirically investigate the algorithm components that were introduced in Sec. 4 and provide supporting empirical evidence for our design decisions.

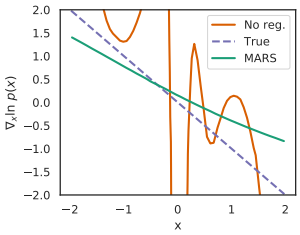
First, we perform a quantitative ablation study where we vary/remove components of our MARS algorithm. We consider MARS 1) with nonparametric score estimator SSGE (Shi et al., 2018) instead of the parametric score network + score matching from Sec. 4, 2) without spectral regularization, and 3) without GP posterior sampling, i.e., using the GP posterior mean instead of samples for the score matching. Table 5 and 6 report the quantitative results of this ablation experiment in terms of the RMSE and the calibration error. In the following, we discuss the three aspects separately:

Parametric score matching outperforms nonparametric score estimation. In Sec. 4 we have introduced a parametric score network which we train with the score matching loss in 1. Alternatively, for each measurement set $\mathbf{X}$, we could apply the nonparametric score estimator SSGE (Shi et al., 2018) to the function values $\tilde{\mathbf{f}}_1^{\mathbf{X}}, ..., \tilde{\mathbf{f}}_n^{\mathbf{X}}$, sampled from the GP posteriors, and directly use the resulting score estimates as prior scores during the fSVGD posterior inference in (4). This alternative approach produces score estimates 'ad-hoc' and does not require us to train an explicit score network. In contrast, our score network is global function which explicitly takes into account the measurement set and thus can exploit similarity structure across $\mathcal{X}$ which SSGE cannot. Additionally, we face the issue that we cannot simply add inductive bias towards higher entropy as we do through the spectral normalization. Overall, the experiment results in Table 5 and 6 suggest that instantiating our general approach with SSGE performs worse than MARS. We can also observe this visually in Fig. 4a where we plot the score estimate of MARS and SSGE on a GP marginal where the true score is known. While the MARS score network slightly over-estimates the variance of true generative-process marginal, SSGE implicitely under-estimates the prior variance which would lead to over-confident predictions. Finally, Table 7 in Appx. D quantitatively shows that, when compared to a variety of nonparametric estimators, our score networks estimates are the most accurate.

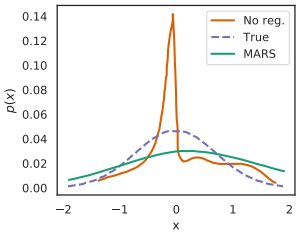
Spectral normalization prevents meta-overfitting. In Sec. 4, we added spectral normalization to our score network. Here, we investigate empirically what happens when we remove the spectral normalization from MARS. Figure 4b and 4c illustrate how our score network over-fits and under-estimates the true data-generating variance when we do not use spectral regularization. Quantitatively, we observe substantial increases in the calibration-errors and a consistent worsening of



(a) SSGE vs. score network



(b) spectral regularization



(c) Corresponding density

Figure 4: (a) Underconfident MARS and overconfident SSGE score predictions on 10 samples from a Gaussian distribution. (b) Score estimates of MARS with and without regularization on 10 function samples from a zero mean GP with SE kernel. (c) Numerically integrated score predictions corresponding to the scores in (b). Overall, SSGE and MARS w/o regularization overfit and underestimate the variance, whereas spectral normalization biases MARS towards higher entropy.

the predictive accuracy. Overall, this highlights the empirical importance of spectral normalization for preventing meta-overfitting and biasing the score estimates towards higher entropy.

Accounting for epistemic uncertainty of the GP interpolators is crucial. In MARS we interpolate each dataset with a GP and, when performing score estimation, use samples from the GP posterior marginals to account for the epistemic uncertainty of the interpolation. Here, we empirically study what happens if we ignore the epistemic uncertainty and just take the each GP's mean predictions. When doing so we observe detrimental effects on the RMSE and calibration error in the majority of the environments in Table 5 and 6. This affirms that propagating the epistemic uncertainty of the interpolation into the score estimates is a critical component of MARS.

D Further experiments

We provide further empirical evidence for the proposed method. In particular, we showcase that:

1. Our parametric score matching approach performs favorably to many nonparametric score estimators.
2. Regularization via spectral normalization does not hinder the flexibility of the network.
3. Sampling from the GP posteriors when training the score network successfully incorporates epistemic uncertainty in areas of the domain where meta-training data is scarce.

D.1 Parametric vs Nonparametric Score Estimation Methods

We start by comparing the performance of MARS to that of several nonparametric score estimators, described in Zhou et al. (2020).

Nonparametric score estimators. There are several theoretically motivated nonparametric score estimation methods with well-understood properties and straightforward implementations. Their simplicity and flexibility make them a popular choice (Coyle et al., 2020; Wang et al., 2019; Deledalle et al., 2014). KEF (Canu & Smola, 2006) performs regularized score matching inside a kernel exponential family; this can further be efficiently approximated by the Nyström method (Sutherland et al., 2018). Liu et al. (2016); Chwialkowski et al. (2016) propose a Stein estimator, and Shi et al. (2018) propose Spectral Stein Gradient Estimator (SSGE) by expanding the score function in terms of the spectral eigenbasis. Together with iterative methods (the ν-method and Landweber iteration (Engl et al., 1996)), these approaches can be naturally unified through a regularized, nonparametric regression framework (Zhou et al., 2020).

MARS vs nonparametric estimators. To compare the performance of MARS to nonparametric methods, we consider score estimation of marginal distributions of Gaussian Processes (GP) (Rasmussen, 2003) and Student's-t processes (TP) (Shah et al., 2014). In both cases, as the stochastic process mean function, we use the sinusoidal mean function $2x + 5sin(2x)$, and for the covariance kernel, we choose the radial basis function kernel, with the lengthscale parameter set to $\ell = 1$. We

	RMSE			
	GP-2D	GP-3D	TP-2D	TP-3D
KEF	8.836 ± 0.000	6.625 ± 0.000	8.852 ± 0.000	13.694 ± 0.000
NKEF	4.322 ± 0.000	4.402 ± 0.000	5.936 ± 0.000	8.402 ± 0.000
nu-method	9.602 ± 0.000	14.833 ± 0.000	8.195 ± 0.000	16.978 ± 0.000
SSGE	1.166 ± 0.000	1.989 ± 0.000	1.753 ± 0.000	4.822 ± 0.000
Stein estimator	0.912 ± 0.000	4.058 ± 0.000	3.519 ± 0.000	7.251 ± 0.000
MARS network	$\mathbf{0.664 \pm 0.179}$	$\mathbf{1.375 \pm 0.215}$	$\mathbf{1.093 \pm 0.241}$	$\mathbf{3.199 \pm 0.418}$

Table 7: RMSE between the true and predicted scores of two/three-dimensional marginal distributions of GP and TP with an RBF kernel and a sinusoidal mean function with a linear trend. MARS score network significantly outperforms all nonparametric score estimators.

	Cosine similarity			
	GP-2D	GP-3D	TP-2D	TP-3D
KEF	0.475 ± 0.000	0.410 ± 0.000	0.532 ± 0.000	0.485 ± 0.000
NKEF	0.556 ± 0.000	0.434 ± 0.000	0.384 ± 0.000	0.318 ± 0.000
nu-method	0.435 ± 0.000	0.206 ± 0.000	0.444 ± 0.000	0.494 ± 0.000
SSGE	$\mathbf{0.943 \pm 0.000}$	0.601 ± 0.000	$\mathbf{0.661 \pm 0.000}$	0.521 ± 0.000
Stein estimator	0.778 ± 0.000	$\mathbf{0.851 \pm 0.000}$	0.626 ± 0.000	0.590 ± 0.000
MARS network	$\mathbf{0.956 \pm 0.110}$	$\mathbf{0.889 \pm 0.048}$	$\mathbf{0.737 \pm 0.097}$	$\mathbf{0.706 \pm 0.061}$

Table 8: Cosine Similarity between the true and predicted scores of two/three-dimensional marginal distributions of GP and TP with an RBF kernel and a sinusoidal mean function with a linear trend. MARS score network performs competitively to nonparametric score estimators.

use Tensorflow Probability's (Dillon et al., 2017) implementation of both GP and TP and set all other parameters to their default values.

We perform four experiments, the first two on estimating the marginal scores of a GP and the last two on the TP. In all experiments, we sample one measurement set $\mathbf{X}$ containing either two ($\mathbf{X} = \{x_1, x_2\}$) or three points ($\mathbf{X} = \{x_1, x_2, x_3\}$) of dimension one, i.e., $x_i \in \mathbb{R}, \forall i$. For all experiments, the x_i follow a uniform distribution: for the GP experiment, $x_i \sim U([-5, 5])$, whereas for the TP experiment, $x_i \sim U([-1, 1]), \forall i$. In the experiments with measurement sets consisting of two points, the score network and the nonparametric estimators are trained on 50 samples from the corresponding *two*-dimensional marginal distributions. For the measurement sets of size 3, we train the score network and nonparametric estimators on 200 samples from the corresponding *three*-dimensional marginal distribution. When evaluating the performance of MARS, we take the average performance over five different seeds after training the network for 2000 iterations in the *two*-dimensional marginal case and 5000 iterations in the *three*-dimensional case. We measure the quality of the score estimates via the RMSE and cosine similarity between the estimates and true scores. The corresponding evaluation results are reported in Tables 7-8.

The RMSE measures deviations of the estimated scores in both direction and magnitude of the gradients. For the SVGD particle estimation, it is more important that the gradients in the vector field point in the correct direction than having the correct magnitude. Thus, we also report the cosine similarity, which only quantifies how well the directions in the estimated vector field match the directions in the true vector field while neglecting errors in the score magnitude. As we can see in Table 7 and 8, MARS consistently outperforms the nonparametric score estimators.

D.2 Flexibility of the regularized network

Recall from section 4 that, in order to prevent overfitting, we perform spectral normalization of the weights of the linear layers by re-parameterizing the weights by $\tilde{\mathbf{W}} := \mathbf{W}/||\mathbf{W}||$. To investigate whether this hinders the flexibility of the network's outputs, we visually examine the network's predictions on the following two tasks, in which we estimate the scores of one-dimensional Laplace and Student's-t distribution.

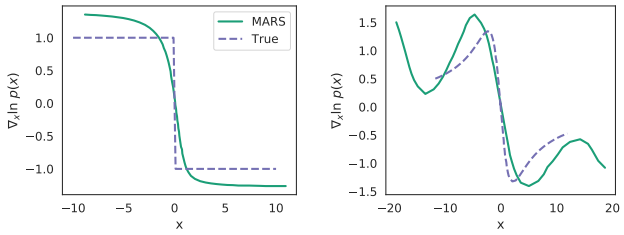


Figure 5: Predictions of the MARS score network, trained on 45 samples from Laplace (left) and Student's-t distribution (right). The network approximates the score functions sufficiently well, showcasing that spectral normalization of linear layers does not hinder its flexibility.

In order to estimate a score $\nabla_x \log p(x)$ for some unknown distribution $p(x)$ (rather than marginal distribution scores $\nabla_{\mathbf{fX}} \ln p(\mathbf{f}^{\mathbf{X}})$ for some unknown stochastic process $p(f)$), we use 45 samples $\mathbf{X} = \{x_1, ..., x_{45}\}$ from: *(i)* one-dimensional Student's-t distribution with $\nu = 5$ degrees of freedom, location parameter $\mu = 0$, and scale parameter $\sigma = 1$, and *(ii)* one-dimensional standard Laplace distribution, i.e., with location parameter $\mu = 0$ and scale parameter $\sigma = 1$.

To perform distributional score estimation, we need to make a slight modification to the original score network architecture; rather than concatenating input-output pairs (i.e., concatenating x_i and $f(x_i)$, as shown in Fig. 3), we use only the distributional samples as inputs. To be more precise, at every iteration, we randomly select (without replacement) $k = 8$ inputs $\{x'_1, ..., x'_k\} \subset \mathbf{X}$, which we input to the score network. We train the network using the regularized score matching loss, i.e., perform spectral normalization of the layers. The network is trained for 2000 iterations, using a learning rate of 0.001. For this experiment, we set the attention embedding dimension to 32, key size to 16, and use 8 attention heads. The remaining hyperparameters are set according to Sec. A.3. The respective predictions are displayed in Figure 5. We observe that the score network approximates the score functions well and that the regularization does not hinder flexibility too much.

D.3 Incorporating uncertainty through GP interpolation

In addition to the ablation study performed in Section C, we visually investigate the implications of sampling functions from the GP posteriors during score matching. In particular, we do so using the sinusoid environment, detailed in Appendix B.1.

Experiment setup. In this experiment, we sample 10 functions from the environment and evaluate each function at five randomly selected inputs in the $[-5, -2] \cup [2, 5]$ range. We investigate whether the proposed GP interpolation method promotes uncertainty in the $[-2, 2]$ part of the domain.

In order to do so, let us recall Algorithm 1, in which we learn the trained score network $\mathbf{s}_\phi$. The algorithm first fits a GP to each of the n datasets $\mathcal{D}_1, ..., \mathcal{D}_n$. Then, at every step, the algorithm samples a measurement set $\mathbf{X} \overset{iid}{\sim} \nu$, and then, for every dataset $\mathcal{D}_i$ (i.e., for every collection of input/output pairs $\mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}}$), the algorithm samples function values from the GP posterior marginal: $\tilde{\mathbf{f}}_i^{\mathbf{X}} \sim p(\mathbf{f}_i^{\mathbf{X}} | \mathbf{X}, \mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})$. We compare this approach to training the score network $\mathbf{s}_\phi$ on the mean predictions of the posterior marginal $p(\mathbf{f}_i^{\mathbf{X}} | \mathbf{X}, \mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})$, i.e., where $\tilde{\mathbf{f}}_i^{\mathbf{X}} = \mu_{p(\mathbf{f}_i^{\mathbf{X}} | \mathbf{X}, \mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})}$. In order to distinguish between the two approaches, we denote the score network trained on the posterior marginal mean predictions with $\mathbf{s}_\phi^\mu$, and use $\tilde{\mathbf{f}}_{\mu\,i}^{\mathbf{X}} := \mu_{p(\mathbf{f}_i^{\mathbf{X}} | \mathbf{X}, \mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})}$ as a shorthand notation for the mean predictions of the posterior marginals. Observe that, for measurement sets $\mathbf{X}$ close to $\mathbf{X}_i^{\mathcal{D}}$, the samples $\tilde{\mathbf{f}}_i^{\mathbf{X}}$ and $\tilde{\mathbf{f}}_{\mu\,i}^{\mathbf{X}}$ will be very similar, whereas when $\mathbf{X}$ is far from $\mathbf{X}_i^{\mathcal{D}}$, the variability of the samples $\tilde{\mathbf{f}}_i^{\mathbf{X}}$ will be much larger. In the following, we showcase that this variability successfully incorporates uncertainty about the areas of the domain $\mathcal{X}$ with little or no data.

In order to compare the two approaches, we visually compare the overall framework when the fSVGD network is trained with $\mathbf{s}_\phi^\mu$ (the score network trained on $\tilde{\mathbf{f}}_{\mu\,i}^{\mathbf{X}}$, the GP mean predictions), and when it is trained with $\mathbf{s}_\phi$ (the score network trained on $\tilde{\mathbf{f}}_i^{\mathbf{X}}$, the samples from the GP posterior marginal), as performed in the original MARS algorithm.

Figure 6: MARS prediction on samples from the sinusoid environment, with no data in the $[-2, 2]$ range. Left: posterior predictions of two randomly selected GPs fitted to the meta-tasks. Middle: fBNN predictions, fitted using $\mathbf{s}_\phi^\mu$, the score estimation network trained on GP mean predictions. Right: fBNN predictions, fitted using $\mathbf{s}_\phi$, the score estimation network trained on samples from the GPs. Training fBNN with $\mathbf{s}_\phi$ (by sampling from the GPs) successfully incorporates uncertainty about the $[-2, 2]$ part of the input domain.

Empirical findings. We plot the corresponding fSVGD-BNN (fBNN) predictions, trained using the two score networks $\mathbf{s}_\phi^\mu$ and $\mathbf{s}_\phi$. Both fBNNs are fitted to four points, where inputs $x_1, ..., x_4$ are sampled uniformly from $[-5, 5]$ (i.e., the whole input domain), and their functional evaluations are obtained according to the sinusoid environment.

The results are visualized in Figure 6. The first two plots on the left in Figure 6 correspond to posterior predictions of two randomly selected GPs fitted to the meta-tasks, where no task contains information in the $[-2, 2]$ input range. The middle plot corresponds to the fBNN network predictions, where the network was trained using the fSVGD algorithm and the score network $\mathbf{s}_\phi^\mu$. The last plot corresponds to the fBNN network predictions, where fBNN is trained using the fSVGD algorithm and the score network $\mathbf{s}_\phi$. We observe a clear difference between the two approaches: MARS (trained using $\mathbf{s}_\phi$) successfully incorporates the epistemic uncertainty in the $[-2, 2]$ part of the input domain into the fBNN posterior, yielding less confident predictions in the area where no data was available during meta-training. In contrast, when we use GP posterior means instead of samples when fitting the score network, the resulting BNN predictions entirely ignore the epistemic uncertainty that arises due to the fact that we don't know the function values in $[-2, 2]$. This may lead to overconfident posterior predictions.