MARS: META-LEARNING AS SCORE MATCHING IN THE FUNCTION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Meta-learning aims to extract useful inductive biases from a set of related datasets. In Bayesian meta-learning, this is typically achieved by constructing a prior distribution over neural network parameters. However, specifying families of computationally viable prior distributions over the high-dimensional neural network parameters is difficult. As a result, existing approaches resort to meta-learning restrictive diagonal Gaussian priors, severely limiting their expressiveness and performance. To circumvent these issues, 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 instead of parameter space priors. 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

Using data from related tasks is of key importance for sample efficiency. Meta-learning attempts to extract prior knowledge (i.e., inductive bias) about the unknown data generation process from these related tasks and embed it into the learner, so that it generalizes better to new learning tasks (Thrun & Pratt, 1998; Vanschoren, 2018). Many meta-learning approaches try to amortize or re-learn the entire inference process (e.g., Santoro et al., 2016; Mishra et al., 2018; Garnelo et al., 2018) or significant parts of it (e.g., Finn et al., 2017; Yoon et al., 2018). As a result, they require large amounts of meta-training data and are prone to meta-overfitting (Qin et al., 2018; Rothfuss et al., 2021a).

The Bayesian framework provides a sound and statistically optimal method for inference, by combining prior knowledge about the data-generating process with new empirical evidence in form of a dataset. In this work, we adopt the Bayesian framework for inference at the task level and only focus on meta-learning informative Bayesian priors. Previous approaches (Amit & Meir, 2018; Rothfuss et al., 2021a) meta-learn Bayesian Neural Network (BNN) prior distributions from a set of related datasets. By meta-learning the prior distribution and applying regularization at the meta-level, they facilitate positive transfer from only a handful of meta-training tasks. However, BNNs lack a parametric family of (meta-)learnable priors over the high-dimensional space of neural network (NN) parameters that is both computationally viable and, at the same time, flexible enough to account for the over-parametrization of NNs. In practice, both approaches use a Gaussian family of priors with diagonal covariance matrix, which is too restrictive to accurately match the complex probabilistic structure of the data-generating process.

To address these shortcomings, we take a new approach to formulating the meta-learning problem and represent prior knowledge in a novel way. We build on recent advances in functional approximate inference for BNNs that perform Bayesian inference in the function space rather than in the parameter space of neural networks (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 source of prior knowledge when performing approximate functional BNN inference on a new target task. This ultimately 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, called *Meta-learning via Attention-based Regularised Score estimation (MARS)*, consistently outperforms previous meta-learners in predictive accuracy and also yields significant improvements in the quality of uncertainty estimates. Notably, MARS enables positive transfer from only a handful of tasks, while maintaining reliable 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 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., 2019b; 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, the 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 form the lack of expressive families of priors for the high-dimensional and complex parameters 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 represented by a NN, instead of a prior distribution. Also related to our stochastic process approach are methods that who 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 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.

3 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 $y_j = f(\mathbf{x}_j) + \epsilon_j$ 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 function parametrized by a NN with weights $\theta \in \Theta$. For regression, where $\mathcal{Y} \subseteq \mathbb{R}$, we can use h_θ to define a conditional distribution over the noisy observations $p(y|\mathbf{x}, \theta) = \mathcal{N}(y|h_\theta(\mathbf{x}), \sigma^2)$, with flexible mean represented by h_θ and observation noise variance σ^2. Given a prior distribution $p(\theta)$ over the model parameters θ, Bayes' theorem yields a posterior distribution $p(\theta|\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) \propto p(\mathbf{y}^{\mathcal{D}}|\mathbf{X}^{\mathcal{D}}, \theta)p(\theta)$, where $p(\mathbf{y}^{\mathcal{D}}|\mathbf{X}^{\mathcal{D}}, \theta) = \prod_{j=1}^m p(y_j|\mathbf{x}_j, \theta)$. For an unseen test point $\mathbf{x}^*$, we compute the predictive distribution, which is defined as $p(y^*|\mathbf{x}^*, \mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) = \int p(y^*|\mathbf{x}^*, \theta)p(\theta|\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}})d\theta$ and obtained by marginalizing out the posterior over parameters θ.

BNN inference in the function space. Posterior inference for BNNs is difficult due to the high-dimensional parameter space Θ and the over-parameterized nature of the NN mapping $h_\theta(x)$. An alternative approach views BNN inference in the function space, i.e., the space of regression functions $h : \mathcal{X} \mapsto \mathcal{Y}$, yielding the Bayes rule $p(h|\mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) \propto p(\mathbf{y}^{\mathcal{D}}|\mathbf{X}^{\mathcal{D}}, h)p(h)$ (Wang et al., 2019; Sun et al., 2019). Here, $p(h)$ is a stochastic process prior with index set $\mathcal{X}$, taking values in $\mathcal{Y}$.

Stochastic processes can be understood as random functions or infinite-dimensional random vectors, and thus, are inherently hard to work with computationally. However, given finite measurement sets $\mathbf{X} := [\mathbf{x}_1, ..., \mathbf{x}_k] \in \mathcal{X}^k, k \in \mathbb{N}$, the stochastic process can by characterized by its corresponding marginal distributions of function values $\rho(\mathbf{h}^{\mathbf{X}}) := \rho(h(\mathbf{x}_1), ...h(\mathbf{x}_k))$ (cf. Kolmogorov Extension Theorem, Oksendal (2013)). Correspondingly, we can break down the functional posterior into a more tractable form by re-phrasing it in terms of posterior marginals based on measurement sets $\mathbf{X}$: $p(\mathbf{h}^{\mathbf{X}}|\mathbf{X}, \mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) \propto p(\mathbf{y}^{\mathcal{D}}|\mathbf{h}^{\mathbf{X}^{\mathcal{D}}})p(\mathbf{h}^{\mathbf{X}})$.

This functional posterior can be tractably approximated to achieve functional BNN inference (Sun et al., 2019; Wang et al., 2019). We briefly describe how this can be done via functional Stein Variational Gradient Descent (fSVGD, Wang et al., 2019). The procedure is explained in more detail in Appendix A.1. The fSVGD algorithm approximates the posterior using a set of L NN parameter particles $\{\theta_1, ..., \theta_L\}$. To optimize posterior approximation, fSVGD iteratively re-samples measurement sets $\mathbf{X}$ from a measurement distribution ν, supported on $\mathcal{X}$, computes SVGD updates (Liu & Wang, 2016) in the function space and projects them back into the parameter space, in order to update the parameter-space particle configuration approximating the posterior. To achieve this, fSVGD uses the functional posterior score, i.e., the gradient of the log-density $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}}|\mathbf{X}, \mathbf{X}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}) = \nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{y}^{\mathcal{D}}|\mathbf{h}^{\mathbf{X}^{\mathcal{D}}}) + \nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$ to guide the particles towards areas of high posterior probability. Formally, for all $l = 1, ..., L$ the particle updates are computed as

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

where $K = [k(\mathbf{h}^{\mathbf{X}}_{\theta^l}, \mathbf{h}^{\mathbf{X}}_{\theta^i})]_{li}$ is the kernel matrix between the function values in the measurement points based on a positive semi-definite kernel function $k(\cdot, \cdot)$. A key insight which we will later draw upon in the paper is that such functional approximate inference techniques only make use of the prior scores and, in principle, do not require a full prior density which integrates to 1.

4 Meta-Learning as Score Estimation in the Function Space

4.1 Problem Statement: Meta-Learning

Meta-learning extracts prior knowledge (i.e., inductive bias) from a set of related learning tasks to accelerate inference on a new and unseen learning task. 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. Following previous work (e.g., Baxter, 2000; Pentina & Lampert, 2014; Rothfuss et al., 2021a), 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 predicitons that generalize better.

4.2 Shortcomings of Meta-Learning Priors in the Parameter Space

Previous work phrases meta-learning as hierarchical Bayesian problems and tries to meta-learn a prior distribution $p(\theta)$ on the NN weight space that resembles $p(f)$ (e.g., Amit & Meir, 2018; Rothfuss et al., 2021a). This approach suffers from the following central issues: Due to the over-parametrized nature of neural networks, there exists a large set of parameters θ that correspond to exactly the same mapping $h_\theta(\cdot)$. This makes specifying a good prior distribution $p(\theta)$ typically very difficult. At the same time, there is a lack of sufficiently rich parametric families of distributions over the high-dimensional parameter space Θ that are also computationally viable. So far, only the very restrictive Gaussian family of priors with diagonal covariance matrices has been employed. As a result, the meta-learned posterior lacks the necessary flexibility to accurately match the complex probabilistic structure of the data-generative process.

4.3 META-LEARNING AS SCORE ESTIMATION ON THE DATA-GENERATING PROCESS

Aiming to address these issues, we acquire prior knowledge in a data-driven manner with a new perspective. We develop a novel a approach to meta-learning which hinges upon three key ideas:

First, we view BNN inference in the function space (see Sec. 3), 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 on the function space. This alleviates the aforementioned problem of over-parametrization which arises when considering priors on the parameter space.

Second, we ask ourselves what is a desirable prior: 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 would want to meta-learn a prior that matches $p(f)$ as closely as possible. Since the meta-training datasets $\mathcal{D}_1, ..., \mathcal{D}_n$ constitute noisy observations of the function draws $f_1, ..., f_n \sim p(f)$, we can use them to estimate the stochastic process marginals $p(\mathbf{f}^{\mathbf{X}})$. Crucially, we tractably represent and estimate the stochastic process $p(f)$ through its finite marginals $p(\mathbf{f}^{\mathbf{X}})$ in measurement sets $\mathbf{X}$.

Our third insight is that all popular approximate inference methods for BNNs only use the prior score, i.e., the gradient of the log-prior, and not the prior distribution itself. In addition, parametrizing and estimating the prior score is computationally easier than the prior distribution itself, since it does not have to integrate to 1, allowing for much more flexible neural network representations. For this reason, our meta-learning approach directly forms an estimate of the scores $s(\mathbf{f}^{\mathbf{X}}, \mathbf{X}) = \nabla_{\mathbf{f}^{\mathbf{X}}} \ln p(\mathbf{f}^{\mathbf{X}})$ of the data-generating process marginals.

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 data-generating stochastic process from 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. In the following section, we discuss in more details how to implement this general approach as a concrete, computationally feasible algorithm with exquisite empirical performance.

5 THE MARS META-LEARNING ALGORITHM

In Section 4, we have motivated and sketched the idea of meta-learning via score estimation of the data-generating process marginals. We now discuss particular challenges and outline the design choices for translating this idea into a practical meta-learning algorithm.

Score estimation of stochastic process marginals has previously been used in the context of existing functional BNN approaches (Wang et al., 2019; Sun et al., 2019). However, they presume a known stochastic process prior with oracle access to arbitrarily many sampled functions that can be evaluated in arbitrary locations of the domain $\mathcal{X}$. Compared to such a stochastic process prior with oracle access, we face two key challenges which make the score estimation problem much harder:

1. We only have access to a finite number of datasets $\{\mathcal{D}_i\}_{i=1}^n$, each corresponding to one function f_i drawn from the data-generating stochastic process. This means, we have so estimate the marginal scores $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$ from only n function samples without over-fitting.
2. Per function f_i, we only have a limited number of m_i noisy function evaluations $\mathbf{y}_i^{\mathcal{D}}$ in input locations $\mathbf{X}_i^{\mathcal{D}}$ which are given to us exogeneously and we have no control over them. However, to perform score estimation for the marginal $p(\mathbf{h}^{\mathbf{X}})$ distributions, for each function, we need a vector of function values $\mathbf{f}_i^{\mathbf{X}}$ evaluated in the measurements points $\mathbf{X}$.

In the following subsections, we present our approach to estimating the stochastic prior scores and discuss how to solve these challenges. In addition, we discuss how to perform functional approximate BNN inference with the meta-learned score estimates. Finally, we compare our approach to several state-of-the-art nonpartametric score estimators and provide the findings in Appendix D.1.

5.1 PARAMETRIC SCORE MATCHING FOR STOCHASTIC PROCESS MARGINALS

Performing functional approximate posterior inference with fSVG updates as in (1) requires estimation of the prior marginal scores $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$ for arbitrary measurement sets $\mathbf{X}$. A key property of stochastic process marginal scores is their permutation equivariance: if we permute the order of the

points in the measurement set, the rows of the score permute in the same manner. Similarly, the measurement sets can be of different sizes, implying score vectors/matrices of different dimensionalities.

Aiming to embed these properties into a parametric model of the prior marginal scores, we use a transformer encoder architecture (Vaswani et al., 2017), that takes as an input a measurement set $\mathbf{X} \in \mathbb{R}^{\dim(\mathcal{X}) \times k}$ of k points and corresponding query function values $\mathbf{h}^{\mathbf{X}} \in \mathbb{R}^{\dim(\mathcal{Y}) \times k}$ and 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 and outputs an estimate of the score, i.e., $\mathbf{s}_\phi(\mathbf{h}^{\mathbf{X}}, \mathbf{X}) \mapsto \tilde{\nabla}_{\mathbf{h}^{\mathbf{X}}} p(\mathbf{h}^{\mathbf{X}})$, a matrix of the same size as the query function values. The model is permutation equivariant w.r.t. the columns that correspond to measurement points and supports inputs / outputs for varying measurement set sizes k. We schematically illustrate the score network architecture in Figure 4 and provide details in Appendix A.3.

We train the score network with a modified version of the score matching loss (Hyvärinen & Dayan, 2005; Song et al., 2020). This loss minimizes the Fisher divergence between the data-generating process marginals and our score network, 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 stochastic process marginals on their measurements sets into account and uses an expectation over randomly sampled measurement sets, similar to the functional approximate inference in Section 3.

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

As noted previously, our meta-training data only provides us with a noisy function evaluations for a limited, exogenously given, set of input locations per task f_i. The input locations $\mathbf{X}_i^{\mathcal{D}}$ per task might even differ in number and be non-overlapping. Thus, we do not have access to the function values $\mathbf{f}_1^{\mathbf{X}}, ..., \mathbf{f}_n^{\mathbf{X}}$ for arbitrary measurement sets $\mathbf{X}$ which we require for our score matching loss in (2). In the following section, we discuss how to resolve this problem.

5.2 Interpolating the Datasets across $\mathcal{X}$

In our meta-learning setup (see Section 4.1), 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 (2), we need the corresponding function values in arbitrary measurement locations in the domain $\mathcal{X}$. While a range of interpolation techniques could be applied, we choose to 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 $\mathbf{X}$. For the GPs, we employ a $5/2$-Matern kernel and choose its lengthscale with 4-fold cross-validation based the log marginal likelihood, averaged across the n tasks.

While we could simply use the GP posterior mean values for computing the score matching loss in 2, this would not reflect the epistemic uncertainty that stems from having only a finite subset of observations at only $\mathbf{X}_i^{\mathcal{D}} \subset \mathcal{X}$ and interpolating them with a GP. Instead, 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 (2). Correspondingly, the score matching loss modifies to

$$\tilde{\mathcal{L}}(\phi) := \mathbb{E}_{\mathbf{X}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p(\tilde{\mathbf{f}}_i^{\mathbf{X}} | \mathbf{X}, \mathbf{X}_i^{\mathcal{D}}, \mathbf{y}_i^{\mathcal{D}})} \left[\mathrm{tr}(\nabla_{\tilde{\mathbf{f}}_i^{\mathbf{X}}} \mathbf{s}_\phi(\tilde{\mathbf{f}}_i^{\mathbf{X}}, \mathbf{X})) + \frac{1}{2} ||\mathbf{s}_\phi(\tilde{\mathbf{f}}_i^{\mathbf{X}}, \mathbf{X})||_2^2 \right] \right] . \tag{3}$$

The further a measurement point x is from the closest point in $\mathbf{X}_i^{\mathcal{D}}$, the higher is the GP's (epistemic) uncertainty of the corresponding function value $f_i(x)$. By sampling from the GP posterior, in expectation, the loss in (3) effectively propagates the uncertainty into the score matching procedure, preventing over-confident score estimates for areas of the $\mathcal{X}$ with scarce data, as illustrated in Figure 6.

5.3 Preventing Meta-Overfitting of the Prior Score Network

The final challenge we need to address is over-fitting to the meta-training tasks (Yin et al., 2020; Rothfuss et al., 2021a). As 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 $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$, making us prone to overfitting the prior score network. If we performed functional BNN inference with such an overfitted score estimate, the BNN's posterior predictions are likely to be over-confident and too biased towards the functions seen during meta-training.

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 $\mathbf{s}_\phi$

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. Spectral normalization controls the Lipschitz constant of the neural network layers by dividing their weight matrix $\mathbf{W}$ by its spectral norm $||\mathbf{W}||$, i.e., re-parametrizing the weights as $\tilde{\mathbf{W}} := \mathbf{W}/||\mathbf{W}||$. Hence, by applying spectral normalization, we bias our score network towards smoother score estimates that correspond to marginal distributions $p(\mathbf{h}^{\mathbf{X}})$ with higher entropy. Empirically we find that spectral normalization effectively combats meta-overfitting and prevents the estimated prior score from inducing over-confident posterior predictions when employed in functional BNNs. In our experiments, we also tried other score regularization methods like gradient penalties (Gulrajani et al., 2017) and adding the sum of spectral norms of each layers' weight matrix as regularization penalty to the loss function, but found spectral normalization to work the best across different environments.

5.4 The full MARS algorithm

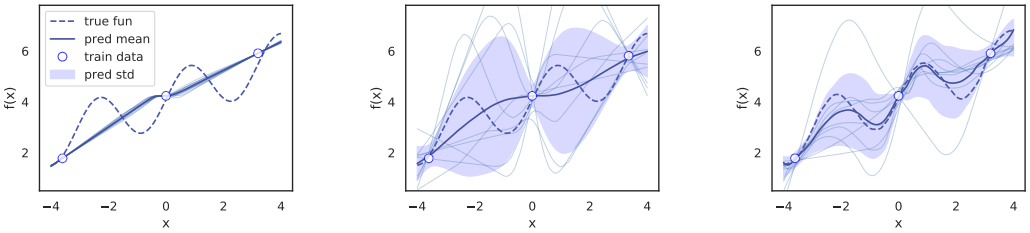
We now summarize our meta-learning algorithm MARS. It consists of two stages:

Stage 1: Meta-Learning the Prior Score Network. After initializing the parameters of the 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 (3). In each iteration, we first sample a measurement set $\mathbf{X}$ i.i.d. from the measurement distribution $\nu = \mathcal{U}(\tilde{\mathcal{X}})$, chosen as uniform distribution over the hypercube $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ conservatively covering the data in $\mathcal{X}$. Based on 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 the score network parameters 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 can infuse the meta-learned inductive bias into the BNN inference by using the prior score network $\mathbf{s}_\phi$ from Stage 1 for the approximate inference. In particular, we can either perform functional VI (Sun et al., 2019) or fSVGD (Wang et al., 2019), using the score network $\mathbf{s}_\phi(\mathbf{h}^{\mathbf{X}}, \mathbf{X})$ predictions as swap-in replacement for the marginal scores $\nabla_{\mathbf{h}^{\mathbf{X}}} \ln p(\mathbf{h}^{\mathbf{X}})$ of a user-specified stochastic process prior. In our experiments, we use fSVGD (see Section 3) as functional approximate inference method. The resulting fSVGD BNN inference procedure with our meta-learned priors score network is summarized in Alg. 2 in Appx. A.

6 Experiments

We provide a detailed benchmark comparison with existing meta-learning, demonstrating that MARS: *(i)* achieves state-of-the-art performance in terms of predictive accuracy, *(ii)* yields well-calibrated uncertainty estimates. In a comprehensive ablation study, we shed more light on the key components of MARS and provide empirical support for our design decisions.



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

Figure 1: BNN posterior predictions with (a) zero-centered Gaussian prior on the NN parameters θ (b) Gaussian process prior and (c) meta-learned MARS prior scores.

	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. Reported are the mean and standard deviation across 5 seeds. MARS consistently yields the most accurate predictions.

6.1 Experiment Setup

Environments. We consider five realistic meta-learning *regression* environments. The first environment corresponds to data of different calibration sessions of the Swiss Free Electron Laser (SwissFEL) (Milne et al., 2017). Here, a task requires predicting the laser's beam intensity based on parameters of its undulators. We also consider data from the *PhysioNet* 2012 challenge, consisting of electronic health measurements from intensive care patients (Silva et al., 2012), in particular the Glasgow Coma Scale (*GCS*) and the hematocrit value (*HCT*). Here, the different tasks correspond to different patients. In addition, we use the Intel Berkeley Lab temperature sensor dataset (*Berkeley-Sensor*) (Madden, 2004), requiring auto-regressive prediction of temperature measurements corresponding to sensors in different locations of the building. Finally, the *Argus-Control* environment requires predicting of the total variation of a robot from its target position based on its PID controller parameters (Rothfuss et al., 2022). Here, different tasks correspond to different step sizes between source and target location. Detailed information regarding the environments can be found in Appendix B.

Baselines. As non-meta-learning baselines, we use a *Vanilla GP* with RBF kernel and a *Vanilla BNN* with a zero-centered, spherical Gaussian prior and SVGD posterior inference (Liu & Wang, 2016). To facilitate a performance comparison to previous meta-learning methods, we report results for model agnostic meta-learning (*MAML*) (Finn et al., 2017), Bayesian MAML (*BMAML*) (Yoon et al., 2018) and neural processes (*NPs*) (Garnelo et al., 2018). In addition, we include two approaches that perform hierarchical Bayesian inference to meta-learn a prior distribution over NN parameters, in particular, *MLAP* (Amit & Meir, 2018) and *PACOH-NN* (Rothfuss et al., 2021a).

	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. Reported are the mean and standard deviation across 5 seeds. MARS provides the best-calibrated uncertainty estimates.

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 3: 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 4: Ablation study results for MARS components in terms of the calibration error.

6.2 Empirical Benchmark Study

Qualitative illustration. Fig. 1 illustrates the 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) and c) a MARS-BNN trained with the meta-learned stochastic prior score based. For the meta-training we use $n = 20$ tasks with each $m = 8$ datapoints, generated from a student-t process sinusoidal mean function $2x + 5\sin(2x)$. The target task is also sampled process. Similar to previous work (e.g., Wenzel et al., 2020; Fortuin et al., 2022) we observe that weight-space BNN priors provide poor inductive bias and yield grossly over-confident uncertainty estimates, whereas approaching the BNN inference problem in the function space (see Fig 1b) results in much better uncertainty estimates. In this case however, the uninformative GP prior does not result in accurate mean predictions. Finally, the meta-learned scores in the function space provide useful inductive biased towards the linear+sinusoidal data-generating pattern, while yielding tight, but reliable confidence intervals.

MARS provides accurate predictions. We perform a comprehensive benchmark study with the environments and baselines introduced in Sec. 6.1. Table 1 reports the root mean squared error (RMSE) on unseen meta-test tasks. MARS yields substantial improvements of the RMSE when compared to the non-meta-learning baselines and significantly outperforms all other meta-learning baselines in the majority of the environments. This shows that MARS is able to reliably acquire useful inductive biases from data. We hypothesize that MARS performs substantially better than PACOH-NN and MLAP which meta-learn restrictive diagonal Gaussian priors over Θ, since it has much more flexibility in expressing prior knowledge the about data-generating process.

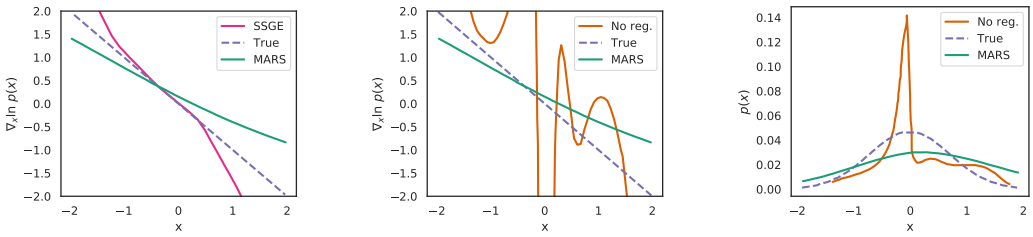
MARS yields well-calibrated uncertainty estimates. We hypothesize that by performing meta-learning in the function space, we avoid the pitfalls from NN over-parametrization which often lead to over-confident uncertainty estimates. To investigate the quality of uncertainty estimates, we compute the *calibration 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[1], and the metric is detailed Appendix C.2. MARS consistently yields the best-calibrated uncertainty estimates, and for some environments reduces the calibration error by more than 30 % when compared to the next best method. This provides further evidence for the soundness and efficacy of our function space approach to meta-learning. Finally, the good calibration performance suggests that the spectral regularization, together with the GP uncertainty sampling in Sec. 5 effectively prevent meta-overfitting and account for epistemic uncertainty.

6.3 Ablation Study

In this section, we empirically investigate the algorithm components that were introduced in Sec. 5 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. 5.1, 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 3 and 4 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:

[1]Note that MAML is not included since it does not provide uncertainty estimates.



(a) SSGE vs. score network (b) spectral regularization (c) Corresponding density

Figure 2: (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.

Parametric score matching outperforms nonparametric score estimation. In Sec. 5.1 we have introduced a parametric score network which we train with the score matching loss in 2. 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 (1). 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 3 and 4 suggest that instantiating our general approach with SSGE performs worse than MARS. We can also observe this visually in Fig. 2a 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. 5.3, we added spectral normalization to our score network. Here, we investigate empirically what happens when we remove the spectral normalization from MARS. Figure 2b and 2c 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 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 3 and 4. This affirms that propagating the epistemic uncertainty of the interpolation into the score estimates is a critical component of MARS.

7 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 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 of 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 opens up many potential extensions towards interactive machine learning where exploration based on epistemic uncertainty is key.

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.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022.

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.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

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`.

Yunxiao Qin, Weiguo Zhang, Chenxu Zhao, Zezheng Wang, Hailin Shi, Guojun Qi, Jingping Shi, and Zhen Lei. Rethink and redesign meta learning. *arXiv preprint arXiv:1812.04955*, 2018.

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, Fabio Ferreira, Simon Boehm, Simon Walther, Maxim Ulrich, Tamim Asfour, and Andreas Krause. Noise Regularization for Conditional Density Estimation. *arXiv preprint arXiv:1907.08982*, 2019a.

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

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.

Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

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.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2020.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. In *International Conference on Learning Representations*, 2020.

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.